A Theory of Active Learning in Dynamic Environments

Andrew Wagenmaker

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2024

Reading Committee:

Kevin Jamieson, Chair

Simon S. Du

Maryam Fazel

Zaid Harchaoui

Program Authorized to Offer Degree:

Computer Science & Engineering

University of Washington

Abstract

A Theory of Active Learning in Dynamic Environments

Andrew Wagenmaker

Chair of the Supervisory Committee:

Kevin Jamieson

Computer Science & Engineering

How should an agent interact with an unknown, dynamic environment in order to collect the information necessary to accomplish its goals? This question is central to the operation of any autonomous system, and answering it effectively is a prerequisite to the development of safe, efficient, and flexible algorithmic agents. While existing work has made progress in addressing this question, it is typically either limited to simple settings, or lacks guarantees strong enough to establish what the most effective approaches are.

This thesis takes steps towards providing a rigorous theoretical answer to this question. In particular, we focus on the setting of dynamic, long-horizon environments, where the actions the agent takes influence not only their current observation, but also future observations. We are interested in eliciting principles of exploration in such settings: *what* aspects of the environment should an agent explore, and *how* can they best explore such aspects? We aim to develop approaches that are *active*—that provide insights into how an agent can interact

with the environment, choosing their actions, to learn as quickly as possible.

We first seek to answer these questions in the setting of continuous control. Here we provide novel algorithmic approaches that establish the instance-optimal rate for both system identification—learning the system's parameters—as well as a general formulation of "decision making"—where the agent wants to learn some decision (e.g. a controller) to minimize some loss. Our results apply to both linear dynamical systems as well as certain classes of nonlinear dynamical systems. To the best of our knowledge, these are the first results to establish the instance-optimal rates for active learning in continuous control.

We then turn to the reinforcement learning (RL) setting. Here our focus is on developing instance-dependent bounds—bounds that adapt to the difficulty of any given instances—and using these bounds to understand what approaches to exploration are most effective in reinforcement learning. We establish novel complexity measures for both tabular and linear RL, which we show can be significantly tighter than existing bounds, and show that existing algorithmic principles are insufficient for achieving the optimal rates, and could be arbitrarily suboptimal.

Finally, we seek to understand not only how we can optimally collect the information necessary accomplish our goals, but how we can efficiently *learn* to collect this information. We argue that answering this question is fundamental for achieving effective, practical approaches. In a general function approximation setting, we establish novel measures of complexity which quantify this cost of learning to explore, and algorithmic approaches which we show achieve the optimal rate. Our results yield exponential improvements over existing bounds in many cases, and provide a nearly complete theory of finite-time instance-optimal learning.

TABLE OF CONTENTS

ACKNOWLEDGMENTS

First and foremost, I would like to thank my advisor Kevin Jamieson for his guidance, insights, and encouragement throughout my PhD. Entering my PhD I had little knowledge of active learning—Kevin's contagious excitement for it, and excitement for research in general, was the source of inspiration for much of this thesis. From the start, Kevin struck a perfect balance between giving me the freedom to pursue the problems I was interested in, while at the same time pushing me to work on really meaningful problems; indeed, his knack for understanding which problems were meaningful has guided much of my thinking over the last several years and greatly shaped this thesis. My perspective and development as a researcher have been deeply influenced by Kevin's mentorship, and I would not be the researcher I am today without his guidance.

The second person I would like to thank is Max Simchowitz. I began collaborating with Max during the second year of my PhD and continued this collaboration for approximately two years, a collaboration that led to four of the chapters in this thesis. I learned so much from working with Max: from incredibly low-level details like how to most effectively use LaTeX macros, to high-level things like how to think about framing my research direction, and everything in-between. His patient guidance and constant insights were instrumental in helping me develop as a researcher, and I am very grateful for all he invested in my work.

I would also like to thank Dylan Foster for hosting me as an intern at MSR, and for letting me in on the DEC work, the product of which is the final chapter of this thesis. Working with Dylan greatly broadened my perspective on learning theory in general, and I have deeply appreciated his guidance and advice both over my internship and beyond.

Towards the end of my PhD, I became increasingly interested in moving beyond pure theory and making things work in practice. Guanya Shi and Abhishek Gupta were instrumental in helping me make this transition. Their perspectives on what real-world problems are important to work on, and how theory might be relevant in such problems, were invaluable to me. While the majority of the work done in collaboration with them did not make it into this thesis, I am very grateful for their mentorship, and have learned so much from my conversations with them.

I have had the opportunity to collaborate with a number of other people during my PhD, and would especially like to thank Simon Du, Aldo Pacchiano, Jamie Morgenstern, Julian Katz-Samuels, and Lalit Jain. In particular, while I've only written two papers with Lalit, his friendship, perspective, and guidance have been a constant source of encouragement throughout my PhD. I would also like to thank my committee—Simon, Maryam Fazel,

Zaid Harchaoui, and Nathan Kutz—for their guidance through this process. My academic siblings—Jennifer, Romain, Yifang, Artin, and Arnab—have been good friends as well as excellent collaborators, and I am thankful to have done my PhD with them.

I would not have made it to where I am today had it not been for the influence of my undergrad advisors, Necmiye Ozay and Raj Nadakuditi, as well as Brian Moore, one of Raj's PhD students who I worked closely with. In particular, I want to thank Necmiye for taking me on us an undergrad with little relevant experience, and guiding me through writing my first paper. Raj's influence was instrumental in leading me to work on machine learning—his course EECS 453 at Michigan, and his excitement for the material, was what first got me interested in machine learning. Furthermore, his and Brian's guidance as research mentors was deeply influential in helping me mature as an independent researcher.

I had the good fortune of making a number of friends outside of the research community over the last six years. In particular, my time in Seattle was shaped to an inordinate degree by the house I lived in, fondly dubbed Ravennesburg Haus, and the people I shared it with: Tom & Huong, Nick, Charles, Patrick, Milo & Jenelle, David, Laura & Aaron, John, Will, Monica, Kevin, Jon, Clayton, Nathan, and Paul. I cannot count the number of good conversations, bonfires, beers (mostly thanks to Milo's brewing prowess), and game nights we had over the last six years, but it was a constant source of joy and stress relief. Within my department, I have greatly appreciated the companionship of Pascal, Romain, and Oscar; I have particularly fond memories of the porch Dominion games during Covid, which were a much-needed source of happiness in a dark time. I would also like to thank my church community at Harbor Anglican Church, especially Fr. Casey for his constant wisdom and guidance in life, Kacy & Troy, Matt, Viktor, Todd, Kelly, the Bettis and Andersen families, and the many other friends I made there. I am also extremely thankful for my longtime friends Jonathan, Austin, and Chris; for their continued friendship and support all these years.

I would not be the person I am today were it not for my family. Especially my parents Trevor & Sandy: they have invested more into my education than anyone; this entire journey would never have started if it was not for the love and support they have given me throughout my life. I am constantly grateful for their encouragement while I pursued a PhD on the other side of the country, and for raising me up in the way I should go. I am also extremely thankful to my siblings—Erin, Alyssa, Jason, and Noelle, as well as Jonathan and Allen—for their friendship and the joy they always bring me. My grandparents Ken & Gisela and Paul & Lois have been a constant source of support and encouragement throughout my life, and I greatly appreciate their wisdom.

Finally, to the One who is before all things and in whom all things hold together: for giving me life, a mind to think, and a passion to understand the world; for saving me when I was lost and leading me on the way to eternity; my Christ, my Lord, my God. Soli Deo Gloria.

Chapter 1

Introduction

From the vantage point of any individual agent, the world often exhibits a significant amount of uncertainty. Rarely does one have complete knowledge of the environment at hand, yet often, for an agent to accomplish a desired goal or objective, some amount of knowledge is required; succeeding at a task of interest may require a sufficiently rich understanding of how the world behaves. In such settings, to accomplish the desired goal, the agent must *learn* about the environment, collecting information to reduce uncertainty enough to infer which actions will achieve their goal.

Collecting sufficient information to determine how to act is a ubiquitous challenge throughout our world, and addressing this challenge is central to developing effective, flexible, and safe autonomous systems. From applications as diverse as robotics and control, scientific discovery and drug design, and recommender systems and online platform optimization, many real-world problems of interest fundamentally require learning how to act in uncertain settings. While much work has been devoted to developing algorithmic solutions to these problems, a common shortcoming of many existing approaches, greatly limiting their applicability to real-world problems, is their large *sample complexity*—the number of environment interactions they require in order to learn to accomplish the desired goal.

How can we reduce the sample complexity of learning, and develop effective real-world solutions to such challenges? A naive solution to collecting the requisite information would be to simply explore an environment uniformly, and use the collected observations to learn to accomplish the desired task. This might be very wasteful, however—some aspects of the environment may be much more relevant than others to accomplishing the desired task, and we should therefore focus our exploration on these aspects. How can we identify which aspects of an environment are most relevant, however? And given such knowledge, how can we most effectively focus our exploration?

This thesis takes steps towards addressing these challenges, and developing a theoretically rigorous understanding of both *what* aspects of an environment are most important to learn about in order to accomplish a desired goal, as well as *how* we can effectively collect the necessary information. We are particularly interested in *interactive* settings, where the agent learns about the environment through interaction with it: the agent takes an action, the environment responds in a particular way, and the agent then gains knowledge about how

the environment behaves. Our focus will be on developing algorithmic principles that guide the agent in how to effectively take actions to elicit useful information—methods for efficient *active learning*. We will frame these questions in the context of "decision making", where our end objective is to learn a "decision" (for example, a mapping from states to actions, a controller or policy) that enables an agent to accomplish a desired task. Our goal is to acquire enough information about the environment in order to find such a decision, and to do so as quickly—with the minimum number of environment interactions, minimum sample complexity—as possible.

We will primarily consider *stochastic, dynamic* settings: stateful, long-horizon environments, where the agent's actions influence not only their immediate observations but also their *state*, and therefore what information they are able to acquire in the future. For example, consider a setting where taking some action at point a causes the agent to move to point b; they are now at point b, and can only make observations on aspects of an environment that live in a neighborhood of point b. As we will see, such settings offer significant challenges not addressed in the existing literature, which we will seek to address here. We emphasize the ubiquity of such settings: the real world is inherently sequential, and understanding sequential settings is therefore required for developing methods applicable to many real world problems of interest.

On Quantifying Sample Complexity: Worst-Case vs Instance-Dependent Guarantees

Our goal in this thesis will be to develop provably correct algorithms for the aforementioned problem, give a precise quantification of their sample complexity, and use this quantification to determine which algorithmic strategies are most effective. How should we think about quantifying the sample complexity of learning, however? Before proceeding, we seek to provide some perspective on this question, as it will frame much of the following discussion.

Historically, much of the work in the learning theory community has focused on obtaining *worst-case* or *minimax* guarantees on sample complexity. Such guarantees are typically formulated by considering some set of instances of interest $\mathcal{M}$—for example, $\mathcal{M}$ might correspond to all environments with at most S states and A actions, and which exhibit some particular, amenable structure. The goal is then to (a) develop an algorithm such that, for *any instance* $M \in \mathcal{M}$, the sample complexity of the algorithm is bounded by some value $C(\mathcal{M})$, and (b) show that for *some* instance $M \in \mathcal{M}$ and *any algorithm*, a sample complexity of $C(\mathcal{M})$ is necessary.

Showing (a) and (b) guarantee that our algorithm performs as well as possible on the *hardest* instance within $\mathcal{M}$—on this instance, it achieves a complexity of $C(\mathcal{M})$, and any algorithm must pay a complexity of $C(\mathcal{M})$. Such guarantees allow us to show, first of all, what classes of problems are efficiently learnable. If $C(\mathcal{M})$ is "small", it guarantees that any instance in $\mathcal{M}$ can be tractably learned. Furthermore, such guarantees help us to see what types of algorithms are effective and allow us to learn on instances in $\mathcal{M}$.

Are such guarantees sufficient for eliciting which strategies are most effective at collecting the necessary information to accomplish a desired goal? A key contribution of this thesis is showing that, in many settings, the answer to this is *no*. Indeed, as we will see, there are settings where an algorithm that is minimax optimal on $\mathcal{M}$ may perform *arbitrarily worse* than the best possible algorithm on any particular instance $M^\star \in \mathcal{M}$. While a minimax-style analysis can show that an algorithm is optimal on the *worst-case* instance within $\mathcal{M}$, there may exist instances within $\mathcal{M}$ that are significantly "easier" than the worst-case instance, and a minimax-style analysis says nothing about the optimality of a given approach on such instances, leaving open the possibility that they are very suboptimal on such instances.

The world is rarely "worst-case". When we deploy our algorithm on a given real-world problem, almost certainly it is not as hard as the hardest possible instance, and if our algorithm was designed for optimal performance on the hardest instance, it follows it could be very suboptimal on the particular instance where we wish to deploy it. Ideally, therefore, we might hope to design algorithms that are optimal on every *particular* instance. Given this, much of this thesis will focus on obtaining *instance-dependent* guarantees: algorithms whose performance provably adapts to the hardness of each particular instance, solving "easy" instances in a smaller number of samples than "hard" instances. Critically, as we will see, performing such a refined analysis will allow us to elicit key algorithmic principles which a minimax analysis typically obscures, motivating algorithms with significantly better performance.

Overview of this Thesis

This thesis covers seven of the author's publications on exploration and active learning in dynamic environments. We will frame our discussion primarily within the context of continuous control and reinforcement learning, both canonical decision making settings. The thesis is organized as follows.

First, in the remainder of this section, we outline the primary settings considered in this thesis, provide several example illustrating key challenges of each setting, and highlight the main contributions of the thesis. Next, Chapter 2 covers related work, and seeks to place this thesis into its broader context within the literature.

In Part I we provide our results on learning and exploration for continuous control. Chapter 3 considers the setting of linear dynamical systems and the problem of active system identification, where the goal is to simply learn an accurate model of the unknown system parameters. In Chapter 4 we take a brief detour from our main results by providing some technical results necessary for Chapters 5 and 6, introducing what we refer to as the *martingale decision making* setting, and providing a generic formulation of "smooth" decision making. In Chapter 5 we instantiate the results of Chapter 4 in the linear dynamical systems setting, providing instance-optimal guarantees on "smooth" decision making, which we show, in particular, encompass the linear quadratic regulator problem. In Chapter 6 we show that these results can be extended to certain classes of nonlinear systems, and derive similar results

in such nonlinear systems. The results of Chapter 3 are originally due to Wagenmaker and Jamieson [2020], Chapters 4 and 5 are due to Wagenmaker et al. [2021], and Chapter 6 are due to Wagenmaker et al. [2024].

Part II covers the reinforcement learning setting, beginning in Chapter 7 with tabular Markov Decision Processes (MDP), where we obtain some of the first instance-dependent guarantees on the policy identification problem in reinforcement learning. Chapter 8 extends the results of Chapter 7 to the function approximation setting, specifically the linear function approximation setting, and in Chapter 9 we consider the *reward-free* problem in the linear function approximation setting, providing the first optimal algorithm for reward-free RL with function approximation. The results of Chapter 7 are originally from Wagenmaker et al. [2022c], Chapter 8 covers the results of Wagenmaker and Jamieson [2022], and Chapter 9 is due to Wagenmaker et al. [2022b].

Finally, in Part III, we consider the problem of *learning to explore*. As we will see, there is an intrinsic cost to answering the "what" and "how" questions of exploration, and in order to perform as effectively as possible, we must not only learn to achieve our goal as efficiently as possible, we must learn to explore as efficiently as possible. These results, initially due to Wagenmaker and Foster [2023], are outlined in Chapter 10, and consider a general function approximation setting.

While we provide proofs for the key results in the main text, for the sake of readability, we defer many of the more technical proofs to the appendix.

Mathematical Notation

We let $\mathbb{R}, \mathbb{C}, \mathbb{N}$ refer to real numbers, complex numbers, and natural numbers, respectively. $\mathcal{S}^{d-1}$ denotes the unit sphere in d dimensions. $\mathcal{B}$ generally denotes a ball, but will be specified within context. $\mathbb{S}^d_{++}$ denotes the set of d-dimensional positive definite matrices, and $\mathbb{S}^d_{+}$ the set of d-dimensional positive semi-definite matrices. Given a matrix $A \in \mathbb{R}^{d_1 \times d_2}$ we let $\|A\|_{\mathrm{F}}$ denote the Frobenius norm of A, and $\|A\|_{\mathrm{op}}$ denote its operator norm. $\mathrm{vec}(\cdot)$ denotes the vectorization of a matrix. $\triangle_{\mathcal{X}}$ denotes the simplex over $\mathcal{X}$. We use standard big-oh notation, which we denote with $\mathcal{O}(\cdot)$, and let $\widetilde{\mathcal{O}}(\cdot)$ suppress logarithmic terms as well. We will use $\lesssim$ and $\gtrsim$ in informal statements to denote inequality up to lower-order terms and potentially other problem-dependent terms. Throughout, c and C denote universal constants. For $n \in \mathbb{N}$, $n \geq 1$, we let $[n] := \{1, 2, \dots, n\}$. ι denotes the imaginary number.

1.1 Active Learning for Continuous Control

Continuous control typically studies the setting of dynamical systems with continuous states, which we can denote generically as:

$$x_{t+1} = f_{\theta_\star}(x_t, u_t, w_t) \tag{1.1.1}$$

for $x_t \in \mathbb{R}^{d_x}$ the state, $u_t \in \mathbb{R}^{d_u}$ the input, $w_t \in \mathbb{R}^{d_w}$ noise, f_θ some parametric form, and $\theta_\star \in \mathbb{R}^{d_\theta}$ the corresponding parameter of the true environment, which we assume is unknown. We consider two particular instantiations of (1.1.1): where f_θ is a linear dynamical system (Chapters 3 and 5), and f_θ a nonlinear dynamical system with observations linear in $\theta_\star$ (Chapter 6). We assume the agent can choose u_t at each timestep, in order to control the evolution of the system, and observes the state x_t.

We will quantify the end goal the agent hopes to achieve in terms of some loss $\mathcal{J}_{\theta_\star}(\mathfrak{a})$, where $\mathfrak{a} \in \mathbb{R}^{d_\mathfrak{a}}$ is some *decision*. For example, $\mathfrak{a}$ may correspond to parameters of a controller, and $\mathcal{J}_{\theta_\star}(\mathfrak{a})$ the loss the controller will incur on the system $\theta_\star$. With this formulation, the objective of the agent is to find some $\mathfrak{a}$ minimizing $\mathcal{J}_{\theta_\star}(\mathfrak{a})$—this corresponds to the agent achieving their goal. We assume the form of our loss $\mathcal{J}_{\theta_\star}(\mathfrak{a})$ is known but, as we do not assume knowledge of $\theta_\star$, the value of the true loss $\mathcal{J}_{\theta_\star}(\mathfrak{a})$ for any particular $\mathfrak{a}$ is in general unknown.

We consider the following interaction protocol:

1. Agent interacts with (1.1.1) for T steps, playing some exploration policy π_{exp} in allowable set of exploration policies Π_{exp} (for example, policies with bounded power, or deemed to be "safe").

2. After T steps, agent proposes some decision $\widehat{\mathfrak{a}}_T$.

3. Agent suffers loss $\mathcal{J}_{\theta_\star}(\widehat{\mathfrak{a}}_T)$.

The goal of the agent is therefore to learn as much useful information about the true system (1.1.1) as possible through the T steps of interaction, so that they may then propose a decision $\widehat{\mathfrak{a}}_T$ which comes as close as possible to minimize the true loss.

1.1.1 Decision Making in Linear Dynamical Systems

In a (discrete-time) linear dynamical system, the dynamics (1.1.1) take the following form:

$$x_{t+1} = A_\star x_t + B_\star u_t + w_t, \quad x_0 \equiv 0, \tag{1.1.2}$$

for $A_\star \in \mathbb{R}^{d_x \times d_x}$ and $B_\star \in \mathbb{R}^{d_x \times d_u}$ the system matrices, and $\theta_\star = (A_\star, B_\star)$ the unknown parameter. For simplicity we assume the process noise w_t is distributed as $w_t \sim \mathcal{N}(0, \sigma_w^2 \cdot I)$ for some $\sigma_w > 0$. Linear dynamical systems are a canonical setting in the controls literature and have been the subject of significant study for over 50 years, with many real-world applications. Furthermore, settings that exhibit some amount of nonlinearity can typically be *linearized* around a point of interesting, rendering (1.1.2) an accurate representation of the dynamics locally. We are motivated by two particular formulations of our goal, $\mathcal{J}_{\theta_\star}(\mathfrak{a})$.

Example 1.1.1 (System Identification)**.** In the system identification problem, the agent's goal is simply to produce an estimate of the system's parameter $\theta_\star$. We can represent this

problem in the framework introduced above by setting our "decision" $\mathfrak{a}$ to θ, and the loss $\mathcal{J}_{\theta_\star}(\mathfrak{a}) = \|\mathfrak{a} - \theta_\star\|_{\mathrm{op}}$. Then finding some $\mathfrak{a}$ which minimizes $\mathcal{J}_{\theta_\star}(\mathfrak{a})$ corresponds to estimating $\theta_\star$ as accurately as possible.

We consider the system identification setting in Chapter 3. While system identification has a long history in the control theory and statistics communities, our results are some of the first to develop provably efficient algorithms for the active system identification problem. Our next example of interest is the classical linear quadratic regulator problem.

Example 1.1.2 (Linear Quadratic Regulator (LQR))**.** In the LQR problem, the agent's objective is to design a policy that minimizes the infinite-horizon cumulative cost, with losses $\ell(x, u) := x^\top R_{\mathbf{x}} x + u^\top R_{\mathbf{u}} u$. The resultant cost function is:

$$\mathcal{J}_{\mathrm{LQR},\theta_\star}[\pi] := \lim_{T\to\infty} \mathbb{E}_{\theta_\star,\pi}\left[\frac{1}{T}\sum_{t=1}^{T} x_t^\top R_{\mathbf{x}} x_t + u_t^\top R_{\mathbf{u}} u_t\right].$$

It is well known that the optimal policies are of the form $u_t = Kx_t$ where $K \in \mathbb{R}^{d_u \times d_x}$; we denote these policies π^K, and let $\mathcal{J}_{\mathrm{LQR},\theta}(K) = \mathcal{J}_{\mathrm{LQR},\theta}[\pi^K]$. Here our decision $\mathfrak{a}$ is the controller K and our loss $\mathcal{J}_\theta$ is $\mathcal{J}_{\mathrm{LQR},\theta}$.

The LQR problem is perhaps the most canonical problem in linear control, and is broadly applicable to a range of real-world settings. We consider the LQR problem in Chapter 5, as well as the setting of learning with general smooth losses $\mathcal{J}$ (which we show the LQR problem is a special case of).

Exploration in Linear Dynamical Systems. How does the information we gain interacting with (1.1.2) help us choose a decision $\widehat{\mathfrak{a}}_T$ that minimizes our loss? And how can we choose our exploration policy π_{exp} to excite (1.1.2) in order to make the loss of $\widehat{\mathfrak{a}}_T$ as small as possible? Folk wisdom in the controls community suggests that all that is necessary is *persistent excitation*—consistently excite all "modes" of (1.1.2) through the T steps of interaction, and use the observed data to estimate the system parameters and propose a decision. Is this the entire story, however?

As an example, consider the following 2D LQR instance:

$$x_{t+1} = \begin{bmatrix}\rho & 0\\ 0 & \rho\end{bmatrix} x_t + \begin{bmatrix}1 & 0\\ 0 & 1\end{bmatrix} u_t + w_t, \quad R_{\mathbf{x}} = \begin{bmatrix}\kappa & 0\\ 0 & 1\end{bmatrix}, \quad R_{\mathbf{u}} = \begin{bmatrix}\mu & 0\\ 0 & \mu\end{bmatrix}, \tag{1.1.3}$$

for some $\rho \in (0, 1), \kappa > 0$, and $\mu > 0$. We see that, in this system, the dynamics and input cost behave identically in each coordinate. However, in the regime where $\kappa \gg 1$, the state cost $R_{\mathbf{x}}$ is significantly larger in the first coordinate than the second.

Now consider the loss of a control policy designed to minimize this cost, where the controller is synthesized on an estimate of the system parameters obtained from interaction with the system. In this case, as the state cost is much larger in the first coordinate than

the second, a small deviation from the optimal control sequence in the first coordinate could be much costlier than a similar deviation in the second coordinate. As estimation error will translate into deviation from the optimal control law, it follows that estimation error in the first coordinate is significantly more costly than estimation error in the second coordinate. This suggests that, while persistent excitation may indeed allow us to learn the parameters of (1.1.3) eventually, and thereby propose a controller minimizing the cost, we may be able to learn significantly more quickly by focusing on our exploration on learning the dynamics in the first coordinate. Indeed, this can be achieved by increasing the input in the first coordinate as much as our constraints allow.

This example suggests that *some parameters of the system may be more important than others* and *we should direct our exploration to quickly learn the most relevant parameters*. Indeed, we show in Chapters 3 and 5 that if we do not take into account this structure and, for example, seek to excite all modes equally well, the rate at which we can learn a decision minimizing $\mathcal{J}_{\theta_\star}(\mathfrak{a})$ could be *arbitrarily worse* than the optimal achievable rate.

How do we achieve this optimal rate? In Part I we provide a general characterization of how uncertainty about the parameter $\theta_\star$ translates into suboptimality in the decision we propose, assuming our loss is "smooth" and we are using the *certainty equivalence* decision rule, which sets $\widehat{\mathfrak{a}}_T$ to the optimal decision on the estimated system. In particular, we show that the suboptimality of such a decision scales with estimation error in a particular, task-dependent norm:

$$\mathcal{J}_{\theta_\star}(\widehat{\mathfrak{a}}_T) - \min_{\mathfrak{a}} \mathcal{J}_{\theta_\star}(\mathfrak{a}) \approx \|\widehat{\theta}_T - \theta_\star\|^2_{\mathcal{H}(\theta_\star)},$$

where $\widehat{\theta}_T$ denotes the estimate of $\theta_\star$, $\widehat{\mathfrak{a}}_T$ is the optimal decision on $\widehat{\theta}_T$, and $\mathcal{H}(\theta_\star)$ measures the *curvature* of the loss function $\mathcal{J}_{\theta_\star}(\mathfrak{a})$ in a neighborhood of the optimal decision. We then show that $\|\widehat{\theta}_T - \theta_\star\|^2_{\mathcal{H}(\theta_\star)}$ can be minimized by choosing our inputs during exploration to minimize:

$$\operatorname{tr}(\mathcal{H}(\theta_\star) \cdot \check{\mathbf{\Lambda}}_T^{-1}) \quad \text{for} \quad \check{\mathbf{\Lambda}}_T = I_{d_x} \otimes \sum_{t=1}^{T} \begin{bmatrix} x_t \\ u_t \end{bmatrix} \begin{bmatrix} x_t \\ u_t \end{bmatrix}^\top.$$

We develop a general algorithmic procedure able to explore so as to minimize this objective, and show that this procedure is computationally efficient, involving only the solution to convex programs, and that it achieves the instance-optimal rate. Our results paint a nearly complete picture of learning in linear dynamical systems. As originally motivated, they provide a precise answer to how we should interact with our environment (in this case (1.1.2)) to collect the necessary information to learn to achieve some goal (in this case quantified by the loss $\mathcal{J}_{\theta_\star}(\mathfrak{a})$).

1.1.2 Decision Making in Nonlinear Systems

While linear systems are able to model a range of real-world settings, many systems of interest are inherently nonlinear, and attempting to represent them with linear dynamics may eliminate key features, rendering such models ineffective for solving downstream tasks. At the same time, considered in full generality, nonlinear systems are notoriously difficult to work with, and are often both computationally and statistically intractable. In Chapter 6 we consider a class of nonlinear systems that seeks to strike a balance between expressivity and tractability. In particular, we consider systems of the form:

$$x_{t+1} = A_\star \boldsymbol{\phi}(x_t, u_t) + w_t, \tag{1.1.4}$$

where $A_\star \in \mathbb{R}^{d_x \times d_{\boldsymbol{\phi}}}$ is the unknown system parameter, and $\boldsymbol{\phi} : \mathbb{R}^{d_x} \times \mathbb{R}^{d_u} \to \mathbb{R}^{d_{\boldsymbol{\phi}}}$ is a known, potentially nonlinear, mapping. In this setting, we are interested in the same high-level goal as the linear setting: finding some decision $\mathfrak{a}$ that minimizes a loss $\mathcal{J}_{\theta_\star}(\mathfrak{a})$ for $\theta_\star = A_\star$.

Systems of the form (1.1.4) are amenable to analysis as the observations we receive—the next state, x_{t+1}—are a linear function of the unknown parameter, $A_\star$, allowing us to apply results from linear estimation to argue about learning $A_\star$. At the same time, systems of this form are very expressive and indeed, given a rich enough featurization $\boldsymbol{\phi}$, are capable of representing any smooth nonlinear system up to arbitrary tolerance. Furthermore, such systems have found recent applications in real-world control settings such as drone control [O'Connell et al., 2022].

While bearing similarity to linear systems in how the unknown parameter relates to the observations, from the perspective of exploration, nonlinear systems of the form (1.1.4) introduce a variety of subtleties and challenges not found in the linear case. First, while settings such as (1.1.3) show that in the linear case there may be a large asymmetry in which parameters are most important, in the nonlinear case, we may have parameters that are truly inactive, and have *no* impact on the dynamics near an equilibrium point of interest whatsoever. For example, consider the following 1D system:

$$x_{t+1} = 0.8 \cdot x_t + u_t - 3\boldsymbol{\phi}_1(x_t) - 3\boldsymbol{\phi}_2(x_t) + w_t$$

where $\boldsymbol{\phi}_1(x) = \max\{1 - 100(x-10)^2, 0\}$ and $\boldsymbol{\phi}_2(x) = \max\{1 - 100(x+10)^2, 0\}$. Assume our goal is to find some control policy which minimizes the cost

$$\text{cost}(x, u) = (x - 10)^2 + 100^{-1} \cdot u^2.$$

Here we see that the nonlinearity $\boldsymbol{\phi}_1(x)$ is only active (nonzero) for x very close to 10, while $\boldsymbol{\phi}_2(x)$ is only active for x very close to -10. As our cost is minimized for $x = 10$, the optimal controller will attempt to steer the state to a neighborhood of $x = 10$. In this case, however, $\boldsymbol{\phi}_2(x)$ is completely inactive and thus the behavior of $\boldsymbol{\phi}_2(x)$ is irrelevant to accomplishing our goal. If we assume the coefficients of $\boldsymbol{\phi}_1$ and $\boldsymbol{\phi}_2$ are unknown and must be learned, this

suggests that we should allocate no input energy to learning the coefficient of ϕ_2. We see then that, even more so in nonlinear than linear systems, it is critical that we determine which parameters are most relevant to the goal at hand, and focus our exploration only on learning these parameters.

A second challenge in nonlinear systems is how to realize an exploration strategy which learns these relevant parameters. In the linear case our methods relied on the fact that, while learning could be sped up by properly choosing our inputs, even "bad" inputs would provide sufficient excitation to ensure *something* was learned. In contrast, in the nonlinear case, "bad" inputs may provide no learning signal at all—for instance, in the example above, we can only learn about the coefficients of ϕ_1 by reaching $x = 10$, which takes a particular sequence of inputs to achieve, and so much more careful planning is required to ensure the necessary information is collected. To address this challenge, we develop a generic technique which reduces the data collection problem (the *experiment design* problem) to policy optimization, and then apply an off-the-shelf policy optimization procedure to solve it.

We present our full results in this setting in Chapter 6. While the nonlinear setting introduces the aforementioned challenges not found in the linear case, and require different algorithmic techniques, our results here bear much resemblance to those presented in Chapter 5—we ultimately show that the suboptimality of our learned decision scales with the estimation error of $A_\star$ in a similar task-dependent norm, $\|\mathrm{vec}(\widehat{A} - A_\star)\|^2_{\mathcal{H}(A_\star)}$, and that this suboptimality can be reduced by exploring so as to collect covariates $\check{\mathbf{\Lambda}}_T$ minimizing $\mathrm{tr}(\mathcal{H}(\theta_\star) \cdot \check{\mathbf{\Lambda}}_T^{-1})$. We present an instance-dependent upper bound and matching lower bound which, to the best of our knowledge, is the first statistically optimal guarantee for systems of the form (1.1.4).

1.2 Exploration in Reinforcement Learning

We turn now to the reinforcement learning (RL) setting, the focus of Part II of this thesis. While there is significant overlap between continuous control and RL, and indeed continuous control can be seen as a particular RL problem, for the purposes of this thesis we make a distinction between these settings, and when referring to RL typically refer to settings with discrete states and actions.

The RL setting is primarily concerned with the study of Markov Decision Processes (MDPs), a generic formulation of decision making in stochastic environments. In an MDP, the agent operates in a state space $\mathcal{S}$, taking actions in some set $\mathcal{A}$. At step h, if the agent is in state s and takes action a, they transition to $s' \sim P_h(\cdot \mid s, a)$, for $P_h(\cdot|s, a) \in \triangle_{\mathcal{S}}$ the transition kernel, and receive reward $R_h(s, a)$. We will be concerned here with the *episodic* setting, where the agent interacts with the MDP for H steps, at which points the process resets and begins anew. We assume the dynamics $\{P_h\}_{h=1}^H$ and rewards $\{R_h\}_{h=1}^H$ are both initially unknown, and the objective of the agent is to learn a policy π, a mapping from states to actions, that achieves maximum expected reward. Here, then, the goal of the agent is

quantified in the reward function, and it must learn to maximize this reward.

We can quantify the performance of a policy in terms of the *value function*, defined as $V_h^\pi(s) := \mathbb{E}_\pi[\sum_{h'=h}^H R_{h'}(s_{h'}, a_{h'})|s_h = s]$, the expected reward that will be achieved playing policy π from step h and state s. Similarly, the Q-value function, $Q_h^\pi(s,a)$, denotes the expected reward that will be obtained if we are in state s at step h, play action a, and then play policy π for the remainder of the episode; formally, $Q_h^\pi(s,a) := \mathbb{E}_\pi[\sum_{h'=h}^H R_{h'}(s_{h'}, a_{h'})|s_h = s, a_h = a]$. The optimal value function is given by $V_h^\star(s) = \sup_\pi V_h^\pi(s)$ and optimal Q-value function by $Q_h^\star(s,a) = \sup_\pi Q_h^\pi(s,a)$, where the suprema is taken over all policies. The *value of a policy* is given by $V_0^\pi = V_1^\pi(s_1)$—the expected reward policy π achieves over an entire episode—and we say a policy π is optimal if $V_0^\pi = V_0^\star$.

PAC Reinforcement Learning. We consider a slightly different interaction protocol in the RL setting as compared to the continuous control setting. While in the continuous control setting we assume the agent interacts with the dynamics for some fixed amount of episodes T, in the RL setting we assume instead that the agent is given two parameters $\epsilon \geq 0$ and $\delta \in (0,1)$, and is tasked with identifying a policy $\widehat{\pi}$ that is ϵ-*optimal* with probability at least $1-\delta$:

$$V_0^\star - V_0^{\widehat{\pi}} \leq \epsilon.$$

This is typically referred to as the PAC RL (Probably-Approximately-Correct) problem, and has been the subject of much attention in the RL community. Given this, we consider the following sequence of interactions:

1. Agent is given parameters $\epsilon \geq 0$ and $\delta \in (0,1)$.

2. Agent interacts with MDP for some number of steps, choosing when to terminate.

3. Agent returns policy $\widehat{\pi}$ such that $V_0^\star - V_0^{\widehat{\pi}} \leq \epsilon$ with probability at least $1-\delta$.

The goal of the agent is to do this as efficiently as possible—to find an ϵ-optimal policy (its "decision") using the smallest number of interactions with the environment possible.

Without imposing additional structure on the MDP, the RL problem is in general intractable. In this thesis, we consider two particular tractable MDP formulations: tabular and linear MDPs.

1.2.1 Instance-Dependent Tabular RL

In the tabular RL setting, the key assumption is that the state and action spaces are finite: $S := |\mathcal{S}| < \infty$ and $A := |\mathcal{A}| < \infty$. We make no other assumptions on the MDP—for instance, states and actions are completely unrelated and knowing the behavior at state s_1 says nothing about the behavior at state s_2. As we assume the number of states and actions are finite, however, there are a finite number of entities that must be learned, making learning tractable.

While the tabular setting is somewhat limited in its ability to model real-world settings of interest, it has received significant attention from the research community as a distillation of more complex and realistic settings which preserves the salient features necessary for understanding effective algorithmic principles. We therefore begin our study of RL here. The majority of existing work in tabular PAC RL has focused on obtaining worst-case guarantees, which in this setting scale with quantities such as S and A. While such bounds show that learning is feasible in tabular RL, they do not offer clear insights into what features of an environment are most important to learn or (as we will see) how we should explore.

If our goal is to learn to accomplish a task of interest, in this case quantified by the reward function, how should we explore in order to quickly learn to do this? For inspiration, we consider the special case of multi-armed bandits (MABs), where $S = H = 1$. In MABs, we simply have A actions ("arms") to choose from at each step—we choose an action, observe a noisy realization of the reward mean, and repeat, with the goal of identifying the best action. If we denote the reward mean for action a as $r(a) := \mathbb{E}[R(a)]$, and the *gap* for action a as $\Delta(a) := \max_{a'} r(a') - r(a)$—the difference between the value of arm a and the value of the best arm—then existing work in MAB shows that the optimal strategy is to pull arm a at a rate proportional to $\frac{1}{\Delta(a)^2}$. Intuitively, for an arm a that is suboptimal but very close to optimal, it will take significantly more observations to distinguish its value from the value of the optimal arm than it would to distinguish a very suboptimal arm, and we therefore must pull it more. This is reflected in the $\frac{1}{\Delta(a)^2}$ rate—we pull suboptimal arms less as we can quickly show they are suboptimal, and allocate the majority of our pulls to arms close to optimal. Indeed, one can show that, to identify the best arm, it suffices to pull each arm approximately $\frac{1}{\Delta(a)^2}$ times, leading to a final sample complexity on order

$$\sum_a \frac{1}{\Delta(a)^2}, \tag{1.2.1}$$

which is known to be instance-optimal.

In the tabular RL setting, where $S > 1$ and $H > 1$, we can imagine running a similar procedure: treating each state as a bandit, and allocating our samples based on the suboptimality of each action in that state. Yet this introduces several questions. First, how do we quantify the suboptimality of an action, given that the return from playing a single action depends not just on that action, but also on all subsequent actions played in the episode? And second, if our goal is exploration, the actions we take affect which states we reach. While we may know which actions we want to sample in a given state, how do we determine which states we should aim to reach, and how do we actually reach these states?

For the first question, existing work has proposed a generalization of the notion of a "gap" to the following [Simchowitz and Jamieson, 2019]:

$$\Delta_h(s, a) := V_h^\star(s) - Q_h^\star(s, a).$$

That is, $\Delta_h(s,a)$ measures the suboptimality of playing action a in (s,h), if we then play the optimal policy. Given this, our exposition of bandits suggests we should play actions in state s proportional to $\frac{1}{\Delta_h(s,a)^2}$. The challenging question, however, is the second: how do we handle the dynamic nature of an MDP during exploration? Assume we are playing some exploration policy π_{exp}, and denote the probability of visiting state s at step h as $w_h^{\pi_{\text{exp}}}(s) := \mathbb{P}^{\pi_{\text{exp}}}[s_h = s]$. Using the complexity given in (1.2.1) for solving the "bandit" at state s, we see that to "solve" state s it will require running π_{exp} for a total of

$$\frac{1}{w_h^{\pi_{\text{exp}}}(s)} \cdot \sum_a \frac{1}{\Delta_h(s,a)^2} \tag{1.2.2}$$

episodes, since on average it takes $\frac{1}{w_h^{\pi_{\text{exp}}}(s)}$ episodes to reach s once. If we wish to "solve" each state, then playing the best-case exploration policy, we might hope to achieve a sample complexity of:

$$\min_{\pi_{\text{exp}}} \max_s \frac{1}{w_h^{\pi_{\text{exp}}}(s)} \cdot \sum_a \frac{1}{\Delta_h(s,a)^2}. \tag{1.2.3}$$

In Chapter 7, we show that this complexity is in fact achievable, and propose an algorithmic procedure which achieves it. Furthermore, we show that, in some sense, it is unimprovable: in general, we cannot obtain a smaller sample complexity on a particular instance. One shortcoming of this complexity is that it corresponds to identifying the *best* policy (i.e. $\epsilon = 0$), while in general in RL we only care about identifying a *good* policy. To address this, we show that this complexity, and our algorithm, can be generalized to the $\epsilon > 0$ case as well.

What does this say about how we should explore in tabular MDPs? The majority of existing work in tabular RL relies on the principle of *optimism*—choosing actions that are the "best possible", given the current observations from the environment, leading them to play "good" actions the majority of the time. Such approaches are provably efficient, and are known to be optimal in a worst-case sense. To understand if these approaches are able to achieve the complexity of (1.2.3), consider the MDP in Figure 1.1. In state s_0, action a_1 is optimal and transitions to state s_1 with probability $1-p$ and state s_2 with probability p. Action a_2 is suboptimal and transitions to state s_2 with probability 1. To learn a good policy, we need to identify the optimal action in both s_1 and s_2. An optimistic algorithm will primarily play a_1 in s_0, as this action is optimal, and it will therefore only reach

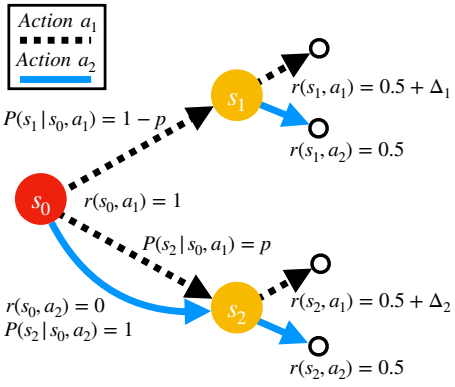

Figure 1.1: A motivating example for tabular RL

s_2 approximately $\mathcal{O}(pK)$ times, after K episode. It follows that a low-regret algorithm will take at least $\Omega(\frac{1}{p\Delta_2^2})$ episodes to learn the optimal action in s_2. In contrast, we could instead play a_2 in s_0, collecting less reward but learning the optimal action in s_2 in only $\Omega(\frac{1}{\Delta_2^2})$ episodes, exactly the complexity reflected in (1.2.3). For small p, this could be arbitrarily better.

This example shows that approaches—such as optimism—which are known to be optimal in a worst-case sense are unable to achieve the complexity (1.2.3), and can in fact be very suboptimal on particular instances, achieving a far worse rate. Instead, to achieve (1.2.3), we must take maximally informative actions, in this case a_2. We emphasize that only by developing a precise understanding of the complexity of learning, e.g. (1.2.3), are we able to arrive at such a conclusion. Our results, therefore, shed light not just on the best achievable complexity, but also on what exploration procedures we should employ to realize such complexities, and what strategies would be suboptimal.

1.2.2 Instance-Dependent Linear RL

While the tabular setting is helpful for eliciting fundamental algorithmic principles, its applicability to real-world problems is limited. In practice, MDPs of interest often have very large, yet highly structured, state spaces—for example, knowledge of the behavior of one state tells how the MDP will behave in "adjacent" states. This has motivated a large amount of work on RL with *function approximation*, which seeks to incorporate such structure into provably efficient approaches.

In Chapter 8 and Chapter 9 we consider, in particular, the setting of RL with *linear* function approximation. While several formulations for RL with linear function approximation exist, we consider the *linear MDP* setting.

Definition 1.2.1 (Linear MDPs [Jin et al., 2020b])**.** We say that an MDP is a *d-dimensional linear MDP*, if there exists some (known) feature map $\boldsymbol{\phi}(s,a) : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$, H (unknown) signed vector-valued measures $\boldsymbol{\mu}_h \in \mathbb{R}^d$ over $\mathcal{S}$, and H (unknown) reward vectors $\boldsymbol{\theta}_h \in \mathbb{R}^d$, such that:

$$P_h(\cdot|s,a) = \langle \boldsymbol{\phi}(s,a), \boldsymbol{\mu}_h(\cdot)\rangle, \quad r_h(s,a) = \langle \boldsymbol{\phi}(s,a), \boldsymbol{\theta}_h\rangle.$$

We will assume $\|\boldsymbol{\phi}(s,a)\|_2 \le 1$ for all s,a; and for all h, $\||\boldsymbol{\mu}_h|(\mathcal{S})\|_2 = \|\int_{s\in\mathcal{S}} |\mathrm{d}\boldsymbol{\mu}_h(s)|\|_2 \le \sqrt{d}$, and $\|\boldsymbol{\theta}_h\|_2 \le \sqrt{d}$.

Linear MDPs encompass, for example, tabular MDPs, but can also model more complex settings, such as feature spaces corresponding to the simplex, or the linear bandit problem. Critically, linear MDPs allow for infinite state-spaces, as well as generalization across states—rather than learning the behavior in particular states, we can learn in the d-dimensional feature space.

Within the linear MDP setting, we consider the same PAC RL problem, and aim to understand how, given the linear structure, what and how we should explore. In the tabular setting, we aimed to solve each state (find a near-optimal action), yet this *action-elimination* strategy is not possible to realize in the linear case as the number of states is potentially infinite. Instead, we aim to learn over the space of *policies*. Given a set of possible policies Π, we seek to gradually eliminate suboptimal policies in Π, focusing our exploration to learn about the remaining "active" policies. We show that this allows us to refine where we explore and focus our exploration only on the most relevant regions, the regions visited by good policies. In Chapter 8 we obtain a complexity somewhat analogous to (1.2.3), but where the gap between actions is replaced by the gap between *policies*.

Reward-Free RL in Linear MDPs. In Chapter 9, we also consider linear MDPs, but with a slightly different objective. In both Chapter 7 and Chapter 8 the goal was to find a policy that maximizes a *single* reward function. In some settings, however, one may not have one reward function they wish to maximize, but instead may wish to learn a set of behaviors that can optimize any given reward function. For example, rather than training an agent to accomplish a particular task, we might want to train them to accomplish any task we tell them to. This setting is typically known as the *unsupervised* or *reward-free* RL setting. Formally, the objective is now to learn a map such that, given any reward function, this map produces a policy that is ϵ-optimal for this reward function.

While the objective here is different, the fundamental questions remain very similar: what and how should we explore in order to learn such a map as efficiently as possible? In Chapter 9 we propose an exploration strategy very similar to that utilized in Chapter 7, which aims to explore to maximize the information gain, and show that this exploration strategy leads to a minimax-optimal sample complexity guarantee. Intuitively, one would expect the reward-free RL setting to be harder than the reward-aware (PAC) RL setting—we are learning a policy to maximize every possible reward, vs only one reward—and indeed, this is known to be the case in tabular RL. Surprisingly, however, our results show that, given an effective exploration strategy, this is not true in the linear MDP setting: reward-free RL is no harder than reward-aware RL. We remark, though, that this conclusion only holds in a minimax sense—the picture is less clear in an instance-dependent setting.

1.3 The Cost of Learning to Explore

In both the RL and continuous control settings, the key to achieving the aforementioned results lies in learning effective *exploration policies*, policies that takes actions to efficiently traverse the environment and collect informative observations. Yet there is a cost to learning these exploration policies. As we will see in the precise statements of our results, in nearly

every situation, our complexities behave like:

$$\underbrace{\text{“leading-order” term}}_{\text{instance-optimal}} \quad + \quad \underbrace{\text{“lower-order” term.}}_{\text{polynomial in problem parameters}}$$

Here the “leading-order” term is the instance-optimal rate, which we can show is unimprovable on our particular instance, and typically scales as $1/T$ (in fixed-budget setting) or $1/\epsilon^2$ (in PAC setting). The “lower-order” term typically scales polynomially in problem parameters and as $1/T^\alpha$ for $\alpha > 1$ or $1/\epsilon^\alpha$ for $\alpha < 2$. This lower-order term is dominated by the leading-order term for large values of T or small values of ϵ, but for small T or large ϵ, it may be larger than the instance-optimal term.

Taking a closer look at the analysis, in nearly every case we find that the lower-order term arises because the algorithm must pay some number of samples to *learn an effective exploration policy*. A priori the algorithm does not know which actions will lead to informative observations, and must take some number of interactions with the environment to determine this. Consider, for example, the problem instance presented in Section 1.2.1 in the tabular RL setting. Here, as we argued, the best exploration policy will play action a_2 some fraction of the time, as this will allow it to efficiently learn the optimal action in state s_2. Yet initially, the algorithm does not know that this action is informative and should be taken—to determine this is a useful action to take, it must first explore the environment to find the useful actions.

While it is known that the instance-optimal term is unavoidable and any algorithm must pay this, prior to the work of this thesis, it was unknown if such a lower-order term for “learning to explore” was necessary and, if it is necessary, what the optimal lower-order term might be. Answering this question is of paramount importance in understanding instance-optimality in decision making. While the results presented in Part I and Part II of this thesis exhibit lower-order terms at most polynomial in problem parameters, in the instance-dependent literature more broadly, it is common to have “lower-order” terms which scale exponentially with problem parameters. In such settings, the leading-order term dominates only asymptotically, making such bounds (and the corresponding algorithms) impractical for virtually all problems of interest. If such lower-order terms are necessary to achieve instance-optimality, then instance-optimality is simply not a practical measure for most problems of interest.

In Part III, we seek to understand this phenomenon, and provide a comprehensive theory of *learning to explore*. We study these questions in as much generality as possible, and consider the *Decision Making with Structured Observations* (DMSO) setting recently introduced by Foster et al. [2021]. This setting provides a general framework for formalizing decision making, and encompasses virtually all of bandits and RL (including function approximation settings). We provide a precise definition of the setting in Chapter 10. In contrast to Parts I and II where our goal was primarily to learn effective policies, in Part III we turn our focus to *regret*, where the goal is to maximize the total reward collected during interaction with the environment. While our objective changes, as we will make clear in Part III, when we are concerned with

instance-optimal regret, the fundamental challenge to achieving instance-optimality still lies in learning an effective exploration policy (the primary difference, in this case, being that this exploration policy must balance exploration with exploitation). Thus, the insights from Part III apply to Parts I and II as well.

Intuitively, the instance-optimal rate quantifies the cost of distinguishing the true instance from every instance with a different optimal decision. We show in Chapter 10 that this same intuition applies to learning exploration policies: to explore optimally, one must distinguish our true instance from every instance that has a different set of optimal *exploration policies*. We quantify the cost necessary to learn this optimal exploration policy in a novel complexity measure which we call the *Allocation Estimation Coefficient* (AEC). We show that, under certain conditions, any algorithm that is instance-optimal must also have learned an optimal exploration policy, and so therefore must have complexity which scales, in addition to the leading-order instance-optimal term, with the AEC.

Inspired by this, we propose an algorithmic procedure which applies in the general DMSO setting, and achieves regret bounded as:

$$\text{“instance”-optimal term} \cdot \log T + \text{AEC} \cdot o(\log T),$$

achieving both the instance-optimal leading order term, and a lower-order term scaling with the AEC. Together, our results show that, to achieve instance-optimality, one must also pay for a term scaling as the AEC in order to learn the optimal allocation, and that our algorithm is optimal, both in the leading-order and lower-order term. We further show that, by instantiating our results in various concrete settings, the AEC in general scales at most polynomially in problem parameters, implying that polynomial lower-order terms are indeed possible in most settings of interest, and yielding an exponential improvement over existing work in a variety of settings.

What do these results suggest about how we should think about exploration? The results of Parts I and II focused on how to explore to learn near-optimal policies as quickly as possible, and provided insights into how and where we should direct our exploration to do this. The results of Part III suggest that, in addition to efficiently exploring to learn the optimal policy, we also must efficiently explore to learn the best *exploration policy*, and that there is a cost to learning this exploration policy. Thus, to return to our overarching goal of understanding how agents should collect observations in unknown environments, an agent must carefully take actions so as to provide information not just about how to accomplish their end goal, but also to learn actions that will help them learn to accomplish their end goal.

Chapter 2

Related Work

Before moving on to the main contributions of this thesis, we seek to place it within the broader literature.

2.1 Related Work in Continuous Control

We first focus on several directions of existing work in the continuous control setting, the focus of Part I of this thesis.

2.1.1 System Identification

There is a large body of classical work in system identification [Ljung, 1998]. Of particular relevance are works which seek to design inputs to excite the system to speed up estimation. In particular, a variety of works aim to design inputs to maximize certain functionals of the Fisher information matrix [Mehra, 1976, Goodwin and Payne, 1977, Jansson and Hjalmarsson, 2005, Gevers et al., 2009, Manchester, 2010, Hägg et al., 2013], or design inputs to meet certain task-specific objectives [Hjalmarsson et al., 1996, Hildebrand and Gevers, 2002, Katselis et al., 2012]. The optimal choice of inputs in general depends on the unknown parameters of the system, and prior work rely on either robust experiment design [Rojas et al., 2007, 2011, Larsson et al., 2012, Hägg et al., 2013] or adaptive experimental design [Lindqvist and Hjalmarsson, 2001, Gerencsér and Hjalmarsson, 2005, Barenthin et al., 2005, Gerencsér et al., 2007, 2009], the method of choice in this thesis, to address this challenge. Past results were typically heuristic, however, and when rigorous bounds are given, they are asymptotic in nature [Gerencsér et al., 2007, 2009]. In contrast, in this thesis we provide finite-time upper and lower bounds which validate the optimality of our approach.

More recently, much attention has been given to obtaining finite-time learning-theoretic-style bounds on the estimation of unknown parameters in a dynamical system, in particular linear systems [Tu et al., 2017, Faradonbeh et al., 2018, Hazan et al., 2018, Hardt et al., 2018, Simchowitz et al., 2018, Sarkar and Rakhlin, 2019, Oymak and Ozay, 2019, Simchowitz et al., 2019, Sarkar et al., 2019, Tsiamis and Pappas, 2019]. While the results in Part I of this thesis can be seen as building on this line of work in many respects, the focus of these works is the passive setting, and no attempt is made to choose inputs to speed up estimation, one goal of

this thesis. Beyond the linear setting, Foster et al. [2020a], Oymak [2019], and Sattar and Oymak [2022] provide formal guarantees on system identification in several different classes of nonlinear systems, yet they only consider noiseless systems, or systems that are significantly easier to excite than systems of the form considered in Chapter 6 (rendering the problem of exploration significantly easier). Of particular relevance to the results of Chapter 6 is Mania et al. [2022], which proposes an active learning approach to identify unknown parameters in (1.1.4), with the goal of minimizing the Euclidean distance in the parameter space. However, as we show, learning a uniformly good model could be significantly worse than learning a model with the goal task in mind, our primary focus. We also mention recent work applying deep learning-based approaches to system identification [Shi et al., 2019, Nguyen-Tuong and Peters, 2011, Brunke et al., 2022, Williams et al., 2017, Shi et al., 2021b], though such approaches lack principled guarantees.

2.1.2 Learning for Control

While the majority of this thesis focuses on the setting where exploration is decoupled from control, it is natural to also ask if we can simultaneously explore while controlling. That is, given an unknown environment, if it is possible to, from the start, attempt to solve the goal task while learning along the way. This is typically referred to as the "adaptive control" problem, and has been studied extensively in the controls literature [Slotine et al., 1991, Åström and Wittenmark, 2013, Feldbaum, 1960, Mesbah, 2018, Bristow et al., 2006, Richards et al., 2021, O'Connell et al., 2022, Boffi et al., 2021]. This line of work typically focuses on obtaining asymptotic results, and is primarily concerned with the stabilization problem, while we are interested in finite-time results, and more general objectives.

More recently, attention has been given towards quantifying the finite-time performance of online algorithms in terms of *regret*, comparing the online performance of an algorithm to the best possible performer in hindsight, where the goal is to minimize some general cost function. Much of the focus here has been on the LQR problem [Abbasi-Yadkori and Szepesvári, 2011, Dean et al., 2020, 2018, Mania et al., 2019, Dean et al., 2019, Cohen et al., 2019, Yu et al., 2020, Ziemann and Sandberg, 2021], with Simchowitz and Foster [2020] ultimately settling the minimax optimal regret in terms of dimension and time horizon. Others have considered regret in online adversarial settings [Agarwal et al., 2019, Simchowitz et al., 2020]. In the nonlinear setting, Kakade et al. [2020] study systems of the form considered in Chapter 6, and obtain sublinear regret guarantees.

2.2 Related Work in Reinforcement Learning

Next we focus on related work in the reinforcement learning setting, considered in Part II, as well as the general decision making setting considered in Part III.

2.2.1 Multi-Armed Bandits

While the main focus of this thesis is in long-horizon decision making problems ($H > 1$), many of our results build on, and are inspired by, the multi-armed bandits literature. The multi-armed bandits problem is a special case of the general RL problem with $H = 1$ and $S = 1$. At every step, the agent has A options (arms) available to them, chooses one of the A arms, and observes a noisy realization of the reward from that arm. While seemingly simple, the multi-armed bandit problem is applicable to a wide range of real-world problems, and has inspired a vast body of work. See Lattimore and Szepesvári [2020] for a complete overview; we highlight only works of particular relevance here.

A significant amount of attention has been devoted to obtaining instance-optimal PAC guarantees in the bandits literature—typically referred to as *best-arm identification*—for both multi-armed (tabular) and structured bandits [Kaufmann et al., 2016, Garivier and Kaufmann, 2016, Russo, 2016, Degenne and Koolen, 2019, Degenne et al., 2019, 2020a]. While many of these works are asymptotic in nature, in specialized settings such as multi-armed bandits [Jamieson et al., 2014, Kaufmann et al., 2016] and linear bandits [Fiez et al., 2019, Katz-Samuels et al., 2020], recent work has shown that the optimal rates are achievable in finite-time.

In the regret minimization setting, the focus of Part III, a variety of minimax regret guarantees have been obtained in settings such as tabular [Auer et al., 2002, Audibert and Bubeck, 2009] and linear bandits [Dani et al., 2008, Abbasi-Yadkori et al., 2011]. For instance-dependent guarantees, many works provide purely asymptotic instance-optimal guarantees for settings such as multi-armed bandits [Lai and Robbins, 1985, Garivier et al., 2016, Lattimore, 2018, Garivier et al., 2019], linear bandits [Lattimore and Szepesvari, 2017, Hao et al., 2020], and general structured bandits [Burnetas and Katehakis, 1996, Magureanu et al., 2014, Combes et al., 2017, Van Parys and Golrezaei, 2023, Degenne et al., 2020b]. While a small number of works [Tirinzoni et al., 2020, Jun and Zhang, 2020, Kirschner et al., 2021] do provide finite-time instance-optimal regret guarantees, they specialize to particular settings such as linear bandits, and it is not clear that their algorithmic approaches generalize.

We remark as well that a small amount of work in the bandits community has been devoted to understanding the complexity of "learning to explore", the focus of Part III of this thesis. While to the best of our knowledge our work is the first to study this question in the general function approximation setting, Simchowitz et al. [2017] and Chen et al. [2017] consider a similar question in the setting of top-k bandits, and Marinov et al. [2022b,a] consider this question in the graph bandits setting.

2.2.2 Tabular RL

Finite-time results on policy identification in tabular MDPs go back to at least the late 90s and early 2000s [Kearns and Singh, 1998, Kearns et al., 1999, Kearns and Singh, 2002, Brafman and Tennenholtz, 2002]. A seminal early work in this area is that of Kakade [2003], which establishes a variety of results, including the performance-difference lemma and

polynomial sample complexities in both the generative and online settings. This work was built upon and refined by a variety of other works over the following decade [Strehl et al., 2006, Auer et al., 2008, Munos and Szepesvári, 2008, Strehl et al., 2009], leading up to works such as [Lattimore and Hutter, 2012, Dann and Brunskill, 2015], which establish sample complexity bounds of $\mathcal{O}(S^2A \cdot \text{poly}(H)/\epsilon^2)$. More recently, [Dann et al., 2017, 2019, Ménard et al., 2021] have proposed algorithms which achieve the optimal dependence on S and A of $\mathcal{O}(SA \cdot \text{poly}(H)/\epsilon^2)$, with [Dann et al., 2019, Ménard et al., 2021] also achieving the optimal H dependence, thereby effectively solving the problem of near-optimal policy identification in the worst case.

The question of regret minimization is intimately related to that of policy identification. Indeed, as shown in Jin et al. [2018], any low-regret algorithm can be used to obtain a near-optimal policy via an online-to-batch conversion. Early examples of low-regret algorithms in tabular MDPs are [Auer and Ortner, 2006, Auer et al., 2008, Azar et al., 2017]. Of note is the work of Auer et al. [2008], which establishes a polynomial regret bound on average-reward MDPs, but that scales with the *diameter* of the MDP—the maximum average number of steps it will take to navigate from any state to any other state. A seminal recent work is that of Zanette and Brunskill [2019], which proposes the EULER algorithm and establishes diameter-free regret bounds scaling as $\mathcal{O}(\sqrt{\mathbb{Q}^* SAHK})$, for $\mathbb{Q}^*$ the maximum variance in next-state value function. Finally, the work of Zhang et al. [2021b] completely eliminates the horizon dependence (assuming properly normalized rewards, and time-homogenous transitions), obtaining regret guarantees which scale purely as $\mathcal{O}(\sqrt{SAK})$. This scaling is optimal, and essentially solves the problem of worst-case regret in tabular RL.

While the problem of obtaining worst-case optimal guarantees in tabular RL is nearly closed, we are only beginning to understand what types of instance-dependent guarantees are possible. In the setting of regret minimization, Ok et al. [2018] provide an algorithm which achieves asymptotically optimal instance-dependent regret. Simchowitz and Jamieson [2019] show that standard optimistic algorithms achieve regret bounded as $\mathcal{O}(\sum_{s,a,h} \frac{\log K}{\Delta_h(s,a)})$, reminiscent of the gap-dependent regret guarantees found in bandits. This result was later refined by [Xu et al., 2021, Dann et al., 2021]. Obtaining instance-dependent guarantees for policy identification has proved more difficult. While several works exist in this space [Zanette et al., 2019, Jonsson et al., 2020, Marjani and Proutiere, 2020, Al Marjani et al., 2021], they all exhibit crucial shortcomings, such as exponential dependence on H [Jonsson et al., 2020], requiring access to a generative model [Zanette et al., 2019, Marjani and Proutiere, 2020], or lacking finite time results and only handling *best* ($\epsilon = 0$) policy identification [Al Marjani et al., 2021]. In the special case of deterministic, tabular MDPs, Tirinzoni et al. [2022] show matching instance-dependent upper and lower bounds; a result that the approach we propose in Chapter 8 is essentially able to match.

2.2.3 RL with Function Approximation

To generalize the tabular setting, the RL community has considered *function approximation*, replacing the tabular model with more powerful models that allow for generalization across states. Such settings have been considered in classical works [Baird, 1995, Bradtke and Barto, 1996, Sutton et al., 1999, Melo and Ribeiro, 2007], yet these works do not provide polynomial sample complexities. More recently, there has been intense interest in obtaining polynomial complexities for general function classes [Jiang et al., 2017, Du et al., 2021, Jin et al., 2021, Foster et al., 2021, 2023], and, in particular, linear function classes [Yang and Wang, 2019, Jin et al., 2020b, Wang et al., 2019, Du et al., 2019, Zanette et al., 2020a,b, Ayoub et al., 2020, Jia et al., 2020, Weisz et al., 2021, Zhou et al., 2021a,b, Zhang et al., 2021c, Wang et al., 2021].

Several works of particular note in this category are [Jin et al., 2020b, Zanette et al., 2020b, Zhou et al., 2021a], which establish regret guarantees in the setting of linear MDPs (a setting considered in this thesis) and the related setting of linear mixture MDPs. Jin et al. [2020b] obtains regret scaling as $\mathcal{O}(\sqrt{d^3H^4K})$, one of the first polynomial complexity bounds in RL with function approximation, using a variant of Q-learning. Zanette et al. [2020b] improves this to $\mathcal{O}(\sqrt{d^2H^4K})$ in the slightly more general low inherent Bellman error setting, though their algorithm is not computationally efficient. Via an online-to-batch conversion, these algorithms achieve PAC complexities of $\mathcal{O}(d^3H^4/\epsilon^2)$ and $\mathcal{O}(d^2H^4/\epsilon^2)$, respectively. While Zanette et al. [2020b] show a lower bound proving that $\Omega(\sqrt{d^2K})$ regret is necessary, prior to the work of this thesis, no lower bounds for PAC RL existed in linear MDPs, and the question of optimal PAC guarantees remained open. In the setting of linear mixture MDPs, Zhou et al. [2021a] show a regret bound of $\mathcal{O}(\sqrt{d^2H^3K})$ and a matching lower bound, yielding one of the first provably tight and computationally efficient algorithms for RL with function approximation.

In the setting of RL with function approximation, prior to the work of this thesis, only a handful of existing works obtain guarantees that would be considered "instance-dependent". Wagenmaker et al. [2022a] (also by the author of this thesis) show a "first-order" regret bound of $\mathcal{O}(\sqrt{d^3H^3V_0^\star K})$, where $V_0^\star$ is the value of the optimal policy on the particular MDP under consideration. He et al. [2021] show that standard optimistic algorithms achieve regret guarantees of $\mathcal{O}(\frac{d^3H^5\log K}{\Delta_{\min}})$ and $\mathcal{O}(\frac{d^2H^5\log^3 K}{\Delta_{\min}})$ in the settings of linear MDPs and linear mixture MDPs, respectively, for $\Delta_{\min}$ the minimum value-function gap. While both these works do obtain instance-dependent results, the instance-dependence is rather coarse, depending on only a single parameter ($V_0^\star$ or $\Delta_{\min}$). One goal of this thesis (in particular Chapter 8) will instead be to obtain more refined instance-dependent guarantees. Several works in the general decision making framework considered in Part III, the DMSO framework, also obtain instance-dependent bounds, but these are purely asymptotic [Komiyama et al., 2015, Dong and Ma, 2022].

The focus of Chapter 9 is the reward-free RL setting. In the tabular setting, it is known that the optimal rate for reward-free RL is $\Theta(S^2A/\epsilon^2)$, a factor of S larger than the optimal

rate for standard PAC RL [Jin et al., 2020a, Ménard et al., 2021, Zhang et al., 2020, Wu et al., 2022]. In the linear MDP setting, existing works on reward-free RL [Zanette et al., 2020c, Wang et al., 2020, Zhang et al., 2021a] achieve complexities of $\mathcal{O}(d^3/\epsilon^2)$, though prior to the work of this thesis, it was not known if this is tight, or if the same separation between reward-free and PAC RL exists in the linear setting.

2.3 Related Work in Experiment Design

Experiment design is a well-developed subfield of statistics which seeks to understand how one should make measurements in order to collect information on some parameter of interest. Consider the basic linear regression setting:

$$z = \langle \theta_\star, x \rangle + w,$$

where we choose $x \in \mathcal{X}$, observe z, and w is some noise. Denote the *design matrix* as $\mathbf{A}(\lambda) := \sum_{x \in \mathcal{X}} \lambda_x x x^\top$, then experiment design optimizes over $\lambda \in \triangle_{\mathcal{X}}$, and chooses measurements $x \sim \lambda$, to optimize some desired objective. Common experiment design criteria include:

- **A-optimal design**: $\min_{\lambda \in \triangle_{\mathcal{X}}} \operatorname{tr}(\mathbf{A}(\lambda)^{-1})$, minimizes the average variance of the estimate of $\theta_\star$.
- **E-optimal design**: $\max_{\lambda \in \triangle_{\mathcal{X}}} \lambda_{\min}(\mathbf{A}(\lambda))$, minimizes the maximum prediction error in the estimate of $\theta_\star$.
- **G-optimal design**: $\min_{\lambda \in \triangle_{\mathcal{X}}} \max_{x \in \mathcal{X}} \|x\|^2_{\mathbf{A}(\lambda)^{-1}}$, minimizes the maximum prediction error for any direction $\mathcal{X}$ in the estimate of $\theta_\star$.
- **XY-optimal design**: $\min_{\lambda \in \triangle_{\mathcal{X}}} \max_{y \in \mathcal{Y}} \|y\|^2_{\mathbf{A}(\lambda)^{-1}}$, a variant of G-optimal design, minimizes the maximum prediction error for any direction $\mathcal{Y}$ in the estimate of $\theta_\star$.

A variety of other criteria exist, and in general these can be extended to nonlinear models as well (with the design matrix replaced by the Fisher information matrix). See Pukelsheim [2006] or Pronzato and Pázman [2013] for a complete overview.

Many of the methods proposed in this thesis can be viewed as experiment designs; indeed, the very question this thesis seeks to answer can be seen as a question of experiment design. In particular, in Chapter 3 we reduce the problem of linear system identification to E-optimal design, Chapters 5 and 6 reduce decision making in dynamical systems to a (weighted) A-optimal design, and Chapter 8 reduces policy learning in MDPs to an XY-design.

Our work deviates from the majority of the experiment design literature in several key ways, however. First, we will be interested primarily with *adaptive* experiment designs—while much of the experiment design literature focuses on *static* designs, which are fixed at the beginning of environment interaction, our methods will rely on designs which change over

time, as more information about the environment is collected. Second, the majority of work on experiment design focuses on single-step settings, while our focus in this thesis is on dynamic, long-horizon settings. Indeed, one contribution of this thesis is to generalize classical experiment design techniques to long-horizon settings.

We highlight several works on experiment design in sequential environments that are particularly relevant. First, the work of Fiez et al. [2019] achieves the instance-optimal rate for best-arm identification in linear bandits and relies on an adaptive XY experiment design-based. The approach of Fiez et al. [2019], as well as the related work of Soare et al. [2014], provides inspiration for the algorithm proposed in Chapter 8—in some sense the approach of Chapter 8 can be seen as a generalization of the RAGE algorithm of Fiez et al. [2019] to problems with horizon greater than 1. Concurrent to the publication of the results of Chapter 8, this algorithm was also extended to the contextual bandit setting in Li et al. [2022].

Also concurrent to the publication of the results of Chapter 8, the work of Mutny et al. [2023] propose an approach to solving experiment design problems in MDPs. To our knowledge, this is the only existing work that directly considers the problem of experiment design in MDPs. However, they make the simplifying assumption that the transition dynamics are *known*, which essentially reduces their problem to a computational one; in contrast, in this thesis, we focus on unknown dynamics. We remark as well that the online experiment design approach of Chapters 6 and 8 is somewhat related to several existing algorithms [Hazan et al., 2019, Zahavy et al., 2021]. Finally, Chaudhuri et al. [2015] gives a non-asymptotic active learning procedure for adaptive maximum likelihood estimation, again adapting the design to the unknown parameter of interest; their setting does not address long-horizon settings, however.

Part I

Active Learning for Continuous Control

Chapter 3

Active Identification of Linear Dynamical Systems

We begin with the system identification problem in the linear dynamical system (LDS) setting. In particular, in this chapter we will consider discrete-time linear dynamical systems with fully observed state of the form:

$$x_{t+1} = A_\star x_t + B_\star u_t + w_t, \quad x_0 \equiv 0 \tag{3.0.1}$$

for $A_\star \in \mathbb{R}^{d_x \times d_x}$ the system matrix, $B_\star \in \mathbb{R}^{d_x \times d_u}$ the control matrix, $x_t \in \mathbb{R}^{d_x}$ the state, $u_t \in \mathbb{R}^{d_u}$ the input, and $w_t \sim \mathcal{N}(0, \sigma_w^2 I)$ process noise. We consider the single trajectory setting: starting from state $x_0 = 0$, we interact with (3.0.1) T steps without resetting. We assume that $B_\star$ is known[1] but that $A_\star$ is unknown, and our goal is to estimate, to identify, $A_\star$ using the minimum number of interactions with (3.0.1) possible. In this chapter we are particularly interested in the *active* identification problem: if we are able to control the input u_t, which inputs should we play in order to estimate $A_\star$ as accurately as possible after T steps of interaction?

The estimation of the unknown parameters of a dynamical system typically falls under the domain of *system identification*. System identification is an entire field unto itself, and the problem of identification of linear dynamical systems has been studied for decades [Ljung, 1998]. Indeed, the particular question studied in this chapter—active system identification—has been considered by many past works, going back to the 1970s [Mehra, 1974, 1976, Goodwin and Payne, 1977, Jansson and Hjalmarsson, 2005, Gerencsér et al., 2007, 2009, Gevers et al., 2009, Manchester, 2010, Katselis et al., 2012, Larsson et al., 2012, Hägg et al., 2013]. These works, however, while providing a variety of different approaches to input design, typically lack formal guarantees. As such, it has remained an open question what the *optimal* achievable rate is for active identification of linear dynamical systems.

The results presented in this chapter are the first to answer this question, establishing the optimal rate for LDS identification in the active setting, and proposing an algorithm

[1] All results in this chapter can be easily extended to the case where $B_\star$ is unknown as well—see Chapter 5 for an example of this.

provably able to achieve this. Towards achieving this, the key technical contribution of this chapter is a novel bound on the estimation error of $A_\star$ when u_t is chosen to be a *periodic* input.

3.1 Preliminaries for Linear Dynamical Systems

Before stating our results, we provide several definitions and assumptions in the LDS setting. Defining $\rho(A)$ as the spectral radius of A—the magnitude of the largest magnitude eigenvalue of A—we make the following assumption on our dynamics:

Assumption 3.1. *Let Θ denote the set of all stable θ: $\Theta := \{\theta = (A, B) : \rho(A) < 1\}$. We assume $\theta_\star \in \Theta$, for $\theta_\star = (A_\star, B_\star)$.*

Assumption 3.1 ensures that, in the absence of control, our dynamics (3.0.1) do not explode. While this assumption is somewhat restrictive, we note that the learning problem is somewhat trivial in the case where the system is unstable, as the exploding state causes the signal-to-noise ratio to increase exponentially.

We also restrict the set of inputs the agent is allowed to play during exploration to those that have bounded power. Without this assumption, the agent may inject inputs of arbitrarily large magnitude into the system, again producing an arbitrarily large signal-to-noise ratio. More importantly, however, in real-world settings, arbitrarily large inputs are typically either impossible to realize, or introduce significant safety concerns. Formally:

Definition 3.1.1 (Power-Constrained Policies)**.** In the linear dynamical system setting, we allow the agent to play all exploration policies π such that $\mathbb{E}^\pi[\sum_{t=1}^T u_t^\top u_t] \le T\gamma^2$, for some $\gamma > 0$. We denote the set of all such policies as Π_{γ^2}.

The $\mathcal{H}_\infty$-norm of a system quantifies the maximum gain of the system (i.e. the maximum amplification of an input), and is a fundamental quantity in the controls literature. We denote the $\mathcal{H}_\infty$-norm of a system with parameter A as:

$$\|A\|_{\mathcal{H}_\infty} := \max_{\omega \in [0, 2\pi]} \|(e^{\iota\omega} I - A)^{-1}\|_{\text{op}},$$

and, for $\theta = (A, B)$, also define:

$$\|\theta\|_{\mathcal{H}_\infty} := \max_{\omega \in [0, 2\pi]} \|(e^{\iota\omega} I - A)^{-1} B\|_{\text{op}}.$$

Throughout, we will define β_∞ as

$$\beta_\infty := c \cdot \Big((1 + \|B_\star\|_{\text{op}}^2)\|A_\star\|_{\mathcal{H}_\infty}^4\Big)\Big(\sqrt{d_x} \log \frac{T}{\delta} + \gamma^2 \|A_\star\|_{\mathcal{H}_\infty}^2\Big).$$

In general the setting of T and δ will be clear from context when using β_∞.

3.1.1 Least Squares Estimation

In this chapter our focus will be on the least squares estimator. Given some trajectory $\boldsymbol{\tau} = (x_0, u_0, x_1, u_1, \ldots, u_T, x_{T+1})$ generated from (3.0.1), the least squares estimate of $A_\star$ is defined as:

$$\widehat{A} = \arg\min_{A} \sum_{t=1}^{T} \|x_{t+1} - Ax_t - B_\star u_t\|_2^2. \tag{3.1.1}$$

By standard analysis, we can write $\widehat{A}$ in closed form as $\widehat{A} = (\sum_{t=1}^T x_t x_t^\top)^{-1} x_t (x_{t+1} - B_\star u_t)^\top$. Thus,

$$\begin{aligned} \widehat{A} - A_\star &= (\textstyle\sum_{t=1}^T x_t x_t^\top)^{-1} x_t (x_{t+1} - B_\star u_t)^\top - A_\star \\ &= (\textstyle\sum_{t=1}^T x_t x_t^\top)^{-1} x_t (A_\star x_t + w_t)^\top - A_\star \\ &= (\textstyle\sum_{t=1}^T x_t x_t^\top)^{-1} x_t w_t^\top. \end{aligned} \tag{3.1.2}$$

Given this, we can bound the estimation error of $\widehat{A}$ as

$$\|\widehat{A} - A_\star\|_{\mathrm{op}} = \|(\textstyle\sum_{t=1}^T x_t x_t^\top)^{-1} x_t w_t^\top\|_{\mathrm{op}} \le \underbrace{\lambda_{\min}(\textstyle\sum_{t=1}^T x_t x_t^\top)^{-1/2}}_{(a)} \cdot \underbrace{\|(\textstyle\sum_{t=1}^T x_t x_t^\top)^{-1/2} x_t w_t^\top\|_{\mathrm{op}}}_{(b)}.$$

Our analysis relies on bounding each of these terms. To bound (b), we apply the following now-standard self-normalized bound with $z_t \leftarrow x_t$ and $\eta_t \leftarrow w_t$.

Proposition 3.1 (Abbasi-Yadkori et al. [2011], Sarkar and Rakhlin [2019]). *Let $(\mathcal{F}_t)_{t=1}^\infty$ be a filtration, $\eta_t \sim \mathcal{N}(0, \sigma^2 I_d)$, and $(z_t)_{t=1}^\infty$ a $\mathbb{R}^d$-valued stochastic process. Denote $V_T := \sum_{t=1}^T x_t x_t^\top + V_0$ for some $V_0 \succ 0$. Then with probability at least $1-\delta$, for all $T \ge 1$, we have:*

$$\left\| \left(\textstyle\sum_{t=1}^T z_t z_t \right)^{-1/2} \sum_{t=1}^T z_t \eta_t^\top \right\|_{\mathrm{op}} \le \sigma \sqrt{8 \log \frac{1}{\delta} + 4 \log \det(V_T V_0^{-1} + I) + 8d \log 5}.$$

In order to estimate $A_\star$ as quickly as possible, the above decomposition then suggests that we must choose our inputs to ensure that (a) is as small as possible—equivalently, that the minimum eigenvalue of the covariates is as large as possible. The primary technical challenge in our analysis comes from lower-bounding (a) when we are driving the system (3.0.1) with some general input signal, $\boldsymbol{u}$, and in showing that we can excite (3.0.1) with the optimal input without perfect knowledge of $A_\star$. We present our solutions to these challenges in Sections 3.2 and 3.3, but first provide some additional notation for the covariance matrices and input signals.

3.1.2 Frequency-Domain Representations and Covariance Matrices

We let bold vectors $\boldsymbol{u} = (u_i)_{i=1}^k \in \mathbb{C}^{kd_u}$ denote sequences of inputs. We say $\boldsymbol{u}$ is mean-zero if $\sum_{i=1}^k u_i = 0$, and denote the discrete-time Fourier transform (DFT) of $\boldsymbol{u}$ as:

$$\check{\boldsymbol{u}} = (\check{u}_\ell)_{\ell=1}^k = \mathfrak{F}(\boldsymbol{u}) \in \mathbb{C}^{kd_u}, \quad \text{where } \check{u}_\ell = \sum_{s=1}^k u_s \exp\left(\frac{2\pi\iota s}{k} \cdot \ell\right). \tag{3.1.3}$$

The mapping $\mathfrak{F}$ is invertible, though in general $\mathfrak{F}^{-1} : \mathbb{C}^{td_u} \to \mathbb{C}^{td_u}$. However, if our frequency-domain representation is *symmetric*, we have that the inverse DFT is purely real.

Definition 3.1.2 (Symmetric Signal)**.** We say that $\check{\boldsymbol{u}} = (\check{u}_\ell)_{\ell=1}^k \in \mathbb{C}^{kd_u}$ is *symmetric* if $\check{u}_\ell = \text{conj}(\check{u}_{k-\ell})$ for $\ell < k$ and $\check{u}_k$ is purely real, where $\text{conj}(\cdot)$ denotes the complex conjugate.

Fact 1. *$\mathfrak{F}^{-1}(\check{u})$ is a vector with real coefficients if and only if $\check{u}$ is symmetric.*

We now consider a convex relation of the outerproduct of this DFT. First, some preliminaries. For a complex vector $z \in \mathbb{C}^d$ (resp. matrix $A \in \mathbb{C}^{d_1 \times d_2}$), let z^{H} (resp. A^{H}) denote its Hermitian adjoint; i.e., the complex conjugate of its transpose. We denote the set of *Hermitian* matrices as $\mathbb{H}^d := \{A \in \mathbb{C}^{d\times d} : A^{\mathsf{H}} = A\}$, and the set of positive-semidefinite Hermitian matrices $\mathbb{H}_+^d := \{A \in \mathbb{H}^d : z^{\mathsf{H}} A z \geq 0, \quad \forall z \in \mathbb{C}^d\}$. Given $\check{u} = (\check{u}_\ell)_{\ell=1}^k \in \mathbb{C}^{kd_u}$, we define its outerproduct as the sequence of complex-rank one Hermitian matrices, $\boldsymbol{U} = (U_\ell)_{\ell=1}^k$, defined by

$$\check{\boldsymbol{u}} \otimes \check{\boldsymbol{u}} := \boldsymbol{U} = (U_\ell)_{\ell=1}^k, \quad \text{where} \quad U_\ell = \check{u}_\ell \check{u}_\ell^{\mathsf{H}} \in \mathbb{H}_+^d.$$

We now define the following set, which relaxes outer products to matrix sequences of the above form, with a total power constraint on their trace:

$$\mathcal{U}_{\gamma^2,k} := \left\{ \boldsymbol{U} = (U_\ell)_{\ell=1}^k \ : \ U_\ell \in \mathbb{H}_+^d, \quad \boldsymbol{U} \text{ is symmetric}, \quad \sum_{\ell=1}^k \text{tr}(U_\ell) \leq k^2\gamma^2 \right\}.$$

Critically, $\mathcal{U}_{\gamma^2,k}$ is convex. We also define $\mathcal{U}_{\gamma^2,k}^0$ to be the restriction of $\mathcal{U}_{\gamma^2,k}$ to signals with $U_k = 0$, corresponding to zero-mean signals. We generalize the definition of symmetric signals here to matrices, defining it identically as we have defined symmetric vector signals. The following class of sequences $\boldsymbol{U}$ are of particular importance.

Definition 3.1.3 (Rank One Relaxation)**.** We say that $\boldsymbol{U} = \{U_\ell : 1 \leq \ell \leq k\} \in \mathcal{U}_{\gamma^2,k}$ is *rank one* if there exists a vector $\check{u} \in \mathbb{C}^{kd_u}$ such that $\boldsymbol{U} = \check{u} \otimes \check{u}$.

Lastly, we define a frequency-domain covariance operator defined on $\boldsymbol{U} \in \mathcal{U}_{\gamma^2,k}$, for $\theta = (A, B)$:

$$\boldsymbol{\Lambda}_k^{\text{freq}}(\theta, \boldsymbol{U}) := \frac{1}{k} \sum_{\ell=1}^k (e^{\iota \frac{2\pi\ell}{k}} I - A)^{-1} B U_\ell B^{\mathsf{H}} (e^{\iota \frac{2\pi\ell}{k}} I - A)^{-\mathsf{H}},$$

$$\mathbf{\Lambda}_{t,k}^{\mathrm{freq}}(\theta, \boldsymbol{U}) := \frac{t}{k}\mathbf{\Lambda}_{k}^{\mathrm{freq}}(\theta, \boldsymbol{U}).$$

We will overload notation, defining

$$\begin{aligned}\mathbf{\Lambda}_{k}^{\mathrm{freq}}(\theta, \check{\boldsymbol{u}}) &= \mathbf{\Lambda}_{k}^{\mathrm{freq}}(\theta, \check{\boldsymbol{u}} \otimes \check{\boldsymbol{u}}), \quad \mathbf{\Lambda}_{t,k}^{\mathrm{freq}}(\theta, \check{\boldsymbol{u}}) = \mathbf{\Lambda}_{t,k}^{\mathrm{freq}}(\theta, \check{\boldsymbol{u}} \otimes \check{\boldsymbol{u}}) \\ \mathbf{\Lambda}_{k}^{\mathrm{freq}}(\theta, \boldsymbol{u}) &= \mathbf{\Lambda}_{k}^{\mathrm{freq}}(\theta, \check{\boldsymbol{u}}), \quad \mathbf{\Lambda}_{t,k}^{\mathrm{freq}}(\theta, \boldsymbol{u}) = \mathbf{\Lambda}_{t,k}^{\mathrm{freq}}(\theta, \check{\boldsymbol{u}}).\end{aligned}$$

for $\boldsymbol{u} = \mathfrak{F}^{-1}(\check{\boldsymbol{u}})$. The following result shows that $\mathbf{\Lambda}_{k}^{\mathrm{freq}}(\theta, \boldsymbol{u})$ corresponds to the *steady-state* covariates of our system when an input $\boldsymbol{u}$ is played.

Proposition 3.2. *Let $\boldsymbol{U} \in \mathcal{U}_{\gamma^2,k}$ be rank one, with $\boldsymbol{U} = \check{\boldsymbol{u}} \otimes \check{\boldsymbol{u}}$, $\check{\boldsymbol{u}} \in \mathbb{C}^{kd_u}$. Let $\boldsymbol{u} = (u_t)_{t=1}^{k} = \mathfrak{F}^{-1}(\check{\boldsymbol{u}})$. Define the extended inputs*

$$\boldsymbol{u}_{1:T}^{\mathrm{ext}} = (u_t^{\mathrm{ext}})_{t\geq 1}, \text{ where } u_t^{\mathrm{ext}} = u_{\mathrm{mod}(t,k)}.$$

Finally, let $x^{\boldsymbol{u}}$ denote the evolution of (3.0.1) when executing the input u_t^{ext}, and with $w_t = 0$ for all t. Then

1. $\frac{1}{k}\mathbf{\Lambda}_{k}^{\mathrm{freq}}(\theta, \boldsymbol{U}) = \lim_{T\to\infty} \frac{1}{T}\sum_{s=0}^{T-1} x_s^{\boldsymbol{u}}(x_s^{\boldsymbol{u}})^\top = \lim_{T\to\infty} \frac{1}{T}\mathbf{\Lambda}_{T}^{\mathrm{in}}(\theta, \boldsymbol{u}^{\mathrm{ext}}, x_0)$.

2. *Let $k' \geq k$ be divisible by k. Let $\boldsymbol{u}' = (u_s^{\mathrm{ext}})_{s=1}^{k'}$, and define the frequency domain quantities*

$$\check{\boldsymbol{u}}' = (\check{u}'_\ell)_{\ell=1}^{k'} = \mathfrak{F}(\boldsymbol{u}'), \quad \boldsymbol{U}' = \check{\boldsymbol{u}}' \otimes \check{\boldsymbol{u}}'.$$

   Then $\mathbf{\Lambda}_{k'}^{\mathrm{freq}}(\theta, \boldsymbol{U}') = \frac{k'}{k}\mathbf{\Lambda}_{k}^{\mathrm{freq}}(\theta, \boldsymbol{U})$.

Noise-Augmented Covariances. We shall also study the covariance matrix that arises from exciting the system with white noise of covariance $\mathbf{\Lambda}_u$, when the process noise has covariance $\mathbf{\Lambda}_w$:

$$\mathbf{\Lambda}_{t}^{\mathrm{noise}}(\theta, \mathbf{\Lambda}_u) := \sum_{s=0}^{t-1} A^s \mathbf{\Lambda}_w (A^s)^\top + \sum_{s=0}^{t-1} A^s B \mathbf{\Lambda}_u B^\top (A^s)^\top.$$

We will overload notation and set

$$\mathbf{\Lambda}_{t}^{\mathrm{noise}}(\theta, \sigma_u) := \mathbf{\Lambda}_{t}^{\mathrm{noise}}(\theta, \sigma_u^2 I).$$

We also define

$$\lambda_{\mathrm{noise}}(\sigma_u) := \lambda_{\min}(\mathbf{\Lambda}_{d_x}^{\mathrm{noise}}(\theta_\star, \sigma_u)), \quad \lambda_{\mathrm{noise}} := \lambda_{\min}(\mathbf{\Lambda}_{d_x}^{\mathrm{noise}}(\theta_\star, \gamma/\sqrt{2d_u}))$$

for $\theta_\star = (A_\star, B_\star)$. Note that

$$\lambda_{\text{noise}}(\sigma_u) = \lambda_{\min}\left(\mathbb{E}\left[x_{d_x} x_{d_x}^\top \mid u_s \sim \mathcal{N}(0, \sigma_u^2 I), s = 0, \ldots, t\right]\right)$$

so it follows that $\lambda_{\text{noise}}(\sigma_u)$ is the minimum eigenvalues of the covariates when we play isotropic noise, and can be thought of as a measure of how easily the system can be excited.

Since the Fourier transform preserves Gaussianity, the relevant covariance matrices become:

$$\begin{aligned}
\mathbf{\Lambda}_{T,k}^{\text{ss}}(\theta, \boldsymbol{U}, \sigma_u) &:= \frac{1}{k}\mathbf{\Lambda}_k^{\text{freq}}(\theta, \boldsymbol{U}) + \frac{1}{T}\sum_{t=1}^{T} \mathbf{\Lambda}_t^{\text{noise}}(\theta, \sigma_u), \\
\mathbf{\Lambda}_k^{\text{ss}}(\theta, \boldsymbol{U}, \sigma_u) &:= \mathbf{\Lambda}_{k,k}^{\text{ss}}(\theta, \boldsymbol{U}, \sigma_u).
\end{aligned}$$

We will also overload notation and denote $\mathbf{\Lambda}_k^{\text{ss}}(A, \boldsymbol{U}, \sigma_u) := \mathbf{\Lambda}_k^{\text{ss}}((A, B_\star), \boldsymbol{U}, \sigma_u)$. If $\boldsymbol{U}$ is rank one, then Proposition 3.2 implies that $\mathbf{\Lambda}_{T,k}^{\text{ss}}(\theta, \boldsymbol{U}, \sigma_u)$ corresponds to the expected steady-state covariates of the noisy system when playing inputs $\boldsymbol{U}$. If $\boldsymbol{U}$ is not rank one, then $\mathbf{\Lambda}_{T,k}^{\text{ss}}(\theta, \boldsymbol{U}, \sigma_u)$ corresponds to the expected steady-state covariates of the noisy system when playing a sequence of inputs formed by decomposing $\boldsymbol{U}$ into rank one inputs, as in Algorithm 3.2.

3.2 Learning with Periodic Inputs in Linear Dynamical Systems

The primary technical contribution of this chapter is the following result, which quantifies the scaling of the covariates in a particular direction v when (3.0.1) is being driven by some periodic input signal $\boldsymbol{u} = (u_t)_{t=1}^k$, that is, when $\boldsymbol{u}$ is played repeatedly on (3.0.1).

Proposition 3.3. *Consider any $v \in \mathcal{S}^{d-1}$ and let x_t evolve according to the dynamical system (3.0.1) under mean-zero input $\boldsymbol{u}$. Let $T_{\text{ss}} \geq 0$ be a value large enough that, for any $T \geq 0$:*

$$\left| \sum_{t=T_{\text{ss}}+T+1}^{T_{\text{ss}}+T+k} (v^\top x_t^{\boldsymbol{u}} - v^\top \bar{x}_{T_{\text{ss}}+T:k}^{\boldsymbol{u}})^2 - v^\top \mathbf{\Lambda}_k^{\text{freq}}(A_\star, \boldsymbol{u}) v \right| \leq \frac{1}{10} v^\top \mathbf{\Lambda}_k^{\text{freq}}(A_\star, \boldsymbol{u}) v \tag{3.2.1}$$

where $\bar{x}_{T:k}^{\boldsymbol{u}} := \frac{1}{k}\sum_{t=T+1}^{T+k} x_t^{\boldsymbol{u}}$. Then we have that

$$\mathbb{P}\left[\sum_{t=T_{\text{ss}}+1}^{T_{\text{ss}}+T} (v^\top x_t)^2 \leq \frac{2}{81} \lfloor T/k \rfloor v^\top \mathbf{\Lambda}_k^{\text{freq}}(A_\star, \boldsymbol{u}) v \right] \leq e^{-\frac{2}{81}\lfloor T/k \rfloor}. \tag{3.2.2}$$

The proof of Proposition 3.3 is somewhat technical, and we defer it to Section 3.4. Given this result and a careful union bound, we obtain our main lower bound on the covariates, $\sum_{t=1}^T x_t x_t^\top$.

Lemma 3.2.1. *Let T_{ss} be the time such that condition (3.2.1) in Proposition 3.3 is met for all $v \in \mathcal{V}$, where*

$$\mathcal{V} = \left\{ v \in \mathcal{S}^{d-1} \ : \ v^\top \mathbf{\Lambda}_k^{\mathrm{noise}}(A_\star, \sigma_u) v \le \frac{1}{k} v^\top \mathbf{\Lambda}_k^{\mathrm{freq}}(A_\star, \boldsymbol{u}) v \right\}.$$

On the event that $\{\sum_{t=1}^{T} x_t x_t^\top \preceq T\bar{\mathbf{\Lambda}}_T\}$ for some $\bar{\mathbf{\Lambda}}_T$, as long as

$$T \ge 2T_{\mathrm{ss}} + c \cdot k \left(d_x + \log\det(\bar{\mathbf{\Lambda}}_T \mathbf{\Lambda}_k^{\mathrm{ss}}(A_\star, \boldsymbol{u}, \sigma_u)^{-1}) + \log\frac{1}{\delta} \right), \tag{3.2.3}$$

we have that, with probability less than δ:

$$\sum_{t=1}^{T} x_t x_t^\top \not\succeq c \cdot T \mathbf{\Lambda}_k^{\mathrm{ss}}(A_\star, \boldsymbol{u}, \sigma_u).$$

By Lemma 3.2.1, Proposition 3.1, and our error decomposition of the least-squares estimator in (3.1.2), we obtain the following result.

Theorem 3.1. *Consider some mean-zero input $\boldsymbol{u} = (u_t)_{t=1}^k$ with period k and satisfying $\sum_{t=1}^k u_t^\top u_t \le k\gamma^2$. Consider playing $\boldsymbol{u}$ repeatedly on (3.0.1) for T steps, and let $\widehat{A}_T$ denote the least-squares estimator of $A_\star$ as given in (3.1.1). Then, as long as*

$$T \ge 2T_{\mathrm{ss}}\left(\frac{1}{10}\lambda_{\mathrm{noise}}(0), k, \sqrt{\beta_\infty T} \right) + c \cdot k \left(d_x \log\frac{\beta_\infty}{\lambda_{\mathrm{noise}}(0)} + \log\frac{1}{\delta} \right),$$

we have that with probability at least $1-\delta$:

$$\|\widehat{A}_T - A_\star\|_{\mathrm{op}} \le C\sigma_w \sqrt{\frac{\log\frac{1}{\delta} + d_x \log\frac{\beta_\infty}{\lambda_{\mathrm{noise}}(0)}}{T \cdot \lambda_{\min}(\mathbf{\Lambda}_k^{\mathrm{ss}}(A_\star, \boldsymbol{u}, 0))}}.$$

We defer the full proof of this result to Section 3.4. Here, $T_{\mathrm{ss}}(\frac{1}{10}\lambda_{\mathrm{noise}}(0), k, \sqrt{\beta_\infty T}) = \mathcal{O}(\|A_\star\|_{\mathcal{H}_\infty}^2 \cdot \log\frac{kT\beta_\infty}{\lambda_{\mathrm{noise}}(0)})$, which we show is a sufficiently large value to ensure that condition (3.2.1) in Proposition 3.3 is met. Note, critically, the $\mathbf{\Lambda}_k^{\mathrm{ss}}(A_\star, \boldsymbol{u}, 0)$ term in the denominator. This term quantifies how the estimation error scales in terms of the interaction between the input and the system—an input $\boldsymbol{u}$ which produces covariates with a larger minimum eigenvalue will achieve a faster rate of estimation, according to Theorem 3.1.

3.3 Optimal Identification of Linear Dynamical Systems

The preceding section provides rates on estimating $A_\star$ when a *fixed* input $\boldsymbol{u}$ is played. In order to identify $A_\star$ as quickly as possible, we wish to first determine, and then play, the *best*

input; that is, the input which achieves the fastest rate of estimation. From Theorem 3.1, we see that we can minimize the estimation error by choosing an input that induces covariates with the maximum possible minimum eigenvalue. The challenge, however, is that, since the system parameters are unknown a priori, we do not know how a given input will excite the system. Our algorithmic approach, given in Algorithm 3.1, adapts the input as more information is gained to ensure the system is optimally excited.

Algorithm 3.1 Optimal Active LDS Identification

1: **Input:** Input power γ^2, initial epoch length $C_{\text{init}} d_u$ ($C_{\text{init}} \in \mathbb{N}$)
2: $T_0 \leftarrow C_{\text{init}} d_u$, $k_0 \leftarrow C_{\text{init}}$, $T \leftarrow T_0$
3: Run system for T_0 steps with $u_t \sim \mathcal{N}(0, \frac{\gamma^2}{d_u} I)$
4: **for** $i = 1, 2, 3, \ldots$ **do**
5: $\quad \widehat{A}_{i-1} \leftarrow \arg\min_A \sum_{t=1}^{T} \|x_{t+1} - Ax_t - B_\star u_t\|_2^2$
6: $\quad T_i \leftarrow T_0 2^i$, $k_i \leftarrow k_0 2^{\lfloor i/4 \rfloor}$, $T \leftarrow T + T_i$
7: $\quad$ Compute input $\boldsymbol{U}_i \leftarrow \arg\max_{\boldsymbol{U} \in \mathcal{U}^0_{\gamma^2/2, k_i}} \lambda_{\min}\left(\boldsymbol{\Lambda}^{\text{ss}}_{k_i}(\widehat{A}_{i-1}, \boldsymbol{U}, \frac{\gamma}{\sqrt{2d_u}})\right)$
8: $\quad (\widetilde{u}^i_t)_{t=1}^{T_i} \leftarrow \texttt{ConstructTimeInput}(\boldsymbol{U}_i, T_i/d_u, k_i)$
9: $\quad$ Run system for T_i steps with $u_t = \widetilde{u}^i_t + u^w_t$, $u^w_t \sim \mathcal{N}(0, \frac{\gamma^2}{2d_u} I)$

Algorithm 3.1 proceeds in epochs, with epoch length increasing exponentially. At each epoch it first obtains an estimate of $A_\star$ using the data collected at previous epochs, then, given this estimate, computes the inputs which will produce the covariates with the largest minimum eigenvalue on the estimated system. After computing these inputs, it plays them for the full epoch and the true system, and the process repeats. We see that Algorithm 3.1 plays inputs from the set $\mathcal{U}_{\gamma^2/2, k_i}$—that is, inputs with period k_i at epoch i. As k_i increases with i, the inputs can be made increasingly expressive. We show that, for large enough k_i, the set $\mathcal{U}_{\gamma^2/2, k_i}$ contains an optimal input.

Computational Efficiency of Algorithm 3.1. Note that the set $\mathcal{U}^0_{\gamma^2/2,k}$ is convex, $\lambda_{\min}(X)$ is concave, and $\boldsymbol{\Lambda}^{\text{ss}}_{k_i}(\widehat{A}_{i-1}, \boldsymbol{U}, \frac{\gamma}{\sqrt{2d_u}})$ is affine in $\boldsymbol{U}$. It follows that the input optimization on Line 7 is a convex optimization problem, and can therefore be efficiently solved by standard SDP solvers. As this is the primary computational burden of Algorithm 3.1, we conclude that Algorithm 3.1 is computationally efficient.

The `ConstructTimeInput` Subroutine. For the constraint set $\mathcal{U}_{\gamma^2/2,k}$ on the input design problem of Line 7 to be convex, we allow our input set to contain inputs that are not rank one. While this relaxation allows our input design problem to be efficiently solved, for a given $\boldsymbol{U} \in \mathcal{U}_{\gamma^2/2,k}$ that is not rank one, it is not clear if $\boldsymbol{U}$ can be implemented in the time domain. Indeed, Proposition 3.2 shows that, if $\boldsymbol{U}$ is rank one, there exists some time domain input

$\boldsymbol{u} = (u_t)_{t=1}^k$ such that

$$\frac{1}{k}\boldsymbol{\Lambda}_k^{\text{freq}}(\theta, \boldsymbol{U}) = \lim_{T\to\infty} \frac{1}{T}\sum_{t=0}^{T-1} x_s^{\boldsymbol{u}}(x_s^{\boldsymbol{u}})^\top$$

which implies that we can approximately realize the response covariates $\boldsymbol{\Lambda}_{T,k}^{\text{ss}}(\theta, \boldsymbol{U}, \gamma/\sqrt{2d_u})$ in the time domain, but this relationship no longer holds if $\boldsymbol{U}$ is not rank one. To remedy this, we propose the following procedure, which decomposes an arbitrary, not necessarily rank one, input $\boldsymbol{U}$ into a sequence of inputs that can be realized in the time domain.

Algorithm 3.2 `ConstructTimeInput`$(\boldsymbol{U}, T, k)$

Denote eigendecompositions $U_\ell = \sum_{j=1}^{d_u} \lambda_{\ell,j} v_{\ell,j} v_{\ell,j}^{\mathsf{H}}, \ell = 1, \dots, k, U_\ell \in \boldsymbol{U}$
$u_t = \mathbf{0} \in \mathbb{R}^{d_u}$ for $t = 1, \dots, d_u T$
for $j = 1, \dots, d_u$ **do**
    $\check{u}_{\ell,j} \leftarrow \sqrt{d_u \lambda_{\ell,j}} v_{\ell,j}$ for $\ell = 1, \dots, k$
    for $n = 1, \dots, T/k$ **do**
        $u_{(j-1)T+(n-1)k+1}, \dots, u_{(j-1)T+nk} \leftarrow \mathfrak{F}^{-1}(\check{u}_{1,j}, \dots, \check{u}_{k,j})$
return $u_1, \dots, u_{d_u T}$

As the following result shows, `ConstructTimeInput` produces a time domain input which realizes the response covariates $\boldsymbol{\Lambda}_k^{\text{freq}}(\theta, \boldsymbol{U})$ for arbitrary $\boldsymbol{U}$.

Proposition 3.4. *Let $\boldsymbol{U} \in \mathcal{U}_{\gamma^2,k}$ not necessarily rank one. Let $\boldsymbol{u}_m = (u_t)_{t=1}^{d_u mk}$ denote the time-domain input returned by calling* `ConstructTimeInput`$(\boldsymbol{U}, mk, k)$ *with m an integer. Then*

$$\frac{1}{k}\boldsymbol{\Lambda}_k^{\text{freq}}(\theta, \boldsymbol{U}) = \lim_{m\to\infty} \frac{1}{d_u mk}\sum_{t=0}^{d_u mk} x_t^{\boldsymbol{u}_m}(x_t^{\boldsymbol{u}_m})^\top$$

and, furthermore, the input satisfies $\sum_{t=1}^{d_u mk} u_t^\top u_t \le d_u mk\gamma^2$, where here $x_t^{\boldsymbol{u}}$ denotes the response of the noiseless dynamical system to input $\boldsymbol{u}$.

Note that the power constraint on the time-domain realization of $\boldsymbol{U}$ given in Proposition 3.4 ensures that the inputs played by Algorithm 3.1 live in Π_{γ^2}, thus satisfying the power constraint given in Definition 3.1.1.

3.3.1 Active Identification of Linear Dynamical Systems

Let:

$$\lambda^\star_{\min} := \limsup_{k\to\infty} \sup_{\boldsymbol{U}\in\mathcal{U}_{\gamma^2,k}} \lambda_{\min}(\boldsymbol{\Lambda}^{\mathrm{ss}}_k(A_\star, \boldsymbol{U}, 0)).$$

That is, $\lambda^\star_{\min}$ denotes the largest achievable steady-state minimum eigenvalue on (3.0.1), given that we are playing inputs in the set $\mathcal{U}_{\gamma^2,k}$. We then have the following result.

Theorem 3.2. *Assume that*

$$T \geq \mathrm{poly}\left(\gamma^2, \sigma_w^2, \|A_\star\|_{\mathcal{H}_\infty}, \|B_\star\|_{\mathrm{op}}, \frac{1}{\lambda_{\mathrm{noise}}}, \log\frac{1}{\delta}\right).$$

Then with probability at least $1-\delta$, the estimate $\widehat{A}_T$ produced after T steps of running Algorithm 3.1 satisfies:

$$\|\widehat{A}_T - A_\star\|_{\mathrm{op}} \leq C\cdot\sigma_w\sqrt{\frac{\log 1/\delta + d_x\log(\beta_\infty/\lambda_{\mathrm{noise}}+1)}{\lambda^\star_{\min}\cdot T}},$$

and, furthermore, the inputs $(u_t)_{t=1}^T$ played by Algorithm 3.1 satisfy $\sum_{t=1}^T \mathbb{E}[u_t^\top u_t] \leq T\gamma^2$.

Theorem 3.2 states that Algorithm 3.1 estimates $A_\star$ at a rate which scales with $\lambda^\star_{\min}$, the *best possible* minimum eigenvalue achievable under the desired input power constraint. The proof of Theorem 3.2 is given in Section 3.5. While Theorem 3.2 achieves a rate scaling with $\lambda^\star_{\min}$, we have not yet shown that this is *necessary*. We next show that this is indeed the case—the rate of estimating $A_\star$ is fundamentally limited by $\lambda^\star_{\min}$. Towards formalizing this, we introduce the following definition.

Definition 3.3.1 ((ϵ,δ)-Locally-Stable Algorithm)**.** We call an algorithm (ϵ,δ)-locally-stable in A if there exists a finite time $T_{\epsilon\delta}$ such that for all $t \geq T_{\epsilon\delta}$ and all $A' \in \mathcal{B}(A, 3\epsilon)$: $\mathbb{P}^{A'}[\|\widehat{A}_t - A'\|_2 \leq \epsilon] \geq 1-\delta$, where here $\mathbb{P}^{A'}$ is the measure induced when the true matrix is A', $\mathcal{B}(A,3\epsilon) := \{A' \in \mathbb{R}^{d\times d} \ : \ \|A - A'\|_{\mathrm{op}} \leq 3\epsilon\}$, and $\widehat{A}_t$ is the estimate obtained by the algorithm after t observations

Intuitively, a decision rule is (ϵ,δ)-locally stable in A if, for every A' "near" A, our decision rule is able to produce an estimate $\widehat{A}_t$ close to A'—that is, for each A', it can correctly solve the estimation problem. We then have the following result.

Theorem 3.3. *For any algorithm (ϵ,δ)-locally-stable in $A_\star$ and which plays inputs in Π_{γ^2}, as long as ϵ,δ satisfy $\frac{1}{\epsilon^2}\cdot\log\frac{1}{\delta} \geq \mathrm{poly}(d_x, d_u, \|A_\star\|_{\mathcal{H}_\infty}, \|B_\star\|_{\mathrm{op}}, \frac{1}{\lambda^\star_{\min}}, \gamma^2, \sigma_w^2)$, we must have:*

$$T_{\epsilon\delta} \geq \frac{1}{\lambda^\star_{\min}}\cdot\frac{\sigma_w^2}{32\epsilon^2}\log\frac{1}{2.4\delta}.$$

Rearranging Theorem 3.2, we see that to ensure $\|\widehat{A}_T - A_\star\|_{\text{op}} \leq \epsilon$ with probability at least $1-\delta$, it suffices if

$$T \geq \frac{C}{\lambda^\star_{\text{min}}} \cdot \frac{\sigma_w^2}{\epsilon^2}\left(\log\frac{1}{\delta} + d_x \log(\beta_\infty/\lambda_{\text{noise}} + 1)\right).$$

Thus, for small enough δ, Theorem 3.2 requires only that $T \geq \frac{C}{\lambda^\star_{\text{min}}} \cdot \frac{\sigma_w^2}{\epsilon^2} \log\frac{1}{\delta}$ in order to produce an ϵ-good estimate of $A_\star$, which, up to constants, precisely matches the rate of Theorem 3.3. Therefore, for small enough δ and goal tolerance ϵ, Theorem 3.2 and therefore Algorithm 3.1 are unimprovable. Note that this optimality holds in a very strong instance-dependent sense. For *any* system $(A_\star, B_\star)$ satisfying Assumption 3.1, Theorem 3.3 provides a lower bound on the sample complexity required to estimate $A_\star$ up to tolerance ϵ, and Theorem 3.2 matches this complexity.

3.3.2 Interpreting the Results

Does active system identification yield an improvement over passive identification? To answer this question, we consider the system with $A_\star = V\Gamma V^\top$ for orthogonal V, real, diagonal $\Gamma \succeq 0$, and $B_\star = I$. Denote the eigenvalues of $A_\star$ as $1 > \lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_d$ and $\lambda = [\lambda_1, \lambda_2, ..., \lambda_d]$. For simplicity assume that $\gamma^2/d_x \geq \sigma_w^2$. The following result compares the performance of Algorithm 3.1 with two approaches which rely on random noise to excite the system.

Corollary 3.4. *Assume that δ is chosen so that $\log\frac{1}{\delta} \gtrsim d_x$. Then for ϵ small enough, Algorithm 3.1 will obtain an estimate $\widehat{A}_T$ satisfying $\|\widehat{A}_T - A_\star\|_{\text{op}} \leq \epsilon$ with probability at least $1-\delta$ for*

$$T = \mathcal{O}\left(\frac{\|\mathbf{1}-\lambda\|_2^2}{\gamma^2} \cdot \frac{\sigma_w^2}{\epsilon^2} \cdot \log\frac{1}{\delta}\right).$$

In contrast, to achieve $\|\widehat{A}_T - A_\star\|_{\text{op}} \leq \epsilon$ with probability $1-\delta$ playing input $u_t \sim \mathcal{N}(0, \frac{\gamma^2}{d_x} I)$, we must have

$$T \geq \Omega\left(\frac{d_x|1-\lambda_d|}{\gamma^2} \cdot \frac{\sigma_w^2}{\epsilon^2} \cdot \log\frac{1}{\delta}\right),$$

and if playing $u_t \sim \mathcal{N}(0, \Sigma)$ for the optimal Σ ensuring our inputs are in Π_{γ^2}, we must have

$$T \geq \Omega\left(\frac{\|\mathbf{1}-\lambda\|_1}{\gamma^2} \cdot \frac{\sigma_w^2}{\epsilon^2} \cdot \log\frac{1}{\delta}\right).$$

From Corollary 3.4, we see that active identification can indeed yield a significant improvement over approaches which are passive, or excite the system with only noise. Indeed, consider the case when $\lambda_i = 1 - 1/d_x$. Then $\|\mathbf{1}-\lambda\|_2^2 = \frac{1}{d_x}$, while $d_x|1-\lambda_d| = \|\mathbf{1}-\lambda\|_1 = 1$,

so Algorithm 3.1 will yield an improvement of a factor of the dimensionality as compared to passive approaches, or approaches which only excite the system via random noise (even the optimal choice of random noise). Note that these approaches would still yield "persistent excitation", ensuring that all modes of the system are excited throughout the trajectory. However, this result shows that, while persistent excitation may be *sufficient* for learning, it can be far from optimal, and that to achieve the optimal rate we must instead tune our input to optimally excite the system.

3.4 Proof of Theorem 3.1

We now prove Theorem 3.1, our main estimation error bound. We start with some additional notation. Throughout this section we will consider inputs $\boldsymbol{u} = (u_t)_{t=0}^{k-1}$ of period k. For $t > k$, we overload notation and let $u_t := u_{\mathrm{mod}(t,k)}$. We will sometimes break up the state into the portion driven by the input $\boldsymbol{u}$, $x_t^{\boldsymbol{u}}$, and the portion driven by the noise, $x_t^{\boldsymbol{w}}$. In particular, we have

$$x_{t+1}^{\boldsymbol{u}} = A_\star x_t^{\boldsymbol{u}} + B_\star u_t, \quad x_{t+1}^{\boldsymbol{w}} = A_\star x_t^{\boldsymbol{w}} + w_t.$$

Due to linearity, $x_t = x_t^{\boldsymbol{u}} + x_t^{\boldsymbol{w}}$. We let $x^{\boldsymbol{u},\mathrm{ss}}$ denote the *steady-state* response. That is, $x_t^{\boldsymbol{u},\mathrm{ss}} = \lim_{i\to\infty} x_{ik+\mathrm{mod}(t,k)}^{\boldsymbol{u}}$, the response of (3.0.1) to playing $\boldsymbol{u}$ once the system has fully mixed and forgotten the initial condition. Unless otherwise noted, we assume that $\boldsymbol{u}$ satisfies the power constraint $\sum_{t=1}^{k} u_t^\top u_t \le k\gamma^2$. Note that, by Parseval's Theorem, we have $\boldsymbol{\Lambda}_k^{\mathrm{freq}}(A_\star, \boldsymbol{u}) = \sum_{t=0}^{k-1} x_t^{\boldsymbol{u},\mathrm{ss}} (x_t^{\boldsymbol{u},\mathrm{ss}})^\top$. Furthermore, we have $x_t^{\boldsymbol{u}} = x_t^{\boldsymbol{u},\mathrm{ss}} + A_\star^t(x_0 - x_0^{\boldsymbol{u},\mathrm{ss}})$, for x_0 the initial state of (3.0.1) (when it is non-zero).

To control the transient behavior of (3.0.1), let:

$$\tau(A,\rho) := \sup\{\|A^k\|_{\mathrm{op}}\rho^{-k} \ : \ k \ge 0\}.$$

$\tau(A,\rho)$ is therefore the smallest value such that $\|A^k\|_{\mathrm{op}} \le \tau(A,\rho)\rho^k$ for all k, and controls the rate at which the system "forgets". We will define

$$\rho_\star := \max\left\{\frac{1}{2}, \frac{2\|A_\star\|_{\mathcal{H}_\infty}\|A_\star\|_{\mathrm{op}}}{1+2\|A_\star\|_{\mathcal{H}_\infty}\|A_\star\|_{\mathrm{op}}}\right\}, \quad \tau_\star := \tau(\widetilde{A}_\star, \rho_\star) \text{ for } \widetilde{A}_\star := \begin{bmatrix} A_\star & B_\star \\ 0 & 0 \end{bmatrix}.$$

The following result relates $\tau_\star$ and $\frac{1}{1-\rho_\star}$ to $\|A_\star\|_{\mathcal{H}_\infty}$ and $\|B_\star\|_{\mathrm{op}}$, which will aid in simplifying our results.

Lemma 3.4.1. *The following upper bounds hold:*

$$\frac{1}{1-\rho_\star} \le 2 + 2\|A_\star\|_{\mathcal{H}_\infty}^2, \quad \tau_\star \le 2(1+2\|B_\star\|_{\mathrm{op}})\|A_\star\|_{\mathcal{H}_\infty}.$$

For initial state x_0 not necessarily zero, we define

$$T_{\mathrm{ss}}(\zeta, k, \|x_0\|_2) := \frac{c}{1-\rho_\star} \cdot \log\left(\frac{k\gamma\tau_\star\|A_\star\|_{\mathcal{H}_\infty}\|B_\star\|_{\mathrm{op}}(1+\|x_0\|_2)}{\zeta(1-\rho_\star)}\right). \tag{3.4.1}$$

We have the following expanded version of Theorem 3.1.

Theorem 3.5. *Consider some set of inputs $(\boldsymbol{u}_i)_{i=1}^p$, such that each $\boldsymbol{u}_i$ has period k and satisfies the power constraint $\sum_{t=1}^k u_{i,t}^\top u_{i,t} \le k\gamma^2$. Assume that we start from some initial state x_0 and play input $\boldsymbol{u}_1$ for T/p steps, then $\boldsymbol{u}_2$ for T/p steps, and so on, combined with input noise distributed as $\mathcal{N}(0, \sigma_u^2 \cdot I)$. Then on the event*

$$\mathcal{E} := \left\{\textstyle\sum_{t=1}^T x_t x_t^\top \preceq T\bar{\boldsymbol{\Lambda}}_T\right\},$$

and as long as

$$\frac{T}{p} \ge 2T_{\mathrm{ss}}\left(\frac{1}{10}\lambda_{\mathrm{noise}}(\sigma_u), k, \sqrt{T\|\bar{\boldsymbol{\Lambda}}_T\|_{\mathrm{op}}}\right) + c\cdot k\left(d_x + \log\det(\bar{\boldsymbol{\Lambda}}_T/\lambda_{\mathrm{noise}}(\sigma_u)) + \log\frac{p}{\delta}\right), \tag{3.4.2}$$

we have that with probability at least $1-\delta$:

$$\|\widehat{A} - A_\star\|_{\mathrm{op}} \le C\sigma_w\sqrt{\frac{\log\frac{1}{\delta} + \log\det(\bar{\boldsymbol{\Lambda}}_T/\lambda_{\mathrm{noise}}(\sigma_u) + I) + d_x}{T\cdot\lambda_{\min}(\boldsymbol{\Lambda}_k^{\mathrm{ss}}(A_\star, \boldsymbol{U}, \sigma_u))}}$$

where $\boldsymbol{U} := \frac{1}{p}\sum_{i=1}^p \check{\boldsymbol{u}}_i \otimes \check{\boldsymbol{u}}_i$.

Note that Theorem 3.1 is a directly corollary of this bound in the case when $p = 1$.

Proof of Theorem 3.5. First define the following events:

$$\mathcal{A} := \left\{\|\widehat{A} - A_\star\|_{\mathrm{op}} \le C\sigma_w\sqrt{\frac{\log\frac{1}{\delta} + \log\det(\bar{\boldsymbol{\Lambda}}_T/\lambda_{\mathrm{noise}}(\sigma_u) + I) + d_x}{T\cdot\lambda_{\min}(\boldsymbol{\Lambda}_k^{\mathrm{ss}}(A_\star, \boldsymbol{U}, \sigma_u))}}\right\}$$

$$\mathcal{E}_1 := \left\{\sum_{t=1}^T x_t x_t^\top \succeq c_1 T\boldsymbol{\Lambda}_k^{\mathrm{ss}}(A_\star, \boldsymbol{U}, \sigma_u)\right\}$$

$$\mathcal{E}_{1,i} := \left\{\sum_{t=T(i-1)/p}^{Ti/p} x_t x_t^\top \succeq c_1\frac{T}{p}\boldsymbol{\Lambda}_k^{\mathrm{ss}}(A_\star, \boldsymbol{u}_i, \sigma_u)\right\}, \quad i = 1, \dots, p,$$

$$\mathcal{E}_2 := \left\{\left\|\left(\sum_{t=1}^T x_t x_t^\top\right)^{-1/2}\sum_{t=1}^T x_t w_t^\top\right\|_{\mathrm{op}} \le c_2\sigma_w\sqrt{\log\frac{1}{\delta} + d_x + \log\det(\bar{\boldsymbol{\Lambda}}_T/\lambda_{\mathrm{noise}}(\sigma_u) + I)}\right\}$$

Our goal now is to bound $\mathbb{P}[\mathcal{A}^c \cap \mathcal{E}]$. To achieve this, we aim to show that each of $\mathcal{E}_1, \mathcal{E}_{1,i}$, and $\mathcal{E}_2$ hold with high probability, and that on these events $\mathcal{A}$ must hold. We have:

$$\mathbb{P}[\mathcal{A}^c \cap \mathcal{E}] \leq \mathbb{P}[\mathcal{A}^c \cap \mathcal{E} \cap \mathcal{E}_1 \cap \mathcal{E}_2] + \mathbb{P}[\mathcal{E} \cap \mathcal{E}_1 \cap \mathcal{E}_2^c] + \mathbb{P}[\mathcal{E} \cap \mathcal{E}_1^c].$$

Note that on $\mathcal{E}$, we can bound $\|x_t\|_2 \leq \sqrt{T\|\bar{\mathbf{\Lambda}}_T\|_{\text{op}}}$ for all t. By Lemma A.1.1, the burn-in condition required by Lemma 3.2.1 will then be met when playing input $\boldsymbol{u}_i$ if

$$\frac{T}{p} \geq \max_{v \in \mathcal{V}_i} 2T_{\text{ss}}\left(\tfrac{1}{10k} v^\top \mathbf{\Lambda}_k^{\text{freq}}(A_\star, \boldsymbol{u})v, k, \sqrt{T\|\bar{\mathbf{\Lambda}}_T\|_{\text{op}}}\right) + ck\left(d_x + \log\det(\bar{\mathbf{\Lambda}}_T \mathbf{\Lambda}_k^{\text{ss}}(A_\star, \boldsymbol{u}_i, \sigma_u)^{-1}) + \log\tfrac{p}{\delta}\right),$$

for $\mathcal{V}_i$ as defined in Lemma 3.2.1, with respect to input $\boldsymbol{u}_i$. By definition of $\mathcal{V}_i$, for each $v \in \mathcal{V}_i$ we have $\frac{1}{k} v^\top \mathbf{\Lambda}_k^{\text{freq}}(A_\star, \boldsymbol{u}_i)v \geq v^\top \mathbf{\Lambda}_k^{\text{noise}}(A_\star, \sigma_u)v \geq \lambda_{\text{noise}}(\sigma_u)$, so $T_{\text{ss}}(\frac{1}{10}\lambda_{\text{noise}}(\sigma_u), k, \sqrt{T\|\bar{\mathbf{\Lambda}}_T\|_{\text{op}}}) \geq T_{\text{ss}}(\frac{1}{10k} v^\top \mathbf{\Lambda}_k^{\text{freq}}(A_\star, \boldsymbol{u})v, k, \sqrt{T\|\bar{\mathbf{\Lambda}}_T\|_{\text{op}}})$. Furthermore, we have $\mathbf{\Lambda}_k^{\text{ss}}(A_\star, \boldsymbol{u}_i, \sigma_u) \succeq \lambda_{\text{noise}}(\sigma_u) \cdot I$. Together, this implies that the burn-in condition required by Lemma 3.2.1 will be met if

$$\frac{T}{p} \geq 2T_{\text{ss}}\left(\frac{1}{10}\lambda_{\text{noise}}(\sigma_u), k, \sqrt{T\|\bar{\mathbf{\Lambda}}_T\|_{\text{op}}}\right) + c \cdot k\left(d_x + \log\det(\bar{\mathbf{\Lambda}}_T/\lambda_{\text{noise}}(\sigma_u)) + \log\frac{p}{\delta}\right).$$

As this is precisely the condition required by (3.4.2), by Lemma 3.2.1 it follows that $\mathbb{P}[\mathcal{E} \cap \mathcal{E}_{1,i}^c] \leq \delta/p$. Union bounding over all i, we have that $\mathbb{P}[\mathcal{E} \cap (\cup_{i=1}^p \mathcal{E}_{1,i}^c)] \leq \delta$. Note that, by definition of $\boldsymbol{U}$, we have

$$\mathbf{\Lambda}_k^{\text{freq}}(A_\star, \boldsymbol{U}, \sigma_u) = \frac{1}{p}\sum_{i=1}^p \mathbf{\Lambda}_k^{\text{freq}}(A_\star, \boldsymbol{u}_i, \sigma_u),$$

which implies that $\mathbf{\Lambda}_k^{\text{ss}}(A_\star, \boldsymbol{U}, \sigma_u) = \frac{1}{p}\sum_{i=1}^p \mathbf{\Lambda}_k^{\text{ss}}(A_\star, \boldsymbol{u}_i, \sigma_u)$. Thus, on $\cap_{i=1}^p \mathcal{E}_{1,i}$, we have $\sum_{t=1}^T x_t x_t^\top \succeq c_1 T \mathbf{\Lambda}_k^{\text{ss}}(A_\star, \boldsymbol{U}, \sigma_u)$. It follows that $\mathcal{E}_1$ holds on $\cap_{i=1}^p \mathcal{E}_{1,i}$, so $\mathbb{P}[\mathcal{E} \cap \mathcal{E}_1^c] \leq \delta$.

By Proposition 3.1 and since $\mathbf{\Lambda}_k^{\text{ss}}(A_\star, \boldsymbol{U}, \sigma_u) \succeq \lambda_{\text{noise}}(\sigma_u) \cdot I$, we have that $\mathbb{P}[\mathcal{E} \cap \mathcal{E}_1 \cap \mathcal{E}_2^c] \leq \delta$. By the error decomposition of the least-squares estimator given in (3.1.2), we see that $\mathbb{P}[\mathcal{A}^c \cap \mathcal{E} \cap \mathcal{E}_1 \cap \mathcal{E}_2] = 0$. The result then follows after rescaling δ. □

3.4.1 Lower Bounding Minimum Eigenvalue of Covariates

The key piece in proving Theorem 3.5 is a high-probability lower bound on the covariates when playing a periodic input. Our main lower bound is encapsulated in Lemma 3.2.1, which is itself a consequence of Proposition 3.3. We prove these results in this section.

Proof of Lemma 3.2.1. The proof of this follows Simchowitz et al. [2018] closely but replacing Proposition 2.5 of Simchowitz et al. [2018] with our Proposition 3.3.

Let $\mathcal{F}_t$ denote the filtration generated by (3.0.1) for noise and inputs up to t. Take some

$s \geq 0$, then:

$$v^\top x_{s+t} \mid \mathcal{F}_s \sim \mathcal{N}\left(v^\top A_\star^t x_s + v^\top x_{s+t}^{\boldsymbol{u}}, v^\top \boldsymbol{\Lambda}_{t-s}^{\text{noise}}(A_\star, \sigma_u) v\right).$$

Given this, we have that x_{s+t} satisfies the $(2k, \boldsymbol{\Lambda}_k^{\text{noise}}(A_\star, \sigma_u), 3/20)$-BMSB condition, as defined in Simchowitz et al. [2018] (see Proposition 3.1 of Simchowitz et al. [2018] for a justification of this). We can then apply Proposition 2.5 of Simchowitz et al. [2018] to get that, for any $v \in \mathcal{S}^{d-1}$:

$$\mathbb{P}\left[\sum_{t=1}^{T}(v^\top x_t)^2 \leq \frac{9}{3200} k \lfloor T/k \rfloor v^\top \boldsymbol{\Lambda}_k^{\text{noise}}(A_\star, \sigma_u) v\right] \leq e^{-\frac{9}{6400}\lfloor T/k \rfloor}.$$

This further implies that

$$\mathbb{P}\left[\sum_{t=1}^{T}(v^\top x_t)^2 \leq \frac{9}{3200} k \lfloor \tfrac{T-T_{\text{ss}}}{k} \rfloor v^\top \boldsymbol{\Lambda}_k^{\text{noise}}(A_\star, \sigma_u) v\right] \leq e^{-\frac{9}{6400}\lfloor \frac{T-T_{\text{ss}}}{k} \rfloor}.$$

By Proposition 3.3, for $v \in \mathcal{V}$, we have that

$$\begin{aligned}\mathbb{P}\left[\sum_{t=1}^{T}(v^\top x_t)^2 \leq \frac{2}{81} \lfloor \tfrac{T-T_{\text{ss}}}{k} \rfloor v^\top \boldsymbol{\Lambda}_k^{\text{freq}}(A_\star, \boldsymbol{u}) v\right] &\leq \mathbb{P}\left[\sum_{t=T_{\text{ss}}}^{T}(v^\top x_t)^2 \leq \frac{2}{81} \lfloor \tfrac{T-T_{\text{ss}}}{k} \rfloor v^\top \boldsymbol{\Lambda}_k^{\text{freq}}(A_\star, \boldsymbol{u}) v\right] \\ &\leq e^{-\frac{2}{81}\lfloor \frac{T-T_{\text{ss}}}{k} \rfloor}.\end{aligned}$$

We see then that, for $v \in \mathcal{V}$,

$$\begin{aligned}\mathbb{P}\left[\sum_{t=1}^{T}(v^\top x_t)^2 \leq \tfrac{9}{6400} \lfloor \tfrac{T-T_{\text{ss}}}{k} \rfloor v^\top \boldsymbol{\Lambda}_k^{\text{ss}}(A_\star, \boldsymbol{u}, \sigma_u) w\right] &\leq \mathbb{P}\left[\left\{\sum_{t=1}^{T}(v^\top x_t)^2 \leq \tfrac{9}{3200} k \lfloor \tfrac{T-T_{\text{ss}}}{k} \rfloor v^\top \boldsymbol{\Lambda}_k^{\text{noise}}(A_\star, \sigma_u) v\right\}\right. \\ &\qquad \left. \cup \left\{\sum_{t=1}^{T}(v^\top x_t)^2 \leq \frac{2}{81} \lfloor \tfrac{T-T_{\text{ss}}}{k} \rfloor v^\top \boldsymbol{\Lambda}_k^{\text{freq}}(A_\star, \boldsymbol{u}) v\right\}\right] \\ &\leq e^{-\frac{9}{6400}\lfloor \frac{T-T_{\text{ss}}}{k} \rfloor} + e^{-\frac{2}{81}\lfloor \frac{T-T_{\text{ss}}}{k} \rfloor} \\ &\leq 2e^{-\frac{9}{6400}\lfloor \frac{T-T_{\text{ss}}}{k} \rfloor}.\end{aligned}$$

For $v \in \mathcal{S}^{d-1} \backslash \mathcal{V}$, we have $v^\top \boldsymbol{\Lambda}_k^{\text{ss}}(A_\star, \boldsymbol{u}, \sigma_u) v \leq 2 v^\top \boldsymbol{\Lambda}_k^{\text{noise}}(A_\star, \sigma_u) v$, so that

$$\begin{aligned}&\mathbb{P}\left[\sum_{t=1}^{T}(v^\top x_t)^2 \leq \frac{9}{6400} k \lfloor \tfrac{T-T_{\text{ss}}}{k} \rfloor v^\top \boldsymbol{\Lambda}_k^{\text{ss}}(A_\star, \boldsymbol{u}, \sigma_u) v\right] \\ &\leq \mathbb{P}\left[\sum_{t=1}^{T}(v^\top x_t)^2 \leq \frac{9}{3200} k \lfloor \tfrac{T-T_{\text{ss}}}{k} \rfloor v^\top \boldsymbol{\Lambda}_k^{\text{noise}}(A_\star, \sigma_u) v\right]\end{aligned}$$

$$\leq e^{-\frac{9}{6400}\lfloor\frac{T-T_{\text{ss}}}{k}\rfloor}.$$

Following the proof of Theorem 2.4 of Simchowitz et al. [2018], let $\mathcal{T}$ be a 1/4-net in the norm $T\bar{\mathbf{\Lambda}}_T$ of $\left\{v \ : \ \frac{9}{6400}k\lfloor\frac{T-T_{\text{ss}}}{k}\rfloor v^\top \mathbf{\Lambda}_k^{\text{ss}}(A_\star, \boldsymbol{u}, \sigma_u)v = 1\right\}$. By Lemma D.1 of Simchowitz et al. [2018], we have that $|\mathcal{T}| \leq c \cdot d + \log\det(\bar{\mathbf{\Lambda}}_T \mathbf{\Lambda}_k^{\text{ss}}(A_\star, \boldsymbol{u}, \sigma_u)^{-1})$. Then by Lemma 4.1 of Simchowitz et al. [2018] we have:

$$\begin{aligned}
&\mathbb{P}\left[\sum_{t=1}^T x_t x_t^\top \not\succeq \frac{9}{12800}k\lfloor\tfrac{T-T_{\text{ss}}}{k}\rfloor v^\top \mathbf{\Lambda}_k^{\text{ss}}(A_\star, \boldsymbol{u}, \sigma_u)v, \sum_{t=1}^T x_t x_t^\top \preceq T\bar{\mathbf{\Lambda}}_T\right] \\
&\leq \mathbb{P}\left[\exists v \in \mathcal{T} \ : \ \sum_{t=1}^T (v^\top x_t)^2 < \frac{9}{6400}k\lfloor\tfrac{T-T_{\text{ss}}}{k}\rfloor v^\top \mathbf{\Lambda}_k^{\text{ss}}(A_\star, \boldsymbol{u}, \sigma_u)v, \sum_{t=1}^T x_t x_t^\top \preceq T\bar{\mathbf{\Lambda}}_T\right] \\
&\leq 2\exp\left(-\frac{9\lfloor\frac{T-T_{\text{ss}}}{k}\rfloor}{6400} + c \cdot d_x + \log\det(\bar{\mathbf{\Lambda}}_T \mathbf{\Lambda}_k^{\text{ss}}(A_\star, \boldsymbol{u}, \sigma_u)^{-1})\right) \\
&\overset{(a)}{\leq} 2\exp\left(-\frac{27(T-T_{\text{ss}})}{25600k} + c \cdot d_x + \log\det(\bar{\mathbf{\Lambda}}_T \mathbf{\Lambda}_k^{\text{ss}}(A_\star, \boldsymbol{u}, \sigma_u)^{-1})\right) \\
&\overset{(b)}{\leq} \delta
\end{aligned}$$

where (a) holds so long as $T \geq T_{\text{ss}} + 4k$, which will be true by (3.2.3), and (b) holds by (3.2.3). Finally, if $T \geq T_{\text{ss}} + 4k$ and (3.2.3) holds, we have $k\lfloor\frac{T-T_{\text{ss}}}{k}\rfloor \geq c \cdot T$, which completes the result. □

Proof of Proposition 3.3. Denote $\alpha := v^\top \mathbf{\Lambda}_k^{\text{freq}}(A_\star, \boldsymbol{u})v = \sum_{t=0}^{k-1}(v^\top x_t^{\boldsymbol{u},\text{ss}})^2$. Let

$$z_t := v^\top x_{T_{\text{ss}}+t} = v^\top(x^{\boldsymbol{u}}_{T_{\text{ss}}+t} + x^{\boldsymbol{w}}_{T_{\text{ss}}+t}), \qquad \mu_t := v^\top x^{\boldsymbol{u}}_{T_{\text{ss}}+t} - v^\top \bar{x}^{\boldsymbol{u}}_{T_{\text{ss}}+\lfloor(t-1)/k\rfloor k:k}.$$

Note that $\sum_{t=1}^k \mu_{jk+t} = 0$ for any $j \geq 0$, and that μ_t is a deterministic quantity. Also define

$$B_j := \mathbb{I}\left[\sum_{i=1}^k z_{jk+i}^2 \geq c_1 \sum_{i=1}^k \mu_{jk+i}^2\right]$$

for some c_1 to be specified. Then $\sum_{i=1}^k z_{jk+i}^2 \geq \left(c_1 \sum_{i=1}^k \mu_{jk+i}^2\right) B_j$. Let $S = \lfloor T/k \rfloor$ and c_2 be some constant to be specified. Our goal is to bound $\mathbb{P}\left[\sum_{t=1}^T z_t^2 \leq c_2 \sum_{t=1}^T \mu_t^2\right]$. First,

$$\mathbb{P}\left[\sum_{t=1}^T z_t^2 \leq c_2 \sum_{t=1}^T \mu_t^2\right] \leq \mathbb{P}\left[\sum_{j=0}^{S-1}\left(c_1 \sum_{i=1}^k \mu_{jk+i}^2\right) B_j \leq c_2 \sum_{t=1}^T \mu_t^2\right] \tag{3.4.3}$$

$$\leq \inf_{\lambda<0} \exp\left\{-\lambda c_2 \sum_{t=1}^{T} \mu_t^2\right\} \mathbb{E}\left[\exp\left\{\lambda \sum_{j=0}^{S-1}\left(c_1 \sum_{i=1}^{k} \mu_{jk+i}^2\right) B_j\right\}\right]$$

where the last inequality is simply Chernoff's bound. Next, let $\mathcal{F}_j$ denote the σ-field generated by $w_0, ..., w_{T_{\text{ss}}+jk}$. We have that:

$$\begin{aligned}
\mathbb{E}[B_j|\mathcal{F}_j] &= \mathbb{P}\left[\sum_{i=1}^{k} z_{jk+i}^2 \geq c_1 \sum_{i=1}^{k} \mu_{jk+i}^2 \mid \mathcal{F}_j\right] \\
&= \mathbb{P}\left[\sum_{i=1}^{k} \left(\mu_{jk+i} + v^\top x_{T_{\text{ss}}+jk+i}^{\boldsymbol{w}} + v^\top \bar{x}_{T_{\text{ss}}+jk}^{\boldsymbol{u}}\right)^2 \geq c_1 \sum_{i=1}^{k} \mu_{jk+i}^2 \mid \mathcal{F}_j\right] \\
&= \mathbb{P}\left[\sum_{i=1}^{k} \left(\mu_{jk+i} + v^\top \sum_{s=0}^{i-1} A_\star^{i-s-1} w_{T_{\text{ss}}+jk+s} + v^\top A_\star^i x_{T_{\text{ss}}+jk}^{\boldsymbol{w}} + v^\top \bar{x}_{T_{\text{ss}}+jk}^{\boldsymbol{u}}\right)^2 \geq c_1 \sum_{i=1}^{k} \mu_{jk+i}^2 \mid \mathcal{F}_j\right]
\end{aligned}$$

where the last equality follows since

$$x_{T_{\text{ss}}+jk+i}^{\boldsymbol{w}} = A_\star^i x_{T_{\text{ss}}+jk}^{\boldsymbol{w}} + \sum_{s=0}^{i-1} A_\star^{i-s-1} w_{T_{\text{ss}}+jk+s}.$$

Note that, conditioned on $\mathcal{F}_j$, $v^\top A_\star^i x_{T_{\text{ss}}+jk}^{\boldsymbol{w}}$ and $v^\top \bar{x}_{T_{\text{ss}}+jk}^{\boldsymbol{u}}$ are deterministic. Further, since w_t is mean 0, $v^\top \sum_{s=0}^{i-1} A_\star^{i-s-1} w_{T_{\text{ss}}+jk+i}$ will simply be a linear combination of mean 0 Gaussians and so will itself be a mean 0 Gaussian. This implies that $\mathbb{P}[v^\top \sum_{s=0}^{i-1} A_\star^{i-s-1} w_{T_{\text{ss}}+jk+i} \geq 0] = 1/2$.

Since we have constructed μ_t in such a way as to be mean zero over a block of length k, for any fixed a:

$$\sum_{i=1}^{k} (\mu_{jk+i} + a)^2 = \sum_{i=1}^{k} \mu_{jk+i}^2 + a \sum_{i=1}^{k} \mu_{jk+i} + a^2 = \sum_{i=1}^{k} \mu_{jk+i}^2 + a^2 \geq \sum_{i=1}^{k} \mu_{jk+i}^2.$$

In particular then

$$\sum_{i=1}^{k} (\mu_{jk+i} + v^\top A_\star^i x_{T_{\text{ss}}+jk}^{\boldsymbol{w}} + v^\top \bar{x}_{T_{\text{ss}}+jk}^{\boldsymbol{u}})^2 \mid \mathcal{F}_j \geq \sum_{i=1}^{k} \mu_{jk+i}^2 \mid \mathcal{F}_j, \tag{3.4.4}$$

which implies

$$\mathbb{P}\left[\sum_{i=1}^{k} \left(\mu_{jk+i} + v^\top \sum_{s=0}^{i-1} A_\star^{i-s-1} w_{T_{\text{ss}}+jk+s} + v^\top A_\star^i x_{T_{\text{ss}}+jk}^{\boldsymbol{w}} + v^\top \bar{x}_{T_{\text{ss}}+jk}^{\boldsymbol{u}}\right)^2 \geq c_1 \sum_{i=1}^{k} \mu_{jk+i}^2 \mid \mathcal{F}_j\right]$$

$$\geq \mathbb{P}\left[\sum_{i=1}^{k}\left(\mu_{jk+i} + v^\top \sum_{s=0}^{i-1} A_\star^{i-s-1} w_{T_{\rm ss}+jk+s} + v^\top A_\star^i x^{\boldsymbol{w}}_{T_{\rm ss}+jk} + v^\top \bar{x}^{\boldsymbol{u}}_{T_{\rm ss}+jk}\right)^2 \right.$$
$$\left. \geq c_1 \sum_{i=1}^{k} (\mu_{jk+i} + v^\top A_\star^i x^{\boldsymbol{w}}_{T_{\rm ss}+jk} + v^\top \bar{x}^{\boldsymbol{u}}_{T_{\rm ss}+jk})^2 \mid \mathcal{F}_j\right]$$
$$\geq \mathbb{P}\left[\sum_{i=1}^{k} \left(\mu_{jk+i} + v^\top A_\star^i x^{\boldsymbol{w}}_{T_{\rm ss}+jk} + v^\top \bar{x}^{\boldsymbol{u}}_{T_{\rm ss}+jk}\right)^2 \mathbb{I}\left[\left|\mu_{jk+i} + v^\top \sum_{s=0}^{i-1} A_\star^{i-s-1} w_{T_{\rm ss}+jk+s}\right.\right.\right.$$
$$\left.\left.\left. + v^\top A_\star^i x^{\boldsymbol{w}}_{T_{\rm ss}+jk} + v^\top \bar{x}^{\boldsymbol{u}}_{T_{\rm ss}+jk}\right| \geq \left|\mu_{jk+i} + v^\top A_\star^i x^{\boldsymbol{w}}_{T_{\rm ss}+jk} + v^\top \bar{x}^{\boldsymbol{u}}_{T_{\rm ss}+jk}\right|\right]\right.$$
$$\left. \geq c_1 \sum_{i=1}^{k} (\mu_{jk+i} + v^\top A_\star^i x^{\boldsymbol{w}}_{T_{\rm ss}+jk} + v^\top \bar{x}^{\boldsymbol{u}}_{T_{\rm ss}+jk})^2 \mid \mathcal{F}_j\right]$$
$$\overset{(a)}{\geq} \frac{1/2 - c_1}{1 - c_1}.$$

Here the last inequality follows by a reverse Markov inequality which states that, for any random variable Z supported in $[0,1]$ almost surely and with $\mathbb{E}[Z] \geq p \in (0,1)$, for all $t \in [0,p]$, $\mathbb{P}[Z \geq t] \geq \frac{p-t}{1-t}$ [Simchowitz et al., 2018]. Noting that, since the noise is 0 mean Gaussian, we have

$$\mathbb{E}\left[\sum_{i=1}^{k} \left(\mu_{jk+i} + v^\top A_\star^i x^{\boldsymbol{w}}_{T_{\rm ss}+jk} + v^\top \bar{x}^{\boldsymbol{u}}_{T_{\rm ss}+jk}\right)^2 \mathbb{I}\left[\left|\mu_{jk+i} + v^\top \sum_{s=0}^{i-1} A_\star^{i-s-1} w_{T_{\rm ss}+jk+s} + v^\top A_\star^i x^{\boldsymbol{w}}_{T_{\rm ss}+jk}\right.\right.\right.$$
$$\left.\left.\left. + v^\top \bar{x}^{\boldsymbol{u}}_{T_{\rm ss}+jk}\right| \geq \left|\mu_{jk+i} + v^\top A_\star^i x^{\boldsymbol{w}}_{T_{\rm ss}+jk} + v^\top \bar{x}^{\boldsymbol{u}}_{T_{\rm ss}+jk}\right|\right] \mid \mathcal{F}_j\right]$$
$$= \sum_{i=1}^{k} \left(\mu_{jk+i} + v^\top A_\star^i x^{\boldsymbol{w}}_{T_{\rm ss}+jk} + v^\top \bar{x}^{\boldsymbol{u}}_{T_{\rm ss}+jk}\right)^2 \mathbb{P}\left[\left|\mu_{jk+i} + v^\top \sum_{s=0}^{i-1} A_\star^{i-s-1} w_{T_{\rm ss}+jk+s} + v^\top A_\star^i x^{\boldsymbol{w}}_{T_{\rm ss}+jk}\right.\right.$$
$$\left.\left. + v^\top \bar{x}^{\boldsymbol{u}}_{T_{\rm ss}+jk}\right| \geq \left|\mu_{jk+i} + v^\top A_\star^i x^{\boldsymbol{w}}_{T_{\rm ss}+jk} + v^\top \bar{x}^{\boldsymbol{u}}_{T_{\rm ss}+jk}\right| \mid \mathcal{F}_j\right]$$
$$= \frac{1}{2} \sum_{i=1}^{k} \left(\mu_{jk+i} + v^\top A_\star^i x^{\boldsymbol{w}}_{T_{\rm ss}+jk} + v^\top \bar{x}^{\boldsymbol{u}}_{T_{\rm ss}+jk}\right)^2.$$

From this (a) follows by simple manipulations. Since we can choose c_1 as we wish, we set it equal to $c_1 = 1/4$ and conclude that

$$\mathbb{E}[B_j \mid \mathcal{F}_j] \geq \frac{1}{3}.$$

Returning to (3.4.3), we can now use this result to bound the expectation. Note that by the tower property of conditional expectation we have

$$\begin{aligned}\mathbb{E}\left[\exp\left\{\lambda\sum_{j=0}^{S-1}\left(c_1\sum_{i=1}^{k}\mu_{jk+i}^2\right)B_j\right\}\right] &= \mathbb{E}\left[\mathbb{E}\left[\exp\left\{\lambda\sum_{j=0}^{S-1}\left(c_1\sum_{i=1}^{k}\mu_{jk+i}^2\right)B_j\right\}\mid\mathcal{F}_{S-1}\right]\right]\\ &= \mathbb{E}\left[\exp\left\{\lambda\sum_{j=0}^{S-2}\left(c_1\sum_{i=1}^{k}\mu_{jk+i}^2\right)B_j\right\}\mathbb{E}\left[\exp\left\{\lambda\left(c_1\sum_{i=1}^{k}\mu_{(S-1)k+i}^2\right)B_{S-1}\right\}\mid\mathcal{F}_{S-1}\right]\right].\end{aligned}$$

Then by what we just proved and applying Hoeffding's Lemma, since $\lambda<0$, we have

$$\mathbb{E}\left[\exp\left\{\lambda\left(c_1\sum_{i=1}^{k}\mu_{(S-1)k+i}^2\right)B_{S-1}\right\}\mid\mathcal{F}_{S-1}\right]\le\exp\left\{\frac{\lambda}{3}\left(c_1\sum_{i=1}^{k}\mu_{(S-1)k+i}^2\right)+\frac{\lambda^2}{8}\left(c_1\sum_{i=1}^{k}\mu_{(S-1)k+i}^2\right)^2\right\}.$$

Repeating this procedure condition on each $\mathcal{F}_i$, we get

$$\mathbb{E}\left[\exp\left\{\lambda\sum_{j=0}^{S-1}\left(c_1\sum_{i=1}^{k}\mu_{jk+i}^2\right)B_j\right\}\right]\le\exp\left\{\frac{\lambda}{3}\sum_{j=0}^{S-1}\left(c_1\sum_{i=1}^{k}\mu_{jk+i}^2\right)+\frac{\lambda^2}{8}\sum_{j=0}^{S-1}\left(c_1\sum_{i=1}^{k}\mu_{jk+i}^2\right)^2\right\}$$

and so

$$\begin{aligned}\mathbb{P}\left[\sum_{t=1}^{T}z_t^2\le c_2\sum_{t=1}^{T}\mu_t^2\right] &\le \inf_{\lambda<0}\exp\left\{-\lambda c_2\sum_{t=1}^{T}\mu_t^2\right\}\exp\left\{\frac{\lambda}{3}\sum_{j=0}^{S-1}\left(c_1\sum_{i=1}^{k}\mu_{jk+i}^2\right)+\frac{\lambda^2}{8}\sum_{j=0}^{S-1}\left(c_1\sum_{i=1}^{k}\mu_{jk+i}^2\right)^2\right\}\\ &= \inf_{\lambda<0}\exp\left\{\lambda\left(\frac{c_1}{3}-c_2\right)\sum_{j=0}^{S-1}\left(\sum_{i=1}^{k}\mu_{jk+i}^2\right)+\frac{\lambda^2}{8}\sum_{j=0}^{S-1}\left(c_1\sum_{i=1}^{k}\mu_{jk+i}^2\right)^2\right\}\\ &\overset{(a)}{\le}\exp\left\{-\frac{2\left(\left(\frac{c_1}{3}-c_2\right)\sum_{j=0}^{S-1}\sum_{i=1}^{k}\mu_{jk+i}^2\right)^2}{\sum_{j=0}^{S-1}\left(c_1\sum_{i=1}^{k}\mu_{jk+i}^2\right)^2}\right\}\\ &\overset{(b)}{=}\exp\left\{-C\frac{\left(\sum_{j=0}^{S-1}\sum_{i=1}^{k}\mu_{jk+i}^2\right)^2}{\sum_{j=0}^{S-1}\left(\sum_{i=1}^{k}\mu_{jk+i}^2\right)^2}\right\}\end{aligned}$$

where (a) follows from choosing the optimal $\lambda<0$ (and assuming c_2 chosen such that $c_1/3-c_2$ is positive), and (b) uses constant $C=2(c_1/3-c_2)^2/c_1^2$. By (3.2.1), we have that

$\sum_{i=1}^{k} \mu_{jk+i}^2 = \alpha + \alpha_j$ for some $|\alpha_j| \le \alpha/10$. Thus:

$$\begin{aligned}
\exp\left\{-C\frac{\left(\sum_{j=0}^{S-1}\sum_{i=1}^{k}\mu_{jk+i}^2\right)^2}{\sum_{j=0}^{S-1}\left(\sum_{i=1}^{k}\mu_{jk+i}^2\right)^2}\right\} &= \exp\left\{-C\frac{\left(\sum_{j=0}^{S-1}\alpha+\alpha_j\right)^2}{\sum_{j=0}^{S-1}(\alpha+\alpha_j)^2}\right\} \\
&\le \exp\left\{-C\frac{\left(\sum_{j=0}^{S-1}9/10\alpha\right)^2}{\sum_{j=0}^{S-1}(11/10)^2\alpha^2}\right\} \\
&= \exp\left\{-C\frac{81}{121}S\right\}
\end{aligned}$$

where the inequality holds from maximizing this expression over α_j. Recalling that $S = \lfloor T/k \rfloor$, we conclude that

$$\mathbb{P}\left[\sum_{t=1}^{T} z_t^2 \le c_2 \sum_{t=1}^{T} \mu_t^2\right] \le \exp\left\{-C\frac{81}{121}\lfloor T/k \rfloor\right\}.$$

It remains then to write this in form of (3.2.2). Plugging in our definitions of μ_t and z_t, we have that the above is equivalent to

$$\begin{aligned}
&\mathbb{P}\left[\sum_{t=T_{\text{ss}}+1}^{T_{\text{ss}}+T}(v^\top x_t)^2 \le c_2\sum_{t=1}^{T}\left(v^\top x^{\boldsymbol{u}}_{T_{\text{ss}}+t} - v^\top \bar{x}^{\boldsymbol{u}}_{T_{\text{ss}}+\lfloor t/k-1\rfloor k:k}\right)^2\right] \le e^{-C\frac{81}{121}\lfloor T/k\rfloor} \\
\overset{(a)}{\Longrightarrow}\ &\mathbb{P}\left[\sum_{t=T_{\text{ss}}+1}^{T_{\text{ss}}+T}(v^\top x_t)^2 \le \frac{c_2}{2}\sum_{t=1}^{\lfloor T/k\rfloor k}\left(v^\top x^{\boldsymbol{u},\text{ss}}_{T_{\text{ss}}+t}\right)^2\right] \le e^{-C\frac{81}{121}\lfloor T/k\rfloor} \\
\overset{(b)}{\Longleftrightarrow}\ &\mathbb{P}\left[\sum_{t=1}^{T}(v^\top x_t)^2 \le \frac{c_2}{2}\lfloor T/k\rfloor v^\top \mathbf{\Lambda}_k^{\text{freq}}(A_\star, \boldsymbol{u})v\right] \le e^{-C\frac{81}{121}\lfloor T/k\rfloor}
\end{aligned}$$

where (a) holds by our assumption on the power (3.2.1) and (b) follows by Parseval's Theorem. Choosing c_2 to balance the constants, we get that

$$\mathbb{P}\left[\sum_{t=T_{\text{ss}}+1}^{T_{\text{ss}}+T}(v^\top x_t)^2 \le \frac{2}{81}\lfloor T/k\rfloor v^\top \mathbf{\Lambda}_k^{\text{freq}}(A_\star, \boldsymbol{U})v\right] \le e^{-\frac{2}{81}\lfloor T/k\rfloor},$$

which completes the proof. □

3.5 Proof of Theorem 3.2

We now prove the main theorem of this chapter, Theorem 3.2. Given Theorem 3.5, our main task is showing that the inputs played by Algorithm 3.1 are approximately optimal; that is,

they approximately maximize the minimum eigenvalue of the covariates on the true system.

Fix an epoch i and let $T = \sum_{j=0}^{i} T_j$. Note that, by the definition of T_i, we will have $T_i = \frac{1}{2}(T + T_0)$. Similarly, $T_{i-1} = \frac{1}{4}(T + T_0)$. Define the following events:

$$\mathcal{E}_1 = \left\{ \|\widehat{A}_{i-1} - A_\star\|_{\mathrm{op}} \le C\sigma_w \sqrt{\frac{\log 1/\delta + d_x \log(\beta_\infty/\lambda_{\mathrm{noise}} + 1)}{T_{i-1}\lambda_{\mathrm{noise}}}} =: \epsilon_{\mathrm{op},i-1} \right\}$$

$$\mathcal{E}_2 = \left\{ \|\widehat{A}_i - A_\star\|_{\mathrm{op}} \le C\sigma_w \sqrt{\frac{\log 1/\delta + d_x \log(\beta_\infty/\lambda_{\mathrm{noise}} + 1)}{T_i \cdot \lambda_{\min}(\mathbf{\Lambda}^{\mathrm{ss}}_{k_i}(A_\star, \boldsymbol{U}_i, \frac{\gamma}{\sqrt{2d_u}}))}} \right\}$$

$$\mathcal{E}_3 := \left\{ \sum_{t=1}^{T} x_t x_t^\top \preceq T\beta_\infty \cdot I \right\}.$$

Events $\mathcal{E}_i$ hold. The following result shows that Algorithm 3.1 is an instance of a *sequential open-loop policy*, denoted as $\Pi^{\mathrm{sol}}_{\gamma^2}$.

Lemma 3.5.1. *Algorithm 3.1 is in $\Pi^{\mathrm{sol}}_{\gamma^2}$ with $\sigma_u = \frac{\gamma}{\sqrt{2d_u}}$ and $n = \mathcal{O}(\log T)$.*

The definition of a sequential open-loop policy is deferred to Chapter 5, but in essence it allows us to obtain a bound on estimation error of $\widehat{A}_{i-1}$ in the form given in $\mathcal{E}_1$. In particular, by Lemma 5.4.1, Lemma 3.5.1 implies that Algorithm 3.1 satisfies Assumption 4.4 with

$$T_{\mathrm{se}} := c_1 d_x \Big((d_x + d_u) \log(\beta_\infty/\lambda_{\mathrm{noise}} + 1) + \log \frac{n}{\delta} \Big), \quad \underline{\lambda} = c_2 \lambda_{\mathrm{noise}}, \quad \bar{\mathbf{\Lambda}}_T = \beta_\infty \cdot I.$$

Thus, as long as $T \ge T_{\mathrm{se}}$ we will have that with probability at least $1-\delta$, $\lambda_{\min}(\Sigma_T) \ge c_2 \lambda_{\mathrm{noise}} T$ and $\sum_{t=1}^{T} x_t x_t^\top \preceq T\beta_\infty I$. We can therefore apply Lemma C.3.2, our operator norm estimation bound, to get that $\mathbb{P}[\mathcal{E}_1^c \cup \mathcal{E}_3^c] \le 2\delta$.

By Theorem 3.5 and given the form of the input returned by calling `ConstructTimeInput` on $\boldsymbol{U}_i$, on the event $\mathcal{E}_3$ and since $\lambda_{\mathrm{noise}} = \lambda_{\mathrm{noise}}(\frac{\gamma}{\sqrt{2d_u}})$ by definition, as long as

$$\frac{T_i}{d_u} \ge 2T_{\mathrm{ss}} \left(\frac{1}{10}\lambda_{\mathrm{noise}}, k_i, \sqrt{\beta_\infty T} \right) + c \cdot k_i \left(d_x \log \frac{T\beta_\infty}{\lambda_{\mathrm{noise}}} + \log \frac{d_u}{\delta} \right)$$

we have that $\mathbb{P}[\mathcal{E}_2] \le \delta$. Note that $k_i \le T_i^{1/4}$ by construction, and furthermore that T_{ss} increases monotonically in k_i. We see then that this condition is implied by

$$T_i \ge 2d_u T_{\mathrm{ss}} \left(\frac{1}{10}\lambda_{\mathrm{noise}}, T, \sqrt{\beta_\infty T} \right) + c d_u^{4/3} \cdot \left(d_x \log \frac{T\beta_\infty}{\lambda_{\mathrm{noise}}} + \log \frac{d_u}{\delta} \right)^{4/3}.$$

Using the definition of T_{ss} given in (3.4.1), we can bound

$$T_{\mathrm{ss}}\left(\frac{1}{10}\lambda_{\mathrm{noise}}, T, \sqrt{\beta_\infty T}\right) \le \frac{c}{1-\rho_\star}\cdot\log\left(\frac{\gamma\beta_\infty\tau_\star\|A_\star\|_{\mathcal{H}_\infty}\|B_\star\|_{\mathrm{op}}T}{\lambda_{\mathrm{noise}}(1-\rho_\star)}\right).$$

Altogether, then, it suffices that

$$T_i \ge \frac{c\cdot d_u}{1-\rho_\star}\cdot\log\left(\frac{\gamma\beta_\infty\tau_\star\|A_\star\|_{\mathcal{H}_\infty}\|B_\star\|_{\mathrm{op}}T}{\lambda_{\mathrm{noise}}(\sigma_u)(1-\rho_\star)}\right) + cd_u^{4/3}\cdot\left(d_x\log\frac{T\beta_\infty}{\lambda_{\mathrm{noise}}} + \log\frac{d_u}{\delta}\right)^{4/3}. \quad (3.5.1)$$

Optimality of Inputs. We next wish to show that the inputs played at round i are nearly optimal. The following result provides a bound on the suboptimaliy of the inputs obtained on some $\widehat{A}$, assuming $\widehat{A}$ is suffficiently close to $A_\star$.

Lemma 3.5.2. *Assume that* $\|\widehat{A}-A_\star\|_{\mathrm{op}} \le \epsilon$ *for*

$$\epsilon \le c\cdot\min\left\{\frac{1-\rho_\star}{\tau_\star}, \frac{1}{\|A_\star\|_{\mathcal{H}_\infty}}, \frac{1}{\gamma^2\|A_\star\|_{\mathcal{H}_\infty}^3\|B_\star\|_{\mathrm{op}}^2}, \frac{(1-\rho_\star)^2}{\sigma_w^2+\sigma_u^2\|B_\star\|_{\mathrm{op}}^2}, \frac{1-\rho_\star}{\sigma_u^2\|B_\star\|_{\mathrm{op}}\tau_\star^2}\right\}$$

and that

$$k \ge c\max\left\{\frac{\|\theta_\star\|_{\mathcal{H}_\infty}\gamma^2}{\lambda_{\mathrm{noise}}(\sigma_u)}, \frac{1}{\|\theta_\star\|_{\mathcal{H}_\infty}}\right\}\cdot\|A_\star\|_{\mathcal{H}_\infty}^2\|B_\star\|_{\mathrm{op}} + \frac{c}{1-\rho_\star}\cdot\log\left(\frac{\sigma_w^2\tau_\star^2+\sigma_u^2\tau_\star^2\|B_\star\|_{\mathrm{op}}^2}{\lambda_{\mathrm{noise}}(\sigma_u)(1-\rho_\star^2)}\right).$$

Then, for any $\bar{k} \ge d_x$, *we have*

$$\lambda_{\min}\left(\frac{1}{k}\mathbf{\Lambda}_k^{\mathrm{freq}}(A_\star, \widehat{\boldsymbol{U}}) + \mathbf{\Lambda}_k^{\mathrm{noise}}(A_\star, \sigma_u)\right) \ge \frac{1}{8}\cdot\max_{\boldsymbol{U}\in\mathcal{U}_{\gamma^2,\bar{k}}}\lambda_{\min}\left(\frac{1}{k}\mathbf{\Lambda}_k^{\mathrm{freq}}(A_\star, \boldsymbol{U}) + \mathbf{\Lambda}_{\bar{k}}^{\mathrm{noise}}(A_\star, 0)\right)$$

for $\widehat{\boldsymbol{U}} = \arg\max_{\boldsymbol{U}\in\mathcal{U}^0_{\gamma^2,k}}\lambda_{\min}(\frac{1}{k}\mathbf{\Lambda}_k^{\mathrm{freq}}(\widehat{A}, \boldsymbol{U}) + \mathbf{\Lambda}_k^{\mathrm{noise}}(\widehat{A}, \sigma_u))$.

By Lemma 3.5.2 and given that $\boldsymbol{U}_i$ is formed identically to $\widehat{\boldsymbol{U}}$ in Lemma 3.5.2 with $\widehat{A} \leftarrow \widehat{A}_{i-1}$, on the event $\mathcal{E}_1$ and as long as T_{i-1} is large enough that

$$\epsilon_{\mathrm{op},i-1} \le c\cdot\min\left\{\frac{1-\rho_\star}{\tau_\star}, \frac{1}{\|A_\star\|_{\mathcal{H}_\infty}}, \frac{1}{\gamma^2\|A_\star\|_{\mathcal{H}_\infty}^3\|B_\star\|_{\mathrm{op}}^2}, \frac{(1-\rho_\star)^2}{\sigma_w^2+\sigma_u^2\|B_\star\|_{\mathrm{op}}^2}, \frac{1-\rho_\star}{\sigma_u^2\|B_\star\|_{\mathrm{op}}\tau_\star^2}\right\} \quad (3.5.2)$$

and

$$k_i \ge \max\left\{\frac{16\pi\|\theta_\star\|_{\mathcal{H}_\infty}\gamma^2}{\lambda_{\mathrm{noise}}(\sigma_u)}, \frac{\pi}{2\|\theta_\star\|_{\mathcal{H}_\infty}}\right\}\cdot\|A_\star\|_{\mathcal{H}_\infty}^2\|B_\star\|_{\mathrm{op}} + \frac{1}{2(1-\rho_\star)}\cdot\log\left(\frac{2(\sigma_w^2\tau_\star^2+\sigma_u^2\tau_\star^2\|B_\star\|_{\mathrm{op}}^2)}{\lambda_{\mathrm{noise}}(\sigma_u)(1-\rho_\star^2)}\right), \quad (3.5.3)$$

we will have that

$$\lambda_{\min}(\mathbf{\Lambda}^{\mathrm{ss}}_{k_i}(A_\star, \boldsymbol{U}_i, \tfrac{\gamma}{\sqrt{2d_u}})) \geq \frac{1}{8} \cdot \max_{\boldsymbol{U} \in \mathcal{U}_{\gamma^2, \bar{k}}} \lambda_{\min}(\mathbf{\Lambda}^{\mathrm{ss}}_{\bar{k}}(A_\star, \boldsymbol{U}, \tfrac{\gamma}{\sqrt{2d_u}}))$$

for any $\bar{k} \geq d_x$. It follows that, if (3.5.3) and (3.5.3) hold:

$$\lambda_{\min}(\mathbf{\Lambda}^{\mathrm{ss}}_{k_i}(A_\star, \boldsymbol{U}_i, \tfrac{\gamma}{\sqrt{2d_u}})) \geq \frac{1}{8} \cdot \limsup_{\bar{k} \to \infty} \max_{\boldsymbol{U} \in \mathcal{U}_{\gamma^2, \bar{k}}} \lambda_{\min}(\mathbf{\Lambda}^{\mathrm{ss}}_{\bar{k}}(A_\star, \boldsymbol{U}, 0)) = \frac{1}{8} \cdot \lambda^\star_{\min}.$$

Thus, on the event $\mathcal{E}_2$ and if (3.5.3) and (3.5.3) hold:

$$\|\widehat{A}_i - A_\star\|_{\mathrm{op}} \leq C \cdot \sqrt{\frac{\log 1/\delta + d_x \log(\beta_\infty/\lambda_{\mathrm{noise}} + 1)}{\lambda^\star_{\min} \cdot T_i}}.$$

To complete the proof, it suffices that we ensure i is large enough that (3.5.1), (3.5.3), and (3.5.3) hold. Recall that Algorithm 3.1 uses $T_i = C_{\mathrm{init}} d_u 2^i$ and $k_i = C_{\mathrm{init}} 2^{\lfloor i/4 \rfloor}$. We then have that

$$\frac{C^{3/4}_{\mathrm{init}}}{d^{1/4}_u} T_i^{1/4} \geq C_{\mathrm{init}} 2^{i/4} \geq k_i \geq C_{\mathrm{init}} 2^{i/4-1} = \frac{C^{3/4}_{\mathrm{init}}}{2d^{1/4}_u} T_i^{1/4}$$

and $T_i \geq T/2$. It follows that all above conditions will be met as long as

$$T \geq \mathrm{poly}\left(\gamma^2, \sigma_w^2, \|A_\star\|_{\mathcal{H}_\infty}, \|B_\star\|_{\mathrm{op}}, \tau_\star, \frac{1}{1-\rho_\star}, \frac{1}{\lambda_{\mathrm{noise}}}, \log\frac{1}{\delta}\right).$$

Finally, we apply Lemma 5.3.1 to bound $\tau_\star, \frac{1}{1-\rho_\star} \leq \mathrm{poly}(\|A_\star\|_{\mathcal{H}_\infty}, \|B_\star\|_{\mathrm{op}})$. The bound on $\|\widehat{A}_i - A_\star\|_{\mathrm{op}}$ follows. The bound on the input power used by Algorithm 3.1 follows from the following result.

Lemma 3.5.3. *Running Algorithm 3.1, we will have* $\frac{1}{T}\sum_{t=1}^{T} \mathbb{E}[u_t^\top u_t] \leq \gamma^2$.

Proof. By Proposition 3.4, we have that $\sum_{t=1}^{T_i} (\widetilde{u}_t^i)^\top \widetilde{u}_t^i \leq T_i \gamma^2/2$. Thus,

$$\sum_{t=1}^{T_i} \mathbb{E}[u_t^\top u_t] \leq T_i\gamma^2/2 + \sum_{t=1}^{T_i} \mathbb{E}[(u_t^w)^\top u_t^w] = T_i\gamma^2/2 + \sum_{t=1}^{T_i} \gamma^2/2 = T_i\gamma^2$$

Thus, the average expected input power for a given epoch is bounded by γ^2. It follows then that, after running for i epochs,

$$\frac{1}{T}\sum_{t=1}^{T} \mathbb{E}[u_t^\top u_t] = \frac{1}{T}\sum_{j=1}^{i}\sum_{t=1}^{T_j} \mathbb{E}[u_{t,j}^\top u_{t,j}] \leq \frac{1}{T}\sum_{j=1}^{i} T_j\gamma^2 = \gamma^2$$

where the last equality follows since, by definition, $T = \sum_{j=1}^{i} T_j$. $\square$

Chapter 4

Certainty Equivalent Decision Making

In the previous chapter, we considered the problem of identifying $A_\star$ in a linear dynamical system. While some applications may call for learning a uniformly good estimate of $A_\star$, in many applications, we want to identify the unknown parameters of our system for a specific purpose—for example, we may wish to obtain a controller which can steer our system to track a desired trajectory. In such settings, while we could learn $A_\star$ uniformly well and then use our knowledge of $A_\star$ to design an effective controller, it may be more effective to focus on learning only the portions of $A_\star$ that are relevant for control. In other words, we want to explore in a *task-dependent* manner, focusing our exploration on the parameters of $A_\star$ relevant for our task.

In this chapter we lay the groundwork for a general theory of task-dependent exploration, formalizing the setting presented in Section 1.1 for arbitrary *smooth* losses. We consider an observation model somewhat more general than the linear dynamical system setting, and assume that we have observations of the form:

$$y_t = \langle \theta_\star, z_t \rangle + w_t, \quad w_t \mid \mathcal{F}_{t-1} \sim \mathcal{N}(0, \sigma_w^2), \quad z_t \text{ is } \mathcal{F}_{t-1}\text{-adapted}, \tag{4.0.1}$$

for $\theta_\star \in \mathbb{R}^{d_\theta}$, a filtration $(\mathcal{F}_t)_{t \geq 1}$, and scalar observations y_t. We allow the distribution of the covariates z_t to be arbitrary: for example, there may be some function $f(\dots)$ of appropriate shape such that $z_t = f(t, z_{1:t-1}, y_{1:t-1}, w_{1:t}, u_{1:t}, \theta_\star)$, for inputs of our choosing $u_{1:t}$. We emphasize the generality of this observation model: (4.0.1) encompasses essentially any setting where we obtain linear observations of the parameter of interest. Indeed, beyond standard linear regression, this setting encompasses linear dynamical systems (as we show in Chapter 5), as well as more general nonlinear systems (Chapter 6).

Our goal in this setting is to learn some *decision* $\mathfrak{a} \in \mathbb{R}^{d_{\mathfrak{a}}}$ minimizing some *loss* $\mathcal{J}_{\theta_\star}(\mathfrak{a}) : \mathbb{R}^{d_{\mathfrak{a}}} \to \mathbb{R}$. For example, $\mathfrak{a}$ may correspond to the parameters of a controller, and $\mathcal{J}_{\theta_\star}(\mathfrak{a})$ the trajectory tracking error on the system with parameter $\theta_\star$ achieved by playing controller $\mathfrak{a}$. In general we assume we do not know $\theta_\star$ a priori, and to find a decision $\mathfrak{a}$ minimizing $\mathcal{J}_{\theta_\star}(\mathfrak{a})$, must estimate relevant features of $\theta_\star$. To this end, we consider the following interaction protocol.

Task-Specific Pure Exploration Problem. The learner's behavior is specified by an

exploration policy $\pi_{\mathrm{exp}} : (x_{1:t}, u_{1:t-1}) \to u_t$ and decision rule dec executed in the dynamics (4.0.1) as follows:

1. For steps $t = 1, \ldots, T$, the learner executes π_{exp} and collects a trajectory $\boldsymbol{\tau} = (y_{1:T}, z_{1:T}, u_{1:T})$, with inputs $u_{1:T}$ satisfying the constraint $\mathbb{E}_{\pi_{\mathrm{exp}}}[\sum_{t=1}^{T} \|u_t\|_2^2] \le T\gamma^2$.

2. The learner proposes a decision $\widehat{\mathfrak{a}} = \mathsf{dec}(\boldsymbol{\tau})$ as a function of $\boldsymbol{\tau}$.

3. The learner suffers loss $\mathcal{J}_{\theta_\star}(\widehat{\mathfrak{a}})$.

Recall from Chapter 3 that we denote all exploration policies satisfying $\mathbb{E}_{\pi_{\mathrm{exp}}}[\sum_{t=1}^{T} \|u_t\|_2^2] \le T\gamma^2$ as Π_{γ^2}. The focus of this chapter is on learning with a *fixed* exploration policy π_{exp}—our goal is to characterize the optimal decision rule dec given data generated by playing π_{exp} on (4.0.1).

Martingale Decision Making. Define the *excess risk* function induced by our loss as:

$$\mathcal{R}(\widehat{\mathfrak{a}}; \theta_\star) := \mathcal{J}_{\theta_\star}(\mathfrak{a}) - \inf_{\mathfrak{a}'} \mathcal{J}_{\theta_\star}(\mathfrak{a}'),$$

and the plug-in optimal decision as:

$$\mathfrak{a}_{\mathrm{opt}}(\theta) := \arg\min_{\mathfrak{a}} \mathcal{R}(\mathfrak{a}; \theta).$$

That is, $\mathfrak{a}_{\mathrm{opt}}(\theta)$ denotes the optimal decision on the model with parameter θ. In order to ensure decision making is tractable, we require that the excess risk function $\mathcal{R}(\cdot;\cdot)$, and the plug-in optimal decision $\mathfrak{a}_{\mathrm{opt}}(\theta)$ are smooth in a certain sense:

Assumption 4.1 (Smooth Decision Making)**.** *There exist $r_{\mathrm{quad}}(\theta_\star) > 0$ and constants $\mu > 0$, $L_{\mathfrak{a}i}, L_{\mathcal{R}i}$, $i \in \{1,2,3\}$, and L_{hess} such that for any θ and $\mathfrak{a}$ satisfying*

$$\|\theta - \theta_\star\|_2 \le r_{\mathrm{quad}}(\theta_\star), \quad \|\mathfrak{a} - \mathfrak{a}_{\mathrm{opt}}(\theta_\star)\|_2 \le L_{\mathfrak{a}1} r_{\mathrm{quad}}(\theta_\star), \tag{4.0.2}$$

the following conditions hold:

- *The optimal action $\mathfrak{a}_{\mathrm{opt}}(\theta)$ is unique, and moreover, there is a parameter μ such that $\mathcal{R}(\mathfrak{a}'; \theta) \ge \frac{\mu}{2}\|\mathfrak{a}' - \mathfrak{a}_{\mathrm{opt}}(\theta)\|_2^2$ for all $\mathfrak{a}' \in \mathbb{R}^{d_{\mathfrak{a}}}$ (not restricted to $\mathfrak{a}'$ satisfying (4.0.2)).*

- *$\|\nabla_{\mathfrak{a}} \mathcal{R}(\mathfrak{a}; \theta)\|_{\mathrm{op}} \le L_{\mathcal{R}1}$, $\|\nabla_{\mathfrak{a}}^2 \mathcal{R}(\mathfrak{a}; \theta)\|_{\mathrm{op}} \le L_{\mathcal{R}2}$, and $\|\nabla_{\mathfrak{a}}^3 \mathcal{R}(\mathfrak{a}; \theta)\|_{\mathrm{op}} \le L_{\mathcal{R}3}$.*

- *$\|\nabla_\theta \mathfrak{a}_{\mathrm{opt}}(\theta)\|_{\mathrm{op}} \le L_{\mathfrak{a}1}$, $\|\nabla_\theta^2 \mathfrak{a}_{\mathrm{opt}}(\theta)\|_{\mathrm{op}} \le L_{\mathfrak{a}2}$, and $\|\nabla_\theta^3 \mathfrak{a}_{\mathrm{opt}}(\theta)[\delta, \delta, \delta]\|_{\mathrm{op}} \le L_{\mathfrak{a}3}$ for all $\delta \in \mathbb{R}^{d_\theta}$ with $\|\delta\|_2 = 1$.*

- *$\nabla_{\mathfrak{a}}^2 \mathcal{R}(\mathfrak{a}; \theta)$ is Lipschitz in θ with Lipschitz constant L_{hess}.*

We also define:

$$L_{\text{quad}} := \frac{1}{6}(L_{\mathcal{R}3}L_{\mathfrak{a}1}^3 + 3L_{\mathcal{R}2}L_{\mathfrak{a}2}L_{\mathfrak{a}1} + L_{\mathcal{R}1}L_{\mathfrak{a}3}), \quad L_{\mathcal{H}} := 6L_{\text{quad}} + L_{\mathcal{R}2}L_{\mathfrak{a}1} + L_{\text{hess}}L_{\mathfrak{a}1}^2. \quad (4.0.3)$$

We are now ready to formally define our general decision making setting.

Definition 4.0.1 (Martingale Decision Making (MDM))**.** Assume our excess risk $\mathcal{R}(\widehat{\mathfrak{a}}; \theta_\star)$ satisfies Assumption 4.1, our observations are generated by (4.0.1), and we follow the task-specific pure exploration protocol stated above. Then we call the problem of choosing a decision $\widehat{\mathfrak{a}}$ to minimize $\mathcal{R}(\widehat{\mathfrak{a}}; \theta_\star)$ *martingale decision making* (MDM).

We emphasize the generality of this set of decision making problems. While we will show that the LQR falls within the MDM setting, many other decision making problems can be cast as an instance of MDM. For example, linear experiment design:

Example 4.0.1 (Linear Experimental Design)**.** Assume that we may choose z_t as we wish in (4.0.1), then this is the standard linear regression setting, and our framework therefore encompasses optimal linear experiment design in arbitrary smooth losses [Pukelsheim, 2006]. For example, with our decision $\mathfrak{a} \leftarrow \theta$, we may consider the A-optimal objective $\mathcal{J}_{\theta_\star}(\theta) = \|\theta - \theta_\star\|_M^2$ for some $M \succeq 0$. Alternatively, we could minimize the negative log-likelihood relative to some reference distribution ν so that $\mathcal{J}_{\theta_\star}(\theta) = \mathbb{E}_{Z\sim\nu, Y\sim p(\cdot|Z,\theta_\star)}[-\log(P(Y|Z,\theta))]$ where $P(Y|Z,\theta)$ is the likelihood of observations such that $y_t \sim p(\cdot|z_t, \theta_\star)$ [Chaudhuri and Mykland, 1993, Chaudhuri et al., 2015, Pronzato and Pázman, 2013]. Non-smooth G-optimal-like objectives such as $\mathcal{J}_{\theta_\star}(\theta) = \max_{x\in\mathcal{X}}\langle x, \theta - \theta_\star\rangle^2$ for some finite set $\mathcal{X} \subset \mathbb{R}^{d_\theta}$ can be captured in our framework by using an approximate smoothed objective such as $\mathcal{J}_{\theta_\star}(\theta) = \frac{1}{\lambda}\log\left(\sum_{x\in\mathcal{X}} e^{\lambda\langle x, \theta-\theta_\star\rangle^2}\right)$ for large λ.

Certainty Equivalence Decision Rule. We are particularly interested in the *certainty equivalence* decision rule [Theil, 1957, Simon, 1956]. Given a trajectory $\boldsymbol{\tau} = (y_{1:T}, z_{1:T}, u_{1:T})$, the least squares estimator of $\theta_\star$ is

$$\widehat{\theta}_{\text{ls}}(\boldsymbol{\tau}) = \arg\min_\theta \sum_{t=1}^T \|y_t - \langle \theta, z_t\rangle\|_2^2.$$

We then define the certainty equivalence decision rule as follows.

Definition 4.0.2 (Certainty Equivalence Decision Rule)**.** The *certainty equivalence* decision rule selects the optimal decision for the least-squares estimate of the dynamics: $\mathsf{ce}(\boldsymbol{\tau}) := \mathfrak{a}_{\text{opt}}(\widehat{\theta}_{\text{ls}}(\boldsymbol{\tau}))$.

4.1 Upper and Lower Bounds on Martingale Decision Making

Our goal is to obtain upper and lower bounds on decision making in the MDM setting. The key insight that lets us obtain such bounds is that, in our smooth decision making setting,

the problem of learning a good decision can be reduced to estimation of $\theta_\star$ in a particular, task-dependent norm. Given this reduction, Theorems 4.3 and 4.4, our upper and lower bounds on decision making, follow from novel upper and lower bounds on regression in general norms, the key technical contribution of this chapter. In what follows, we outline the reduction from decision making to estimation of $\theta_\star$, state our bounds on estimating and $\theta_\star$ in general norms, and provide formal results on decision making in MDM.

4.1.1 From Decision Making to Regression

To begin, under Assumption 4.1, our key smoothness assumption, we have that the following Lipschitz conditions hold.

Proposition 4.1. *Assume that $\mathcal{R}, \mathfrak{a}_{\mathrm{opt}}$ satisfy Assumption 4.1. Then for any model $\theta \in \mathbb{R}^{d_\theta}$ and action $\mathfrak{a} \in \mathbb{R}^{d_{\mathfrak{a}}}$ satisfying (4.0.2), it holds that*

- *$\nabla_{\mathfrak{a}}^{(i)} \mathcal{R}(\mathfrak{a}; \theta)$ is Lipschitz in the operator norm with Lipschitz constant $L_{\mathcal{R}(i+1)}$ for $i = 0, 1, 2$*
- *$\nabla_{\theta}^{(i)} \mathfrak{a}_{\mathrm{opt}}(\theta)$ is Lipschitz in the operator norm with Lipschitz constant $L_{\mathfrak{a}(i+1)}$, for $i = 0, 1$.*

In the above, we adopted the convention $\nabla_x^{(0)} f(x) = f(x)$.

The next step is to relate smooth decision making to M-norm estimation. We begin by introducing the relevant gradients and Hessians, and in particular, the *model-task Hessian* $\mathcal{H}(\theta)$.

Definition 4.1.1 (Key Gradients and Hessians)**.** For some $\theta_\star$ and function $\mathcal{R}, \mathfrak{a}_{\mathrm{opt}}$, let:

- $M_{\mathfrak{a}}(\theta_\star) := \nabla_{\mathfrak{a}}^2 \mathcal{R}(\mathfrak{a}; \theta_\star)$ at $\mathfrak{a} = \mathfrak{a}_{\mathrm{opt}}(\theta_\star)$.
- $G_{\mathfrak{a}}(\theta_\star) := \nabla_\theta \mathfrak{a}_{\mathrm{opt}}(\theta)$ at $\theta = \theta_\star$.
- $\mathcal{H}(\theta_\star) = \nabla_\theta^2 \mathcal{R}(\mathfrak{a}_{\mathrm{opt}}(\theta); \theta_\star)$ at $\theta = \theta_\star$. In particular, $\mathcal{H}(\theta_\star) = G_{\mathfrak{a}}(\theta_\star)^\top M_{\mathfrak{a}}(\theta_\star) G_{\mathfrak{a}}(\theta_\star)$.

The following result utilizes Assumption 4.1 to guarantee that $\mathcal{H}(\theta)$ is itself a smooth map, and that the norm induced by $\mathcal{H}(\theta_\star)$ can be used to approximate $\mathcal{R}(\mathfrak{a}_{\mathrm{opt}}(\widehat{\theta}); \theta_\star)$, both of which are critical pieces in our analysis.

Proposition 4.2. *Assume that $\mathcal{R}, \mathfrak{a}_{\mathrm{opt}}$ satisfy Assumption 4.1 and that $\widehat{\theta}$ satisfies $\|\widehat{\theta} - \theta_\star\|_2 \le r_{\mathrm{quad}}(\theta_\star)$. Then the following hold:*

$$\left|\mathcal{R}(\mathfrak{a}_{\mathrm{opt}}(\widehat{\theta}); \theta_\star) - \|\widehat{\theta} - \theta_\star\|_{\mathcal{H}(\theta_\star)}^2\right| \le L_{\mathrm{quad}} \cdot \|\widehat{\theta} - \theta_\star\|_2^3, \quad \|\mathcal{H}(\theta_\star) - \mathcal{H}(\widehat{\theta})\|_{\mathrm{op}} \le L_{\mathcal{H}} \cdot \|\widehat{\theta} - \theta_\star\|_2.$$

for L_{quad} and $L_{\mathcal{H}}$ as given in (4.0.3).

Note that this immediately implies an upper bound on decision making with the certainty-equivalent decision rule:

$$\mathcal{R}(\mathfrak{a}_{\text{opt}}(\widehat{\theta}); \theta_\star) \le \|\widehat{\theta} - \theta_\star\|^2_{\mathcal{H}(\theta_\star)} + L_{\text{quad}} \cdot \|\widehat{\theta} - \theta_\star\|_2^3.$$

Thus, if we can control the estimation error in the $\mathcal{H}(\theta_\star)$ norm, we can provide a bound on the suboptimality of our decision. The argument for the lower bound is somewhat more subtle, as we wish to obtain a lower bound on *all* possible decision rules, not just the certainty equivalent rule. The following result shows that we can also lower bound the suboptimality by the estimation error in the $\mathcal{H}(\theta_\star)$ norm.

Lemma 4.1.1. *Assume that the excess risk $\mathcal{R}$ and optimal-decision function $\mathfrak{a}_{\text{opt}}$ satisfy Assumption 4.1. Let $r > 0$ be a radius parameter satisfying*

$$r \le r_{\text{quad}}(\theta_\star)/4$$

and define the associated balls $\mathcal{B}_T(\theta_\star) := \{\theta : \|\theta - \theta_\star\|_2 \le r\}$. Then,

$$\min_{\widehat{\mathfrak{a}}} \max_{\theta \in \mathcal{B}_T(\theta_\star)} \mathbb{E}_{\theta, \pi_{\text{exp}}}[\mathcal{R}(\widehat{\mathfrak{a}}; \theta)] \ge \min \left\{ \min_{\widehat{\theta}} \max_{\theta \in \mathcal{B}_T(\theta_\star)} \frac{1}{2} \mathbb{E}_{\theta, \pi_{\text{exp}}} \left[\|\widehat{\theta} - \theta\|^2_{\mathcal{H}(\theta_\star)} \right] - C_1 r^3, \mu L_{\mathfrak{a}1}^2 r^2 \right\}$$

where we define the constant C_1, for a universal numerical constant c_1,

$$C_1 = c_1 \Big(L_{\mathfrak{a}1} L_{\mathfrak{a}2} L_{\mathcal{R}2} + L_{\mathfrak{a}1}^3 L_{\mathcal{R}3} + L_{\text{hess}} \Big).$$

Proof Sketch. Our goal is to show that $\mathbb{E}_{\theta, \pi_{\text{exp}}}[\mathcal{R}(\widehat{\mathfrak{a}}; \theta)]$ can be lower bounded by the estimation error of θ in the $\|\cdot\|_{\mathcal{H}(\theta_\star)}$ norm. While Proposition 4.2, shows that this equivalence is true when $\widehat{\mathfrak{a}}$ is the certainty equivalence estimate, here we want to show that this is true for *any* estimate. To this end, we define

$$\begin{aligned} \delta_\star(\widehat{\mathfrak{a}}) &:= \arg\min_{\delta} \|M_{\mathfrak{a}}(\theta_\star)^{1/2} \left(\widehat{\mathfrak{a}} - \mathfrak{a}_{\text{opt}}(\theta_\star) - G_{\mathfrak{a}}(\theta_\star)\delta\right) \|_2 \\ &= (M_{\mathfrak{a}}(\theta_\star)^{1/2} G_{\mathfrak{a}}(\theta_\star))^\dagger M_{\mathfrak{a}}(\theta_\star)^{1/2} (\widehat{\mathfrak{a}} - \mathfrak{a}_{\text{opt}}(\theta_\star)) \end{aligned}$$

and define the induced estimate

$$\widehat{\theta}(\widehat{\mathfrak{a}}) := \theta_\star + \delta_\star(\widehat{\mathfrak{a}})$$

Intuitively, we would expect $\widehat{\theta}(\widehat{\mathfrak{a}})$ to be close to θ in an appropriate metric if our decision $\widehat{\mathfrak{a}}$ achieves a small excess risk, $\mathcal{R}(\widehat{\mathfrak{a}}; \theta)$. By carefully Taylor expanding both $\mathcal{R}(\widehat{\mathfrak{a}}; \theta)$ and $\mathfrak{a}_{\text{opt}}(\theta)$, we show that this is the case, writing the excess risk $\mathcal{R}(\widehat{\mathfrak{a}}; \theta)$ as the sum of $\|\widehat{\theta}(\widehat{\mathfrak{a}}) - \theta\|^2_{\mathcal{H}(\theta_\star)}$ and a $\mathcal{O}(r^3)$ term. This implies that if $\mathcal{R}(\widehat{\mathfrak{a}}; \theta)$ is small, $\|\widehat{\theta}(\widehat{\mathfrak{a}}) - \theta\|^2_{\mathcal{H}(\theta_\star)}$ will also be small, which reduces the problem of minimizing $\mathcal{R}(\widehat{\mathfrak{a}}; \theta)$ to that of estimating θ in the $\|\cdot\|_{\mathcal{H}(\theta_\star)}$ norm.

As $\widehat{\theta}(\widehat{\mathfrak{a}})$ is a particular estimator of θ given our trajectory, it follows that the resulting loss is lower bounded by minimizing over all estimators, $\widehat{\theta}$, which gives the result. We defer the details of this argument to Appendix B.1.1. □

4.1.2 Upper and Lower Bounds on M-norm Regression

The above results show that in our decision making setting, both upper and lower bounds on decision making scale with the estimation error of $\theta_\star$ in the $\mathcal{H}(\theta_\star)$-norm. Obtaining tight upper and lower bounds on decision making therefore reduces to obtaining tight upper and lower bounds on regression in the $\mathcal{H}(\theta_\star)$-norm. In this section, we state our results on regression in general norms. To the best of our knowledge, these results are the most general bounds on linear regression in the literature.

We will consider the setting of regression in some general M-norm. That is, our aim is to produce an estimator $\widehat{\theta} \in \mathbb{R}^{d_\theta}$ so as to minimize the following weighted estimation loss:

$$\mathcal{R}_{\mathrm{est}}(\widehat{\theta}; \theta) := \|\widehat{\theta} - \theta\|_M^2, \quad M \succeq 0. \tag{4.1.1}$$

In this section, we state our upper and lower bounds on $\mathcal{R}_{\mathrm{est}}(\widehat{\theta}; \theta)$. Throughout, we let $\mathbb{E}_\theta$ and $\mathbb{P}_\theta$ denote probabilities and expectations with respect to (4.0.1) when $\theta_\star = \theta$.

We will show that the least squares estimator

$$\widehat{\theta}_{\mathrm{ls}} := \left(\sum_{t=1}^T z_t z_t^\top\right)^{-1} \sum_{t=1}^T z_t y_t \tag{4.1.2}$$

is the optimal estimator of $\theta_\star$ for the risk in (4.1.1), in a very strong, instance dependent sense. Throughout, the central object of our analysis is the random covariance matrix:

$$\mathbf{\Sigma}_T := \sum_{t=1}^T z_t z_t^\top.$$

Let us start with the lower bound. We will call an estimator $\widehat{\theta}$ *measurable* if $\widehat{\theta}$ is a measurable function of the covariates and responses $(y_t, z_t : 1 \le t \le T)$, and possibly some internal randomness.

Theorem 4.1. *Let $\widehat{\theta}$ be an arbitrary measurable estimator. Moreover, fix a covariance parameter $\mathbf{\Lambda} \in \mathbb{S}_{++}^{d_\theta}$, nominal instance $\theta_0 \in \mathbb{R}^{d_\theta}$, and radius $r \ge \sqrt{5\mathrm{tr}(\mathbf{\Lambda}^{-1})}$. Let $\mathcal{B} := \{\theta : \|\theta - \theta_0\|_2 \le r\}$ denote a Euclidean ball around θ_0. Then, it holds that*

$$\inf_{\widehat{\theta}} \max_{\theta \in \mathcal{B}} \mathbb{E}_\theta[\mathcal{R}_{\mathrm{est}}(\widehat{\theta}; \theta)] \ge \sigma_w^2 \cdot \min_{\theta \in \mathcal{B}} \mathrm{tr}\left(M \cdot (\mathbb{E}_\theta[\mathbf{\Sigma}_T] + \mathbf{\Lambda})^{-1}\right) - \Psi(r; \mathbf{\Lambda}, M),$$

where $\Psi(r; \mathbf{\Lambda}, M) := \frac{32\|M\|_{\mathrm{op}}}{\lambda_{\min}(\mathbf{\Lambda})} \exp(-\frac{r^2}{5}\lambda_{\min}(\mathbf{\Lambda}))$.

In our applications, we shall choose r sufficiently large and $\mathbf{\Lambda}$ sufficiently small so that

the lower bound reads

$$\inf_{\widehat{\theta}} \max_{\theta \in \mathcal{B}} \mathbb{E}_\theta[\mathcal{R}_{\mathrm{est}}(\widehat{\theta}; \theta)] \gtrsim \sigma_w^2 \cdot \min_{\theta \in \mathcal{B}} \mathrm{tr}\left(M \mathbb{E}_\theta[\mathbf{\Sigma}_T]^{-1}\right); \tag{4.1.3}$$

in other words, that the M-weighted trace of the inverse covariance matrix lower bounds the risk. Even though the right-hand side considers the minimum over $\theta \in \mathcal{B}$, the radius r of $\mathcal{B}$ can be chosen small enough that this quantity does not vary significantly, under certain regularity conditions. For a sense of scaling $\mathbb{E}_\theta[\mathbf{\Sigma}_T]$ will typically scale like $\Omega(T)$, by choosing $\mathbf{\Lambda} \preceq o(T)$, and $r^2 \propto \mathrm{tr}(\mathbf{\Lambda}^{-1}) \log^2(T)$, the term $\Psi(r, \mathbf{\Lambda}, M)$ vanishes as $T^{-\omega(1)}$, and the approximation (4.1.3) holds. Moreover, since this scaling of r vanishes at a rate of $\log^2 T/\sqrt{T}$, r is small enough so as to ensure $\mathbb{E}_\theta[\mathbf{\Sigma}_T]$ does not vary significantly on $\mathcal{B}$.

Remark 4.1.1 (Comparison to Previous Lower Bounds)**.** Lower bounds for regression are typically derived from the Cramer-Rao bound (e.g, in Chaudhuri et al. [2015]), which applies only to unbiased estimators, and does not rule out more efficient estimation by allowing bias. In contrast, our work provides an *unconditional* information theoretic lower bound, derived from a closed-form computation of an expected Bayes risk in linear regression with a Gaussian prior (Theorem 4.1). This technique is similar in spirit to the Van Trees inequality [Gill et al., 1995] which was used in concurrent work to understand instance-optimal regret in LQR when $A_\star$ is known but $B_\star$ is not [Ziemann and Sandberg, 2021]. Another common technique for adaptive estimation lower bounds is Assouad's method [Arias-Castro et al., 2012, Simchowitz and Foster, 2020], though this method typically yields worst-case (not sharp, instance-dependent) lower bounds.

We turn now to our upper bound.

Theorem 4.2. *Fix any matrices $\mathbf{\Lambda} \in \mathbb{S}_{++}^{d_\theta}, M \in \mathbb{S}_{+}^{d_\theta}$, with $M \neq 0$. Given a parameter $\beta \in (0, 1/4)$, define the event*

$$\mathcal{E} := \{\|\mathbf{\Sigma}_T - \mathbf{\Lambda}\|_{\mathrm{op}} \leq \beta \lambda_{\min}(\mathbf{\Lambda})\}.$$

Then, if $\mathcal{E}$ holds, the following holds with probability $1 - \delta$:

$$\|\widehat{\theta}_{\mathrm{ls}} - \theta_\star\|_M^2 \leq 5(1 + \alpha) \cdot \sigma_w^2 \log \frac{6 d_\theta}{\delta} \cdot \mathrm{tr}(M \mathbf{\Lambda}^{-1}),$$

where $\alpha := 26\beta^2 \lambda_{\max}(\mathbf{\Lambda}) \mathrm{tr}(\mathbf{\Lambda}^{-1})$.

Together, Theorem 4.1 and Theorem 4.2 provide nearly matching upper and lower bounds on M-norm regression, with both scaling as $\mathrm{tr}(M \cdot \mathbf{\Lambda}^{-1})$, for $\mathbf{\Lambda}$ behaving as the covariates.

4.1.3 Lower Bounds for Decision Making in MDM

We have shown that the upper and lower bounds on decision making in MDM scale with $\|\widehat{\theta} - \theta_\star\|_{\mathcal{H}(\theta_\star)}^2$, and that $\|\widehat{\theta} - \theta_\star\|_{\mathcal{H}(\theta_\star)}^2$ scales as, approximately, $\mathrm{tr}(\mathcal{H}(\theta_\star) \cdot \mathbf{\Sigma}_T^{-1})$, where $\mathbf{\Sigma}_T$

denotes the covariance used to obtain $\widehat{\theta}$. Given this, we are now ready to present our final upper and lower bounds in the MDM setting.

Throughout this section, for a fixed exploration policy π_{exp}, we let $\mathbf{\Lambda}_T$ denote the expected covariates obtained on (4.0.1):

$$\mathbf{\Lambda}_T(\pi_{\mathrm{exp}}; \theta) := \frac{1}{T}\mathbb{E}_{\theta, \pi_{\mathrm{exp}}}\left[\mathbf{\Sigma}_T\right].$$

We also define the *idealized risk* as follows.

Definition 4.1.2 (Idealized Risk)**.** We define the idealized risk for policy π_{exp} as:

$$\Phi_T(\pi_{\mathrm{exp}}; \theta_\star) := \mathrm{tr}(\mathcal{H}(\theta_\star)\mathbf{\Lambda}_T(\pi_{\mathrm{exp}}; \theta_\star)^{-1}),$$

where we recall that $\mathcal{H}(\theta_\star) := \nabla^2_\theta \mathcal{R}(\mathfrak{a}_{\mathrm{opt}}(\theta); \theta_\star)\big|_{\theta=\theta_\star}$.

In proving our bounds, we assume we are playing a fixed exploration policy π_{exp}. We will state our lower bound in terms of the *local minimax risk* under π_{exp}:

Definition 4.1.3 (Decision Making Local Minimax Risk)**.** Let $\mathcal{B}$ denote a subset of instances. The T-sample local minimax decision risk on $\mathcal{B}$ under exploration policy π_{exp} is

$$\mathfrak{M}_{\pi_{\mathrm{exp}}}(\mathcal{R}; \mathcal{B}) := \min_{\mathsf{dec}} \max_{\theta \in \mathcal{B}} \mathbb{E}_{\boldsymbol{\tau} \sim \theta, \pi_{\mathrm{exp}}}\left[\mathcal{R}(\mathsf{dec}(\boldsymbol{\tau}); \theta)\right]$$

where the minimization is over all maps from trajectories to decisions. Typically, we shall let $\mathcal{B}$ take the form $\mathcal{B}_2(r; \theta_\star) := \{\theta : \|\theta - \theta_\star\|_2 \le r\}$.

By choosing $\mathcal{B}$ to contain only instances close to $\theta_\star$, the local minimax risk captures the difficulty of decision making on the specific instance $\theta_\star$, yielding an effectively instance-specific lower bound. For our lower bound to hold, the covariance matrices in question must satisfy two rather mild regularity conditions.

Assumption 4.2 (Sufficient Excitation)**.** *For some $\underline{\lambda} > 0$ independent of T, and under our exploration policy $\pi_{\mathrm{exp}} \in \Pi_{\gamma^2}$:*

$$\lambda_{\min}(\mathbf{\Lambda}_T(\pi_{\mathrm{exp}}; \theta_\star)) \ge \underline{\lambda}.$$

In the special case of linear dynamical systems, Assumption 4.2 can be enforced by adding a small amount of white noise to any exploration policy, and the budget constraint can still be met by scaling down inputs by a constant factor.

Assumption 4.3 (Smooth Response)**.** *There exist parameters $r_{\mathrm{cov}}(\theta_\star) > 0$, $L_{\mathrm{cov}}(\theta_\star, \gamma^2) > 0$, $C_{\mathrm{cov}} > 0$, $c_{\mathrm{cov}} > 0$, and $\alpha > 0$ such that, under our exploration policy $\pi_{\mathrm{exp}} \in \Pi_{\gamma^2}$, for all θ satisfying $\|\theta - \theta_\star\|_2 \le r_{\mathrm{cov}}(\theta_\star)$, we have:*

$$\mathbf{\Lambda}_T(\pi_{\mathrm{exp}}; \theta) \preceq c_{\mathrm{cov}}\mathbf{\Lambda}_T(\pi_{\mathrm{exp}}; \theta_\star) + \left(L_{\mathrm{cov}}(\theta_\star, \gamma^2) \cdot \|\theta - \theta_\star\|_2 + \frac{C_{\mathrm{cov}}}{T^\alpha}\right) \cdot I.$$

Intuitively, Assumption 4.3 says that the covariance matrices do not vary too much in the ground truth instances. Under these assumptions, we obtain the following lower bound.

Theorem 4.3. *Assume we are in the* MDM *setting, that $\mathcal{R}$ satisfies Assumption 4.1, our exploration policy $\pi_{\mathrm{exp}} \in \Pi_{\gamma^2}$ satisfies Assumption 4.2 and Assumption 4.3, and suppose that the time horizon T satisfies*

$$\underline{\lambda} T \geq \max\left\{ \left(\frac{80 d_\theta}{r_{\mathrm{quad}}(\theta_\star)^2}\right)^{6/5}, \left(\frac{\sigma_w^2 L_{\mathcal{R}2}}{5\mu}\right)^6, \left(\frac{L_{\mathrm{cov}}(\theta_\star,\gamma^2)\sqrt{5d_\theta}}{c_{\mathrm{cov}}\underline{\lambda}}\right)^{12/5}, \left(\frac{2C_{\mathrm{cov}}}{c_{\mathrm{cov}}\underline{\lambda}^{1-\alpha}}\right)^{1/\alpha}, \left(\frac{5d_\theta}{r_{\mathrm{cov}}(\theta_\star)^2}\right)^{6/5} \right\}$$

Then, defining the localizing ball $\mathcal{B}_T := \{\theta : \|\theta - \theta_\star\|_2^2 \leq 5d_\theta/(\underline{\lambda}T)^{5/6}\}$, we have

$$\mathfrak{M}_{\pi_{\mathrm{exp}}}(\mathcal{R}; \mathcal{B}_T) = \min_{\mathsf{dec}} \max_{\theta \in \mathcal{B}_T} \mathbb{E}_{\boldsymbol{\tau}\sim\theta,\pi_{\mathrm{exp}}}[\mathcal{R}(\mathsf{dec}(\boldsymbol{\tau}); \theta)] \geq \frac{\sigma_w^2}{1+2c_{\mathrm{cov}}} \cdot \frac{\Phi_T(\pi_{\mathrm{exp}}; \theta_\star)}{T} - \frac{C_{\mathrm{lb}}}{(\underline{\lambda}T)^{5/4}}$$

where $C_{\mathrm{lb}} = c_1((L_{\mathfrak{a}1}L_{\mathfrak{a}2}L_{\mathcal{R}2} + L_{\mathfrak{a}1}^3 L_{\mathcal{R}3} + L_{\mathrm{hess}})d_\theta^{3/2} + L_{\mathfrak{a}1}^2 L_{\mathcal{R}2})$ for a universal constant c_1.

Theorem 4.3 shows that, for *any* decision rule, the loss of the learned decision scales, up to constants, as at least $\frac{\sigma_w^2}{T} \cdot \Phi_T(\pi_{\mathrm{exp}}; \theta_\star)$. This result is itself a corollary of a more general result, Theorem 4.5, which provides a lower bound without Assumption 4.2 or Assumption 4.3. We prove this result in Appendix B.1. We emphasize again the generality of this result: Theorem 4.3 holds for *any* loss satisfying Assumption 4.1 so long as our observations follow (4.0.1).

4.1.4 Upper Bounds for Certainty Equivalence Decision Making in MDM

We next consider upper bounds on decision making in the MDM setting when we are playing a fixed exploration policy π_{exp}. The following is a sufficient assumption on π_{exp} to guarantee the efficiency of certainty equivalence decision making. Recall that $\boldsymbol{\Sigma}_T$ denotes the random covariates.

Assumption 4.4 (Exploration Policy Regularity)**.** *We assume that the true instance $\theta_\star$ and policy $\pi_{\mathrm{exp}} \in \Pi_{\gamma^2}$ satisfy the following regularity conditions:*

- *There exists some time $T_{\mathrm{se}}(\pi_{\mathrm{exp}})$ such that for any $T \geq T_{\mathrm{se}}(\pi_{\mathrm{exp}})$ the system is sufficiently excited. That is, if $T \geq T_{\mathrm{se}}(\pi_{\mathrm{exp}})$:*

$$\mathbb{P}_{\theta_\star,\pi_{\mathrm{exp}}}\left[\lambda_{\min}(\boldsymbol{\Sigma}_T) \geq \underline{\lambda} T, \boldsymbol{\Sigma}_T \preceq T\bar{\boldsymbol{\Lambda}}_T\right] \geq 1 - \delta$$

  for deterministic $\underline{\lambda} > 0$ and $\bar{\boldsymbol{\Lambda}}_T \succeq 0$.

- *There exists some time $T_{\mathrm{con}}(\pi_{\mathrm{exp}})$ such that, for any $T \geq T_{\mathrm{con}}(\pi_{\mathrm{exp}})$, the covariates have*

concentrated to their mean. That is, if $T \geq T_{\mathrm{con}}(\pi_{\mathrm{exp}})$*:*

$$\mathbb{P}_{\theta_\star,\pi_{\mathrm{exp}}}\left[\|\mathbf{\Sigma}_T - \mathbb{E}_{\theta_\star,\pi_{\mathrm{exp}}}[\mathbf{\Sigma}_T]\|_{\mathrm{op}} \leq \tfrac{C_{\mathrm{con}}}{T^\alpha}\lambda_{\min}(\mathbb{E}_{\theta_\star,\pi_{\mathrm{exp}}}[\mathbf{\Sigma}_T])\right] \geq 1-\delta$$

for deterministic $C_{\mathrm{con}} > 0$ *and* $\alpha > 0$*.*

The following result precisely quantifies the loss of the certainty equivalence decision-making rule under this assumption on the policy.

Theorem 4.4. *Assume we are in the* MDM *setting with some loss* $\mathcal{R}$ *satisfying Assumption 4.1 and exploration policy* $\pi_{\mathrm{exp}} \in \Pi_{\gamma^2}$ *which satisfies Assumption 4.4 with minimal times* $T_{\mathrm{con}}(\pi_{\mathrm{exp}})$ *and* $T_{\mathrm{se}}(\pi_{\mathrm{exp}})$ *and covariance lower bound* $\underline{\lambda} > 0$*. If*

$$T > \max\left\{T_{\mathrm{con}}(\pi_{\mathrm{exp}}), T_{\mathrm{se}}(\pi_{\mathrm{exp}}), (4C_{\mathrm{con}})^{1/\alpha}, \tfrac{c_1(\log(1/\delta)+d_\theta+\log\det(\bar{\mathbf{\Lambda}}_T/\underline{\lambda}+I))}{\underline{\lambda} r_{\mathrm{quad}}(\theta_\star)^2}\right\} \tag{4.1.4}$$

then for $\delta \in (0,1/2)$*, with probability* $1-\delta$*, the certainty equivalence decision rule achieves the following rate,*

$$\mathcal{R}(\mathfrak{a}_{\mathrm{opt}}(\widehat{\theta}_{\mathrm{ls}});\theta_\star) \leq 5\sigma_w^2 \log\frac{24 d_\theta}{\delta}\cdot\frac{\Phi_T(\pi_{\mathrm{exp}};\theta_\star)}{T} + \frac{C_{\mathrm{ce},1}}{T^{3/2}} + \frac{C_{\mathrm{ce},2}}{T^{1+2\alpha}}$$

where we let c_1, c_2, c_3 *be universal numerical constants and set*

$$C_{\mathrm{ce},1} := \tfrac{c_2 L_{\mathrm{quad}}}{\underline{\lambda}^{3/2}}\Big(\log\tfrac{1}{\delta} + d_\theta + \log\det(\bar{\mathbf{\Lambda}}_T/\underline{\lambda}+I)\Big)^{3/2}, \quad C_{\mathrm{ce},2} := \tfrac{c_3\sigma_w^2 C_{\mathrm{con}}^2 d_\theta \mathrm{tr}(\mathcal{H}(\theta_\star))}{\underline{\lambda}}\log\frac{d_\theta}{\delta}.$$

Theorem 4.4 shows that, up to constants and log factors, the loss achieved by the certainty equivalence decision rule scales as $\frac{\sigma_w^2}{T}\cdot\Phi_T(\pi_{\mathrm{exp}};\theta_\star)$. Note that this precisely matches the lower bound given in Theorem 4.3. This shows that the certainty equivalence decision rule is instance optimal for *any* decision making problem in the MDM setting.

The burn-in time (4.1.4) and lower-order terms have transparent interpretations. For the burn-in, the requirement that T be larger than $T_{\mathrm{con}}(\pi_{\mathrm{exp}})$ and $T_{\mathrm{se}}(\pi_{\mathrm{exp}})$ is necessary to ensure that the concentration and excitation events stated in Assumption 4.4 hold with high probability. The requirement that T be larger that $(4C_{\mathrm{con}})^{1/\alpha}$ is necessary to ensure that the covariates have concentrated enough for our M-norm estimation bound, Theorem 4.2, to hold. Finally, the last term in the burn-in ensures that our estimate $\widehat{\theta}_{\mathrm{ls}}$ is in a ball of radius $r_{\mathrm{quad}}(\theta_\star)$ around $\theta_\star$, which allows us to approximate $\mathcal{R}(\mathfrak{a}_{\mathrm{opt}}(\widehat{\theta}_{\mathrm{ls}});\theta_\star)$ by a quadratic. The lower order terms similarly yield intuitive explanations. $C_{\mathrm{ce},1}/T^{3/2}$ quantifies the additional loss due to the error in our quadratic approximation of $\mathcal{R}(\mathfrak{a}_{\mathrm{opt}}(\widehat{\theta}_{\mathrm{ls}});\theta_\star)$, while $C_{\mathrm{ce},2}/T^{1+2\alpha}$ is due to the lower order term given in our M-norm estimation bound, Theorem 4.2.

4.2 Proof of Theorem 4.3 and Theorem 4.4

We next provide the proofs of Theorem 4.3 and Theorem 4.4.

4.2.1 Proof of Theorem 4.3

Theorem 4.3 is based on the following more general result, and indeed, can be derived as a simple corollary of this result under Assumption 4.2 and Assumption 4.3

Theorem 4.5. *Suppose the smoothness assumption Assumption 4.1 holds with its stated smoothness parameters. In addition, fix a regularization parameter $\lambda > 0$, and suppose that T satisfies*

$$\lambda T \geq \max\left\{\left(\frac{80 d_\theta}{r_{\mathrm{quad}}(\theta_\star)^2}\right)^{6/5}, \left(\frac{\sigma_w^2 L_{\mathcal{R}2}}{5\mu}\right)^6\right\}$$

Finally, define the localizing ball $\mathcal{B}_T := \{\theta : \|\theta - \theta_\star\|_2^2 \leq 5 d_\theta/(\lambda T)^{5/6}\}$. Then, for any $\theta_0 \in \mathcal{B}_T(\theta_\star)$,

$$\min_{\widehat{\mathfrak{a}}} \max_{\theta \in \mathcal{B}_T} \mathbb{E}_{\theta, \pi_{\mathrm{exp}}}[\mathcal{R}(\widehat{\mathfrak{a}}; \theta)] \geq \sigma_w^2 \min_{\theta \in \mathcal{B}_T} \mathrm{tr}\left(\mathcal{H}(\theta_\star)\left(\mathbb{E}_{\theta, \pi_{\mathrm{exp}}}[\mathbf{\Sigma}_T] + \lambda T \cdot I\right)^{-1}\right) - \frac{C_2}{(\lambda T)^{5/4}}.$$

where we have defined the constant, for a universal numerical constant c_2,

$$C_2 = c_2\Big((L_{\mathfrak{a}1} L_{\mathfrak{a}2} L_{\mathcal{R}2} + L_{\mathfrak{a}1}^3 L_{\mathcal{R}3} + L_{\mathrm{hess}}) d_\theta^{3/2} + L_{\mathfrak{a}1}^2 L_{\mathcal{R}2}\Big).$$

Proof. For simplicity, we shall write $r^2 = 5 d_\theta/(\lambda T)^{5/6}$. The result follows by instantiating Theorem 4.1 to lower bound:

$$\min_{\widehat{\theta}} \max_{\theta : \|\theta - \theta_\star\|_2 \leq r} \mathbb{E}_{\theta, \pi_{\mathrm{exp}}}[\|\widehat{\theta} - \theta\|_{M(\theta_\star)}^2],$$

via the simplification provided by Lemma 4.1.1. Apply Theorem 4.1 with parameters $\mathbf{\Lambda} = \lambda T \cdot I$, so that $\mathrm{tr}(\mathbf{\Lambda}^{-1}) = \frac{d_\theta}{\lambda T}$ and $\lambda_{\min}(\mathbf{\Lambda}) = \lambda T$. Then, the remainder term Ψ from that theorem is bounded by

$$\Psi(r; \mathbf{\Lambda}, \mathcal{H}(\theta_\star)) \leq \frac{32 \|\mathcal{H}(\theta_\star)\|_{\mathrm{op}}}{\lambda T} \exp\left(-\frac{1}{5} r^2 \lambda T\right)$$

Selecting $r^2 = 5 d_\theta/(\lambda T)^{5/6}$ yields

$$\Psi(r; \mathbf{\Lambda}, \mathcal{H}(\theta_\star)) \leq \frac{32 \|\mathcal{H}(\theta_\star)\|_{\mathrm{op}}}{\lambda T} \exp\left(-d_\theta (\lambda T)^{1/6}\right)$$

Noting that $r^2 \geq 5\mathrm{tr}(\mathbf{\Lambda})$, Theorem 4.1 yields that

$$\begin{aligned}
&\min_{\widehat{\theta}} \max_{\theta:\|\theta-\theta_\star\|_2\leq r} \mathbb{E}_{\theta,\pi_{\mathrm{exp}}}[\|\widehat{\theta}-\theta\|^2_{\mathcal{H}(\theta_\star)}] \\
&\geq \sigma_w^2 \min_{\theta:\|\theta-\theta_\star\|_2\leq r} \mathrm{tr}\left(\mathcal{H}(\theta_\star)\left(\mathbb{E}_{\theta,\pi_{\mathrm{exp}}}[\mathbf{\Sigma}_T]+\lambda T\cdot I\right)^{-1}\right) - \frac{32\|\mathcal{H}(\theta_\star)\|_{\mathrm{op}}}{\lambda T}\exp\left(-d_\theta(\lambda T)^{1/6}\right) \\
&\geq \sigma_w^2 \min_{\theta:\|\theta-\theta_\star\|_2\leq r} \mathrm{tr}\left(\mathcal{H}(\theta_\star)\left(\mathbb{E}_{\theta,\pi_{\mathrm{exp}}}[\mathbf{\Sigma}_T]+\lambda T\cdot I\right)^{-1}\right) - \frac{32L_{\mathfrak{a}1}^2 L_{\mathcal{R}2}}{\lambda T}\exp\left(-d_\theta(\lambda T)^{1/6}\right),
\end{aligned}$$

where in the last line, we invoked Assumption 4.1 to obtain

$$\|\mathcal{H}(\theta_\star)\|_{\mathrm{op}} = \|G_{\mathfrak{a}}(\theta_\star)^\top M_{\mathfrak{a}}(\theta_\star)G_{\mathfrak{a}}(\theta_\star)\|_{\mathrm{op}} \leq L_{\mathfrak{a}1}^2 L_{\mathcal{R}2}$$

Now, observe that our condition on λT, namely $\lambda T \geq (80d_\theta/r_{\mathrm{quad}}(\theta_\star)^2)^{6/5}$, implies that $r = (5d_\theta/(\lambda T)^{5/6})^{1/2} \leq \frac{1}{4}r_{\mathrm{quad}}(\theta_\star)$. Hence, we can apply Lemma 4.1.1 to obtain

$$\begin{aligned}
\min_{\widehat{\mathfrak{a}}} \max_{\theta:\|\theta-\theta_\star\|_2\leq r} \mathbb{E}[\mathcal{R}(\widehat{\mathfrak{a}};\theta)] \geq \min\Bigg\{ &5\mu L_{\mathfrak{a}1}^2 r^2, \sigma_w^2 \min_{\theta\|\theta-\theta_\star\|_2\leq r} \mathrm{tr}\left(\mathcal{H}(\theta_\star)\left(\mathbb{E}_{\theta,\pi_{\mathrm{exp}}}[\mathbf{\Sigma}_T]+\lambda T I\right)^{-1}\right) \\
&- C_1 r^3 - \frac{32L_{\mathfrak{a}1}^2 L_{\mathcal{R}2}}{\lambda T}\exp\left(-d_\theta(\lambda T)^{1/6}\right)\Bigg\}
\end{aligned}$$

for C_1 as in Lemma 4.1.1. To conclude, we consolidate

$$C_1 r^3 + \frac{32L_{\mathfrak{a}1}^2 L_{\mathcal{R}2}}{\lambda T}\exp\left(-d_\theta(\lambda T)^{1/6}\right) = \frac{C_1(5d_\theta)^{3/2} + 32L_{\mathfrak{a}1}^2 L_{\mathcal{R}2}(\lambda T)^{1/4}\exp\left(-d_\theta(\lambda T)^{1/6}\right)}{(\lambda T)^{5/4}}.$$

Observing that $(\lambda T)^{1/4}\exp\left(-d_\theta(\lambda T)^{1/6}\right)$ is bounded above by a universal constant (since $d_\theta \geq 1$, and for all $d \geq 1$, $\max_{x\geq 0} xe^{-dx} \leq \max_{x\geq 0} xe^{-x}$ is bounded), the above is at most

$$\mathcal{O}\left(\frac{C_1 d_\theta^{3/2} + L_{\mathfrak{a}1}^2 L_{\mathcal{R}2}}{(\lambda T)^{5/4}}\right) = \frac{C_2}{(\lambda T)^{5/4}},$$

for C_2 as in the statement of the result. To conclude, it suffices to show that for our choice of r, we have

$$5\mu L_{\mathfrak{a}1}^2 r^2 \geq \sigma_w^2 \min_{\theta\|\theta-\theta_\star\|_2\leq r} \mathrm{tr}\left(\mathcal{H}(\theta_\star)\left(\mathbb{E}_{\theta,\pi_{\mathrm{exp}}}[\mathbf{\Sigma}_T]+\lambda T I\right)^{-1}\right).$$

Lower bounding $\mathbb{E}_{\theta,\pi_{\mathrm{exp}}}[\mathbf{\Sigma}_T] + \lambda T I \succeq \lambda T I$, upper bounding $\mathrm{tr}(\mathcal{H}(\theta_\star)) \leq d_\theta\|\mathcal{H}(\theta_\star)\|_{\mathrm{op}} \leq$

$d_\theta L_{\mathfrak{a}1}^2 L_{\mathcal{R}2}$, and substituting in the choice of r^2, it is enough that

$$\mu L_{\mathfrak{a}1}^2 \cdot (5d_\theta/(\lambda T)^{5/6}) \geq \sigma_w^2 \frac{d_\theta L_{\mathfrak{a}1}^2 L_{\mathcal{R}2}}{\lambda T} \iff (\lambda T)^{1/6} \geq \frac{\sigma_w^2 L_{\mathcal{R}2}}{5\mu},$$

which is satisfied for our choice of λ. □

Proof of Theorem 4.3. We apply Theorem 4.5 with $\lambda = \underline{\lambda}$, which is greater than 0 by Assumption 4.2. Then,

$$\operatorname{tr}\left(\mathcal{H}(\theta_\star)\left(\mathbb{E}_{\theta,\pi_{\mathrm{exp}}}[\boldsymbol{\Sigma}_T] + \underline{\lambda} T I\right)^{-1}\right) \geq \operatorname{tr}\left(\mathcal{H}(\theta_\star)\left(\mathbb{E}_{\theta,\pi_{\mathrm{exp}}}[\boldsymbol{\Sigma}_T] + \mathbb{E}_{\theta_\star,\pi_{\mathrm{exp}}}[\boldsymbol{\Sigma}_T]\right)^{-1}\right)$$

Under Assumption 4.3, for any θ satisfying $\|\theta - \theta_\star\|_2^2 \leq 5d_\theta/(\underline{\lambda}T)^{5/6}$ and as long as $5d_\theta/(\underline{\lambda}T)^{5/6} \leq r_{\mathrm{cov}}(\theta_\star)^2$, we have

$$\mathbb{E}_{\theta,\pi_{\mathrm{exp}}}[\boldsymbol{\Sigma}_T] \preceq c_{\mathrm{cov}} \mathbb{E}_{\theta_\star,\pi_{\mathrm{exp}}}[\boldsymbol{\Sigma}_T] + \left(\frac{L_{\mathrm{cov}}(\theta_\star,\gamma^2)\sqrt{5d_\theta}T^{7/12}}{\underline{\lambda}^{5/12}} + C_{\mathrm{cov}} T^{1-\alpha}\right) \cdot I$$

Therefore, since our alternate instances, $\theta \in \mathcal{B}_T$, do satisfy $\|\theta - \theta_\star\|_2^2 \leq 5d_\theta/(\underline{\lambda}T)^{5/6}$, if T is large enough that

$$\frac{L_{\mathrm{cov}}(\theta_\star,\gamma^2)\sqrt{5d_\theta}T^{7/12}}{\underline{\lambda}^{5/12}} + C_{\mathrm{cov}} T^{1-\alpha} \leq c_{\mathrm{cov}} T \underline{\lambda}$$

we will have, for all $\theta \in \mathcal{B}_T$,

$$\begin{aligned}\mathbb{E}_{\theta,\pi_{\mathrm{exp}}}[\boldsymbol{\Sigma}_T] &\preceq c_{\mathrm{cov}} \mathbb{E}_{\theta_\star,\pi_{\mathrm{exp}}}[\boldsymbol{\Sigma}_T] + \left(\frac{L_{\mathrm{cov}}(\theta_\star,\gamma^2)\sqrt{5d_\theta}T^{7/12}}{\underline{\lambda}^{5/12}} + C_{\mathrm{cov}} T^{1-\alpha}\right) \cdot I \\ &\preceq c_{\mathrm{cov}} \mathbb{E}_{\theta_\star,\pi_{\mathrm{exp}}}[\boldsymbol{\Sigma}_T] + c_{\mathrm{cov}} T \underline{\lambda} \cdot I \preceq 2 c_{\mathrm{cov}} \mathbb{E}_{\theta_\star,\pi_{\mathrm{exp}}}[\boldsymbol{\Sigma}_T]\end{aligned}$$

The result then follows from Theorem 4.5 and simple manipulations. □

4.2.2 Proof of Theorem 4.4

We define the following events.

$$\mathcal{A} = \left\{\mathcal{R}(\mathfrak{a}_{\mathrm{opt}}(\widehat{\theta}_{\mathrm{ls}}); \theta_\star) \leq 5\sigma_w^2 \operatorname{tr}(\mathcal{H}(\theta_\star)\mathbb{E}_{\theta_\star,\pi_{\mathrm{exp}}}[\boldsymbol{\Sigma}_T]^{-1}) \log\frac{6d_\theta}{\delta} + \frac{C_1}{T^{3/2}} + \frac{C_2}{T^{1+2\alpha}}\right\}$$

(Good event)

$$\mathcal{E}_1 = \{\lambda_{\min}(\Sigma_T) \geq \underline{\lambda} T, \Sigma_T \preceq T\bar{\boldsymbol{\Lambda}}_T\}$$ (Sufficient excitation)

$$\mathcal{E}_2 = \{\|\widehat{\theta}_{\mathrm{ls}} - \theta_\star\|_2 \leq r_{\mathrm{quad}}(\theta_\star)\}$$ (Quadratic approximation regime)

$$\mathcal{E}_3 = \{\|\Sigma_T - \mathbb{E}_{\theta_\star,\pi_{\mathrm{exp}}}[\Sigma_T]\|_{\mathrm{op}} \leq \tfrac{C_{\mathrm{con}}}{T^\alpha}\lambda_{\min}(\mathbb{E}_{\theta_\star,\pi_{\mathrm{exp}}}[\Sigma_T])\}$$ (Concentration of covariates)

We would like to show that $\mathcal{A}$ holds with high probability. The following is trivial:

$$\mathbb{P}[\mathcal{A}^c] \le \mathbb{P}[\mathcal{A}^c \cap \mathcal{E}_1 \cap \mathcal{E}_2 \cap \mathcal{E}_3] + \mathbb{P}[\mathcal{E}_1^c] + \mathbb{P}[\mathcal{E}_1 \cap \mathcal{E}_2^c] + \mathbb{P}[\mathcal{E}_3^c].$$

Events $\mathcal{E}_i$ hold with high probability. We now show that the events $\mathcal{E}_1, \mathcal{E}_2$, and $\mathcal{E}_3$ hold with high probability. Since π_{exp} satisfies Assumption 4.4, we will have $\mathbb{P}[\mathcal{E}_1^c] \le \delta$ and $\mathbb{P}[\mathcal{E}_3^c] \le \delta$ as long as

$$T \ge T_{\mathrm{se}}(\pi_{\mathrm{exp}}), \quad T \ge T_{\mathrm{con}}(\pi_{\mathrm{exp}}). \tag{4.2.1}$$

By Lemma B.2.1, on the event $\mathcal{E}_1$, with probability at least $1-\delta$,

$$\|\widehat{\theta}_{\mathrm{ls}} - \theta_\star\|_2 \le C\sqrt{\frac{\log(1/\delta) + d_\theta + \log\det(\bar{\mathbf{\Lambda}}_T/\underline{\lambda} + I)}{\underline{\lambda} T}},$$

so as long as

$$T \ge \frac{C(\log(1/\delta) + d_\theta + \log\det(\bar{\mathbf{\Lambda}}_T/\underline{\lambda} + I))}{\underline{\lambda} r_{\mathrm{quad}}(\theta_\star)^2}, \tag{4.2.2}$$

we will have

$$\|\widehat{\theta}_{\mathrm{ls}} - \theta_\star\|_2 \le C\sqrt{\frac{\log(1/\delta) + d_x + \log\det(\bar{\mathbf{\Lambda}}_T/\underline{\lambda} + I)}{\underline{\lambda} T}} \le r_{\mathrm{quad}}(\theta_\star).$$

Thus, $\mathbb{P}[\mathcal{E}_1 \cap \mathcal{E}_2^c] \le \delta$.

Events $\mathcal{E}_i$ imply good event holds. We now consider the event $\mathcal{E}_1 \cap \mathcal{E}_2 \cap \mathcal{E}_3$. By Proposition 4.2, since $\mathcal{R}$ satisfies Assumption 4.1, on this event we have

$$\mathcal{R}(\mathfrak{a}_{\mathrm{opt}}(\widehat{\theta}_{\mathrm{ls}}); \theta_\star) \le \|\widehat{\theta}_{\mathrm{ls}} - \theta_\star\|^2_{\mathcal{H}(\theta_\star)} + L_{\mathrm{quad}}\|\widehat{\theta}_{\mathrm{ls}} - \theta_\star\|_2^3 \le \|\widehat{\theta}_{\mathrm{ls}} - \theta_\star\|^2_{\mathcal{H}(\theta_\star)} + C_1/T^{3/2}$$

where the last inequality follows by the bound on $\|\widehat{\theta}_{\mathrm{ls}} - \theta_\star\|_2$ shown above for

$$C_1 := CL_{\mathrm{quad}}\frac{(\log(1/\delta) + d_\theta + \log\det(\bar{\mathbf{\Lambda}}_T/\underline{\lambda} + I))^{3/2}}{\underline{\lambda}^{3/2}}.$$

By Theorem 4.2, on the event $\mathcal{E}_3$ and if T is large enough so that

$$C_{\mathrm{con}}/T^\alpha < 1/4 \tag{4.2.3}$$

and since $\|\boldsymbol{\Sigma}_T - \mathbb{E}_{\theta_\star,\pi_{\mathrm{exp}}}[\boldsymbol{\Sigma}_T]\|_{\mathrm{op}}$, with probability at least $1-\delta$:

$$\begin{aligned}\|\widehat{\theta}_{\mathrm{ls}} - \theta_\star\|^2_{\mathcal{H}(\theta_\star)} &\le 5\sigma_w^2 \mathrm{tr}(\mathcal{H}(\theta_\star)\mathbb{E}_{\theta_\star,\pi_{\mathrm{exp}}}[\boldsymbol{\Sigma}_T]^{-1}) \log\frac{6d_\theta}{\delta} \\ &\quad + \frac{130\sigma_w^2 C_{\mathrm{con}}^2}{T^{2\alpha}} \lambda_{\min}(\mathbb{E}_{\theta_\star,\pi_{\mathrm{exp}}}[\boldsymbol{\Sigma}_T]) \mathrm{tr}(\mathbb{E}_{\theta_\star,\pi_{\mathrm{exp}}}[\boldsymbol{\Sigma}_T]^{-1}) \mathrm{tr}(\mathcal{H}(\theta_\star)\mathbb{E}_{\theta_\star,\pi_{\mathrm{exp}}}[\boldsymbol{\Sigma}_T]^{-1}) \log\frac{6d_\theta}{\delta} \\ &\le 5\sigma_w^2 \mathrm{tr}(\mathcal{H}(\theta_\star)\mathbb{E}_{\theta_\star,\pi_{\mathrm{exp}}}[\boldsymbol{\Sigma}_T]^{-1}) \log\frac{6d_\theta}{\delta} + \frac{260\sigma_w^2 C_{\mathrm{con}}^2 d_\theta \mathrm{tr}(\mathcal{H}(\theta_\star))}{\underline{\lambda} T^{1+2\alpha}} \log\frac{6d_\theta}{\delta}\end{aligned}$$

where the final inequality follows since

$$\mathbb{E}_{\theta_\star,\pi_{\mathrm{exp}}}[\boldsymbol{\Sigma}_T] \succeq \mathbb{E}_{\theta_\star,\pi_{\mathrm{exp}}}[\mathbb{I}\{\mathcal{E}_1\}\boldsymbol{\Sigma}_T] \ge \mathbb{P}[\mathcal{E}_1]\underline{\lambda} T \cdot I \ge \frac{1}{2}\underline{\lambda} T \cdot I$$

Thus, $\mathbb{P}[\mathcal{E}^c \cap \mathcal{E}_1 \cap \mathcal{E}_2 \cap \mathcal{E}_3] \le \delta$ with

$$C_2 := \frac{260\sigma_w^2 C_{\mathrm{con}}^2 d_\theta \mathrm{tr}(\mathcal{H}(\theta_\star))}{\underline{\lambda}} \log\frac{6d_\theta}{\delta}$$

so it follows that $\mathbb{P}[\mathcal{A}^c] \le 4\delta$. The final result then follows by rescaling δ and so long as T is large enough that (4.2.1), (4.2.2), and (4.2.3) hold, which will be the case if (4.1.4) holds.

4.3 Proof of Theorem 4.1 and Theorem 4.2

To conclude our discussion of the MDM setting, we provide proofs for our upper and lower bounds on regression in general norms.

4.3.1 Proof of M-norm Regression Lower Bound (Theorem 4.1)

We now prove Theorem 4.1. Without loss of generality, set $\sigma_w^2 = 1$. The proof of the lower bound is a Gaussian-specialization of the Van Trees inequality (see, e.g. Gill et al. [1995]), a Bayes-risk lower bound which considers the risk of estimating a quantity under a certain prior. For our prior, we use a normal distribution, which we truncate to a radius r. In what follows, we set

$$\Gamma := \boldsymbol{\Lambda}^{-1} \in \mathbb{S}^{d_\theta}_{++}.$$

We let $\mathcal{N}_{\mathrm{tr}}$ denote the following *truncated normal* distribution: the distribution of $Z \sim \mathcal{N}(\theta_0, \Gamma)$, conditioned on the event $\|Z - \theta_0\|_2 \le r$. We further define the full data $\mathfrak{D}_T := (y_{1:T}, z_{1:T})$, and let:

- $\mathcal{D}_{\mathrm{full}}(\mathfrak{D}_T)$ denote the posterior of θ given $\mathfrak{D}_T$ and when the prior is $\mathcal{N}(\theta_0, \Gamma)$;
- $\mathcal{D}_{\mathrm{trunc}}(\mathfrak{D}_T)$ denote the distribution of $\theta \mid \mathfrak{D}_T$.

Throughout, we assume that our posited estimator $\widehat{\theta}$ is a deterministic function of $\mathfrak{D}_T$; this is without loss of generality for a Bayes-risk lower bound. Then, since the distribution $\mathcal{N}_{\text{tr}}$ is supported on the ball $\mathcal{B} := \{\theta : \|\theta - \theta_0\|_2 \le r\}$, for any such estimator $\widehat{\theta}$:

$$\max_{\theta \in \mathcal{B}} \mathbb{E}_\theta[\mathcal{R}_{\text{est}}(\widehat{\theta}; \theta)] \ge \mathbb{E}_{\theta \sim \mathcal{N}_{\text{tr}}} \mathbb{E}_\theta[\mathcal{R}_{\text{est}}(\widehat{\theta}; \theta)] = \mathbb{E}_{\theta \sim \mathcal{N}_{\text{tr}}} \mathbb{E}_{\mathfrak{D}_T \sim \theta}[\|\widehat{\theta}(\mathfrak{D}_T) - \theta\|_2^2].$$

We can rewrite the above via the following lemma:

Lemma 4.3.1. *Let X, Y be abstract random variables, with $X \sim \mathbb{P}_X$, $Y \sim \mathbb{P}_{Y|X}(\cdot \mid X)$, and let $f(x, y)$ be an integrable function. Moreover, suppose that $X \mid Y$ has density $\mathbb{P}_{X|Y}(\cdot \mid Y)$. Then*

$$\mathbb{E}_{X \sim \mathbb{P}_X} \mathbb{E}_{Y \sim \mathbb{P}_{Y|X}(\cdot|X)}[f(X, Y)] = \mathbb{E}_{X \sim \mathbb{P}_X} \mathbb{E}_{Y \sim \mathbb{P}_{Y|X}(\cdot|X)} \mathbb{E}_{X' \sim \mathbb{P}_{X|Y}(\cdot|Y)}[f(X', Y)].$$

By Lemma 4.3.1, the above is equal to

$$\mathbb{E}_{\theta \sim \mathcal{N}_{\text{tr}}} \mathbb{E}_\theta[\mathcal{R}_{\text{est}}(\widehat{\theta}; \theta)] = \mathbb{E}_{\theta \sim \mathcal{N}_{\text{tr}}} \mathbb{E}_{\mathfrak{D}_T \sim \theta} \mathbb{E}_{\theta' \sim \mathcal{D}_{\text{trunc}}(\mathfrak{D}_T)}[\|\widehat{\theta}(\mathfrak{D}_T) - \theta'\|_M^2]. \tag{4.3.1}$$

Note that for a random vector Z and any fixed a, $\mathbb{E}[\|Z - a\|_M^2] \ge \mathbb{E}[\|Z - \mathbb{E}[Z]\|_M^2]$. Denoting the event $\mathcal{E} := \{\|\theta' - \theta_0\|_2 \le r\}$, regardless of the choice of $\widehat{\theta}$ we can lower bound (4.3.1) by

$$\begin{aligned}
(4.3.1) &\ge \mathbb{E}_{\theta \sim \mathcal{N}_{\text{tr}}} \mathbb{E}_{\mathfrak{D}_T \sim \theta} \mathbb{E}_{\theta' \sim \mathcal{D}_{\text{trunc}}(\mathfrak{D}_T)}[\|\theta' - \mathbb{E}_{\theta' \sim \mathcal{D}_{\text{trunc}}(\mathfrak{D}_T)}[\theta']\|_M^2] \\
&= \mathbb{E}_{\theta \sim \mathcal{N}_{\text{tr}}} \mathbb{E}_{\mathfrak{D}_T \sim \theta} \mathbb{E}_{\theta' \sim \mathcal{D}_{\text{full}}(\mathfrak{D}_T)} \left[\left\| \theta' - \mathbb{E}_{\theta' \sim \mathcal{D}_{\text{full}}(\mathfrak{D}_T)}[\theta' \mid \mathcal{E}] \right\|_M^2 \mid \mathcal{E} \right].
\end{aligned}$$

To handle this expression, we use the following technical lemma:

Lemma 4.3.2. *Consider a square-integrable random vector $Z \in \mathbb{R}^{d_\theta}$, fixed $\mu \in \mathbb{R}^{d_\theta}$, $r \ge 0$. Define the event $\mathcal{E} := \{\|Z - \mu\|_2 \le r\}$. Then,*

$$\mathbb{E}[\|Z - \mathbb{E}[Z \mid \mathcal{E}]\|_M^2 \mid \mathcal{E}] \ge \mathbb{E}[\|Z - \mathbb{E}[Z]\|_M^2] - 4\|M\|_{\text{op}} \mathbb{E}[\mathbb{I}\{\mathcal{E}^c\} \|Z - \mu\|_2^2].$$

Instantiating Lemma 4.3.2 gives:

$$\begin{aligned}
(4.3.1) \ge\ & \mathbb{E}_{\theta \sim \mathcal{N}_{\text{tr}}} \mathbb{E}_{\mathfrak{D}_T \sim \theta} \mathbb{E}_{\theta' \sim \mathcal{D}_{\text{full}}(\mathfrak{D}_T)} \left[\|\theta' - \mathbb{E}_{\theta'' \sim \mathcal{D}_{\text{full}}(\mathfrak{D}_T)}[\theta'']\|_M^2 \right] \\
& - 4\|M\|_{\text{op}}^2 \mathbb{E}_{\theta \sim \mathcal{N}_{\text{tr}}} \mathbb{E}_{\mathfrak{D}_T \sim \theta} \mathbb{E}_{\theta' \sim \mathcal{D}_{\text{full}}(\mathfrak{D}_T)} \mathbb{E}_{\theta' \sim \mathcal{D}_{\text{full}}(\mathfrak{D}_T)} \left[\|\theta' - \theta_0\|_2^2 \cdot \mathbb{I}\{\|\theta' - \theta_0\|_2^2 > r^2\} \right].
\end{aligned}$$

Hence, retracing our steps thus far, we have shown that for any $\widehat{\theta}$,

$$\begin{aligned}
\max_{\theta \in \mathcal{B}} \mathbb{E}_\theta \mathcal{R}_{\text{est}}(\widehat{\theta}; \theta) \ge\ & \underbrace{\mathbb{E}_{\theta \sim \mathcal{N}_{\text{tr}}} \mathbb{E}_{\mathfrak{D}_T \sim \theta} \mathbb{E}_{\theta' \sim \mathcal{D}_{\text{full}}(\mathfrak{D}_T)} \left[\|\theta' - \mathbb{E}_{\theta' \sim \mathcal{D}_{\text{full}}(\mathfrak{D}_T)}[\theta']\|_M^2 \right]}_{\text{(a)}} \\
& - 4\|M\|_{\text{op}}^2 \cdot \underbrace{\left(\mathbb{E}_{\theta \sim \mathcal{N}_{\text{tr}}} \mathbb{E}_{\mathfrak{D}_T \sim \theta} \mathbb{E}_{\theta' \sim \mathcal{D}_{\text{full}}(\mathfrak{D}_T)} \left[\|\theta' - \theta_0\|_2^2 \cdot \mathbb{I}\{\|\theta' - \theta_0\|_2^2 > r^2\} \right] \right)}_{\text{(b)}}.
\end{aligned} \tag{4.3.2}$$

Let us control the two resulting terms.

Controlling Term (a)**.** First, we bound the dominant term (a):

Lemma 4.3.3. *The following identity holds:*

$$\mathbb{E}_{\theta' \sim \mathcal{D}_{\text{full}}(\mathfrak{D}_T)} \left[\|\theta' - \mathbb{E}_{\theta'' \sim \mathcal{D}_{\text{full}}(\mathfrak{D}_T)}[\theta'']\|_M^2 \right] = \sigma_w^2 \text{tr}(M^{1/2}(\mathbf{\Sigma}_T + \Gamma^{-1})^{-1} M^{1/2}).$$

As a direct consequence of the above lemma, we find that

$$\begin{aligned} \text{(a)} &= \mathbb{E}_{\theta \sim \mathcal{N}_{\text{tr}}} \mathbb{E}_{\mathfrak{D}_T \sim \theta} \mathbb{E}_{\theta' \sim \mathcal{D}_{\text{full}}(\mathfrak{D}_T)} \mathbb{E}_{\theta'} \left[\sigma_w^2 \text{tr}(M^{1/2}(\mathbf{\Sigma}_T + \Gamma^{-1})^{-1} M^{1/2}) \right] \\ &= \mathbb{E}_{\theta \sim \mathcal{N}_{\text{tr}}} \mathbb{E}_{\theta} \left[\sigma_w^2 \text{tr}(M^{1/2}(\mathbf{\Sigma}_T + \Gamma^{-1})^{-1} M^{1/2}) \right], \end{aligned} \tag{4.3.3}$$

where in the last line, we have invoked Lemma 4.3.1.

Controlling Term (b)**.** Let $p_r := \mathbb{P}_{\theta \sim \mathcal{N}(\theta_0, \Gamma)}[\|\theta - \theta_0\| > r]$ denote the probability that θ lies within the truncation region. Then, for any nonnegative function $f(\theta) \geq 0$,

$$\mathbb{E}_{\theta \sim \mathcal{N}_{\text{tr}}}[f(\theta)] = \frac{1}{1 - p_r} \mathbb{E}_{\theta \sim \mathcal{N}(\theta_0, \Gamma)}[f(\theta) \cdot \mathbb{I}\{\|\theta - \theta_0\| \leq r\}] \leq \frac{1}{1 - p_r} \mathbb{E}_{\theta \sim \mathcal{N}(\theta_0, \Gamma)}[f(\theta)].$$

Thus,

$$\begin{aligned} \text{(b)} &= \mathbb{E}_{\theta \sim \mathcal{N}_{\text{tr}}} \mathbb{E}_{\mathfrak{D}_T \sim \theta} \mathbb{E}_{\theta' \sim \mathcal{D}_{\text{full}}(\mathfrak{D}_T)} \left[\|\theta' - \theta_0\|_2^2 \cdot \mathbb{I}\{\|\theta' - \theta_0\|_2^2 > r^2\} \right] \\ &\leq \frac{1}{1 - p_r} \mathbb{E}_{\theta \sim \mathcal{N}(\theta_0, \Gamma)} \mathbb{E}_{\mathfrak{D}_T \sim \theta} \mathbb{E}_{\theta' \sim \mathcal{D}_{\text{full}}(\mathfrak{D}_T)} \left[\|\theta' - \theta_0\|_2^2 \cdot \mathbb{I}\{\|\theta' - \theta_0\|_2^2 > r^2\} \right]. \end{aligned} \tag{4.3.4}$$

By Lemma 4.3.1, we have

$$(4.3.4) = \frac{1}{1 - p_r} \mathbb{E}_{\theta \sim \mathcal{N}(\theta_0, \Gamma)} \left[\|\theta - \theta_0\|_2^2 \cdot \mathbb{I}\{\|\theta - \theta_0\|_2^2 > r\} \right].$$

We now bound the above. Note that $\theta - \theta_0$ has the same distribution as $\Gamma^{1/2}\mathbf{g}$, where $\mathbf{g} \sim \mathcal{N}(0, I_d)$. Hence, $p_r = \mathbb{P}_{\mathbf{g} \sim \mathcal{N}(0, I_d)}[\mathbf{g}^\top \Gamma \mathbf{g} > r^2]$. For $r^2 \geq 2\text{tr}(\Gamma)$, Markov's inequality therefore implies $p_r \leq 1/2$, so that $\frac{1}{1-p_r} \leq 2$. Moreover, by the same change of variables,

$$\mathbb{E}_{\theta \sim \mathcal{N}(\theta_0, \Gamma)} \left[\|\theta - \theta_0\|_2^2 \cdot \mathbb{I}\{\|\theta - \theta_0\|_2^2 > r^2\} \right] = \mathbb{E}_{\mathbf{g} \sim \mathcal{N}(0, I_d)} \left[\mathbf{g}^\top \Gamma \mathbf{g} \cdot \mathbb{I}\{\mathbf{g}^\top \Gamma \mathbf{g} > r^2\} \right],$$

Thus, for $r^2 \geq 2\text{tr}(\Gamma) = 2\text{tr}(\mathbf{\Lambda}^{-1})$ (which will be true since by assumption $r \geq \sqrt{5\text{tr}(\mathbf{\Lambda}^{-1})}$), it holds that

$$\text{(b)} \leq 2\mathbb{E}_{\mathbf{g} \sim \mathcal{N}(0, I_d)} \left[\mathbf{g}^\top \Gamma \mathbf{g} \cdot \mathbb{I}\{\mathbf{g}^\top \Gamma \mathbf{g} > r^2\} \right] = 2 \int_{r^2}^{\infty} \mathbb{P}[\mathbf{g}^\top \Gamma \mathbf{g} > \tau] d\tau.$$

We now invoke a coarse consequence of the Hanson-Wright inequality:

Lemma 4.3.4. *For any $u \geq 4\mathrm{tr}(\Gamma)$, we have $\mathbb{P}[\mathbf{g}^\top \Gamma \mathbf{g} > \mathrm{tr}(\Gamma) + u] \leq e^{-\frac{u}{4\|\Gamma\|_{\mathrm{op}}}}$.*

Hence, under the assumption of the theorem, $r^2 \geq 5\mathrm{tr}(\mathbf{\Lambda}^{-1}) = 5\mathrm{tr}(\Gamma)$, we may bound

$$\begin{aligned}
\text{(b)} &\leq 2\int_{r^2}^{\infty} \mathbb{P}[\mathbf{g}^\top \Gamma \mathbf{g} > \tau] d\tau \\
&= 2\int_{r^2 - \mathrm{tr}(\Gamma)}^{\infty} \mathbb{P}[\mathbf{g}^\top \Gamma \mathbf{g} > \mathrm{tr}(\Gamma) + \tau] d\tau \\
&\leq 2\int_{r^2 - \mathrm{tr}(\Gamma)}^{\infty} e^{-\frac{u}{4\|\Gamma\|_{\mathrm{op}}}} = 8\|\Gamma\|_{\mathrm{op}} e^{-\frac{r^2 - \mathrm{tr}(\Gamma)}{4\|\Gamma\|_{\mathrm{op}}}} \\
&\leq 8\|\Gamma\|_{\mathrm{op}} e^{-\frac{r^2}{5\|\Gamma\|_{\mathrm{op}}}}. \qquad (4.3.5)
\end{aligned}$$

Concluding the Proof. Combining (4.3.2), (4.3.3), and (4.3.5), we have

$$\max_{\theta \in \mathcal{B}} \mathbb{E}_\theta[\mathcal{R}_{\mathrm{est}}(\widehat{\theta}; \theta)] \geq \mathbb{E}_{\theta \sim \mathcal{N}_{\mathrm{tr}}} \mathbb{E}_\theta \left[\sigma_w^2 \mathrm{tr}(M^{1/2}(\mathbf{\Sigma}_T + \Gamma^{-1})^{-1} M^{1/2})\right] - 32\|\Gamma\|_{\mathrm{op}}\|M\|_{\mathrm{op}} e^{-\frac{r^2}{5\|\Gamma\|_{\mathrm{op}}}}.$$

Since $\Gamma = \mathbf{\Lambda}^{-1}$, we have $32\|\Gamma\|_{\mathrm{op}}\|M\|_{\mathrm{op}} e^{-\frac{r^2}{5\|\Gamma\|_{\mathrm{op}}}} = \frac{32\|M\|_{\mathrm{op}}}{\lambda_{\min}(\mathbf{\Lambda})} \exp(-\frac{r^2}{5}\lambda_{\min}(\mathbf{\Lambda})) =: \Psi(r; \mathbf{\Lambda}, M)$. Noting that $X \mapsto \mathrm{tr}(X^{-1})$ is a convex function (on the domain of positive-definite matrices), and since convexity is preserved under affine transformation, Jensen's inequality gives that

$$\mathbb{E}_{\theta \sim \mathcal{N}_{\mathrm{tr}}} \mathbb{E}_\theta \left[\sigma_w^2 \mathrm{tr}(M^{1/2}(\mathbf{\Sigma}_T + \Gamma^{-1})^{-1} M^{1/2})\right] \geq \mathbb{E}_{\theta \sim \mathcal{N}_{\mathrm{tr}}} \left[\sigma_w^2 \mathrm{tr}(M^{1/2}(\mathbb{E}_\theta[\mathbf{\Sigma}_T] + \Gamma^{-1})^{-1} M^{1/2})\right].$$

Substituting in $\mathbf{\Lambda} = \Gamma^{-1}$, noting that the distribution $\mathcal{N}_{\mathrm{tr}}$ is supported on the ball $\mathcal{B}$, and that the above holds for any deterministic estimator $\widehat{\theta}$ concludes the bound.

4.3.2 Proof of M-norm Regression Upper Bound (Theorem 4.2)

Let β be a parameter to be tuned, and set. Since M may not be full rank, we consider a perturbation

$$N := M + \zeta I, \quad \zeta = \mathrm{tr}(M\mathbf{\Lambda}^{-1})/\mathrm{tr}(\mathbf{\Lambda}^{-1}),$$

where $\zeta > 0$ is to be chosen. We further define:

- $\mathbf{Z} \in \mathbb{R}^{T \times d_\theta}$ denote the matrix whose rows are $z_t^\top$.
- $\mathbf{w} \in \mathbb{R}^T$ as the vector whose entries are w_t.
- $\widehat{\theta}_{\mathrm{ls}}$ the least squares estimate of $\theta_\star$ defined in (4.1.2) .
- We let $v_1, \ldots, v_d$ be the eigenvectors of $N^{1/2}\mathbf{\Lambda}^{-1}N^{1/2}$ and $\lambda_j := \lambda_j(N^{1/2}\mathbf{\Lambda}^{-1}N^{1/2})$, which we note are deterministic.

The error of the least-squares estimate is then,

$$\|\widehat{\theta}_{\mathrm{ls}} - \theta_\star\|_M^2 = \|(\mathbf{Z}^\top \mathbf{Z})^{-1}\mathbf{Z}^\top \mathbf{w}\|_M^2.$$

Since $\epsilon < \lambda_{\min}(\mathbf{\Lambda})/4$, we can apply Lemma C.3.1 to get

$$\|(\mathbf{Z}^\top \mathbf{Z})^{-1} - \mathbf{\Lambda}^{-1}\|_{\mathrm{op}} \leq \frac{\epsilon}{\lambda_{\min}(\mathbf{\Lambda})(\lambda_{\min}(\mathbf{\Lambda}) - \epsilon)} < \frac{2\epsilon}{\lambda_{\min}(\mathbf{\Lambda})^2}.$$

We now invoke the following lemma, controlling the relation of (weighted) squares of matrices in the PSD order:

Lemma 4.3.5. *Let $A, B, M \succeq 0$ and $C = A - B$. Then, $AMA + 7CMC \succeq BMB/2$.*

Instantiating Lemma 4.3.5 with $A = \mathbf{\Lambda}^{-1}$, $B = (\mathbf{Z}^\top \mathbf{Z})^{-1}$, and $M = N$, we have

$$\mathbf{\Lambda}^{-1} N \mathbf{\Lambda}^{-1} + \frac{13\|N\|_{\mathrm{op}}\epsilon^2}{\lambda_{\min}(\mathbf{\Lambda})^4} I \succeq (\mathbf{Z}^\top \mathbf{Z})^{-1} N (\mathbf{Z}^\top \mathbf{Z})^{-1}.$$

Suppose that ϵ is chosen sufficiently small that, for a constant α to be specified:

$$\alpha \mathbf{\Lambda}^{-1} N \mathbf{\Lambda}^{-1} \succeq \frac{13\|N\|_{\mathrm{op}}\epsilon^2}{\lambda_{\min}(\mathbf{\Lambda})^4}; \tag{4.3.6}$$

we shall revisit this point at the end of the proof. Then,

$$\begin{aligned}
\|(\mathbf{Z}^\top \mathbf{Z})^{-1}\mathbf{Z}^\top \mathbf{w}\|_M^2 &\leq \|(\mathbf{Z}^\top \mathbf{Z})^{-1}\mathbf{Z}^\top \mathbf{w}\|_N^2 \\
&= \mathbf{w}^\top \mathbf{Z}(\mathbf{Z}^\top \mathbf{Z})^{-1} N (\mathbf{Z}^\top \mathbf{Z})^{-1}\mathbf{Z}^\top \mathbf{w} \\
&\leq (1+\alpha)\mathbf{w}^\top \mathbf{Z}\mathbf{\Lambda}^{-1} N \mathbf{\Lambda}^{-1}\mathbf{Z}^\top \mathbf{w} \\
&= (1+\alpha)\|N^{1/2}\mathbf{\Lambda}^{-1}\mathbf{Z}^\top \mathbf{w}\|_2^2 \\
&= (1+\alpha)\sum_{j=1}^{d_\theta} (v_j^\top N^{1/2}\mathbf{\Lambda}^{-1} N^{1/2} N^{-1/2}\mathbf{Z}^\top \mathbf{w})^2 \\
&= (1+\alpha)\sum_{j=1}^{d_\theta} \lambda_j^2 (v_j^\top N^{-1/2}\mathbf{Z}^\top \mathbf{w})^2 \\
&= (1+\alpha)\sum_{j=1}^{d_\theta} \lambda_j^2 (\mathbf{Z}_j^\top \mathbf{w})^2,
\end{aligned} \tag{4.3.7}$$

where we define $\mathbf{Z}_j := \mathbf{Z}N^{-1/2}v_j \in \mathbb{R}^T$. For any $\tau > 0$, we have

$$(\mathbf{Z}_j^\top \mathbf{w})^2 = \left(\sum_{t=1}^T \mathbf{Z}_{jt}\mathbf{w}_t\right)^2 = \left(\sum_{t=1}^T \mathbf{Z}_{jt}^2 + \tau\lambda_j^{-1}\right)\left(\left(\sum_{t=1}^T \mathbf{Z}_{jt}^2 + \tau\lambda_j^{-1}\right)^{-1/2} \cdot \sum_{t=1}^T \mathbf{Z}_{jt}\mathbf{w}_t\right)^2$$
$$= (\|\mathbf{Z}_j\|_2^2 + \tau\lambda_j^{-1}) \cdot \left\|\sum_{t=1}^T \mathbf{Z}_{jt}\mathbf{w}_t\right\|^2_{(\sum_{t=1}^T \mathbf{Z}_{jt}^2 + \tau\lambda_j^{-1})^{-1}}.$$

Applying Proposition 3.1 with a union bound over indices $j \in [d_\theta]$, it holds with probability $1-\delta$ for all $j \in [d_\theta]$ simultaneously and for any fixed $\tau > 0$:

$$(\mathbf{Z}_j^\top \mathbf{w})^2 \le 2\sigma_w^2(\|\mathbf{Z}_j\|_2^2 + \tau\lambda_j^{-1}) \log \frac{d_\theta(\|\mathbf{Z}_j\|_2^2 + \tau\lambda_j^{-1})}{\tau\lambda_j^{-1}\delta}.$$

In addition, note that for any $\tau \ge \beta$,

$$\|\mathbf{Z}_j\|_2^2 = v_j^\top N^{-1/2}\mathbf{Z}^\top\mathbf{Z}N^{-1/2}v_j \le v_j^\top N^{-1/2}(\mathbf{\Lambda} + \epsilon I)N^{-1/2}v_j$$
$$\le (1+\beta)v_j^\top N^{-1/2}\mathbf{\Lambda}N^{-1/2}v_j = (1+\beta)\lambda_j^{-1} \le (1+\tau)\lambda_j^{-1}.$$

Hence, with probability $1-\delta$ the following holds for all $j \in [d_\theta]$ simultaneously:

$$(\mathbf{Z}_j^\top \mathbf{w})^2 \le \frac{2(1+2\tau)\sigma_w^2}{\lambda_j} \log \frac{d_\theta(\tau^{-1}+2)}{\delta}.$$

Then, combining with (4.3.7), we have that with probability $1-\delta$:

$$\|(\mathbf{Z}^\top\mathbf{Z})^{-1}\mathbf{Z}^\top\mathbf{w}\|_M^2 \le 2(1+\alpha)(1+\tau)\sigma_w^2 \sum_{j=1}^{d_\theta} \lambda_j \log \frac{d_\theta(\tau^{-1}+2)}{\delta}$$
$$\le 2(1+\alpha)(1+\tau)\sigma_w^2 \log \frac{d_\theta(\tau^{-1}+2)}{\delta} \cdot \mathrm{tr}(N^{1/2}\mathbf{\Lambda}^{-1}N^{1/2}).$$

Finally, we can simplify $\mathrm{tr}(N^{1/2}\mathbf{\Lambda}^{-1}N^{1/2}) = \mathrm{tr}(N\mathbf{\Lambda}^{-1}) = \mathrm{tr}((M+\zeta I)\mathbf{\Lambda}^{-1}) \le 2\mathrm{tr}(M\mathbf{\Lambda}^{-1})$ for our choice of $\zeta = \mathrm{tr}(M\mathbf{\Lambda}^{-1})/\mathrm{tr}(\mathbf{\Lambda}^{-1})$; thus, choosing $\tau \ge 1/4 \ge \beta$ (recall the assumption, $\beta \le 1/4$), we have with probability at least $1-\delta$:

$$\|(\mathbf{Z}^\top\mathbf{Z})^{-1}\mathbf{Z}^\top\mathbf{w}\|_M^2 \le 5\sigma_w^2(1+\alpha)\log\frac{6d_\theta}{\delta} \cdot \mathrm{tr}(M\mathbf{\Lambda}^{-1}). \tag{4.3.8}$$

To conclude, let us find a suitable constant α satisfying (4.3.6). Recall that we require

$\alpha \mathbf{\Lambda}^{-1} N \mathbf{\Lambda}^{-1} \succeq \frac{13\|N\|_{\text{op}}\epsilon^2}{\lambda_{\min}(\mathbf{\Lambda})^4}$. Since $N \succeq \zeta \cdot I$, and $\epsilon \leq \beta\lambda_{\min}(\mathbf{\Lambda})$, we want

$$\alpha\zeta \geq 13\beta^2\|N\|_{\text{op}} = 13\beta^2(\|M\|_{\text{op}} + \zeta)$$

Recalling $\zeta = \text{tr}(M\mathbf{\Lambda}^{-1})/\text{tr}(\mathbf{\Lambda}^{-1}) \leq \|M\|_{\text{op}}$, we can chose $\alpha \geq \frac{26\beta^2\|M\|_{\text{op}}\text{tr}(\mathbf{\Lambda}^{-1})}{\text{tr}(M\mathbf{\Lambda}^{-1})}$. In particular, since $\text{tr}(M\mathbf{\Lambda}^{-1}) \geq \|M\|_{\text{op}}\lambda_{\min}(\mathbf{\Lambda}^{-1})$, we can take

$$\alpha = \frac{26\beta^2\|M\|_{\text{op}}\text{tr}(\mathbf{\Lambda}^{-1})}{\|M\|_{\text{op}}\lambda_{\min}(\mathbf{\Lambda})^{-1}} = 26\beta^2\lambda_{\max}(\mathbf{\Lambda})\text{tr}(\mathbf{\Lambda}^{-1}).$$

Chapter 5

Task-Dependent Exploration in Linear Dynamical Systems

We return now to the linear dynamical system setting considered in Chapter 3, but, rather than identifying $A_\star$, we consider the decision making objective of Chapter 4. In particular, we study dynamics of the form

$$x_{t+1} = A_\star x_t + B_\star u_t + w_t, \quad x_0 \equiv 0, \tag{5.0.1}$$

and aim to find a decision $\mathfrak{a} \in \mathbb{R}^{d_\mathfrak{a}}$ that minimizes $\mathcal{J}_{\theta_\star}(\mathfrak{a})$, for $\theta_\star = (A_\star, B_\star)$. We assume that $\theta_\star$ is unknown, and that $\mathcal{J}_{\theta_\star}(\mathfrak{a})$ satisfies the smoothness assumptions given in Chapter 4. As we will see, (5.0.1) is an instance of the general linear observation model of Chapter 4, (4.0.1), allowing the results from Chapter 4 to be applied to linear dynamical systems. Formally, we refer to this as the *linear dynamical decision making* (LDDM) setting:

Definition 5.0.1 (Linear Dynamical Decision Making (LDDM))**.** Assume our excess risk $\mathcal{R}(\widehat{\mathfrak{a}}; \theta_\star)$ satisfies Assumption 4.1 and that our observations are generated by a linear dynamical system, (5.0.1). Then we call the problem of choosing a decision $\widehat{\mathfrak{a}}$ to minimize $\mathcal{R}(\widehat{\mathfrak{a}}; \theta_\star)$ *linear dynamical decision making* (LDDM).

The results of Chapter 4 show that the certainty equivalence decision rule is optimal in a strong, instance-dependent sense when any fixed exploration policy π_{exp} is used to collect data. While characterizing the optimal decision rule under a *fixed* exploration policy, these results fail to characterize the *optimal* exploration policy. In this chapter, we seek to fill this gap, and provide a tight characterization of the optimal achievable rate for the best decision rule and exploration policy. That is, we seek to understand how we can best excite an unknown linear dynamical system if our end goal is not to identify the system, but to learn a decision minimizing the given loss on the system.

While our setting is quite broad and encompasses a variety of settings of interest, a key motivation is the problem of learning an effective controller to, for example, minimize some cost on the system (5.0.1). In such settings, we let $\mathfrak{a}$ denote the parameters of our controller, and $\mathcal{J}_{\theta_\star}(\mathfrak{a})$ denotes the cost of incurred by playing the controller with parameter $\mathfrak{a}$ on the system $\theta_\star$. With this $\mathfrak{a}$ and $\mathcal{J}_{\theta_\star}(\mathfrak{a})$, our goal is to estimate $\theta_\star$ well enough so that we can

learn a controller able to effectively control it. A key instance of this control problem which our model captures is the LQR problem, as stated in Section 1.1. We recall the definition of the LQR problem below.

Example 5.0.1 (Linear Quadratic Regulator (LQR))**.** In the LQR problem, the agent's objective is to design a policy that minimizes the infinite-horizon cumulative cost, with losses $\ell(x,u) := x^\top R_{\mathbf{x}} x + u^\top R_{\mathbf{u}} u$. The resultant cost function is:

$$\mathcal{J}_{\mathrm{LQR},\theta_\star}[\pi] := \lim_{T\to\infty} \mathbb{E}_{\theta_\star,\pi}\left[\frac{1}{T}\sum_{t=1}^{T} x_t^\top R_{\mathbf{x}} x_t + u_t^\top R_{\mathbf{u}} u_t\right].$$

It is well known that the optimal policies are of the form $u_t = Kx_t$ where $K \in \mathbb{R}^{d_u \times d_x}$; we denote these policies π^K, and let $\mathcal{J}_{\mathrm{LQR},\theta}(K) = \mathcal{J}_{\mathrm{LQR},\theta}[\pi^K]$. Here our decision $\mathfrak{a}$ is the controller K and our loss $\mathcal{J}_\theta$ is $\mathcal{J}_{\mathrm{LQR},\theta}$.

In the following, we consider general smooth losses, but return the the LQR setting in Section 5.2.

5.1 Optimal Decision Making in Linear Dynamical Systems

Unless otherwise noted, throughout the remainder of this chapter we adopt the LDS notation from Chapter 3, and the decision making notation given in Chapter 4. We specialize the covariance definitions given in Chapter 4 to linear dynamical systems as follows:

$$\mathbf{\Lambda}_T(\pi;\theta) := \frac{1}{T}\mathbb{E}_{\theta,\pi}\left[\sum_{t=1}^{T}\begin{bmatrix}x_t\\u_t\end{bmatrix}\begin{bmatrix}x_t\\u_t\end{bmatrix}^\top\right], \qquad \check{\mathbf{\Lambda}}_T(\pi;\theta) := I_{d_x} \otimes \mathbf{\Lambda}_T(\pi;\theta), \tag{5.1.1}$$

where $\otimes$ denotes the Kronecker product. In the following, for any covariance matrix $\mathbf{\Lambda}$ defined in Chapter 3 (e.g. $\mathbf{\Lambda}^{\mathrm{ss}}$), we let $\check{\mathbf{\Lambda}} := I_{d_x} \otimes \mathbf{\Lambda}$ denote the Kronecker product of $\mathbf{\Lambda}$ with I_{d_x} (e.g. $\check{\mathbf{\Lambda}}^{\mathrm{ss}} := I_{d_x} \otimes \mathbf{\Lambda}^{\mathrm{ss}}$). As in Chapter 3, during exploration we require that our exploration policy is in Π_{γ^2}, that is, $\mathbb{E}_{\theta_\star,\pi}[\sum_{t=1}^{T} u_t^\top u_t] \le T\gamma^2$, and also assume that our system's dynamics are stable.

Assumption 5.1. *Let Θ denote the set of all stable θ: $\Theta := \{\theta = (A,B) : \rho(A) < 1\}$. We assume that $\theta_\star \in \Theta$.*

Connection Between (5.0.1) and MDM Setting. In order to apply the results of Chapter 4 to the LDDM setting, we first argue that linear dynamical systems are a special case of the observation model given in Chapter 4, (4.0.1). Recall that this observation model takes the form (modifying several pieces of notation in (4.0.1) to avoid confusion with the LDDM setting):

$$y_{\mathfrak{t}} = \langle \mu^\star, z_{\mathfrak{t}} \rangle + \eta_{\mathfrak{t}}, \quad \eta_{\mathfrak{t}} \mid \mathcal{F}_{\mathfrak{t}-1} \sim \mathcal{N}(0,\sigma^2), \quad z_{\mathfrak{t}} \text{ is } \mathcal{F}_{\mathfrak{t}-1}\text{-adapted.} \tag{5.1.2}$$

Now let $\mu^\star = \mathrm{vec}(\begin{bmatrix}A_\star, B_\star\end{bmatrix}^\top)$, and, for any t and $i \in [d_x]$, let $\mathtt{t} = (t, i)$ and $z_\mathtt{t} = [\mathbf{0}_{(d_x+d_u)(i-1)}, x_t, u_t, \mathbf{0}_{(d_x+d_u)(d_x-i)}] \in \mathbb{R}^{d_x(d_x+d_u)}$ where $\mathbf{0}_d$ denotes the zero vector of length d. Then we see that

$$[x_{t+1}]_i = \langle \mu^\star, z_\mathtt{t} \rangle + [w_t]_i.$$

Setting $y_\mathtt{t} = [x_{t+1}]_i$ and $\eta_\mathtt{t} = [w_t]_i$, it is clear that this follows the observation model of the MDM setting with $d_\mu = d_x(d_x + d_u)$ and $\mathtt{T} = d_x T$, for $\mathtt{T}$ the time horizon in MDM. It is also straightforward to see that the measurability assumptions of the MDM setting are satisfied by this and, furthermore, we see that the definition of $\check{\mathbf{\Lambda}}_T$ given in (5.1.1) coincides with the definition of $\mathbf{\Lambda}_T$ given in Chapter 5, with the above choice of $z_\mathtt{t}$. Given this, in the following, we are able to apply results from Chapter 4 in the LDDM setting.

Optimal Complexity Measures in LDDM. From Chapter 4, we know that the decision making loss when exploring with policy π_{exp} is governed by

$$\Phi_T(\pi_{\mathrm{exp}}; \theta_\star) = \mathrm{tr}(\mathcal{H}(\theta_\star) \cdot \check{\mathbf{\Lambda}}_T(\pi_{\mathrm{exp}}; \theta_\star)^{-1}).$$

It stands to reason that the optimal choice of exploration policy procedure seeks to minimize this quantity. Towards making this formal, we introduce several quantities describing the optimal complexity.

Definition 5.1.1 (Optimal Risk)**.** We define:

$$\Phi_{\mathrm{opt}}(\gamma^2; \theta_\star) := \liminf_{T\to\infty} \inf_{\pi_{\mathrm{exp}} \in \Pi_{\gamma^2}} \Phi_T(\pi_{\mathrm{exp}}; \theta_\star),$$

the risk obtained by the policy minimizing the complexity $\Phi_T(\pi_{\mathrm{exp}}; \theta_\star)$.

Definition 5.1.2 (Exploration Local Minimax Risk)**.** Let $\mathcal{B} \subset \Theta$ denote a subset of instances. The T-sample local minimax exploration risk on $\mathcal{B}$ with budget γ^2 is

$$\mathfrak{M}_{\gamma^2}(\mathcal{R}; \mathcal{B}) := \min_{\pi_{\mathrm{exp}} \in \Pi_{\gamma^2}} \min_{\mathsf{dec}} \max_{\theta \in \mathcal{B}} \mathbb{E}_{\boldsymbol{\tau} \sim \theta, \pi_{\mathrm{exp}}}[\mathcal{R}_\theta(\mathsf{dec}(\boldsymbol{\tau}))].$$

Given these measures of complexity, we are now ready to state our main results in the LDDM setting.

5.1.1 Lower Bound for Optimal Decision Making in LDDM

In order to obtain a lower bound in the LDDM setting, we require that a certain level of excitation of the system is possible. Similar to Chapter 3, here we define:

$$\lambda_{\mathrm{noise}}(\sigma_u) := \min\left\{\lambda_{\min}\left(\sigma_w^2 \textstyle\sum_{s=0}^{d_x-1} A_\star^s (A_\star^s)^\top + \sigma_u^2 \sum_{s=0}^{d_x-1} A_\star^s B_\star B_\star^\top (A_\star^s)^\top\right), \sigma_u^2\right\}$$

and $\lambda_{\text{noise}} := \lambda_{\text{noise}}(\gamma/\sqrt{2d_u})$. We make the following assumption.

Assumption 5.2. *$\theta_\star$ and σ_w are such that $\lambda_{\text{noise}} > 0$. In particular, it suffices that $\sigma_w > 0$, or the system is controllable.*

We then have the following.

Theorem 5.1. *Assume we are in the* LDDM *setting and consider a loss function $\mathcal{J}_\theta(\mathfrak{a}) : \mathbb{R}^{d_{\mathfrak{a}}} \to \mathbb{R}$ with induced excess risk $\mathcal{R}(\mathfrak{a}; \theta_\star) := \mathcal{J}_{\theta_\star}(\mathfrak{a}) - \inf_{\mathfrak{a}'} \mathcal{J}_{\theta_\star}(\mathfrak{a}')$. Fix a model $\theta_\star$ and time horizon T. Suppose that*

- *$\mathcal{R}$ satisfies the smoothness condition, Assumption 4.1.*
- *The model $\theta_\star$ satisfies the excitation assumption Assumption 5.2 with parameter $\lambda_{\text{noise}} > 0$.*
- *The time horizon satisfies $T \geq \max\left\{C_{\text{lb}}^{\text{init}}, \left(\frac{80(d_x^2+d_xd_u)}{(\lambda_{\text{noise}})^{5/6} r_{\text{quad}}(\theta_\star)^2}\right)^{6/5}, \left(\frac{\sigma_w^2 L_{\mathcal{R}2}}{5\mu}\right)^6\right\}$.*

Finally, define the localized ball of instances

$$\mathcal{B}_T := \{\|\theta - \theta_\star\|_{\text{F}}^2 \leq 5(d_x^2 + d_xd_u)/(\lambda_{\text{noise}} T^{5/6})\}$$

Then, any decision rule $\mathsf{dec}(\boldsymbol{\tau})$ suffers the following lower bound :

$$\mathfrak{M}_{\gamma^2}(\mathcal{R}; \mathcal{B}_T) = \min_{\pi_{\text{exp}} \in \Pi_{\gamma^2}} \min_{\mathsf{dec}} \max_{\theta \in \mathcal{B}_T} \mathbb{E}_{\boldsymbol{\tau} \sim \theta, \pi_{\text{exp}}}[\mathcal{R}(\mathsf{dec}(\boldsymbol{\tau}); \theta)] \geq \frac{\sigma_w^2}{64} \cdot \frac{\Phi_{\text{opt}}(\gamma^2; \theta_\star)}{T} - \frac{C_{\text{lb}}}{(\lambda_{\text{noise}} T)^{5/4}},$$

where above:

$$C_{\text{lb}}^{\text{init}} := \text{poly}\left(d_x, d_u, \|B_\star\|_{\text{op}}, \|A_\star\|_{\mathcal{H}_\infty}, \gamma^2, \sigma_w^2, \tfrac{1}{\lambda_{\text{noise}}}, \log T\right),$$
$$C_{\text{lb}} := c_1\Big((L_{\mathfrak{a}1} L_{\mathfrak{a}2} L_{\mathcal{R}2} + L_{\mathfrak{a}1}^3 L_{\mathcal{R}3} + L_{\text{hess}})(d_x^2 + d_xd_u)^{3/2} + L_{\mathfrak{a}1}^2 L_{\mathcal{R}2}\Big).$$

We emphasize that this result lower bounds the achievable rate for *any* exploration policy with bounded power, $\pi_{\text{exp}} \in \Pi_{\gamma^2}$, and *any* decision rule. As such, it provides a lower bound on optimal decision making in LDDM.

5.1.2 Optimal Decision Making in LDDM

With our lower bound Theorem 5.1 in hand, we turn now to obtaining an upper bound in the LDDM setting. Before stating our upper bound, we first present our algorithm, TOPLE.

Algorithm 5.1 Task **OP**tima**L** **E**xperiment Design (TOPLE)

1: **Input:** Input power γ^2, initial epoch length $C_{\mathrm{init}} d_u$ ($C_{\mathrm{init}} \in \mathbb{N}$)
2: $T_0 \leftarrow C_{\mathrm{init}} d_u$, $k_0 \leftarrow C_{\mathrm{init}}$, $T \leftarrow T_0$
3: Run system for T_0 steps with $u_t \sim \mathcal{N}(0, \frac{\gamma^2}{d_u} I)$
4: **for** $i = 1, 2, 3, \ldots$ **do**
5: $\quad \widehat{\theta}_{i-1} \leftarrow \arg\min_\theta \sum_{t=1}^{T} \|x_{t+1} - \theta[x_t; u_t]\|_2^2$
6: $\quad T_i \leftarrow T_0 2^i$, $k_i \leftarrow k_0 2^{\lfloor i/4 \rfloor}$, $T \leftarrow T + T_i$
7: $\quad$ Compute input $\boldsymbol{U}_i \leftarrow \arg\min_{\boldsymbol{U} \in \mathcal{U}_{\gamma^2/2, k_i}} \mathrm{tr}\left(\mathcal{H}(\widehat{\theta}_{i-1}) \cdot \check{\boldsymbol{\Lambda}}^{\mathrm{ss}}_{T_i, T_i/d_u}(\widehat{\theta}_{i-1}, \boldsymbol{U}, \gamma/\sqrt{2d_u})^{-1}\right)$
8: $\quad (\widetilde{u}_t^i)_{t=1}^{T_i} \leftarrow$ `ConstructTimeInput`$(\boldsymbol{U}_i, T_i/d_u, k_i)$ (Algorithm 3.2)
9: $\quad$ Run system for T_i steps with $u_t = \widetilde{u}_t^i + u_t^w$, $u_t^w \sim \mathcal{N}(0, \frac{\gamma^2}{2d_u} I)$

TOPLE proceeds in epochs. At each epoch it chooses a set of exploration inputs (an exploration policy) that minimize the certainty-equivalence design objective, $\Phi_T(\pi_{\mathrm{exp}}; \widehat{\theta})$, based on the estimate $\widehat{\theta}$ of the system's parameters produced in the previous epoch. As the estimate of $\theta_\star$ is refined, the exploration policy is improved, and ultimately achieves near-optimal excitation of the system for the task of interest. Note that the form of TOPLE is extremely similar to Algorithm 3.1—indeed, they are identical except for the choice of exploration inputs on Line 7.

Remark 5.1.1 (Computational Efficiency of TOPLE)**.** Note that the set $\mathcal{U}_{\gamma^2,k}$ is convex, and that the objective is also convex, due to the convexity of $\mathrm{tr}(X^{-1})$ and since $\boldsymbol{\Lambda}^{\mathrm{ss}}_{t,t/d_u}(\widehat{\theta}, \boldsymbol{U}, \gamma/\sqrt{2d_u})$ is affine in $\boldsymbol{U}$. It follows that the optimization on Line 7 can be efficiently solved with any SDP solver. The structure of $\mathcal{U}_{\gamma^2,k}$, however, allows for an even more efficient solution. Note that any $\boldsymbol{U}$ can be projected onto $\mathcal{U}_{\gamma^2,k}$ by computing the SVD of each $U_\ell \in \boldsymbol{U}$, an operation which takes time $\mathcal{O}(kd_u^3)$. Therefore, the optimization on Line 7 can be efficiently solved by running the following projected gradient descent update:

$$\boldsymbol{U}_{j+1} \leftarrow \boldsymbol{U}_j - \eta \nabla_{\boldsymbol{U}} \mathrm{tr}\big(\mathcal{H}(\widehat{\theta}) \cdot \boldsymbol{\Lambda}^{\mathrm{ss}}_{t,t/d_u}(\widehat{\theta}, \boldsymbol{U}_j, \gamma/\sqrt{2d_u})^{-1}\big)$$
$$\boldsymbol{U}_{j+1} \leftarrow \mathrm{proj}(\boldsymbol{U}_{j+1}; \mathcal{U}_{\gamma^2/2,k})$$

where $\mathrm{proj}(\boldsymbol{U}_{j+1}; \mathcal{U}_{\gamma^2/2,k})$ denotes the projection of $\boldsymbol{U}_{j+1}$ onto $\mathcal{U}_{\gamma^2/2,k}$. It follows that TOPLE is computationally efficient.

Main Upper Bound in LDDM. The following assumption quantifies how large T must be to guarantee we achieve the optimal rate.

Assumption 5.3 (Sufficiently Large T)**.** *T is large enough that the burn-in time of Corol-*

lary 5.4, (5.4.1), *is met with* $n = c_1 \log T$ *and* $\lambda_{\mathrm{noise}}(\sigma_u) = \lambda_{\mathrm{noise}}$, *and*

$$T \geq \max\left\{ \frac{C^{\mathrm{init}}_{\textsc{Tople}} \sqrt{d_x}(\sigma_w^2 + 1)}{\lambda_{\mathrm{noise}}}, \frac{c_2(\log \frac{1}{\delta} + d \log(\beta_\infty/\lambda_{\mathrm{noise}} + 1))}{\min\{r_{\mathrm{cov}}(\theta_\star)^2, d_x^{-1} r_{\mathrm{quad}}(\theta_\star)^2, \lambda_{\mathrm{noise}}^2 L_{\mathrm{cov}}(\theta_\star, \gamma^2)^{-2}\}\lambda_{\mathrm{noise}}} \right\} \tag{5.1.3}$$

where

$$C^{\mathrm{init}}_{\textsc{Tople}} = \mathrm{poly}\left(\|A_\star\|_{\mathcal{H}_\infty}, \|B_\star\|_{\mathrm{op}}, d_u, \gamma^2, \log \frac{1}{\delta}, C_{\mathrm{init}} \right),$$

$r_{\mathrm{cov}}(\theta_\star) = C_{\mathrm{sys}}^{-1}$ *is defined as in Lemma 5.3.7 for some* $C_{\mathrm{sys}} = \mathrm{poly}(\|A_\star\|_{\mathcal{H}_\infty}, \|B_\star\|_{\mathrm{op}})$, $d = d_x + d_u$, *and* c_1, c_2 *are universal numerical constants.*

Then we have the following theorem, upper bounding the loss achieved by TOPLE.

Theorem 5.2. *Assume we are in the* LDDM *setting, that* $\mathcal{R}$ *satisfies Assumption 4.1,* $\delta \in (0, 1/3)$, *and that* T *is large enough for Assumption 5.3 to hold. Then with probability at least* $1 - \delta$, *the estimate* $\widehat{\theta}_T$ *produced by Algorithm 5.1 satisfies:*

$$\mathcal{R}(\mathfrak{a}_{\mathrm{opt}}(\widehat{\theta}_T); \theta_\star) \leq 480\sigma_w^2 \log\frac{72(d_x^2 + d_x d_u)}{\delta} \cdot \frac{\Phi_{\mathrm{opt}}(\gamma^2; \theta_\star)}{T} + \frac{C_{\mathrm{ce},1} + C_{\textsc{Tople},1}}{T^{3/2}} + \frac{C_{\mathrm{ce},2} + C_{\textsc{Tople},2}}{T^2}$$

and, furthermore, $\mathbb{E}[\sum_{t=1}^T u_t^\top u_t] \leq T\gamma^2$. *Here* $C_{\mathrm{ce},1}$ *and* $C_{\mathrm{ce},2}$ *are defined as in Corollary 5.4,*

$$C_{\textsc{Tople},1} := c_1 \left(\tfrac{\sqrt{d_x} d^2 L_{\mathrm{hess}}}{(\lambda_{\mathrm{noise}})^{3/2}} + \tfrac{L_{\mathrm{cov}}(\theta_\star, \gamma^2)\mathrm{tr}(\mathcal{H}(\theta_\star))}{(\lambda_{\mathrm{noise}})^{5/2}} \right) \sqrt{\log(1/\delta) + d\log(\beta_\infty/\lambda_{\mathrm{noise}} + 1)},$$
$$C_{\textsc{Tople},2} := c_2 \tfrac{d^2 L_{\mathrm{cov}}(\theta_\star, \gamma^2) L_{\mathrm{hess}}}{(\lambda_{\mathrm{noise}})^3} \Big(\log(1/\delta) + d\log(\beta_\infty/\lambda_{\mathrm{noise}} + 1) \Big),$$

$L_{\mathrm{cov}}(\theta_\star, \gamma^2) = C_{\mathrm{sys}}(\sigma_w^2 + \gamma^2)$ *is defined as in Lemma 5.3.7,* $d := d_x + d_u$, *and* c_1, c_2 *are universal numerical constants.*

We note that this upper bound matches the lower bound on LDDM decision making given in Theorem 5.1, up to constants. The additional burn-in and lower-order terms are required to quantify how close to optimal the inputs being played are. In particular, when T satisfies (5.1.3), we are able to show that the inputs TOPLE plays achieve near-optimal performance.

Together, Theorems 5.1 and 5.2 provide a complete instance-dependent characterization of smooth decision making in linear dynamical systems. We emphasize that the only assumptions needed for Theorems 5.1 and 5.2 to hold are that our system, $\theta_\star$, is stable and sufficiently excitable, and that the loss we are considering is sufficiently smooth. For any system and any loss satisfying these minimal assumptions, Theorems 5.1 and 5.2 show that certainty equivalence decision making is instance-wise optimal, and that TOPLE hits this optimal rate. Furthermore, while TOPLE relies on an experiment design over open-loop inputs in order to

compute the optimal exploration policy, its sample complexity is also optimal over algorithms which incorporate feedback.

5.2 Instance-Optimal Linear Quadratic Regulator

We now return to the LQR example introduced in Example 5.0.1, giving a corollary to Theorem 5.2 in this setting, and providing several examples showing that, in the LQR problem, task-directed exploration yields a provable gain over more naive exploration strategies.

5.2.1 Instance-Optimal LQR Synthesis

Consider the pure exploration LQR problem stated in Example 5.0.1. We define

$$\mathcal{R}_{\text{LQR}}(K;\theta_\star) := \mathcal{J}_{\text{LQR},\theta_\star}(K) - \min_K \mathcal{J}_{\text{LQR},\theta_\star}(K)$$

where $\mathcal{J}_{\text{LQR},\theta_\star}(K)$ is given in Example 5.0.1. The *discrete algebraic Ricatti equation*, defined for some (A, B), is given as:

$$P = A^\top PA - A^\top PB(R_{\mathbf{u}} + B^\top PB)B^\top PA + R_{\mathbf{x}}.$$

If θ is stabilizable and $R_{\mathbf{x}}, R_{\mathbf{u}} \succ 0$, it is a well-known fact that this has a unique solution, $P \succeq 0$. We denote the solution for the instance $\theta_\star = (A_\star, B_\star)$ by $P_\star$, and let $\Phi_{\text{LQR}}(\gamma^2;\theta_\star) := \Phi_{\text{opt}}(\gamma^2;\theta_\star)$ in the case when our loss is the LQR loss, $\mathcal{J}_{\theta_\star} = \mathcal{J}_{\text{LQR},\theta_\star}$. As we show in Theorem C.2, the LQR problem satisfies Assumption 4.1. Given these definitions, the following corollary shows the performance of TOPLE on the pure exploration LQR problem.

Corollary 5.3. *As long as $T \geq C_{LQR}(d_x \log^2 T + d_x^2)$, with probability at least $1-\delta$, TOPLE achieves the following rate for the LQR problem:*

$$\mathcal{R}_{LQR}(\mathsf{ce}(\boldsymbol{\tau});\theta_\star) \leq C \cdot \sigma_w^2 \log(\tfrac{d_x}{\delta}) \cdot \frac{\Phi_{LQR}(\gamma^2;\theta_\star)}{T} + \tfrac{C_{LQR} d_x^5}{T^{3/2}}.$$

Furthermore, any algorithm must incur the following loss:

$$\mathfrak{M}_{\gamma^2}(\mathcal{R}_{LQR};\mathcal{B}_T) \geq \frac{\sigma_w^2}{64} \cdot \frac{\Phi_{LQR}(\gamma^2;\theta_\star)}{T} - \tfrac{C_{LQR} d_x^5}{T^{5/4}}$$

where $\mathcal{B}_T$ is as in Theorem 5.1 and $C_{LQR} = C'_{LQR}/\min\{\sigma_w^6, \gamma^6/d_u^3, 1\}$ for C'_{LQR} polynomial in $\|P_\star\|_{\text{op}}$, $\|B_\star\|_{\text{op}}$, $\|B_\star\|_{\text{op}}^{-1}$, $\|A_\star\|_{\mathcal{H}_\infty}$, $\|R_{\mathbf{u}}\|_{\text{op}}$, γ^2, σ_w^2, d_u, $\log\log T$, and $\log\frac{1}{\delta}$.

As this result shows, TOPLE is instance-optimal for the LQR problem, with sample complexity governed by the constant $\Phi_{\text{LQR}}(\gamma^2;\theta_\star)$. To the best of our knowledge, this is the first (and only) algorithm provably instance-optimal for LQR—albeit in the pure-exploration LQR setting.

5.2.2 Task-Guided Exploration yields Provable Gains

We turn now to several examples which illustrate that taking into account the task of interest when performing exploration yields provable gains over task-agnostic exploration schemes. We focus on the LQR setting and compare against the following natural exploration baselines:

- **System Identification in Operator Norm** (Chapter 3): Let π_{op} denote the exploration policy that is optimal for estimating $\theta_\star = (A_\star, B_\star)$ under the operator norm $\|\widehat{\theta} - \theta_\star\|_{\mathrm{op}}$. Explicitly, $\pi_{\mathrm{op}} = \arg\max_{\pi \in \Pi_{\gamma^2}} \lambda_{\min}(\mathbf{\Lambda}_T(\pi; \theta_\star))$.

- **System Identification in Frobenius Norm**: Let π_{fro} denote the exploration policy that is optimal for estimating $\theta_\star = (A_\star, B_\star)$ under the Frobenius norm: $\pi_{\mathrm{fro}} = \arg\min_{\pi \in \Pi_{\gamma^2}} \mathrm{tr}(\mathbf{\Lambda}_T(\pi; \theta_\star)^{-1})$.

- **Task-Optimal Gaussian Noise**: Let π_{noise} denote the exploration policy such that $\pi_{\mathrm{noise}} =: \pi_{\mathrm{noise}}(\Gamma_\star)$ where $\pi_{\mathrm{noise}}(\Gamma_\star)$ plays the inputs $u_t \sim \mathcal{N}(0, \Gamma_\star)$ and

$$\Gamma_\star = \arg \min_{\Gamma : \mathrm{tr}(\Gamma) \le \gamma^2} \mathrm{tr}(\mathcal{H}(\theta_\star) \check{\mathbf{\Lambda}}_T(\pi_{\mathrm{noise}}(\Gamma); \theta_\star)^{-1}).$$

In stating our results, we overload notation and let $\mathcal{R}_{\mathrm{LQR}}(\pi_{\mathrm{exp}}; \theta_\star) = \mathcal{R}_{\mathrm{LQR}}(\mathsf{ce}(\boldsymbol{\tau}); \theta_\star)$ for $\boldsymbol{\tau} \sim \pi_{\mathrm{exp}}, \theta_\star$. We are concerned primarily in how the complexity scales with the dimension, d_x, and $\frac{1}{1-\rho}$ where ρ is the spectral radius of the system, and use $\Theta(\cdot)$ and $\mathcal{O}(\cdot)$ to suppress lower order dependence on these terms. Our first example shows that, if $(A_\star, B_\star)$ is properly structured, TOPLE achieves a tighter scaling in $\frac{1}{1-\rho}$ than all naive exploration approaches.

Proposition 5.1. *Consider the system $A_\star = \rho \mathbf{e}_1 \mathbf{e}_1^\top$, $B_\star = bI$, $R_{\mathbf{x}} = \kappa I$, and $R_{\mathbf{u}} = \mu I$. There exist values of b, κ, μ, and σ_w such that the loss of TOPLE, optimal operator norm identification, optimal Frobenius norm identification, and optimally exciting Gaussian noise have the following scalings:*

$$\begin{aligned}
\mathcal{R}_{LQR}(\textsc{Topple}; \theta_\star) &= \mathcal{O}\left(\tfrac{d_x^2}{(1-\rho)^2} \tfrac{\sigma_w^2}{\gamma^2 T}\right) \\
\mathcal{R}_{LQR}(\pi_{\mathrm{fro}}; \theta_\star) &= \Theta\left(\left(\tfrac{d_x^2}{(1-\rho)^2} + \tfrac{d_x}{(1-\rho)^{5/2}}\right) \tfrac{\sigma_w^2}{\gamma^2 T}\right) \\
\mathcal{R}_{LQR}(\pi_{\mathrm{op}}; \theta_\star) &= \Theta\left(\left(\tfrac{d_x^2}{(1-\rho)^2} + \tfrac{d_x}{(1-\rho)^{3}}\right) \tfrac{\sigma_w^2}{\gamma^2 T}\right) \\
\mathcal{R}_{LQR}(\pi_{\mathrm{noise}}; \theta_\star) &= \Theta\left(\left(\tfrac{(1-\rho)^{-1} + d_x^4 (1-\rho)}{(1-\rho)^2}\right) \tfrac{\sigma_w^2}{\gamma^2 T}\right).
\end{aligned}$$

TOPLE achieves the optimal scaling in $\frac{1}{1-\rho}$ and, as $\rho \to 1$, will outperform other approaches by an arbitrarily large factor. In addition, we note that Frobenius norm identification outperforms operator norm identification for this task. Intuitively, on this instance, the first coordinate is easily excited and π_{op} and π_{fro} will therefore devote the majority of their

energy to reducing the uncertainty in the remaining coordinates. However, the LQR cost will primarily be incurred in the first coordinate due to the same effect—this coordinate is easily excited and therefore the first coordinate of the state grows at a much faster rate. As such, the task-optimal allocation does the opposite of π_{op} and π_{fro} and seeks to learn the first coordinate more precisely than the remaining coordinates so as to mitigate this growth.

In our next example, our system behaves isotropically but our costs are non-isotropic. As a result, certain directions incur greater cost than others, and the task-optimal allocation seeks to primarily reduce uncertainty in these directions.

Proposition 5.2. *Consider the system $A_\star = \rho I, B_\star = I, R_{\mathbf{x}} = I + \kappa e_1 e_1^\top$ and $R_{\mathbf{u}} = \mu I$. Then there exists a choice of μ, κ, and σ_w such that*

$$\begin{aligned}
\mathcal{R}_{LQR}(\textsc{Tople}; \theta_\star) &= \mathcal{O}\left(\tfrac{1}{(1-\rho)^4}\tfrac{\sigma_w^2}{\gamma^2 T}\right) \\
\mathcal{R}_{LQR}(\pi_{\mathrm{op}}; \theta_\star) = \mathcal{R}_{LQR}(\pi_{\mathrm{fro}}; \theta_\star) &= \Theta\left(\tfrac{d_x}{(1-\rho)^4}\tfrac{\sigma_w^2}{\gamma^2 T}\right) \\
\mathcal{R}_{LQR}(\pi_{\mathrm{noise}}; \theta_\star) &= \Theta\left(\tfrac{d_x^2}{(1-\rho)^4}\tfrac{\sigma_w^2}{\gamma^2 T}\right).
\end{aligned}$$

We note that TOPLE improves on task-agnostic exploration by a factor of at least the dimensionality. These examples make clear that, in the setting of a linear dynamical system, when our goal is to perform a specific task, exploration agnostic to this task can be arbitrarily suboptimal.

5.2.3 Suboptimality of Low-Regret Algorithms

In contrast to our pure-exploration setting, where we do not incur cost during exploration, a significant body of work exists on regret-minimization for the *online* LQR problem with unknown $A_\star, B_\star$. Here the goal is to choose a *low regret* policy π_{lr} so as to minimize

$$\mathrm{Reg}_T := \mathbb{E}_{\theta_\star, \pi_{\mathrm{lr}}}\left[\textstyle\sum_{t=1}^T \ell(x_t, u_t)\right] - T \cdot \min_K \mathcal{J}_{\mathrm{LQR}, \theta_\star}(K)$$

for $\ell(x_t, u_t)$ as defined in Example 5.0.1. While our objectives differ, it would seem a natural strategy to run a low-regret algorithm for T steps to obtain a controller K_{lr}, and then evaluate the cost $\mathcal{J}_{\mathrm{LQR}, \theta_\star}$ on this K_{lr}. The following result shows that there is a fundamental tradeoff between regret and estimation; in particular, the optimal $\Theta(\sqrt{T})$ (see Simchowitz and Foster [2020]) regret translates to a (very suboptimal) $\Omega(1/\sqrt{T})$ excess risk $\mathcal{R}_{\mathrm{LQR}}(K_{\mathrm{lr}}; \theta_\star)$.

Proposition 5.3 (Suboptimality of Low Regret, Informal)**.** *For sufficiently large T and any regret bound $R \in [\sqrt{T}, T]$, any policy π_{lr} with regret $\mathbb{E}_{\pi_{\mathrm{lr}}, \theta_\star}[\mathrm{Reg}_T] \le R$ which returns a controller K_{lr} as a function of its trajectory must have $\mathbb{E}_{\theta_\star, \pi_{\mathrm{lr}}}[\mathcal{R}_{LQR}(K_{\mathrm{lr}}; \theta_\star)] = \Omega(\frac{d_u^2 d_x}{R})$.*

In particular, Proposition 5.3 implies that popular low-regret strategies, such as optimism-in-the-face-of-uncertainty [Abbasi-Yadkori and Szepesvári, 2011, Abeille and Lazaric, 2020]

or certainty equivalence online control [Mania et al., 2019, Simchowitz and Foster, 2020], are highly suboptimal in our setting. The key intuition behind the proof is that low regret algorithms converge to inputs $u_t \approx K_\star x_t$ approaching the optimal control policy; in doing so, they under-explore directions *perpendicular* to the hyperplane $\{(x, u) : u = K_\star x\}$, which are necessary for identifying the optimal control policy.

5.3 Proof of Theorem 5.1

We next proceed to the proof of Theorem 5.1. Before proving Theorem 5.1, we introduce additional notation used throughout the proofs of the results in Section 5.1.

5.3.1 Additional Notation for Linear Dynamical Systems

Covariance Notation. At the center of our analysis are the covariance matrices that arise from excitation of the linear system with a certain input. For an input sequence $\boldsymbol{u} := (u_1, \dots, u_t) \in \mathbb{R}^{td_u}$, we define the open loop input covariance

$$\mathbf{\Lambda}_t^{\mathrm{in}}(\theta, \boldsymbol{u}, x_0) := \sum_{s=0}^{t-1} x_s^{\boldsymbol{u}}(x_s^{\boldsymbol{u}})^\top \quad \text{where} \quad x_{s+1}^{\boldsymbol{u}} = Ax_s^{\boldsymbol{u}} + Bu_s, \quad x_0^{\boldsymbol{u}} = x_0.$$

We overload notation so that the above is also defined when $\boldsymbol{u} = (u_s)_{s=1}^{t'}$ for $t' \geq t$, or even infinite sequences $\boldsymbol{u} = (u_s)_{s \geq 1}$. In addition, if $\boldsymbol{u} = (u_s)_{s=1}^{t'}$ for $t' < t$, we define $\mathbf{\Lambda}_t^{\mathrm{in}}(\theta, \boldsymbol{u}, x_0)$ to be the open loop covariance when playing $\boldsymbol{u}$ periodically: that is, the input $u_t = u_{\mathrm{mod}(t,t')}$. Recall that:

$$\mathbf{\Lambda}_t^{\mathrm{noise}}(\theta, \Gamma_u) := \sum_{s=0}^{t-1} A^s \Gamma_w (A^s)^\top + \sum_{s=0}^{t-1} A^s B \Gamma_u B^\top (A^s)^\top$$

and observe that we can equivalently define

$$\mathbf{\Lambda}_t^{\mathrm{noise}}(\theta, \Gamma_u) = \mathbb{E}\left[x_t x_t^\top \mid u_s \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \Gamma_u), w_s \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \Gamma_w), s \leq t, x_0 = 0\right]$$

We also define the following, which corresponds to the total expected average covariates starting from some state x_0 and playing any input $u_t = \widetilde{u}_t + u_t^w$, where $\boldsymbol{u} = (\widetilde{u}_t)_{t=1}^k$ and $u_t^w \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma_u^2 I)$:

$$\mathbf{\Lambda}_T(\theta, \boldsymbol{u}, \sigma_u, x_0) := \frac{1}{T}\mathbf{\Lambda}_T^{\mathrm{in}}(\theta, \boldsymbol{u}, x_0) + \frac{1}{T}\sum_{t=1}^{T} \mathbf{\Lambda}_t^{\mathrm{noise}}(\theta, \sigma_u)$$

We set:

$$\widetilde{\mathbf{\Lambda}}_t^{\text{noise}} := \begin{bmatrix} \sigma_w^2 \sum_{s=0}^{t-1} A_\star^s (A_\star^s)^\top + \frac{\gamma^2}{2d_u} \sum_{s=0}^{t-1} A_\star^s B_\star B_\star^\top (A_\star^s)^\top & 0 \\ 0 & \frac{\gamma^2}{2d_u} I \end{bmatrix}.$$

Lifted Dynamical System. We set

$$\widetilde{\theta} := (\widetilde{A}, \widetilde{B}), \quad \widetilde{A} := \begin{bmatrix} A & B \\ 0 & 0 \end{bmatrix}, \quad \widetilde{B} := \begin{bmatrix} 0 \\ I \end{bmatrix} \tag{5.3.1}$$

and in particular $\widetilde{\theta}_\star := (\widetilde{A}_\star, \widetilde{B}_\star)$. Then consider the dynamical system:

$$z_{t+1} = \widetilde{A}_\star z_t + \widetilde{B}_\star u_{t+1} + \begin{bmatrix} w_t \\ 0 \end{bmatrix} \tag{5.3.2}$$

Note that $z_t = [x_t; u_t]$, where x_t is the state of (5.0.1). It follows that

$$\Sigma_T := \sum_{t=1}^{T} z_t z_t^\top = \sum_{t=1}^{T} \begin{bmatrix} x_t \\ u_t \end{bmatrix} \begin{bmatrix} x_t \\ u_t \end{bmatrix}^\top$$

so a bound on the covariates of the system (5.3.2) can be directly applied to the state-input covariates from (5.0.1). For subsequent results, we will use $z_t := [x_t; u_t]$.

Key Parameters in the Analysis. For any $\theta = (A, B)$, the $\mathcal{H}_\infty$ norm of θ is defined as:

$$\|\theta\|_{\mathcal{H}_\infty} := \max_{\omega \in [0, 2\pi]} \|(e^{\iota\omega} I - A)^{-1} B\|_{\text{op}}.$$

Lemma 5.3.1. *The following upper bound hold:* $\|\widetilde{\theta}_\star\|_{\mathcal{H}_\infty} \le 1 + (1 + \|B_\star\|_{\text{op}}) \|A_\star\|_{\mathcal{H}_\infty}$.

Recall the definition of $\tau(\theta, \rho)$ given in Section 3.4. We can relate the value of τ for a lifted system $\widetilde{\theta}$ to the original system θ.

Lemma 5.3.2. *Let $\widetilde{A}$ be defined as in (5.3.1). Then $\tau(\widetilde{A}, \rho) \le (1 + \rho^{-1} \|B\|_{\text{op}}) \tau(A, \rho)$.*

We introduce the following constants to control the smoothness of the covariates:

$$r_{\text{cov}}(\theta_\star) := \min \left\{ \frac{1 - \rho_\star}{2\tau_\star}, \frac{1}{2\|A_\star\|_{\mathcal{H}_\infty}}, 1 \right\},$$

$$L_{\text{cov}}(\theta_\star, \gamma^2) := \frac{8(\sigma_w^2 + \sigma_u^2 \|B_\star\|_{\text{op}}^2) \tau_\star^3}{(1 - \rho_\star^2)^2} + \frac{4\sigma_u^2 (\|B_\star\|_{\text{op}} + 1) \tau_\star}{1 - \rho_\star^2} + 34\gamma^2 \|A_\star\|_{\mathcal{H}_\infty}^3 (\|B_\star\|_{\text{op}} + 1)^2.$$

Lemma 5.3.7 implies that, if

$$\|\theta - \theta_\star\|_{\text{op}} \le r_{\text{cov}}(\theta_\star),$$

then for any $u \in \mathcal{U}_{\gamma^2,k}$, if T_2 is divisible by k,

$$\|\check{\mathbf{\Lambda}}^{\mathrm{ss}}_{T_1,T_2}(\theta,u,\sigma_u) - \check{\mathbf{\Lambda}}^{\mathrm{ss}}_{T_1,T_2}(\theta_\star,u,\sigma_u)\|_{\mathrm{op}} \le L_{\mathrm{cov}}(\theta_\star,\gamma^2)\cdot\|\theta-\theta_\star\|_{\mathrm{op}}.$$

This holds regardless of the loss $\mathcal{R}$.

Finally, in our analysis it will be convenient to work with a slightly different definition of the optimal risk, which we define as:

$$\Phi^{\mathrm{ss}}_{\mathrm{opt}}(\gamma^2;\theta_\star) := \liminf_{T\to\infty} \min_{u\in\mathcal{U}_{\gamma^2,T}} \mathrm{tr}\Big(\mathcal{H}(\theta_\star)\check{\mathbf{\Lambda}}^{\mathrm{ss}}_{T,T}(\widetilde{\theta}_\star,u,0)^{-1}\Big)$$

As the following result shows, Φ_{opt} and $\Phi^{\mathrm{ss}}_{\mathrm{opt}}$ are equal up to absolute constants.

Lemma 5.3.3. *$\Phi_{\mathrm{opt}}(\gamma^2;\theta_\star)$ and $\Phi^{\mathrm{ss}}_{\mathrm{opt}}(\gamma^2;\theta_\star)$ are equal up to constants:*

$$\frac{1}{4}\Phi_{\mathrm{opt}}(\gamma^2;\theta_\star) \le \Phi^{\mathrm{ss}}_{\mathrm{opt}}(\gamma^2;\theta_\star) \le 16\Phi_{\mathrm{opt}}(\gamma^2;\theta_\star).$$

5.3.2 Proof of Theorem 5.1

The outline of the proof is as follows:

1. Apply Theorem 4.5 to show that

   $$\min_{\widehat{\mathfrak{a}}}\max_{\theta\in\mathcal{B}_T}\mathbb{E}_{\theta,\pi_{\mathrm{exp}}}[\mathcal{R}(\widehat{\mathfrak{a}}_T;\theta)] \ge \min_{\theta:\|\theta-\theta_0\|_F^2\le 5(d_x^2+d_xd_u)/(\lambda T^{5/6})} \mathbb{E}\left[\mathrm{tr}\left(\mathcal{H}(\theta_\star)\left(\mathbb{E}_{\theta,\pi_{\mathrm{exp}}}[\mathbf{\Sigma}_T]+\lambda T\cdot I\right)^{-1}\right)\right]$$

   for a particular choice of λ.

2. Apply Lemma 5.3.5 to show that, for any policy π_{exp} and any θ, there exists a *periodic* policy π'_{exp} such that

   $$\mathbb{E}_{\theta,\pi_{\mathrm{exp}}}[\mathbf{\Sigma}_T] \preceq \mathbb{E}_{\theta,\pi'_{\mathrm{exp}}}[\mathbf{\Sigma}_{c_1T}] + c_2.$$

3. Given that π'_{exp} is periodic, apply Lemma 5.3.6 to show that we can upper bound the expected covariates by the expected *steady-state* covariates:

   $$\mathbb{E}_{\theta,\pi'_{\mathrm{exp}}}[\mathbf{\Sigma}_{c_1T}] \preceq \mathbb{E}_{\theta,\pi'_{\mathrm{exp}}}[\mathbf{\Lambda}^{\mathrm{freq}}_{c_1T}(\widetilde{\theta},\boldsymbol{U})] + c_3.$$

4. Use the frequency-domain representation to show that there exists a non-random input $\boldsymbol{U}'$ that meets the power constraint and achieves the same steady state covariates:

   $$\mathbb{E}_{\theta,\pi'_{\mathrm{exp}}}[\mathbf{\Lambda}^{\mathrm{freq}}_{c_1T}(\widetilde{\theta},\boldsymbol{U})] = \mathbf{\Lambda}^{\mathrm{freq}}_{c_1T}(\widetilde{\theta},\boldsymbol{U}').$$

5. Apply the perturbation bound for the steady state covariates given in Lemma 5.3.7 to show that, for any θ in our set, we can upper bound the covariates on θ by the

covariates on $\theta_\star$:

$$\mathbf{\Lambda}^{\text{freq}}_{c_1 T}(\widetilde{\theta}, \boldsymbol{U}') \preceq \mathbf{\Lambda}^{\text{freq}}_{c_1 T}(\widetilde{\theta}_\star, \boldsymbol{U}') + c_4.$$

6. Finally, we conclude the proof by optimizing over π'_{exp} to obtain a lower bound scaling as $\Phi_{\text{opt}}(\gamma^2; \theta_\star)$.

Throughout the proof, we assume expectations are taken with respect to θ and π_{exp}, and therefore write $\mathbb{E}[\cdot]$ in place of $\mathbb{E}_{\theta, \pi_{\text{exp}}}[\cdot]$. As stated in Section 5.1, linear dynamical systems are simply an instance of vector regression and we can therefore apply the results of Chapter 4 in this setting.

Applying Theorem 4.5: Define

$$\lambda^\star_{\text{min},\infty} := \limsup_{T' \to \infty} \max_{\boldsymbol{U} \in \mathcal{U}_{\gamma^2, T'}} \frac{1}{10T'} \lambda_{\min}(\mathbf{\Lambda}^{\text{ss}}_{T'}(\widetilde{\theta}_\star, \boldsymbol{U}, 0)).$$

The following result gives a lower bound on this quantity.

Lemma 5.3.4. *We have:*

$$\min\left\{\lambda_{\min}\left(\sigma_w^2 \sum_{t=0}^{d_x} A_\star^t (A_\star^t)^\top + \sigma_u^2 \sum_{t=0}^{d_x} A_\star^t B_\star B_\star^\top (A_\star^t)^\top\right), \sigma_u^2\right\} \le \limsup_{T \to \infty} \sup_{\boldsymbol{U} \in \mathcal{U}_{\gamma^2, T}} 2\lambda_{\min}(\mathbf{\Lambda}^{\text{ss}}_T(\widetilde{\theta}_\star, \boldsymbol{U})).$$

Under Assumption 5.2 and by Lemma 5.3.4, we will have that $\lambda^\star_{\text{min},\infty} > 0$. Then the first conclusion of Theorem 4.5 holds with $\lambda = \lambda^\star_{\text{min},\infty}$. That is, if T is large enough that the burn-in of Theorem 4.5 is met, we will have

$$\min_{\widehat{\mathfrak{a}}} \max_{\theta \in \mathcal{B}_T} \mathbb{E}_{\theta, \pi_{\text{exp}}}[\mathcal{R}(\widehat{\mathfrak{a}}_T; \theta)] \ge \min_{\theta : \|\theta - \theta_0\|_F^2 \le 5(d_x^2 + d_x d_u)/(\lambda^\star_{\text{min},\infty} T^{5/6})} \mathbb{E}\left[\operatorname{tr}\left(\mathcal{H}(\theta_\star)\left(\mathbb{E}[\mathbf{\Sigma}_T] + \lambda^\star_{\text{min},\infty} T \cdot I\right)^{-1}\right)\right].$$

Sufficiency of periodic policies: Our goal is to lower bound

$$\min_{\theta : \|\theta - \theta_0\|_F^2 \le 5(d_x^2 + d_x d_u)/(\lambda^\star_{\text{min},\infty} T^{5/6})} \mathbb{E}\left[\operatorname{tr}\left(\mathcal{H}(\theta_\star)\left(\mathbb{E}[\mathbf{\Sigma}_T] + \lambda^\star_{\text{min},\infty} T \cdot I\right)^{-1}\right)\right].$$

Fix some θ such that $\|\theta_0 - \theta\|_F^2 \le 5(d_x^2 + d_x d_u)/(\lambda^\star_{\text{min},\infty} T^{5/6})$, and consider the extended system $\widetilde{\theta}$, as defined in (5.3.1). Let $z_t^{\boldsymbol{u}}$ denote the component of the state of $\widetilde{\theta}$ driven by both the random and deterministic components of the input and $z_t^{\boldsymbol{w}}$ the component driven by the process noise. Then $z_t = z_t^{\boldsymbol{u}} + z_t^{\boldsymbol{w}}$, so

$$\sum_{t=1}^T z_t z_t^\top = \sum_{t=1}^T (z_t^{\boldsymbol{u}} + z_t^{\boldsymbol{w}})(z_t^{\boldsymbol{u}} + z_t^{\boldsymbol{w}})^\top \preceq 2 \sum_{t=1}^T \left(z_t^{\boldsymbol{u}} (z_t^{\boldsymbol{u}})^\top + z_t^{\boldsymbol{w}} (z_t^{\boldsymbol{w}})^\top\right)$$

Therefore,

$$\mathbb{E}[\Sigma_T] \preceq 2\mathbb{E}[\textstyle\sum_{t=1}^{T} z_t^{\boldsymbol{u}}(z_t^{\boldsymbol{u}})^\top] + 2\sum_{t=1}^{T} \boldsymbol{\Lambda}_t^{\mathrm{noise}}(\widetilde{\theta}, 0).$$

The following result shows that we can upper bound these covariates by some covariates generated by a periodic input satisfying the same power constraint.

Lemma 5.3.5. *Fix some $\rho \geq \rho(A)$, parameter $\epsilon > 0$, and let*

$$H_\epsilon := \left\lceil \log\left(\frac{\epsilon(1-\rho^2)}{8\|B\|_{\mathrm{op}}^2 \tau(A,\rho)^3 \gamma^2 T^2}\right) / \log \rho \right\rceil \quad \text{and} \quad T_\epsilon := 2H_\epsilon((d_x^2 + d_x)/2 + 1). \tag{5.3.3}$$

Consider some input $\{u_t\}_{t=0}^{T-1}$ satisfying $\mathbb{E}[\sum_{t=0}^{T-1} u_t^\top u_t] \leq T\gamma^2$. Then there exists an input $\{\widetilde{u}_t\}_{t=0}^{T_\epsilon - 1}$ such that

$$\mathbb{E}\left[\sum_{t=0}^{T_\epsilon - 1} \widetilde{u}_t^\top \widetilde{u}_t\right] \leq T_\epsilon \gamma^2$$

and extending to times $t \geq T_\epsilon - 1$ via a periodic signal $\widetilde{u}_t = \widetilde{u}_{\mathrm{mod}(t,T_\epsilon)}$, where equality here holds almost surely, satisfies

$$\mathbb{E}\left[\sum_{t=1}^{T} x_t^{\boldsymbol{u}}(x_t^{\boldsymbol{u}})^\top\right] \preceq \mathbb{E}\left[\sum_{t=1}^{2T+3T_\epsilon/2} x_t^{\widetilde{\boldsymbol{u}}}(x_t^{\widetilde{\boldsymbol{u}}})^\top\right] + 5\epsilon I,$$

where above, $x^{\boldsymbol{u}}$ are the states under the initial inputs (u_t), and $x^{\widetilde{\boldsymbol{u}}}$ are the iterates under $(\tilde{u}_t)$, and where we take $x_0^{\boldsymbol{u}} = 0$ in both.

By the power constraint on u_t and Lemma 5.3.5,

$$\mathbb{E}\sum_{t=1}^{T} z_t^{\boldsymbol{u}}(z_t^{\boldsymbol{u}})^\top \preceq \mathbb{E}\sum_{t=1}^{2T+T_\epsilon} z_t^{\widetilde{\boldsymbol{u}}}(z_t^{\widetilde{\boldsymbol{u}}})^\top + 5\epsilon I \preceq \mathbb{E}\sum_{t=1}^{4T} z_t^{\widetilde{\boldsymbol{u}}}(z_t^{\widetilde{\boldsymbol{u}}})^\top + 5\epsilon I$$

for some input $\widetilde{\boldsymbol{u}}$ with period $k_{\widetilde{\boldsymbol{u}}} := T_\epsilon = 2H((d_x^2+d_x)/2+1)$ satisfying $\mathbb{E}[\sum_{t=0}^{k_{\widetilde{\boldsymbol{u}}}-1} \widetilde{u}_t^\top \widetilde{u}_t] \leq k_{\widetilde{\boldsymbol{u}}}\gamma^2$, and

$$H = \left\lceil \log\left(\frac{\epsilon(1-\rho^2)}{8\tau(\widetilde{A},\rho)^3\gamma^2 T^2}\right) / \log\rho \right\rceil.$$

The final inequality follows by upper bounding $T_\epsilon \leq 2T$, which will hold by our definition of H and assumption on the size of T. Choosing $\epsilon = \lambda_{\mathrm{min},\infty}^\star T/5$, we can upper bound

$$\mathbb{E}[\textstyle\sum_{t=1}^{T} z_t^{\boldsymbol{u}}(z_t^{\boldsymbol{u}})^\top] + \sum_{t=1}^{T} \boldsymbol{\Lambda}_t^{\mathrm{noise}}(\widetilde{\theta}, 0) \preceq \mathbb{E}\sum_{t=1}^{4T} z_t^{\widetilde{\boldsymbol{u}}}(z_t^{\widetilde{\boldsymbol{u}}})^\top + \sum_{t=1}^{T} \boldsymbol{\Lambda}_t^{\mathrm{noise}}(\widetilde{\theta}, 0) + \lambda_{\mathrm{min},\infty}^\star T \cdot I$$

From time domain to frequency domain: Next, the following result lets us move from the time domain to the frequency domain.

Lemma 5.3.6. *Let $\{u_t\}_{t=0}^{k-1}$ be a signal with $\mathbb{E}[\sum_{t=0}^{k-1} u_t^\top u_t] \le k\gamma^2$. Consider playing u_t periodically for T steps on system $\theta = (A, B)$ with no process noise, where we assume $x_0 = 0$ and set $u_t = u_{\mathrm{mod}(t,k)}$ almost surely. Then,*

$$\left\|\mathbb{E}\sum_{t=0}^{T} x_t x_t^\top - \mathbb{E}\frac{1}{T}\sum_{t=1}^{T}(e^{\iota\frac{2\pi t}{T}}I - A)^{-1}B\check{u}_t\check{u}_t^{\mathsf{H}}B^{\mathsf{H}}(e^{\iota\frac{2\pi t}{T}}I - A)^{-\mathsf{H}}\right\|_{\mathrm{op}} \le \frac{\tau(A,\rho)\|\theta\|_{\mathcal{H}_\infty}^2\|B\|_{\mathrm{op}}\sqrt{T+1}k\gamma^2}{1-\rho^k} + \frac{\tau(A,\rho)^2\|\theta\|_{\mathcal{H}_\infty}^2\|B\|_{\mathrm{op}}^2k^2\gamma^2}{(1-\rho^k)^2}.$$

where $(\check{u}_t)_{t=1}^T = \mathfrak{F}^{-1}((u_t)_{t=1}^T)$.

The conditions of Lemma 5.3.6 are met for the above $\widetilde{\boldsymbol{u}}$ with $k \leftarrow 2H((d_x^2+d_x)/2+1)$, so it follows that

$$\mathbb{E}\sum_{t=1}^{4T} z_t^{\widetilde{\boldsymbol{u}}}(z_t^{\widetilde{\boldsymbol{u}}})^\top \preceq \mathbb{E}\boldsymbol{\Lambda}_{4T}^{\mathrm{freq}}(\widetilde{\theta},\widetilde{\boldsymbol{u}}) + \left(\frac{\tau(\widetilde{A},\rho)(2H((d_x^2+d_x)/2+1))\sqrt{4T+1}}{1-\rho^k} + \frac{\tau(\widetilde{A},\rho)^2(2H((d_x^2+d_x)/2+1))^2}{(1-\rho^k)^2}\right)\|\widetilde{\theta}\|_{\mathcal{H}_\infty}^2\gamma^2\cdot I.$$

Then if

$$T \ge \frac{1}{\lambda_{\mathrm{min},\infty}^\star}\left(\frac{\tau(\widetilde{A},\rho)(2H((d_x^2+d_x)/2+1))\sqrt{4T+1}}{1-\rho^k} + \frac{\tau(\widetilde{A},\rho)^2(2H((d_x^2+d_x)/2+1))^2}{(1-\rho^k)^2}\right)\|\widetilde{\theta}\|_{\mathcal{H}_\infty}^2\gamma^2, \tag{5.3.4}$$

we can upper bound

$$\mathbb{E}\sum_{t=1}^{4T} z_t^{\widetilde{\boldsymbol{u}}}(z_t^{\widetilde{\boldsymbol{u}}})^\top + \textstyle\sum_{t=1}^T\boldsymbol{\Lambda}_t^{\mathrm{noise}}(\widetilde{\theta},0) + \lambda_{\mathrm{min},\infty}^\star T\cdot I \preceq \mathbb{E}\boldsymbol{\Lambda}_{4T}^{\mathrm{freq}}(\widetilde{\theta},\widetilde{\boldsymbol{u}}) + \textstyle\sum_{t=1}^T\boldsymbol{\Lambda}_t^{\mathrm{noise}}(\widetilde{\theta},0) + 2\lambda_{\mathrm{min},\infty}^\star T\cdot I.$$

Sufficiency of deterministic inputs: Let $\widetilde{\mathcal{U}}_{\gamma^2,k_{\widetilde{\boldsymbol{u}}}}$ denote the set of inputs with average expected power bounded by γ^2 and period $k_{\widetilde{\boldsymbol{u}}}$. Then we have shown that

$$\begin{aligned}
&\mathrm{tr}\Big(\mathcal{H}(\theta_\star)(\mathbb{E}[\boldsymbol{\Sigma}_T] + \lambda_{\mathrm{min},\infty}^\star T\cdot I)^{-1}\Big)\\
&\ge \frac{1}{2}\mathrm{tr}\Big(\mathcal{H}(\theta_\star)\left(\mathbb{E}\check{\boldsymbol{\Lambda}}_{4T}^{\mathrm{in,ss}}(\widetilde{\theta},\widetilde{\boldsymbol{u}}) + \textstyle\sum_{t=1}^T\check{\boldsymbol{\Lambda}}_t^{\mathrm{noise}}(\widetilde{\theta},0) + 3\lambda_{\mathrm{min},\infty}^\star T\cdot I\right)^{-1}\Big)\\
&\ge \frac{1}{2}\min_{\boldsymbol{u}\in\widetilde{\mathcal{U}}_{\gamma^2,k_{\widetilde{\boldsymbol{u}}}}}\mathrm{tr}\Big(\mathcal{H}(\theta_\star)\left(\mathbb{E}\check{\boldsymbol{\Lambda}}_{4T}^{\mathrm{in,ss}}(\widetilde{\theta},\boldsymbol{u}) + \textstyle\sum_{t=1}^T\check{\boldsymbol{\Lambda}}_t^{\mathrm{noise}}(\widetilde{\theta},0) + 3\lambda_{\mathrm{min},\infty}^\star T\cdot I\right)^{-1}\Big)
\end{aligned}$$

By definition of $\mathbf{\Lambda}^{\text{freq}}$ and for any $\boldsymbol{u} \in \widetilde{\mathcal{U}}_{\gamma^2,k_{\widetilde{\boldsymbol{u}}}}$, using that $\check{\boldsymbol{u}} = \mathfrak{F}(\boldsymbol{u})$,

$$\begin{aligned}\mathbb{E}[\mathbf{\Lambda}^{\text{freq}}_{4T}(\widetilde{\theta}, \boldsymbol{u})] &= \mathbb{E}\left[\frac{4T}{k_{\widetilde{\boldsymbol{u}}}}\frac{1}{k_{\widetilde{\boldsymbol{u}}}}\sum_{t=1}^{k_{\widetilde{\boldsymbol{u}}}}(e^{\iota\frac{2\pi t}{k_{\widetilde{\boldsymbol{u}}}}}I - \widetilde{A})^{-1}\widetilde{B}\check{u}_t\check{u}_t^{\mathsf{H}}\widetilde{B}^{\mathsf{H}}(e^{\iota\frac{2\pi t}{k_{\widetilde{\boldsymbol{u}}}}}I - \widetilde{A})^{-\mathsf{H}}\right] \\ &= \frac{4T}{k_{\widetilde{\boldsymbol{u}}}}\frac{1}{k_{\widetilde{\boldsymbol{u}}}}\sum_{t=1}^{k_{\widetilde{\boldsymbol{u}}}}(e^{\iota\frac{2\pi t}{k_{\widetilde{\boldsymbol{u}}}}}I - \widetilde{A})^{-1}\widetilde{B}\mathbb{E}[\check{u}_t\check{u}_t^{\mathsf{H}}]\widetilde{B}^{\mathsf{H}}(e^{\iota\frac{2\pi t}{k_{\widetilde{\boldsymbol{u}}}}}I - \widetilde{A})^{-\mathsf{H}}.\end{aligned}$$

Define $U_t := \mathbb{E}[\check{u}_t\check{u}_t^{\mathsf{H}}]$. By Parseval's Theorem, and the power constraint on $\boldsymbol{u}$, we have

$$\textstyle\sum_{t=1}^{k_{\widetilde{\boldsymbol{u}}}}\operatorname{tr}(U_t) = \mathbb{E}[\sum_{t=1}^{k_{\widetilde{\boldsymbol{u}}}}\check{u}_t\check{u}_t^{\mathsf{H}}] = \mathbb{E}[k_{\widetilde{\boldsymbol{u}}}\sum_{t=0}^{k_{\widetilde{\boldsymbol{u}}}}u_t^\top u_t] \le k_{\widetilde{\boldsymbol{u}}}^2\gamma^2.$$

Thus, optimizing over over the (possibly random) input $\boldsymbol{u}$, is equivalent to optimizing over PSD matrices U_t that satisfy this trace constraint. Therefore,

$$\begin{aligned}&\frac{1}{2}\min_{\boldsymbol{u}\in\widetilde{\mathcal{U}}_{\gamma^2,k_{\widetilde{\boldsymbol{u}}}}}\operatorname{tr}\left(\mathcal{H}(\theta_\star)\left(\mathbb{E}\check{\mathbf{\Lambda}}^{\text{in,ss}}_{4T}(\widetilde{\theta},\boldsymbol{u}) + \textstyle\sum_{t=1}^{T}\check{\mathbf{\Lambda}}^{\text{noise}}_t(\widetilde{\theta},0) + 3\lambda^\star_{\min,\infty}T\cdot I\right)^{-1}\right) \\ &\quad\ge \frac{1}{2}\min_{\boldsymbol{u}\in\mathcal{U}_{\gamma^2,k_{\widetilde{\boldsymbol{u}}}}}\operatorname{tr}\left(\mathcal{H}(\theta_\star)\left(\check{\mathbf{\Lambda}}^{\text{in,ss}}_{4T}(\widetilde{\theta},\boldsymbol{u}) + \textstyle\sum_{t=1}^{T}\check{\mathbf{\Lambda}}^{\text{noise}}_t(\widetilde{\theta},0) + 3\lambda^\star_{\min,\infty}T\cdot I\right)^{-1}\right) \\ &\quad\ge \frac{1}{8T}\min_{\boldsymbol{u}\in\mathcal{U}_{\gamma^2,k_{\widetilde{\boldsymbol{u}}}}}\operatorname{tr}\left(\mathcal{H}(\theta_\star)\left(\check{\mathbf{\Lambda}}^{\text{ss}}_{4T}(\widetilde{\theta},\boldsymbol{u},0) + 3\lambda^\star_{\min,\infty}\cdot I\right)^{-1}\right) \\ &\quad\ge \frac{1}{8T}\min_{\boldsymbol{U}\in\mathcal{U}_{\gamma^2,4T}}\operatorname{tr}\left(\mathcal{H}(\theta_\star)\left(\check{\mathbf{\Lambda}}^{\text{ss}}_{4T}(\widetilde{\theta},\boldsymbol{U},0) + 3\lambda^\star_{\min,\infty}\cdot I\right)^{-1}\right)\end{aligned}$$

where the constraint set in the second minimization is simply the set defined in (3.1.2), and we can thus drop the expectation.

From θ to $\theta_\star$: We next want to show that, for θ close enough to $\theta_\star$, the covariates generated on θ are close to those generated on $\theta_\star$. The following result provides such a bound.

Lemma 5.3.7. *For all $\boldsymbol{U} \in \mathcal{U}_{\gamma^2,k}$ and all θ with*

$$\|\theta - \theta_\star\|_{\text{op}} \le \min\left\{\frac{1-\rho}{2\tau(A_\star,\rho)}, \frac{1}{2\|A_\star\|_{\mathcal{H}_\infty}}, 1\right\} =: r_{\text{cov}}(\theta_\star), \tag{5.3.5}$$

if T_2 is divisible by k, then:

$$\begin{aligned}\|\check{\mathbf{\Lambda}}^{\text{ss}}_{T_1,T_2}(\theta,\boldsymbol{U},\sigma_u) - \check{\mathbf{\Lambda}}^{\text{ss}}_{T_1,T_2}(\theta_\star,\boldsymbol{U},\sigma_u)\|_{\text{op}} \le \Bigg(&\frac{8(\sigma_w^2+\sigma_u^2\|B_\star\|_{\text{op}}^2)\tau(A_\star,\rho)^3}{(1-\rho^2)^2} + \frac{4\sigma_u^2(\|B_\star\|_{\text{op}}+1)\tau(A_\star,\rho)^2}{1-\rho^2} \\ &+ 34\gamma^2\|A_\star\|_{\mathcal{H}_\infty}^3(\|B_\star\|_{\text{op}}+1)^2\Bigg)\|\theta-\theta_\star\|_{\text{op}} =: L_{\text{cov}}(\theta_\star,\gamma^2)\cdot\|\theta-\theta_\star\|_{\text{op}}.\end{aligned} \tag{5.3.6}$$

By assumption

$$\|\theta-\theta_\star\|_{\text{op}} \le \|\theta-\theta_\star\|_F \le \|\theta-\theta_0\|_F + \|\theta_0-\theta_\star\|_F \le 2\sqrt{5(d_x^2+d_xd_u)}/(\sqrt{\lambda^\star_{\min,\infty}}T^{5/12})$$

so if

$$2\sqrt{5(d_x^2+d_xd_u)}/(\sqrt{\lambda^\star_{\min,\infty}}T^{5/12}) \le r_{\text{cov}}(\theta_\star) \tag{5.3.7}$$

we are in the domain of Lemma 5.3.7 and

$$\begin{aligned}\|\check{\mathbf{\Lambda}}^{\text{ss}}_{4T}(\widetilde{\theta},\boldsymbol{U},0)-\check{\mathbf{\Lambda}}^{\text{ss}}_{4T}(\widetilde{\theta}_\star,\boldsymbol{U},0)\|_{\text{op}} &\le L_{\text{cov}}(\theta_\star,\gamma^2)\cdot\|\widetilde{\theta}-\widetilde{\theta}_\star\|_{\text{op}} = L_{\text{cov}}(\theta_\star,\gamma^2)\cdot\|\theta-\theta_\star\|_{\text{op}}\\ &\le \frac{L_{\text{cov}}(\theta_\star,\gamma^2)2\sqrt{5(d_x^2+d_xd_u)}}{\sqrt{\lambda^\star_{\min,\infty}}T^{5/12}},\end{aligned}$$

It follows that as long as

$$\frac{L_{\text{cov}}(\theta_\star,\gamma^2)2\sqrt{5(d_x^2+d_xd_u)}}{\sqrt{\lambda^\star_{\min,\infty}}T^{5/12}} \le \lambda^\star_{\min,\infty} \tag{5.3.8}$$

then

$$\check{\mathbf{\Lambda}}^{\text{ss}}_{4T}(\widetilde{\theta},\boldsymbol{U},0) \preceq \check{\mathbf{\Lambda}}^{\text{ss}}_{4T}(\widetilde{\theta}_\star,\boldsymbol{U},0) + \lambda^\star_{\min,\infty}\cdot I,$$

and thus,

$$\begin{aligned}&\frac{1}{8T}\min_{\boldsymbol{U}\in\mathcal{U}_{\gamma^2,4T}} \text{tr}\Big(\mathcal{H}(\theta_\star)\left(\check{\mathbf{\Lambda}}^{\text{ss}}_{4T}(\widetilde{\theta},\boldsymbol{U},0)+3\lambda^\star_{\min,\infty}\cdot I\right)^{-1}\Big)\\ &\quad\ge \frac{1}{8T}\min_{\boldsymbol{U}\in\mathcal{U}_{\gamma^2,4T}} \text{tr}\Big(\mathcal{H}(\theta_\star)\left(\check{\mathbf{\Lambda}}^{\text{ss}}_{4T}(\widetilde{\theta}_\star,\boldsymbol{U},0)+4\lambda^\star_{\min,\infty}\cdot I\right)^{-1}\Big).\end{aligned}$$

Concluding the lower bound: Next, the following lemmas show that for large enough time horizon, the covariates can approximate the infinite-horizon covariates.

Lemma 5.3.8. *Fix any $\bar{k}$ and input $\boldsymbol{U}^\star \in \mathcal{U}_{\gamma^2,\bar{k}}$. Then for any:*

$$k \ge \max\left\{\frac{4\pi\|\theta\|_{\mathcal{H}_\infty}\gamma^2}{\epsilon}, \frac{\pi}{2\|\theta\|_{\mathcal{H}_\infty}}\right\}\left(\max_{\omega\in[0,2\pi]}\|(e^{\iota\omega}I-A)^{-2}B\|_{\text{op}}\right)$$

there exists an input $\boldsymbol{U}\in\mathcal{U}_{\gamma^2,k}$ such that:

$$\left\|\frac{1}{k}\mathbf{\Lambda}^{\text{freq}}_k(\theta,\boldsymbol{U}) - \frac{1}{\bar{k}}\mathbf{\Lambda}^{\text{freq}}_{\bar{k}}(\theta,\boldsymbol{U}^\star)\right\|_{\text{op}} \le \epsilon.$$

Lemma 5.3.9. *If*

$$T \geq \max\left\{\frac{8\tau(A,\rho)^2(\sigma_w^2+\sigma_u^2\|B\|_{\mathrm{op}}^2)}{(1-\rho^2)^2}\frac{1}{\epsilon}, \log\left(\frac{(1-\rho^2)^2\epsilon}{8\tau(A,\rho)^2(\sigma_w^2+\sigma_u^2\|B\|_{\mathrm{op}}^2)}\right)\frac{1}{2\log\rho}\right\}$$

then, for any $T' \geq T$,

$$\left\|\frac{1}{T}\sum_{t=1}^{T}\mathbf{\Lambda}_t^{\mathrm{noise}}(\theta,\sigma_u) - \frac{1}{T'}\sum_{t=1}^{T'}\mathbf{\Lambda}_t^{\mathrm{noise}}(\theta,\sigma_u)\right\|_{\mathrm{op}} \leq \epsilon.$$

By Lemma 5.3.8, so long as

$$4T \geq \max\left\{\frac{8\pi\|\widetilde{\theta}_\star\|_{\mathcal{H}_\infty}\gamma^2}{\lambda^\star_{\mathrm{min},\infty}}, \frac{\pi}{2\|\widetilde{\theta}_\star\|_{\mathcal{H}_\infty}}\right\}\left(\max_{\omega\in[0,2\pi]}\|(e^{\iota\omega}I-\widetilde{A}_\star)^{-2}\widetilde{B}_\star\|_{\mathrm{op}}\right), \tag{5.3.9}$$

then for any $T' \geq 4T$ and $\boldsymbol{U}^\star \in \mathcal{U}_{\gamma^2,4T}$, there exists a $\boldsymbol{U}' \in \mathcal{U}_{\gamma^2,T'}$ such that

$$\left\|\frac{1}{4T}\mathbf{\Lambda}_{4T}^{\mathrm{freq}}(\widetilde{\theta}_\star,\boldsymbol{U}') - \frac{1}{T'}\mathbf{\Lambda}_{T'}^{\mathrm{freq}}(\widetilde{\theta}_\star,\boldsymbol{U}^\star)\right\|_{\mathrm{op}} \leq \frac{1}{2}\lambda^\star_{\mathrm{min},\infty}.$$

Furthermore, by Lemma 5.3.9, if

$$4T \geq \max\left\{\frac{16\tau(\widetilde{A}_\star,\rho)^2(\sigma_w^2+\gamma^2/d_u)}{(1-\rho^2)^2\lambda^\star_{\mathrm{min},\infty}}, \log\left(\frac{(1-\rho^2)^2\lambda^\star_{\mathrm{min},\infty}}{16\tau(\widetilde{A}_\star,\rho)^2(\sigma_w^2+\gamma^2/d_u)}\right)\frac{1}{2\log\rho}\right\}, \tag{5.3.10}$$

then, for any $T' \geq 4T$,

$$\left\|\frac{1}{4T}\sum_{t=1}^{4T}\mathbf{\Lambda}_t^{\mathrm{noise}}(\widetilde{\theta}_\star,\gamma/\sqrt{d_u}) - \frac{1}{T'}\sum_{t=1}^{T'}\mathbf{\Lambda}_t^{\mathrm{noise}}(\widetilde{\theta}_\star,\gamma/\sqrt{d_u})\right\|_{\mathrm{op}} \leq \frac{1}{2}\lambda^\star_{\mathrm{min},\infty}.$$

By what we've just shown, for any $T' \geq 4T$:

$$\check{\mathbf{\Lambda}}_{4T}^{\mathrm{ss}}(\widetilde{\theta}_\star,\boldsymbol{U}',0) \preceq \check{\mathbf{\Lambda}}_{T'}^{\mathrm{ss}}(\widetilde{\theta}_\star,\boldsymbol{U}^\star,0) + \lambda^\star_{\mathrm{min},\infty}\cdot I.$$

Thus,

$$\begin{aligned}
&\frac{1}{8T}\min_{\boldsymbol{U}\in\mathcal{U}_{\gamma^2,4T}}\mathrm{tr}\left(\mathcal{H}(\theta_\star)\left(\check{\mathbf{\Lambda}}_{4T}^{\mathrm{ss}}(\widetilde{\theta}_\star,\boldsymbol{U},0)+4\lambda^\star_{\mathrm{min},\infty}\cdot I\right)^{-1}\right)\\
&\qquad\geq \liminf_{T'\to\infty}\frac{1}{8T}\min_{\boldsymbol{U}\in\mathcal{U}_{\gamma^2,T'}}\mathrm{tr}\left(\mathcal{H}(\theta_\star)\left(\check{\mathbf{\Lambda}}_{T'}^{\mathrm{ss}}(\widetilde{\theta}_\star,\boldsymbol{U},0)+5\lambda^\star_{\mathrm{min},\infty}\cdot I\right)^{-1}\right).
\end{aligned}$$

Let

$$\lambda^\star_{\min,T} := \inf_{T' \geq T} \max_{\boldsymbol{U} \in \mathcal{U}_{\gamma^2,T'}} \frac{1}{10T'} \lambda_{\min}(\boldsymbol{\Lambda}^{\mathrm{freq}}_{T'}(\widetilde{\theta}_\star, \boldsymbol{U})).$$

Note that by Lemma 5.3.8, so long as

$$T \geq \max\left\{ \frac{8\pi \|\widetilde{\theta}_\star\|_{\mathcal{H}_\infty} \gamma^2}{\lambda^\star_{\min,\infty}}, \frac{\pi}{2\|\widetilde{\theta}_\star\|_{\mathcal{H}_\infty}} \right\} \left(\max_{\omega \in [0,2\pi]} \|(e^{\iota\omega} I - \widetilde{A}_\star)^{-2} \widetilde{B}_\star\|_{\mathrm{op}} \right), \tag{5.3.11}$$

then for any $T' \geq T$ and $\boldsymbol{U}^\star \in \mathcal{U}_{\gamma^2,T}$, there exists a $\boldsymbol{U}' \in \mathcal{U}_{\gamma^2,T'}$ such that

$$\left\| \frac{1}{T} \boldsymbol{\Lambda}^{\mathrm{freq}}_T(\widetilde{\theta}_\star, \boldsymbol{U}') - \frac{1}{T'} \boldsymbol{\Lambda}^{\mathrm{freq}}_{T'}(\widetilde{\theta}_\star, \boldsymbol{U}^\star) \right\|_{\mathrm{op}} \leq \frac{1}{2} \lambda^\star_{\min,\infty}.$$

This implies that so long as T satisfies (5.3.11), we will have $\lambda^\star_{\min,T} \geq \frac{1}{2}\lambda^\star_{\min,\infty}$. By definition of $\lambda^\star_{\min,T}$, for any $T' \geq T$ there exists some input $\boldsymbol{U}'' \in \mathcal{U}_{\gamma^2,T'}$ such that $\lambda_{\min}(\boldsymbol{\Lambda}^{\mathrm{ss}}_{T'}(\widetilde{\theta}_\star, \boldsymbol{U}'', 0)) \geq 10\lambda^\star_{\min,T}$. It follows that for any $T' \geq T$,

$$\begin{aligned}
&\min_{\boldsymbol{U} \in \mathcal{U}_{\gamma^2,T'}} \mathrm{tr}\left(\mathcal{H}(\theta_\star) \left(\check{\boldsymbol{\Lambda}}^{\mathrm{ss}}_{T'}(\widetilde{\theta}_\star, \boldsymbol{U}, 0) + 5\lambda^\star_{\min,\infty} \cdot I \right)^{-1} \right) \\
&\quad \geq \min_{\boldsymbol{U} \in \mathcal{U}_{\gamma^2,T'}} \mathrm{tr}\left(\mathcal{H}(\theta_\star) \left(\check{\boldsymbol{\Lambda}}^{\mathrm{ss}}_{T'}(\widetilde{\theta}_\star, \boldsymbol{U}, 0) + 10\lambda^\star_{\min,T} \cdot I \right)^{-1} \right) \\
&\quad \geq \min_{\boldsymbol{U} \in \mathcal{U}_{\gamma^2,T'}} \mathrm{tr}\left(\mathcal{H}(\theta_\star) \left(\check{\boldsymbol{\Lambda}}^{\mathrm{ss}}_{T'}(\widetilde{\theta}_\star, \boldsymbol{U}, 0) + \check{\boldsymbol{\Lambda}}^{\mathrm{ss}}_{T'}(\widetilde{\theta}_\star, \boldsymbol{U}'', 0) \right)^{-1} \right) \\
&\quad \geq \min_{\boldsymbol{U} \in \mathcal{U}_{2\gamma^2,T'}} \mathrm{tr}\left(\mathcal{H}(\theta_\star) \left(\check{\boldsymbol{\Lambda}}^{\mathrm{ss}}_{T'}(\widetilde{\theta}_\star, \boldsymbol{U}, 0) \right)^{-1} \right) \\
&\quad \geq \frac{1}{2} \min_{\boldsymbol{U} \in \mathcal{U}_{\gamma^2,T'}} \mathrm{tr}\left(\mathcal{H}(\theta_\star) \left(\check{\boldsymbol{\Lambda}}^{\mathrm{ss}}_{T'}(\widetilde{\theta}_\star, \boldsymbol{U}, 0) \right)^{-1} \right).
\end{aligned}$$

This implies that

$$\begin{aligned}
&\liminf_{T' \to \infty} \frac{1}{8T} \min_{\boldsymbol{U} \in \mathcal{U}_{\gamma^2,T'}} \mathrm{tr}\left(\mathcal{H}(\theta_\star) \left(\check{\boldsymbol{\Lambda}}^{\mathrm{ss}}_{T'}(\widetilde{\theta}_\star, \boldsymbol{U}, 0) + 5\lambda^\star_{\min,\infty} \cdot I \right)^{-1} \right) \\
&\qquad \geq \liminf_{T' \to \infty} \frac{1}{16T} \min_{\boldsymbol{U} \in \mathcal{U}_{\gamma^2,T'}} \mathrm{tr}\left(\mathcal{H}(\theta_\star) \left(\check{\boldsymbol{\Lambda}}^{\mathrm{ss}}_{T'}(\widetilde{\theta}_\star, \boldsymbol{U}, 0) \right)^{-1} \right) = \frac{1}{16T} \Phi^{\mathrm{ss}}_{\mathrm{opt}}(\gamma^2; \theta_\star).
\end{aligned}$$

Putting everything together, Theorem 4.5 and what we have shown imply that as long as T is large enough so that the burn-in of Theorem 4.5 is met, $T \geq H((d_x^2 + d_x)/2 + 1)$ and

(5.3.4), (5.3.7), (5.3.8), (5.3.9), (5.3.10), and (5.3.11) hold, we will have

$$\min_{\widehat{\mathfrak{a}}} \max_{\theta:\|\theta-\theta_0\|_2^2 \le 5(d_x^2+d_x d_u)/(\lambda^\star_{\min,\infty} T^{5/6})} \mathbb{E}[\mathcal{R}(\widehat{\mathfrak{a}};\theta)] \ge \frac{\sigma_w^2}{16T}\Phi^{\mathrm{ss}}_{\mathrm{opt}}(\gamma^2;\theta_\star) - \frac{C_1}{(\lambda^\star_{\min,\infty}T)^{5/4}}$$

where $C_2 = \mathcal{O}\Big((L_{\mathfrak{a}1}L_{\mathfrak{a}2}L_{\mathcal{R}2} + L_{\mathfrak{a}1}^3 L_{\mathcal{R}3} + L_{\mathrm{hess}})(d_x^2+d_x d_u)^{3/2} + L_{\mathfrak{a}1}^2 L_{\mathcal{R}2}\Big)$. Finally we can lower bound $\Phi^{\mathrm{ss}}_{\mathrm{opt}}(\gamma^2;\theta_\star)$ with $\Phi_{\mathrm{opt}}(\gamma^2;\theta_\star)/4$ by Lemma 5.3.3.

Simplifying the Burn-In Time: It remains to simplify the bound. First, note that by Lemma F.9 of Wagenmaker and Jamieson [2020], as long as $\|\widetilde{\theta}-\widetilde{\theta}_\star\|_{\mathrm{op}} \le c/\|\widetilde{\theta}_\star\|_{\mathcal{H}_\infty}$, we will have that $\|\widetilde{\theta}\|_{\mathcal{H}_\infty}$ and $\|\widetilde{\theta}_\star\|_{\mathcal{H}_\infty}$ are within a constant factor of each other. Next, note that Proposition C.1 implies that, so long as $\|\widetilde{\theta}-\widetilde{\theta}_\star\|_{\mathrm{op}} \le \epsilon$, $\|\widetilde{A}^k\|_{\mathrm{op}} \le \tau(\widetilde{A}_\star,\rho)(\rho+\tau(\widetilde{A}_\star,\rho)\epsilon)^k$. This implies that

$$\tau(\widetilde{A},\rho+\tau(\widetilde{A}_\star,\rho)\epsilon) = \sup_k \|\widetilde{A}^k\|_{\mathrm{op}}(\rho+\tau(\widetilde{A}_\star,\rho)\epsilon)^{-k} \le \tau(\widetilde{A}_\star,\rho).$$

As long as $\epsilon < (1-\rho_\star)/(2\tau(\widetilde{A}_\star,\rho))$ we can then choose $\rho = \rho_\star + \tau(\widetilde{A}_\star,\rho_\star)\epsilon$ which will allow us to upper bound

$$\frac{\tau(\widetilde{A},\rho)^n}{(1-\rho^m)^p} \le \frac{c\tau(\widetilde{A}_\star,\rho_\star)^n}{(1-\rho_\star^m)^p}.$$

As we have assumed $\|\theta-\theta_0\|_F, \|\theta_\star-\theta_0\|_F \le \sqrt{5(d_x^2+d_x d_u)}/(\sqrt{\lambda^\star_{\min,\infty}}T^{5/12})$, we can upper bound $\|\widetilde{\theta}-\widetilde{\theta}_\star\|_{\mathrm{op}} \le \|\theta-\theta_\star\|_F \le 2\sqrt{5(d_x^2+d_x d_u)}/(\sqrt{\lambda^\star_{\min,\infty}}T^{5/12})$. Some algebra, Lemma 5.3.2, and the definition of $L_{\mathrm{cov}}(\theta_\star,\gamma^2)$ and $r_{\mathrm{cov}}(\theta_\star)$ then gives that as long as

$$T \ge \mathrm{poly}\left(\frac{1}{1-\rho_\star}, \tau_\star, \|B_\star\|_{\mathrm{op}}, d_x, d_u, \|\widetilde{\theta}_\star\|_{\mathcal{H}_\infty}, \gamma^2, \sigma_w^2, \frac{1}{\lambda^\star_{\min,\infty}}, \log T\right),$$

these bounds on $\|\widetilde{\theta}\|_{\mathcal{H}_\infty}$ and $\tau(\widetilde{A},\rho)$ will hold, $T \ge H((d_x^2+d_x)/2+1)$ and (5.3.4), (5.3.7), (5.3.8), (5.3.9), (5.3.10), and (5.3.11) hold. Finally, we use Lemma 5.3.4 to replace $\lambda^\star_{\min,\infty}$ with λ_{noise}, and Lemma 5.3.1 to upper bound $\tau_\star$, $\frac{1}{1-\rho_\star}$ and $\|\widetilde{\theta}_\star\|_{\mathcal{H}_\infty}$ by $\mathrm{poly}(\|B_\star\|_{\mathrm{op}}, \|A_\star\|_{\mathcal{H}_\infty})$.

5.4 Proof of Theorem 5.2

Towards proving Theorem 5.2, we first specialize our results on certainty-equivalence decision making from Chapter 4 to the LDDM setting. We then prove Theorem 5.2 in Section 5.4.2.

5.4.1 Certainty Equivalence Decision Making in LDDM

While Theorem 4.4 holds in a very general setting, our optimal decision making algorithm, TOPLE, applies only to the LDDM setting, and uses a highly structured set of policies. In

order to facilitate the analysis of TOPLE, it is helpful to obtain a corollary of Theorem 4.4 in this more restricted setting. Towards making this precise, we introduce a set of policies in the LDDM setting, *sequential-open loop* policies, which we show contains TOPLE. Before formally defining these policies, we recall that

$$\beta_\infty := c\Big((1 + \|B_\star\|_{\mathrm{op}}^2)\|A_\star\|_{\mathcal{H}_\infty}^4\Big)\Big(\sqrt{d_x}\log\frac{T}{\delta} + \gamma^2\|A_\star\|_{\mathcal{H}_\infty}^2\Big).$$

By Lemma C.2.2, $\bar{\mathbf{\Lambda}}_T := T\beta_\infty \cdot I$ is a high probability upper bound on the covariates, assuming that $\pi_{\mathrm{exp}} \in \Pi_{\gamma^2}^{\mathrm{sol}}$, where we define $\Pi_{\gamma^2}^{\mathrm{sol}}$, the set of sequential open-loop policies, as:

Definition 5.4.1 (Sequential Open-Loop Policies)**.** We define a *sequential open-loop policy* to be an exploration policy $\pi_{\mathrm{exp}} \in \Pi_{\gamma^2}$ satisfying the following conditions:

- **(Open-Loop Gaussian)** There exist deterministic times $\{\bar{t}_0, \bar{t}_1, \ldots, \bar{t}_{n-1}, \bar{t}_n\} \subseteq [T]$ with $\bar{t}_0 = 0, \bar{t}_n = T, \bar{t}_{i+1} \geq \bar{t}_i$, such that, for $t \in \{\bar{t}_i, \ldots, \bar{t}_{i+1} - 1\}$:

  $$u_t | \mathcal{F}_{\bar{t}_i} \sim \mathcal{N}(\widetilde{u}_t, \Gamma_{u,i})$$

  for $\mathcal{F}_{\bar{t}_i}$ measurable $\widetilde{u}_t$ and $\Gamma_{u,i} \succeq 0$ satisfying:

  $$\sum_{t=0}^{T-1} \widetilde{u}_t^\top \widetilde{u}_t \leq T\gamma^2, \quad \mathrm{tr}(\Gamma_{u,i}) \leq \gamma^2, \quad \lambda_{\min}(\Gamma_{u,i}) \geq \sigma_u^2$$

  almost surely, for deterministic σ_u.

- **(Low-Switching)** For any t, there exists some epoch i such that $|\{t, \ldots, t + T_{\mathrm{se}}(\pi_{\mathrm{exp}})\} \cap \{\bar{t}_i, \ldots, \bar{t}_{i+1} - 1\}| \geq \frac{1}{2}T_{\mathrm{se}}(\pi_{\mathrm{exp}})$ where

  $$T_{\mathrm{se}}(\pi_{\mathrm{exp}}) := c_1 d_x\Big((d_x + d_u)\log(\beta_\infty/\lambda_{\mathrm{noise}}(\sigma_u) + 1) + \log\frac{n}{\delta}\Big).$$

  In words, at least half of any length $T_{\mathrm{se}}(\pi_{\mathrm{exp}})$ interval is contained in a single epoch.

We make several comments on this definition.

- Any policy $\pi_{\mathrm{exp}} \in \Pi_{\gamma^2}^{\mathrm{sol}}$ satisfies Assumption 4.4, which we prove in Section 4.2.2.
- TOPLE, an optimal policy (up to constants), is in $\Pi_{\gamma^2}^{\mathrm{sol}}$, with $n = \mathcal{O}(\log T)$.
- The assumption that $u_t \mid \mathcal{F}_{T_i}$ be Gaussian is for simplicity of analysis, and in general is not necessary—the noise could take different sub-Gaussian distributions if desired.

The following result instantiates Theorem 4.4 with any policy $\pi_{\mathrm{exp}} \in \Pi_{\gamma^2}^{\mathrm{sol}}$, and assuming we are in the LDDM setting.

Corollary 5.4. *Assume we are in the* LDDM *setting and consider some loss $\mathcal{R}$ satisfying Assumption 4.1, stable system $\theta_\star$, and exploration policy $\pi_{\mathrm{exp}} \in \Pi^{\mathrm{sol}}_{\gamma^2}$. If*

$$T \geq c_1 \left(\frac{C_{\mathrm{sys}} n^2 (\sigma_w^4 + \gamma^4)}{\lambda_{\mathrm{noise}}(\sigma_u)^2} + \frac{d_x}{\lambda_{\mathrm{noise}}(\sigma_u) r_{\mathrm{quad}}(\theta_\star)^2} + \frac{\sqrt{d_x}\sigma_w^2}{\gamma^2} + n + d_x \right) \left(\log \tfrac{n}{\delta} + d \log(\beta_\infty / \lambda_{\mathrm{noise}}(\sigma_u) + 3) \right) \tag{5.4.1}$$

then for $\delta \in (0, 1/3)$, with probability $1 - \delta$:

$$\mathcal{R}(\mathfrak{a}_{\mathrm{opt}}(\widehat{\theta}_{\mathrm{ls}}); \theta_\star) \leq 5\sigma_w^2 \log \frac{24(d_x^2 + d_x d_u)}{\delta} \cdot \frac{\Phi_T(\pi_{\mathrm{exp}}; \theta_\star)}{T} + \frac{C_{\mathrm{ce},1}}{T^{3/2}} + \frac{C_{\mathrm{ce},2}}{T^2}$$

where $C_{\mathrm{sys}} = \mathrm{poly}(\|B_\star\|_{\mathrm{op}}, \|A_\star\|_{\mathcal{H}_\infty})$, $d := d_x + d_u$, universal numerical constants c_1, c_2, and

$$C_{\mathrm{ce},1} := \frac{c_2 L_{\mathrm{quad}}}{\lambda_{\mathrm{noise}}(\sigma_u)^{3/2}} \left(\log \tfrac{1}{\delta} + d_x d \log(\beta_\infty / \lambda_{\mathrm{noise}}(\sigma_u) + 3) \right)^{3/2}, \quad C_{\mathrm{ce},2} := \frac{C_{\mathrm{sys}} \sigma_w^2 (\sigma_w^4 + \gamma^4) \mathrm{tr}(\mathcal{H}(\theta_\star)) d^3 n}{\lambda_{\mathrm{noise}}(\sigma_u)^3} \log^2 \tfrac{dn}{\delta}.$$

Proof. This is a direct consequence of Theorem 4.4 and the following result.

Lemma 5.4.1. *Any policy $\pi_{\mathrm{exp}} \in \Pi^{\mathrm{sol}}_{\gamma^2}$ meets Assumption 4.4 on some system $\theta_\star$ with*

$$\begin{aligned}
T_{\mathrm{se}}(\pi_{\mathrm{exp}}) &= c_1 d_x \left((d_x + d_u) \log(\beta_\infty / \lambda_{\min}(\mathbf{\Lambda}^{\mathrm{noise}}_{d_x}(\widetilde{\theta}_\star, \sigma_u)) + 1) + \log \frac{n}{\delta} \right) \\
\underline{\lambda} &= c_2 \lambda_{\min}(\mathbf{\Lambda}^{\mathrm{noise}}_{d_x}(\widetilde{\theta}_\star, \sigma_u)) \\
\bar{\mathbf{\Lambda}}_T &= \beta_\infty \cdot I \\
T_{\mathrm{con}}(\pi_{\mathrm{exp}}) &= \max \left\{ T_{\mathrm{se}}(\pi_{\mathrm{exp}}), (n + \sqrt{d_x}\sigma_w^2/\gamma^2) \log \frac{2(n+1)^2}{\delta} \right\} \\
C_{\mathrm{con}} &= n \frac{c_3 \tau_\star^3 (\sigma_w^2 + \gamma^2)}{(1 - \rho_\star)^{5/2} \lambda_{\min}(\mathbf{\Lambda}^{\mathrm{noise}}_{d_x}(\widetilde{\theta}_\star, \sigma_u))} \sqrt{\log \frac{n}{\delta} + d_x + d_u} \\
\alpha &= 1/2
\end{aligned}$$

for universal constants c_1, c_2, c_3.

The stated results follows from some algebra to simplify terms and using Lemma 5.3.1 to upper bound $\mathrm{poly}(\tau_\star, \frac{1}{1-\rho_\star})$ terms by $\mathrm{poly}(\|B_\star\|_{\mathrm{op}}, \|A_\star\|_{\mathcal{H}_\infty})$, and setting $d_\theta = d_x^2 + d_x d_u$, the dimensionality of (A, B). □

5.4.2 Proof of Theorem 5.2

Fix an epoch i and let $T = \sum_{j=0}^{i} T_j$. Note that, by the definition of T_i, we will have $T_i = \frac{1}{2}(T + T_0)$. Similarly, $T_{i-1} = \frac{1}{4}(T + T_0)$. Define the following events.

$$\mathcal{E}_1 = \left\{ \|\widehat{\theta}_{i-1} - \theta_\star\|_{\mathrm{op}} \le C\sqrt{\frac{\log(1/\delta) + (d_x + d_u)\log(\beta_\infty/\lambda_{\mathrm{noise}} + 1)}{T_{i-1}\lambda_{\mathrm{noise}}}} =: \epsilon_{\mathrm{op},i-1} \right\}$$

$$\mathcal{E}_2 = \left\{ \mathcal{R}(\mathfrak{a}_{\mathrm{opt}}(\widehat{\theta}_i); \theta_\star) \le 5\sigma_w^2 \left(\mathcal{H}(\theta_\star)\mathbb{E}_{\theta_\star}[\mathbf{\Sigma}_T]^{-1}\right) \log \frac{24(d_x^2 + d_x d_u)}{\delta} + \frac{C_1}{T^{3/2}} + \frac{C_2}{T^2} \right\}$$

$$\mathcal{E}_3 = \left\{ \|z_{T-T_i}\|_2^2 \le \frac{4\tau_\star^2\gamma^2 k_i^2}{1-\rho_\star^{k_i}} + 8\sqrt{\|\widetilde{\mathbf{\Lambda}}_T^{\mathrm{noise}}\|_F^2 \log\frac{2}{\delta}} + 8\|\widetilde{\mathbf{\Lambda}}_T^{\mathrm{noise}}\|_{\mathrm{op}} \log\frac{2}{\delta} \right\}$$

for C_1, C_2 as defined in Corollary 5.4.

Events $\mathcal{E}_i$ hold. By Lemma 3.5.1, we know that TOPLE $\in \Pi_{\gamma^2}^{\mathrm{sol}}$. By Lemma 5.4.1, this implies that TOPLE satisfies Assumption 4.4 with

$$T_{\mathrm{se}}(\textsc{Topple}) = c_1 d_x\Big((d_x + d_u)\log(\beta_\infty/\lambda_{\mathrm{noise}} + 1) + \log\frac{n}{\delta}\Big), \quad \underline{\lambda} = c_2\lambda_{\mathrm{noise}}, \quad \bar{\mathbf{\Lambda}}_T = \beta_\infty \cdot I$$

Thus, as long as

$$T \ge T_{\mathrm{se}}(\textsc{Topple}) \tag{5.4.2}$$

we will have that with probability at least $1-\delta$, $\lambda_{\min}(\Sigma_T) \ge c_2\lambda_{\mathrm{noise}} T$ and $\Sigma_T \preceq T\beta_\infty I$. We can therefore apply Lemma C.3.2, our operator norm estimation bound, to get that $\mathbb{P}[\mathcal{E}_1^c] \le \delta$. Furthermore, by Corollary 5.4, we will have, as long as T is large enough for the burn-in, (5.4.1) to be met, that $\mathbb{P}[\mathcal{E}_2^c] \le \delta$. To show that $\mathcal{E}_3$ occurs with high probability, we break up the state into two components: $z_t^{\boldsymbol{u}}$, the portion of the state driven by $\widetilde{u}_t$, and $z_t^{\boldsymbol{w}}$, the portion of the state driven by the input noise and process noise. As the structure of TOPLE is identical to that of the algorithm in Chapter 3, Lemma A.1.2 gives that

$$\|z_{T-T_i}^{\boldsymbol{u}}\|_2^2 \le \frac{4\tau(\widetilde{A}_\star, \rho)^2 k_i^2 \gamma^2}{(1-\rho^{k_i})^2}$$

and we choose $\rho = \rho_\star$. Crucially for subsequent steps, this scales as k_i^2 instead of T, which is the scaling we would obtain applying Lemma C.2.7 would scale. Next, applying Lemma C.2.8 gives that, with probability $1-\delta$,

$$\|z_{T-T_i}^{\boldsymbol{w}}\|_2^2 \le 4\sqrt{\|\widetilde{\mathbf{\Lambda}}_T^{\mathrm{noise}}\|_F^2 \log\frac{2}{\delta}} + 4\|\widetilde{\mathbf{\Lambda}}_T^{\mathrm{noise}}\|_{\mathrm{op}} \log\frac{2}{\delta}.$$

Note that we can apply Lemma C.2.8 since the input noise variance is deterministically fixed for all epochs, and by upper bounding the state bound for epochs $i \geq 1$ by the state bound that would hold if we always set the input noise to have variance γ^2/d_u. This implies $\mathbb{P}[\mathcal{E}_3^c] \leq \delta$. Altogether then, we have that $\mathbb{P}[\mathcal{E}_1 \cap \mathcal{E}_2 \cap \mathcal{E}_3] \geq 1 - 3\delta$.

Events $\mathcal{E}_i$ imply optimal inputs. We now assume that $\mathcal{E}_1 \cap \mathcal{E}_2 \cap \mathcal{E}_3$ holds. The following result shows that, for sufficiently large T_i and $\widehat{\theta}_{i-1}$ sufficiently accurate, the exploration inputs played by Algorithm 5.1 are near-optimal.

Lemma 5.4.2. *Fix an epoch i of Algorithm 5.1 and let k_i denote the discretization level of the input set at that epoch, $\mathcal{U}_{\gamma^2/2,k_i}$, and assume that $T_i/(d_u k_i)$ is an integer. Let k be any value satisfying $d_x \leq k \leq T_i/2$ and take $\epsilon \in (0,1)$. Then, as long as*

$$T_i \geq \max\left\{ \frac{2d_u\tau_\star^3(2k_i\gamma + \|z_{T-T_i}\|_2)((4+2\tau_\star)k_i\gamma + \tau_\star\|z_{T-T_i}\|_2))}{(1-\rho_\star^{k_i})^2(1-\rho_\star)\lambda_{\text{noise}}\epsilon}, \frac{16\tau_\star^2(\sigma_w^2 + \gamma^2/(2d_u))}{(1-\rho_\star^2)^2\lambda_{\text{noise}}\epsilon} \right.$$
$$\left. \left(\frac{\tau_\star^2 k_i^2 d_u\gamma^2}{(1-\rho_\star^{k_i})^2} + \frac{\tau_\star k_i\sqrt{d_u}\gamma^2\sqrt{T_i}}{1-\rho_\star^{k_i}}\right)\frac{16d_u\|\widetilde{\theta}_\star\|_{\mathcal{H}_\infty}^2}{\lambda_{\text{noise}}\epsilon}, \log\left(\frac{(1-\rho_\star^2)^2\lambda_{\text{noise}}\epsilon}{16\tau_\star^2(\sigma_w^2+\gamma^2/(2d_u))}\right)\frac{1}{2\log\rho_\star}\right\},$$

$$k_i \geq \max\left\{\frac{8\pi\|\widetilde{\theta}_\star\|_{\mathcal{H}_\infty}\gamma^2}{\lambda_{\text{noise}}\epsilon}, \frac{\pi}{2\|\widetilde{\theta}_\star\|_{\mathcal{H}_\infty}}\right\}\left(\max_{\omega\in[0,2\pi]}\|(e^{\iota\omega}I - \widetilde{A}_\star)^{-2}\widetilde{B}\|_{\text{op}}\right),$$

$$\|\widehat{\theta}_{i-1} - \theta_\star\|_{\text{op}} \leq \epsilon_{\text{op},i-1} \leq \min\left\{r_{\text{cov}}(\theta_\star), \frac{r_{\text{quad}}(\theta_\star)}{\sqrt{d_x}}, \frac{\lambda_{\text{noise}}}{4L_{\text{cov}}(\theta_\star,\gamma^2)}\right\},$$

we have, for any $T' \geq T_i$:

$$\text{tr}\Big(\mathcal{H}(\theta_\star)(\mathbb{E}_{\theta_\star,\boldsymbol{U}_i,T_i}[I_{d_x}\otimes\textstyle\sum_{t=T-T_i}^{T} z_t z_t^\top])^{-1}\Big) \leq \min_{\boldsymbol{U}\in\mathcal{U}_{\gamma^2,T'}} \frac{2\text{tr}\left(\mathcal{H}(\theta_\star)\check{\boldsymbol{\Lambda}}^{\text{ss}}_{T',T'}(\widetilde{\theta}_\star,\boldsymbol{U},0)^{-1}\right)}{(1-\epsilon)^3 T_i} + \frac{C_{\text{exp}}}{(1-\epsilon)^2}$$

where

$$C_{\text{exp}} = \left(\frac{4(d_x^{5/2}+d_x^{3/2}d_u)L_{\text{hess}}}{\lambda_{\text{noise}}} + \frac{8L_{\text{cov}}(\theta_\star,\gamma^2)\text{tr}(\mathcal{H}(\theta_\star))}{\lambda_{\text{noise}}^2}\right)\frac{\epsilon_{\text{op},i-1}}{T_i} + \frac{8(d_x^2+d_x d_u)L_{\text{cov}}(\theta_\star,\gamma^2)L_{\text{hess}}}{\lambda_{\text{noise}}^2}\frac{\epsilon_{\text{op},i-1}^2}{T_i}.$$

Assume that T_i is large enough that

$$\epsilon_{\text{op},i-1} \leq \min\{r_{\text{cov}}(\theta_\star), r_{\text{quad}}(\theta_\star)/\sqrt{d_x}, \lambda_{\text{noise}}/(4L_{\text{cov}}(\theta_\star,\gamma^2))\}.$$

Then, as long as,

$$T_i \geq \text{poly}\left(\frac{1}{1-\rho_\star}, \tau_\star, \|\widetilde{\theta}_\star\|_{\mathcal{H}_\infty}\right)\frac{d_u\|z_{T-T_i}\|_2^2 + d_u^{3/2}\gamma^2 k_i\sqrt{T_i} + d_u^2\gamma^2k_i^2 + \sigma_w^2}{\lambda_{\text{noise}}}, \tag{5.4.3}$$

$$k_i \geq \max\left\{\frac{80\pi\|\widetilde{\theta}_\star\|_{\mathcal{H}_\infty}\gamma^2}{\lambda_{\text{noise}}}, \frac{\pi}{2\|\widetilde{\theta}_\star\|_{\mathcal{H}_\infty}}\right\}\left(\max_{\omega\in[0,2\pi]}\|(e^{\iota\omega}I - \widetilde{A}_\star)^{-2}\widetilde{B}\|_{\text{op}}\right), \tag{5.4.4}$$

we can apply Lemma 5.4.2, which gives that the performance achieved by $\boldsymbol{U}_i$ is nearly optimal. That is, for any $T' \geq T_i$,

$$\begin{aligned}\operatorname{tr}(\mathcal{H}(\theta_\star)(\mathbb{E}_{\theta_\star}[\boldsymbol{\Sigma}_T])^{-1}) &\leq \operatorname{tr}(\mathcal{H}(\theta_\star)(\mathbb{E}_{\theta_\star}[I_{d_x}\otimes\textstyle\sum_{t=T-T_i}^{T} z_t z_t^\top])^{-1}) \\ &\leq \min_{\boldsymbol{u}\in\mathcal{U}_{\gamma^2,T'}} \frac{3\operatorname{tr}\left(\mathcal{H}(\theta_\star)\check{\boldsymbol{\Lambda}}^{\text{ss}}_{T',T'}(\widetilde{\theta}_\star,\boldsymbol{u},0)^{-1}\right)}{T_i} + 2C_{\text{exp}}\end{aligned}$$

for C_{exp} as above and we have chosen ϵ such that $1/(1-\epsilon)^3 = 3/2$. Recall that Algorithm 5.1 uses $T_i = C_{\text{init}}d_u 2^i$ and $k_i = C_{\text{init}}2^{\lfloor i/4\rfloor}$. We then have that:

$$\frac{C_{\text{init}}^{3/4}}{d_u^{1/4}}T_i^{1/4} \geq C_{\text{init}}2^{i/4} \geq k_i \geq C_{\text{init}}2^{i/4-1} = \frac{C_{\text{init}}^{3/4}}{2d_u^{1/4}}T_i^{1/4}.$$

On event $\mathcal{E}_3$, which upper bounds $\|z_{T-T_i}\|_2^2$, it follows that (5.4.3) and (5.4.4) hold as long as

$$\begin{aligned}T_i &\geq \frac{\text{poly}\left(\frac{1}{1-\rho_\star}, \tau_\star, \|\widetilde{\theta}_\star\|_{\mathcal{H}_\infty}\right)}{\lambda_{\text{noise}}}\Bigg(d_u^{5/4}\gamma^2 C_{\text{init}}^{3/4}T_i^{3/4} + d_u^{3/2}\gamma^2 C_{\text{init}}^{3/2}\sqrt{T_i} \\ &\qquad + \sigma_w^2 + \sqrt{\|\boldsymbol{\Lambda}^{\text{noise}}_{T-T_i}\|_F^2\log\frac{2}{\delta}} + \|\boldsymbol{\Lambda}^{\text{noise}}_{T-T_i}\|_{\text{op}}\log\frac{2}{\delta}\Bigg) \\ T_i^{1/4} &\geq \max\left\{\frac{80\pi\|\widetilde{\theta}_\star\|_{\mathcal{H}_\infty}\gamma^2}{\lambda_{\text{noise}}}, \frac{\pi}{2\|\widetilde{\theta}_\star\|_{\mathcal{H}_\infty}}\right\}\frac{2d_u^{1/4}\left(\max_{\omega\in[0,2\pi]}\|(e^{\iota\omega}I - \widetilde{A}_\star)^{-2}\widetilde{B}\|_{\text{op}}\right)}{C_{\text{init}}^{3/4}}.\end{aligned}$$

We have then shown that, on the event $\mathcal{E}_1\cap\mathcal{E}_2\cap\mathcal{E}_3$ and assuming T is large enough to meet the burn-ins stated above, we have, for any T',

$$\begin{aligned}\mathcal{R}(\mathfrak{a}_{\text{opt}}(\widehat{\theta}_i);\theta_\star) &\leq 5\sigma_w^2\left(\mathcal{H}(\theta_\star)\mathbb{E}_{\theta_\star}[\boldsymbol{\Sigma}_T]^{-1}\right)\log\frac{18(d_x^2+d_xd_u)}{\delta} + \frac{C_1}{T^{3/2}} + \frac{C_2}{T^2} \\ &\leq 15\sigma_w^2\min_{\boldsymbol{u}\in\mathcal{U}_{\gamma^2,T'}}\frac{\operatorname{tr}\left(\mathcal{H}(\theta_\star)\check{\boldsymbol{\Lambda}}^{\text{ss}}_{T',T'}(\widetilde{\theta}_\star,\boldsymbol{u},0)^{-1}\right)}{T_i}\log\frac{24(d_x^2+d_xd_u)}{\delta} + \frac{C_1}{T^{3/2}} + \frac{C_2}{T^2} \\ &\qquad + 10\sigma_w^2 C_{\text{exp}}\log\frac{24(d_x^2+d_xd_u)}{\delta}.\end{aligned}$$

As this bound hold for any $T' \geq T_i$, we take $\liminf_{T'\to\infty}$, to obtain

$$\mathcal{R}(\mathfrak{a}_{\text{opt}}(\widehat{\theta}_i);\theta_\star) \leq \frac{15\sigma_w^2\Phi^{\text{ss}}_{\text{opt}}(\gamma^2;\theta_\star)}{T_i}\log\frac{24(d_x^2+d_xd_u)}{\delta} + \frac{C_1}{T^{3/2}} + \frac{C_2}{T^2} + 10\sigma_w^2 C_{\text{exp}}\log\frac{24(d_x^2+d_xd_u)}{\delta}$$

We can also upper bound $\Phi^{\text{ss}}_{\text{opt}}(\gamma^2;\theta_\star)$ by $16\Phi_{\text{opt}}(\gamma^2;\theta_\star)$ via Lemma 5.3.3. On $\mathcal{E}_1 \cap \mathcal{E}_2 \cap \mathcal{E}_4$, it is easy to see $C_{\text{exp}} = C_4/T^{3/2} + C_5/T^2$ for some C_4, C_5. The conclusion then follows by rescaling δ, and since $T_i \geq T/2$. The fact that the average expected power of the inputs is bounded by γ^2 follows by Lemma 3.5.3. Finally, some algebra shows that the burn-in times stated above are all met as long as Assumption 5.3 holds.

Chapter 6

Task-Dependent Exploration in Nonlinear Dynamical Systems

The focus of this chapter is on decision making in nonlinear dynamical systems. In particular, we consider again the decision making setting of Chapter 4 but, rather than instantiating it in the setting of linear dynamical systems as in Chapter 5, we consider a certain class of nonlinear dynamical systems parameterized as:

$$x_{h+1} = A_\star \boldsymbol{\phi}(x_h, u_h) + w_h. \tag{6.0.1}$$

Here $x_h \in \mathbb{R}^{d_x}$ denotes the state of the system, $u_h \in \mathcal{U} \subseteq \mathbb{R}^{d_u}$ the input, $w_h \sim \mathcal{N}(0, \sigma_w^2 \cdot I)$ random noise, $\boldsymbol{\phi}(\cdot, \cdot) \in \mathbb{R}^{d_{\boldsymbol{\phi}}}$ a (possibly nonlinear, known) feature map, and $A_\star \in \mathbb{R}^{d_x \times d_{\boldsymbol{\phi}}}$ the (unknown) system parameter. As in Chapter 5, our goal in this setting will be to explore in a task-dependent manner—to play inputs in (6.0.1) that allow us to learn the optimal decision as quickly as possible.

While we restrict our dynamics to a particular parametric form (6.0.1), systems of this form are able to model a variety of real-world settings that linear systems are not capable of modeling, from drones to robotic systems, and more [Shi et al., 2021a, O'Connell et al., 2022, Boffi et al., 2021, Song and Sun, 2021, Richards et al., 2021]. Indeed, we can model any sufficiently regular, smooth nonlinear system $f(x, u)$ in the form (6.0.1) by choosing $\boldsymbol{\phi}$ to correspond to sufficiently expressive random features [Rahimi and Recht, 2008]. Critically, however, despite its expressivity, we obtain linear observations of the unknown parameter in (6.0.1), $A_\star$, which allows us to apply our results from the MDM setting of Chapter 4 to quantify the statistical complexity of learning in (6.0.1).

To motivate the need for effective exploration in nonlinear systems, and highlight key differences and challenges between linear and nonlinear systems, consider an expanded version of the example from Section 1.1.2 with dynamics given by:

$$x_{h+1} = a_1 x_h + a_2 u_h + \sum_{i=1}^{10} a_{i+2} \boldsymbol{\phi}_i(x_h) + w_h$$

where $\boldsymbol{\phi}_i(x) = \max\{1 - 100(x - c_i)^2, 0\}$ for some c_i. We choose $a_1 = 0.8, a_2 = 1$, and $a_3 = \ldots = a_{12} = -3$. We assume $a_{1:12}$ are unknown, $(\boldsymbol{\phi}_i)_{i=1}^{10}$ is known, and set

$$\text{cost}(x, u) = (x - c_1)^2 + 100^{-1} \cdot u^2.$$

Our goal is to learn a controller which minimizes the cost on the true system. With this choice of cost, the optimal controller will attempt to direct the state x to the equilibrium point c_1 and maintain this position. Note that, with our choice of $\boldsymbol{\phi}_i$, $\boldsymbol{\phi}_i(x) \neq 0$ only when x is very close to c_i. This renders the parameters $a_{4:12}$ irrelevant to learning the optimal controller, since $\boldsymbol{\phi}_2, \ldots, \boldsymbol{\phi}_{10}$ will be inactive if we are playing optimally, but learning a_3 is critical to performing optimally, as its value significantly changes the dynamics at the goal state.

We illustrate the result of running on this system in Figure 6.1, comparing our proposed task-dependent exploration approach (`Task-Driven Exploration`, Algorithm 6.1) to the approach which chooses $u_h \sim \mathcal{N}(0, \sigma_u^2)$ (`Random Exploration`), and the approach proposed in Mania et al. [2022] (`Uniform Exploration`) which seeks to explore so as to estimate $a_{1:12}$ uniformly well. As can be seen, neither of these latter two approaches are able to learn a good controller, while our approach easily finds a near-optimal controller. The failure modes of each of these approaches is somewhat different. Here `Random Exploration` fails since the chance of reaching the point $x_h \sim c_1$ is extremely small if the input is random noise—reaching c_1 requires playing a particular sequence of actions which are very unlikely to be played if u_h is chosen randomly. The `Uniform Exploration` approach does, in contrast, plan and, given enough time, is guaranteed to estimate all parameters accurately. However, as it aims to estimate all parameters uniformly well, it will attempt to estimate $a_{4:12}$ accurately despite their irrelevance to control, which will slow down the rate at which it is able to estimate a_3. Only our approach, which both plans and takes into account the cost while exploring, is able to reach a_3 enough times to efficiently estimate it, and learn a good controller.

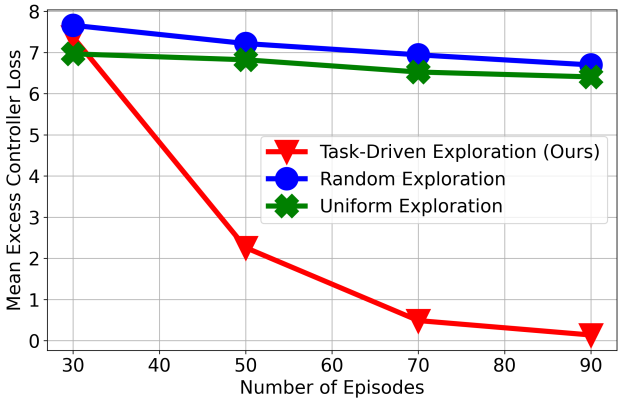

Figure 6.1: Performance on Motivating Example

This example illustrates that it is critical both to explore efficiently, and also to let the objective—learning a good controller—guide this exploration. We emphasize that the behavior in this example is only exhibited in nonlinear systems—though taking into account the task while exploring in linear systems does yield provable improvements, as shown in Chapter 5, even playing random noise allows every direction to be learned in such systems. In nonlinear systems, however, this is not the case—one may fail to learn completely unless

careful planning is performed.

6.1 Preliminaries for Nonlinear Dynamical Systems

In contrast with the previous chapters, in this chapter we consider the *episodic* setting, where the learner interacts with the environment for H steps, at which point the system resets. We assume each episode starts from a fixed start state x_1. We also assume $\|A_\star\|_{\mathrm{op}} \leq B_A$ for some known $B_A > 0$.

We consider a particular form of the decision making setting of Chapter 4 where the goal of the learner is to find a policy (controller) $\pi = (\pi_h)_{h=1}^H$ which achieves minimal cost on (6.0.1), for the cost defined by some (known) function $(\mathrm{cost}_h(\cdot,\cdot))_{h=1}^H$, with $\mathrm{cost}_h : \mathbb{R}^{d_x} \times \mathcal{U} \to \mathbb{R}_+$. For a given policy π, we then define the loss on the system with parameter A as

$$\mathcal{J}_A(\pi) := \mathbb{E}_{A,\pi}\left[\sum_{h=1}^{H} \mathrm{cost}_h(x_h, u_h)\right].$$

We consider an interaction protocol essentially identical to that considered in Chapters 4 and 5:

1. Learner interacts with system (6.0.1) for T episodes, at every episode playing an exploration policy $\pi_{\mathrm{exp}} \in \Pi_{\mathrm{exp}}$.

2. After T episodes, the learner proposes a policy $\widehat{\pi}_T \in \Pi^\star$.

3. The learner suffers cost $\mathcal{J}_{A_\star}(\widehat{\pi}_T)$.

The goal of the learner is therefore first to explore and, after T episodes of exploration, to propose its best guess at the optimal controller for (6.0.1), $\widehat{\pi}_T$. Here we take Π_{exp} to be a (known) set of admissible exploration policies (for example, we could have $\Pi_{\mathrm{exp}} = \Pi_{\gamma^2}$, policies with bounded input power, but allow for more general policy classes), and $\Pi^\star$ a (known) set of admissible control policies. We assume that $\Pi^\star$ has the following parametric form:

$$\Pi^\star = \{\pi^{\mathfrak{a}} \ : \ \mathfrak{a} \in \mathbb{R}^{d_{\mathfrak{a}}}\},$$

where $\mathfrak{a}$ denotes the parameters of the policy. Given this, we overload notation and denote $\mathcal{J}_A(\mathfrak{a}) := \mathcal{J}_A(\pi^{\mathfrak{a}})$, and in general use $\mathfrak{a}$ and $\pi^{\mathfrak{a}}$ interchangeably. We assume that policies in $\Pi^\star$ are deterministic—that is, they are deterministic mappings from histories to inputs—but allow for randomized policies in Π_{exp}. Policies may be either open- or closed-loop. Note that we do not assume $\Pi^\star = \Pi_{\mathrm{exp}}$—in general Π_{exp} need not be equal to $\Pi^\star$.

Connection Between (6.0.1) and MDM Setting We next describe how (6.0.1) can be mapped to the MDM setting of Chapter 4. Assume that we have run (6.0.1) for T episodes, and collected observations $\{(x_1^t, u_1^t, x_2^t, \dots, x_h^t, u_h^t, x_{h+1}^t)\}_{t=1}^T$. Now let $\theta_\star := \mathrm{vec}(A_\star)$. Furthermore, for any t, h, and $i \in [d_x]$, let $\mathtt{t} = (t, h, i)$ and $z_\mathtt{t} = [\mathbf{0}_{d_{\boldsymbol{\phi}}(i-1)}, \boldsymbol{\phi}(x_h^t, u_h^t), \mathbf{0}_{d_{\boldsymbol{\phi}}(d_x - i)}] \in \mathbb{R}^{d_x d_{\boldsymbol{\phi}}}$ where $\mathbf{0}_d$ denotes the zero vector of length d. Then we see that

$$[x_{h+1}^t]_i = \langle \theta_\star, z_\mathtt{t} \rangle + [w_h^t]_i.$$

Setting $y_\mathtt{t} = [x_{h+1}^t]_i$ and $w_\mathtt{t} = [w_h^t]_i$, it is clear that this follows the observation model of the MDM setting with $d_{\boldsymbol{\theta}} = d_{\boldsymbol{\phi}} d_x$ and $\mathtt{T} = d_x T H$. It is also straightforward to see that the measurability assumptions of the setting of MDM are satisfied by this.

System Notation. Before proceeding, we introduce several additional pieces of notation. First, we let $\mathcal{T}$ denote the space of all possible state-input trajectories, $\mathcal{T} \subseteq (\mathbb{R}^{d_x} \times \mathcal{U})^H \times \mathbb{R}^{d_x}$, and, for any $\boldsymbol{\tau} \in \mathcal{T}$, let $\boldsymbol{\tau}_{1:h}$ denote the first h states and inputs in $\boldsymbol{\tau}$. Second, for any policy π, we denote

$$\boldsymbol{\Lambda}_{A,\pi} := \mathbb{E}_{A,\pi}\left[\sum_{h=1}^H \boldsymbol{\phi}(x_h, u_h)\boldsymbol{\phi}(x_h, u_h)^\top\right]$$

the expected covariance induced by playing π on system A. In particular, we set $\boldsymbol{\Lambda}_\pi := \boldsymbol{\Lambda}_{A_\star,\pi}$. We also denote $\check{\boldsymbol{\Lambda}} := I_{d_x} \otimes \boldsymbol{\Lambda}$ the Kronecker product of I_{d_x} and $\boldsymbol{\Lambda}$. Finally, we let $\boldsymbol{\Omega}$ denote the set of all possible covariance matrices induced by playing mixtures of policies in Π_{exp}:

$$\boldsymbol{\Omega} := \left\{\mathbb{E}_{\pi\sim\omega}[\boldsymbol{\Lambda}_\pi] \ : \ \omega \in \triangle_{\Pi_{\mathrm{exp}}}\right\},$$

where $\triangle_{\Pi_{\mathrm{exp}}}$ denotes the set of distributions over Π_{exp}.

6.1.1 Regularity Assumptions

In order to make learning in (6.0.1) tractable, we need several regularity assumptions. We first introduce assumptions on the boundedness of the feature map $\boldsymbol{\phi}$, the boundedness of the cost, and the achievable minimum eigenvalue.

Assumption 6.1 (Bounded Features)**.** *For all $x \in \mathbb{R}^{d_x}$ and $u \in \mathcal{U}$, we have $\|\boldsymbol{\phi}(x, u)\|_2 \le B_{\boldsymbol{\phi}}$.*

Assumption 6.2 (Bounded Cost)**.** *There exists some $r_{\mathrm{cost}}(A_\star) > 0$ such that, for all $A \in \mathcal{B}_{\mathrm{F}}(A_\star; r_{\mathrm{cost}}(A_\star))$ and all $\pi \in \Pi^\star$, we have $\mathbb{E}_{A,\pi}[(\sum_{h=1}^H \mathrm{cost}_h(x_h, u_h))^2] \le L_{\mathrm{cost}}$.*

Assumption 6.3 (Uniform Feature Excitation)**.** *There exists $\omega \in \triangle_{\Pi_{\mathrm{exp}}}$ such that*

$$\lambda_{\min}(\mathbb{E}_{\pi_{\mathrm{exp}}\sim\omega}[\boldsymbol{\Lambda}_{\pi_{\mathrm{exp}}}]) \ge \bar{\lambda}_{\min}$$

for some $\bar{\lambda}_{\min} > 0$.

We remark that these assumptions have appeared before in work on systems of the form (6.0.1) [Mania et al., 2022, Kakade et al., 2020]. In order to precisely quantify the optimal rates of learning, we require that our system satisfy certain smoothness assumptions. First, we require that $\boldsymbol{\phi}(\cdot, \cdot)$ is differentiable in its second argument.

Assumption 6.4 (Smooth Nonlinearity)**.** *For all $x \in \mathbb{R}^{d_x}$ and $u \in \mathcal{U}$, $\boldsymbol{\phi}(x, u)$ is four-times differentiable in u. Furthermore, $\|\nabla_u^{(i)} \boldsymbol{\phi}(x, u)\|_{\mathrm{op}} \leq L_{\boldsymbol{\phi}}$, $\forall i \in \{1, 2, 3, 4\}$, $x \in \mathbb{R}^{d_x}$, and $u \in \mathcal{U}$.*

We assume that the parameterization of our policies, $\pi^{\mathfrak{a}}$, is smooth in the following sense:

Assumption 6.5 (Smooth Controller Class)**.** *$\pi_h^{\mathfrak{a}}(\boldsymbol{\tau}_{1:h})$ is four-times differentiable in $\mathfrak{a}$ for all $\boldsymbol{\tau} \in \mathcal{T}$ and $h \in [H]$. Furthermore, $\|\nabla_{\mathfrak{a}}^{(i)} \pi_h^{\mathfrak{a}}(\boldsymbol{\tau}_{1:h})\|_{\mathrm{op}} \leq L_{\mathfrak{a}}$ for $\forall i \in \{1, 2, 3, 4\}$, $\mathfrak{a} \in \mathbb{R}^{d_{\mathfrak{a}}}$, and $\boldsymbol{\tau} \in \mathcal{T}$.*

Assumption 6.5 is satisfied for commonly considered classes of controllers, such as linear controllers, but is also satisfied by more complex classes such as neural network controllers. While the learner may propose any $\widehat{\pi}_T \in \Pi^\star$, we are particularly interested in the certainty equivalence decision rule, where the controller parameters are chosen as:

$$\mathfrak{a}_{\mathrm{opt}}(\widehat{A}) := \operatorname*{arg\,min}_{\mathfrak{a} \in \mathbb{R}^{d_{\mathfrak{a}}}} \mathcal{J}_{\widehat{A}}(\mathfrak{a}). \tag{6.1.1}$$

To ensure that $\mathfrak{a}_{\mathrm{opt}}(A)$ is well-defined and sufficiently regular, we make the following assumption.

Assumption 6.6 (Unique Optimal Controller)**.** *We assume that the global minimum of $\mathcal{J}_{A_\star}(\mathfrak{a})$, $\mathfrak{a}_{\mathrm{opt}}(A_\star)$, is unique, and that $\nabla_{\mathfrak{a}}^2 \mathcal{J}_{A_\star}(\mathfrak{a})|_{\mathfrak{a}=\mathfrak{a}_{\mathrm{opt}}(A_\star)} \succ 0$.*

In settings where the globally optimal decision cannot be efficiently computed, (6.1.1) can be replaced with a locally optimal decision, and Assumption 6.6 can be replaced by assuming the differentiability of $\mathfrak{a}_{\mathrm{opt}}(A)$ with respect to A for A near $A_\star$. For ease of exposition, in this thesis we assume that Assumption 6.6 holds and that $\mathfrak{a}_{\mathrm{opt}}(A)$ is defined as in (6.1.1) (see Wagenmaker et al. [2024] for the extension). With these definitions and under Assumptions 6.1, 6.2, 6.4 and 6.5, we can show that $\mathcal{J}_{A_\star}(\mathfrak{a})$ is differentiable in $\mathfrak{a}$ and, combined with Assumption 6.6, that $\mathfrak{a}_{\mathrm{opt}}(A)$ is differentiable in A, for $A \in \mathcal{B}_{\mathrm{F}}(A_\star; r_{\mathfrak{a}}(A_\star))$ and some $r_{\mathfrak{a}}(A_\star) > 0$. We let $L_{\pi_\star}$ denote an upper bound on the norm of the derivatives of $\mathfrak{a}_{\mathrm{opt}}(A)$. We always take $B_{\boldsymbol{\phi}}, L_{\mathrm{cost}}, L_{\boldsymbol{\phi}}, L_{\mathfrak{a}}, L_{\pi_\star} \geq 1$.

The above assumptions allow us to show that a (slightly relaxed) version of Assumption 4.1 holds in the setting of (6.0.1). As we will see, given this, and the above argument that (6.0.1) maps to the observation model of the MDM setting, the complexity of learning in nonlinear systems scales similarly to the complexities given in Chapters 4 and 5.

6.2 Optimal Decision Making in Nonlinear Systems

Assume that we have collected observations $\{\{(x_h^t, u_h^t)\}_{h=1}^H\}_{t=1}^T$ and define the covariates

$$\mathbf{\Lambda}_T = \sum_{t=1}^{T}\sum_{h=1}^{H} \boldsymbol{\phi}(x_h^t, u_h^t)\boldsymbol{\phi}(x_h^t, u_h^t)^\top.$$

Let $\widehat{A}$ denote the least-squares estimate of our unknown parameter:

$$\widehat{A} = \arg\min_A \sum_{t=1}^{T}\sum_{h=1}^{H} \|x_{h+1}^t - A\boldsymbol{\phi}(x_h^t, u_h^t)\|_2^2.$$

By the results of Chapter 4, we can immediately bound the suboptimality of $\pi_\star(\widehat{A})$ as:

$$\mathcal{J}_{A_\star}(\mathfrak{a}_{\mathrm{opt}}(\widehat{A})) - \mathcal{J}_{A_\star}(\mathfrak{a}_{\mathrm{opt}}(A_\star)) \lesssim \sigma_w^2 \cdot \mathrm{tr}(\mathcal{H}(A_\star)\cdot \check{\mathbf{\Lambda}}_T^{-1}),$$

where $\mathcal{H}(A_\star) := \nabla_A^2 \mathcal{J}_{A_\star}(\mathfrak{a}_{\mathrm{opt}}(A))|_{A=A_\star}$. Just as in Chapter 5, this bound directly motivates our algorithmic approach: explore to collect covariates $\mathbf{\Lambda}_T$ minimizing $\mathrm{tr}(\mathcal{H}(A_\star)\check{\mathbf{\Lambda}}_T^{-1})$. We present our main algorithm in Algorithm 6.1.

Algorithm 6.1 Optimal Exploration in Nonlinear Systems

inputs: number of episodes to run T, cost function $(\mathrm{cost}_h)_{h=1}^H$, confidence δ, control policies $\Pi^\star$, exploration policies Π_{exp}
$\widehat{A}^1 \leftarrow$ anything, $\ell_T \leftarrow \lceil \log_2 T/8 \rceil$
for $\ell = 1, 2, 3, \ldots, \ell_T$ **do**
  $\widehat{A}^\ell \leftarrow \arg\min_A \sum_{h=1}^H \sum_{(x_{h+1},u_h,x_h)\in\mathfrak{D}_{\ell-1}} \|x_{h+1} - A\boldsymbol{\phi}(x_h, u_h)\|_2^2$
  $T_\ell \leftarrow 2^\ell, \delta_\ell \leftarrow \delta/12\ell^2$
  $\Pi_\ell \leftarrow$ LearnExpΠ$(\mathcal{H}(\widehat{A}^\ell), T_\ell, \delta_\ell, \Pi_{\mathrm{exp}})$ (Algorithm D.2)
  Rerun each policy in Π_ℓ $N_\ell = \lceil T_\ell/|\Pi_\ell| \rceil$ times, denote collected data $\mathfrak{D}_\ell$
return $\widehat{\mathfrak{a}}_T \leftarrow \mathfrak{a}_{\mathrm{opt}}(\widehat{A}^{\ell_T+1})$

Algorithm 6.1 proceeds in epochs of exponentially increasing length. At each epoch it first approximates $\mathcal{H}(A_\star)$ by computing the model-task Hessian of the estimated system, $\widehat{A}^\ell$. Using this approximation of $\mathcal{H}(A_\star)$, it seeks to explore to minimize $\mathrm{tr}(\mathcal{H}(\widehat{A}^\ell)\check{\mathbf{\Lambda}}_T^{-1})$. This exploration routine is encapsulated in the LearnExpΠ function, an adaptive experiment-design routine described in more detail in Section 6.3. LearnExpΠ returns a set of exploration policies, Π_ℓ, which we run to collect data $\mathfrak{D}_\ell$. As we will show, the collected covariates, $\mathbf{\Lambda}_\ell := \sum_{h=1}^H \sum_{(u_h,x_h)\in\mathfrak{D}_\ell} \boldsymbol{\phi}(x_h, u_h)\boldsymbol{\phi}(x_h, u_h)^\top$, satisfy

$$\mathrm{tr}(\mathcal{H}(\widehat{A}^\ell)\check{\mathbf{\Lambda}}_\ell^{-1}) \lesssim T_\ell^{-1} \cdot \min_{\mathbf{\Lambda}\in\mathbf{\Omega}} \mathrm{tr}(\mathcal{H}(\widehat{A}^\ell)\check{\mathbf{\Lambda}}^{-1}),$$

which implies that LearnExpΠ collects data minimizing $\mathrm{tr}(\mathcal{H}(\widehat{A}^\ell)\check{\mathbf{\Lambda}}_\ell^{-1})$ at a near-optimal rate. Given the data $\mathfrak{D}_\ell$, we form the least-squares estimate of $A_\star$, $\widehat{A}^{\ell+1}$, and the process repeats. After running for T episodes, the certainty-equivalence controller on the last estimate obtained, $\widehat{\mathfrak{a}}_T = \mathfrak{a}_{\mathrm{opt}}(\widehat{A}^{\ell_T+1})$, is returned.

Note that, while Algorithm 6.1 is similar in structure to the algorithmic approaches presented in Chapter 3 and Chapter 5, it deviates significantly in the strategy employed to compute effective exploration policies. In both Chapter 3 and Chapter 5, we chose the inputs to be the inputs which would optimally excite the estimated system. The analysis employed to show the effectiveness of this approach, however, relied critically on the linearity assumption, and does not immediately extend to nonlinear systems. To circumvent this, in the nonlinear setting we propose an exploration strategy which reduces exploration to policy optimization, and is amenable to nonlinear systems. This is encapsulated in the LearnExpΠ routine, described in more detail in Section 6.3.

The following result bounds the suboptimality of $\widehat{\mathfrak{a}}_T$ returned by Algorithm 6.1 as compared to $\mathfrak{a}_{\mathrm{opt}}(A_\star)$.

Theorem 6.1. *Assume Assumptions 6.1 to 6.6 hold. Then if*

$$T \geq C_{\mathrm{poly}} \cdot \max\left\{1, r_{\mathrm{cost}}(A_\star)^{-2}, r_{\mathfrak{a}}(A_\star)^{-2}\right\}, \tag{6.2.1}$$

with probability at least $1-\delta$, Algorithm 6.1 plays exploration policies $\pi_{\mathrm{exp}} \in \Pi_{\mathrm{exp}}$ at every episode, runs for at most T episodes, and the controller $\widehat{\mathfrak{a}}_T$ returned Algorithm 6.1 satisfies, with probability at least $1-\delta$:

$$\mathcal{J}_{A_\star}(\widehat{\mathfrak{a}}_T) - \mathcal{J}_{A_\star}(\mathfrak{a}_{\mathrm{opt}}(A_\star)) \leq \frac{\sigma_w^2}{T} \cdot \min_{\mathbf{\Lambda}\in\mathbf{\Omega}} \mathrm{tr}\left(\mathcal{H}(A_\star)\check{\mathbf{\Lambda}}^{-1}\right) \cdot C \log \frac{d_x d_{\boldsymbol{\phi}}}{\delta} + \frac{C_{\mathrm{lot}}}{T^{3/2}}$$

for C a universal constant and

$$C_{\mathrm{poly}} := \mathrm{poly}\left(d_{\boldsymbol{\phi}}, d_x, H, B_A, B_{\boldsymbol{\phi}}, L_{\boldsymbol{\phi}}, L_{\mathfrak{a}}, L_{\mathrm{cost}}, L_{\pi_\star}, \sigma_w, \sigma_w^{-1}, \tfrac{1}{\lambda_{\min}}, \log \tfrac{T}{\delta}\right).$$

Theorem 6.1 shows that Algorithm 6.1 is able to explore so as to optimally minimize the exploration loss $\mathrm{tr}(\mathcal{H}(A_\star)\check{\mathbf{\Lambda}}_T^{-1})$, up to a lower-order term scaling as $T^{-3/2}$ and polynomially in system parameters.

Remark 6.2.1 (Computational Efficiency of Algorithm 6.1)**.** The primary computational burden of Algorithm 6.1 is in the computation of $\mathcal{H}(\widehat{A}^\ell)$—which involves differentiating $\mathfrak{a}_{\mathrm{opt}}(\widehat{A}^\ell)$—the computation of $\mathfrak{a}_{\mathrm{opt}}(\widehat{A}^{\ell_T+1})$, and the LearnExpΠ subroutine. In general, if we define $\mathfrak{a}_{\mathrm{opt}}(A)$ as in (6.1.1), it may not be efficiently computable, as it involves solving a possibly non-convex optimization problem, but in certain cases may admit an efficient solution. We discuss the computational efficiency of DynamicOED in more detail in Section 6.3, but note that in general it may not be computationally efficient as it relies on calls to the LC^3 algorithm of Kakade et al. [2020], which requires access to a computational oracle. Despite

these computational challenges, in Section 6.4 we demonstrate that in practice, by making several reasonable approximations, Algorithm 6.1 can be implemented efficiently, and that this efficient implementation performs very well on realistic systems.

6.2.1 Lower Bounds on Learning Controllers

Our goal is to show that, up to constants and lower-order terms, the bound given in Theorem 6.1 is not improvable, regardless of which controller estimate we use. To obtain such lower bounds, we need several additional assumptions:

Assumption 6.7. *There exists some $r_\mu(A_\star) > 0$ such that, for all $A \in \mathcal{B}_{\mathrm{F}}(A_\star, r_\mu(A_\star))$, $\mathfrak{a}_{\mathrm{opt}}(A)$ is unique and, furthermore, there exists some $\mu > 0$ such that*

$$\mathcal{J}_A(\mathfrak{a}) \geq \mathcal{J}_A(\mathfrak{a}_{\mathrm{opt}}(A)) + \tfrac{\mu}{2}\|\mathfrak{a} - \mathfrak{a}_{\mathrm{opt}}(A)\|_2^2.$$

Assumption 6.7 will be satisfied in cases where $\mathcal{J}_A(\mathfrak{a})$ is strongly convex in $\mathfrak{a}$, but may hold even when this is not the case. Intuitively, it requires that our controller class is not overparameterized—moving $\mathfrak{a}$ away from its optimal value will cause the loss to increase. We will additionally make the following regularity assumptions on policies in Π_{exp} and their induced covariates set, $\mathbf{\Omega}$:

Assumption 6.8. *There exists some $\underline{\lambda} > 0$ such that, for each $\mathbf{\Lambda} \in \mathbf{\Omega}$, we have $\lambda_{\min}(\mathbf{\Lambda}) \geq \underline{\lambda}$.*

Assumption 6.8 requires that *every* exploration policy we consider excites all directions in $\boldsymbol{\phi}$ space (in contrast, Assumption 6.3 only assumes there *exists* some distribution over policies in Π_{exp} which excite all directions). We remark that this assumption is relatively mild if Assumption 6.3 holds. As we show in Appendix D.3.1, under Assumption 6.3, a mixture over policies, ω, satisfying $\lambda_{\min}(\mathbb{E}_{\pi\sim\omega}[\mathbf{\Lambda}_\pi]) > 0$ can be learned using only a number of samples scaling polynomially in problem parameters. Given ω, a policy class Π_{exp} satisfying Assumption 6.8 can be obtained by simply mixing ω with every other exploration policy. We are now ready to state our main lower bound.

Theorem 6.2. *Under Assumptions 6.1, 6.2 and 6.4 to 6.8, as long as $T \geq C_{\mathrm{lb}}$, for any $\omega_{\mathrm{exp}} \in \triangle_{\Pi_{\mathrm{exp}}}$, we have*

$$\min_{\widehat{\mathfrak{a}}} \max_{A\in\mathcal{B}_T} \mathbb{E}_{\mathfrak{D}_T\sim A,\omega_{\mathrm{exp}}}[\mathcal{J}_A(\widehat{\mathfrak{a}}(\mathfrak{D}_T)) - \mathcal{J}_A(\mathfrak{a}_{\mathrm{opt}}(A))] \geq \frac{\sigma_w^2}{3T} \cdot \min_{\mathbf{\Lambda}\in\mathbf{\Omega}} \mathrm{tr}(\mathcal{H}(A_\star)\check{\mathbf{\Lambda}}^{-1}) - \frac{C_{\mathrm{lb}}}{T^{5/4}}$$

for $\mathcal{B}_T := \{A \ : \ \|A - A_\star\|_{\mathrm{F}}^2 \leq 5d_x d_{\boldsymbol{\phi}}/(\underline{\lambda} d_x T H)^{5/6}\}$, $\mathbb{E}_{\mathfrak{D}_T\sim A,\omega_{\mathrm{exp}}}[\cdot] = \mathbb{E}_{\pi\sim\omega_{\mathrm{exp}}}[\mathbb{E}_{\mathfrak{D}_T\sim A,\pi}[\cdot]]$ denotes the expectation over trajectories generated by running policies π drawn according to ω_{exp} on system A for T episodes, $\widehat{\mathfrak{a}}$ any mapping from observations to policies in $\Pi^\star$, and

$$C_{\mathrm{lb}} := \mathrm{poly}\left(d_{\boldsymbol{\phi}}, d_x, H, \|A_\star\|_{\mathrm{op}}, B_{\boldsymbol{\phi}}, L_{\boldsymbol{\phi}}, L_{\mathfrak{a}}, L_{\mathrm{cost}}, L_{\pi_\star}, \sigma_w, \sigma_w^{-1}, \tfrac{1}{\underline{\lambda}}, \tfrac{1}{\mu}, \tfrac{1}{r_{\mathrm{cost}}(A_\star)}, \tfrac{1}{r_{\mathfrak{a}}(A_\star)}, \tfrac{1}{r_\mu(A_\star)}\right).$$

This result is an immediate corollary of Theorem 4.3. Note that this lower bound holds for *any* $A_\star$ and mapping $\boldsymbol{\phi}$, as long as our assumptions are met. Up to constants and lower-order terms, the scaling of Theorem 6.2 matches that of Theorem 6.1—both scale with $\min_{\boldsymbol{\Lambda}\in\boldsymbol{\Omega}} \operatorname{tr}(\mathcal{H}(A_\star)\check{\boldsymbol{\Lambda}}^{-1})$—which implies that Algorithm 6.1 is indeed optimal (under certain additional regularity conditions). To the best of our knowledge, this is the first result characterizing the optimal statistical rates for learning in nonlinear dynamical systems. We emphasize that Theorem 6.2 holds for *any* decision rule $\widehat{\mathfrak{a}}$—it does not require that we use the certainty equivalence decision rule. As Algorithm 6.1 does rely on certainty equivalence, this result also implies that the certainty equivalence decision rule is optimal for (certain classes of) nonlinear dynamical systems.

6.3 Optimal Experiment Design in Arbitrary Dynamical Systems

We turn now to the LearnExpΠ routine, which is the key algorithmic tool we use to prove Theorem 6.1. LearnExpΠ itself relies on a separate routine, DynamicOED, which is a general reduction from policy optimization to optimal experiment design in arbitrary dynamical systems. We focus our discussion in this section on DynamicOED, as this is the key algorithmic component—LearnExpΠ is a simple corollary of DynamicOED, and is defined explicitly in Algorithm D.2.

To illustrate the generality of this reduction, in this section we consider the following system:

$$x_{h+1} = f_h(x_h, u_h, w_h), \quad h = 1, 2, \dots, H, \tag{6.3.1}$$

where $x_h \in \mathcal{X}$ denotes the state, $u_h \in \mathcal{U}$ the input, and $w_h \in \mathcal{W}$ the noise. We take the dynamics $(f_h)_{h=1}^H$ to be unknown and arbitrary.

We assume there is some known featurization of our system that is of interest, $\boldsymbol{\psi} : \mathcal{T} \to \mathbb{R}^{d_\psi}$, and an experiment design objective on this featurization, $\Phi : \mathbb{R}^{d_\psi} \to \mathbb{R}$. We also define $\mathcal{T} = (\mathcal{X} \times \mathcal{U})^H \times \mathcal{X}$ the space of possible state-input trajectories, $\boldsymbol{\tau} = (x_0, u_0, x_1, \dots, x_{H-1}, u_{H-1}, x_H) \in \mathcal{T}$, and assume that $\boldsymbol{\psi}$ can be decomposed additively as $\boldsymbol{\psi}(\boldsymbol{\tau}) = \sum_{h=1}^H \boldsymbol{\psi}_h(x_h, u_h)$. Our goal is to collect some set of trajectories $\{\boldsymbol{\tau}_t\}_{t=1}^T$ which minimizes Φ:

$$\Phi\left(\tfrac{1}{T}\textstyle\sum_{t=1}^T \boldsymbol{\psi}(\boldsymbol{\tau}_t)\right).$$

As an example, if $\boldsymbol{\psi}_h(x_h, u_h) = \boldsymbol{\phi}(x_h, u_h)\boldsymbol{\phi}(x_h, u_h)^\top$ for some $\boldsymbol{\phi}$, and $\Phi(\boldsymbol{\Lambda}) = \log\det(\boldsymbol{\Lambda})$, this reduces to D-optimal design, and if $\Phi(\boldsymbol{\Lambda}) = \operatorname{tr}(\mathcal{H} \cdot \boldsymbol{\Lambda}^{-1})$, the setting considered in Section 6.2, this reduces to weighted A-optimal design.

As before, we will be interested in defining optimal exploration with respect to some set

of exploration policies, Π_{exp}. Let

$$\boldsymbol{\Omega}_{\boldsymbol{\psi}} := \{\mathbb{E}_{\pi\sim\omega}[\mathbb{E}_{f,\pi}[\boldsymbol{\psi}(\boldsymbol{\tau})]] \ : \ \omega \in \triangle_{\Pi}\}$$

denote the space of expected value of $\boldsymbol{\psi}(\boldsymbol{\tau})$ for mixtures of policies in Π_{exp}. To distinguish elements $\boldsymbol{\Lambda} \in \boldsymbol{\Omega}$ from elements in $\boldsymbol{\Omega}_{\boldsymbol{\psi}}$, we will let $\boldsymbol{\Gamma} = \mathbb{E}_{\pi\sim\omega}[\mathbb{E}_{f,\pi}[\boldsymbol{\psi}(\boldsymbol{\tau})]]$ refer to elements of $\boldsymbol{\Omega}_{\boldsymbol{\psi}}$, and in particular define $\boldsymbol{\Gamma}_\pi := \mathbb{E}_{f,\pi}[\boldsymbol{\psi}(\boldsymbol{\tau})]$ (where it is assumed that the expectation is collected over trajectories on (6.3.1)). We will usually denote unnormalized sums of features, e.g. $\sum_{t=1}^{T} \boldsymbol{\psi}(\boldsymbol{\tau}_t)$, with $\boldsymbol{\Sigma}$. We also define $\widehat{\boldsymbol{\Omega}}_{\boldsymbol{\psi}}$ to be the space of all possible combinations of $\boldsymbol{\psi}(\boldsymbol{\tau})$:

$$\widehat{\boldsymbol{\Omega}}_{\boldsymbol{\psi}} := \{\mathbb{E}_{\boldsymbol{\tau}\sim\omega}[\boldsymbol{\psi}(\boldsymbol{\tau})] \ : \ \omega \in \triangle_{\mathcal{T}}\}.$$

To facilitate efficient experiment design in this setting, we will make the following assumption on Φ.

Assumption 6.9 (Regularity of Φ)**.** *We make the following assumptions:*

1. *Φ is convex, differentiable, and β-smooth in the norm $\|\cdot\|$:*

$$\|\nabla_{\boldsymbol{\Gamma}}\Phi(\boldsymbol{\Gamma}) - \nabla_{\boldsymbol{\Gamma}'}\Phi(\boldsymbol{\Gamma}')\|_* \le \beta \cdot \|\boldsymbol{\Gamma} - \boldsymbol{\Gamma}'\|, \quad \forall \boldsymbol{\Gamma}, \boldsymbol{\Gamma}' \in \widehat{\boldsymbol{\Omega}}_{\boldsymbol{\psi}}$$

   for $\|\cdot\|_*$ the dual norm of $\|\cdot\|$.

2. *There exists some $M < \infty$ satisfying*

$$\sup_{\boldsymbol{\Gamma}\in\widehat{\boldsymbol{\Omega}}_{\boldsymbol{\psi}}} \sup_{\boldsymbol{\tau}\in\mathcal{T}} |\langle \nabla_{\boldsymbol{\Gamma}}\Phi(\boldsymbol{\Gamma}), \boldsymbol{\psi}(\boldsymbol{\tau})\rangle| \le M.$$

The key algorithmic assumption we make is access to a regret minimization oracle on (6.3.1).

Assumption 6.10 (Regret Minimization Oracle)**.** *Let $\text{cost}_h(\boldsymbol{\tau}) = \langle Q_h, \boldsymbol{\psi}_h(\boldsymbol{\tau})\rangle$ for some $Q_h \in \mathbb{R}^d$, and $\text{cost}(\boldsymbol{\tau}) = \sum_{h=1}^{H} \text{cost}_h(\boldsymbol{\tau})$ the total cost of trajectory $\boldsymbol{\tau}$. We assume we have access to some learner $\mathbb{A}_{\mathcal{R}}$ which, in the setting when $|\text{cost}(\boldsymbol{\tau})| \le 1$ for all $\boldsymbol{\tau} \in \mathcal{T}$, is able to achieve low regret on $\{\text{cost}_h(\cdot,\cdot)\}_{h=1}^{H}$ with respect to policy class Π_{exp}. That is, with probability at least $1-\delta$:*

$$\sum_{t=1}^{T} \mathbb{E}_{f,\pi_t}[\text{cost}(\boldsymbol{\tau}_t)] - T \cdot \inf_{\pi\in\Pi_{\text{exp}}} \mathbb{E}_{f,\pi}[\text{cost}(\boldsymbol{\tau})] \le C_{\mathcal{R}} \cdot \log^{p_{\mathcal{R}}} \frac{T}{\delta} \cdot T^{\alpha}$$

for some $C_{\mathcal{R}} > 0$, $p_{\mathcal{R}} > 0$, and $\alpha \in (0,1)$, and where π_t is the policy $\mathbb{A}_{\mathcal{R}}$ plays at episode t.

Note that the regret minimization algorithm satisfying Assumption 6.10 may be arbitrary. For example, for linear systems, we could apply provably efficient algorithms for the Linear Quadratic Regulator [Simchowitz and Foster, 2020, Mania et al., 2019]; for nonlinear systems of the form (6.0.1) we could apply the LC^3 algorithm of [Kakade et al., 2020]; for more general settings of reinforcement learning with function approximation, algorithms such as BiLin-UCB [Du et al., 2021] or E2D [Foster et al., 2021] could be applied. In practice, though they may not formally satisfy the guarantee of Assumption 6.10, deep RL approaches could be used. We define DynamicOED, our main experiment design routine, in Algorithm 6.2.

Algorithm 6.2 Dynamic Optimal Experiment Design (DynamicOED)

input: objective Φ, number of iterates N, episodes per iterate K, confidence δ, regret minimization algorithm $\mathbb{A}_{\mathcal{R}}$, exploration policies Π_{exp}

Play any policy $\pi_{\mathrm{exp}} \in \Pi_{\mathrm{exp}}$ for K episodes, collect trajectories $\mathfrak{D}_0 = \{\boldsymbol{\tau}_k^0\}_{k=1}^K$, set $\boldsymbol{\Gamma}_0 \leftarrow K^{-1}\sum_{k=1}^K \boldsymbol{\psi}(\boldsymbol{\tau}_k^0)$

for $n = 1, 2, \ldots, N$ **do**

  Set $\gamma_n \leftarrow \frac{1}{n+1}$

  Run $\mathbb{A}_{\mathcal{R}}$ on cost

$$\mathrm{cost}_h^n(\boldsymbol{\tau}) \leftarrow \frac{1}{M}\langle \Xi_n, \boldsymbol{\psi}_h(x_h, u_h)\rangle \quad \text{for} \quad \Xi_n \leftarrow \nabla_{\boldsymbol{\Gamma}}\Phi(\boldsymbol{\Gamma})|_{\boldsymbol{\Gamma}=\boldsymbol{\Gamma}_{n-1}}$$

  for K episodes, collect trajectories $\mathfrak{D}_n = \{\boldsymbol{\tau}_k^n\}_{k=1}^K$, denote policies run as Π_n

  $\boldsymbol{\Gamma}_n \leftarrow (1-\gamma_n)\boldsymbol{\Gamma}_{n-1} + \gamma_n K^{-1}\sum_{k=1}^K \boldsymbol{\psi}(\boldsymbol{\tau}_k^n)$

return $(N+1)K\boldsymbol{\Gamma}_N$, $\cup_{n=0}^N \Pi_n$, $\cup_{n=0}^N \mathfrak{D}_n$

We have the following result.

Theorem 6.3. *Let Assumption 6.9 hold, and assume that we have access to a learner $\mathbb{A}_{\mathcal{R}}$ satisfying Assumption 6.10. Fix $N, K > 0$. Then, with probability at least $1-\delta$, DynamicOED runs for at most $(N+1)K$ episodes, and collects a dataset satisfying $\mathfrak{D} = \{\{\boldsymbol{\tau}_k^n\}_{k=1}^K\}_{n=0}^N$ satisfying*

$$\Phi\left(\frac{1}{K(N+1)}\sum_{n=0}^{N}\sum_{k=1}^{K}\boldsymbol{\psi}(\boldsymbol{\tau}_k^n)\right) - \min_{\boldsymbol{\Gamma}\in\boldsymbol{\Omega}_\psi}\Phi(\boldsymbol{\Gamma}) \le \frac{\beta R^2(\log N+1)}{2(N+1)} + M\cdot\left(C_{\mathcal{R}}\log^{p_{\mathcal{R}}}\frac{2NK}{\delta}\cdot K^{\alpha-1} + \sqrt{\frac{8\log(4N/\delta)}{K}}\right)$$

where $R = \sup_{\boldsymbol{\Gamma},\boldsymbol{\Gamma}'\in\widehat{\boldsymbol{\Omega}}_\psi}\|\boldsymbol{\Gamma}-\boldsymbol{\Gamma}'\|$.

Theorem 6.3 shows that, given access only to a regret minimization oracle, it is possible to solve experiment design problems on arbitrary dynamical systems. Applying Theorem 6.3 with a proper setting of N and K in DynamicOED yields the following result.

Corollary 6.4. *Under Assumption 6.9, and assuming we have access to a learner $\mathbb{A}_{\mathcal{R}}$ satisfying Assumption 6.10 with $\alpha = 1/2$, there exists settings of N and K such that* DynamicOED *runs for T episodes on (6.3.1), and with probability at least $1-\delta$ collects data $\{\boldsymbol{\tau}_t\}_{t\in[T]}$ satisfying*

$$\Phi\left(\frac{1}{T}\cdot\sum_{t=1}^{T}\boldsymbol{\psi}(\boldsymbol{\tau}_t)\right) - \min_{\boldsymbol{\Gamma}\in\boldsymbol{\Omega}_{\boldsymbol{\psi}}}\Phi(\boldsymbol{\Gamma}) \le C\cdot\frac{\beta R^2\log T + M(C_{\mathcal{R}}\log^{p_{\mathcal{R}}}\frac{T}{\delta}+\log^{1/2}\frac{T}{\delta})}{T^{1/3}}.$$

6.3.1 Overview of DynamicOED Algorithm

Conceptually, DynamicOED runs a variant of conditional gradient descent on the objective $\Phi(\boldsymbol{\Gamma})$. At each iteration, n, it computes the gradient of the loss at the current iterate, $\Xi_n \leftarrow \nabla_{\boldsymbol{\Gamma}}\Phi(\boldsymbol{\Gamma})|_{\boldsymbol{\Gamma}=\boldsymbol{\Gamma}_{n-1}}$. To run a standard gradient descent algorithm on this objective, we would simply update $\boldsymbol{\Gamma}_{n-1}$ by taking a step in the direction Ξ_n. However, our objective is to minimize Φ over the constraint set, $\boldsymbol{\Omega}_{\boldsymbol{\psi}}$. Thus, rather than taking a step in the direction Ξ_n, we wish to take a step in the direction of steepest descent *within the constraint set*.

The challenge is that the constraint set in our setting, $\boldsymbol{\Omega}_{\boldsymbol{\psi}}$, is *unknown*, as it depends on the expectation over trajectories induced on the unknown dynamics $(f_h)_{h=1}^H$, and therefore we cannot directly compute this steepest descent direction. The key observation is that the computation of this steepest descent direction is equivalent to solving:

$$\arg\min_{\pi_{\mathrm{exp}}\in\Pi_{\mathrm{exp}}}\mathbb{E}_{f,\pi_{\mathrm{exp}}}\left[\sum_{h=1}^{H}\boldsymbol{\phi}(x_h,u_h)^\top(\Xi_n)\boldsymbol{\phi}(x_h,u_h)\right].$$

This is simply a policy optimization problem, however, and can be solved approximately by $\mathbb{A}_{\mathcal{R}}$ under Assumption 6.10. Thus, in the call to $\mathbb{A}_{\mathcal{R}}$ on Line 5, we approximate the steepest descent direction, and on Line 6 update $\boldsymbol{\Gamma}_{n-1}$ in this direction. Convergence of this procedure to the optimal value, $\min_{\boldsymbol{\Gamma}\in\boldsymbol{\Omega}_{\boldsymbol{\psi}}}\Phi(\boldsymbol{\Gamma})$, can then be shown by the standard analysis of conditional gradient descent. We remark that, under Assumption 6.9 and Assumption 6.10, this argument is completely generic and does not require that our system, (6.3.1), exhibit any additional properties.

Overview of LearnExpΠ. Under certain conditions, it can be shown that, if the exploration policies DynamicOED runs to collect $\mathfrak{D}$ are *rerun*, the newly collected data satisfies a similar guarantee as Theorem 6.3. This lets us run DynamicOED to learn an approximate solution of $\min_{\boldsymbol{\Gamma}\in\boldsymbol{\Omega}}\Phi(\boldsymbol{\Gamma})$, and then rerun the learned policies as many times as desired to collect additional data approximately minimizing Φ. In settings such as those considered in Section 6.2, where we want to ultimately learn an accurate estimate of $A_\star$, by rerunning these policies we can increase the amount of informative data we collect, and improve our accuracy in the estimate of $A_\star$.

LearnExpΠ (Algorithm D.2) instantiates this procedure. In essence, LearnExpΠ

repeatedly calls DynamicOED, gradually increasing the settings of N and K until certain conditions are met which guarantee that the policies returned by DynamicOED will collect sufficiently informative data when rerun. Once these conditions are reached, LearnExpΠ returns these policies—the policies computed by DynamicOED—and terminates. We provide a precise definition of DynamicOED in Appendix D.3.2.

6.3.2 From Corollary 6.4 to Theorem 6.1

In Algorithm 6.1, our goal is to collect covariates, $\check{\mathbf{\Lambda}}_{T_\ell}^{-1}$, such that $\mathrm{tr}(\mathcal{H}(\widehat{A}^\ell)\check{\mathbf{\Lambda}}_{T_\ell}^{-1})$ is as small as possible. To achieve this, we apply LearnExpΠ, which relies on DynamicOED, to the objective $\Phi_\ell(\mathbf{\Lambda}) = \mathrm{tr}(\mathcal{H}(\widehat{A}^\ell)\check{\mathbf{\Lambda}}^{-1})$, with Assumption 6.10 instantiated by the LC^3 algorithm of Kakade et al. [2020]. By the guarantee given in Corollary 6.4, after running for a number of episodes N which scales polynomially in problem parameters, DynamicOED will collect covariates $\mathbf{\Lambda}_N$ such that $\Phi_\ell(\frac{1}{N}\mathbf{\Lambda}_N) \le 2 \cdot \min_{\mathbf{\Lambda}\in\mathbf{\Omega}} \Phi_\ell(\mathbf{\Lambda})$, which implies $\mathrm{tr}(\mathcal{H}(\widehat{A}^\ell)\check{\mathbf{\Lambda}}_N^{-1}) \le \frac{2}{N} \cdot \min_{\mathbf{\Lambda}\in\mathbf{\Omega}} \mathrm{tr}(\mathcal{H}(\widehat{A}^\ell)\check{\mathbf{\Lambda}}^{-1})$. By rerunning the policies DynamicOED used to collect this $\mathbf{\Lambda}_N$ for T_ℓ/N additional times, we can ensure $\mathrm{tr}(\mathcal{H}(\widehat{A}^\ell)\check{\mathbf{\Lambda}}_{T_\ell}^{-1}) \lesssim \frac{1}{T_\ell} \cdot \min_{\mathbf{\Lambda}\in\mathbf{\Omega}} \mathrm{tr}(\mathcal{H}(\widehat{A}^\ell)\check{\mathbf{\Lambda}}^{-1})$ as desired.

We remark that the LC^3 algorithm requires access to a computation oracle. As the focus of this work is primarily statistical, we leave addressing this computational challenge for future work. Furthermore, as we show in the following section, computationally efficient, sampling-based implementations of our approach are very effective in practice. We remark as well that the objective we ultimately care about minimizing is $\mathrm{tr}(\mathcal{H}(A_\star)\check{\mathbf{\Lambda}}^{-1})$. As we show, by including a small amount of uniform exploration, we can ensure that $\mathcal{H}(\widehat{A}^\ell)$ is not too far from $\mathcal{H}(A_\star)$, and so the suboptimality incurred optimizing $\mathrm{tr}(\mathcal{H}(\widehat{A}^\ell)\check{\mathbf{\Lambda}}^{-1})$ instead of $\mathrm{tr}(\mathcal{H}(A_\star)\check{\mathbf{\Lambda}}^{-1})$ only contributes to the lower-order terms of the final guarantee in Theorem 6.1.

6.4 Experimental Results

Finally, we demonstrate the effectiveness of our proposed approach (Algorithm 6.1, the `Task-Driven Exploration` method in Figures 6.1 to 6.3) on several systems motivated by robotic applications. We compare Algorithm 6.1 with an approach that plays $u_h \sim \mathcal{N}(0, \sigma_u^2 \cdot I)$ (`Gaussian Exploration`), and an approach inspired by Mania et al. [2022] (`Uniform Exploration`), which seeks to estimate $A_\star$ uniformly well, playing inputs that reduce $\|\widehat{A} - A_\star\|_{\mathrm{op}}$.

To benchmark the performance of these approaches, we consider an affine system with dynamics corresponding to that of a simplified 3-D drone (i.e., 3-D double integrator with a gravity term), and a nonlinear system with dynamics corresponding to that of a 2-D car. For both systems, we choose $H = 50$, and plot the value of $\mathcal{J}_{A_\star}(\widehat{\mathfrak{a}}_t) - \mathcal{J}_{A_\star}(\mathfrak{a}_{\mathrm{opt}}(A_\star))$ for $\widehat{\pi}_t$ the certainty-equivalence controller computed on the estimate of the system obtained at time t. For the drone, we let $\Pi^\star$ be the class of linear-affine feedback controllers, and for the car, $\Pi^\star$ is a set of nonlinear controllers with dimension 4. While the optimal controller for

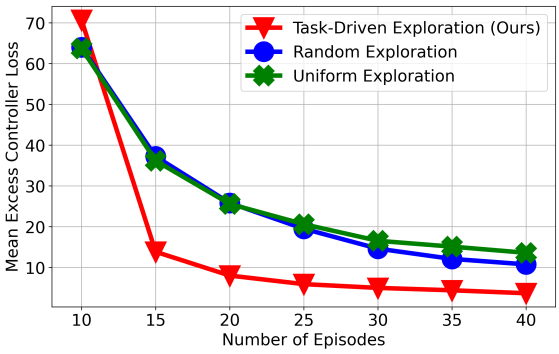

Figure 6.2: Performance on Drone

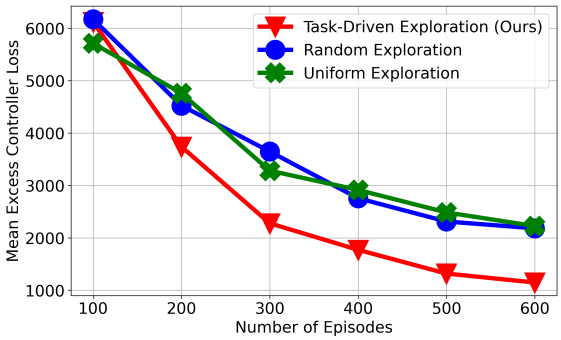

Figure 6.3: Performance on Car

the drone can be computed in closed-form, for the car we rely on a sampling-based routine to find an approximately optimal controller. The model-task hessian $\mathcal{H}(\widehat{A}^{\ell})$ is computed via automatic differentiation. For the exploration policies of `Task-Driven Exploration` and `Uniform Exploration`, Π_{exp}, we rely on MPC-style sampling based methods. For all approaches, we require that $\mathbb{E}_{A,\pi_{\mathrm{exp}}}[\sum_{h=1}^{H} \|u_h\|_2^2] \leq \gamma^2$ for some $\gamma^2 > 0$ and all $\pi_{\mathrm{exp}} \in \Pi_{\mathrm{exp}}$. On all examples, we implement DynamicOED with $\mathbb{A}_{\mathcal{R}}$ a posterior sampling-inspired version of the LC^3 of Kakade et al. [2020]. Figures 6.1 and 6.3 shows performance averaged over 100 trials, and Figure 6.2 over 200 trials.

As illustrated in Figures 6.1 to 6.3, our approach yields a non-trivial gain over existing approaches on all systems. In particular, in Figures 6.2 and 6.3 it improves on the sample complexity of existing approaches by roughly a factor of 2—for example, in the drone system, reaching excess controller cost of 10 after less than 20 episodes, as compared to over 40 episodes for existing approaches.

6.5 Proof of Theorem 6.1

We proceed to the proof of Theorem 6.1. The following result provides estimation error bounds on the estimate $\widehat{A}$ produced at each epoch of Algorithm D.2, and is the starting point of our analysis.

Lemma 6.5.1. *Consider running Algorithm D.2 (LearnExpΠ) with weight matrix $\mathcal{H}$, parameter $\widetilde{N}$, $\mathbb{A}_{\mathcal{R}}$ the LC^3 algorithm of Kakade et al. [2020], and confidence δ, and rerunning each policy in Π_{out}, the policies returned by LearnExpΠ, $N \leq \widetilde{N}$ times. Then, under Assumptions 6.1 and 6.3, with probability at least $1 - 6\delta$:*

$$\|\mathrm{vec}(\widehat{A} - A_\star)\|_{\mathcal{H}}^2 \leq \frac{60}{NT_{\mathrm{out}}} \cdot \min_{\boldsymbol{\Lambda} \in \boldsymbol{\Omega}} \mathrm{tr}\left(\mathcal{H}\boldsymbol{\Lambda}^{-1}\right) \cdot \sigma_w^2 \log \frac{6d_x d_{\boldsymbol{\phi}}}{\delta}$$

$$+ \operatorname{poly}\left(d_{\boldsymbol{\phi}}, d_x, \frac{1}{\bar{\lambda}_{\min}}, B_A, B_{\boldsymbol{\phi}}, \log\frac{1}{\sigma_w}, H, \|\mathcal{H}\|_{\mathrm{op}}, \log\frac{\widetilde{N}}{\delta}\right) \cdot \frac{1}{N^2}$$

where $\widehat{A}$ *denotes the least-squares estimate of* $A_\star$ *obtained on the data generated by rerunning* Π_{out} *and* $T_{\mathrm{out}} := |\Pi_{\mathrm{out}}|$*. In addition, we have*

$$\|\widehat{A} - A_\star\|_{\mathrm{F}} \le \operatorname{poly}\left(d_{\boldsymbol{\phi}}, d_x, \frac{1}{\bar{\lambda}_{\min}}, B_A, B_{\boldsymbol{\phi}}, \sigma_w, H, \log\frac{\widetilde{N}}{\delta}\right) \cdot \frac{1}{\sqrt{N}}$$

and

$$T_{\mathrm{out}} \le \operatorname{poly}\left(d_{\boldsymbol{\phi}}, d_x, \frac{1}{\bar{\lambda}_{\min}}, B_A, B_{\boldsymbol{\phi}}, \log\frac{1}{\sigma_w}, H, \log\frac{\widetilde{N}}{\delta}\right)$$

and the total number of episodes collected by this procedure is bounded by $NT_{\mathrm{out}} + (16 + 2\log(T_{\mathrm{out}})) \cdot T_{\mathrm{out}}$*.*

Let $\mathcal{E}_\ell$ denote that the good event of Lemma 6.5.1 holds at round ℓ which, by Lemma 6.5.1, occurs with probability at least $1-6\delta_\ell$, since we call LearnExpΠ with confidence $\delta_\ell = \delta/12\ell^2$. By our setting of δ_ℓ, we have that the total failure probability of $\mathcal{E}_\ell$ for all ℓ is bounded as

$$\sum_{\ell=1}^{\infty} 6 \cdot \frac{\delta}{12\ell^2} \le \delta.$$

Henceforth we assume that $\mathcal{E} := \cap_\ell \mathcal{E}_\ell$ holds. Let $\widehat{A} := \widehat{A}^{\ell_T+1}$, and $\widehat{A}^- := \widehat{A}^{\ell_T}$.

Bounding the Number of Episodes. Denote $T_\ell^{\mathrm{oed}} = |\Pi_\ell|$. Note that by construction we always have $N_\ell T_\ell^{\mathrm{oed}} \ge T_\ell$. By Lemma D.1.2, as long as (6.2.1) is met, we can bound the total number of episodes collected up to and including round ℓ by $4T_\ell$ for $\ell \in \{\ell_T, \ell_T - 1\}$. We therefore, in the following, will make use of the fact that

$$c \cdot T \le N_\ell K_\ell \le c' \cdot T$$

for $\ell \in \{\ell_T, \ell_T - 1\}$ and absolute constants c, c'. Furthermore, it also follows from this that the total number of episodes run by Algorithm 6.1 is bounded by

$$4T_{\ell_T} = 4 \cdot 2^{\lceil \log T/8 \rceil} \le 8 \cdot \frac{T}{8} = T,$$

so Algorithm 6.1 runs for at most T episodes.

Approximating the Controller Loss. Let $r_{\mathrm{est}}(A_\star) := \min\{1, r_{\mathrm{cost}}(A_\star), r_{\mathfrak{a}}(A_\star)\}$, for $r_{\mathrm{cost}}(A_\star)$ as in Assumption 6.2 and $r_{\mathfrak{a}}(A_\star)$ as in Assumption 6.6. By Lemma D.2.4, un-

der Assumptions 6.1, 6.2 and 6.4 to 6.6, as long as $\widehat{A} \in \mathcal{B}_{\mathrm{F}}(A_\star; r_{\mathrm{est}}(A_\star))$, we have

$$\begin{aligned}\mathcal{J}_{A_\star}(\widehat{\pi}_T) - \mathcal{J}_{A_\star}(\pi_\star(A_\star)) &\le \|\mathrm{vec}(\widehat{A} - A_\star)\|^2_{\mathcal{H}(A_\star)} \\ &\quad + \mathrm{poly}(L_{\pi_\star}, \|A_\star\|_{\mathrm{op}}, L_{\boldsymbol{\phi}}, L_{\mathfrak{a}}, L_{\mathrm{cost}}, \sigma_w^{-1}, H, d_x) \cdot \|\widehat{A} - A_\star\|^3_{\mathrm{op}}.\end{aligned}$$

Furthermore, we can bound

$$\begin{aligned}\|\mathrm{vec}(\widehat{A} - A_\star)\|^2_{\mathcal{H}(A_\star)} &= \|\mathrm{vec}(\widehat{A} - A_\star)\|^2_{\mathcal{H}(\widehat{A}^-)} + \mathrm{vec}(\widehat{A} - A_\star)^\top (\mathcal{H}(A_\star) - \mathcal{H}(\widehat{A}^-))\mathrm{vec}(\widehat{A} - A_\star) \\ &\le (\widehat{A} - A_\star)^\top \mathcal{H}(\widehat{A}^-)(\widehat{A} - A_\star) + \|\widehat{A}^- - A_\star\|^2_{\mathrm{op}} \|\mathcal{H}(A_\star) - \mathcal{H}(\widehat{A}^-)\|_{\mathrm{op}}.\end{aligned}$$

Bounding the Hessian Estimation Error. On $\mathcal{E}$, by Lemma 6.5.1 and as long as (6.2.1) is met, we have (note that at the final epoch, the plug-in estimator $\mathcal{H}(\widehat{A}^-)$ is given as input to LearnExpΠ):

$$\begin{aligned}\|\mathrm{vec}(\widehat{A} - A_\star)\|^2_{\mathcal{H}(\widehat{A}^-)} &\le \frac{60}{N_{\ell_T} T^{\mathrm{oed}}_{\ell_T}} \cdot \min_{\boldsymbol{\Lambda} \in \boldsymbol{\Omega}} \mathrm{tr}\left(\mathcal{H}(\widehat{A}^-) \check{\boldsymbol{\Lambda}}^{-1}\right) \cdot \sigma_w^2 \log \frac{6 d_x d_{\boldsymbol{\phi}}}{\delta} \\ &\quad + \mathrm{poly}\left(d_{\boldsymbol{\phi}}, d_x, \frac{1}{\bar{\lambda}_{\min}}, B_A, B_{\boldsymbol{\phi}}, \log \frac{1}{\sigma_w}, H, \|\mathcal{H}(\widehat{A}^-)\|_{\mathrm{op}}, \log \frac{T_{\ell_T}}{\delta}\right) \cdot \frac{1}{N^2_{\ell_T}} \\ &\le \frac{C}{T} \cdot \min_{\boldsymbol{\Lambda} \in \boldsymbol{\Omega}} \mathrm{tr}\left(\mathcal{H}(\widehat{A}^-) \check{\boldsymbol{\Lambda}}^{-1}\right) \cdot \sigma_w^2 \log \frac{6 d_x d_{\boldsymbol{\phi}}}{\delta} \\ &\quad + \mathrm{poly}\left(d_{\boldsymbol{\phi}}, d_x, \frac{1}{\bar{\lambda}_{\min}}, B_A, B_{\boldsymbol{\phi}}, \log \frac{1}{\sigma_w}, H, \|\mathcal{H}(\widehat{A}^-)\|_{\mathrm{op}}, \log \frac{T}{\delta}\right) \cdot \frac{1}{T^2}\end{aligned}$$

where the last line uses that $N_{\ell_T} T^{\mathrm{oed}}_{\ell_T}$ is within a constant of T, and that $T^{\mathrm{oed}}_{\ell_T}$ can be bounded by

$$\mathrm{poly}\left(d_{\boldsymbol{\phi}}, d_x, \frac{1}{\bar{\lambda}_{\min}}, B_A, B_{\boldsymbol{\phi}}, \log \frac{1}{\sigma_w}, H, \log \frac{T}{\delta}\right)$$

by Lemma 6.5.1. We next wish to show that $\mathrm{tr}(\mathcal{H}(\widehat{A}^-)\check{\boldsymbol{\Lambda}}^{-1})$ and $\mathrm{tr}(\mathcal{H}(A_\star)\check{\boldsymbol{\Lambda}}^{-1})$ are close. The following result shows that this is the case, assuming $\mathcal{H}(\widehat{A}^-)$ and $\mathcal{H}(A_\star)$ are close.

Lemma 6.5.2. *Under Assumption 6.3, for any $\mathcal{H}, \mathcal{H}'$, we can bound*

$$\min_{\check{\boldsymbol{\Lambda}} \in \boldsymbol{\Omega}} \mathrm{tr}(\mathcal{H} \check{\boldsymbol{\Lambda}}^{-1}) \le \min_{\check{\boldsymbol{\Lambda}} \in \boldsymbol{\Omega}} 2\mathrm{tr}(\mathcal{H}' \check{\boldsymbol{\Lambda}}^{-1}) + \frac{2 d_x d_{\boldsymbol{\phi}}}{\bar{\lambda}_{\min}} \cdot \|\mathcal{H} - \mathcal{H}'\|_{\mathrm{op}}.$$

By Lemma 6.5.2 we can bound

$$\min_{\boldsymbol{\Lambda} \in \boldsymbol{\Omega}} \mathrm{tr}\left(\mathcal{H}(\widehat{A}^-) \check{\boldsymbol{\Lambda}}^{-1}\right) \le \min_{\boldsymbol{\Lambda} \in \boldsymbol{\Omega}} 2\mathrm{tr}\left(\mathcal{H}(A_\star) \check{\boldsymbol{\Lambda}}^{-1}\right) + \frac{2 d_x d_{\boldsymbol{\phi}}}{\bar{\lambda}_{\min}} \cdot \|\mathcal{H}(A_\star) - \mathcal{H}(\widehat{A}^-)\|_{\mathrm{op}}$$

and by Lemma D.2.5, under Assumptions 6.1, 6.2 and 6.4 to 6.6, and as long as $\widehat{A}^- \in \mathcal{B}_{\mathrm{F}}(A_\star; r_{\mathrm{est}}(A_\star))$, we can bound

$$\|\mathcal{H}(A_\star) - \mathcal{H}(\widehat{A}^-)\|_{\mathrm{op}} \le \mathrm{poly}(L_{\pi_\star}, \|A_\star\|_{\mathrm{op}}, L_{\boldsymbol{\phi}}, L_{\mathfrak{a}}, L_{\mathrm{cost}}, \sigma_w^{-1}, H, d_x) \cdot \|\widehat{A}^- - A_\star\|_{\mathrm{op}}.$$

Let

$$C_{\mathrm{lot}} := \mathrm{poly}\left(L_{\pi_\star}, B_A, B_{\boldsymbol{\phi}}, L_{\boldsymbol{\phi}}, L_{\mathfrak{a}}, L_{\mathrm{cost}}, d_x, d_{\boldsymbol{\phi}}, \frac{1}{\bar{\lambda}_{\min}}, \sigma_w, \sigma_w^{-1}, H, \|\mathcal{H}(A_\star)\|_{\mathrm{op}}, \log\frac{T}{\delta}\right)$$

denote some lower-order constant, whose precise polynomial dependence may change from line to line. On $\mathcal{E}$, by Lemma 6.5.1 we can bound

$$\|\widehat{A} - A_\star\|_{\mathrm{op}}, \|\widehat{A}^- - A_\star\|_{\mathrm{op}} \le C_{\mathrm{lot}} \cdot \frac{1}{\sqrt{T}}, \tag{6.5.1}$$

and so assuming the burn-in (6.2.1) is met, we can bound $\|\widehat{A}^- - A_\star\|_{\mathrm{op}} \le \frac{1}{2} r_{\mathrm{est}}(A_\star)$ and $\|\widehat{A} - A_\star\|_{\mathrm{op}} \le \frac{1}{2} r_{\mathrm{est}}(A_\star)$. This then implies that

$$\|\mathcal{H}(A_\star) - \mathcal{H}(\widehat{A}^-)\|_{\mathrm{op}} \le C_{\mathrm{lot}} \cdot \frac{1}{\sqrt{T}},$$

so in particular we can bound

$$\begin{aligned}
&\mathrm{poly}\left(d_{\boldsymbol{\phi}}, d_x, \frac{1}{\bar{\lambda}_{\min}}, B_A, B_{\boldsymbol{\phi}}, H, \log\frac{1}{\sigma_w}, \|\mathcal{H}(\widehat{A}^-)\|_{\mathrm{op}}, \log\frac{T}{\delta}\right) \\
&\le \mathrm{poly}\left(d_{\boldsymbol{\phi}}, d_x, \frac{1}{\bar{\lambda}_{\min}}, B_A, B_{\boldsymbol{\phi}}, H, \log\frac{1}{\sigma_w}, \|\mathcal{H}(A_\star)\|_{\mathrm{op}}, \log\frac{T}{\delta}\right).
\end{aligned}$$

We have therefore shown that

$$\begin{aligned}
\mathcal{J}_{A_\star}(\widehat{\pi}_T) - \mathcal{J}_{A_\star}(\pi_\star(A_\star)) &\le \frac{C}{T} \cdot \min_{\boldsymbol{\Lambda} \in \boldsymbol{\Omega}} \mathrm{tr}\left(\mathcal{H}(A_\star)\check{\boldsymbol{\Lambda}}^{-1}\right) \cdot \sigma_w^2 \log\frac{6 d_x d_{\boldsymbol{\phi}}}{\delta} \\
&\quad + C_{\mathrm{lot}} \cdot \left(\|\widehat{A} - A_\star\|_{\mathrm{op}}^3 + \|\widehat{A}^- - A_\star\|_{\mathrm{op}}^3 + \frac{1}{T^2} + \frac{1}{T} \cdot \|\widehat{A}^- - A_\star\|_{\mathrm{op}}\right) \\
&\le \frac{C}{T} \cdot \min_{\boldsymbol{\Lambda} \in \boldsymbol{\Omega}} \mathrm{tr}\left(\mathcal{H}(A_\star)\check{\boldsymbol{\Lambda}}^{-1}\right) \cdot \sigma_w^2 \log\frac{6 d_x d_{\boldsymbol{\phi}}}{\delta} + \frac{C_{\mathrm{lot}}}{T^{3/2}},
\end{aligned}$$

The final result follows from using Lemma D.2.6 to bound

$$\|\mathcal{H}(A_\star)\|_{\mathrm{op}} \le \mathrm{poly}(\|A_\star\|_{\mathrm{op}}, B_{\boldsymbol{\phi}}, L_{\boldsymbol{\phi}}, L_{\mathfrak{a}}, L_{\mathrm{cost}}, L_{\pi_\star}, \sigma_w^{-1}, H, d_x).$$

6.6 Proof of Theorem 6.3

As noted, the DynamicOED algorithm is based on conditional gradient descent, also known as the Frank-Wolfe algorithm [Frank and Wolfe, 1956]. In this section we first show that the Frank-Wolfe algorithm can be run with the iterate only approximately the minimizer of the gradient, and then in Section 6.6.2 use this to prove Theorem 6.3.

6.6.1 Approximate Frank-Wolfe

Algorithm 6.3 Approximate Frank-Wolfe

input: function to optimize g, number of iterations to run N, starting iterate x_1
for $t = 1, 2, \ldots, N$ **do**
    Set $\gamma_n \leftarrow \frac{1}{n+1}$
    Choose $\boldsymbol{y}_n$ to be any point such that

$$\nabla g(\boldsymbol{z}_n)^\top \boldsymbol{y}_n \leq \min_{\boldsymbol{y} \in \mathcal{Z}} \nabla g(\boldsymbol{z}_n)^\top \boldsymbol{y} + \epsilon_n$$

    $\boldsymbol{z}_{n+1} \leftarrow (1 - \gamma_n)\boldsymbol{z}_n + \gamma_n \boldsymbol{y}_n$
return x_{N+1}

Lemma 6.6.1. *Consider running Algorithm 6.3 with some convex function f that is β-smooth with respect to some norm $\|\cdot\|$, assume that $\boldsymbol{y}_n \in \mathcal{Y}$ for some $\mathcal{Y}$ and all n, and let $R := \sup_{\boldsymbol{z}, \boldsymbol{y} \in \mathcal{Z} \cup \mathcal{Y}} \|\boldsymbol{z} - \boldsymbol{y}\|$. Then for $N \geq 2$, we have*

$$g(\boldsymbol{z}_{N+1}) - \min_{\boldsymbol{z} \in \mathcal{Z}} g(\boldsymbol{z}) \leq \frac{\beta R^2 (\log N + 1)}{2(N+1)} + \frac{1}{N+1} \sum_{n=1}^{N} \epsilon_n.$$

Proof. Let $\boldsymbol{z}^\star = \arg\min_{\boldsymbol{z} \in \mathcal{Z}} g(\boldsymbol{z})$. Using that g is β-smooth, the definition of $\boldsymbol{y}_s$, and the convexity of g, we have that for any s,

$$\begin{aligned}
g(\boldsymbol{z}_{s+1}) - g(\boldsymbol{z}_s) &\leq \nabla g(\boldsymbol{z}_s)^\top (\boldsymbol{z}_{s+1} - \boldsymbol{z}_s) + \frac{\beta}{2} \|\boldsymbol{z}_{s+1} - \boldsymbol{z}_s\|^2 \\
&\leq \gamma_s \nabla g(\boldsymbol{z}_s)^\top (\boldsymbol{y}_s - \boldsymbol{z}_s) + \frac{\beta}{2} \gamma_s^2 R^2 \\
&\leq \gamma_s \nabla g(\boldsymbol{z}_s)^\top (\boldsymbol{z}^\star - \boldsymbol{z}_s) + \gamma_s \epsilon_s + \frac{\beta}{2} \gamma_s^2 R^2 \\
&\leq \gamma_s (g(\boldsymbol{z}^\star) - g(\boldsymbol{z}_s)) + \gamma_s \epsilon_s + \frac{\beta}{2} \gamma_s^2 R^2.
\end{aligned}$$

Letting $\delta_s = g(\boldsymbol{z}_s) - g(\boldsymbol{z}^\star)$, this implies that

$$\delta_{s+1} \leq (1-\gamma_s)\delta_s + \gamma_s\epsilon_s + \frac{\beta}{2}\gamma_s^2 R^2.$$

Unrolling this backwards gives

$$\begin{aligned}
\delta_{N+1} &\leq (1-\gamma_N)\delta_N + \gamma_N\epsilon_N + \frac{\beta}{2}\gamma_N^2 R^2 \\
&\leq (1-\gamma_N)(1-\gamma_{N-1})\delta_{N-1} + (1-\gamma_N)(\gamma_{N-1}\epsilon_{N-1} + \frac{\beta}{2}\gamma_{N-1}^2 R^2) + \gamma_N\epsilon_N + \frac{\beta}{2}\gamma_N^2 R^2 \\
&\leq \sum_{n=1}^{N}\left(\prod_{s=n+1}^{N}(1-\gamma_s)\right)(\gamma_n\epsilon_n + \frac{\beta}{2}\gamma_n^2 R^2).
\end{aligned}$$

We can write

$$\prod_{s=n+1}^{N}(1-\gamma_s) = \prod_{s=n+1}^{N}\frac{s}{s+1} = \frac{n+1}{N+1}$$

so

$$\begin{aligned}
\sum_{n=1}^{N}\left(\prod_{s=n+1}^{N}(1-\gamma_s)\right)\frac{\beta}{2}\gamma_n^2 R^2 &= \sum_{n=1}^{N}\frac{n+1}{N+1}\frac{\beta}{2}\frac{1}{(n+1)^2}R^2 \\
&= \frac{\beta R^2}{2(N+1)}\sum_{n=1}^{N}\frac{1}{n+1} \\
&\leq \frac{\beta R^2(\log N + 1)}{2(N+1)}
\end{aligned}$$

and

$$\begin{aligned}
\sum_{n=1}^{N}\left(\prod_{s=n+1}^{N}(1-\gamma_s)\right)\gamma_n\epsilon_n &= \sum_{n=1}^{N}\frac{n+1}{N+1}\frac{1}{n+1}\epsilon_n \\
&= \frac{1}{N+1}\sum_{n=1}^{N}\epsilon_n
\end{aligned}$$

which proves the result.

□

Lemma 6.6.2. *When running Algorithm 6.3, we have*

$$\boldsymbol{z}_{N+1} = \frac{1}{N+1}\left(\sum_{n=1}^{N} \boldsymbol{y}_n + \boldsymbol{z}_1\right).$$

Proof. We have:

$$\begin{aligned}
\boldsymbol{z}_{N+1} &= \sum_{n=1}^{N}\left(\prod_{s=n+1}^{N}(1-\gamma_s)\right)\gamma_n \boldsymbol{y}_n + \left(\prod_{s=1}^{N}(1-\gamma_s)\right)\boldsymbol{z}_1 \\
&= \sum_{n=1}^{N}\frac{n+1}{N+1}\frac{1}{n+1}\boldsymbol{y}_t + \frac{1}{N+1}\boldsymbol{z}_1 \\
&= \frac{1}{N+1}\sum_{n=1}^{N}\boldsymbol{y}_n + \frac{1}{N+1}\boldsymbol{z}_1.
\end{aligned}$$

□

6.6.2 Proof of Theorem 6.3

Proof of Theorem 6.3. By our assumption on $\mathbb{A}_{\mathcal{R}}$, Assumption 6.10, we have that, at round n, with probability at least $1-\delta/2N$,

$$\sum_{k=1}^{K}\mathbb{E}_{\pi_k}[\mathrm{cost}^n(\boldsymbol{\tau}_k)] - K\cdot\inf_{\pi\in\Pi_{\mathrm{exp}}}\mathbb{E}_{\pi}\left[\mathrm{cost}^n(\boldsymbol{\tau})\right] \leq C_{\mathcal{R}}\log^{p_{\mathcal{R}}}\frac{2NK}{\delta}\cdot K^{\alpha}$$

where we have used that, under Assumption 6.9 and by the definition of $\mathrm{cost}^n_h(\boldsymbol{\tau})$, $|\mathrm{cost}^n(\boldsymbol{\tau})| \leq 1$ for all $\boldsymbol{\tau}\in\mathcal{T}$. This implies that

$$\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}_{\pi_k}[\langle\Xi_n, \boldsymbol{\psi}(\boldsymbol{\tau}_k)\rangle] \leq M\cdot\inf_{\pi\in\Pi_{\mathrm{exp}}}\mathbb{E}_{\pi}\left[\mathrm{cost}^n(\boldsymbol{\tau})\right] + MC_{\mathcal{R}}\log^{p_{\mathcal{R}}}\frac{2NK}{\delta}\cdot K^{\alpha-1}.$$

Furthermore, by Azuma-Hoeffding and under Assumption 6.9, we have that, with probability at least $1-\delta/2N$,

$$\left|\frac{1}{K}\sum_{k=1}^{K}\langle\Xi_n, \boldsymbol{\psi}(\boldsymbol{\tau}_k)\rangle - \frac{1}{K}\sum_{k=1}^{K}\mathbb{E}_{\pi_k}[\langle\Xi_n, \boldsymbol{\psi}(\boldsymbol{\tau}_k)\rangle]\right| \leq \sqrt{\frac{8M^2\log(4N/\delta)}{K}}.$$

This implies that

$$\frac{1}{K}\sum_{k=1}^{K}\langle \Xi_n, \boldsymbol{\psi}(\boldsymbol{\tau}_k)\rangle \leq M \cdot \inf_{\pi\in\Pi_{\text{exp}}} \mathbb{E}_\pi\left[\text{cost}^n(\boldsymbol{\tau})\right] + M \cdot \left(C_{\mathcal{R}} \log^{p_{\mathcal{R}}} \frac{K}{\delta} \cdot K^{\alpha-1} + \sqrt{\frac{8\log(4N/\delta)}{K}}\right). \tag{6.6.1}$$

Note that

$$\langle \Xi_n, \boldsymbol{\psi}(\boldsymbol{\tau}_k)\rangle = \langle \nabla_{\boldsymbol{\Gamma}}\Phi(\boldsymbol{\Gamma})|_{\boldsymbol{\Gamma}=\boldsymbol{\Gamma}_n}, \boldsymbol{\psi}(\boldsymbol{\tau})\rangle,$$

and that for any $\boldsymbol{\Gamma} \in \boldsymbol{\Omega}_{\boldsymbol{\psi}}$, we have

$$\langle \nabla_{\boldsymbol{\Gamma}}\Phi(\boldsymbol{\Gamma})|_{\boldsymbol{\Gamma}=\boldsymbol{\Gamma}_n}, \boldsymbol{\Gamma}\rangle = \mathbb{E}_{\pi\sim\omega}[\mathbb{E}_\pi[\langle \nabla_{\boldsymbol{\Gamma}}\Phi(\boldsymbol{\Gamma})|_{\boldsymbol{\Gamma}=\boldsymbol{\Gamma}_n}, \boldsymbol{\psi}(\boldsymbol{\tau})\rangle]]$$

for some ω. This implies that

$$\begin{aligned}
\inf_{\boldsymbol{\Gamma}\in\boldsymbol{\Omega}_{\boldsymbol{\psi}}} \langle \nabla_{\boldsymbol{\Gamma}}\Phi(\boldsymbol{\Gamma})|_{\boldsymbol{\Gamma}=\boldsymbol{\Gamma}_n}, \boldsymbol{\Gamma}\rangle &= \inf_{\omega\in\triangle_\Pi} \mathbb{E}_{\pi\sim\omega}[\mathbb{E}_\pi[\langle \nabla_{\boldsymbol{\Gamma}}\Phi(\boldsymbol{\Gamma})|_{\boldsymbol{\Gamma}=\boldsymbol{\Gamma}_n}, \boldsymbol{\psi}(\boldsymbol{\tau})\rangle]] \\
&= \inf_{\pi\in\Pi_{\text{exp}}} \mathbb{E}_\pi[\langle \Xi_n, \boldsymbol{\psi}(\boldsymbol{\tau})\rangle] \\
&= M \cdot \inf_{\pi\in\Pi_{\text{exp}}} \mathbb{E}_\pi[\text{cost}^n(\boldsymbol{\tau})].
\end{aligned}$$

By (6.6.1) above, we have that

$$\frac{1}{K}\sum_{k=1}^{K}\boldsymbol{\psi}(\boldsymbol{\tau}_k)$$

is an approximate minimizer of $M \cdot \sup_{\pi\in\Pi_{\text{exp}}} \mathbb{E}_\pi[\text{cost}^n(\boldsymbol{\tau})]$, with approximation tolerance $M(C_{\mathcal{R}} \log^{p_{\mathcal{R}}} \frac{2NK}{\delta} \cdot K^{\alpha-1} + \sqrt{\frac{8\log(4N/\delta)}{K}})$. We can therefore apply Lemma 6.6.1 with

$$\epsilon_n = M \cdot \left(C_{\mathcal{R}} \log^{p_{\mathcal{R}}} \frac{K}{\delta} \cdot K^{\alpha-1} + \sqrt{\frac{8\log(4N/\delta)}{K}}\right)$$

to get that

$$\Phi(\boldsymbol{\Gamma}_{N+1}) - \min_{\boldsymbol{\Gamma}\in\boldsymbol{\Omega}_{\boldsymbol{\psi}}} \Phi(\boldsymbol{\Gamma}) \leq \frac{\beta R^2(\log N + 1)}{2(N+1)} + M \cdot \left(C_{\mathcal{R}} \log^{p_{\mathcal{R}}} \frac{2NK}{\delta} \cdot K^{\alpha-1} + \sqrt{\frac{8\log(4N/\delta)}{K}}\right).$$

The result then follows since $\boldsymbol{\Gamma}_{N+1} = \frac{1}{K(N+1)}\sum_{n=0}^{N}\sum_{k=1}^{K}\boldsymbol{\psi}(\boldsymbol{\tau}_k^n)$ by Lemma 6.6.2.

□

In this work we are particularly interested in the case where $\boldsymbol{\psi}(\boldsymbol{\tau}) \in \mathcal{S}_+^{d_\psi}$. We encapsulate this in the following assumption.

Assumption 6.11 (Matrix Experiment Design)**.** *We assume that $\boldsymbol{\psi}(\boldsymbol{\tau}) \in \mathcal{S}_+^{d_\psi}$ and that, for all $\boldsymbol{\tau} \in \mathcal{T}$, $\operatorname{tr}(\boldsymbol{\psi}(\boldsymbol{\tau})) \leq D$ for some $D > 0$.*

The following corollary instantiates Theorem 6.3 under Assumption 6.11 with objective $\Phi(\boldsymbol{\Gamma}) = \operatorname{tr}\left(\mathcal{H}(\boldsymbol{\Gamma} + \boldsymbol{\Gamma}_0)^{-1}\right)$, the objective considered in Algorithm 6.1.

Corollary 6.5. *Consider the objective*

$$\Phi(\boldsymbol{\Gamma}) = \operatorname{tr}\left(\mathcal{H} \cdot (\boldsymbol{\Gamma} + \boldsymbol{\Gamma}_0)^{-1}\right)$$

and assume that $\mathcal{H} \succeq 0$ and Assumption 6.10 holds with $\alpha = 1/2$ and Assumption 6.11 holds. Fix N, K, let $T := (N+1)K$, and consider running Algorithm 6.2 on this objective. Then Algorithm 6.2 will run for at most T episodes, and, with probability at least $1-\delta$, will return data satisfying

$$\operatorname{tr}\left(\mathcal{H}\left(\sum_{t=1}^{T} \boldsymbol{\psi}(\boldsymbol{\tau}_t) + T\boldsymbol{\Gamma}_0\right)^{-1}\right) \leq \frac{1}{T} \cdot \min_{\boldsymbol{\Gamma} \in \boldsymbol{\Omega}_\psi} \operatorname{tr}\left(\mathcal{H}(\boldsymbol{\Gamma} + \boldsymbol{\Gamma}_0)^{-1}\right) + \frac{8D^4 \|\mathcal{H}\|_{\mathrm{op}} \|\boldsymbol{\Gamma}_0^{-1}\|_{\mathrm{op}}^3}{T(N+1)} + \frac{8D\|\mathcal{H}\|_{\mathrm{op}}\|\boldsymbol{\Gamma}_0^{-1}\|_{\mathrm{op}}^2(\log^{1/2}\frac{4T}{\delta} + C_{\mathcal{R}}\log^{p_{\mathcal{R}}}\frac{2T}{\delta})}{T\sqrt{K}}.$$

Proof. By Theorem 6.3, for any setting of N and K, we have that with probability at least $1-\delta$:

$$\operatorname{tr}\left(\mathcal{H}\left(\frac{1}{K(N+1)}\sum_{n=0}^{N}\sum_{k=1}^{K} \boldsymbol{\psi}(\boldsymbol{\tau}_k^n) + \boldsymbol{\Gamma}_0\right)^{-1}\right) - \min_{\boldsymbol{\Gamma} \in \boldsymbol{\Omega}} \operatorname{tr}\left(\mathcal{H}(\boldsymbol{\Gamma} + \boldsymbol{\Gamma}_0)^{-1}\right) \leq \frac{\beta R^2 \log N}{N+1} + \frac{MC_{\mathcal{R}}\log^{p_{\mathcal{R}}}\frac{2NK}{\delta}}{K^{1-\alpha}} + M\sqrt{\frac{8\log\frac{4N}{\delta}}{K}}$$

which implies

$$\operatorname{tr}\left(\mathcal{H}\left(\sum_{n=0}^{N}\sum_{k=1}^{K} \boldsymbol{\psi}(\boldsymbol{\tau}_k^n) + T\boldsymbol{\Gamma}_0\right)^{-1}\right) - \frac{\min_{\boldsymbol{\Gamma} \in \boldsymbol{\Omega}} \operatorname{tr}\left(\mathcal{H}(\boldsymbol{\Gamma} + \boldsymbol{\Gamma}_0)^{-1}\right)}{T} \leq \frac{\beta R^2 \log N}{T(N+1)} + \frac{MC_{\mathcal{R}}\log^{p_{\mathcal{R}}}\frac{2NK}{\delta}}{T\sqrt{K}} + M\frac{1}{T}\sqrt{\frac{8\log\frac{4N}{\delta}}{K}}$$

This gives

$$\operatorname{tr}\left(\mathcal{H}\left(\sum_{n=0}^{N}\sum_{k=1}^{K}\boldsymbol{\psi}(\boldsymbol{\tau}_k^n)+T\boldsymbol{\Gamma}_0\right)^{-1}\right) \le \frac{\min_{\boldsymbol{\Gamma}\in\boldsymbol{\Omega}}\operatorname{tr}\left(\mathcal{H}(\boldsymbol{\Gamma}+\boldsymbol{\Gamma}_0)^{-1}\right)}{T} + \frac{\beta R^2\log T}{T(N+1)} + \frac{M(3\log^{1/2}\frac{4T}{\delta}+C_{\mathcal{R}}\log^{p_{\mathcal{R}}}\frac{2T}{\delta})}{T\sqrt{K}}.$$

It then remains to bound R, β, and M. By Lemma F.4.6, we have that

$$\nabla_{\boldsymbol{\Gamma}}\Phi(\boldsymbol{\Gamma})[\widetilde{\boldsymbol{\Gamma}}] = -\operatorname{tr}\left(\mathcal{H}(\boldsymbol{\Gamma}+\boldsymbol{\Gamma}_0)^{-1}\widetilde{\boldsymbol{\Gamma}}(\boldsymbol{\Gamma}+\boldsymbol{\Gamma}_0)^{-1}\right).$$

We can then compute the second derivative as, using Lemma F.4.6:

$$\begin{aligned}\nabla^2_{\boldsymbol{\Gamma}}\Phi(\boldsymbol{\Gamma})[\widetilde{\boldsymbol{\Gamma}},\bar{\boldsymbol{\Gamma}}] &= \frac{\mathrm{d}}{\mathrm{d}t}\left[-\operatorname{tr}\left(\mathcal{H}(\boldsymbol{\Gamma}+\boldsymbol{\Gamma}_0+t\bar{\boldsymbol{\Gamma}})^{-1}\widetilde{\boldsymbol{\Gamma}}(\boldsymbol{\Gamma}+\boldsymbol{\Gamma}_0+t\bar{\boldsymbol{\Gamma}})^{-1}\right)\right]\\ &= \operatorname{tr}\left(\mathcal{H}(\boldsymbol{\Gamma}+\boldsymbol{\Gamma}_0)^{-1}\bar{\boldsymbol{\Gamma}}(\boldsymbol{\Gamma}+\boldsymbol{\Gamma}_0)^{-1}\widetilde{\boldsymbol{\Gamma}}(\boldsymbol{\Gamma}+\boldsymbol{\Gamma}_0)^{-1}\right)\\ &\quad+\operatorname{tr}\left(\mathcal{H}(\boldsymbol{\Gamma}+\boldsymbol{\Gamma}_0)^{-1}\widetilde{\boldsymbol{\Gamma}}(\boldsymbol{\Gamma}+\boldsymbol{\Gamma}_0)^{-1}\bar{\boldsymbol{\Gamma}}(\boldsymbol{\Gamma}+\boldsymbol{\Gamma}_0)^{-1}\right).\end{aligned}$$

Recall that M is any bound on

$$\sup_{\boldsymbol{\Gamma}\in\widehat{\boldsymbol{\Omega}}_{\boldsymbol{\psi}}}\sup_{\boldsymbol{\tau}\in\mathcal{T}}|\langle\nabla_{\boldsymbol{\Gamma}}\Phi(\boldsymbol{\Gamma}),\boldsymbol{\psi}(\boldsymbol{\tau})\rangle|.$$

By the above computation of the gradient, we can bound this as

$$\begin{aligned}\sup_{\boldsymbol{\Gamma}\in\widehat{\boldsymbol{\Omega}}_{\boldsymbol{\psi}}}\sup_{\boldsymbol{\tau}\in\mathcal{T}}|\langle\nabla_{\boldsymbol{\Gamma}}\Phi(\boldsymbol{\Gamma}),\boldsymbol{\psi}(\boldsymbol{\tau})\rangle| &\le \sup_{\boldsymbol{\Gamma}\in\widehat{\boldsymbol{\Omega}}_{\boldsymbol{\psi}}}\sup_{\boldsymbol{\tau}\in\mathcal{T}}\left|\operatorname{tr}\left(\mathcal{H}(\boldsymbol{\Gamma}+\boldsymbol{\Gamma}_0)^{-1}\boldsymbol{\psi}(\boldsymbol{\tau})(\boldsymbol{\Gamma}+\boldsymbol{\Gamma}_0)^{-1}\right)\right|\\ &\le \|\mathcal{H}\|_{\mathrm{op}}\|\boldsymbol{\Gamma}_0^{-1}\|_{\mathrm{op}}^2\cdot\sup_{\boldsymbol{\tau}\in\mathcal{T}}\operatorname{tr}(\boldsymbol{\psi}(\boldsymbol{\tau}))\\ &\le D\|\mathcal{H}\|_{\mathrm{op}}\|\boldsymbol{\Gamma}_0^{-1}\|_{\mathrm{op}}^2.\end{aligned} \tag{6.6.2}$$

To bound β, by the Mean Value Theorem it suffices to bound the operator norm of $\nabla^2_{\boldsymbol{\Gamma}}\Phi(\boldsymbol{\Gamma})$. Using the expression above, we can bound this as

$$\begin{aligned}\sup_{\widetilde{\boldsymbol{\Gamma}},\bar{\boldsymbol{\Gamma}}\in\widehat{\boldsymbol{\Omega}}_{\boldsymbol{\psi}}}|\nabla^2_{\boldsymbol{\Gamma}}\Phi(\boldsymbol{\Gamma})[\boldsymbol{\psi}(\boldsymbol{\tau}_1),\boldsymbol{\psi}(\boldsymbol{\tau}_2)]| &\le 2\|\mathcal{H}\|_{\mathrm{op}}\|\boldsymbol{\Gamma}_0^{-1}\|_{\mathrm{op}}^3\cdot\sup_{\widetilde{\boldsymbol{\Gamma}},\bar{\boldsymbol{\Gamma}}\in\widehat{\boldsymbol{\Omega}}_{\boldsymbol{\psi}}}\operatorname{tr}(\widetilde{\boldsymbol{\Gamma}}\bar{\boldsymbol{\Gamma}})\\ &\le 2D^2\|\mathcal{H}\|_{\mathrm{op}}\|\boldsymbol{\Gamma}_0^{-1}\|_{\mathrm{op}}^3.\end{aligned}$$

Finally, it's straightforward to bound $R \le 2D$. Putting all of this together gives the result. □

Part II

Exploration in Reinforcement Learning

Chapter 7

Instance-Dependent Tabular Reinforcement Learning

We turn our attention now from settings with continuous states and actions, the focus of Part I, to settings with discrete states and actions. We first consider the simplest possible discrete state-action setting: tabular Markov Decision Processes (MDPs). In tabular MDPs, we assume there are a finite number of states, S, and actions, A, but that states and actions are otherwise unrelated (i.e. knowledge of the behavior of the environment at state s_1 provides no information about the behavior at state s_2). Our goal is to find a policy π—a mapping from states to actions—which maximizes some reward when played on the true environment. We focus on the pure exploration or PAC, setting, where the agent first explores the environment for some number steps and then uses the collect observations to determine an effective reward-maximizing policy.

In Part I we provided a precise characterization of a similar problem—learning an optimal decision—in the continuous control setting. There we obtained complexities that scaled with instance-dependent measures, quantifying the complexity in terms of the relationship between the curvature of our particular loss of interest, $\mathcal{H}(\theta_\star)$, and the covariates realizable on our particular system. More importantly, however, by developing a precise understanding of the complexity of learning, we were able to show which approaches were optimal (e.g. task-dependent exploration) and which weren't (e.g. persistent excitation).

We pursue a similar direction in the RL setting. We aim to obtain instance-dependent bounds on the complexity of learning, and use this characterization to elicit which learning strategies are most effective. While the RL setting, and particularly the tabular RL setting, has been given a significant amount of attention in the literature, prior to the work presented in this chapter, obtaining such instance-dependent bounds has proved elusive. Instead, the majority of the work has focused on obtaining minimax, worst-case complexity bounds which, in tabular RL, typically scale with S, A, and H, the horizon. In this chapter we present some of the first instead-dependent bounds on pure-exploration RL, and argue that existing minimax characterizations are insufficient for fully characterizing which approaches to learning in MDPs are optimal.

7.1 Preliminaries for Markov Decision Processes

We study finite-horizon, time inhomogeneous Markov Decision Processes denoted by the tuple $M = (\mathcal{S}, \mathcal{A}, H, \{P_h\}_{h=1}^H, P_0, \{R_h\}_{h=1}^H)$. Here $\mathcal{S}$ is the set of states ($S := |\mathcal{S}|$), $\mathcal{A}$ the set of actions ($A := |\mathcal{A}|$), H the horizon, $P_h : \mathcal{S} \times \mathcal{A} \to \triangle(\mathcal{S})$ the transition kernel at step h, $P_0 \in \triangle(\mathcal{S})$ the initial state distribution, and $R_h : \mathcal{S} \times \mathcal{A} \to \triangle([0,1])$ the reward distribution, with $r_h(s,a) = \mathbb{E}[R_h(s,a)]$. We assume that $\{P_h\}_{h=1}^H, P_0$, and $\{R_h\}_{h=1}^H$ are all initially unknown to the learner, and denote the true instance as $M^\star$.

An episode is a trajectory of state-action-reward triples, $\{(s_h, a_h, R_h)\}_{h=1}^H$, where $s_1 \sim P_0$, $s_{h+1} \sim P_h(\cdot|s_h, a_h)$, and $R_h \sim R_h(s_h, a_h)$. After H steps of interaction, the MDP restarts and the process repeats. A *policy* π is a mapping from states to actions: $\pi : \mathcal{S} \times [H] \to \triangle(\mathcal{A})$. $\pi_h(a|s)$ denotes the probability that π chooses a at (s,h). If for all (s,h), $\pi_h(a|s) = 1$ for some a, we say π is a *deterministic policy* and denote $\pi_h(s)$ the action it chooses at (s,h). Otherwise we say π is a *stochastic policy*.

Given a policy π, the Q-value function, $Q^\pi : \mathcal{S} \times \mathcal{A} \times [H] \to [0,H]$, denotes the expected reward obtained by playing action a in state s at time h, and then playing π for all subsequent time. Formally, it is defined as

$$Q_h^\pi(s,a) := \mathbb{E}_\pi \left[\sum_{h'=h}^{H} R_{h'}(s_{h'}, a_{h'}) | s_h = s, a_h = a \right].$$

We also define the value function, $V^\pi : \mathcal{S} \times [H] \to [0,H]$, as $V_h^\pi(s) := \mathbb{E}_{a\sim\pi_h(s)}[Q_h^\pi(s,a)]$. The Q-function satisfies the Bellman equation:

$$Q_h^\pi(s,a) = r_h(s,a) + \sum_{s'} P_h(s'|s,a) V_{h+1}^\pi(s').$$

We let $V_{H+1}^\pi(s) = 0$ and $Q_{H+1}^\pi(s,a) = 0$. We define the optimal Q-function as $Q_h^\star(s,a) := \sup_\pi Q_h^\pi(s,a)$, $V_h^\star(s) := \sup_\pi V_h^\pi(s)$, and let $\pi^\star$ denote an optimal policy. $V_0^\pi := \sum_s P_0(s) V_1^\pi(s)$ denotes the *value* of a policy, the expected reward it will obtain, and $V_0^\star := \sup_\pi V_0^\pi$.

Optimal Actions and Effective Gap. Critical to our analysis is the concept of a *suboptimality gap*. In particular, we will define the suboptimality gap as:

$$\Delta_h(s,a) := V_h^\star(s) - Q_h^\star(s,a).$$

In words, $\Delta_h(s,a)$ denotes the suboptimality of taking action a instead of an optimal action in (s,h), and then playing the optimal policy henceforth. We also let $\Delta_h^\pi(s,a) := \max_{a'} Q_h^\pi(s,a') - Q_h^\pi(s,a)$.

We say a is optimal at (s,h) if $\Delta_h(s,a) = 0$ (at least one such action is guaranteed to exist). We say a is the unique optimal action at (s,h) if $\Delta_h(s,a) = 0$, but $\Delta_h(s,a') > 0$ for all other $a \neq a'$, and say a is a non-unique optimal action if there exists another a' for which

$\Delta_h(s,a) = \Delta_h(s,a') = 0$. We say $\mathcal{M}$ has unique optimal actions if, for all (s,h), there is a unique optimal action a.

We now construct an effective gap $\widetilde{\Delta}_h(s,a)$ which coincides with $\Delta_h(s,a)$ for suboptimal actions, but is possibly non-zero if the optimal action is unique. Formally, at a particular (s,h), we denote the minimum non-zero gap as

$$\Delta_{\min}(s,h) := \min_{a:\Delta_h(s,a)>0} \Delta_h(s,a).$$

The effective gap is then defined as follows:

$$\widetilde{\Delta}_h(s,a) := \begin{cases} \Delta_h(s,a) & \Delta_h(s,a) > 0 \\ \Delta_{\min}(s,h) & a \text{ is the unique action at } s,h \text{ for which } \Delta_h(s,a) = 0 \\ 0 & a \text{ is a non-unique action at } s,h \text{ for which } \Delta_h(s,a) = 0. \end{cases}$$

Finally, we introduce the idea of a state-action visitation distribution. We define

$$w_h^\pi(s,a) := \mathbb{P}_\pi[s_h = s, a_h = a], \quad w_h^\pi(s) := \mathbb{P}_\pi[s_h = s].$$

Note that $w_h^\pi(s,a) = \pi_h(a|s) w_h^\pi(s)$. We denote the *maximum reachability* of a state s at time h by:

$$W_h(s) := \sup_\pi w_h^\pi(s).$$

In words, $W_h(s)$ is the maximum probability with which we could hope to reach s at time h.

PAC Reinforcement Learning Problem. The pure exploration setting considered in this chapter is typically referred to as *PAC RL* (probably approximately correct RL). Formally, in PAC RL, the goal is to, with probability $1-\delta$, identify a policy $\widehat{\pi}$ such that

$$V_0^\star - V_0^{\widehat{\pi}} \le \epsilon \tag{7.1.1}$$

using as few episodes as possible. We say that a policy satisfying (7.1.1) is ϵ-*optimal* and that an algorithm which returns a policy satisfying (7.1.1) with probability at least $1-\delta$ is (ϵ,δ)-PAC. Note that our goal is to find a single policy not a distribution over policies.

7.2 Near-Optimal Policy Identification in Tabular MDPs

Before stating our main result, we introduce our new notion of sample complexity for MDPs.

Definition 7.2.1 (Gap-Visitation Complexity)**.** For a given MDP M, we define the *gap-*

visitation complexity as:

$$\mathcal{C}(M,\epsilon) := \sum_{h=1}^{H} \inf_{\pi} \max_{s,a} \min\left\{ \frac{1}{w_h^\pi(s,a)\widetilde{\Delta}_h(s,a)^2}, \frac{W_h(s)^2}{w_h^\pi(s,a)\epsilon^2} \right\} + \frac{H^2|\mathrm{OPT}(\epsilon)|}{\epsilon^2}.$$

where the infimum is over all policies, both deterministic and stochastic, and:

$$\mathrm{OPT}(\epsilon) := \left\{(s,a,h) \;:\; \epsilon \geq W_h(s)\widetilde{\Delta}_h(s,a)/3\right\}.$$

In the special case when M has unique optimal actions, we define the *best-policy gap-visitation complexity* as:

$$\mathcal{C}^\star(M) := \sum_{h=1}^{H} \inf_{\pi} \max_{s,a} \frac{1}{w_h^\pi(s,a)\Delta_h(s,a)^2}.$$

Since $w_h^\pi(s,a) = \pi_h(a|s)w_h^\pi(s)$, as long as $w_h^\pi(s) > 0$ for some π, we can always choose our policy such that all actions are supported and $w_h^\pi(s,a) > 0$ for all a[1]. Recall that we have defined $\widetilde{\Delta}_h(s,a)$ so that $\widetilde{\Delta}_h(s,a) > 0$ for all a as long as (s,h) has a unique optimal action. This implies that as $\epsilon \to 0$, if M has unique optimal actions, $|\mathrm{OPT}(\epsilon)| \to 0$. Given this new notion of complexity, we are now ready to state our main result.

Theorem 7.1. *There exists an (ϵ,δ)-PAC algorithm, MOCA, which, with probability at least $1-\delta$, terminates after running for at most*

$$\mathcal{C}(M^\star,\epsilon)\cdot H^2 c_\epsilon \log\tfrac{1}{\delta} + \tfrac{C_{LOT}(\epsilon)}{\epsilon}$$

episodes and returns an ϵ-optimal policy, for lower-order term $C_{LOT}(\epsilon) = \mathrm{poly}(S,A,H,\log\frac{1}{\epsilon},\log\frac{1}{\delta})$ and $c_\epsilon = \mathrm{poly}\log(SAH/\epsilon)$. Furthermore, if

$$\epsilon < \epsilon^\star := \min\{\min_{s,a,h} W_h(s)\Delta_h(s,a)/3, 2H^2 S \min_{s,h} W_h(s)\}$$

and $M^\star$ has unique optimal actions, MOCA terminates after at most

$$\mathcal{C}^\star(M^\star)\cdot H^2 c_{\epsilon^\star} \log\tfrac{1}{\delta} + \tfrac{C_{LOT}(\epsilon^\star)}{\epsilon^\star}$$

episodes and returns $\pi^\star$, the optimal policy, with probability $1-\delta$.

In addition, MOCA is computationally efficient with computational cost scaling polynomially in problem parameters. We provide a proof of Theorem 7.1 and outline of the MOCA algorithm in Section 7.4.

[1]Here, we adopt the convention that, in the trivial case $W_h(s) = 0$ (and thus $w_h^\pi(s,a) = 0$), $\frac{W_h(s)^2}{w_h^\pi(s,a)\epsilon^2}$ evaluates to 0.

7.2.1 Interpreting the Complexity

Intuitively, the first term in the gap-visitation complexity quantifies how quickly we can eliminate all actions at least $\epsilon/W_h(s)$-suboptimal for all s and h, given that we must explore in our particular MDP. For a given s and h, if we play policy π for K episodes, we will reach (s,h) on average $w_h^\pi(s)K$ times. Thus, if we imagine that there is a bandit at (s,h), to eliminate action a will require that we run for at least $\frac{1}{w_h^\pi(s,a)\Delta_h(s,a)^2}$ episodes. The following result makes this rigorous—up to H factors, a complexity of $\mathcal{O}(\mathcal{C}^\star(M^\star)\cdot\log 1/\delta)$, which MOCA achieves, cannot be improved on in general for best-policy identification.

Proposition 7.1. *Fix some $S>1, A>1, H>1$, $\bar{h}\in[H]$, transition kernels $\{P_h\}_{h=1}^{\bar{h}-1}$, and gaps $\{\mathsf{gap}(s,a)\}_{s\in[S],a\in[A-1]}\subseteq(0,1/2)^{SA}$. Then there exists some MDP $M^\star$ with S states, A actions, horizon H, transition kernel P_h for $h\leq\bar{h}-1$, and gaps*

$$\Delta_{\bar{h}}(s,a)=\mathsf{gap}(s,a),\quad \forall s\in\mathcal{S},a\in\mathcal{A},a\neq\pi^\star_{\bar{h}}(s),\quad \Delta_h(s,a)\geq 1,\quad \forall s\in\mathcal{S},a\in\mathcal{A},h\neq\bar{h},$$

such that any $(0,\delta)$-PAC algorithm with stopping time K_δ requires:

$$\mathbb{E}_{M^\star}[K_\delta]\gtrsim\inf_\pi\max_{s,a}\frac{1}{w_{\bar{h}}^\pi(s,a)\Delta_{\bar{h}}(s,a)^2}\cdot\log\tfrac{1}{2.4\delta}.$$

In this instance, as $\Delta_h(s,a)\geq 1$ for $h\neq\bar{h}$, assuming $\{P_h\}_{h=1}^{\bar{h}-1}$ is chosen such that $W_h(s)$ is not too small for each s and $h\leq\bar{h}$, we will have that $\mathcal{C}^\star(M^\star)=\mathcal{O}(\inf_\pi\max_{s,a}\frac{1}{w_{\bar{h}}^\pi(s,a)\Delta_{\bar{h}}(s,a)^2})$, so Proposition 7.1 implies that we must have $\mathbb{E}_{M^\star}[K_\delta]\geq\Omega(\mathcal{C}^\star(M^\star)\cdot\log 1/\delta)$, matching the upper bound given in Theorem 7.1 up to H factors.

The second term in $\mathcal{C}(M^\star,\epsilon)$, $H^2|\mathrm{OPT}(\epsilon)|/\epsilon^2$, captures the complexity of ensuring that, after eliminating $\epsilon/W_h(s)$-suboptimal actions, sufficient exploration is performed to guarantee the returned policy is ϵ-optimal. While this will be no worse than H^3SA/ϵ^2, it could be much better, if in our MDP the number of (s,a,h) with $\widetilde{\Delta}_h(s,a)\lesssim\epsilon/W_h(s)$ is small (note that in the case when $M^\star$ has unique optimal actions, since $\widetilde{\Delta}_h(s,a)\geq\Delta_{\min}(s,h)$ by definition for all (s,a,h), $\mathrm{OPT}(\epsilon)$ will only contain states for which the minimum *non-zero* gap is less than $\epsilon/W_h(s)$). We next obtain the following bounds on $\mathcal{C}(M^\star,\epsilon)$, providing an interpretation of $\mathcal{C}(M^\star,\epsilon)$ in terms of the maximum reachability, and illustrating $\mathcal{C}(M^\star,\epsilon)$ is no larger than the minimax optimal complexity. This implies MOCA is nearly worst-case optimal, matching the lower bound of $\Omega(\frac{SAH^2}{\epsilon^2}\cdot\log 1/\delta)$ from Dann and Brunskill [2015] up to H and log factors[2].

Proposition 7.2. *The following bounds hold:*

1. $\mathcal{C}(M^\star,\epsilon)\leq\frac{H^3SA}{\epsilon^2}$

[2]This lower bound is for the stationary setting. As noted in Ménard et al. [2021], one would expect a lower bound of $\Omega(\frac{SAH^3}{\epsilon^2}\cdot\log 1/\delta)$ in the non-stationary setting, implying MOCA is H^2 off the lower bound.

2. $\mathcal{C}(M^\star, \epsilon) \leq \sum_{h=1}^{H} \sum_{s,a} \min\{\frac{1}{W_h(s)\widetilde{\Delta}_h(s,a)^2}, \frac{W_h(s)}{\epsilon^2}\} + \frac{H^2|\mathrm{OPT}(\epsilon)|}{\epsilon^2}$

3. $\mathcal{C}(M^\star, \epsilon) \leq \sum_{h=1}^{H} \sum_{s,a} \frac{1}{\epsilon \max\{\widetilde{\Delta}_h(s,a), \epsilon\}} + \frac{H^2|\mathrm{OPT}(\epsilon)|}{\epsilon^2}$.

In the special case of multi-armed and contextual bandits, the gap-visitation complexity simplifies considerably.

Proposition 7.3. *If $M^\star$ is a multi-armed bandit, then*

$$\mathcal{C}(M^\star, \epsilon) = \sum_a \min\left\{\frac{1}{\widetilde{\Delta}(a)^2}, \frac{1}{\epsilon^2}\right\}, \quad \mathcal{C}^\star(M^\star) = \sum_{a:\Delta(a)>0} \frac{1}{\Delta(a)^2}.$$

Furthermore, if $M^\star$ is a contextual bandit, then

$$\mathcal{C}^\star(M^\star) = \max_s \frac{1}{W(s)} \sum_a \frac{1}{\Delta(s,a)^2}.$$

The values given here are known to be the optimal problem-dependent constants for both best arm identification and (ϵ, δ)-PAC for multi-armed bandits [Kaufmann et al., 2016, Degenne and Koolen, 2019]. To our knowledge, the lower bound for best-policy identification in contextual bandits has never been formally stated, yet it is obvious it will take the form of $\mathcal{C}^\star(M^\star)$ given here. It follows that in the special cases of multi-armed bandits and contextual bandits, MOCA is instance-optimal, up to logarithmic factors and lower-order terms.

Several additional interpretations of the gap-visitation complexity are given in Appendix E.1.1. The above results show that the gap-visitation complexity cleanly interpolates between the worst-case optimal rate for (ϵ, δ)-PAC, and, in certain MDPs, the instance-optimal rate for best-policy identification. In between these extremes, it captures an intuitive sense of instance-dependence. As we will show in the following section, this instance-dependence can offer significant improvements over worst-case optimal approaches.

7.3 Low-Regret Algorithms are Suboptimal for PAC RL

Obtaining instance-dependent complexity guarantees of the form Theorem 7.1 typically requires exploring near-optimally—collecting maximally informative observations from the environment as quickly as possible. In this section, we seek to understand what implications Theorem 7.1 has on thinking about exploration in tabular RL.

Much of the existing literature on tabular RL relies on the principle of *optimism*—at every episode, actions are played which have the "best possible" value, given the current observations one has acquired. This approach is known to achieve sublinear regret and, via an *online-to-batch conversion*, can learn an ϵ-optimal policy at the minimax optimal rate. In this section, we seek to understand if such an approach is instance-optimal. Can it achieve the

same complexity as Theorem 7.1, and existing analysis is simply loose? Or is it fundamentally unable to achieve this measure?

Formally, we consider the family of all low-regret algorithms (optimistic algorithms, and beyond).

Definition 7.3.1 (Low-Regret Algorithm)**.** We say an algorithm $\mathcal{R}$ is a *low-regret algorithm* if it has expected regret bounded as $\text{Regret}(K) = \sum_{k=1}^{K} \mathbb{E}_{\mathcal{R}}[V_0^\star - V_0^{\pi_k}] \leq C_1 K^\alpha + C_2$, for some constants C_1, C_2, $\alpha \in (0,1)$, and where π_k is the policy $\mathcal{R}$ plays at episode k.

Protocol 7.3.1 (Low-Regret to PAC)**.** We consider the following procedure:

1. Learner runs low-regret algorithm $\mathcal{R}$ satisfying Definition 7.3.1 for K episodes, collects data $\mathfrak{D}_{\mathcal{R}}(K)$.

2. Using $\mathfrak{D}_{\mathcal{R}}(K)$ any way it wishes, the learner proposes a (possibly stochastic) policy $\widehat{\pi}$.

As a first example, consider the MDP in Figure 7.1. In state s_0, action a_1 is optimal and transitions to state s_1 with probability $1-p$ and state s_2 with probability p. Action a_2 is suboptimal and transitions to state s_2 with probability 1. To learn a good policy, we need to identify the optimal action in both s_1 and s_2. An optimistic or low-regret algorithm will primarily play a_1 in s_0, as this action is optimal, and it will therefore only reach s_2 approximately $\mathcal{O}(pK)$ times. It follows that a low-regret algorithm will take at least $\Omega(\frac{1}{p\Delta_2^2})$ episodes to learn the optimal action in s_2. In contrast, we could instead play a_2 in s_0, collecting less reward but learning the optimal action in s_2 in only $\Omega(\frac{1}{\Delta_2^2})$ episodes. For small p, this could be arbitrarily better. The following result makes this formal, illustrating that for identifying good policies in MDPs, existing low-regret and optimistic approaches can be highly suboptimal, and more intentional exploration procedures are needed.

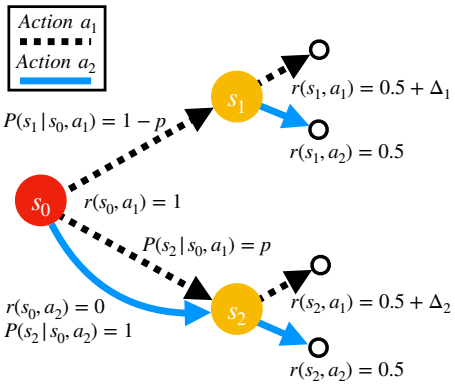

Figure 7.1: A motivating example

Proposition 7.4 (Informal)**.** *On the example in Figure 7.1, any algorithm following Protocol 7.3.1 must run for at least* $K \geq \Omega\big(\frac{\log 1/\delta}{\Delta_1^2} + \frac{\log 1/\delta}{p\Delta_2^2}\big)$ *episodes to identify the optimal policy, while* MOCA *will terminate and output the optimal policy after only* $K \leq \mathcal{O}\big(\frac{\log 1/\delta}{\Delta_1^2} + \frac{\log 1/\delta}{\Delta_2^2}\big)$ *episodes.*

Proposition 7.4 formally shows that, in this example, any low-regret algorithm is unable to achieve the complexity given in Theorem 7.1. We next consider a second example, to show that this effect extends to cases where $\epsilon > 0$.

Instance Class 7.3.1. Given a number of states $S \in \mathbb{N}$, consider an MDP with horizon $H = 2$, S states, and $S + 1$ actions, defined as in Figure 7.2.

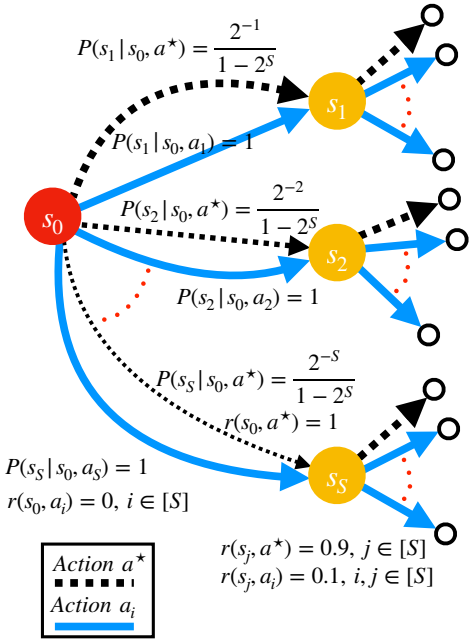

Figure 7.2: MDP from Instance Class 7.3.1

Similar to the example considered in Proposition 7.4, here $a^\star$ is the optimal action in every state, yet in state s_0, taking action a_i is much more informative. The following result shows that this structure results in poor performance for low-regret algorithms.

Proposition 7.5 (Informal)**.** *For the MDP in Instance Class 7.3.1 with S states and small enough ϵ, to find an ϵ-optimal policy with probability $1 - \delta$ any learner executing Protocol 7.3.1 with a low-regret algorithm satisfying Definition 7.3.1 must collect at least $\Omega(\frac{S \log 1/\delta}{\epsilon})$ episodes. In contrast, on this example $\mathcal{C}^\star(M^\star) = \mathcal{O}(S^2)$ and $\epsilon^\star = 1/3$, so, for $\epsilon \leq 1/3$, with probability $1 - \delta$, MOCA terminates and output $\pi^\star$ in $\widetilde{\mathcal{O}}(\mathrm{poly}(S))$ episodes.*

In particular, this example shows that there is an exponential separation between low-regret algorithms and MOCA. For exponentially small ϵ, learning the optimal policy following Protocol 7.3.1 takes $\widetilde{\Omega}(2^S)$ samples, yet MOCA finds the optimal policy in $\widetilde{\mathcal{O}}(\mathrm{poly}(S))$ samples.

Proposition 7.5, as well as Proposition 7.4, imply that the true complexity of finding a good policy is often much smaller than the complexity of finding a good policy *given that we explore to minimize regret*. As noted, the key piece in this example, and the example of Proposition 7.4, is that the optimal action in the initial state is very uninformative—if we want to learn the optimal action in a *subsequent* state, we should not take the optimal action in the initial state, but should instead take an action that leads us to the subsequent state with high probability. Nearly all existing works rely on algorithms which play policies which *converge to a good policy*. For instance-dependent PAC RL, instead of *playing* good policies, our examples show that an algorithm ought to explore efficiently, possibly taking very suboptimal actions in the process, and ultimately *recommending* a good policy. This shortcoming of greedy algorithms motivates our design of MOCA, where we seek to incorporate this insight.

While it is known that low-regret algorithms are minimax optimal for PAC RL, these instances show that running a low-regret algorithm and then an online-to-batch procedure is suboptimal by an arbitrarily large factor for PAC RL. We conclude that minimax optimality is far from being the complete story for PAC RL, and that if our goal is to simply identify a good policy, we can do much better than running a low-regret algorithm.

7.4 Algorithm and Proof of Theorem 7.1

We turn now to the definition of our algorithm, MOCA, and provide a proof of Theorem 7.1. We first provide some intuition for MOCA in Section 7.4.1 and give an overview of our core exploration procedure in Section 7.4.2, before stating the full definition of MOCA and giving the proof of Theorem 7.1 in Section 7.4.3. Omitted technical details from the proof are deferred to Appendix E.2.

7.4.1 Algorithm Intuition

At a high level, MOCA operates by treating every state as an individual bandit, and running an action elimination-style algorithm at each state [Even-Dar et al., 2006]. Unlike low-regret algorithms, MOCA aggressively directs its exploration to reach uncertain states as quickly as possible. The sequential structure of an MDP introduces several unique challenges, upon which we expand below.

Compounding Errors. In a standard bandit, from the perspective of the learner, the value of a particular action is determined solely by the environment. However, in an MDP, the value of an action a at state s and time h depends not only on the environment, but also on the actions the learner chooses to play in subsequent steps. If we run some policy $\widehat{\pi}$ after reaching (s, h), though we may be able to identify the optimal action to play at (s, h) *given that we then play* $\widehat{\pi}$, if $\widehat{\pi}$ is suboptimal, this action may also be suboptimal. The following result, a direct consequence of the celebrated performance-difference lemma [Kakade, 2003], is a key piece in our analysis, allowing us to effectively handle the compounding nature of errors, and may be of independent interest.

Proposition 7.6. *Assume that for each h and s, $\widehat{\pi}$ plays an action which satisfies $\max_a Q_h^{\widehat{\pi}}(s, a) - Q_h^{\widehat{\pi}}(s, \widehat{\pi}_h(s)) \le \epsilon_h(s)$. Then the suboptimality of $\widehat{\pi}$ is bounded as:*

$$V_0^\star - V_0^{\widehat{\pi}} \le \sum_{h=1}^{H} \sup_{\pi} \sum_{s} w_h^\pi(s)\epsilon_h(s).$$

In particular, if $\epsilon_h(s) \le \epsilon/H$ for all s and h, then we guarantee $V_0^\star - V_0^{\widehat{\pi}} \le \epsilon$. Proposition 7.6 in fact holds with $w_h^\pi(s)$ replaced by $w_h^\star(s)$, the visitation probability under the optimal policy. We choose to work instead with the (looser) bound stated in Proposition 7.6 as we do not in general know $\pi^\star$ and, as we will see, can more easily control the visitations under this "worst-case" policy.

Intuitively, Proposition 7.6 says that it is sufficient to learn an action in each state that performs well *as compared to the best action one could take given that $\widehat{\pi}$ is played in subsequent steps*. This motivates the basic premise of our algorithm. We proceed backwards, first learning near-optimal actions in every (s, H), which gives us $\widehat{\pi}_H$. We then continue on to level $H - 1$ where, after playing an action a, we play $\widehat{\pi}_H$. This gives us an unbiased estimate

of $Q^{\widehat{\pi}}_{H-1}(s,a)$, and allows us to determine actions that are near-optimal at stage $H-1$ if we play $\widehat{\pi}$ at stage H. We repeat this process backwards: at stage h, after playing action a, we play $\{\widehat{\pi}_{h'}\}_{h'=h+1}^{H}$, yielding an unbiased estimate of $Q^{\widehat{\pi}}_h(s,a)$, the "reward" of action a, and allowing us to eliminate actions that are suboptimal, given that we play $\widehat{\pi}$ in subsequent steps.

Our approach relies on a *Monte Carlo* estimate of the value of a particular action a at a given (s,h). Rather than attempting to compute this value using knowledge of the MDP, or relying on a bootstrapped estimator, we simply play the policy and observe the reward obtained. As the rewards are bounded, concentration applies, allowing us to efficiently estimate $Q^{\widehat{\pi}}_h(s,a)$, and turning the learning problem at a given (s,h) into nothing more than a bandit problem. We note that the Monte Carlo technique has previously proven useful for attaining refined gap-dependent guarantees in the regret setting [Xu et al., 2021].

Balancing Suboptimality and Reachability. To perform the above procedure efficiently, we must guarantee that we can reach every (s,h) enough times to eliminate suboptimal actions. Bear in mind the weighting of each suboptimality, $\epsilon_h(s)$, in Proposition 7.6: for a given (s,h), knowing an $\epsilon_h(s)$-optimal action in (s,h) will only add at most $\sup_\pi w^\pi_h(s)\epsilon_h(s) =: W_h(s)\epsilon_h(s)$ to the total suboptimality. Thus, we only need to learn good actions in each state in proportion to how easily that state may be reached.

In particular, if we play the policy achieving $w^\pi_h(s) = W_h(s)$ for K episodes, we will reach (s,h) $W_h(s)K$ times on average. By standard bandit sample complexities, we would expect it to take on order $\frac{A}{\epsilon_h(s)^2}$ samples to learn an $\epsilon_h(s)$-optimal action at (s,h), so it follows that the total number of episodes we would need to run would be $K \gtrsim \frac{A}{W_h(s)\epsilon_h(s)^2}$. However, if we set $\epsilon_h(s) \sim \beta\epsilon/W_h(s)$, which will ensure that the suboptimality of our policy is proportional to ϵ and does not scale with the reachability, we will only require $K \gtrsim \frac{AW_h(s)}{\beta^2\epsilon^2}$. We see then that the difficulty of reaching a state to explore it is balanced by the fact that such a state does not contribute significantly to the total suboptimality.

Navigating the MDP by Grouping States. Naively performing the above strategy could result in a sample complexity very suboptimal in its dependence on S. Indeed, to ensure our final policy is ϵ-suboptimal, we would need to choose $\beta \sim (SH)^{-1}$, since in this case we can only bound the suboptimality term from Proposition 7.6 as

$$\sum_{h=1}^{H} \sup_\pi \sum_s w^\pi_h(s)\epsilon_h(s) \lesssim SH\beta\epsilon.$$

This would give us a sample complexity scaling as a suboptimal S^3. To overcome this, we propose an exploration procedure which *groups states*—instead of exploring each state individually, in a given rollout it seeks to reach any number of states which are "nearby", in the sense that a single policy may reach any of them with similar probability.

To make this practical, we take inspiration from the algorithm of Zhang et al. [2020]—

designed for the so-called "reward-free" learning setting [Jin et al., 2020a], where the agent seeks merely to learn policies which traverse all reachable states—which is itself inspired by the classical RMAX algorithm [Brafman and Tennenholtz, 2002]. We modify the true reward function, giving a reward of "1" to any (s, a, h) pair we wish to visit, and otherwise setting the reward to "0". We then run a (variance-sensitive) regret minimizing algorithm, EULER [Zanette and Brunskill, 2019], on this modified reward function to generate a set of policies that can effectively traverse the MDP to visit the desired states. Critically, we show that the complexity of generating these policies amounts to a lower-order term—it is easier to learn to explore an MDP than to learn a good policy on it. Furthermore, grouping states allows us to obtain the optimal worst-case dependence on S and A.

7.4.2 Efficiently Exploring and Eliminating Actions

Before describing our main algorithm MOCA, we first outline the key sub-routines it utilizes.

7.4.2.1 `Learn2Explore` *Overview*

Our key exploration procedure is the `Learn2Explore` routine. `Learn2Explore` implements the navigation procedure described in Section 7.4.1. In particular, it takes as input a set $\mathcal{X} \subseteq \mathcal{S} \times \mathcal{A}$, and returns a partition $\{\mathcal{X}_j\}_j, \mathcal{X}_j \subseteq \mathcal{X}$, set of policies $\{\Pi_j\}_j$, and values $\{N_j\}_j$. These sets satisfy the following property.

Theorem 7.2 (Performance of `Learn2Explore`, informal)**.** *With high probability, the partition* $\{\mathcal{X}_j\}_j$ *returned by* `Learn2Explore` *satisfies*

$$\sup_{\pi} \sum_{(s,a)\in\mathcal{X}_j} w_h^{\pi}(s, a) \leq 2^{-j+1},$$

Moreover, the policy classes Π_j *are such that, by executing a single trajectory of each* $\pi \in \Pi_j$ *once, we visit every* $(s, a) \in \mathcal{X}_j$ *at least* $\frac{1}{2}N_j$ *times, where*

$$N_j = \mathcal{O}\left(\frac{2^{-j}|\Pi_j|}{|\mathcal{X}\setminus \cup_{j'=1}^{j-1} \mathcal{X}_{j'}|}\right), \quad |\Pi_j| = K_j(\tfrac{\delta}{\lceil \log(1/\epsilon_{\mathsf{L2E}})\rceil}, \delta_{\mathrm{samp}}) = \mathcal{O}(2^j S^3 A^2 H^4 \log^3 1/\delta) \quad (7.4.1)$$

Furthermore, if `Learn2Explore` *is run with tolerance* ϵ_{L2E}*, it will terminate after running for at most* $\mathrm{poly}(S, A, H, \log 1/\delta, \log 1/\epsilon_{\mathsf{L2E}}) \cdot \frac{1}{\epsilon_{\mathsf{L2E}}}$ *episodes.*

In other words, the sets $\mathcal{X}_j$ are groupings of "nearby" states that are increasingly difficult to reach, and the sets Π_j give a policy cover which navigates to each $(s, a) \in \mathcal{X}_j$. In addition, as $N_j = \mathcal{O}(2^{-j}|\Pi_j|/|\mathcal{X}\setminus \cup_{j'=1}^{j-1} \mathcal{X}_{j'}|)$, if we wish to collect n samples from each $(s, a) \in \mathcal{X}_j$, it will only require running for

$$\mathcal{O}\left(|\Pi_j| \cdot \frac{n}{2^{-j}|\Pi_j|/|\mathcal{X}\setminus \cup_{j'=1}^{j-1} \mathcal{X}_{j'}|}\right) = \mathcal{O}\left(2^j|\mathcal{X}\setminus \cup_{j'=1}^{j-1} \mathcal{X}_{j'}| \cdot n\right) \leq \mathcal{O}\left(2^j SAn\right)$$

Algorithm 7.1 `Learn2Explore`

function `Learn2Explore`(active set $\mathcal{X} \subseteq \mathcal{S} \times \mathcal{A}$, step h, confidence δ, sampling confidence δ_{samp}, tolerance $\epsilon_{\texttt{L2E}}$)
  if $|\mathcal{X}| = 0$ **then return** $\{(\emptyset, \emptyset, 0, 0)\}_{j=1}^{\lceil \log(1/\epsilon_{\texttt{L2E}}) \rceil}$
  for $j = 1, \dots, \lceil \log(1/\epsilon_{\texttt{L2E}}) \rceil$ **do**
    $K_j \leftarrow K_j(\frac{\delta}{\lceil \log(1/\epsilon_{\texttt{L2E}}) \rceil}, \delta_{\text{samp}})$ as defined in (7.4.1), $M_j \leftarrow |\mathcal{X}|$, $N_j \leftarrow K_j/(4|\mathcal{X}| \cdot 2^j)$
    $\mathcal{X}_j, \Pi_j \leftarrow$ `FindExplorableSets`$(\mathcal{X}, h, \delta, K_j, N_j)$
    $\mathcal{X} \leftarrow \mathcal{X} \backslash \mathcal{X}_j$
  return $\{(\mathcal{X}_j, \Pi_j, N_j, M_j)\}_{j=1}^{\lceil \log(1/\epsilon_{\texttt{L2E}}) \rceil}$

function `FindExplorableSets`(active set $\mathcal{X} \subseteq \mathcal{S} \times \mathcal{A}$, step h, confidence δ, epochs to run K, samples to collect N)
  Set $r_h^1(s,a) \leftarrow 1$ for $(s,a) \in \mathcal{X}$ and 0 otherwise, $N(s,a,h) \leftarrow 0$, $\mathcal{Y} \leftarrow \emptyset$, $\Pi \leftarrow \emptyset$, $j \leftarrow 1$
  for $k = 1, 2, \dots, K$ **do**
    `// Euler is as defined in Zanette and Brunskill [2019]`
    Run EULER on reward function r_h^j, get trajectory $\{(s_h^k, a_h^k, h)\}_{h=1}^H$ and policy π_k
    $N(s_h^k, a_h^k) \leftarrow N(s_h^k, a_h^k) + 1$, $\Pi \leftarrow \Pi \cup \pi_k$
    if $N(s_h^k, a_h^k) \geq N$, $(s_h^k, a_h^k) \in \mathcal{X}$, and $(s_h^k, a_h^k) \notin \mathcal{Y}$ **then**
      $\mathcal{Y} \leftarrow \mathcal{Y} \cup (s_h^k, a_h^k)$
      $r_h^{j+1}(s,a) \leftarrow 1$ for $(s,a) \in \mathcal{X} \backslash \mathcal{Y}$ and 0 otherwise
      $j \leftarrow j + 1$
      Restart EULER
  return $\mathcal{Y}, \Pi$

episodes. Thus, if we choose n so that it is proportional to the reachability of $\mathcal{X}_j$—for example, $n \sim 2^{-j}/\epsilon^2$—the total number of episodes that must be run to collect n samples is no more than $\mathcal{O}(\frac{SA}{\epsilon^2})$ (this can tightened to a term behaving in some cases as $\mathcal{O}(\frac{|\mathcal{X}_j|}{\epsilon^2})$). As we noted in Section 7.4.1, it suffices to collect samples from every state in proportion with its reachability, which, combined with this fact, allows our exploration to be performed efficiently. `Learn2Explore` is the backbone of our sample collection procedure and is called both in Line 4 of `Moca-SE` as well as in `CollectSamples`. We provide the full statement of Theorem 7.2 in Appendix E.3.

7.4.2.2 Helper Function Descriptions

We next turn to two additional functions utilized by MOCA—`CollectSamples`, which calls `Learn2Explore` in order to collect observations from the MDP, and `EliminateActions`, which takes as input the collected data and eliminates suboptimal actions.

Algorithm 7.2 MOCA Helper Functions

function `CollectSamples`(active set $\mathcal{X}$, allocation $\{n_j\}_{j=1}^{\lceil \log 1/\epsilon_{\mathsf{cs}} \rceil}$, step h, policy $\widehat{\pi}$, tolerance δ_{cs}, precision ϵ_{cs})
  $\{(\mathcal{X}_j, \Pi_j, N_j)\}_{j=1}^{\lceil \log 1/\epsilon_{\mathsf{cs}} \rceil} \leftarrow$ `Learn2Explore`$(\mathcal{X}, h, \delta_{\mathsf{cs}}, \frac{\delta_{\mathsf{cs}}}{\lceil \log 1/\epsilon_{\mathsf{cs}} \rceil \max_j n_j}, \epsilon_{\mathsf{cs}})$, $\mathfrak{D} \leftarrow \emptyset$
  for $j = 1, \dots, \lceil \log 1/\epsilon_{\mathsf{cs}} \rceil$ **do**
    for $\pi \in \Pi_j$ **do**
      Run π for $T = \lceil 2n_j/N_j \rceil$ times up to level h, then play $\widehat{\pi}$
      Collect reward rollouts $\mathfrak{D} \leftarrow \mathfrak{D} \cup \{s_h^t, a_h^t, \mathring{Q}_h^{\widehat{\pi},t}(s_h^t, a_h^t) := \sum_{h'=h}^{H} R_{h'}^t\}_{t=1}^{T}$
  return $\mathfrak{D}$, $\{\mathcal{X}_j\}_{j=1}^{\lceil \log 1/\epsilon_{\mathsf{cs}} \rceil}$

function `EliminateActions`(active set $\mathcal{X}$, partition $\{\mathcal{X}_j\}_{j=1}^{k}$, dataset $\mathfrak{D}$, active actions $\{\mathcal{A}_h(s)\}_{s \in \mathcal{Z}}$, level h, thresholds $\{\gamma_j\}_{j=1}^{k}$)
  for $(s,a) \in \mathcal{X}$ **do**
    $N_h(s,a) \leftarrow \sum_{(s_h^t, a_h^t, \mathring{Q}_h^{\widehat{\pi},t}(s_h^t,a_h^t)) \in \mathfrak{D}} \mathbb{I}\{(s_h^t, a_h^t) = (s,a)\}$
    $\widehat{Q}_h^{\widehat{\pi}}(s,a) \leftarrow \frac{1}{N_h(s,a)} \sum_{(s_h^t, a_h^t, \mathring{Q}_h^{\widehat{\pi},t}(s_h^t,a_h^t)) \in \mathfrak{D}} \mathbb{I}\{(s_h^t, a_h^t) = (s,a)\} \cdot \mathring{Q}_h^{\widehat{\pi},t}(s_h^t, a_h^t)$
  for $j = 1, \dots, k$ **do**
    for s s.t. $\exists a$ with $(s,a) \in \mathcal{X}_j$ **do**
      $j(s) \leftarrow \arg\max_{j'} j'$ s.t. $\exists a', (s,a') \in \mathcal{X}_{j'}$
      $\mathcal{A}_h(s) \leftarrow \{a \in \mathcal{A}_h(s) \ : \ \max_{a' \in \mathcal{A}_h(s)} \widehat{Q}_h^{\widehat{\pi}}(s,a') - \widehat{Q}_h^{\widehat{\pi}}(s,a) \le \gamma_{j(s)}\}$
  return $\{\mathcal{A}_h(s)\}_{s \in \mathcal{Z}}$

Description of `CollectSamples`. `CollectSamples` takes as input a set $\mathcal{X} \subseteq \mathcal{S} \times \mathcal{A}$, an allocation $\{n_j\}_j$, a timestep h, and a policy $\widehat{\pi}$. In short, `CollectSamples` first calls `Learn2Explore` on $\mathcal{X}$ to obtain a partition $\{\mathcal{X}_j\}_j$, and then reruns the policies returned by `Learn2Explore` enough times to ensure that every $(s,a) \in \mathcal{X}_j$ is reached at least n_j times at timestep h. After reaching (s,a,h), $\widehat{\pi}$ is played, to obtain a Monte Carlo estimate $\mathring{Q}_h^{\widehat{\pi},t}(s,a)$ of $Q_h^{\widehat{\pi}}(s,a)$. `CollectSamples` then returns the data collected and the partition returned by `Learn2Explore`.

Description of `EliminateActions`. `EliminateActions` takes as input a set $\mathcal{X} \subseteq \mathcal{S} \times \mathcal{A}$, a partition of this set $\{\mathcal{X}_j\}_j$, a dataset $\mathfrak{D}$ generated by `CollectSamples`, a set of active actions $\{\mathcal{A}_h(s)\}_s$, a timestep h, and a threshold $\{\gamma_j\}_j$. For each $(s,a) \in \mathcal{X}$, it forms an estimate of $Q_h^{\widehat{\pi}}(s,a)$ from the rollouts in $\mathfrak{D}$. Given these estimates, for s such that there exists a with $(s,a) \in \mathcal{X}_j$, it removes actions from $\mathcal{A}_h(s)$ that are more than $\gamma_{j(s)}$-suboptimal.

7.4.3 Detailed Algorithm Description and Proof of Theorem 7.1

Finally, we outline our main algorithm MOCA and provide a proof of Theorem 7.1. We will argue the correctness of MOCA on two events: $\mathcal{E}_{\text{exp}}$, which is the event that every call

to `Learn2Explore` succeeds, and $\mathcal{E}_{\text{est}}$, which is the event that the estimates calculated in `EliminateActions` are accurate. We can think of $\mathcal{E}_{\text{exp}}$ as the event on which we *explore* successfully—we reach every state the desired number of times—and $\mathcal{E}_{\text{est}}$ the event on which we *estimate* correctly—our Monte Carlo estimates of $Q_h^{\widehat{\pi}}(s,a)$ concentrate. The following lemma shows that these events hold with high probability.

Lemma 7.4.1. *If we run* MOCA, $\mathbb{P}[\mathcal{E}_{\text{exp}} \cap \mathcal{E}_{\text{est}}] \geq 1 - \delta_{\text{tol}}$.

We formally define these events in Appendix E.2, and provide proofs to the lemmas in this section.

7.4.3.1 `Moca-SE` *Overview*

Given this description of `Learn2Explore`, we are ready to describe the `Moca-SE` (single-epoch MOCA) procedure. Assume that we run `Moca-SE` with tolerance ϵ and confidence δ. We begin by calling `Learn2Explore` on Line 4, which allows us to form an estimate of $W_h(s)$, the maximum reachability of (s,h). This in turn allows us to determine which states are efficiently reachable. We let $\mathcal{Z}_h$ denote the set of all such efficiently reachable states at stage h: $W_h(s) \geq \frac{\epsilon}{2H^2S}, \forall s \in \mathcal{Z}_h$. All other states have little effect on the performance of any policy and can henceforth be ignored. The following claim shows that our estimate of $W_h(s)$ is in fact accurate for $s \in \mathcal{Z}_h$.

Claim 1. *If running* `Moca-SE`*, on* $\mathcal{E}_{\text{est}} \cap \mathcal{E}_{\text{exp}}$ *we have* $\widehat{W}_h(s) \leq W_h(s) \leq 32\widehat{W}_h(s)$ *for all* $s \in \mathcal{Z}_h$.

We then proceed to our main loop over h in Line 7. For a fixed h, we loop over i and form the partition $\mathcal{Z}_{hi}$ which contains all $s \in \mathcal{Z}_h$ with $\widehat{W}_h(s) \sim 2^{-i}$. Given $\mathcal{Z}_{hi}$, we next loop over ℓ, and for each ℓ aim to eliminate actions from $\mathcal{Z}_{hi}$ that are more than $\epsilon_\ell = H2^{-\ell}$-suboptimal. We define $\mathcal{Z}_{hi}^\ell \subseteq \mathcal{S} \times \mathcal{A}$ as the set of (s,a) for $s \in \mathcal{Z}_{hi}$, and a we have not yet determined are $\epsilon_{\ell-1}/W_h(s)$-suboptimal. To collect a sufficient number of samples from each $(s,a) \in \mathcal{Z}_{hi}^\ell$ in order to eliminate suboptimal actions, we run `CollectSamples` on $\mathcal{Z}_{hi}^\ell$ and seek to collect $n_{ij}^\ell = \mathcal{O}(H^2/(2^{2i}\epsilon_\ell^2)) = \mathcal{O}(H^2 W_h(s)^2/\epsilon_\ell^2)$ from each $(s,a) \in \mathcal{Z}_{hi}^\ell$.

Note that every $(s,a) \in \mathcal{Z}_{hi}^\ell$ has similar maximum reachability, $W_h(s) \sim 2^{-i}$, determined by index i. Nevertheless, as outlined in Section 7.4.1, to obtain the proper scaling in S, we may still need to group states in a way that allows nearby states to be explored effectively. Calling `Learn2Explore` in `CollectSamples` does just this, efficiently traversing the MDP to guarantee enough samples are collected from all states in tandem.

After running `CollectSamples`, we run `EliminateActions` to eliminate suboptimal actions, yielding a set of candidate $\epsilon_\ell/W_h(s)$-suboptimal actions for each (s,h), denoted $\mathcal{A}_h^\ell(s)$. The following result shows that this procedure does indeed winnow out sufficiently suboptimal actions.

Algorithm 7.3 Monte Carlo Action Elimination - Single Epoch (Moca-SE(ϵ, δ, FinalRound))

input: tolerance ϵ, confidence δ, final round flag FinalRound
initialize $\epsilon_{\text{exp}} \leftarrow \frac{\epsilon}{2H^2S}$, $\mathcal{Z}_h \leftarrow \emptyset$, $\varsigma_{\text{exp}} = \lceil \log \frac{1}{\epsilon_{\text{exp}}} \rceil$
for each (s,h) **do** // loop over all s,h to learn maximum reachability
  $\{(\mathcal{X}_j^{sh}, \Pi_j^{sh}, N_j^{sh})\}_{j=1}^{\varsigma_{\text{exp}}} \leftarrow$ Learn2Explore$(\{(s,a)\}, h, \frac{\delta}{SH}, \frac{1}{2}, \epsilon_{\text{exp}})$ for arbitrary $a \in \mathcal{A}$
  if $\mathcal{X}_j^{sh} = \{(s,a)\}$ for $j \in [\varsigma_{\text{exp}}]$ **then** $\widehat{W}_h(s) \leftarrow \frac{N_j^{sh}}{2|\Pi_j^{sh}|} = \frac{1}{16 \cdot 2^j}$, $\mathcal{Z}_h \leftarrow \mathcal{Z}_h \cup \{s\}$
set $\varsigma_\epsilon \leftarrow \lceil \log \frac{64}{H^2 S \epsilon} \rceil$, $\varsigma_\delta \leftarrow \log \frac{SAH\varsigma_\epsilon(\ell_\epsilon+1)}{\delta}$, $\ell_\epsilon \leftarrow \lceil \log \frac{H}{\epsilon} \rceil$, $\widehat{\pi}_h(s) \leftarrow$ arbitrary action, $\mathcal{A}_h^0(s) \leftarrow \mathcal{A}$.
for $h = H, H-1, \ldots, 1$ **do** // loop over horizon
  for $i = 1, 2, \ldots, \varsigma_\epsilon$ **do** // loop over estimated maximum reachability
    $\mathcal{Z}_{hi} \leftarrow \{s \in \mathcal{Z}_h \ : \ \widehat{W}_h(s) \in [2^{-i}, 2^{-i+1}]\}$
    for $\ell = 1, \ldots, \ell_\epsilon$ **do** // loop over tolerance ϵ_ℓ
      $\epsilon_\ell \leftarrow H2^{-\ell}$, $\mathcal{Z}_{hi}^\ell \leftarrow \{(s,a) \ : \ s \in \mathcal{Z}_{hi}, a \in \mathcal{A}_h^{\ell-1}(s), |\mathcal{A}_h^{\ell-1}(s)| > 1\}$
      $n_{ij}^\ell \leftarrow \frac{2^{18}H^2\varsigma_\delta}{2^{2i}\epsilon_\ell^2}$, $\gamma_{ij}^\ell \leftarrow \frac{2^i\epsilon_\ell}{2^8}$ for $j = 1, \ldots, \varsigma_\epsilon$
      $\mathfrak{D}_{hi}^\ell, \{\mathcal{X}_{hij}^\ell\}_{j=1}^{\varsigma_\epsilon} \leftarrow$ CollectSamples$(\mathcal{Z}_{hi}^\ell, \{n_{ij}^\ell\}_{j=1}^{\varsigma_\epsilon}, h, \widehat{\pi}, \frac{\delta}{H\varsigma_\epsilon\ell_\epsilon}, \frac{\epsilon_{\text{exp}}}{32})$
      $\{\mathcal{A}_h^\ell(s)\}_{s \in \mathcal{Z}_{hi}} \leftarrow$ EliminateActions$(\mathcal{Z}_{hi}^\ell, \{\mathcal{X}_{hij}^\ell\}_{j=1}^{\varsigma_\epsilon}, \mathfrak{D}_{hi}^\ell, \{\mathcal{A}_h^{\ell-1}(s)\}_{s \in \mathcal{Z}_{hi}}, h, \{\gamma_{ij}^\ell\}_{j=1}^{\varsigma_\epsilon})$
  if FinalRound is true **then** // ensure $\widehat{\pi}$ ϵ-optimal
    $\mathcal{Z}_h^{\ell_\epsilon+1} \leftarrow \{(s,a) \ : \ s \in \mathcal{Z}_h, a \in \mathcal{A}_h^{\ell_\epsilon}(s), |\mathcal{A}_h^{\ell_\epsilon}(s)| > 1\}$
    $n_j^{\ell_\epsilon+1} \leftarrow \frac{64H^4\varsigma_\delta\varsigma_\epsilon^2 2^{2(-j+1)}}{\epsilon^2}$, $\gamma_j^{\ell_\epsilon+1} \leftarrow \frac{\epsilon}{4H\varsigma_\epsilon 2^{-j+1}}$ for $j = 1, \ldots, \varsigma_\epsilon$
    $\mathfrak{D}_h^{\ell_\epsilon+1}, \{\mathcal{X}_{hj}^{\ell_\epsilon+1}\}_{j=1}^{\varsigma_\epsilon} \leftarrow$ CollectSamples$(\mathcal{Z}_h^{\ell_\epsilon+1}, \{n_j^{\ell_\epsilon+1}\}_{j=1}^{\varsigma_\epsilon}, h, \widehat{\pi}, \frac{\delta}{H}, \frac{\epsilon_{\text{exp}}}{32})$
    $\{\mathcal{A}_h^{\ell_\epsilon+1}(s)\}_{s \in \mathcal{Z}_h^{\ell_\epsilon+1}} \leftarrow$ EliminateActions$(\mathcal{Z}_h^{\ell_\epsilon+1}, \{\mathcal{X}_{hj}^{\ell_\epsilon+1}\}_{j=1}^{\varsigma_\epsilon}, \mathfrak{D}_h^{\ell_\epsilon+1}, \{\mathcal{A}_h^{\ell_\epsilon}(s)\}_{s \in \mathcal{Z}_h^{\ell_\epsilon+1}}, h, \{\gamma_j^{\ell_\epsilon+1}\}_{j=1}^{\varsigma_\epsilon})$
  else
    $\mathcal{A}_h^{\ell_\epsilon+1}(s) \leftarrow \mathcal{A}_h^{\ell_\epsilon}(s)$ for all $s \in \mathcal{Z}_h$
  Set $\widehat{\pi}_h(s)$ to any action in $\mathcal{A}_h^{\ell_\epsilon+1}(s)$ for all $s \in \mathcal{Z}_h$
return $\widehat{\pi}$, $\max_{s,h} |\mathcal{A}_h^{\ell_\epsilon+1}(s)|$

Lemma 7.4.2. *On the event $\mathcal{E}_{\text{est}} \cap \mathcal{E}_{\text{exp}}$, if* Moca-SE *is run with tolerance ϵ, for any $h \in [H]$ and $\ell \in [\ell_\epsilon + 1]$, if $|\mathcal{A}_h^\ell(s)| = 1$, then for $a \in \mathcal{A}_h^\ell(s)$,*

$$\max_{a'} Q_h^{\widehat{\pi}}(s,a') - Q_h^{\widehat{\pi}}(s,a) = 0.$$

Furthermore, if $|\mathcal{A}_h^\ell(s)| > 1$, $\ell \le \ell_\epsilon$, and $s \in \mathcal{Z}_{hi}$ for some i, then any $a \in \mathcal{A}_h^\ell(s)$ satisfies

$$\Delta_h(s,a) \le \frac{3\epsilon_\ell}{2W_h(s)}.$$

Finally, if $|\mathcal{A}_h^{\ell_\epsilon+1}(s)| > 1$ *and* $s \in \mathcal{Z}_h$*, then any* $a \in \mathcal{A}_h^{\ell_\epsilon+1}(s)$ *satisfies*

$$\Delta_h^{\widehat{\pi}}(s,a) \le \frac{\epsilon}{2H\varsigma_\epsilon \cdot 2^{-j(s)+1}}$$

where $j(s) = \arg\max_j j$ *s.t.* $\exists a', (s,a') \in \mathcal{X}_{hj}^{\ell_\epsilon+1}$.

The guarantee follows by verifying that our exploration collects enough samples to ensure the confidence intervals on $Q_h^{\widehat{\pi}}(s,a)$ have width $\mathcal{O}(\epsilon_\ell / W_h(s))$. Furthermore, using properties of `Learn2Explore` given in Theorem 7.2, we can bound the sample complexity of this procedure, which yields a dominant term reminiscent of our sample complexity measure, $\mathcal{C}(M^\star, \epsilon)$ in Definition 7.2.1.

Lemma 7.4.3. *On the event* $\mathcal{E}_{\text{est}} \cap \mathcal{E}_{\text{exp}}$*, for a given* h *and* i*, the loop over* ℓ *on Line 10 of* `Moca-SE` *will take at most*

$$cH^2\varsigma_\delta\varsigma_\epsilon^2\ell_\epsilon \inf_\pi \max_{s\in\mathcal{Z}_{hi}} \max_a \min\left\{\frac{1}{w_h^\pi(s,a)\widetilde{\Delta}_h(s,a)^2}, \frac{W_h(s)^2}{w_h^\pi(s,a)\epsilon^2}\right\}$$

episodes. Furthermore, the total complexity of calling `Moca-SE` *with* `FinalRound = False` *is bounded by:*

$$H^2c\varsigma_\delta\varsigma_\epsilon^3\ell_\epsilon \cdot \sum_{h=1}^H \inf_\pi \max_{s,a} \min\left\{\frac{1}{w_h^\pi(s,a)\widetilde{\Delta}_h(s,a)^2}, \frac{W_h(s)^2}{w_h^\pi(s,a)\epsilon^2}\right\} + \frac{\mathrm{poly}(S,A,H,\log 1/\epsilon, \log 1/\delta)}{\epsilon}$$

for a universal constant c.

Proof Sketch of Lemma 7.4.3. In order to collect at least n_{ij}^ℓ samples from each $(s,a) \in \mathcal{X}_{hij}^\ell$, Theorem 7.2 ensures that it suffices to run for $\mathcal{O}(|\Pi_j| n_{ij}^\ell / N_j) \approx \mathcal{O}(2^j |\mathcal{X}_{hij}^\ell| n_{ij}^\ell)$ episodes; thus, we can collect n_{ij}^ℓ samples from each $(s,a) \in \mathcal{Z}_{hi}^\ell$ with only $\mathcal{O}(\sum_j 2^j |\mathcal{X}_{hij}^\ell| n_{ij}^\ell)$ episodes.

Theorem 7.2 also shows that $\mathcal{X}_{hij}^\ell$ satisfies

$$\sup_\pi \min_{(s,a)\in\mathcal{X}_{hij}^\ell} |\mathcal{X}_{hij}^\ell| w_h^\pi(s,a) \le \sup_\pi \sum_{(s,a)\in\mathcal{X}_{hij}^\ell} w_h^\pi(s,a) \le 2^{-j+1},$$

which upper bounds the $\mathcal{O}(\sum_j 2^j |\mathcal{X}_{hij}^\ell| n_{ij}^\ell)$ episodes required by $\mathcal{O}(\sum_j \inf_\pi \max_{(s,a)\in\mathcal{X}_{hij}^\ell} \frac{n_{ij}^\ell}{w_h^\pi(s,a)})$. As all $(s,a), (s',a') \in \mathcal{X}_{hij}^\ell$ satisfy $W_h(s) \approx W_h(s')$ by construction, $n_{ij}^\ell = \mathcal{O}(H^2 W_h(s)^2 / \epsilon_\ell^2)$ for any $(s,a) \in \mathcal{X}_{hij}^\ell$, so the sample complexity reduces to $\mathcal{O}(H^2 \sum_j \inf_\pi \max_{(s,a)\in\mathcal{X}_{hij}^\ell} \frac{W_h(s)^2}{w_h^\pi(s,a)\epsilon_\ell^2})$. Finally, since actions in stage ℓ are only active if their gap is less than $W_h(s)/\epsilon_\ell$, we obtain Lemma 7.4.3. □

The `FinalRound` flag. Single-epoch MOCA is called multiple times by our main algorithm (Algorithm 7.4), each with geometrically decreasing tolerance ϵ. For all but the smallest such ϵ, `Moca-SE` is run with `FinalRound` = `False`, and terminates after the previously described loop over h, i, ℓ terminates. The last call to `Moca-SE` constitutes the "final round", where we set `FinalRound` = `True`; this calls `CollectSamples` and `EliminateActions` one more time for each h.

While the loop with the `FinalRound` = `False` is able to eliminate suboptimal actions, it does not shrink the action set enough to guarantee that the returned policy is ϵ-optimal. In particular, while each (s, h) pair upon entering this final-round loop is sub-optimal by at most $\epsilon_h(s) = \mathcal{O}(\epsilon/W_h(s))$, Proposition 7.6 suggests that we actually need $\epsilon_h(s) \leq \mathcal{O}(\epsilon/H \cdot \sup_\pi \sum_{s' \in \mathcal{X}} w_h^\pi(s'))$. To remedy this, `FinalRound` = `True` invokes a final step to ensure the latter bound holds. Critically, while in the previous step we only sampled (s, a) in proportion with $W_h(s)^2$, the individual maximum reachability of that state, in this step we sample each (s, a) in proportion with the *reachability of the partition containing* (s, a). *This subtlety is indispensable for attaining our instance-dependent sample complexity.*

In other words, after forming our set $\mathcal{Z}_h^{\ell_\epsilon+1}$ of active states and actions corresponding to the minimal error-resolution index $\ell = \ell_\epsilon$ (from the previous argument, this will only contain states we have not determined the optimal action for and actions that satisfy $\Delta_h(s, a) \leq \frac{3\epsilon}{2W_h(s)}$) and partitioning it into $\{\mathcal{X}_{hj}^{\ell_\epsilon+1}\}_j$ by calling `Learn2Explore`, we seek to collect $\mathcal{O}(H^4 2^{-2j}/\epsilon^2)$ from every $(s, a) \in \mathcal{X}_{hj}^{\ell_\epsilon+1}$. By Theorem 7.2, $\mathcal{X}_{hj}^{\ell_\epsilon+1}$ satisfies $\sup_\pi \sum_{(s,a)\in\mathcal{X}_{hj}^{\ell_\epsilon+1}} w_h^\pi(s, a) \leq 2^{-j+1}$, so sampling (s, a) $\mathcal{O}(H^4 2^{-2j}/\epsilon^2)$ times means we sample it in proportion to its group reachability squared. By the final conclusion of Lemma 7.4.2, we can cleanly bound the suboptimality of actions remaining after this step and, as the following lemma shows, the number of samples used by this procedure.

Lemma 7.4.4. *On the event $\mathcal{E}_{\text{est}} \cap \mathcal{E}_{\text{exp}}$, if* `Moca-SE` *is called with* `FinalRound` = `True`*, the procedure within the if statement on Line 15 will terminate after collecting at most*

$$\frac{cH^4 \varsigma_\delta \varsigma_\epsilon^2}{\epsilon^2} |\mathcal{Z}_h^{\ell_\epsilon+1}| + \frac{\text{poly}(S, A, H, \log 1/\delta, \log 1/\epsilon)}{\epsilon}$$

episodes. Furthermore, the total complexity of calling `Moca-SE` *with* `FinalRound` = `True` *is bounded by:*

$$H^2 c \varsigma_\delta \varsigma_\epsilon^3 \ell_\epsilon \cdot \mathcal{C}(M^\star, \epsilon) + \frac{\text{poly}(S, A, H, \log 1/\epsilon, \log 1/\delta)}{\epsilon}$$

for a universal constant c.

Critically, as noted above, $\mathcal{Z}_h^{\ell_\epsilon+1}$ will only contain near-optimal actions and unsolved states, so its cardinality could be much less than SA. Finally, a simple calculation combining Lemma 7.4.2 and Proposition 7.6 gives the following result.

Lemma 7.4.5. *On the event $\mathcal{E}_{\text{est}} \cap \mathcal{E}_{\text{exp}}$, if* `Moca-SE` *is run with tolerance ϵ and* `FinalRound` $=$ `true`*, then the policy $\widehat{\pi}$ returned by* `Moca-SE` *is ϵ-suboptimal.*

7.4.3.2 Putting everything together: MOCA and proving Theorem 7.1

Algorithm 7.4 MOnte **C**arlo **A**ction Elimination (MOCA)

input: tolerance ϵ_{tol}, confidence δ_{tol}
$\mathcal{A}_h^0(s) \leftarrow \mathcal{A}$ for all s, h
for $m = 1, \dots, \lceil \log(H/\epsilon_{\text{tol}}) \rceil - 1$ **do**
  $\epsilon_{\text{tol}(m)} \leftarrow H2^{-m}, \delta_{\text{tol}(m)} \leftarrow \frac{\delta_{\text{tol}}}{36m^2}$
  $\widehat{\pi}^m$, `MaxOpt` $\leftarrow$ `Moca-SE`$(\epsilon_{\text{tol}(m)}, \delta_{\text{tol}(m)},$ `false`$)$
  if `MaxOpt` $= 1$ **then**
    return $\widehat{\pi}^m$
$\widehat{\pi}$, `MaxOpt` $\leftarrow$ `Moca-SE`$(\epsilon_{\text{tol}}, \frac{\delta_{\text{tol}}}{36\lceil \log(H/\epsilon_{\text{tol}}) \rceil^2},$ `true`$)$
return $\widehat{\pi}$

We turn now to our main algorithm, MOCA. MOCA takes as input a tolerance ϵ_{tol} and confidence δ_{tol}. Were our goal simply to find an ϵ_{tol}-optimal policy, from the above argument we could call `Moca-SE` with tolerance ϵ_{tol} and `FinalRound` $=$ `True`. However, if ϵ_{tol} is small enough that `Moca-SE` identifies the optimal action in *every* state, this may result in overexploring—since once we have identified the optimal action in every state we can terminate and output the optimal policy. To remedy this, we instead call `Moca-SE` with exponentially decreasing tolerance and `FinalRound` $=$ `False`. If it returns a set of actions for every s, h with $|\mathcal{A}_h(s)| = 1$, we can guarantee we have identified the optimal policy, and simply terminate without overexploring. Note also in this stage, since `FinalRound` $=$ `False`, we do not pay for the $\widetilde{\mathcal{O}}(\frac{H^4}{\epsilon^2}|\mathcal{Z}_h^{\ell_\epsilon+1}|)$ term. If this condition is never met, we simply call `Moca-SE` a final time at the end with `FinalRound` $=$ `True` to ensure the policy we return is ϵ_{tol}-optimal. A formal proof of Theorem 7.1 follows directly from this argument.

Proof of Theorem 7.1. Note that $\mathbb{P}[\mathcal{E}_{\text{est}} \cap \mathcal{E}_{\text{exp}}] \geq 1 - \delta$ by Lemma 7.4.1. We will assume for the remainder of the proof that this event holds.

Case 1: $\epsilon_{\text{tol}} \geq \min\{\min_{s,a,h} W_h(s)\widetilde{\Delta}_h(s,a)/3, 2H^2 S \min_{s,h} W_h(s)\}$**.** In this case, that the policy returned is ϵ_{tol}-optimal is guaranteed by Lemma 7.4.5 since the final call to `Moca-SE` is run with `FinalRound` $=$ `true`. To bound the sample complexity, we can then simply combine Lemma 7.4.3 and Lemma 7.4.4, which gives that the total sample complexity is bounded as (using that $\epsilon_{\text{tol}(m)} \geq \epsilon_{\text{tol}}$ and that $\delta_{\text{tol}(m)} \geq \delta_{\text{tol}}/(36\lceil \log H/\epsilon_{\text{tol}} \rceil^2) =: \delta'$):

$$\sum_{m=1}^{\lceil \log H/\epsilon_{\text{tol}} \rceil - 1} H^2 c\varsigma_{\delta_{\text{tol}(m)}} \varsigma^3_{\epsilon_{\text{tol}(m)}} \ell_{\epsilon_{\text{tol}(m)}} \cdot \sum_{h=1}^{H} \inf_{\pi} \max_{s,a} \min \left\{ \frac{1}{w_h^\pi(s,a)\widetilde{\Delta}_h(s,a)^2}, \frac{W_h(s)^2}{w_h^\pi(s,a)\epsilon^2_{\text{tol}(m)}} \right\}$$

$$+ H^2 c\varsigma_{\delta_{\mathrm{tol}(m)}}\varsigma^3_{\epsilon_{\mathrm{tol}}}\ell_{\epsilon_{\mathrm{tol}}} \cdot \sum_{h=1}^{H} \inf_{\pi} \max_{s,a} \min\left\{\frac{1}{w_h^\pi(s,a)\widetilde{\Delta}_h(s,a)^2}, \frac{W_h(s)^2}{w_h^\pi(s,a)\epsilon_{\mathrm{tol}}^2}\right\}$$
$$+ \frac{cH^4\varsigma_{\delta_{\mathrm{tol}}}\varsigma^2_{\epsilon_{\mathrm{tol}}}|\mathrm{OPT}(\epsilon_{\mathrm{tol}})|}{\epsilon_{\mathrm{tol}}^2} + \frac{\lceil \log H/\epsilon_{\mathrm{tol}}\rceil \cdot \mathrm{poly}(S,A,H,\log 1/\epsilon_{\mathrm{tol}},\log 1/\delta_{\mathrm{tol}})}{\epsilon_{\mathrm{tol}}}.$$

This can be upper bounded as

$$\lceil \log H/\epsilon_{\mathrm{tol}}\rceil \cdot H^2 c\varsigma_{\delta'}\varsigma^3_{\epsilon_{\mathrm{tol}}}\ell_{\epsilon_{\mathrm{tol}}} \cdot \sum_{h=1}^{H} \inf_{\pi} \max_{s,a} \min\left\{\frac{1}{w_h^\pi(s,a)\widetilde{\Delta}_h(s,a)^2}, \frac{W_h(s)^2}{w_h^\pi(s,a)\epsilon_{\mathrm{tol}}^2}\right\}$$
$$+ \frac{cH^4\varsigma_{\delta'}\varsigma^2_{\epsilon_{\mathrm{tol}}}|\mathrm{OPT}(\epsilon_{\mathrm{tol}})|}{\epsilon_{\mathrm{tol}}^2} + \frac{\mathrm{poly}(S,A,H,\log 1/\epsilon_{\mathrm{tol}},\log 1/\delta_{\mathrm{tol}})}{\epsilon_{\mathrm{tol}}}.$$

This and the definition of $\mathcal{C}(M^\star,\epsilon)$ gives the first conclusion of Theorem 7.1.

Case 2: $\epsilon_{\mathrm{tol}} < \min\{\min_{s,a,h} W_h(s)\widetilde{\Delta}_h(s,a)/3, 2H^2S\min_{s,h} W_h(s)\}$**.** As we showed in the proof of Lemma 7.4.4, we will have that $\mathcal{Z}_h^{\ell_\epsilon+1} \subseteq \mathrm{OPT}(\epsilon,h)$. Therefore, if for all (s,a), $\widetilde{\Delta}_h(s,a) > 3\epsilon/W_h(s)$, we will have that $|\mathcal{Z}_h^{\ell_\epsilon+1}| = 0$, which implies that for every $s \in \mathcal{Z}_h$, $|\mathcal{A}_h^{\ell_\epsilon}(s)| = 1$. Furthermore, on $\mathcal{E}_{\mathrm{exp}}$, we will have that $\mathcal{Z}_h = \mathcal{S}\times\mathcal{A}$ if $\frac{\epsilon}{2H^2S} < \min_s W_h(s)$. If each of these conditions hold for all h, then the returned sets $\mathcal{A}_h^{\ell_\epsilon+1}(s)$ will satisfy $|\mathcal{A}_h^{\ell_\epsilon+1}(s)|$ for all s and h.

It follows then that if $\epsilon_{\mathrm{tol}} < \min\{\min_{s,a,h} W_h(s)\widetilde{\Delta}_h(s,a)/3, 2H^2S\min_{s,h} W_h(s)\}$, either $\epsilon_{\mathrm{tol}(m)} < \min\{\min_{s,a,h} W_h(s)\widetilde{\Delta}_h(s,a)/3, 2H^2S\min_{s,h} W_h(s)\}$ for some m, in which case the above condition will be met, and the termination criteria on Line 6 of MOCA will be satisfied, or

$$\epsilon_{\mathrm{tol}(m)} \geq \min\{\min_{s,a,h} W_h(s)\widetilde{\Delta}_h(s,a)/3, 2H^2S\min_{s,h} W_h(s)\},$$

and MOCA will reach the final call of `Moca-SE` with `FinalRound` = `True`. In the former case, letting $\bar{m}$ denote the value of m at which MOCA terminates, the total sample complexity will be bounded as, using the same argument as in Case 1,

$$\sum_{m=1}^{\bar{m}} H^2 c\varsigma_{\delta_{\mathrm{tol}(\bar{m})}}\varsigma^3_{\epsilon_{\mathrm{tol}(\bar{m})}}\ell_{\epsilon_{\mathrm{tol}(\bar{m})}} \cdot \sum_{h=1}^{H} \inf_{\pi} \max_{s,a} \min\left\{\frac{1}{w_h^\pi(s,a)\widetilde{\Delta}_h(s,a)^2}, \frac{W_h(s)^2}{w_h^\pi(s,a)\epsilon_{\mathrm{tol}(m)}^2}\right\}$$
$$+ \frac{\bar{m}\cdot\mathrm{poly}(S,A,H,\log 1/\epsilon_{\mathrm{tol}(\bar{m})},\log 1/\delta_{\mathrm{tol}(\bar{m})})}{\epsilon_{\mathrm{tol}(\bar{m})}}$$
$$\leq \bar{m}H^2 c\varsigma_{\delta_{\mathrm{tol}(\bar{m})}}\varsigma^3_{\epsilon_{\mathrm{tol}(\bar{m})}}\ell_{\epsilon_{\mathrm{tol}(\bar{m})}} \cdot \sum_{h=1}^{H} \inf_{\pi} \max_{s,a} \min\left\{\frac{1}{w_h^\pi(s,a)\widetilde{\Delta}_h(s,a)^2}, \frac{W_h(s)^2}{w_h^\pi(s,a)\epsilon_{\mathrm{tol}(\bar{m})}^2}\right\}$$
$$+ \frac{\mathrm{poly}(S,A,H,\log 1/\epsilon_{\mathrm{tol}(\bar{m})},\log 1/\delta_{\mathrm{tol}(\bar{m})})}{\epsilon_{\mathrm{tol}(\bar{m})}}$$

and note that $\epsilon_{\mathrm{tol}(\bar{m}-1)} \geq \min\{\min_{s,a,h} W_h(s)\widetilde{\Delta}_h(s,a)/3, 2H^2 S \min_{s,h} W_h(s)\}$, since we did not terminate at round $\bar{m}-1$, implying that $\epsilon_{\mathrm{tol}(\bar{m})} \geq 2\min\{\min_{s,a,h} W_h(s)\widetilde{\Delta}_h(s,a)/3, 2H^2 S \min_{s,h} W_h(s)\}$. Note also that $\delta_{\mathrm{tol}(\bar{m})} = \frac{\delta}{36\log^2 \epsilon_{\mathrm{tol}(\bar{m})}}$, so we can also bound

$$\log 1/\delta_{\mathrm{tol}(\bar{m})} \leq \mathcal{O}(\log 1/\delta_{\mathrm{tol}} + \log\log(2\min\{\min_{s,a,h} W_h(s)\widetilde{\Delta}_h(s,a)/3, 2H^2 S \min_{s,h} W_h(s)\})).$$

Together these give the bound stated in Theorem 7.1.

In the latter case, when we do not terminate early at Line 6, the same sample complexity bound applies but with $\epsilon_{\mathrm{tol}(\bar{m})}$ replaced by ϵ_{tol}, since if $|\mathcal{Z}_h^{\ell_\epsilon+1}| = 0$, the final call to `CollectSamples` in Line 18 of `Moca-SE` will not collect any samples. As before, in this case we can lower bound

$$\epsilon_{\mathrm{tol}} \geq 2\min\{\min_{s,a,h} W_h(s)\widetilde{\Delta}_h(s,a)/3, 2H^2 S \min_{s,h} W_h(s)\}$$

from which the bound follows.

It remains to show that $\widehat{\pi} = \pi^\star$. This follows inductively from Lemma 7.4.2 since if $|\mathcal{A}_H^\ell(s)| = 1$, this implies that for $a \in \mathcal{A}_H^\ell(s)$, $a = \pi_H^\star(s)$. Then if we assume that $\widehat{\pi}_{h'}(s) = \pi_{h'}^\star(s)$ for all s and $h' > h$, if $|\mathcal{A}_h^\ell(s)| = 1$ this implies that for $a \in \mathcal{A}_h^\ell(s)$, $a = \pi_h^\star(s)$ since, by Lemma 7.4.2, in this case

$$\max_{a'} Q_h^{\widehat{\pi}}(s,a') - Q_h^{\widehat{\pi}}(s,a) = 0$$

but $Q_h^{\widehat{\pi}}(s,a'') = Q_h^\star(s,a'')$. Thus, it follows that $\widehat{\pi} = \pi^\star$, which completes the proof.

□

Chapter 8

Instance-Dependent Linear Reinforcement Learning

While a canonical setting in the reinforcement learning literature, tabular MDPs are severely limited in what real-world environments they can conveniently represent. In real-world settings, it is rarely the case that states and actions are completely unrelated—more often, there is sufficient structure in the problem that knowing the behavior at one state provides information on the behavior at another state. Ignoring this information-sharing, as the tabular assumption does, results in sample complexities significantly larger than we might otherwise achieve.

Inspired by this, much work has been done over the last several years to move beyond the tabular assumption and develop sample efficient RL algorithms in settings that require function approximation [Jiang et al., 2017, Jin et al., 2020b, Zhou et al., 2021a, Du et al., 2021, Jin et al., 2021, Foster et al., 2021]. A particular setting that has received significant attention is that of MDPs with *linear* function approximation—where it is assumed that the MDP can be parametrized linearly in some featurization of the state-action space [Yang and Wang, 2019, Jin et al., 2020b, Wang et al., 2019, Du et al., 2019, Zanette et al., 2020a,b, Ayoub et al., 2020, Jia et al., 2020, Weisz et al., 2021, Zhou et al., 2021a,b, Zhang et al., 2021c, Wang et al., 2021]. In such settings, the number of states could be infinite—making learning in a tabular representation of such an MDP intractable—yet the linear structure allows information to be shared between states, making learning tractable when this structure is taken into account. While significant progress has been made understanding the worst-case complexity of RL with linear function approximation, little progress has been made understanding the instance-dependent complexity of learning.

In this chapter we provide the first significant results in this direction, extending the results of Chapter 7 to MDPs with linear function approximation. As in Chapter 7 we consider the PAC setting, where the goal is to identify an ϵ-optimal policy in the minimum amount of samples possible. While there exist several formulations of MDPs with linear function approximation, we consider in particular the *linear MDP* setting of Jin et al. [2020b]. As we will see, though our results are similar in spirit to that of Chapter 7, the algorithmic techniques are significantly different—relying, in fact, on the same online experiment design

techniques as in Chapter 6—and the final complexity differs significantly as well.

8.1 Reinforcement Learning with Linear Function Approximation

Unless otherwise noted, we adopt the same MDP notation in this chapter as in Chapter 7, though we make a slight modification to the setting of Chapter 7 by assuming that each episode begins from the same fixed state instead of being drawn from some distribution. Here we consider MDPs with linear structure, defined as follows.

Definition 8.1.1 (Linear MDPs [Jin et al., 2020b])**.** We say that an MDP is a *d-dimensional linear MDP*, if there exists some (known) feature map $\boldsymbol{\phi}(s,a) : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$, H (unknown) signed vector-valued measures $\boldsymbol{\mu}_h \in \mathbb{R}^d$ over $\mathcal{S}$, and H (unknown) reward vectors $\boldsymbol{\theta}_h \in \mathbb{R}^d$, such that:

$$P_h(\cdot|s,a) = \langle \boldsymbol{\phi}(s,a), \boldsymbol{\mu}_h(\cdot) \rangle, \quad \mathbb{E}[R_h(s,a)] = \langle \boldsymbol{\phi}(s,a), \boldsymbol{\theta}_h \rangle.$$

We will assume $\|\boldsymbol{\phi}(s,a)\|_2 \leq 1$ for all s, a; and for all h, $\||\boldsymbol{\mu}_h|(\mathcal{S})\|_2 = \| \int_{s \in \mathcal{S}} |\mathrm{d}\boldsymbol{\mu}_h(s)| \|_2 \leq \sqrt{d}$ and $\|\boldsymbol{\theta}_h\|_2 \leq \sqrt{d}$.

Linear MDPs encompass, for example, tabular MDPs, but can also model more complex settings, such as feature spaces corresponding to the simplex [Jin et al., 2020b], or the linear bandit problem. Critically, linear MDPs allow for infinite state-spaces, as well as generalization across states—rather than learning the behavior in particular states, we can learn in the d-dimensional feature space. Note that the standard definition of linear MDPs, for example as given in Jin et al. [2020b], assumes the rewards are deterministic, while we assume the rewards are random but that their means are linear. We still assume, however, that the random rewards, $r_h(s,a)$, are contained in $[0,1]$ almost surely.

For a given policy π, we define the *feature-visitation* at step h, the expected feature vector policy π encounters at step h, as $\boldsymbol{\phi}_{\pi,h} := \mathbb{E}_\pi[\boldsymbol{\phi}(s_h, a_h)]$. Note that this is a direct generalization of state-visitations in tabular RL—if our MDP is in fact tabular, $[\boldsymbol{\phi}_{\pi,h}]_{s,a} = \mathbb{P}_\pi[s_h = s, a_h = a]$, so the feature visitation vector corresponds directly to the state visitations. Note also that we can write the value of a policy as $V_0^\pi = \sum_{h=1}^H \langle \boldsymbol{\phi}_{\pi,h}, \boldsymbol{\theta}_h \rangle$. Denote the average feature vector induced by π in a particular state s as $\boldsymbol{\phi}_{\pi,h}(s) = \mathbb{E}_{a \sim \pi_h(\cdot|s)}[\boldsymbol{\phi}(s,a)]$. We also let $\boldsymbol{\phi}_{h,\tau} := \boldsymbol{\phi}(s_{h,\tau}, a_{h,\tau})$ denote the feature vector encountered at step h of episode τ (and similarly define $R_{h,\tau}$). We define $\boldsymbol{\Lambda}_{\pi,h} := \mathbb{E}_\pi[\boldsymbol{\phi}(s_h,a_h)\boldsymbol{\phi}(s_h,a_h)^\top]$, the expected covariance of policy π at step h, $\bar{\lambda}_{\min,h} = \sup_\pi \lambda_{\min}(\boldsymbol{\Lambda}_{\pi,h})$ the largest achievable minimum eigenvalue at step h, and $\bar{\lambda}_{\min} = \min_h \bar{\lambda}_{\min,h}$. We will make the following assumption.

Assumption 8.1 (Full Rank Covariates)**.** *In our MDP,* $\bar{\lambda}_{\min} > 0$.

We remark that Assumption 8.1 is analogous to other explorability assumptions found in the RL with function approximation literature [Zanette et al., 2020c, Hao et al., 2021, Agarwal et al., 2021].

To reduce uncertainty in directions of interest, we will be interested in optimizing over the set of all realizable covariance matrices on our particular MDP. To this end, define

$$\mathbf{\Omega}_h := \{\mathbb{E}_{\pi\sim\omega}[\mathbf{\Lambda}_{\pi,h}] \ : \ \omega \in \mathbf{\Omega}_\pi\} \tag{8.1.1}$$

for $\mathbf{\Omega}_\pi$ the set of all valid distributions over Markovian policies (both deterministic and stochastic). $\mathbf{\Omega}_h$ is, then, the set of all covariance matrices realizable by distributions over policies at step h.

While in Chapter 7 we consider state-action gaps, in this chapter the more relevant notion of a gap is that of a *policy* gap. For some set of policies Π (which may not contain an optimal policy), we let $V_0^\star(\Pi) := \max_{\pi\in\Pi} V_0^\pi$. We then define the minimum policy gap as:

$$\Delta_{\min}^{\Pi} := \begin{cases} V_0^\star(\Pi) - \max_{\pi\in\Pi: V_0^\pi < V_0^\star(\Pi)} V_0^\pi & |\{\pi \in \Pi : V_0^\pi = V_0^\star(\Pi)\}| = 1 \\ 0 & \text{o.w.} \end{cases}$$

That is, $\Delta_{\min}^{\Pi}$ is the gap between the values of the best policy and next best policy in Π if the best policy in Π is unique, and otherwise 0.

8.2 Near-Optimal Policy Identification in Linear MDPs

We are now ready to state our algorithm, PEDEL.

PEDEL Description. PEDEL is a *policy-elimination*-style algorithm. It takes as input some set of policies, Π, and proceeds in epochs, maintaining a set of *active* policies, Π_ℓ, such that all $\pi \in \Pi_\ell$ are guaranteed to satisfy $V_0^\pi \geq V_0^\star(\Pi) - 4\epsilon_\ell$, for $\epsilon_\ell = 2^{-\ell}$. After running for $\lceil \log \frac{4}{\epsilon} \rceil$ epochs, it returns any of the remaining active policies, which will be guaranteed to have value at least $V_0^\star(\Pi) - \epsilon$.

Similar to Chapter 7, in order to ensure Π_ℓ only contains $4\epsilon_\ell$-optimal policies, sufficient exploration must be performed at every epoch to refine the estimate of each policy's value. In particular, one can show that, if we have collected some covariates $\mathbf{\Lambda}_{h,\ell}$, the uncertainty in our estimate of the value of policy π at step h scales as $\|\widehat{\boldsymbol{\phi}}_{\pi,h}^\ell\|_{\mathbf{\Lambda}_{h,\ell}^{-1}}$, for $\widehat{\boldsymbol{\phi}}_{\pi,h}^\ell$ the estimated feature-visitation for policy π at epoch ℓ. To reduce our uncertainty at each round, we would therefore like to collect covariates such that $\|\widehat{\boldsymbol{\phi}}_{\pi,h}^\ell\|_{\mathbf{\Lambda}_{h,\ell}^{-1}} \lesssim \epsilon_\ell$. Collecting covariates which satisfy this using the minimum number of episodes of exploration possible involves solving the experiment design:

$$\inf_{\mathbf{\Lambda}_{\text{exp}}\in\mathbf{\Omega}_h} \max_{\pi\in\Pi_\ell} \|\widehat{\boldsymbol{\phi}}_{\pi,h}^\ell\|_{\mathbf{\Lambda}_{\text{exp}}^{-1}}^2. \tag{8.2.1}$$

Note that this design has the form of an XY-experiment design [Soare et al., 2014]. Solving (8.2.1) will produce covariance $\mathbf{\Lambda}_{\text{exp}}$ which reduces uncertainty in relevant feature directions. However, to solve this design we require knowledge of which covariance matrices are realizable

Algorithm 8.1 **P**olicy Learning via **E**xperiment **De**sign in **L**inear MDPs (PEDEL)

input: tolerance ϵ, confidence δ, policy set Π
$\ell_0 \leftarrow \lceil \log_2 \frac{d^{3/2}}{H} \rceil$, $\Pi_{\ell_0} \leftarrow \Pi$, $\widehat{\boldsymbol{\phi}}^1_{\pi,1} \leftarrow \mathbb{E}_{a\sim\pi_1(\cdot|s_1)}[\boldsymbol{\phi}(s_1,a)], \forall \pi \in \Pi$
for $\ell = \ell_0, \ell_0+1, \ldots, \lceil \log \frac{4}{\epsilon} \rceil$ **do**
  $\epsilon_\ell \leftarrow 2^{-\ell}$, $\beta_\ell \leftarrow 64H^4 \log \frac{4H^2|\Pi_\ell|\ell^2}{\delta}$
  for $h = 1, 2, \ldots, H$ **do**
    Solve (8.2.1) by running procedure of Theorem 8.5 with parameters

$$\epsilon_{\text{exp}} \leftarrow \frac{\epsilon_\ell^2}{\beta_\ell}, \quad \delta \leftarrow \frac{\delta}{2H\ell^2}, \quad \underline{\lambda} \leftarrow \log \frac{4H^2|\Pi_\ell|\ell^2}{\delta}, \quad \Phi \leftarrow \Phi_{h,\ell} := \{\widehat{\boldsymbol{\phi}}^\ell_{\pi,h} : \pi \in \Pi_\ell\}$$

    in order to collect data $\{(s_{h,\tau}, a_{h,\tau}, R_{h,\tau}, s_{h+1,\tau})\}_{\tau=1}^{K_{h,\ell}}$, satisfying

$$\max_{\pi\in\Pi_\ell} \|\widehat{\boldsymbol{\phi}}^\ell_{\pi,h}\|^2_{\boldsymbol{\Lambda}^{-1}_{h,\ell}} \le \frac{\epsilon_\ell^2}{\beta_\ell} \quad \text{for} \quad \boldsymbol{\Lambda}_{h,\ell} \leftarrow \sum_{\tau=1}^{K_{h,\ell}} \boldsymbol{\phi}(s_{h,\tau}, a_{h,\tau})\boldsymbol{\phi}(s_{h,\tau}, a_{h,\tau})^\top + \frac{1}{d}\cdot I$$

    for $\pi \in \Pi_\ell$ **do** `// Estimate feature-visitations for active policies`
      $\widehat{\boldsymbol{\phi}}^\ell_{\pi,h+1} \leftarrow \Big(\sum_{\tau=1}^{K_{h,\ell}} \boldsymbol{\phi}_{\pi,h+1}(s_{h+1,\tau})\boldsymbol{\phi}_{h,\tau}^\top \boldsymbol{\Lambda}^{-1}_{h,\ell}\Big)\widehat{\boldsymbol{\phi}}^\ell_{\pi,h}$
    $\widehat{\boldsymbol{\theta}}^\ell_h \leftarrow \boldsymbol{\Lambda}^{-1}_{h,\ell}\sum_{\tau=1}^{K_{h,\ell}} \boldsymbol{\phi}_{h,\tau} r_{h,\tau}$ `// Estimate reward vectors`
  `// Remove provably suboptimal policies from active policy set`

$$\Pi_{\ell+1} \leftarrow \Pi_\ell \backslash \Big\{\pi \in \Pi_\ell \ : \ \widehat{V}_0^\pi < \sup_{\pi'\in\Pi_\ell} \widehat{V}_0^{\pi'} - 2\epsilon_\ell\Big\} \quad \text{for} \quad \widehat{V}_0^\pi := \textstyle\sum_{h=1}^H \langle \widehat{\boldsymbol{\phi}}^\ell_{\pi,h}, \widehat{\boldsymbol{\theta}}^\ell_h\rangle$$

  if $|\Pi_{\ell+1}| = 1$ **then return** $\pi \in \Pi_{\ell+1}$
return any $\pi \in \Pi_{\ell+1}$

on our particular MDP. In general we do not know the MDP's dynamics, and therefore do not have access to this knowledge. To overcome this and solve (8.2.1), we apply a technique very similar to the DYNAMICOED technique employed in Chapter 6.

Estimating Feature-Visitations. We remark briefly on the estimation of the feature-visitations on Line 8. If we assume that $\{\boldsymbol{\phi}_{h,\tau}\}_{\tau=1}^{K_{h,\ell}}$ is fixed and that all randomness is due to $s_{h+1,\tau}$, then it is easy to see that, using the structure present in linear MDPs as given in Definition 8.1.1,

$$\begin{aligned}\mathbb{E}[\textstyle\sum_{\tau=1}^{K_{h,\ell}} \boldsymbol{\phi}_{\pi,h+1}(s_{h+1,\tau})\boldsymbol{\phi}_{h,\tau}^\top \boldsymbol{\Lambda}^{-1}_{h,\ell}] &= \textstyle\sum_{\tau=1}^{K_{h,\ell}} \big(\int \boldsymbol{\phi}_{\pi,h+1}(s)\mathrm{d}\boldsymbol{\mu}_h(s)^\top \boldsymbol{\phi}_{h,\tau}\big)\boldsymbol{\phi}_{h,\tau}^\top \boldsymbol{\Lambda}^{-1}_{h,\ell} \\ &= \textstyle\int \boldsymbol{\phi}_{\pi,h+1}(s)\mathrm{d}\boldsymbol{\mu}_h(s)^\top.\end{aligned}$$

By Definition 8.1.1, we have $\boldsymbol{\phi}_{\pi,h+1} = \left(\int \boldsymbol{\phi}_{\pi,h+1}(s)\mathrm{d}\boldsymbol{\mu}_h(s)^\top\right)\boldsymbol{\phi}_{\pi,h}$. Comparing these, we see that our estimator of $\boldsymbol{\phi}_{\pi,h+1}$ on Line 8 is (conditioned on $\{\boldsymbol{\phi}_{h,\tau}\}_{\tau=1}^{K_{h,\ell}}$) unbiased, assuming $\widehat{\boldsymbol{\phi}}^{\ell}_{\pi,h} \approx \boldsymbol{\phi}_{\pi,h}$.

8.2.1 Main Results

We have the following result on the performance of PEDEL.

Theorem 8.1. *Consider running PEDEL with some set of Markovian policies Π on any linear MDP satisfying Definition 8.1.1 and Assumption 8.1. Then with probability at least $1-\delta$, PEDEL outputs a policy $\widehat{\pi} \in \Pi$ such that $V_0^{\widehat{\pi}} \geq V_0^\star(\Pi) - \epsilon$, and runs for at most*

$$C_0 H^4 \cdot \sum_{h=1}^{H} \inf_{\boldsymbol{\Lambda}_{\mathrm{exp}} \in \boldsymbol{\Omega}_h} \max_{\pi \in \Pi} \frac{\|\boldsymbol{\phi}_{\pi,h}\|^2_{\boldsymbol{\Lambda}_{\mathrm{exp}}^{-1}}}{\max\{V_0^\star(\Pi) - V_0^\pi, \Delta^{\Pi}_{\mathrm{min}}, \epsilon\}^2} \cdot \left(\log|\Pi| + \log\frac{1}{\delta}\right) + C_1$$

episodes, with $C_0 = \log\frac{1}{\epsilon} \cdot \mathrm{poly}\log(H, \log\frac{1}{\epsilon})$, $C_1 = \mathrm{poly}\big(d, H, \frac{1}{\lambda_{\mathrm{min}}}, \log\frac{1}{\delta}, \log\frac{1}{\epsilon}, \log|\Pi|\big)$, and $\boldsymbol{\Omega}_h$ the set of covariance matrices realizable on our MDP, as defined in (8.1.1).

The proof of Theorem 8.1 is given in Section 8.3. Theorem 8.1 quantifies, in a precise instance-dependent way, the complexity of identifying a policy $\widehat{\pi}$ with value at most a factor of ϵ from the value of the optimal policy in Π. At a high level, the complexity measure can be thought of as quantifying the difficulty of exploring the MDP so as to eliminate suboptimal policies. $\|\boldsymbol{\phi}_{\pi,h}\|_{\boldsymbol{\Lambda}_{\mathrm{exp}}^{-1}}$ quantifies the difficulty of collecting data necessary to eliminate policy π (in particular, $\boldsymbol{\phi}_{\pi,h}$ is the direction we need to explore to learn about the performance of policy π, and $\boldsymbol{\Lambda}_{\mathrm{exp}}^{-1}$ quantifies how easily we can reach directions in the MDP to do this), while $V_0^\star(\Pi) - V_0^\pi$ quantifies how suboptimal policy π is, and therefore how many samples are needed to distinguish it from the optimal policy in the class. Thus, rather than scaling only with factors such as d and ϵ, our complexity scales with instance-dependent quantities—the covariance matrices we can obtain and the feature vectors we expect to observe on our particular MDP, and the policy gaps on our MDP. Our complexity measure has a close resemblance to the complexity measure for best-arm identification in linear bandits found in Fiez et al. [2019], but generalizes it to problems with long horizon where navigation is required.

Theorem 8.1 holds for an arbitrary set of policies, yet, in general, we are interested in learning a policy which has value within a factor of ϵ of the value of the *optimal* policy on the MDP, $V_0^\star$. Such a guarantee is immediately attainable by applying Theorem 8.1 with a policy set Π such that $\sup_{\pi\in\Pi} V_0^\pi \geq V_0^\star - \epsilon$. The following result shows that it is possible to construct such a set of policies, and therefore learn a *globally* near-optimal policy.

Corollary 8.2. *There exists a set of policies Π_ϵ such that $\log|\Pi_\epsilon| \leq \widetilde{\mathcal{O}}(dH \cdot \log 1/\epsilon)$ and, for any linear MDP satisfying Definition 8.1.1, $\sup_{\pi\in\Pi_\epsilon} V_0^\pi \geq V_0^\star - \epsilon$. If we run PEDEL with*

$\Pi \leftarrow \Pi_\epsilon$, *then with probability at least* $1-\delta$, *it returns a policy* $\widehat{\pi}$ *such that* $V_0^{\widehat{\pi}} \geq V_0^\star - 2\epsilon$, *and runs for at most*

$$C_0 H^4 \cdot \sum_{h=1}^{H} \inf_{\mathbf{\Lambda}_{\mathrm{exp}} \in \mathbf{\Omega}_h} \max_{\pi \in \Pi_\epsilon} \frac{\|\boldsymbol{\phi}_{\pi,h}\|^2_{\mathbf{\Lambda}_{\mathrm{exp}}^{-1}}}{\max\{V_0^\star - V_0^\pi, \epsilon\}^2} \cdot \left(dH + \log\frac{1}{\delta}\right) + C_1$$

episodes, for $C_0 = \mathrm{poly}\log(d, H, \frac{1}{\epsilon})$.

Note that the policy set constructed in Corollary 8.2, Π_ϵ, is guaranteed to contain an ϵ-optimal policy for *any* linear MDP. Thus, this result states that without any prior knowledge of our MDP, PEDEL can be applied to find an ϵ-optimal policy. While Theorem 8.1 and Corollary 8.2 quantify the instance-dependent complexity of learning, it is natural to ask what the *worst-case* complexity of PEDEL is. The following result provides such a bound.

Corollary 8.3. *For any linear MDP satisfying Definition 8.1.1,* $\inf_{\mathbf{\Lambda}_{\mathrm{exp}} \in \mathbf{\Omega}_h} \max_{\pi \in \Pi_\epsilon} \|\boldsymbol{\phi}_{\pi,h}\|^2_{\mathbf{\Lambda}_{\mathrm{exp}}^{-1}} \leq d$, *so the sample complexity of Algorithm 8.1 when run with* $\Pi \leftarrow \Pi_\epsilon$ *is no larger than*

$$\widetilde{\mathcal{O}}\left(\frac{dH^5(dH + \log 1/\delta)}{\epsilon^2} + C_1\right).$$

Corollary 8.3 shows that PEDEL has worst-case optimal dimension dependence, matching the lower bound of $\Omega(d^2H^2/\epsilon^2)$ which we will see in Chapter 9, up to H and log factors.

8.2.2 Low-Regret Algorithms are Suboptimal for PAC RL in Large State-Spaces

Similar to Chapter 7, we next seek to understand what our instance-dependent analysis implies about how we should learn in RL with function approximation. We first show that there are problems on which the instance-dependent complexity of PEDEL improves on the worst-case lower bound shown in Chapter 9, thereby demonstrating that we do indeed obtain favorable complexities on "easy" instances.

Proposition 8.1. *For any* $d > 2$, *there exists a* d*-dimensional linear MDP with* $H = 2$ *such that with probability* $1-\delta$, PEDEL *identifies an* ϵ*-optimal policy on this MDP after running for only* $\widetilde{\mathcal{O}}\big(\frac{\log d/\delta}{\epsilon^2} + \mathrm{poly}(d, \log\frac{1}{\delta}, \log\frac{1}{\epsilon})\big)$ *episodes.*

The complexity given in Proposition 8.1 is a factor of d^2 better than the worst-case lower bound of $\Omega(d^2/\epsilon^2)$. While this shows that PEDEL yields a significant improvement over existing worst-case lower bounds on favorable instances, it is natural to ask whether the same complexity is attainable with existing algorithms, perhaps by applying a tighter analysis. Towards answering this, we consider the same online-to-batch learning protocol as Chapter 7.

Definition 8.2.1 (Low-Regret Algorithm)**.** We say that an algorithm is a *low-regret algorithm* if its expected regret is bounded as, for all K:

$$\mathbb{E}[\mathcal{R}_K] = \textstyle\sum_{k=1}^K \mathbb{E}[V_0^\star - V_0^{\pi_k}] \leq \mathcal{C}_1 K^\alpha + \mathcal{C}_2$$

for some constants $\mathcal{C}_1, \mathcal{C}_2$, and $\alpha \in (0, 1)$.

Protocol 8.2.1 (Online-to-Batch Learning)**.** The online-to-batch protocol proceeds as follows:

1. The learner plays a low-regret algorithm satisfying Definition 8.2.1 for K episodes.

2. The learner stops at a (possibly random) time K, and, using the observations it has collected in any way it wishes, outputs a policy $\widehat{\pi}$ it believes is ϵ-optimal.

In general, by applying online-to-batch learning, one can convert a regret guarantee of $\mathcal{C}_1 K^\alpha + \mathcal{C}_2$ to a PAC complexity of $\mathcal{O}((\frac{\mathcal{C}_1}{\epsilon})^{\frac{1}{1-\alpha}} + \frac{\mathcal{C}_2}{\epsilon})$ [Jin et al., 2018], allowing low-regret algorithms such as that of Zanette et al. [2020b] to obtain the minimax-optimal PAC complexity of $\mathcal{O}(d^2H^4/\epsilon^2)$. The following result shows, however, that this protocol is unable to obtain the instance-optimal rate.

Proposition 8.2. *On the instance of Proposition 8.1, for small enough ϵ, any learner that is (ϵ, δ)-PAC on the set of linear MDPs satisfying Definition 8.1.1 and following Protocol 8.2.1 with stopping time K must have $\mathbb{E}[K] \geq \Omega\big(\frac{d \cdot \log 1/\delta}{\epsilon^2}\big)$.*

Together, Proposition 8.1 and Proposition 8.2 show that running a low-regret algorithm to learn a near-optimal policy in a linear MDP is provably suboptimal—at least a factor of d worse than the instance-dependent rate obtained by PEDEL. This result complements the results presented in Section 7.3, showing that in settings of RL with function approximation, low-regret algorithms are insufficient for obtaining optimal PAC sample complexity. As standard optimistic algorithms are also low-regret, this result implies that all such optimistic algorithms are also suboptimal.

8.2.3 Tabular and Deterministic MDPs

The following result shows that the complexity PEDEL obtains on tabular MDPs and the gap-visitation complexity from Chapter 7 do not have a clear ordering.

Proposition 8.3. *Fix any $\epsilon \in (0, 1/2)$ and $S \geq \log_2(1/\epsilon)$. Then there exist tabular MDPs $\mathcal{M}_1$ and $\mathcal{M}_2$, each with $H = 2$, S states, and $\mathcal{O}(S)$ actions, such that:*

- *On $\mathcal{M}_1$, the complexity bound of PEDEL given in Theorem 8.1 scales as $\mathrm{poly}(S, \log 1/\delta)$, while the gap-visitation complexity scales as $\Omega(1/\epsilon^2)$.*

- *On $\mathcal{M}_2$, the complexity bound of PEDEL given in Theorem 8.1 scales as $\Omega(1/\epsilon^2)$, while the gap-visitation complexity scales as $\mathrm{poly}(S, \log 1/\delta)$.*

The lack of ordering between the two complexity measures arises because, on some problem instances, it is easier to learn in policy-space (as PEDEL does), while on other instances,

it is easier to learn near-optimal actions on individual states directly, and then synthesize these actions into a near-optimal policy (the approach MOCA takes). This difference arises because, in the former instance, the minimum *policy gap* is large ($V_0^\star - V_0^\pi = \Omega(1)$ for every deterministic policy $\pi \neq \pi^\star$), while in the latter instance, the minimum policy gap is small, but all value-function gaps are large, satisfying $\Delta_h(s,a) = \Omega(1)$ for all $a \neq \arg\max_{a \in \mathcal{A}} Q_h^\star(s,a)$ and all s and h. Thus, on the former instance, it is much easier to learn over the space of policies, while on the latter it is much easier to learn optimal actions in individual states[1]. Resolving this discrepancy with an algorithm able to achieve the "best-of-both-worlds" is an interesting direction for future work.

Deterministic MDPs. Finally, we turn to the simplified setting of tabular, deterministic MDPs. Here, for each (s,a,h), there exists some s' such that $P_h(s'|s,a) = 1$. We still allow the rewards to be random, however, so the agent must still learn in order to find a near-optimal policy. Following the same notation as the recent work of Tirinzoni et al. [2022], let $\Pi_{sah} = \{\pi \text{ deterministic } : s_h^\pi = s, a_h^\pi = a\}$, where s_h^π and a_h^π are the state and action policy π will be in at step h (note that these quantities are well-defined quantities for deterministic policies). Also define the *deterministic return gap* as $\bar{\Delta}_h(s,a) := V_0^\star - \max_{\pi \in \Pi_{sah}} V_0^\pi$, and let $\bar{\Delta}_{\min} := \min_{s,a,h:\bar{\Delta}_h(s,a)>0} \bar{\Delta}_h(s,a)$ in the case when there exists a unique optimal deterministic policy, and $\bar{\Delta}_{\min} := 0$ otherwise. We obtain the following.

Corollary 8.4. *In the setting of tabular, deterministic MDPs,* PEDEL *outputs an ϵ-optimal policy with probability at least $1-\delta$, and has sample complexity bounded as*

$$\widetilde{\mathcal{O}}\left(H^4 \cdot \sum_{h=1}^{H} \sum_{s,a} \frac{1}{\max\{\bar{\Delta}_h(s,a), \bar{\Delta}_{\min}, \epsilon\}^2} \cdot (H + \log\tfrac{1}{\delta}) + \text{poly}\left(S, A, H, \log\tfrac{1}{\delta}, \log\tfrac{1}{\epsilon}\right) \right).$$

Up to H and log factors and lower-order terms, the rate given in Corollary 8.4 matches the instance-dependent lower bound given in Tirinzoni et al. [2022]. Thus, we conclude that, in the setting of tabular, deterministic MDPs, PEDEL is (nearly) instance-optimal. While Tirinzoni et al. [2022] also obtain instance-optimality in this setting, their algorithm and analysis are specialized to tabular, deterministic MDPs—in contrast, PEDEL requires no modification from its standard operation.

8.3 Proof of Theorem 8.1

We next provide a proof of Theorem 8.1. Towards proving Theorem 8.1, in Section 8.3.1 we first establish some bounds on the estimation error of the feature-visitations, $\boldsymbol{\phi}_{\pi,h}$. Using these bounds, in Section 8.3.2 we show that with high probability the policy-elimination procedure

[1] We remark that, in Proposition 8.3, $\mathcal{M}_1$ is a deterministic MDP, which PEDEL takes advantage of to obtain a tighter complexity. It is unclear if a similar result can be shown to hold for $\mathcal{M}_1$ a stochastic MDP.

of PEDEL succeeds in producing an ϵ-optimal policy, and provide per-epoch bounds on the number of samples collected. Section 8.3.3 then combines these results to prove Theorem 8.1.

Throughout this section, assuming we have run for some number of episodes K, we let $(\mathcal{F}_\tau)_{\tau=1}^K$ the filtration on this, with $\mathcal{F}_\tau$ the filtration up to and including episode τ. We also let $\mathcal{F}_{\tau,h}$ denote the filtration on all episodes $\tau' < \tau$, and on steps $h' = 1, \dots, h$ of episode τ.

8.3.1 Estimating Feature-Visitations and Rewards

Define:

$$\mathcal{T}_{\pi,h} := \int \boldsymbol{\phi}_{\pi,h}(s) \mathrm{d}\boldsymbol{\mu}_{h-1}(s)^\top.$$

Note that $\boldsymbol{\phi}_{\pi,h} = \mathcal{T}_{\pi,h}\boldsymbol{\phi}_{\pi,h-1}$. Our estimator of $\boldsymbol{\phi}_{\pi,h}$ relies on estimating $\mathcal{T}_{\pi,h}$. The following result bounds the estimation error of $\mathcal{T}_{\pi,h}$, which is critical to obtaining an estimation error bound on $\boldsymbol{\phi}_{\pi,h}$.

Lemma 8.3.1. *Assume that we have collected some data $\{(s_{h-1,\tau}, a_{h-1,\tau}, s_{h,\tau})\}_{\tau=1}^K$, where, for each τ', $s_{h,\tau'}|\mathcal{F}_{h-1,\tau'}$ is independent of $\{(s_{h-1,\tau}, a_{h-1,\tau}, s_{h,\tau})\}_{\tau \neq \tau'}$. Denote $\boldsymbol{\phi}_{h-1,\tau} = \boldsymbol{\phi}(s_{h-1,\tau}, a_{h-1,\tau})$ and $\boldsymbol{\Lambda}_{h-1} = \sum_{\tau=1}^K \boldsymbol{\phi}_{h-1,\tau}\boldsymbol{\phi}_{h-1,\tau}^\top + \lambda I$. Fix π and let*

$$\widehat{\mathcal{T}}_{\pi,h} = \left(\sum_{\tau=1}^K \boldsymbol{\phi}_{\pi,h}(s_{h,\tau})\boldsymbol{\phi}_{h-1,\tau}^\top\right)\boldsymbol{\Lambda}_{h-1}^{-1}.$$

Fix $\boldsymbol{v} \in \mathbb{R}^d$ satisfying $|\boldsymbol{v}^\top \boldsymbol{\phi}_{\pi,h}(s)| \leq 1$ for all s and $\boldsymbol{u} \in \mathbb{R}^d$. Then with probability at least $1-\delta$, we can bound

$$|\boldsymbol{v}^\top(\mathcal{T}_{\pi,h} - \widehat{\mathcal{T}}_{\pi,h})\boldsymbol{u}| \leq \left(2\sqrt{\log 2/\delta} + \frac{\log 2/\delta}{\sqrt{\lambda_{\min}(\boldsymbol{\Lambda}_{h-1})}} + \sqrt{\lambda}\|\mathcal{T}_{\pi,h}^\top \boldsymbol{v}\|_2\right) \cdot \|\boldsymbol{u}\|_{\boldsymbol{\Lambda}_{h-1}^{-1}}.$$

Proof. Let $\mathfrak{D} = \{(s_{h-1,\tau}, a_{h-1,\tau})\}_{\tau=1}^K$, our data collected at step $h-1$. Then by our assumption on the independence of $s_{h,\tau}$, we have that $s_{h,\tau}|\mathcal{F}_{h-1,\tau}$ has the same distribution as $s_{h,\tau}|(\mathcal{F}_{h-1,\tau}, \mathfrak{D})$. Conditioning on $\mathfrak{D}$, the $\boldsymbol{\phi}_{h-1,\tau}$ vectors are fixed, so $\boldsymbol{\Lambda}_{h-1}$ is also fixed. Note that

$$\begin{aligned}
\mathcal{T}_{\pi,h} &= \int \boldsymbol{\phi}_{\pi,h}(s)\mathrm{d}\boldsymbol{\mu}_{h-1}(s)^\top \\
&= \int \boldsymbol{\phi}_{\pi,h}(s)\mathrm{d}\boldsymbol{\mu}_{h-1}(s)^\top \left(\sum_{\tau=1}^K \boldsymbol{\phi}_{h-1,\tau}\boldsymbol{\phi}_{h-1,\tau}^\top\right)\boldsymbol{\Lambda}_{h-1}^{-1} + \lambda \int \boldsymbol{\phi}_{\pi,h}(s)\mathrm{d}\boldsymbol{\mu}_{h-1}(s)^\top \boldsymbol{\Lambda}_{h-1}^{-1} \\
&= \sum_{\tau=1}^K \left(\int \boldsymbol{\phi}_{\pi,h}(s)\mathrm{d}\boldsymbol{\mu}_{h-1}(s)^\top \boldsymbol{\phi}_{h-1,\tau}\right)\boldsymbol{\phi}_{h-1,\tau}^\top \boldsymbol{\Lambda}_{h-1}^{-1} + \lambda \int \boldsymbol{\phi}_{\pi,h}(s)\mathrm{d}\boldsymbol{\mu}_{h-1}(s)^\top \boldsymbol{\Lambda}_{h-1}^{-1}
\end{aligned}$$

$$= \sum_{\tau=1}^{K} \mathbb{E}[\boldsymbol{\phi}_{\pi,h}(s_{h,\tau})|\mathcal{F}_{h-1,\tau}]\boldsymbol{\phi}_{h-1,\tau}^\top \boldsymbol{\Lambda}_{h-1}^{-1} + \lambda \int \boldsymbol{\phi}_{\pi,h}(s)\mathrm{d}\boldsymbol{\mu}_{h-1}(s)^\top \boldsymbol{\Lambda}_{h-1}^{-1}$$
$$= \sum_{\tau=1}^{K} \mathbb{E}[\boldsymbol{\phi}_{\pi,h}(s_{h,\tau})|\mathcal{F}_{h-1,\tau}]\boldsymbol{\phi}_{h-1,\tau}^\top \boldsymbol{\Lambda}_{h-1}^{-1} + \lambda \mathcal{T}_{\pi,h}\boldsymbol{\Lambda}_{h-1}^{-1}$$

so

$$|\boldsymbol{v}^\top(\mathcal{T}_{\pi,h} - \widehat{\mathcal{T}}_{\pi,h})\boldsymbol{u}| \leq \underbrace{\left|\sum_{\tau=1}^{K} \boldsymbol{v}^\top \left(\mathbb{E}[\boldsymbol{\phi}_{\pi,h}(s_{h,\tau})|\mathcal{F}_{h-1,\tau}] - \boldsymbol{\phi}_{\pi,h}(s_{h,\tau})\right)\boldsymbol{\phi}_{h-1,\tau}^\top \boldsymbol{\Lambda}_{h-1}^{-1}\boldsymbol{u}\right|}_{(a)} + \underbrace{\left|\lambda \boldsymbol{v}^\top \mathcal{T}_{\pi,h}\boldsymbol{\Lambda}_{h-1}^{-1}\boldsymbol{u}\right|}_{(b)}.$$

Conditioned on $\mathfrak{D}$, (a) is simply the sum of mean 0 random variables, where the τth random variable has magnitude bounded as

$$\begin{aligned}
|\boldsymbol{v}^\top \left(\mathbb{E}[\boldsymbol{\phi}_{\pi,h}(s_{h,\tau})|\mathcal{F}_{h-1,\tau}] - \boldsymbol{\phi}_{\pi,h}(s_{h,\tau})\right)\boldsymbol{\phi}_{h-1,\tau}^\top \boldsymbol{\Lambda}_{h-1}^{-1}\boldsymbol{u}| &\leq 2|\boldsymbol{\phi}_{h-1,\tau}^\top \boldsymbol{\Lambda}_{h-1}^{-1}\boldsymbol{u}| \\
&\leq 2\|\boldsymbol{\phi}_{h-1,\tau}\|_{\boldsymbol{\Lambda}_{h-1}^{-1}}\|\boldsymbol{u}\|_{\boldsymbol{\Lambda}_{h-1}^{-1}} \\
&\leq 2\|\boldsymbol{u}\|_{\boldsymbol{\Lambda}_{h-1}^{-1}}/\sqrt{\lambda_{\min}(\boldsymbol{\Lambda}_{h-1})}
\end{aligned}$$

Furthermore, the variance of each term in (a) is bounded as

$$\begin{aligned}
&\mathrm{Var}\left[\boldsymbol{v}^\top \left(\mathbb{E}[\boldsymbol{\phi}_{\pi,h}(s_{h,\tau})|\mathcal{F}_{h-1,\tau}] - \boldsymbol{\phi}_{\pi,h}(s_{h,\tau})\right)\boldsymbol{\phi}_{h-1,\tau}^\top \boldsymbol{\Lambda}_{h-1}^{-1}\boldsymbol{u}|\mathcal{F}_{h-1}\right] \\
&= \mathbb{E}\left[\left(\boldsymbol{v}^\top \left(\mathbb{E}[\boldsymbol{\phi}_{\pi,h}(s_{h,\tau})|\mathcal{F}_{h-1,\tau}] - \boldsymbol{\phi}_{\pi,h}(s_{h,\tau})\right)\boldsymbol{\phi}_{h-1,\tau}^\top \boldsymbol{\Lambda}_{h-1}^{-1}\boldsymbol{u}\right)^2 |\mathcal{F}_{h-1}\right] \\
&\leq \boldsymbol{u}^\top \boldsymbol{\Lambda}_{h-1}^{-1}\boldsymbol{\phi}_{h-1,\tau}\boldsymbol{\phi}_{h-1,\tau}^\top \boldsymbol{\Lambda}_{h-1}^{-1}\boldsymbol{u}.
\end{aligned}$$

It follows that, by Bernstein's Inequality, we can bound, with probability at least $1-\delta$ conditioned on $\mathfrak{D}$:

$$\begin{aligned}
(a) &\leq 2\sqrt{\sum_{\tau=1}^{K} \boldsymbol{u}^\top \boldsymbol{\Lambda}_{h-1}^{-1}\boldsymbol{\phi}_{h-1,\tau}\boldsymbol{\phi}_{h-1,\tau}^\top \boldsymbol{\Lambda}_{h-1}^{-1}\boldsymbol{u} \cdot \log\frac{2}{\delta}} + \frac{2\|\boldsymbol{u}\|_{\boldsymbol{\Lambda}_{h-1}^{-1}}}{\sqrt{\lambda_{\min}(\boldsymbol{\Lambda}_{h-1})}} \cdot \log\frac{2}{\delta} \\
&\leq 2(\sqrt{\log 2/\delta} + \frac{\log 2/\delta}{\sqrt{\lambda_{\min}(\boldsymbol{\Lambda}_{h-1})}}) \cdot \|\boldsymbol{u}\|_{\boldsymbol{\Lambda}_{h-1}^{-1}}.
\end{aligned}$$

In other words,

$$\mathbb{P}\left[(a) \geq 2(\sqrt{\log 2/\delta} + \frac{\log 2/\delta}{\sqrt{\lambda_{\min}(\boldsymbol{\Lambda}_{h-1})}}) \cdot \|\boldsymbol{u}\|_{\boldsymbol{\Lambda}_{h-1}^{-1}} | \mathfrak{D}\right] \leq \delta$$

so, by the law of total probability, for any distribution F over $\mathfrak{D}$,

$$\begin{aligned}
&\mathbb{P}\left[(a) \geq 2(\sqrt{\log 2/\delta} + \min\{1, \lambda^{-1}\}\log 2/\delta) \cdot \|\boldsymbol{u}\|_{\boldsymbol{\Lambda}_{h-1}^{-1}}\right] \\
&= \int \mathbb{P}\left[(a) \geq 2(\sqrt{\log 2/\delta} + \min\{1, \lambda^{-1}\}\log 2/\delta) \cdot \|\boldsymbol{u}\|_{\boldsymbol{\Lambda}_{h-1}^{-1}} | \mathfrak{D}\right] \mathrm{d}F(\mathfrak{D}) \\
&\leq \delta \int \mathrm{d}F(\mathfrak{D}) \\
&= \delta.
\end{aligned}$$

We can also bound

$$(b) \leq \sqrt{\lambda}\|\boldsymbol{u}\|_{\boldsymbol{\Lambda}_{h-1}^{-1}} \|\mathcal{T}_{\pi,h}^\top \boldsymbol{v}\|_2.$$

Combining these gives the result. □

Lemma 8.3.2. *Fix π and let*

$$\widehat{\boldsymbol{\phi}}_{\pi,h} = \widehat{\mathcal{T}}_{\pi,h}\widehat{\mathcal{T}}_{\pi,h-1}\ldots\widehat{\mathcal{T}}_{\pi,2}\widehat{\mathcal{T}}_{\pi,1}\boldsymbol{\phi}_{\pi,0}.$$

Fix $\boldsymbol{u} \in \mathcal{S}^{d-1}$ or $\boldsymbol{u}$ a valid reward vector as defined by Definition 8.1.1. Then with probability at least $1-\delta$:

$$|\langle \boldsymbol{u}, \boldsymbol{\phi}_{\pi,h} - \widehat{\boldsymbol{\phi}}_{\pi,h}\rangle| \leq \sum_{i=1}^{h-1}\left(2\sqrt{\log\frac{2H}{\delta}} + \frac{\log\frac{2H}{\delta}}{\sqrt{\lambda_{\min}(\boldsymbol{\Lambda}_i)}} + \sqrt{d\lambda}\right) \cdot \|\widehat{\boldsymbol{\phi}}_{\pi,i}\|_{\boldsymbol{\Lambda}_i^{-1}}.$$

Proof. Note that

$$\begin{aligned}
\boldsymbol{\phi}_{\pi,h} - \widehat{\boldsymbol{\phi}}_{\pi,h} &= \mathcal{T}_{\pi,h}\boldsymbol{\phi}_{\pi,h-1} - \widehat{\mathcal{T}}_{\pi,h}\widehat{\boldsymbol{\phi}}_{\pi,h-1} \\
&= \mathcal{T}_{\pi,h}(\boldsymbol{\phi}_{\pi,h-1} - \widehat{\boldsymbol{\phi}}_{\pi,h-1}) + (\mathcal{T}_{\pi,h} - \widehat{\mathcal{T}}_{\pi,h})\widehat{\boldsymbol{\phi}}_{\pi,h-1}.
\end{aligned}$$

Thus, unrolling this all the way back, we get

$$\boldsymbol{\phi}_{\pi,h} - \widehat{\boldsymbol{\phi}}_{\pi,h} = \sum_{i=0}^{h-2}\left(\prod_{j=h-i+1}^{h}\mathcal{T}_{\pi,j}\right)(\mathcal{T}_{\pi,h-i} - \widehat{\mathcal{T}}_{\pi,h-i})\widehat{\boldsymbol{\phi}}_{\pi,h-i-1}$$

where we order the product $\prod_{j=h-i+1}^{h}\mathcal{T}_{\pi,j} = \mathcal{T}_{\pi,h}\mathcal{T}_{\pi,h-1}\ldots\mathcal{T}_{\pi,h-i+1}$. It follows that

$$|\langle \boldsymbol{u}, \boldsymbol{\phi}_{\pi,h} - \widehat{\boldsymbol{\phi}}_{\pi,h}\rangle| \leq \sum_{i=0}^{h-2}\left|\boldsymbol{u}^\top\left(\prod_{j=h-i+1}^{h}\mathcal{T}_{\pi,j}\right)(\mathcal{T}_{\pi,h-i} - \widehat{\mathcal{T}}_{\pi,h-i})\widehat{\boldsymbol{\phi}}_{\pi,h-i-1}\right|.$$

Denote $\boldsymbol{v}_i := \boldsymbol{u}^\top \left(\prod_{j=h-i+1}^{h} \mathcal{T}_{\pi,j}\right)$. By Lemma F.1.3 and our assumption on $\boldsymbol{u}$, we can bound $\|\boldsymbol{v}_i\|_2 \le \sqrt{d}$ and also have that for all s, a, $|\boldsymbol{v}_i^\top \boldsymbol{\phi}(s,a)| \le 1$, which implies

$$|\boldsymbol{v}_i^\top \boldsymbol{\phi}_{\pi,j}(s)| = \left|\sum_{a\in\mathcal{A}} \boldsymbol{v}_i^\top \boldsymbol{\phi}(s,a)\pi_h(a|s)\right| \le \sum_{a\in\mathcal{A}} \pi_h(a|s) = 1.$$

We can therefore apply Lemma 8.3.1 to get that, with probability at least $1-\delta$, for all i,

$$\left|\boldsymbol{v}_i^\top (\mathcal{T}_{\pi,h-i} - \widehat{\mathcal{T}}_{\pi,h-i})\widehat{\boldsymbol{\phi}}_{\pi,h-i-1}\right| \le \left(2\sqrt{\log\frac{2H}{\delta}} + \frac{\log\frac{2H}{\delta}}{\sqrt{\lambda_{\min}(\boldsymbol{\Lambda}_{h-i-1})}} + \sqrt{\lambda}\|\mathcal{T}_{\pi,h}^\top \boldsymbol{v}_i\|_2\right) \cdot \|\widehat{\boldsymbol{\phi}}_{\pi,h-i-1}\|_{\boldsymbol{\Lambda}_{h-i-1}^{-1}}$$

By Lemma F.1.3, the definition of $\boldsymbol{v}_i$, and our assumption on $\boldsymbol{u}$, we can bound $\|\mathcal{T}_{\pi,h}^\top \boldsymbol{v}_i\|_2 \le \sqrt{d}$. Summing over i proves the result. □

8.3.2 Correctness and Sample Complexity of PEDEL

Using the above estimation error bounds, we now seek to show the correctness of PEDEL; that with high probability it succeeds in producing an ϵ-optimal policy. The following results define several "good events"—that we estimate the expected feature vectors and rewards accurately, and that we explore effectively—and some implications of these good events. Lemma 8.3.7 then shows that on these good events, PEDEL indeed produces an ϵ-optimal policy.

Lemma 8.3.3. *Let $\mathcal{E}_{\text{est}}^{\ell,h}$ denote the event on which, for all $\pi \in \Pi_\ell$:*

$$|\langle \boldsymbol{\theta}_{h+1}, \widehat{\boldsymbol{\phi}}_{\pi,h+1}^{\ell} - \boldsymbol{\phi}_{\pi,h+1}\rangle| \le \sum_{i=1}^{h}\left(3\sqrt{\log\frac{4H^2|\Pi_\ell|\ell^2}{\delta}} + \frac{\log\frac{4H^2|\Pi_\ell|\ell^2}{\delta}}{\sqrt{\lambda_{\min}(\boldsymbol{\Lambda}_{i,\ell})}}\right) \cdot \|\widehat{\boldsymbol{\phi}}_{\pi,i}^{\ell}\|_{\boldsymbol{\Lambda}_{i,\ell}^{-1}},$$

$$\|\widehat{\boldsymbol{\phi}}_{\pi,h+1}^{\ell} - \boldsymbol{\phi}_{\pi,h+1}\|_2 \le d\sum_{i=1}^{h}\left(3\sqrt{\log\frac{4H^2 d|\Pi_\ell|\ell^2}{\delta}} + \frac{\log\frac{4H^2 d|\Pi_\ell|\ell^2}{\delta}}{\sqrt{\lambda_{\min}(\boldsymbol{\Lambda}_{i,\ell})}}\right) \cdot \|\widehat{\boldsymbol{\phi}}_{\pi,i}^{\ell}\|_{\boldsymbol{\Lambda}_{i,\ell}^{-1}},$$

$$|\langle \widehat{\boldsymbol{\phi}}_{\pi,h}^{\ell}, \widehat{\boldsymbol{\theta}}_h - \boldsymbol{\theta}_h\rangle| \le \left(2\sqrt{\log\frac{4H^2|\Pi_\ell|\ell^2}{\delta}} + \frac{\log\frac{4H^2|\Pi_\ell|\ell^2}{\delta}}{\sqrt{\lambda_{\min}(\boldsymbol{\Lambda}_{h,\ell})}}\right) \cdot \|\widehat{\boldsymbol{\phi}}_{\pi,h}^{\ell}\|_{\boldsymbol{\Lambda}_{h,\ell}^{-1}}.$$

Then $\mathbb{P}[(\mathcal{E}_{\text{est}}^{\ell,h})^c] \le \frac{\delta}{2H\ell^2}$.

Proof. The result follows by Lemma 8.3.2, Lemma F.3.1, and Lemma F.3.2, and setting $\lambda = 1/d$. □

Lemma 8.3.4. *Let $\mathcal{E}_{\text{exp}}^{\ell,h}$ denote the event on which:*

- *The exploration procedure on Line 6 terminates after running for at most*

$$C \cdot \frac{\inf_{\mathbf{\Lambda} \in \mathbf{\Omega}_h} \max_{\boldsymbol{\phi} \in \Phi_{\ell,h}} \|\boldsymbol{\phi}\|^2_{\mathbf{A}(\mathbf{\Lambda})^{-1}}}{\epsilon_\ell^2/\beta_\ell} + \mathrm{poly}\left(d, H, \log \frac{\ell^2}{\delta}, \frac{1}{\bar{\lambda}_{\min}}, \log |\Pi_\ell|\right)$$

  episodes.

- *The covariates returned by Line 6 for any (h, ℓ), $\mathbf{\Lambda}_{h,\ell}$, satisfy*

$$\max_{\boldsymbol{\phi} \in \Phi_{\ell,h}} \|\boldsymbol{\phi}\|^2_{\mathbf{\Lambda}_{h,\ell}^{-1}} \le \frac{\epsilon_\ell^2}{\beta_\ell}, \qquad \lambda_{\min}(\mathbf{\Lambda}_{h,\ell}) \ge \log \frac{4H^2|\Pi_\ell|\ell^2}{\delta}.$$

Then $\mathbb{P}[(\mathcal{E}^{\ell,h}_{\mathrm{exp}})^c \cap \mathcal{E}^{\ell,h-1}_{\mathrm{est}} \cap (\cap_{i=1}^{h-1} \mathcal{E}^{\ell,i}_{\mathrm{exp}})] \le \frac{\delta}{2H\ell^2}$.

Proof. By Lemma 8.3.5, on the event $\mathcal{E}^{\ell,h-1}_{\mathrm{est}} \cap (\cap_{i=1}^{h-1} \mathcal{E}^{\ell,i}_{\mathrm{exp}})$ we can bound $\|\widehat{\boldsymbol{\phi}}^\ell_{\pi,h} - \boldsymbol{\phi}_{\pi,h}\|_2 \le d\epsilon_\ell/2H$. By Lemma F.1.1, we can lower bound $\|\boldsymbol{\phi}_{\pi,h}\|_2 \ge 1/\sqrt{d}$. By the reverse triangle inequality,

$$\|\widehat{\boldsymbol{\phi}}^\ell_{\pi,h}\|_2 \ge \|\boldsymbol{\phi}_{\pi,h}\|_2 - \|\widehat{\boldsymbol{\phi}}^\ell_{\pi,h} - \boldsymbol{\phi}_{\pi,h}\|_2 \ge 1/\sqrt{d} - d\epsilon_\ell/2H.$$

It follows that as long as $\epsilon_\ell \le H/d^{3/2}$, that we can lower bound $\|\widehat{\boldsymbol{\phi}}^\ell_{\pi,h}\|_2 \ge 1/(2\sqrt{d})$. Since we start ℓ at $\ell = \lceil \log_2 \frac{d^{3/2}}{H} \rceil$, we will have that $\epsilon_\ell = 2^{-\ell} \le H/d^{3/2}$.

The result then follows by applying Theorem 8.5 with our chosen parameters and $\gamma_\Phi \leftarrow 1/(2\sqrt{d})$.

□

Lemma 8.3.5. *On the event $\mathcal{E}^{\ell,h}_{\mathrm{est}} \cap (\cap_{i=1}^{h} \mathcal{E}^{\ell,i}_{\mathrm{exp}})$, for all $\pi \in \Pi_\ell$:*

$$\begin{aligned}
&|\langle \boldsymbol{\theta}_{h+1}, \widehat{\boldsymbol{\phi}}^\ell_{\pi,h+1} - \boldsymbol{\phi}_{\pi,h+1} \rangle| \le \epsilon_\ell/2H, \\
&\|\widehat{\boldsymbol{\phi}}^\ell_{\pi,h+1} - \boldsymbol{\phi}_{\pi,h+1}\|_2 \le d\epsilon_\ell/2H, \\
&|\langle \widehat{\boldsymbol{\phi}}^\ell_{\pi,h}, \widehat{\boldsymbol{\theta}}_h - \boldsymbol{\theta}_h \rangle| \le \epsilon_\ell/2H.
\end{aligned}$$

Proof. On $\mathcal{E}^{\ell,i}_{\mathrm{exp}}$, we can lower bound

$$\lambda_{\min}(\mathbf{\Lambda}_{i,\ell}) \ge \log \frac{4H^2|\Pi_\ell|\ell^2}{\delta}$$

which implies

$$3\sqrt{\log \frac{4H^2|\Pi_\ell|\ell^2}{\delta}} + \frac{\log \frac{4H^2|\Pi_\ell|\ell^2}{\delta}}{\sqrt{\lambda_{\min}(\mathbf{\Lambda}_{i,\ell})}} \le 4\sqrt{\log \frac{4H^2|\Pi_\ell|\ell^2}{\delta}}.$$

Furthermore, on $\mathcal{E}_{\text{exp}}^{\ell,i}$, $\|\widehat{\boldsymbol{\phi}}_{\pi,i}^{\ell}\|_{\boldsymbol{\Lambda}_{i,\ell}^{-1}} \le \frac{\epsilon_\ell}{\sqrt{\beta_\ell}}$. Since $\beta_\ell = 64H^4 \log \frac{4H^2|\Pi_\ell|\ell^2}{\delta}$, on $\mathcal{E}_{\text{est}}^{\ell,h}$, we can then upper bound

$$\begin{aligned}
|\langle \boldsymbol{\theta}_{h+1}, \widehat{\boldsymbol{\phi}}_{\pi,h+1}^{\ell} - \boldsymbol{\phi}_{\pi,h+1}\rangle| &\le \sum_{i=1}^{h} \left(3\sqrt{\log \frac{4H^2|\Pi_\ell|\ell^2}{\delta}} + \frac{\log \frac{4H^2|\Pi_\ell|\ell^2}{\delta}}{\sqrt{\lambda_{\min}(\boldsymbol{\Lambda}_{i,\ell})}}\right) \cdot \|\widehat{\boldsymbol{\phi}}_{\pi,i}^{\ell}\|_{\boldsymbol{\Lambda}_{i,\ell}^{-1}} \\
&\le H4\sqrt{\log \frac{4H^2|\Pi_\ell|\ell^2}{\delta}} \frac{\epsilon_\ell}{\sqrt{\beta_\ell}} \\
&\le \epsilon_\ell/2H.
\end{aligned}$$

The same calculation gives the bounds on $\|\widehat{\boldsymbol{\phi}}_{\pi,h}^{\ell} - \boldsymbol{\phi}_{\pi,h}\|_2$ and $|\langle \widehat{\boldsymbol{\phi}}_{\pi,h}^{\ell}, \widehat{\boldsymbol{\theta}}_h - \boldsymbol{\theta}_h\rangle|$. □

Lemma 8.3.6. *Define* $\mathcal{E}_{\text{exp}} = \cap_\ell \cap_h \mathcal{E}_{\text{exp}}^{\ell,h}$ *and* $\mathcal{E}_{\text{est}} = \cap_\ell \cap_h \mathcal{E}_{\text{est}}^{\ell,h}$. *Then* $\mathbb{P}[\mathcal{E}_{\text{est}} \cap \mathcal{E}_{\text{exp}}] \ge 1 - 2\delta$ *and on* $\mathcal{E}_{\text{est}} \cap \mathcal{E}_{\text{exp}}$, *for all* h, ℓ, *and* $\pi \in \Pi_\ell$,

$$\begin{aligned}
&|\langle \boldsymbol{\theta}_{h+1}, \widehat{\boldsymbol{\phi}}_{\pi,h+1}^{\ell} - \boldsymbol{\phi}_{\pi,h+1}\rangle| \le \epsilon_\ell/2H, \\
&\|\widehat{\boldsymbol{\phi}}_{\pi,h+1}^{\ell} - \boldsymbol{\phi}_{\pi,h+1}\|_2 \le d\epsilon_\ell/2H, \\
&|\langle \widehat{\boldsymbol{\phi}}_{\pi,h}^{\ell}, \widehat{\boldsymbol{\theta}}_h - \boldsymbol{\theta}_h\rangle| \le \epsilon_\ell/2H.
\end{aligned}$$

Proof. Clearly,

$$\begin{aligned}
\mathcal{E}_{\text{est}}^c \cup \mathcal{E}_{\text{exp}}^c &= \bigcup_{\ell=\ell_0}^{\lceil \log 4/\epsilon\rceil} \bigcup_{h=1}^{H} ((\mathcal{E}_{\text{est}}^{\ell,h})^c \cup (\mathcal{E}_{\text{exp}}^{\ell,h})^c) \\
&= \bigcup_{\ell=\ell_0}^{\lceil \log 4/\epsilon\rceil} \bigcup_{h=1}^{H} (\mathcal{E}_{\text{est}}^{\ell,h})^c \backslash \left((\mathcal{E}_{\text{est}}^{\ell,h-1})^c \cup (\cup_{i=1}^{h-1} (\mathcal{E}_{\text{exp}}^{\ell,i})^c)\right) \cup \bigcup_{\ell=\ell_0}^{\lceil \log 4/\epsilon\rceil} \bigcup_{h=1}^{H} (\mathcal{E}_{\text{exp}}^{\ell,h})^c \\
&= \bigcup_{\ell=\ell_0}^{\lceil \log 4/\epsilon\rceil} \bigcup_{h=1}^{H} (\mathcal{E}_{\text{est}}^{\ell,h})^c \cap \left(\mathcal{E}_{\text{est}}^{\ell,h-1} \cap (\cup_{i=1}^{h-1} \mathcal{E}_{\text{exp}}^{\ell,i})\right) \cup \bigcup_{\ell=\ell_0}^{\lceil \log 4/\epsilon\rceil} \bigcup_{h=1}^{H} (\mathcal{E}_{\text{exp}}^{\ell,h})^c.
\end{aligned}$$

The first conclusion follows by Lemma 8.3.3, Lemma 8.3.4, and since we can bound

$$\sum_{\ell} \sum_{h=1}^{H} 2 \cdot \frac{\delta}{2H\ell^2} \le \frac{\pi^2}{6}\delta \le 2\delta.$$

The second conclusion follows by Lemma 8.3.5. □

Lemma 8.3.7. *On the event* $\mathcal{E}_{\text{est}} \cap \mathcal{E}_{\text{exp}}$, *for all* $\ell > \ell_0$, *every policy* $\pi \in \Pi_\ell$ *satisfies* $V_0^\star(\Pi) - V_0^\pi \le 4\epsilon_\ell$ *and* $\widetilde{\pi}^\star \in \Pi_\ell$, *for* $\widetilde{\pi}^\star = \arg\max_{\pi \in \Pi} V_0^\pi$.

Proof. The value of a policy π is given by

$$\sum_{h=1}^{H} \langle \boldsymbol{\theta}_h, \boldsymbol{\phi}_{\pi,h} \rangle.$$

By Lemma 8.3.6, for all $\pi \in \Pi_\ell$ we can bound

$$|\langle \widehat{\boldsymbol{\theta}}_h, \widehat{\boldsymbol{\phi}}^\ell_{\pi,h} \rangle - \langle \boldsymbol{\theta}_h, \boldsymbol{\phi}_{\pi,h} \rangle| \le |\langle \widehat{\boldsymbol{\theta}}_h - \boldsymbol{\theta}_h, \widehat{\boldsymbol{\phi}}^\ell_{\pi,h} \rangle| + |\langle \boldsymbol{\theta}_h, \widehat{\boldsymbol{\phi}}^\ell_{\pi,h} - \boldsymbol{\phi}_{\pi,h} \rangle| \le \epsilon_\ell / 2H + \epsilon_\ell / 2H = \epsilon_\ell / H.$$

Thus,

$$\left| \sum_{h=1}^{H} \langle \widehat{\boldsymbol{\theta}}^\ell_h, \widehat{\boldsymbol{\phi}}^\ell_{\pi,h} \rangle - \sum_{h=1}^{H} \langle \boldsymbol{\theta}_h, \boldsymbol{\phi}_{\pi,h} \rangle \right| \le \epsilon_\ell.$$

We will only include $\pi \in \Pi_{\ell+1}$ if $\pi \in \Pi_\ell$ and

$$\sum_{h=1}^{H} \langle \widehat{\boldsymbol{\phi}}^\ell_{\pi,h}, \widehat{\boldsymbol{\theta}}^\ell_h \rangle \ge \sup_{\pi' \in \Pi_\ell} \sum_{h=1}^{H} \langle \widehat{\boldsymbol{\phi}}^\ell_{\pi',h}, \widehat{\boldsymbol{\theta}}^\ell_h \rangle - 2\epsilon_\ell.$$

Using the estimation error given above, this implies that for any $\pi \in \Pi_\ell$,

$$V_0^\pi = \sum_{h=1}^{H} \langle \boldsymbol{\theta}_h, \boldsymbol{\phi}_{\pi,h} \rangle \ge \sup_{\pi' \in \Pi_\ell} \sum_{h=1}^{H} \langle \boldsymbol{\theta}_h, \boldsymbol{\phi}_{\pi',h} \rangle - 4\epsilon_\ell = \sup_{\pi' \in \Pi_\ell} V_0^{\pi'} - 4\epsilon_\ell.$$

Both claims then follow if we can show $\widetilde{\pi}^\star$ is always contained in the active set. Assume that $\widetilde{\pi}^\star \in \Pi_\ell$. Then

$$\sum_{h=1}^{H} \langle \widehat{\boldsymbol{\phi}}^\ell_{\widetilde{\pi}^\star,h}, \widehat{\boldsymbol{\theta}}^\ell_h \rangle \ge V_0^{\widetilde{\pi}^\star} - \epsilon_\ell, \quad \sup_{\pi' \in \Pi_\ell} \sum_{h=1}^{H} \langle \widehat{\boldsymbol{\phi}}^\ell_{\pi',h}, \widehat{\boldsymbol{\theta}}^\ell_h \rangle \le \sup_{\pi' \in \Pi_\ell} \sum_{h=1}^{H} \langle \boldsymbol{\phi}_{\pi',h}, \boldsymbol{\theta}_h \rangle + \epsilon_\ell = V_0^{\widetilde{\pi}^\star} + \epsilon_\ell.$$

Rearranging this gives

$$\sum_{h=1}^{H} \langle \widehat{\boldsymbol{\phi}}^\ell_{\widetilde{\pi}^\star,h}, \widehat{\boldsymbol{\theta}}^\ell_h \rangle \ge \sup_{\pi' \in \Pi_\ell} \sum_{h=1}^{H} \langle \widehat{\boldsymbol{\phi}}^\ell_{\pi',h}, \widehat{\boldsymbol{\theta}}^\ell_h \rangle - 2\epsilon_\ell$$

so $\widetilde{\pi}^\star \in \Pi_{\ell+1}$. □

8.3.3 Proof of Theorem 8.1

Finally, we combine the above results in Lemma 8.3.8 to obtain a proof of Theorem 8.1.

Lemma 8.3.8. *With probability at least $1 - 2\delta$, Algorithm 8.1 will terminate after collecting*

at most

$$CH^4\sum_{h=1}^{H}\sum_{\ell=\ell_0+1}^{\varsigma_0}\frac{\inf_{\boldsymbol{\Lambda}\in\boldsymbol{\Omega}_h}\max_{\pi\in\Pi(4\epsilon_\ell)}\|\boldsymbol{\phi}_{\pi,h}\|^2_{\boldsymbol{\Lambda}^{-1}}}{\epsilon_\ell^2}\cdot\log\frac{H|\Pi(4\epsilon_\ell)|\log\frac{1}{\epsilon}}{\delta}$$
$$+\operatorname{poly}\left(d,H,\frac{1}{\bar{\lambda}_{\min}},\log\frac{1}{\delta},\log|\Pi|,\log\frac{1}{\epsilon}\right)+CH^4\sum_{h=1}^{H}\frac{\inf_{\boldsymbol{\Lambda}\in\boldsymbol{\Omega}_h}\max_{\pi\in\Pi}\|\boldsymbol{\phi}_{\pi,h}\|^2_{\boldsymbol{\Lambda}^{-1}}}{\epsilon_{\ell_0}^2}\cdot\log\frac{H|\Pi|\log(1/\epsilon)}{\delta}$$

episodes for $\varsigma_0:=\min\{\lceil\log\frac{4}{\epsilon}\rceil,\log\frac{4}{\Delta_{\min}(\Pi)}\}$*, and will output a policy* $\widehat{\pi}$ *such that*

$$V_0^{\widehat{\pi}}\geq\max_{\pi\in\Pi}V_0^{\pi}-\epsilon,$$

where here $\Pi(4\epsilon_\ell)=\{\pi\in\Pi\ :\ V_0^{\pi}\geq\max_{\pi\in\Pi}V_0^{\pi}-4\epsilon_\ell\}$.

Proof. By Lemma 8.3.6 the event $\mathcal{E}_{\text{est}}\cap\mathcal{E}_{\text{exp}}$ occurs with probability at least $1-2\delta$. Henceforth we assume we are on this event.

Correctness follows by Lemma 8.3.7, since upon termination, Π_ℓ will only contain policies π satisfying $V_0^{\pi}\geq\max_{\pi\in\Pi}V_0^{\pi}-\epsilon$ (and will contain at least 1 policy since $\widetilde{\pi}^\star\in\Pi_\ell$ for all ℓ). Furthermore, by Lemma 8.3.7, if $4\epsilon_\ell<\Delta_{\min}(\Pi)$, we must have that $\Pi_\ell=\{\widetilde{\pi}^\star\}$, and will therefore terminate on Line 11 since $|\Pi_\ell|=1$. Thus, we can bound the number of number of epochs by

$$\varsigma_0:=\min\{\lceil\log\frac{4}{\epsilon}\rceil,\log\frac{4}{\Delta_{\min}(\Pi)}\}.$$

By Lemma 8.3.4, the total number of episodes collected is bounded by

$$\sum_{h=1}^{H}\sum_{\ell=1}^{\varsigma_0}C\cdot\frac{\inf_{\boldsymbol{\Lambda}\in\boldsymbol{\Omega}_h}\max_{\boldsymbol{\phi}\in\Phi_{\ell,h}}\|\boldsymbol{\phi}\|^2_{\mathbf{A}(\boldsymbol{\Lambda})^{-1}}}{\epsilon_\ell^2/\beta_\ell}+\operatorname{poly}\left(d,H,\log\frac{1}{\delta},\frac{1}{\bar{\lambda}_{\min}},\log|\Pi|,\log\frac{1}{\epsilon}\right)$$
$$\leq\sum_{h=1}^{H}\sum_{\ell=1}^{\varsigma_0}C\cdot\frac{\inf_{\boldsymbol{\Lambda}\in\boldsymbol{\Omega}_h}\max_{\boldsymbol{\phi}\in\Phi_{\ell,h}}\|\boldsymbol{\phi}\|^2_{\mathbf{A}(\boldsymbol{\Lambda})^{-1}}}{\epsilon_\ell^2}\cdot H^4\log\frac{H|\Pi_\ell|\log(1/\epsilon)}{\delta}$$
$$+\operatorname{poly}\left(d,H,\log\frac{1}{\delta},\frac{1}{\bar{\lambda}_{\min}},\log|\Pi|,\log\frac{1}{\epsilon}\right).$$

On $\mathcal{E}_{\text{est}}\cap\mathcal{E}_{\text{exp}}$, by Lemma 8.3.6, for each $\pi\in\Pi_\ell$, we have $\|\widehat{\boldsymbol{\phi}}^{\ell}_{\pi,h}-\boldsymbol{\phi}_{\pi,h}\|_2\leq d\epsilon_\ell/2H$. As $\Phi_{\ell,h}=\{\widehat{\boldsymbol{\phi}}^{\ell}_{\pi,h}:\pi\in\Pi_\ell\}$, it follows that we can upper bound

$$\inf_{\boldsymbol{\Lambda}\in\boldsymbol{\Omega}_h}\max_{\boldsymbol{\phi}\in\Phi_{\ell,h}}\|\boldsymbol{\phi}\|^2_{\mathbf{A}(\boldsymbol{\Lambda})^{-1}}=\inf_{\boldsymbol{\Lambda}\in\boldsymbol{\Omega}_h}\max_{\pi\in\Pi_\ell}\|\widehat{\boldsymbol{\phi}}^{\ell}_{\pi,h}\|^2_{\mathbf{A}(\boldsymbol{\Lambda})^{-1}}$$
$$\leq\inf_{\boldsymbol{\Lambda}\in\boldsymbol{\Omega}_h}\max_{\pi\in\Pi_\ell}(2\|\boldsymbol{\phi}_{\pi,h}\|^2_{\mathbf{A}(\boldsymbol{\Lambda})^{-1}}+2\|\widehat{\boldsymbol{\phi}}^{\ell}_{\pi,h}-\boldsymbol{\phi}_{\pi,h}\|^2_{\mathbf{A}(\boldsymbol{\Lambda})^{-1}})$$

$$
\begin{aligned}
&\leq \inf_{\mathbf{\Lambda}\in\mathbf{\Omega}_h} \max_{\pi\in\Pi_\ell} (2\|\boldsymbol{\phi}_{\pi,h}\|^2_{\mathbf{A}(\mathbf{\Lambda})^{-1}} + \frac{d^2\epsilon_\ell^2}{2H^2\lambda_{\min}(\mathbf{A}(\mathbf{\Lambda}))}) \\
&\leq \inf_{\mathbf{\Lambda}\in\mathbf{\Omega}_h} \max_{\pi\in\Pi_\ell} 4\|\boldsymbol{\phi}_{\pi,h}\|^2_{\mathbf{A}(\mathbf{\Lambda})^{-1}} + \inf_\pi \frac{d^2\epsilon_\ell^2}{H^2\lambda_{\min}(\mathbf{A}(\mathbf{\Lambda}))} \\
&\leq \inf_{\mathbf{\Lambda}\in\mathbf{\Omega}_h} \max_{\pi\in\Pi_\ell} 4\|\boldsymbol{\phi}_{\pi,h}\|^2_{\mathbf{\Lambda}^{-1}} + \frac{d^2\epsilon_\ell^2}{H^2\bar{\lambda}_{\min}}
\end{aligned}
$$

so

$$
\frac{\inf_{\mathbf{\Lambda}\in\mathbf{\Omega}_h} \max_{\boldsymbol{\phi}\in\Phi_{\ell,h}} \|\boldsymbol{\phi}\|^2_{\mathbf{A}(\mathbf{\Lambda})^{-1}}}{\epsilon_\ell^2} \leq \frac{\inf_{\mathbf{\Lambda}\in\mathbf{\Omega}_h} \max_{\pi\in\Pi_\ell} 4\|\boldsymbol{\phi}_{\pi,h}\|^2_{\mathbf{\Lambda}^{-1}}}{\epsilon_\ell^2} + \frac{d^2}{H^2\bar{\lambda}_{\min}}.
$$

Note also that, by Lemma 8.3.7, for $\ell > \ell_0$, every policy $\pi \in \Pi_\ell$ will be $4\epsilon_\ell$ optimal, so we therefore have

$$
\Pi_\ell \subseteq \{\pi \in \Pi \ : \ V_0^\pi \geq V_0^{\widetilde{\pi}^\star} - 4\epsilon_\ell\} =: \Pi(4\epsilon_\ell).
$$

Putting this together, we can upper bound the complexity by

$$
\begin{aligned}
&\sum_{h=1}^H \sum_{\ell=\ell_0+1}^{\varsigma_0} C \cdot \frac{\inf_{\mathbf{\Lambda}\in\mathbf{\Omega}_h} \max_{\pi\in\Pi(4\epsilon_\ell)} \|\boldsymbol{\phi}_{\pi,h}\|^2_{\mathbf{\Lambda}^{-1}}}{\epsilon_\ell^2} \cdot H^4 \log \frac{H|\Pi(4\epsilon_\ell)|\log(1/\epsilon)}{\delta} \\
&+ \mathrm{poly}\left(d, H, \log\frac{1}{\delta}, \frac{1}{\bar{\lambda}_{\min}}, \log|\Pi|, \log\frac{1}{\epsilon}\right) + C \cdot \frac{\inf_{\mathbf{\Lambda}\in\mathbf{\Omega}_h} \max_{\pi\in\Pi} \|\boldsymbol{\phi}_{\pi,h}\|^2_{\mathbf{\Lambda}^{-1}}}{\epsilon_{\ell_0}^2} \cdot H^4 \log \frac{H|\Pi|\log(1/\epsilon)}{\delta}.
\end{aligned}
$$

□

Proof of Theorem 8.1. By the definition of $\Pi(4\epsilon_\ell)$, for each $\pi \in \Pi(4\epsilon_\ell)$ we have

$$
\epsilon_\ell^2 = \frac{1}{16}\left((V_0^\star(\Pi) - V_0^\pi)^2 \vee (4\epsilon_\ell)^2\right).
$$

We can therefore upper bound

$$
\begin{aligned}
&\sum_{\ell=\ell_0+1}^{\varsigma_0} \frac{\inf_{\mathbf{\Lambda}\in\mathbf{\Omega}_h} \max_{\pi\in\Pi(4\epsilon_\ell)} \|\boldsymbol{\phi}_{\pi,h}\|^2_{\mathbf{\Lambda}^{-1}}}{\epsilon_\ell^2} \cdot \log \frac{H|\Pi(4\epsilon_\ell)|\log\frac{1}{\epsilon}}{\delta} \\
&\leq C \sum_{\ell=\ell_0+1}^{\varsigma_0} \inf_{\mathbf{\Lambda}\in\mathbf{\Omega}_h} \max_{\pi\in\Pi(4\epsilon_\ell)} \frac{\|\boldsymbol{\phi}_{\pi,h}\|^2_{\mathbf{\Lambda}^{-1}}}{(V_0^\star(\Pi) - V_0^\pi)^2 \vee \epsilon_\ell^2} \cdot \log \frac{H|\Pi(4\epsilon_\ell)|\log\frac{1}{\epsilon}}{\delta} \\
&\leq C \log\frac{1}{\epsilon} \cdot \inf_{\mathbf{\Lambda}\in\mathbf{\Omega}_h} \max_{\pi\in\Pi} \frac{\|\boldsymbol{\phi}_{\pi,h}\|^2_{\mathbf{\Lambda}^{-1}}}{(V_0^\star(\Pi) - V_0^\pi)^2 \vee \epsilon^2 \vee \Delta_{\min}(\Pi)^2} \cdot \log \frac{H|\Pi|\log\frac{1}{\epsilon}}{\delta}.
\end{aligned}
$$

Furthermore, since $\ell_0 = \lceil \log_2 d^{3/2}/H \rceil$, using Lemma F.3.3 we can also bound

$$C \cdot \frac{\inf_{\mathbf{\Lambda} \in \mathbf{\Omega}_h} \max_{\pi \in \Pi} \|\boldsymbol{\phi}_{\pi,h}\|^2_{\mathbf{\Lambda}^{-1}}}{\epsilon^2_{\ell_0}} \cdot H^4 \log \frac{H|\Pi| \log(1/\epsilon)}{\delta} \leq \text{poly}\left(d, H, \log 1/\delta, \log|\Pi|, \log 1/\epsilon\right).$$

The result then follows by Lemma 8.3.8. □

8.4 XY-Optimal Design

Key to achieving the rate of Theorem 8.1 is efficiently collecting data to ensure the condition on Line 6 of Algorithm 8.1, $\max_{\pi \in \Pi_\ell} \|\widehat{\boldsymbol{\phi}}^\ell_{\pi,h}\|^2_{\mathbf{\Lambda}^{-1}_{h,\ell}} \leq \frac{\epsilon^2_\ell}{\beta_\ell}$, is met by collecting the minimal amount of samples possible. In this section we argue that this can be achieved by applying the DynamicCOED algorithm of Chapter 6 to this problem.

In particular, we consider running the following procedure, which relies on the DynamicCOED algorithm of Chapter 6 to minimize a sequence of objective $(f_i)_i$.

Algorithm 8.2 Collect Optimal Covariates (OptCov)

1: **input**: functions to optimize $(f_i)_i$, constraint tolerance ϵ, confidence δ
2: **for** $i = 1, 2, 3, \ldots$ **do**
3: $N_i \leftarrow 2^i$, $K_i \leftarrow 2^{2i}$
`// Force algorithm from Wagenmaker et al. [2022a], Π_exp all possible policies on linear MDP`
4: $\widehat{\mathbf{\Lambda}}_i, \mathfrak{D}_i \leftarrow$ DynamicCOED $(f_i, N_i - 1, K_i, \frac{\delta}{2i^2}, \text{Force}, \Pi_{\text{exp}})$ (Algorithm 6.2)
5: **if** $f_i(\widehat{\mathbf{\Lambda}}_i) \leq K_i N_i \epsilon$ **then**
6: **return** $\widehat{\mathbf{\Lambda}}_i$, $K_i N_i$, $\mathfrak{D}_i$

The following result shows that this procedure succeeds, for a generic set of objectives satisfying Assumption 6.9.

Lemma 8.4.1. *Let $(f_i)_i$ denote some sequence of functions which satisfy Assumption 6.9 with constants (β_i, M_i) and assume $\beta_i \geq 1$. Let (β, M) be some values such that $\beta_i \leq \beta, M_i \leq M$ for al i, and let f be some function such that $f_i(\mathbf{\Lambda}) \leq f(\mathbf{\Lambda})$ for all i and $\mathbf{\Lambda} \succeq 0$. Denote $f_{\min}$ a lower bound on all f_i: $\min_i \inf_{\mathbf{\Lambda} \in \mathbf{\Omega}} f_i(\mathbf{\Lambda}) \geq f_{\min}$.*

Then, if we run Algorithm 8.2 on $(f_i)_i$ with constraint tolerance ϵ and confidence δ, we have that with probability at least $1 - \delta$, it will run for at most

$$\frac{19 \cdot \min_{\mathbf{\Lambda} \in \mathbf{\Omega}} f(\mathbf{\Lambda})}{\epsilon} + \text{poly}\left(\beta, R, M, d, H, f^{-1}_{\min}, \log \frac{1}{\delta}\right)$$

episodes, and will return data $\{\boldsymbol{\phi}_\tau\}_{\tau=1}^T$ with covariance $\widehat{\mathbf{\Sigma}}_T = \sum_{\tau=1}^T \boldsymbol{\phi}_\tau \boldsymbol{\phi}_\tau^\top$ such that

$$f_{\widehat{i}}(T^{-1}\widehat{\mathbf{\Sigma}}_T) \leq T\epsilon,$$

where $\widehat{i}$ is the iteration on which OptCov *terminates.*

Optimally collecting data at each epoch can be thought of as optimizing the function

$$\mathsf{XY}_{\text{opt}}(\mathbf{\Lambda}) = \max_{\boldsymbol{\phi}\in\Phi} \|\boldsymbol{\phi}\|^2_{\mathbf{A}(\mathbf{\Lambda})^{-1}} \quad \text{for} \quad \mathbf{A}(\mathbf{\Lambda}) = \mathbf{\Lambda} + \mathbf{\Lambda}_0$$

for $\mathbf{\Lambda} \in \mathbf{\Omega}$ and with $\mathbf{\Lambda}_0 \succ 0$ some fixed regularizer. We refer to this as an XY-*Optimal Design*. We would like to apply Lemma 8.4.1 to optimally collect covariates minimizing this objective. $\mathsf{XY}_{\text{opt}}(\mathbf{\Lambda})$, however, is not smooth, as required by Lemma 8.4.1, so we relax it to the following:

$$\widetilde{\mathsf{XY}}_{\text{opt}}(\mathbf{\Lambda}) := \text{LogSumExp}\left(\{e^{\eta\|\boldsymbol{\phi}\|^2_{\mathbf{A}(\mathbf{\Lambda})^{-1}}}\}_{\boldsymbol{\phi}\in\Phi}; \eta\right) = \frac{1}{\eta}\log\left(\sum_{\boldsymbol{\phi}\in\Phi} e^{\eta\|\boldsymbol{\phi}\|^2_{\mathbf{A}(\mathbf{\Lambda})^{-1}}}\right). \tag{8.4.1}$$

We show in Appendix F.4.1 that $\widetilde{\mathsf{XY}}_{\text{opt}}(\mathbf{\Lambda})$ satisfies the smoothness conditions required in Lemma 8.4.1, and that for properly chosen η the relaxation does not significantly affect the minimization of the original objective. Given this, we then instantiate OptCov and Lemma 8.4.1 to minimize the precise objective required by Pedel.

Theorem 8.5. *Consider running* OptCov *with some ϵ_{exp} and functions*

$$f_i(\mathbf{\Lambda}) \leftarrow \widetilde{\mathsf{XY}}_{\text{opt}}(\mathbf{\Lambda})$$

for $\mathbf{\Lambda}_0 \leftarrow (T_iK_i)^{-1}\mathbf{\Sigma}_i =: \mathbf{\Lambda}_i$ and

$$\eta_i = \frac{2}{\gamma_\Phi}\cdot(1+\|\mathbf{\Lambda}_i\|_{\text{op}})\cdot\log|\Phi|, \quad \beta_i = 2\|\mathbf{\Lambda}_i^{-1}\|^3_{\text{op}}(1+\eta_i\|\mathbf{\Lambda}_i^{-1}\|_{\text{op}}), \quad M_i = \|\mathbf{\Lambda}_i^{-1}\|^2_{\text{op}}$$

where $\gamma_\Phi := \max_{\boldsymbol{\phi}\in\Phi}\|\boldsymbol{\phi}\|_2$ and $\mathbf{\Sigma}_i$ is the matrix returned by running the procedure of Lemma F.4.7 with parameters $N \leftarrow T_iK_i$, $\delta \leftarrow \delta/(2i^2)$, and some $\underline{\lambda} \geq 0$, in order to collect full-rank covariates. Then with probability $1-2\delta$, this procedure will collect at most

$$40\cdot\frac{\inf_{\mathbf{\Lambda}\in\mathbf{\Omega}}\max_{\boldsymbol{\phi}\in\Phi}\|\boldsymbol{\phi}\|^2_{\mathbf{A}(\mathbf{\Lambda})^{-1}}}{\epsilon_{\text{exp}}} + \text{poly}\left(d, H, \log\frac{1}{\delta}, \frac{1}{\bar{\lambda}_{\min}}, \frac{1}{\gamma_\Phi}, \underline{\lambda}, \log|\Phi|, \log\frac{1}{\epsilon_{\text{exp}}}\right)$$

episodes, where

$$\mathbf{A}(\mathbf{\Lambda}) = \mathbf{\Lambda} + \text{poly}\left(d, \frac{1}{\bar{\lambda}_{\min}}, H, \log\frac{\lambda}{\delta}\right)^{-1}\cdot I,$$

and will produce covariates $\widehat{\mathbf{\Sigma}} + \mathbf{\Sigma}_i$ such that

$$\max_{\boldsymbol{\phi}\in\Phi}\|\boldsymbol{\phi}\|^2_{(\widehat{\mathbf{\Sigma}}+\mathbf{\Sigma}_i)^{-1}} \leq \epsilon_{\text{exp}} \quad \textit{and} \quad \lambda_{\min}(\widehat{\mathbf{\Sigma}}+\mathbf{\Sigma}_i) \geq \max\{d\log 1/\delta, \underline{\lambda}\}.$$

Proof. Note that the total failure probability of our calls to the procedure of Lemma F.4.7 is

at most

$$\sum_{i=1}^{\infty} \frac{\delta}{2i^2} = \frac{\pi^2}{12}\delta \le \delta.$$

For the remainder of the proof, we will then assume that we are on the success event defined in Lemma F.4.7.

By Lemma F.4.5, $f_i(\mathbf{\Lambda})$ satisfies Assumption 6.9 with constants

$$\beta_i = 2\|\mathbf{\Lambda}_i^{-1}\|_{\mathrm{op}}^3(1 + \eta_i\|\mathbf{\Lambda}_i^{-1}\|_{\mathrm{op}}), \quad M_i = \|\mathbf{\Lambda}_i^{-1}\|_{\mathrm{op}}^2$$

for $\mathbf{\Lambda}_i \leftarrow (T_iK_i)^{-1}\mathbf{\Sigma}_i$.

By Lemma F.4.7, on the success event of Lemma F.4.7 we have that

$$\lambda_{\min}(\mathbf{\Lambda}_i) \ge \frac{1}{C^i_{\lambda_{\min}}} + \max\{d\log 1/\delta, \underline{\lambda}\}/(T_iK_i)$$

for $C^i_{\lambda_{\min}} = \mathrm{poly}\left(d, \frac{1}{\underline{\lambda}_{\min}}, H, \log\frac{\underline{\lambda}}{\delta}, i\right)$. Assume that the termination condition of OptCov for $\widehat{i}$ satisfying

$$\widehat{i} \le \log\Bigg(\underbrace{\mathrm{poly}\left(\frac{1}{\epsilon_{\exp}}, d, H, \log 1/\delta, \frac{1}{\underline{\lambda}_{\min}}, \frac{1}{\gamma_\Phi}, \underline{\lambda}, \log|\Phi|\right)}_{=:\varsigma}\Bigg). \tag{8.4.2}$$

We assume this holds and justify it at the conclusion of the proof. Then given this bound on $\widehat{i}$ and lower bound on $\lambda_{\min}(\mathbf{\Lambda}_i)$, we can bound, for all i:

$$M_i, \beta_i \le \mathrm{poly}\left(d, H, \log 1/\delta, \frac{1}{\underline{\lambda}_{\min}}, \frac{1}{\gamma_\Phi}, \underline{\lambda}, \log|\Phi|\right) =: M, \beta.$$

Now take $f(\mathbf{\Lambda}) \leftarrow \widetilde{\mathsf{XY}}_{\mathrm{opt}}(\mathbf{\Lambda}; \eta, \mathbf{\Lambda}_0)$ with $\mathbf{\Lambda}_0 \leftarrow 1/\mathrm{poly}(d, \frac{1}{\underline{\lambda}_{\min}}, H, \log\frac{\underline{\lambda}}{\delta}) \cdot I$ and $\eta \leftarrow \frac{2\log|\Phi|}{\gamma_\Phi} \cdot (1 + \|\mathbf{\Lambda}_0\|_{\mathrm{op}})$. Note that in this case, we have $\|\mathbf{\Lambda}_0\|_{\mathrm{op}} \le \lambda_{\min}(\mathbf{\Lambda}_i)$ for all i, so $\mathbf{\Lambda}_0 \preceq \mathbf{\Lambda}_i$ and $\eta \le \eta_i$. By the construction of $\widetilde{\mathsf{XY}}_{\mathrm{opt}}$ and Lemma F.4.2, it follows that $f(\mathbf{\Lambda}) \ge f_i(\mathbf{\Lambda})$ for all $\mathbf{\Lambda} \succeq 0$, so this is a valid choice of f, as required by Lemma 8.4.1. Furthermore, we can set $R = 2$, since $\|\mathbb{E}_{(s,a)\sim\omega}[\boldsymbol{\phi}(s,a)\boldsymbol{\phi}(s,a)^\top]\|_{\mathrm{F}} \le 1$ for all $\omega \in \triangle_{\mathcal{S}\times\mathcal{A}}$.

To apply Lemma 8.4.1, it remains only to find a suitable value of $f_{\min}$. By Lemma F.4.1 and Lemma F.4.3, we can lower bound f_i by $\frac{\gamma_\Phi}{1+\|\mathbf{\Lambda}_i\|_{\mathrm{op}}}$. By Lemma F.4.7, we can lower bound

$$\frac{\gamma_\Phi}{1 + \|\mathbf{\Lambda}_i\|_{\mathrm{op}}} \ge \frac{\gamma_\Phi}{2 + \mathrm{poly}(d, \frac{1}{\underline{\lambda}_{\min}}, H, \log\frac{1}{\delta}, \underline{\lambda})}.$$

We then take this as our choice of $f_{\min}$.

We can now apply Lemma 8.4.1 and get that with probability at least $1-\delta$, OptCov will terminate in

$$N \leq \frac{19 \inf_{\mathbf{\Lambda} \in \mathbf{\Omega}} f(\mathbf{\Lambda})}{\epsilon_{\text{exp}}} + \text{poly}\left(d, H, \log \frac{1}{\delta}, \frac{1}{\bar{\lambda}_{\min}}, \frac{1}{\gamma_{\Phi}}, \underline{\lambda}, \log |\Phi|\right)$$

episodes, and will return (time-normalized) covariates $\widehat{\mathbf{\Lambda}}$ such that

$$f_{\widehat{i}}(\widehat{\mathbf{\Lambda}}) \leq N \epsilon_{\text{exp}}.$$

By Lemma F.4.4, our choice of η and $\mathbf{\Lambda}_0$, we can upper bound

$$\inf_{\mathbf{\Lambda} \in \mathbf{\Omega}} f(\mathbf{\Lambda}) \leq 2 \inf_{\mathbf{\Lambda} \in \mathbf{\Omega}} \max_{\boldsymbol{\phi} \in \Phi} \|\boldsymbol{\phi}\|^2_{\mathbf{A}(\mathbf{\Lambda})^{-1}}$$

where here $\mathbf{A}(\mathbf{\Lambda}) = \mathbf{\Lambda} + \mathbf{\Lambda}_0$. Furthermore, by Lemma F.4.1 we have

$$\max_{\boldsymbol{\phi} \in \Phi} \|\boldsymbol{\phi}\|^2_{(\widehat{\mathbf{\Sigma}} + \mathbf{\Lambda}_0)^{-1}} \leq f_{\widehat{i}}(\widehat{\mathbf{\Lambda}}).$$

The final upper bound on the number of episodes collected and the lower bound on the minimum eigenvalue of the covariates follows from Lemma F.4.7.

It remains to justify our bound on $\widehat{i}$, (8.4.2). Note that by definition of OptCov, if we run for a total of $\bar{N}$ episodes, we can bound $\widehat{i} \leq \frac{1}{3}\log_2(\bar{N})$. However, we see that the bound on $\widehat{i}$ given in (8.4.2) upper bounds $\frac{1}{3}\log_2(\bar{N})$ for $\bar{N}$ the upper bound on the number of samples collected by OptCov stated above. Thus, our bound on $\widehat{i}$ is valid. □

Chapter 9

Optimal Reward-Free Reinforcement Learning

The previous two chapters have considered the problem of finding a policy to maximize a fixed reward function. In many scenarios, however, we may not have a single reward we wish to maximize, but may wish to learn enough about the environment to be able to maximize any reward function presented to the agent at deployment. In such settings, our goal, instead of learning a single policy π, is to learn a mapping which takes as input a reward function r, and outputs a policy $\pi(r)$ maximizing this reward.

We call the former, single-reward setting the *reward-aware* setting, while the latter setting we call the *reward-free* setting. In the reward-aware setting, as we saw with the algorithmic approaches of Chapter 7 and Chapter 8, we could focus our exploration on regions most relevant to the particular reward of interest. In contrast, in the reward-free setting, we must explore the entire space sufficiently such that, if some region of the space were to become relevant under some reward r, we would know how to behave there.

We consider again the linear MDP setting. Instead of focusing on obtaining instance-dependent guarantees, however, we will be satisfied with minimax guarantees. Our goal will be to understand the relationship between the reward-aware and reward-free setting, and the role of exploration in achieving tight bounds in reward free RL. As we will see, while the reward-aware and reward-free settings differ in where we should focus our exploration, the key to achieving tight bounds in each setting is to explore efficiently. Indeed, the exploration procedure proposed in this chapter is very similar to that considered in Chapter 7.

The reward-free setting has been considered in both tabular and linear MDPs in the past. In the tabular setting, it is widely known that the optimal minimax sample complexity for reward-free RL is (ignoring horizon factors) $\Theta(S^2A/\epsilon^2)$ [Jin et al., 2018]. In contrast, in the reward-aware setting, the optimal minimax sample complexity scales as $\mathcal{O}(SA/\epsilon^2)$ [Dann et al., 2019, Ménard et al., 2021], a factor of S less than the optimal reward-free setting. This implies that, in tabular RL, reward-aware RL is *easier* than reward-free RL—it is easier to learn a policy maximizing a single reward than to learn a policy maximizing any number of yet-to-be-revealed rewards. In the linear MDP setting, while a variety of works have studied the reward-free problem, and obtained sample complexities scaling polynomially in d, H, and

$1/\epsilon$ [Zanette et al., 2020c, Wang et al., 2020, Zhang et al., 2021a], prior to the results given in this chapter, the optimal rate was unknown. In this chapter we seek to close this gap, and determine the optimal rate for reward-free RL in linear MDPs.

We formally define the reward-free RL setting as follows.

Reward Free RL. The reward-free RL setting stands in contrast to the reward-aware RL setting in that the learner does not observe the rewards (or have any access to the reward function) while exploring. The reward-free RL setting proceeds in two phases:

1. Given (ϵ, δ), without observing any rewards, the learner explores an MDP for K episodes, where K is of the learner's choosing, collecting trajectories
$$\{(s_{1,k}, a_{1,k}, s_{2,k}, a_{2,k}, \ldots, s_{H,k}, a_{H,k})\}_{k=1}^{K}.$$
2. The learner outputs a map $\widehat{\pi}(\cdot)$ which takes as input a reward function and outputs a policy such that $V_0^{\star}(r) - V_0^{\widehat{\pi}(r)}(r) \leq \epsilon$ for all valid reward functions r simultaneously, with probability at least $1 - \delta$.

As in reward-aware RL, the goal is to obtain a procedure able to provide the above guarantee with probability at least $1 - \delta$, and do so using as few episodes as possible.

For simplicity, in this chapter, we assume all reward functions are deterministic, and denote reward functions by r. We state our main results on reward-free RL in Section 9.2 but first describe our exploration procedure, which is the key tool in achieving tight bounds on reward-free RL.

9.1 Efficient Exploration in Linear MDPs

Our main algorithmic technique is an exploration procedure able to efficiently generate "covering trajectories" in linear MDPs. Before presenting our MDP covering algorithm, CoverTraj, we describe its key subroutine, Egs.

Explore Goal Set (Egs) Subroutine. Egs takes as input a target set $\mathcal{X}$ to be explored; we fix $\mathcal{X}$ in what follows. It then creates an "exploration reward function"—which places a high reward on regions of $\mathcal{X}$ for which we have large uncertainty—and then runs a regret minimization algorithm on this reward function to direct exploration to these regions. More specifically, we define the reward function

$$r(\boldsymbol{\phi}; \boldsymbol{\Lambda}) \leftarrow \begin{cases} 1 & \|\boldsymbol{\phi}\|^2_{\boldsymbol{\Lambda}^{-1}} > \boldsymbol{\gamma}^2, \boldsymbol{\phi} \in \mathcal{X} \\ \frac{1}{\boldsymbol{\gamma}^2}\|\boldsymbol{\phi}\|^2_{\boldsymbol{\Lambda}^{-1}} & \|\boldsymbol{\phi}\|^2_{\boldsymbol{\Lambda}^{-1}} \leq \boldsymbol{\gamma}^2, \boldsymbol{\phi} \in \mathcal{X} \\ 0 & \boldsymbol{\phi} \notin \mathcal{X} \end{cases}. \tag{9.1.1}$$

Algorithm 9.1 Explore Goal Set (EGS)

1: **input:** goal set $\mathcal{X} \subseteq \mathbb{R}^d$, step h, number of episodes K, collection tolerance $\boldsymbol{\gamma}^2$, regret minimization algorithm REGMIN
2: Run REGMIN for K episodes, using reward $r_h^k(s,a) := r(\boldsymbol{\phi}(s,a); \boldsymbol{\Lambda}_{h,k-1})$ at episode k, for r defined as in (9.1.1), $\boldsymbol{\Lambda}_{h,k-1} = I + \sum_{\tau=1}^{k-1} \boldsymbol{\phi}_{h,\tau}\boldsymbol{\phi}_{h,\tau}^\top$, and $r_{h'}^k(s,a) = 0$ for $h' \neq h$
3: Collect transitions observed at step h

$$\mathcal{D} \leftarrow \{(s_{h,\tau}, a_{h,\tau}, s_{h+1,\tau})\}_{\tau=1}^K$$

4: **return** $\{\boldsymbol{\phi} \in \mathcal{X} \ : \ \boldsymbol{\phi}^\top \boldsymbol{\Lambda}_{h,K}^{-1} \boldsymbol{\phi} \leq \boldsymbol{\gamma}^2\}$, $\mathcal{D}$, $\boldsymbol{\Lambda}_{h,K}$

At the kth episode, we instantiate our reward as $r_h^k(s,a) := r(\boldsymbol{\phi}(s,a); \boldsymbol{\Lambda}_{h,k-1})$, where $\boldsymbol{\Lambda}_{h,k-1} = I + \sum_{\tau=1}^{k-1} \boldsymbol{\phi}_{h,\tau}\boldsymbol{\phi}_{h,\tau}^\top$ are the covariates collected on the first $k-1$ episodes. This captures the fact that we have collected some information over the first $k-1$ episodes, so our uncertainty in direction $\boldsymbol{\phi}$ scales as $\|\boldsymbol{\phi}\|_{\boldsymbol{\Lambda}_{h,k-1}^{-1}}$. By choosing our reward function to increase as $\|\boldsymbol{\phi}\|_{\boldsymbol{\Lambda}_{h,k-1}^{-1}}$ increases (for $\boldsymbol{\phi} \in \mathcal{X}$), we incentive exploring directions in $\boldsymbol{\phi} \in \mathcal{X}$ with large uncertainty. EGS also takes as input a scalar tolerance $\boldsymbol{\gamma}^2$ and, after running for K episodes, returns the set of $\boldsymbol{\phi} \in \mathcal{X}$ which have been explored up to tolerance $\boldsymbol{\gamma}^2$.

In order to efficiently collect data, we require a regret-minimization algorithm which achieves low regret with respect to a *time-varying* reward function, r^k. We define regret with respect to such a reward function as

$$\mathcal{R}_K := \sum_{k=1}^K [V_0^\star(r^k) - V_0^{\pi_k}(r^k)].$$

To achieve the optimal scaling in d, REGMIN must attain *first order* regret in the following sense.

Definition 9.1.1 (First-Order Regret Minimization Algorithm)**.** Consider a time-varying reward function r^k that is $\mathcal{F}_{k-1}$-measurable, satisfies $r_h^k(s,a) \in [0,1]$, and is non-increasing in k (that is, $r_h^k(s,a) \leq r_h^{k-1}(s,a)$ for all s, a, h, k). Then we call a regret minimization algorithm REGMIN a *first-order regret minimization algorithm* if it achieves regret bounded as, with probability at least $1-\delta$,

$$\mathcal{R}_K \leq \sqrt{\mathcal{C}_1 V_0^\star(r^1) K \cdot \log^{p_1}(HK/\delta)} + \mathcal{C}_2 \log^{p_2}(HK/\delta)$$

for constants $\mathcal{C}_1, \mathcal{C}_2, p_1, p_2$ which do not depend on K.

In general, we can think of $\mathcal{C}_1$ and $\mathcal{C}_2$ as $\mathrm{poly}(d, H)$ and p_1 and p_2 as absolute constants. The following definition will be helpful in quantifying how easily a set of directions are to reach.

Definition 9.1.2 (Set Visitation)**.** For any $\mathcal{X} \subseteq \mathbb{R}^d$ and policy π, we let

$$w_h^\pi(\mathcal{X}) := \mathbb{P}_\pi[\boldsymbol{\phi}(s_h, a_h) \in \mathcal{X}]$$

denote the probability of visiting $\mathcal{X}$ under π at step h. Furthermore, we will say that $\mathcal{X}$ is *c-unreachable* if $\sup_\pi w_h^\pi(\mathcal{X}) \le c$.

We are now ready to present our main exploration algorithm, CoverTraj, in Algorithm 9.2.

Algorithm 9.2 Collect Covering Trajectories (CoverTraj)

input: step h, confidence δ, number of epochs m, tolerance vector $\boldsymbol{\gamma}^2 \in [0,1]^m$, regret minimization algorithm RegMin (default: Force, see Section 9.1.1)
$\mathcal{X} \leftarrow \mathcal{B}^d$
for $i = 1, 2, \ldots, m$ **do**
    Set K_i as in (9.1.2)
    $\mathcal{X}_i, \mathcal{D}_i, \boldsymbol{\Lambda}_i \leftarrow \text{Egs}(\mathcal{X}, h, K_i, \boldsymbol{\gamma}_i^2, \text{RegMin})$
    $\mathcal{X} \leftarrow \mathcal{X} \backslash \mathcal{X}_i$
return $\{(\mathcal{X}_i, \mathcal{D}_i, \boldsymbol{\Lambda}_i)\}_{i=1}^m$

CoverTraj Algorithm. CoverTraj partitions the feature space by repeatedly calling Egs while exponentially increasing the number of episodes Egs is run for. As the number of episodes Egs is run for increases, Egs is able to reach harder and harder to reach feature directions, while ensuring easier to reach directions have been explored up to their desired tolerance $\boldsymbol{\gamma}_i^2$. Specifically, at each epoch i, Egs runs for

$$K_i \leftarrow \widetilde{\mathcal{O}}\left(2^i \cdot \max\left\{\frac{d}{\boldsymbol{\gamma}_i^2} \log \tfrac{2^i}{\boldsymbol{\gamma}_i^2}, \mathcal{C}_1 (mp_1)^{p_1} \log^{p_1} \tfrac{1}{\delta}, \mathcal{C}_2 (mp_2)^{p_2} \log^{p_2} \tfrac{1}{\delta}\right\}\right) \tag{9.1.2}$$

episodes. CoverTraj enjoys the following guarantee.

Theorem 9.1. *Fix $h \in [H]$, $\delta > 0$, $m \ge 1$, and set $\boldsymbol{\gamma}^2 \in [0,1]^m$ to desired tolerances. Consider running CoverTraj with a regret minimization algorithm RegMin satisfying Definition 9.1.1, and let $\{(\mathcal{X}_i, \mathcal{D}_i, \boldsymbol{\Lambda}_i)\}_{i=1}^m$ denote the arguments returned. Then, with probability at least $1 - \delta$, for each $i \in [m]$ simultaneously:*

$$\sup_\pi w_h^\pi(\mathcal{X}_i) \le 2^{-i+1} \quad \text{and} \quad \boldsymbol{\phi}^\top \boldsymbol{\Lambda}_i^{-1} \boldsymbol{\phi} \le \boldsymbol{\gamma}_i^2, \forall \boldsymbol{\phi} \in \mathcal{X}_i$$

and, on the same event, it also holds that

$$\sup_\pi w_h^\pi(\mathcal{B}^d \backslash \cup_{i=1}^m \mathcal{X}_i) \le 2^{-m}.$$

Furthermore, CoverTraj terminates after at most $\sum_{i=1}^{m} K_i$ *episodes, for* K_i *as in* (9.1.2).

Theorem 9.1 shows that CoverTraj partitions the feature space into sets $\mathcal{X}_i$ such that $\mathcal{X}_i$ is 2^{-i+1}-unreachable, and where we have learned every $\boldsymbol{\phi} \in \mathcal{X}_i$ up to tolerance $\boldsymbol{\gamma}_i^2$: $\|\boldsymbol{\phi}\|^2_{\boldsymbol{\Lambda}_i^{-1}} \le \boldsymbol{\gamma}_i^2$. Furthermore, it takes roughly $2^i \cdot \frac{d}{\boldsymbol{\gamma}_i^2}$ episodes of exploration to accomplish this, which is the intuitively correct rate, as the following example illustrates.

Example 9.1.1 (Tabular MDPs)**.** Consider a tabular MDP with $H = 2$, state space $\mathcal{S}$ with $|\mathcal{S}| < \infty$, and A actions (we think of $A \ll |\mathcal{S}|$ as an absolute constant). Assume we always start in state s_0 and can break the state space into sets $\mathcal{S}_1, \dots, \mathcal{S}_A$ such that $|\mathcal{S}_1| = \mathcal{O}(|\mathcal{S}|/A)$ and where, if we take action a_i, we will end up in $\mathcal{S}_i$ with probability 2^{-i} (with equal probability of being in any particular state within $\mathcal{S}_i$), and s_0 with probability $1 - 2^{-i}$. Representing this as a linear MDP, we have $d = A|\mathcal{S}|$ and $\boldsymbol{\phi}(s, a) = \boldsymbol{e}_{sa}$, a standard basis vector. As such, $\boldsymbol{\Lambda}_K$, the covariates collected over K episodes, is diagonal with $[\boldsymbol{\Lambda}_K]_{sa} = N(s, a)$ the number of visits to s, a.

Now assume we want to ensure that $\boldsymbol{\phi}(s, a)^\top \boldsymbol{\Lambda}_K^{-1} \boldsymbol{\phi}(s, a) \le \boldsymbol{\gamma}_i^2$ for some $\boldsymbol{\gamma}_i^2$, each $s \in \mathcal{S}_i$, and all $a \in [A]$. For any given $s \in \mathcal{S}_i$, if we take action a_i in state s_0, we will arrive in state s with probability $2^{-i}/|\mathcal{S}_i|$. Thus, in expectation, to collect N samples from s, we must run for at least

$$\frac{N}{2^{-i}/|\mathcal{S}_i|} = 2^i |\mathcal{S}_i| N$$

episodes. It follows that to collect N samples from each $s \in \mathcal{S}_i$ and all $a \in [A]$, it will take at least $2^i |\mathcal{S}_i| AN$ episodes. Furthermore, by the above observation that $[\boldsymbol{\Lambda}_K]_{sa} = N(s, a)$, achieving $\boldsymbol{\phi}(s, a)^\top \boldsymbol{\Lambda}_K^{-1} \boldsymbol{\phi}(s, a) \le \boldsymbol{\gamma}_i^2$ is equivalent to setting $N = 1/\boldsymbol{\gamma}_i^2$. Since each $\mathcal{S}_i$ can only be reached independently of the other $\mathcal{S}_j$, $j \ne i$, this implies that to meet our objective we must run for at least

$$\sum_{i=1}^{A} 2^i \cdot \frac{|\mathcal{S}_i| A}{\boldsymbol{\gamma}_i^2} = \mathcal{O}\Big(\sum_{i=1}^{A} 2^i \cdot \frac{d}{\boldsymbol{\gamma}_i^2} \Big)$$

episodes. This recovers the complexity CoverTraj gets as given in Theorem 9.1 (for $\boldsymbol{\gamma}_i^2$ sufficiently small).

9.1.1 Instantiating CoverTraj with Force

A recent work, Wagenmaker et al. [2022a], provides a first-order regret minimization algorithm for linear MDPs, Force. The following result shows the complexity of CoverTraj when instantiated with Force.

Corollary 9.2. *When running CoverTraj with RegMin set to the computationally efficient version of Force [Wagenmaker et al., 2022a], all the guarantees of Theorem 9.1 hold, and*

the sample complexity can be bounded as

$$\widetilde{\mathcal{O}}\left(\sum_{i=1}^{m} 2^i \cdot \max\left\{\frac{d}{\boldsymbol{\gamma}_i^2}\log\frac{2^i}{\boldsymbol{\gamma}_i^2}, d^4H^3m^3\log^{7/2}\frac{1}{\delta}\right\}\right)$$

where the $\widetilde{\mathcal{O}}(\cdot)$ hides absolute constants and logarithmic terms. Furthermore, when run with Force, CoverTraj *is computationally efficient with computational cost scaling polynomially in problem parameters.*

Thus, CoverTraj may be instantiated in a computationally efficient way. We make several additional comments on the computational complexity of this approach.

Computational Complexity. Note that the primary computational cost of CoverTraj is incurred in running RegMin, and in evaluating the reward function. As the sets $\mathcal{X}_i$ can be parameterized as ellipsoids, it can be efficiently checked if a given $\boldsymbol{\phi} \in \mathbb{R}^d$ is in $\mathcal{X}_i$, which makes evaluating the reward function efficient. Furthermore, Force only needs access to the reward function at points encountered on each trajectory, so it only need evaluate the reward at polynomially many points. As noted, a computationally efficient version of Force exists (assuming $|\mathcal{A}|$ is not too large), which is what we employ here, making the entire procedure computationally efficient.

9.2 Worst-Case Optimal Reward-Free RL

We next present our main upper bound on reward-free RL, and then provide a lower bound on reward-aware RL—giving a lower bound on reward-free RL as an immediate corollary.

9.2.1 Upper Bounds on Reward Free RL in Linear MDPs

We present our main algorithm, RFLin-Explore, as Algorithm 9.3.

Algorithm 9.3 Reward Free Linear RL: Exploration (RFLin-Explore)

input: tolerance ϵ, confidence δ
$K_{\max} \leftarrow \mathrm{poly}(1/\epsilon, d, H, \log 1/\delta)$
$\beta \leftarrow cH\sqrt{d\log(1+dHK_{\max}) + \log H/\delta}$
$\varsigma_\epsilon \leftarrow \lceil \log_2(4\beta H/\epsilon)\rceil$
$\boldsymbol{\gamma}^2 \in \mathbb{R}^{\varsigma_\epsilon}$ with $\boldsymbol{\gamma}_i^2 \leftarrow 2^{2i}\cdot\frac{\epsilon^2}{64H^2\varsigma_\epsilon^2\beta^2}$ for $i \in [\varsigma_\epsilon]$.
for $h = 1, \ldots, H$ **do**
    $\{(\mathcal{X}_{hi}, \mathcal{D}_{hi}, \boldsymbol{\Lambda}_{hi})\}_{i=1}^{\varsigma_\epsilon} \leftarrow$ CoverTraj$(h, \frac{\delta}{H}, \boldsymbol{\gamma}^2, \varsigma_\epsilon)$
$\mathfrak{D} \leftarrow \{\{(\mathcal{X}_{hi}, \mathcal{D}_{hi}, \boldsymbol{\Lambda}_{hi})\}_{i=1}^{\varsigma_\epsilon}\}_{h=1}^{H}$
return RFLin-Plan$(\cdot\,; \epsilon, \delta, \mathfrak{D})$

RFLin-Explore explores by calling CoverTraj for each h, with tolerance set to $\boldsymbol{\gamma}_i = \mathcal{O}(2^i\epsilon)$. RFLin-Explore returns RFLin-Plan$(\cdot\,; \epsilon, \delta, \mathfrak{D})$ which defines a mapping from reward functions to policies. This mapping is parameterized by the data returned by CoverTraj and, given a reward function as input, outputs a policy it believes is near-optimal for this reward function. RFLin-Plan itself runs a simple least squares value iteration procedure to compute the policy. We define RFLin-Plan in Algorithm 9.4.

Algorithm 9.4 Reward Free Linear RL: Planning (RFLin-Plan)

input: reward functions r
parameters: tolerance ϵ, confidence δ, data $\{\{(\mathcal{X}_{hi}, \mathcal{D}_{hi}, \boldsymbol{\Lambda}_{hi})\}_{i=1}^{\varsigma_\epsilon}\}_{h=1}^{H}$
$K_{\max} \leftarrow \text{poly}(1/\epsilon, d, H, \log 1/\delta)$
$\beta \leftarrow cH\sqrt{d\log(1 + dHK_{\max}) + \log H/\delta}$
for $h = H, H-1, \ldots, 1$ **do**
  $\widehat{\boldsymbol{w}}_h \leftarrow \arg\min_{\boldsymbol{w}} \sum_{i=1}^{\varsigma_\epsilon}\sum_{\tau=1}^{K_i}\left(\boldsymbol{w}^\top \boldsymbol{\phi}_{h,\tau}^i - r_h(s_{h,\tau}^i, a_{h,\tau}^i) - V_{h+1}(s_{h+1,\tau}^i)\right)^2 + \|\boldsymbol{w}\|_2^2$
  $\boldsymbol{\Lambda}_h = I + \sum_{i=1}^{\varsigma_\epsilon}\sum_{\tau=1}^{K_i}\boldsymbol{\phi}_{h,\tau}^i(\boldsymbol{\phi}_{h,\tau}^i)^\top$
  $Q_h(\cdot,\cdot) \leftarrow \min\{\langle\boldsymbol{\phi}(\cdot,\cdot), \widehat{\boldsymbol{w}}_h\rangle + \beta\|\boldsymbol{\phi}(\cdot,\cdot)\|_{\boldsymbol{\Lambda}_h^{-1}}, H\}$
  $V_h(\cdot) \leftarrow \max_a Q_h(\cdot, a)$
  $\widehat{\pi}_h(\cdot) \leftarrow \arg\max_a Q_h(\cdot, a)$
return $\{\widehat{\pi}_h\}_{h=1}^{H}$

We then have the following result.

Theorem 9.3. *Consider running RFLin-Explore with tolerance $\epsilon > 0$ and confidence $\delta > 0$, and let RFLin-Plan$(\cdot)$ denote its output. Then with probability at least $1-\delta$, for an arbitrary number of deterministic reward functions r satisfying Definition 8.1.1, RFLin-Plan(r) returns a policy $\widehat{\pi}$ which is ϵ-optimal with respect to r. Furthermore, this procedure collects at most*

$$\widetilde{\mathcal{O}}\left(\frac{dH^5(d+\log 1/\delta)}{\epsilon^2} + \frac{d^{9/2}H^6\log^4(1/\delta)}{\epsilon}\right)$$

episodes, where $\widetilde{\mathcal{O}}(\cdot)$ hides absolute constants and terms $\text{poly}\log(d, H, 1/\epsilon)$ *and* $\text{poly}\log\log(1/\delta)$.

For ϵ sufficiently small, exploring via RFLin-Explore and planning via RFLin-Plan learns an ϵ-optimal policy for an arbitrarily large number of reward functions using only $\widetilde{\mathcal{O}}(\frac{d^2H^5}{\epsilon^2})$ episodes. This improves on both existing reward-free algorithms for linear MDPs by a factor of d [Zanette et al., 2020c, Wang et al., 2020].

Remark 9.2.1 (Computational Efficiency)**.** In MDPs with a finite number of actions, both RFLin-Explore and RFLin-Plan are computationally efficient, with computational cost scaling polynomially in $d, H, 1/\epsilon, \log 1/\delta$, and $|\mathcal{A}|$. The primary computational cost of RFLin-Explore is due to CoverTraj, which is computationally efficient, while the computational cost of RFLin-Plan is due primarily to solving a least-squares problem.

9.2.1.1 *Deriving Theorem 9.3 from* CoverTraj

We briefly sketch out how one may apply CoverTraj to obtain Theorem 9.3 and defer the full proof to Section 9.4. Consider running RFLin-Explore and then calling RFLin-Plan. Let V_h denote the estimates maintained by RFLin-Plan. We first apply the following self-normalized bound.

Lemma 9.2.1. *With high probability,*

$$\left\| \sum_{i=1}^{\varsigma_\epsilon} \sum_{\tau=1}^{K_i} \boldsymbol{\phi}^i_{h,\tau} [V_{h+1}(s^i_{h+1,\tau}) - \mathbb{E}_h[V_{h+1}](s^i_{h,\tau}, a^i_{h,\tau})] \right\|_{\boldsymbol{\Lambda}_h^{-1}} \leq cH\sqrt{d\log(1+dHK_{\max}) + \log H/\delta} = \beta.$$

Critically, in contrast to the self-normalized bound used in Jin et al. [2020b], ours scales as $\widetilde{\mathcal{O}}(H\sqrt{d})$ instead of $\widetilde{\mathcal{O}}(Hd)$. As our exploration procedure explores each $h \in [H]$ separately, V_{h+1} is *uncorrelated* with $\{\{(s^i_{h,\tau}, a^i_{h,\tau}, s^i_{h+1,\tau})\}_{\tau=1}^{K_i}\}_{i=1}^{\varsigma_\epsilon}$, and we can avoid the union bound over $\boldsymbol{\Lambda}_{h+1}$ that is necessary in Jin et al. [2020b], saving us a factor of $\sqrt{d}$.

Given Lemma 9.2.1, using an argument similar to Jin et al. [2020b], one can show that

$$V_h(s_h) \leq r_h(s_h, \widehat{\pi}_h(s_h)) + \mathbb{E}_h[V_{h+1}](s_h, \widehat{\pi}_h(s_h)) + 2\beta\|\boldsymbol{\phi}(s_h, \widehat{\pi}_h(s_h))\|_{\boldsymbol{\Lambda}_h^{-1}}$$

and, furthermore, that $V_0 \geq V_0^\star$. Some algebra shows that we can then bound the suboptimality of $\widehat{\pi}$, the policy returned by RFLin-Plan, as

$$V_0^\star - V_0^{\widehat{\pi}} \leq V_0 - V_0^{\widehat{\pi}} \leq 2\beta \sum_{h=1}^{H} \mathbb{E}_{\widehat{\pi}}[\|\boldsymbol{\phi}(s_h, \widehat{\pi}_h(s_h))\|_{\boldsymbol{\Lambda}_h^{-1}}].$$

However, it is easy to see that

$$\begin{aligned}\sum_{h=1}^{H} \mathbb{E}_{\widehat{\pi}}[\|\boldsymbol{\phi}(s_h, \widehat{\pi}_h(s_h))\|_{\boldsymbol{\Lambda}_h^{-1}}] &\leq \sum_{h=1}^{H} \sup_{\pi} \mathbb{E}_\pi[\|\boldsymbol{\phi}(s_h, a_h)\|_{\boldsymbol{\Lambda}_h^{-1}}] \\ &\leq \sum_{h=1}^{H} \sum_{i=1}^{\varsigma_\epsilon+1} \sup_{\boldsymbol{\phi} \in \mathcal{X}_{hi}} \|\boldsymbol{\phi}\|_{\boldsymbol{\Lambda}_h^{-1}} \cdot \sup_{\pi} \mathbb{E}_\pi[\mathbb{I}\{\boldsymbol{\phi}(s_h, a_h) \in \mathcal{X}_{hi}\}].\end{aligned}$$

Recall that RFLin-Explore calls CoverTraj for each $h \in [H]$ with tolerances $\boldsymbol{\gamma}_i^2 = \mathcal{O}(\frac{2^{2i}\epsilon^2}{dH^4})$. Applying Theorem 9.1, we then have that $\sup_\pi \mathbb{E}_\pi[\mathbb{I}\{\boldsymbol{\phi}(s_h, a_h) \in \mathcal{X}_{hi}\}] = \sup_\pi w_h^\pi(\mathcal{X}_{hi}) \leq 2^{-i+1}$ and $\sup_{\boldsymbol{\phi} \in \mathcal{X}_{hi}} \|\boldsymbol{\phi}\|_{\boldsymbol{\Lambda}_h^{-1}} \leq \sqrt{\boldsymbol{\gamma}_i^2} = \mathcal{O}(2^i\epsilon/\sqrt{d}H^2)$. Thus, since $\beta = \widetilde{\mathcal{O}}(\sqrt{d}H)$, we can bound the total suboptimality as

$$V_0^\star - V_0^{\widehat{\pi}} \leq 2\beta \sum_{h=1}^{H} \mathbb{E}_{\widehat{\pi}}[\|\boldsymbol{\phi}(s_h, \widehat{\pi}_h(s_h))\|_{\boldsymbol{\Lambda}_h^{-1}}]$$

$$\leq \widetilde{\mathcal{O}}(\sqrt{d}H) \cdot \sum_{h=1}^{H} \sum_{i=1}^{\varsigma_\epsilon+1} \mathcal{O}(2^i \epsilon/\sqrt{d}H^2) \cdot 2^{-i+1}$$
$$= \widetilde{\mathcal{O}}(\epsilon)$$

so it follows that $\widehat{\pi}$ is ϵ-optimal.

We remark briefly on our choice of $\boldsymbol{\gamma}_i^2$. Note that as i increases, the probability of reaching $\mathcal{X}_{hi}$ decreases exponentially in i, $\sup_\pi \mathbb{E}_\pi[\mathbb{I}\{\boldsymbol{\phi}(s_h, a_h) \in \mathcal{X}_{hi}\}] = \sup_\pi w_h^\pi(\mathcal{X}_{hi}) \leq 2^{-i+1}$. Thus, if our "uncertainty" over $\mathcal{X}_{hi}$ scales as $\mathcal{O}(2^i)$, the net contribution of $\mathcal{X}_{hi}$ to the suboptimality scales as $\mathcal{O}(1)$. As we can bound our uncertainty over $\mathcal{X}_{hi}$ as $\sup_{\boldsymbol{\phi} \in \mathcal{X}_{hi}} \|\boldsymbol{\phi}\|_{\boldsymbol{\Lambda}_h^{-1}} \leq \sqrt{\boldsymbol{\gamma}_i^2}$, our choice of $\boldsymbol{\gamma}_i^2 = \mathcal{O}(\frac{2^{2i}\epsilon^2}{dH^4})$ results in a contribution to the suboptimality of $\mathcal{O}(\epsilon/\sqrt{d}H^2)$ for $\mathcal{X}_{hi}$—our choice of $\boldsymbol{\gamma}_i^2$ cancels the contribution of how easily $\mathcal{X}_{hi}$ can be reached. In other words, we only need to decrease uncertainty for a given $\mathcal{X}_{hi}$ in proportion to how easily that $\mathcal{X}_{hi}$ can be reached.

Furthermore, our choice of $\boldsymbol{\gamma}_i^2$ allows us to efficiently collect the samples necessary to reduce uncertainty. The leading-order term in the complexity given in Corollary 9.2 scales as, given our choice of $\boldsymbol{\gamma}_i^2$:

$$\widetilde{\mathcal{O}}\left(\sum_{i=1}^{m} \frac{2^i d}{\boldsymbol{\gamma}_i^2}\right) = \widetilde{\mathcal{O}}\left(\sum_{i=1}^{m} \frac{d^2H^4}{2^i \cdot \epsilon^2}\right) = \widetilde{\mathcal{O}}\left(\frac{d^2H^4}{\epsilon^2}\right),$$

Repeated for each h this yields the complexity given in Theorem 9.3. The key property we exploit here is that CoverTraj collects samples from a given $\mathcal{X}_{hi}$ at a rate inversely proportional to how easily $\mathcal{X}_{hi}$ can be reached—it takes on order $\mathcal{O}(2^i)$ episodes for CoverTraj to collect a sample from $\mathcal{X}_{hi}$, while the probability of reaching $\mathcal{X}_{hi}$ (for any algorithm) is at most on order 2^{-i}. Thus, as our choice of $\boldsymbol{\gamma}_i^2$ only guarantees that we reduce uncertainty for $\mathcal{X}_{hi}$ in proportion with how easily $\mathcal{X}_{hi}$ can be reached, we have that the complexity of collecting the necessary samples is not prohibitively large, and in particular does not scale with the difficulty of reaching $\mathcal{X}_{hi}$.

Remark 9.2.2 (Necessity of First-Order Regret)**.** Suppose that we instantiate CoverTraj with a regret minimization algorithm which only achieves minimax regret, $\mathcal{R}_K \leq \widetilde{\mathcal{O}}(\sqrt{\mathcal{C}_1 K})$, rather than first-order regret. In order to guarantee $\sup_\pi w_h^\pi(\mathcal{X}_{i+1}) \leq 2^{-i+2}$, we must show that $\mathcal{O}(2^{-i}K_i) \geq \mathcal{R}_{K_i}$, which ensures we have reached the "difficult to reach" states. Using a minimax regret algorithm we therefore need $K_i \geq \widetilde{\mathcal{O}}(2^{2i}\mathcal{C}_1)$, while for a first-order algorithm, assuming $\sup_\pi w_h^\pi(\mathcal{X}_i) \leq 2^{-i+1}$, our choice of reward function gives $V_0^\star(r^1) \leq \mathcal{O}(2^{-i})$, so we need $\mathcal{O}(2^{-i}K_i) \geq \widetilde{\mathcal{O}}(\sqrt{\mathcal{C}_1 2^{-i} K_i}) \iff K_i \geq \widetilde{\mathcal{O}}(2^i \mathcal{C}_1)$. This makes a critical difference when applying CoverTraj to reward-free RL in RFLin-Explore, since in RFLin-Explore we set $m = \varsigma_\epsilon = \mathcal{O}(\log(\sqrt{d}H^2/\epsilon))$. With this choice of m, a non-first-order algorithm would have complexity $\widetilde{\mathcal{O}}(\sum_{i=1}^m 2^{2i} \cdot \mathcal{C}_1) = \widetilde{\mathcal{O}}(2^{2m} \cdot \mathcal{C}_1) = \widetilde{\mathcal{O}}(\mathcal{C}_1 \frac{d^2H^4}{\epsilon^2})$. As the best known minimax regret algorithm has $\mathcal{C}_1 = d^2H^4$, we obtain a (very suboptimal) final complexity of $\widetilde{\mathcal{O}}(\frac{d^4H^8}{\epsilon^2})$.

On the other hand, a first-order algorithm attains $\widetilde{\mathcal{O}}(\sum_{i=1}^{m} 2^i \cdot \mathcal{C}_1) = \widetilde{\mathcal{O}}(2^m \cdot \mathcal{C}_1) = \widetilde{\mathcal{O}}(\mathcal{C}_1 \frac{dH^2}{\epsilon})$.

9.2.2 Lower Bounds on Reward-Aware RL in Linear MDPs

To our knowledge, Theorem 9.3 is the first result to show that computationally efficient d^2 complexity is possible in linear MDPs for any of the reward-free, reward-aware, or regret minimization settings. We next show a somewhat surprising result: reward-free RL is no harder than reward-aware RL in linear MDPs, up to horizon factors. To this end, we show a lower bound on reward-aware RL in linear MDPs.

Theorem 9.4. *Fix $\epsilon > 0$, $d > 1$, $H > 1$, and $K \geq d^2$. Consider running a (possibly adaptive) algorithm for K episodes in a $(d+1)$-dimensional linear MDP with horizon H, which stops at (a possibly random stopping time) τ and outputs a guess at an ϵ-optimal policy, $\widehat{\pi}$. Let $\mathcal{E}$ be the event*

$$\mathcal{E} := \{\tau \leq K \text{ and } \widehat{\pi} \text{ is } \epsilon\text{-optimal}\}.$$

Then there is a universal constant $c > 0$ such that unless

$$K \geq c \cdot \frac{d^2 H^2}{\epsilon^2},$$

there exists a linear MDP $\mathcal{M}$ for which $\mathbb{P}_{\mathcal{M}}[\mathcal{E}^c] \geq 1/10$.

As Theorem 9.4 shows, a $\Omega(d^2H^2/\epsilon^2)$ dependence is necessary for learning an ϵ-optimal policy in linear MDPs. Note that this lower bound holds for learning a *single* policy given access to the reward function during exploration (indeed, it holds in the case when the learner has oracle access to the rewards)—the reward-aware RL setting.

In contrast, Theorem 9.3 shows that we can learn ϵ-optimal policies for *all valid reward functions simultaneously*, without having access to the reward functions during exploration, using only $\widetilde{\mathcal{O}}(d^2H^5/\epsilon^2)$ episodes. In other words, up to H factors, in a worst-case sense, *reward-free RL is no harder than reward-aware RL in linear MDPs*. This is in contrast to the tabular setting, where it is well known that the optimal rate for reward-aware RL is $\Theta(SA/\epsilon^2)$ while the optimal rate for reward-free RL is $\Theta(S^2A/\epsilon^2)$.

9.3 Proof of Theorem 9.1

We next provide the proof of Theorem 9.1. Here, instead of setting $\mathbf{\Lambda}_{h,k-1} = I + \sum_{\tau=1}^{k-1} \boldsymbol{\phi}_{h,\tau} \boldsymbol{\phi}_{h,\tau}^\top$ as in EGS, we prove the result for the setting of $\mathbf{\Lambda}_{h,k-1} = \lambda I + \sum_{\tau=1}^{k-1} \boldsymbol{\phi}_{h,\tau} \boldsymbol{\phi}_{h,\tau}^\top$ in EGS, for some $\lambda > 0$ (for convenience we also assume $\lambda \leq \max\{\mathcal{C}_1, \mathcal{C}_2, 1\}$, though this can be relaxed if desired).

In CoverTraj, the precise setting of K_i is:

$$\left\lceil 2^i \cdot \max\left\{ 2^{10+p_1}\mathcal{C}_1 p_1^{p_1}\log^{p_1}(\frac{2^{i+12}p_1 m\mathcal{C}_1 H}{\delta}), 2^{4+p_2}\mathcal{C}_2 p_2^{p_2}\log^{p_2}(\frac{2^{i+6}p_2 m\mathcal{C}_2 H}{\delta}), \frac{24d}{\boldsymbol{\gamma}_i^2}\log\frac{48\cdot 2^i d/\lambda}{\boldsymbol{\gamma}_i^2}\right\}\right\rceil.$$

The key piece in proving Theorem 9.1 is the following result.

Lemma 9.3.1. *Consider running Algorithm 9.1 with $\boldsymbol{\gamma} \in (0,1]$ and some regret-minimization algorithm RegMin satisfying Definition 9.1.1, and with input set $\mathcal{X}$ satisfying*

$$\sup_{\pi} w_h^{\pi}(\mathcal{X}) \le 2^{-i}.$$

Assume also that K is chosen to satisfy

$$\begin{aligned} K \ge \Bigg\lceil 2^i \cdot \max\Big\{ & 1024\mathcal{C}_1\cdot(2p_1)^{p_1}\log^{p_1}\cdot[4096\cdot 2^i p_1\mathcal{C}_1 H/\delta], \\ & 16\mathcal{C}_2\cdot(2p_2)^{p_2}\log^{p_2}\cdot[64\cdot 2^i p_2\mathcal{C}_2 H/\delta], \frac{24d}{\boldsymbol{\gamma}^2}\log\frac{48\cdot 2^i d}{\boldsymbol{\gamma}^2}\Big\}\Bigg\rceil. \end{aligned} \tag{9.3.1}$$

Let $\widetilde{\mathcal{X}} \subseteq \mathbb{R}^d$ denote the set returned by Algorithm 9.1 defined as

$$\widetilde{\mathcal{X}} = \{\boldsymbol{\phi} \in \mathcal{X} \ : \ \boldsymbol{\phi}^\top \boldsymbol{\Lambda}_{h,K}^{-1}\boldsymbol{\phi} \le \boldsymbol{\gamma}^2\}.$$

Then, with probability at least $1-\delta$,

$$\sup_{\pi} w_h^{\pi}(\mathcal{X}\backslash\widetilde{\mathcal{X}}) \le 2^{-i-1}.$$

Proof. First note that the reward sequence used in Algorithm 9.1 satisfies the conditions of Definition 9.1.1. Thus, we have that, with probability at least $1-\delta$,

$$\sum_{k=1}^{K}[V_0^\star(r^k) - V_0^{\pi_k}(r^k)] \le \sqrt{\mathcal{C}_1 V_0^\star(r^1)K\cdot\log^{p_1}(HK/\delta)} + \mathcal{C}_2\log^{p_2}(HK/\delta). \tag{9.3.2}$$

For simplicity we will assume that $\mathcal{C}_1, \mathcal{C}_2, p_1, p_2 \ge 1$ (since if this is not true, for example if $\mathcal{C}_1 < 1$, (9.3.2) still holds with $\mathcal{C}_1$ replaced by $\max\{\mathcal{C}_1, 1\}$).

Relating $\sum_{k=1}^K V_0^{\pi_k}(r^k)$ to Random Reward. Note that $r_h^k(s,a) \le 1$ for all s, a, and since the reward is non-zero only at step h, $\mathbb{E}_{\pi_k}[r_h^k(s_h^k, a_h^k)] = V_0^{\pi_k}(r^k)$ and $V_0^{\pi_k}(r_k) \le 1$. Then, since the reward is non-increasing,

$$\mathbb{E}[(r_h^k(s_h^k,a_h^k) - V_0^{\pi_k}(r^k))^2|\mathcal{F}_{k-1}] \le 2V_0^{\pi_k}(r^k) \le 2V_0^\star(r^1).$$

By Freedman's inequality, Lemma F.1.13, it follows that with probability at least $1-\delta$,

$$\left|\sum_{k=1}^{K}[r_h^k(s_h^k,a_h^k)-V_0^{\pi_k}(r^k)]\right| \le \sqrt{8V_0^\star(r^1)K\log 1/\delta}+\log 1/\delta.$$

Thus, union bounding over this event and the event of (9.3.2), we have with probability $1-\delta$ that

$$\begin{aligned}\sum_{k=1}^{K} r_h^k(s_h^k,a_h^k) &\ge \sum_{k=1}^{K} V_0^{\pi_k}(r^k) - \sqrt{8V_0^\star(r^1)K\log 2/\delta} - \log 2/\delta \\ &\ge \sum_{k=1}^{K} V_0^\star(r^k) - \sqrt{16\mathcal{C}_1 V_0^\star(r^1)K\cdot\log^{p_1}(2HK/\delta)} - 2\mathcal{C}_2\cdot\log^{p_2}(2HK/\delta) \quad (9.3.3)\end{aligned}$$

where we have used that $\mathcal{C}_1,\mathcal{C}_2,p_1,p_2\ge 1$ to group terms.

Proof by Contradiction. Our goal is use (9.3.4) to reach a contradiction and show that, for our choice of K, $V_0^\star(r^K)\le 2^{-i-1}$. To set up the argument, assume for the sake of contradiction that $V_0^\star(r^K)>2^{-i-1}$. Since r^k is non-increasing, this implies that $V_0^\star(r^k)>2^{-i-1}$ for all $k\in[K]$. Then we can lower bound (9.3.3) as

$$(9.3.3) > K2^{-i-1} - \sqrt{16\mathcal{C}_1 V_0^\star(r^1)K\cdot\log^{p_1}(2HK/\delta)} - 2\mathcal{C}_2\cdot\log^{p_2}(2HK/\delta). \quad (9.3.4)$$

Upper Bounding the Reward. We next upper bound the total reward that can be obtained:

$$\begin{aligned}\sum_{k=1}^{K} r_h^k(s_h^k,a_h^k) &= \sum_{k=1}^{K}\min\{1,\boldsymbol{\gamma}^{-2}\|\boldsymbol{\phi}(s_h^k,a_h^k)\|^2_{\boldsymbol{\Lambda}_{h,k-1}^{-1}}\}\cdot\mathbb{I}\{\boldsymbol{\phi}(s_h^k,a_h^k)\in\mathcal{X}\} \\ &\le \frac{1}{\boldsymbol{\gamma}^2}\cdot\sum_{k=1}^{K}\min\{1,\|\boldsymbol{\phi}(s_h^k,a_h^k)\|^2_{\boldsymbol{\Lambda}_{h,k-1}^{-1}}\}\cdot\mathbb{I}\{\boldsymbol{\phi}(s_h^k,a_h^k)\in\mathcal{X}\} \\ &\le \frac{1}{\boldsymbol{\gamma}^2}\cdot\sum_{k=1}^{K}\min\{1,\|\boldsymbol{\phi}(s_h^k,a_h^k)\|^2_{\boldsymbol{\Lambda}_{h,k-1}^{-1}}\} \\ &\le \frac{1}{\boldsymbol{\gamma}^2}\cdot 2d\log(1+K/d\lambda)\end{aligned} \quad (9.3.5)$$

where we have used that $\boldsymbol{\gamma}\le 1$, and where the last inequality follows by the Elliptic Potential Lemma, Lemma G.1.1, since $\|\boldsymbol{\phi}(s_h^k,a_h^k)\|_2\le 1$ by assumption, and we normalize $\boldsymbol{\Lambda}_{h,k}$ by λI.

Lower Bounding the Reward. By assumption, we have that $\sup_\pi w_h^\pi(\mathcal{X}) \le 2^{-i}$. This implies that

$$V_0^\star(r^1) = \sup_\pi \mathbb{E}_\pi[r_h^1(s_h, a_h)] \le \sup_\pi \mathbb{E}_\pi[\mathbb{I}\{\boldsymbol{\phi}(s_h, a_h) \in \mathcal{X}\}] \le 2^{-i}$$

where the last inequality follows since $\sup_\pi \mathbb{E}_\pi[\mathbb{I}\{\boldsymbol{\phi}(s_h, a_h) \in \mathcal{X}\}] = \sup_\pi w_h^\pi(\mathcal{X})$ by definition. Then,

$$(9.3.4) \ge K2^{-i-1} - \sqrt{2^{-i} \cdot 16\mathcal{C}_1 K \cdot \log^{p_1}(2HK/\delta)} - 2\mathcal{C}_2 \cdot \log^{p_2}(2HK/\delta). \tag{9.3.6}$$

Assume that K is chosen such that

$$K \ge 1024 \cdot 2^i \mathcal{C}_1 (2p_1)^{p_1} \log^{p_1}[4096 \cdot 2^i p_1 \mathcal{C}_1 H/\delta] \tag{9.3.7}$$

then by Lemma G.1.2 we will have

$$\frac{1}{4} K 2^{-i-1} - \sqrt{2^{-i} \cdot 16\mathcal{C}_1 K \cdot \log^{p_1}(2HK/\delta)} \ge 0.$$

Similarly, if

$$K \ge 16 \cdot 2^i \mathcal{C}_2 (2p_2)^{p_2} \log^{p_2}[64 \cdot 2^i p_2 \mathcal{C}_2 H/\delta] \tag{9.3.8}$$

then

$$\frac{1}{4} K 2^{-i-1} - 2\mathcal{C}_2 \cdot \log^{p_2}(2HK/\delta) \ge 0.$$

As (9.3.1) requires that K is chosen so as to satisfy both (9.3.7) and (9.3.8), it follows that

$$(9.3.6) \ge \frac{1}{2} K 2^{-i-1}.$$

However, (9.3.1) also gives that $K \ge \frac{2^i \cdot 24d}{\gamma^2} \log \frac{2^i \cdot 48d/\lambda}{\gamma^2}$. Lemma G.1.2 then implies that

$$K \ge \frac{2^i \cdot 12d}{\gamma^2} \log(2K/\lambda)$$

so that, stringing together the above inequalities,

$$\sum_{k=1}^K r_h^k(s_h^k, a_h^k) \ge (9.3.3) > (9.3.4) \ge (9.3.6) \ge \frac{1}{2} K 2^{-i-1} \ge \frac{3d}{\gamma^2} \log(2K/\lambda). \tag{9.3.9}$$

Concluding the Proof. Combining Equations (9.3.5) and (9.3.9), we have shown that

$$\frac{2d}{\gamma^2}\log(1+K/d\lambda) \geq \sum_{k=1}^{K} r_h^k(s_h^k, a_h^k) > (9.3.6) \geq \frac{3d}{\gamma^2}\log(2K/\lambda).$$

This is a contradiction, since $\frac{2d}{\gamma^2}\log(1+K/d\lambda) \leq \frac{3d}{\gamma^2}\log(2K/\lambda)$. Thus, with probability $1-\delta$, we must have that $V_1^\star(r^K) \leq 2^{-i-1}$. The conclusion that $\sup_\pi w_h^\pi(\mathcal{X}\backslash\widetilde{\mathcal{X}}) \leq 2^{-i-1}$ follows on this event since

$$\begin{aligned}
V_0^\star(r^K) &= \sup_\pi \mathbb{E}_\pi[\mathbb{I}\{\boldsymbol{\phi}(s_h,a_h)^\top \boldsymbol{\Lambda}_{h,K-1}^{-1}\boldsymbol{\phi}(s_h,a_h) > \boldsymbol{\gamma}^2, \boldsymbol{\phi}(s_h,a_h) \in \mathcal{X}\}] \\
&\quad + \mathbb{E}_\pi[\boldsymbol{\gamma}^{-2}\boldsymbol{\phi}(s_h,a_h)^\top \boldsymbol{\Lambda}_{h,K-1}^{-1}\boldsymbol{\phi}(s_h,a_h) \cdot \mathbb{I}\{\boldsymbol{\phi}(s_h,a_h)^\top \boldsymbol{\Lambda}_{h,K-1}^{-1}\boldsymbol{\phi}(s_h,a_h) \leq \boldsymbol{\gamma}^2, \boldsymbol{\phi}(s_h,a_h) \in \mathcal{X}\}] \\
&\geq \sup_\pi \mathbb{E}_\pi[\mathbb{I}\{\boldsymbol{\phi}(s_h,a_h)^\top \boldsymbol{\Lambda}_{h,K}^{-1}\boldsymbol{\phi}(s_h,a_h) > \boldsymbol{\gamma}^2, \boldsymbol{\phi}(s_h,a_h) \in \mathcal{X}\}] \\
&= \sup_\pi \mathbb{E}_\pi[\mathbb{I}\{\boldsymbol{\phi}(s_h,a_h) \in \mathcal{X}\backslash\widetilde{\mathcal{X}}\}]
\end{aligned}$$

where the final equality follows by definition of $\widetilde{\mathcal{X}}$.

□

Proof of Theorem 9.1. Theorem 9.1 is a direct consequence of Lemma 9.3.1. For $i = 1$, it is clearly the case that

$$\sup_\pi w_h^\pi(\mathcal{X}) \leq 2^{-i+1} = 1,$$

so Lemma 9.3.1 and our choice of K_1 gives that with probability at least $1-\delta/m$,

$$\sup_\pi w_h^\pi(\mathcal{X}\backslash\mathcal{X}_1) \leq 2^{-i} \quad \text{and} \quad \boldsymbol{\phi}^\top \boldsymbol{\Lambda}_1^{-1}\boldsymbol{\phi} \leq \boldsymbol{\gamma}_1^2, \forall \boldsymbol{\phi} \in \mathcal{X}_1.$$

Now assume that for some i,

$$\sup_\pi w_h^\pi(\mathcal{X}) \leq 2^{-i+1},$$

then again Lemma 9.3.1 and our choice of K_i gives that with probability at least $1-\delta/m$,

$$\sup_\pi w_h^\pi(\mathcal{X}\backslash\mathcal{X}_i) \leq 2^{-i} \quad \text{and} \quad \boldsymbol{\phi}^\top \boldsymbol{\Lambda}_i^{-1}\boldsymbol{\phi} \leq \boldsymbol{\gamma}_i^2, \forall \boldsymbol{\phi} \in \mathcal{X}_i.$$

The result follows by union bounding over the success event of Lemma 9.3.1 holding for each $i \in [m]$. The final conclusion holds since after epoch m, $\mathcal{X} = \mathcal{B}^d \backslash \cup_{i=1}^m \mathcal{X}_i$. □

9.4 Proof of Theorem 9.3

Let us first establish notation. Recall the episode magnitudes K_i from Algorithm 9.2. We define

$$K_{\mathrm{tot}} := \sum_{i=1}^{\varsigma_\epsilon} K_i, \qquad \text{where } \varsigma_\epsilon := \lceil \log_2(\frac{\epsilon}{4\beta H}) \rceil$$

We assume that our parameters $K_{\max}$ and β are chosen such that

$$\beta \geq \widetilde{\beta} := cH\sqrt{d\log(1+dHK_{\mathrm{tot}}) + \log H/\delta}, \quad K_{\max} \geq K_{\mathrm{tot}}.$$

Throughout this section we will consider an arbitrary reward function satisfying Definition 8.1.1, and will let $V^\star, V^\pi, Q^\star, Q^\pi$ denote the value functions for $\pi^\star$ and π, respectively, with respect to r. Similarly, we will let V and Q refer to the value function estimates maintained by RFLin-Plan when run with r as an input.

Proof of Theorem 9.3. First we establish ϵ-suboptimality, then we address sample complexity.

ϵ-suboptimality. Fix some reward function r satisfying Definition 8.1.1. Let $\widehat{\pi}$ denote the policy returned by RFLin-Plan(r). As noted above, let $V^\star, V^\pi, V, Q^\star, Q^\pi, Q$ refer to the value functions with respect to r.

Let $\mathcal{E}_Q$ denote the high-probability event from Lemma G.2.4 (the full version of Lemma 9.2.1), which holds with probability at least $1-\delta$. The following two results establish that on $\mathcal{E}_Q$ V_0 is an optimistic estimate of $V_0^\star$, and shows a bounds on the estimation error of $\widehat{\boldsymbol{w}}_h$.

Lemma 9.4.1. *On the event $\mathcal{E}_Q$ and assuming that $\beta \geq \widetilde{\beta}$, it holds that $Q_h(s,a) \geq Q_h^\star(s,a)$ for all s, a, h.*

Lemma 9.4.2. *On the event $\mathcal{E}_Q$, for all s, a, h and all r satisfying Definition 8.1.1,*

$$|\langle \boldsymbol{\phi}(s,a), \widehat{\boldsymbol{w}}_h\rangle - r_h(s,a) - \mathbb{E}_h[V_{h+1}](s,a)| \leq \widetilde{\beta}\|\boldsymbol{\phi}(s,a)\|_{\boldsymbol{\Lambda}_h^{-1}}$$

where $\widetilde{\beta} = cH\sqrt{d\log(1+dHK_{\mathrm{tot}}) + \log H/\delta}$.

By Lemma 9.4.1 and the choice of parameter $\beta \geq \widetilde{\beta}$, the following holds on $\mathcal{E}_Q$:

$$V_0^\star - V_0^{\widehat{\pi}} \leq V_0 - V_0^{\widehat{\pi}}.$$

Moreover, by Lemma 9.4.2 and, again using $\beta \geq \widetilde{\beta}$, we have

$$|\langle \boldsymbol{\phi}(s,a), \widehat{\boldsymbol{w}}_h\rangle - r_h(s,a) - \mathbb{E}_h[V_{h+1}](s,a)| \leq \widetilde{\beta}\|\boldsymbol{\phi}(s,a)\|_{\boldsymbol{\Lambda}_h^{-1}} \leq \beta\|\boldsymbol{\phi}(s,a)\|_{\boldsymbol{\Lambda}_h^{-1}}.$$

Thus, by definition of Q, we have

$$\begin{aligned} V_h(s_h) = Q_h(s_h, \widehat{\pi}_h(s_h)) &\leq \langle \boldsymbol{\phi}(s_h, \widehat{\pi}_h(s_h)), \widehat{\boldsymbol{w}}_h \rangle + \beta \|\boldsymbol{\phi}(s_h, \widehat{\pi}_h(s_h))\|_{\boldsymbol{\Lambda}_h^{-1}} \\ &\leq r_h(s_h, \widehat{\pi}_h(s_h)) + \mathbb{E}_h[V_{h+1}](s_h, \widehat{\pi}_h(s_h)) + 2\beta \|\boldsymbol{\phi}(s_h, \widehat{\pi}_h(s_h))\|_{\boldsymbol{\Lambda}_h^{-1}}. \end{aligned}$$

Similarly, by the Bellman Equation, we have

$$V_h^{\widehat{\pi}}(s_h) = Q_h^{\widehat{\pi}}(s_h, \widehat{\pi}_h(s_h)) = r_h(s_h, \widehat{\pi}_h(s_h)) + \mathbb{E}_h[V_{h+1}^{\widehat{\pi}}](s_h, \widehat{\pi}_h(s_h)).$$

It follows that

$$V_h(s_h) - V_h^{\widehat{\pi}}(s_h) \leq \mathbb{E}_h[V_{h+1} - V_{h+1}^{\widehat{\pi}}](s_h, \widehat{\pi}_h(s_h)) + 2\beta \|\boldsymbol{\phi}(s_h, \widehat{\pi}_h(s_h))\|_{\boldsymbol{\Lambda}_h^{-1}}.$$

Thus, unrolling this backwards gives (since $\widehat{\pi}$ is deterministic):

$$\begin{aligned} V_0 - V_0^{\widehat{\pi}} &\leq \mathbb{E}_1[V_2 - V_2^{\widehat{\pi}}](s_1, \widehat{\pi}_1(s_1)) + 2\beta \|\boldsymbol{\phi}(s_1, \widehat{\pi}_1(s_1))\|_{\boldsymbol{\Lambda}_1^{-1}} \\ &\leq \mathbb{E}_1[\mathbb{E}_2[V_3 - V_3^{\widehat{\pi}}](s_2, \widehat{\pi}_2(s_2))](s_1, \widehat{\pi}_1(s_1)) + 2\beta \mathbb{E}_1[\|\boldsymbol{\phi}(s_2, \widehat{\pi}_2(s_2))\|_{\boldsymbol{\Lambda}_2^{-1}}](s_1, \widehat{\pi}_1(s_1)) \\ &\qquad + 2\beta \|\boldsymbol{\phi}(s_1, \widehat{\pi}_1(s_1))\|_{\boldsymbol{\Lambda}_1^{-1}} \\ &= \mathbb{E}_{\widehat{\pi}}[V_3(s_3) - V_3^{\widehat{\pi}}(s_3)] + 2\beta \sum_{h=1}^{2} \mathbb{E}_{\widehat{\pi}}[\|\boldsymbol{\phi}(s_h, \widehat{\pi}_h(s_h))\|_{\boldsymbol{\Lambda}_h^{-1}}] \\ &\vdots \\ &\leq 2\beta \sum_{h=1}^{H} \mathbb{E}_{\widehat{\pi}}[\|\boldsymbol{\phi}(s_h, \widehat{\pi}_h(s_h))\|_{\boldsymbol{\Lambda}_h^{-1}}]. \end{aligned}$$

We then upper bound

$$\begin{aligned} 2\beta \sum_{h=1}^{H} \mathbb{E}_{\widehat{\pi}}[\|\boldsymbol{\phi}(s_h, \widehat{\pi}_h(s_h))\|_{\boldsymbol{\Lambda}_h^{-1}}] &\leq 2\beta \sum_{h=1}^{H} \sup_{\pi} \mathbb{E}_{\pi}[\|\boldsymbol{\phi}(s_h, a_h)\|_{\boldsymbol{\Lambda}_h^{-1}}] \\ &\leq 2\beta \sum_{h=1}^{H} \sum_{i=1}^{\varsigma_\epsilon + 1} \sup_{\pi} \mathbb{E}_{\pi}[\|\boldsymbol{\phi}(s_h, a_h)\|_{\boldsymbol{\Lambda}_h^{-1}} \cdot \mathbb{I}\{\boldsymbol{\phi}(s_h, a_h) \in \mathcal{X}_{h,i}\}] \\ &\leq 2\beta \sum_{h=1}^{H} \sum_{i=1}^{\varsigma_\epsilon + 1} \sup_{\boldsymbol{\phi} \in \mathcal{X}_{h,i}} \|\boldsymbol{\phi}\|_{\boldsymbol{\Lambda}_h^{-1}} \cdot \sup_{\pi} \mathbb{E}_{\pi}[\mathbb{I}\{\boldsymbol{\phi}(s_h, a_h) \in \mathcal{X}_{h,i}\}]. \end{aligned}$$

By Theorem 9.1 and a union bound over H, we will have that, with probability at least $1 - \delta$, for all $h \in [H]$ and $i \in [\varsigma_\epsilon]$ simultaneously,

$$\sup_{\pi} \mathbb{E}_{\pi}[\mathbb{I}\{\boldsymbol{\phi}(s_h, a_h) \in \mathcal{X}_{h,i}\}] = \sup_{\pi} w_h^{\pi}(\mathcal{X}_{h,i}) \leq 2^{-i+1}.$$

Furthermore, Theorem 9.1 also gives

$$\sup_{\boldsymbol{\phi}\in\mathcal{X}_{h,i}} \|\boldsymbol{\phi}\|_{\boldsymbol{\Lambda}_h^{-1}} \le \sup_{\boldsymbol{\phi}\in\mathcal{X}_{h,i}} \|\boldsymbol{\phi}\|_{\boldsymbol{\Lambda}_{h,i}^{-1}} \le \sqrt{\boldsymbol{\gamma}_i^2} = \frac{2^i\epsilon}{8H\varsigma_\epsilon\beta}$$

where the final equality follows by our setting of $\boldsymbol{\gamma}_i^2 = \frac{2^{2i}\epsilon^2}{64H^2\varsigma_\epsilon^2\beta^2}$. Finally, one last invocation of Theorem 9.1, followed by the choice of ς_ϵ, gives that

$$\sup_\pi \mathbb{E}_\pi[\mathbb{I}\{\boldsymbol{\phi}(s_h,a_h)\in\mathcal{X}_{h,\varsigma_\epsilon+1}\}] = \sup_\pi w_h^\pi(\mathcal{X}_{h,\varsigma_\epsilon+1}) \le 2^{-\varsigma_\epsilon} \le \frac{\epsilon}{4\beta H}$$

Lastly, observe that $\sup_{\boldsymbol{\phi}\in\mathcal{X}_{h,\varsigma_\epsilon+1}} \|\boldsymbol{\phi}\|_{\boldsymbol{\Lambda}_h^{-1}} \le 1$ always holds, since $\boldsymbol{\Lambda}_h \succeq I$ and $\|\boldsymbol{\phi}\|_2 \le 1$. This gives that

$$\begin{aligned}2\beta\sum_{h=1}^{H}\sum_{i=1}^{\varsigma_\epsilon+1}\sup_{\boldsymbol{\phi}\in\mathcal{X}_{h,i}}\|\boldsymbol{\phi}\|_{\boldsymbol{\Lambda}_h^{-1}}\cdot\sup_\pi\mathbb{E}_\pi[\mathbb{I}\{\boldsymbol{\phi}(s_h,a_h)\in\mathcal{X}_{h,i}\}] &\le 2\beta\sum_{h=1}^{H}\sum_{i=1}^{\varsigma_\epsilon}2^{-i+1}\frac{2^i\epsilon}{8H\varsigma_\epsilon\beta}+\frac{\epsilon}{2}\\ &= \frac{\epsilon}{2}+\frac{\epsilon}{2} = \epsilon\end{aligned}$$

so our policy $\widehat{\pi}$ is ϵ-optimal for the reward function r. However, since r was arbitrary, the above holds for all r satisfying Definition 8.1.1.

Sample Complexity. It remains to bound the sample complexity. We note that we ran for a total of $\sum_{h=1}^{H}\sum_{i=1}^{\varsigma_\epsilon} K_i = HK_{\text{tot}}$ episodes, where again we recall

$$\begin{aligned}K_i = \Bigg\lceil 2^i\cdot\max\Big\{&1024\mathcal{C}_1(2p_1)^{p_1}\log^{p_1}[4096\cdot 2^i p_1\varsigma_\epsilon\mathcal{C}_1H^2/\delta],\\ &16\mathcal{C}_2(2p_2)^{p_2}\log^{p_2}[64\cdot 2^i p_2\varsigma_\epsilon\mathcal{C}_2H^2/\delta],\frac{24d}{\boldsymbol{\gamma}_i^2}\log\frac{48\cdot 2^i d}{\boldsymbol{\gamma}_i^2}\Big\}\Bigg\rceil.\end{aligned}$$

As we use FORCE as our regret minimization algorithm, using the values of $\mathcal{C}_1,\mathcal{C}_2,p_1,p_2$ given in Lemma G.1.4 we can bound (for a universal constant c)

$$\begin{aligned}K_i &\le c2^i d^4H^3\log^{3/2}(\sqrt{d}/\boldsymbol{\gamma}_i^2)i^3\log^{7/2}\left(\varsigma_\epsilon dH\log^{3/2}(\sqrt{d}/\boldsymbol{\gamma}_i^2)/\delta\right)+2^i\frac{24d}{\boldsymbol{\gamma}_i^2}\log\frac{48\cdot 2^i d}{\boldsymbol{\gamma}_i^2}\\ &\le 2^i d^4H^3\log^{7/2}(1/\delta)\cdot\operatorname{poly}\log(d,H,1/\epsilon,\log 1/\delta)+\frac{2^i d}{\boldsymbol{\gamma}_i^2}\cdot\operatorname{poly}\log(d,H,1/\epsilon,\log 1/\delta)\end{aligned}$$

where the last inequality follows by our choice of $\boldsymbol{\gamma}_i^2 = 2^{2i}\cdot\frac{\epsilon^2}{64H^2\varsigma_\epsilon^2\beta^2}$ and $\varsigma_\epsilon = \lceil\log_2(4\beta H/\epsilon)\rceil$, and upper bounding $\operatorname{poly}(i)$ factors by $\operatorname{poly}(\varsigma_\epsilon)$. Let $\varsigma = \operatorname{poly}\log(d,H,1/\epsilon,\log 1/\delta)$ (the precise setting of which may change from line to line). Then we can bound the sample

complexity by:

$$\begin{aligned}\sum_{h=1}^{H}\sum_{i=1}^{\varsigma_\epsilon} K_i &\le \varsigma H \sum_{i=1}^{\varsigma_\epsilon}\left(2^i d^4 H^3 \log^{7/2}(1/\delta) + \frac{2^i d}{\boldsymbol{\gamma}_i^2}\right) \\ &\le \varsigma H \sum_{i=1}^{\varsigma_\epsilon}\left(2^i d^4 H^3 \log^{7/2}(1/\delta) + \frac{2^i d H^2 \beta^2}{2^{2i}\epsilon^2}\right) \\ &\le \frac{d^4 H^5 \beta \log^{7/2}(1/\delta)\varsigma}{\epsilon} + \frac{dH^3\beta^2\varsigma}{\epsilon^2} \\ &\le \frac{d^{9/2} H^6 \log^4(1/\delta)\varsigma}{\epsilon} + \frac{dH^5(d+\log 1/\delta)\varsigma}{\epsilon^2}.\end{aligned}$$

Selecting β. It remains to show that $K_{\max} \ge K_{\mathrm{tot}}$, which will imply that our setting of β in RFLin-Plan satisfies $\beta \ge \widetilde{\beta}$. However, this follows directly by the above complexity bound, which is polynomial in all arguments, justifying our choice of $K_{\max}$.

□

9.5 Proof of Theorem 9.4

We first establish a lower bound on adaptive linear regression, then show that this implies a lower bound on PAC policy identification in linear bandits, and finally argue that linear MDPs can represent linear bandits, giving a lower bound on PAC policy identification in linear MDPs.

Theorem 9.5 (Lower Bound on Adaptive Linear Regression)**.** *Let $\Phi = \mathcal{S}^{d-1}$ and $\Theta = \{-\mu, \mu\}^d$ for some $\mu \in (0, \frac{1}{20\sqrt{d}}]$. Consider the query model where at every step $t = 1, \dots, T$ we choose some $\boldsymbol{\phi}_t \in \Phi$, and observe*

$$y_t \sim \mathrm{Bernoulli}(1/2 + \langle \boldsymbol{\theta}, \boldsymbol{\phi}_t \rangle) \tag{9.5.1}$$

for $\boldsymbol{\theta} \in \Theta$. Then

$$\inf_{\widehat{\boldsymbol{\theta}}, \pi} \max_{\boldsymbol{\theta} \in \Theta} \mathbb{E}_{\boldsymbol{\theta}}[\|\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}\|_2^2] \ge \frac{d\mu^2}{2}\left(1 - \sqrt{\frac{20T\mu^2}{d}}\right).$$

where the infimum is over all measurable estimators $\widehat{\boldsymbol{\theta}}$ and measurable (but possibly adaptive) query rules π, and $\mathbb{E}_{\boldsymbol{\theta}}[\cdot]$ denotes the expectation over the randomness in the observations and decision rules if $\boldsymbol{\theta}$ is the true instance. In particular, if $T \ge d^2$, choosing $\mu = \sqrt{d/700T}$, we get

$$\inf_{\widehat{\boldsymbol{\theta}}, \pi} \max_{\boldsymbol{\theta} \in \Theta} \mathbb{E}_{\boldsymbol{\theta}}[\|\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}\|_2^2] \ge 0.00059 \cdot d^2/T.$$

Proof. The proof of this result follows closely the proof of Theorem 3 of Shamir [2013]. Clearly,

$$\begin{aligned}\max_{\boldsymbol{\theta}\in\Theta}\mathbb{E}_{\boldsymbol{\theta}}[\|\boldsymbol{\theta}-\widehat{\boldsymbol{\theta}}\|_2^2] &\geq \mathbb{E}_{\boldsymbol{\theta}\sim\mathrm{unif}(\Theta)}\mathbb{E}_{\boldsymbol{\theta}}[\|\boldsymbol{\theta}-\widehat{\boldsymbol{\theta}}\|_2^2] \\ &= \mathbb{E}_{\boldsymbol{\theta}\sim\mathrm{unif}(\Theta)}\mathbb{E}_{\boldsymbol{\theta}}\left[\sum_{i=1}^{d}(\boldsymbol{\theta}_i-\widehat{\boldsymbol{\theta}}_i)^2\right] \\ &\geq \mathbb{E}_{\boldsymbol{\theta}\sim\mathrm{unif}(\Theta)}\mathbb{E}_{\boldsymbol{\theta}}\left[\mu^2\sum_{i=1}^{d}\mathbb{I}\{\boldsymbol{\theta}_i\widehat{\boldsymbol{\theta}}_i<0\}\right].\end{aligned}$$

As in Shamir [2013], we assume that the query strategy is deterministic conditioned on the past: $\boldsymbol{\phi}_t$ is a deterministic function of $y_1,\ldots,y_{t-1},\boldsymbol{\phi}_1,\ldots,\boldsymbol{\phi}_{t-1}$, which is without loss of generality. We then need the following result.

Lemma 9.5.1 (Lemma 4 of Shamir [2013])**.** *Let $\boldsymbol{\theta}$ be a random vector, none of whose coordinates is supported on 0, and let $y_1,y_2,\ldots,y_T$ be a sequence of query values obtained by a deterministic strategy returning a point $\widehat{\boldsymbol{\theta}}$ (that is, $\boldsymbol{\phi}_t$ is a deterministic function of $y_1,\ldots,y_{t-1},\boldsymbol{\phi}_1,\ldots,\boldsymbol{\phi}_{t-1}$, and $\widehat{\boldsymbol{\theta}}$ is a deterministic function of $y_1,\ldots,y_T$). Then we have*

$$\mathbb{E}_{\boldsymbol{\theta}\sim\mathrm{unif}(\Theta)}\mathbb{E}_{\boldsymbol{\theta}}\left[\sum_{i=1}^{d}\mathbb{I}\{\boldsymbol{\theta}_i\widehat{\boldsymbol{\theta}}_i<0\}\right]\geq\frac{d}{2}\left(1-\sqrt{\frac{1}{d}\sum_{i=1}^{d}\sum_{t=1}^{T}U_{t,i}}\right)$$

where

$$U_{t,i}=\sup_{\boldsymbol{\theta}_j,j\neq i}\mathrm{KL}\Big(\mathbb{P}(y_t|\boldsymbol{\theta}_i>0,\{\boldsymbol{\theta}_j\}_{j\neq i},\{y_s\}_{s=1}^{t-1})||\mathbb{P}(y_t|\boldsymbol{\theta}_i<0,\{\boldsymbol{\theta}_j\}_{j\neq i},\{y_s\}_{s=1}^{t-1})\Big).$$

In our setting, y_t is distributed as in (9.5.1). Thus, we have that

$$U_{t,i}=\sup_{\boldsymbol{\theta}_j,j\neq i}\mathrm{KL}\Big(\mathrm{Bernoulli}(1/2+\sum_{j\neq i}\boldsymbol{\theta}_j\boldsymbol{\phi}_{t,j}+\mu\boldsymbol{\phi}_{t,i})||\mathrm{Bernoulli}(1/2+\sum_{j\neq i}\boldsymbol{\theta}_j\boldsymbol{\phi}_{t,j}-\mu\boldsymbol{\phi}_{t,i})\Big).$$

Note that:

Lemma 9.5.2 (Lemma 2.7 of Tsybakov [2009])**.**

$$\mathrm{KL}(\mathrm{Bernoulli}(p)||\mathrm{Bernoulli}(q))\leq\frac{(p-q)^2}{q(1-q)}.$$

Thus, by Lemma 9.5.2, we can bound

$$U_{t,i} \le \frac{(2\mu\boldsymbol{\phi}_{t,i})^2}{(1/2 + \sum_{j\neq i}\boldsymbol{\theta}_j\boldsymbol{\phi}_{t,j} - \mu\boldsymbol{\phi}_{t,i})(1 - 1/2 - \sum_{j\neq i}\boldsymbol{\theta}_j\boldsymbol{\phi}_{t,j} + \mu\boldsymbol{\phi}_{t,i})} \le 20\mu^2\boldsymbol{\phi}_{t,i}^2$$

where the last inequality follows by our assumption that $\mu \le \frac{1}{20\sqrt{d}}$ and since $\boldsymbol{\phi}_t \in \mathcal{S}^{d-1}$, which implies that $1/2 + \sum_{j\neq i}\boldsymbol{\theta}_j\boldsymbol{\phi}_{t,j} - \mu\boldsymbol{\phi}_{t,i} \ge 9/20$ and $1 - 1/2 - \sum_{j\neq i}\boldsymbol{\theta}_j\boldsymbol{\phi}_{t,j} + \mu\boldsymbol{\phi}_{t,i} \ge 9/20$.

By Lemma 9.5.1 and this calculation, we can lower bound

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{\theta}\sim\mathrm{unif}(\Theta)}\mathbb{E}_{\boldsymbol{\theta}}\left[\mu^2\sum_{i=1}^{d}\mathbb{I}\{\boldsymbol{\theta}_i\widehat{\boldsymbol{\theta}}_i < 0\}\right] &\ge \frac{d\mu^2}{2}\left(1 - \sqrt{\frac{1}{d}\sum_{i=1}^{d}\sum_{t=1}^{T}U_{t,i}}\right) \\
&\ge \frac{d\mu^2}{2}\left(1 - \sqrt{\frac{1}{d}\sum_{i=1}^{d}\sum_{t=1}^{T}20\mu^2\boldsymbol{\phi}_{t,i}^2}\right) \\
&= \frac{d\mu^2}{2}\left(1 - \sqrt{\frac{20T\mu^2}{d}}\right)
\end{aligned}$$

where the final equality follows since $\boldsymbol{\phi}_t \in \mathcal{S}^{d-1}$. This proves the first conclusion. The second conclusion follows by our choice of μ. □

9.5.1 Lower Bounds on Linear Bandits

Definition 9.5.1 (Linear Bandits and ϵ-Optimal Arms)**.** Consider the linear bandit setting parameterized by some $\boldsymbol{\theta} \in \mathbb{R}^d$ and $\Phi \subseteq \mathbb{R}^d$, where at every step k the learner chooses $\boldsymbol{\phi}_k \in \Phi$ and observes

$$y_k \sim \mathrm{Bernoulli}(\langle\boldsymbol{\theta}, \boldsymbol{\phi}_k\rangle + 1/2).$$

We say an arm $\boldsymbol{\phi} \in \Phi$ is ϵ-optimal if

$$\langle\boldsymbol{\theta}, \boldsymbol{\phi}\rangle + 1/2 \ge \sup_{\boldsymbol{\phi}'\in\Phi}\langle\boldsymbol{\theta}, \boldsymbol{\phi}'\rangle + 1/2 - \epsilon$$

and a policy $\pi \in \triangle_\Phi$ is ϵ-optimal if

$$\mathbb{E}_{\boldsymbol{\phi}\sim\pi}[\langle\boldsymbol{\theta}, \boldsymbol{\phi}\rangle] + 1/2 \ge \sup_{\boldsymbol{\phi}'\in\Phi}\langle\boldsymbol{\theta}, \boldsymbol{\phi}'\rangle + 1/2 - \epsilon.$$

Lemma 9.5.3. *Fix $\epsilon > 0$, $d > 1$, and $K \ge d^2$. Let $\boldsymbol{\theta} \in \Theta = \{-\sqrt{d/700K}, \sqrt{d/700K}\}^d$ and $\Phi = \mathcal{S}^{d-1}$. Consider running a (possibly adaptive) algorithm in the linear bandit setting of Definition 9.5.1 which stops at (a possibly random stopping time) τ and outputs a guess at*

an ϵ-optimal policy, $\widehat{\pi} \in \triangle_\Phi$. Let $\mathcal{E}$ be the event

$$\mathcal{E} := \{\tau \le K \text{ and } \widehat{\pi} \text{ is } \epsilon\text{-optimal}\}.$$

Then unless $K \ge c \cdot \frac{d^2}{\epsilon^2}$ for a universal $c > 0$, there exists $\boldsymbol{\theta} \in \Theta$ for which $\mathbb{P}_{\boldsymbol{\theta}}[\mathcal{E}^c] \ge 1/10$; i.e., with constant probability either $\widehat{\pi}$ is not ϵ-optimal or more than K samples are collected.

Proof. We consider the setting of instances in Theorem 9.5 with $\Phi = \mathcal{S}^{d-1}$ and $\Theta = \{-\mu, \mu\}^d$. We first show how estimation error relates to finding ϵ-good arms. Let $\boldsymbol{\phi}^\star(\boldsymbol{\theta})$ denote the optimal arm for $\boldsymbol{\theta}$. Clearly, $\boldsymbol{\phi}^\star(\boldsymbol{\theta}) = \boldsymbol{\theta}/\|\boldsymbol{\theta}\|_2 = \boldsymbol{\theta}/(\sqrt{d}\mu)$ and $\boldsymbol{\phi}^\star(\boldsymbol{\theta})^\top \boldsymbol{\theta} = \sqrt{d}\mu$. Now consider a distribution $\pi \in \triangle_\Phi$. By definition, we have that π is ϵ-optimal if

$$\mathbb{E}_{\boldsymbol{\phi}\sim\pi}[\boldsymbol{\theta}^\top \boldsymbol{\phi}] = \boldsymbol{\theta}^\top \mathbb{E}_{\boldsymbol{\phi}\sim\pi}[\boldsymbol{\phi}] \ge \sqrt{d}\mu - \epsilon.$$

For a policy π, let $\boldsymbol{\phi}_\pi := \mathbb{E}_{\boldsymbol{\phi}\sim\pi}[\boldsymbol{\phi}]$. Note that by the convexity of norms and Jensen's inequality:

$$\|\boldsymbol{\phi}_\pi\|_2^2 \le \mathbb{E}_{\boldsymbol{\phi}\sim\pi}[\|\boldsymbol{\phi}\|_2^2] = 1.$$

Write $\boldsymbol{\phi}_\pi = \boldsymbol{\phi}^\star(\boldsymbol{\theta}) + \Delta$. Then,

$$1 \ge \|\boldsymbol{\phi}^\star(\boldsymbol{\theta}) + \Delta\|_2^2 = 1 + \|\Delta\|_2^2 + 2\boldsymbol{\phi}^\star(\boldsymbol{\theta})^\top \Delta \implies \boldsymbol{\phi}^\star(\boldsymbol{\theta})^\top \Delta \le -\frac{1}{2}\|\Delta\|_2^2$$
$$\implies \boldsymbol{\theta}^\top \Delta \le -\frac{\sqrt{d}\mu}{2}\|\Delta\|_2^2.$$

However, $\boldsymbol{\phi}_\pi^\top \boldsymbol{\theta} = \boldsymbol{\phi}^\star(\boldsymbol{\theta})^\top \boldsymbol{\theta} + \Delta^\top \boldsymbol{\theta}$, so if π is ϵ-optimal, we have $\Delta^\top \boldsymbol{\theta} \ge -\epsilon$ which implies

$$-\frac{\sqrt{d}\mu}{2}\|\Delta\|_2^2 \ge -\epsilon.$$

In other words, if π is ϵ-optimal for $\boldsymbol{\theta}$, then $\boldsymbol{\phi}_\pi = \boldsymbol{\phi}^\star(\boldsymbol{\theta}) + \Delta$ for some Δ with $\|\Delta\|_2^2 \le \frac{2\epsilon}{\sqrt{d}\mu}$, so

$$\boldsymbol{\phi}_\pi = \boldsymbol{\phi}^\star(\boldsymbol{\theta}) + \Delta = \frac{\boldsymbol{\theta}}{\sqrt{d}\mu} + \Delta \implies \boldsymbol{\theta} = \sqrt{d}\mu(\boldsymbol{\phi}_\pi - \Delta).$$

Now assume that we have a $\widehat{\pi}$ which is ϵ-optimal, and denote $\widehat{\boldsymbol{\phi}} := \boldsymbol{\phi}_{\widehat{\pi}}$. Let $\boldsymbol{\theta} = \sqrt{d}\mu(\widehat{\boldsymbol{\phi}} - \Delta)$ as above. Define the following estimator

$$\widehat{\boldsymbol{\theta}} = \begin{cases} \boldsymbol{\theta}' & \text{if } \exists \boldsymbol{\theta}' \in \Theta \text{ with } \boldsymbol{\theta}' = \sqrt{d}\mu(\widehat{\boldsymbol{\phi}} - \Delta') \text{ for some } \Delta' \in \mathbb{R}^d, \|\Delta'\|_2^2 \le \frac{2\epsilon}{\sqrt{d}\mu} \\ \text{any } \boldsymbol{\theta}' \in \Theta & \text{otherwise} \end{cases}$$

If $\widehat{\boldsymbol{\phi}}$ is actually ϵ-optimal for some $\boldsymbol{\theta} \in \Theta$, then the first condition is met and

$$\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\|_2 = \|\sqrt{d}\mu(\widehat{\boldsymbol{\phi}} - \Delta') - \sqrt{d}\mu(\widehat{\boldsymbol{\phi}} - \Delta)\|_2 \leq 2\sqrt{d}\mu\sqrt{\frac{2\epsilon}{\sqrt{d}\mu}} = \sqrt{8\sqrt{d}\mu\epsilon}.$$

Thus, if we can find an ϵ-good arm for $\boldsymbol{\theta}$, we can estimate $\boldsymbol{\theta}$ up to tolerance $\sqrt{8\sqrt{d}\mu\epsilon}$.

Now set $\mu = \sqrt{d/700K}$. Let $\widehat{\boldsymbol{\theta}}$ denote the estimator constructed from $\widehat{\boldsymbol{\phi}}$ as outlined above. Let $\mathcal{E}$ be the event that $\widehat{\boldsymbol{\phi}}$ is ϵ-good for $\boldsymbol{\theta}$ and note that regardless of whether or not we are on $\mathcal{E}$, $\|\widehat{\boldsymbol{\theta}}\|_2 \leq d/\sqrt{700K}$. Thus,

$$\begin{aligned}\mathbb{E}_{\boldsymbol{\theta}}[\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\|_2^2] &= \mathbb{E}_{\boldsymbol{\theta}}[\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\|_2^2 \cdot \mathbb{I}\{\mathcal{E}\} + \|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\|_2^2 \cdot \mathbb{I}\{\mathcal{E}^c\}] \\ &\leq \frac{8d\epsilon}{\sqrt{700K}} + \frac{4d^2}{700K}\mathbb{P}_{\boldsymbol{\theta}}[\mathcal{E}^c]\end{aligned}$$

where the first inequality follows by our observation above that any ϵ-good $\widehat{\boldsymbol{\phi}}$ yields an estimate of $\boldsymbol{\theta}$ up to tolerance $\sqrt{2d\epsilon/\sqrt{K}}$.

However, by what we have shown above, there exists some $\boldsymbol{\theta}$ such that if we collect (no more than) K samples,

$$\mathbb{E}_{\boldsymbol{\theta}}[\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\|_2^2] \geq 0.00059 \cdot d^2/K.$$

This is a contradiction unless

$$\frac{8d\epsilon}{\sqrt{700K}} + \frac{4d^2}{700K}\mathbb{P}_{\boldsymbol{\theta}}[\mathcal{E}^c] \geq 0.00059\frac{d^2}{K} \iff \mathbb{P}_{\boldsymbol{\theta}}[\mathcal{E}^c] \geq 0.10325 - \frac{2\sqrt{700}\epsilon\sqrt{K}}{d}.$$

It follows that if

$$0.10325 - \frac{2\sqrt{700}\epsilon\sqrt{K}}{d} \geq 0.1 \iff (\frac{0.00325}{2\sqrt{700}})^2 \cdot \frac{d^2}{\epsilon^2} \geq K$$

we have that $\mathbb{P}_{\boldsymbol{\theta}}[\mathcal{E}^c] \geq 0.1$. □

9.5.2 Mapping to Linear MDPs

We next show that the linear bandit instance of Lemma 9.5.3 with parameter $\boldsymbol{\theta}$ (for $\boldsymbol{\theta} \in \Theta$ as in Lemma 9.5.3) can be mapped to a linear MDP with state space $\mathcal{S} = \{s_0, s_1, \bar{s}_2, \ldots, \bar{s}_{d+1}\}$, action space $\mathcal{A} = \mathcal{S}^{d-1} \cup \{\boldsymbol{e}_{d+1}/2\}$, parameters

$$\boldsymbol{\theta}_1 = \mathbf{0}, \quad \boldsymbol{\theta}_h = \boldsymbol{e}_1, h \geq 2$$

$$\boldsymbol{\mu}_1(s_1) = [2\boldsymbol{\theta}, 1], \quad \boldsymbol{\mu}_1(\bar{s}_i) = \frac{1}{d}[-2\boldsymbol{\theta}, 1], \quad \boldsymbol{\mu}_h(s_i) = \boldsymbol{e}_i, h \geq 2$$

and feature vectors

$$\boldsymbol{\phi}(s_0, \boldsymbol{e}_{d+1}/2) = \boldsymbol{e}_{d+1}/2, \quad \boldsymbol{\phi}(s_0, \widetilde{\boldsymbol{\phi}}) = [\widetilde{\boldsymbol{\phi}}/2, 1/2], \quad \forall \widetilde{\boldsymbol{\phi}} \in \mathcal{S}^{d-1}$$
$$\boldsymbol{\phi}(s_1, \widetilde{\boldsymbol{\phi}}) = \boldsymbol{e}_1, \quad \boldsymbol{\phi}(\bar{s}_i, \widetilde{\boldsymbol{\phi}}) = \boldsymbol{e}_i, i \geq 2, \quad \forall \widetilde{\boldsymbol{\phi}} \in \mathcal{A}.$$

Note that, if we take action $\widetilde{\boldsymbol{\phi}}$ in state s_0, our expected episode reward is

$$P_1(s_1|s_0, \widetilde{\boldsymbol{\phi}}) \cdot H + \sum_{i=2}^{d+1} P_1(\bar{s}_i|s_0, \widetilde{\boldsymbol{\phi}}) \cdot 0 = H(\langle \boldsymbol{\theta}, \widetilde{\boldsymbol{\phi}} \rangle + 1/2)$$

since we always acquire a reward of 1 in any state s_1, and a reward of 0 in any state $\bar{s}_i$, and the reward distribution is Bernoulli.

Lemma 9.5.4. *The MDP constructed above is a valid linear MDP as defined in Definition 8.1.1.*

Proof. For $\widetilde{\boldsymbol{\phi}} \in \mathcal{S}^{d-1}$ we have,

$$P_1(s_1|s_0, \widetilde{\boldsymbol{\phi}}) = \langle \boldsymbol{\phi}(s_0, \widetilde{\boldsymbol{\phi}}), \boldsymbol{\mu}_1(s_1) \rangle = \langle \boldsymbol{\theta}, \widetilde{\boldsymbol{\phi}} \rangle + 1/2 \geq 0$$
$$P_1(\bar{s}_i|s_0, \widetilde{\boldsymbol{\phi}}) = \langle \boldsymbol{\phi}(s_0, \widetilde{\boldsymbol{\phi}}), \boldsymbol{\mu}_1(\bar{s}_i) \rangle = \frac{1}{d}(-\langle \boldsymbol{\theta}, \widetilde{\boldsymbol{\phi}} \rangle + 1/2) \geq 0$$

where the inequality follows since $|\langle \boldsymbol{\theta}, \widetilde{\boldsymbol{\phi}} \rangle| \leq \mathcal{O}(1/d)$ for all $\widetilde{\boldsymbol{\phi}} \in \mathcal{S}^{d-1}$, by the choice of $\boldsymbol{\theta}$ in Lemma 9.5.3, and our condition that $K \geq d^2$. In addition,

$$P_1(s_1|s_0, \widetilde{\boldsymbol{\phi}}) + \sum_{i=2}^{d+1} P_1(\bar{s}_i|s_0, \widetilde{\boldsymbol{\phi}}) = \langle \boldsymbol{\theta}_\star, \widetilde{\boldsymbol{\phi}} \rangle + 1/2 + d \cdot \frac{1}{d}(-\langle \boldsymbol{\theta}_\star, \widetilde{\boldsymbol{\phi}} \rangle + 1/2) = 1.$$

Thus, $P_1(\cdot|s_0, \widetilde{\boldsymbol{\phi}})$ is a valid probability distribution for $\widetilde{\boldsymbol{\phi}} \in \mathcal{S}^{d-1}$. A similar calculation shows the same for action $\boldsymbol{e}_{d+1}/2$. It is obvious that $P_h(\cdot|s, \widetilde{\boldsymbol{\phi}})$ is a valid distribution for all s and $\widetilde{\boldsymbol{\phi}} \in \mathcal{A}$.

It remains to check the normalization bounds. Clearly, by our construction of the feature vectors, $\|\boldsymbol{\phi}(s, a)\|_2 \leq 1$ for all s and a. It is also obvious that $\|\boldsymbol{\theta}_1\|_2 \leq \sqrt{d}$ and $\|\boldsymbol{\theta}_h\|_2 \leq \sqrt{d}$. Finally,

$$\||\boldsymbol{\mu}_1(\mathcal{S})|\|_2 = \left\| \sum_{s \in \mathcal{S} \setminus s_0} |\boldsymbol{\mu}_1(s)| \right\|_2 = \|[2\boldsymbol{\theta}, 1] + d \cdot \frac{1}{d}[2\boldsymbol{\theta}, 1]\|_2 \leq \sqrt{d}.$$

Thus, all normalization bounds are met, so this is a valid linear MDP. □

Lower bounding the performance of low-regret algorithms. Assume that we have access to the linear bandit instance constructed in Lemma 9.5.3. That is, at every timestep t we can choose an arm $\widetilde{\boldsymbol{\phi}}_t \in \mathcal{S}^{d-1}$ and obtain and observe reward $y_t \sim \text{Bernoulli}(\langle \boldsymbol{\theta}, \widetilde{\boldsymbol{\phi}}_t \rangle + 1/2)$. Using the mapping above, we can use this bandit to simulate a linear MDP as follows:

1. Start in state s_0 and choose any action $\widetilde{\boldsymbol{\phi}}_t \in \mathcal{A}$

2. Play action $\widetilde{\boldsymbol{\phi}}_t$ in our linear bandit. If reward obtained is $y_t = 1$, then in the MDP transition to any of the states s_1. If the reward obtained is $y_t = 0$ transition to any of the states $\bar{s}_2, \ldots, \bar{s}_{d+1}$, each with probability $1/d$. If the chosen action was $\widetilde{\boldsymbol{\phi}}_t = \boldsymbol{e}_{d+1}/2$, then play any action in the linear bandit and transition to state s_1 with probability $1/2$ and $\bar{s}_2, \ldots, \bar{s}_{d+1}$ with probability $1/2d$, regardless of y_t

3. For the next $H - 1$ steps, take any action in the state in which you end up, transition back to the same state with probability 1, and receive reward of 1 if you are in s_1, and reward of 0 if you are in $\bar{s}_2, \ldots, \bar{s}_{d+1}$.

Note that this MDP has precisely the transition and reward structure as the MDP constructed above.

Lemma 9.5.5. *Assume π is ϵ-optimal in the MDP constructed above. Then $\pi_1(\cdot|s_0)$ is ϵ/H-optimal on the linear bandit instance with parameter $\boldsymbol{\theta}$ and action set $\mathcal{S}^{d-1}$.*

Proof. Note that the value of π in the linear MDP is given by

$$V_0^\pi = H \cdot \left(\sum_{\widetilde{\boldsymbol{\phi}} \in \mathcal{S}^{d-1}} \pi_1(\widetilde{\boldsymbol{\phi}}|s_0)(\langle \widetilde{\boldsymbol{\phi}}, \boldsymbol{\theta} \rangle + \frac{1}{2}) + \frac{1}{2}\pi_1(\boldsymbol{e}_{d+1}/2|s_0) \right) = H \cdot \left(\sum_{\widetilde{\boldsymbol{\phi}} \in \mathcal{S}^{d-1}} \pi_1(\widetilde{\boldsymbol{\phi}}|s_0)\langle \widetilde{\boldsymbol{\phi}}, \boldsymbol{\theta} \rangle + \frac{1}{2} \right)$$

and the optimal policy is $\pi_1(a^\star|s_0) = 1$, for $a^\star = \arg\max_{\widetilde{\boldsymbol{\phi}} \in \mathcal{S}^{d-1}} \langle \widetilde{\boldsymbol{\phi}}, \boldsymbol{\theta} \rangle$, and has value $V_0^\star = \langle a^\star, \boldsymbol{\theta}_\star \rangle + 1/2$. It follows that if π is ϵ-optimal, then

$$\begin{aligned} H \cdot \left(\sum_{\widetilde{\boldsymbol{\phi}} \in \mathcal{S}^{d-1}} \pi_1(\widetilde{\boldsymbol{\phi}}|s_0)\langle \widetilde{\boldsymbol{\phi}}, \boldsymbol{\theta} \rangle + 1/2 \right) &\geq H \cdot (\langle a^\star, \boldsymbol{\theta} \rangle + 1/2) - \epsilon \\ \iff \sum_{\widetilde{\boldsymbol{\phi}} \in \mathcal{S}^{d-1}} \pi_1(\widetilde{\boldsymbol{\phi}}|s_0)\langle \widetilde{\boldsymbol{\phi}}, \boldsymbol{\theta} \rangle + 1/2 &\geq \langle a^\star, \boldsymbol{\theta} \rangle + 1/2 - \epsilon/H \end{aligned}$$

This implies that $\pi_1(\cdot|s_0)$ is ϵ/H-optimal for the linear bandit with parameter $\boldsymbol{\theta}$ and action set $\mathcal{S}^{d-1}$ (note that any mass $\pi(\cdot|s_0)$ places on arm $\boldsymbol{e}_{d+1}/2$ can be instead allocated to any arm in $\mathcal{S}^{d-1}$ without changing this result). □

Proof of Theorem 9.4. Consider running the above procedure for some number of steps. By Lemma 9.5.5, if we can identify an ϵ-optimal policy in this MDP, we can use it to determine an ϵ/H-optimal policy on our bandit instance. As we have used no extra information other than samples from the linear bandit to construct this, it follows that to find an ϵ/H-optimal policy in the MDP, we must take at least the number of samples prescribed by Lemma 9.5.3 for $\epsilon \leftarrow \epsilon/H$. □

Part III

Learning to Explore

Chapter 10

Towards a Non-Asymptotic Theory of Instance-Optimality in Decision Making

In Parts I and II we, broadly, focused on obtaining instance-dependent guarantees on the sample complexity of linear near-optimal policies in various settings. In each example, we saw that the key to obtaining tight sample complexity bounds lay in exploring efficiently: identifying which aspects of the environment were most important to explore, and how we could explore these regions most efficiently.

In each of these examples, while we were able to obtain leading-order instance-dependent complexity terms which were optimal or nearly optimal, our complexity bounds also exhibited potentially large lower-order terms scaling polynomially in problem parameters. While these terms are truly lower-order for T large enough or ϵ small enough, for small T or large ϵ they may dominate the complexity. In every case, these lower-order terms arose because our approaches required that we *learn to explore*. In order to achieve the instance-optimal rate, we must collect the most informative observations as quickly as possible, but we must learn which observations are most informative and how to collect these observations, which requires some amount of interaction with the environment.

While the approaches proposed in Parts I and II were carefully designed to ensure that these lower-order terms were not too large—truly lower-order for reasonable settings of T and ϵ—in the broader literature on instance-optimal decision making, this is not the case. Similar "lower-order" terms, while perhaps lower-order in terms of T, can be exponentially large in problem parameters, leading them to dominate the complexity bounds in all regimes of practical interest.

The focus of this chapter is in understanding this phenomenon. Is the cost to "learning-to-explore" necessary if we wish to obtain instance-optimality? If so, what is the minimum amount of interactions required to learn how to explore and thereby achieve the instance-optimal rate; what is the best-case "lower-order" term? We consider these questions in the regret minimization setting rather than the policy identification setting of Parts I and II, yet the insights developed here extend to the settings of Parts I and II as well. We also move to a general *interactive decision making* setting, first proposed in Foster et al. [2021], which encompasses both bandit and RL settings, including those with function approximation.

This chapter proceeds as follows. In Section 10.1, we introduce our general decision making setting, provide a concrete example illustrating that the cost of learning to explore is necessary, and introduce a novel notion of complexity, the AEC, which quantifies this cost of learning to explore. In Section 10.2 we present our main algorithm and upper bound, and in Section 10.3 instantiate this upper bound in a variety of settings of interest, providing polynomial lower-order terms in all cases (in some cases yielding an exponential improvement over existing results). Finally, in Section 10.4 we provide lower bounds on the cost of learning to explore, and argue that the complexity we achieve is essentially unimprovable in both its leading-order and lower-order terms.

10.1 Formal Setting and Background

We first formally introduce the general decision making setting we consider in Section 10.1.1, then present an overview of the key concepts for defining instance optimal regret in Section 10.1.2. In Section 10.1.3 we provide a motivating example, illustrating that there is indeed a cost for learning-to-explore, and in Section 10.1.4 introduce a novel complexity measure, the AEC, which quantifies the cost of learning to explore.

10.1.1 Interactive Decision Making

We adopt the *Decision Making with Structured Observations* (DMSO) framework of Foster et al. [2021], which is a general setting for interactive decision making that encompasses bandit problems (structured, contextual, and so forth) and reinforcement learning with function approximation.

The DMSO framework is specified by a *decision space* Π, *reward space* $\mathcal{R} \subseteq \mathbb{R}$, and *observation space* $\mathcal{O}$. The learner is given access to a (known) *model class* $\mathcal{M} \subset (\Pi \to \triangle_{\mathcal{R}\times\mathcal{O}})$, and it is assumed there exists some true model $M^\star \in \mathcal{M}$, unknown to the learner, which represents the underlying environment. Formally, we make the following assumption.

Assumption 10.1 (Realizability)**.** *The learner has access to a model class $\mathcal{M}$ containing the true model $M^\star$.*

The learning protocol consists of T rounds. For each round $t = 1, \dots, T$:

1. The learner selects a *decision* $\pi^t \in \Pi$.
2. The learner receives a reward $r^t \in \mathcal{R}$ and observation $o^t \in \mathcal{O}$ sampled via $(r^t, o^t) \sim M^\star(\pi^t)$, and observes (r^t, o^t).

We can think of the model class $\mathcal{M}$ as representing the learner's prior knowledge about the decision making problem, and it allows one to appeal to estimation and function approximation. For structured bandit problems, for example, models correspond to reward

distributions, and $\mathcal{M}$ encodes structure in the reward landscape. For reinforcement learning problems, models correspond to Markov decision processes (MDPs), and $\mathcal{M}$ typically encodes structure in value functions or transition probabilities. We refer the reader to Section 10.3.1 and Section 10.3.2 for concrete examples of how standard decision making settings can be instantiated within the DMSO framework, and refer to Foster et al. [2021] for further background.

For a model $M \in \mathcal{M}$, $\mathbb{E}^{M,\pi}[\cdot]$ denotes the expectation under the process $(r, o) \sim M(\pi)$, $f^M(\pi) := \mathbb{E}^{M,\pi}[r]$ denotes the mean reward function, $\pi_M \in \arg\max_{\pi\in\Pi} f^M(\pi)$ denotes any optimal decision, and $\boldsymbol{\pi}_M = \arg\max_{\pi\in\Pi} f^M(\pi)$ denotes the *set* of all optimal decisions. Similarly, when the algorithm is clear from context, $\mathbb{E}^M[\cdot]$ and $\mathbb{P}^M[\cdot]$ refer to the expectation and probability measure, respectively, induced over histories, $\mathcal{H}^t = (\pi^1, r^1, o^1), \ldots, (\pi^t, r^t, o^t)$, on M. We overload notation somewhat and use $\mathbb{P}^{M,\pi}[\cdot]$ to refer to the conditional density over $\mathcal{R} \times \mathcal{O}$ induced by M. While we do not assume the *random* rewards are strictly bounded (allowing, for example, Gaussian rewards), we make the following assumption on the reward means.

Assumption 10.2 (Bounded Reward Means)**.** *For each $M \in \mathcal{M}$ and $\pi \in \Pi$, we have $f^M(\pi) \in [0, 1]$.*

Throughout this work, we also make the following assumption on the uniqueness of the optimal action of $M^\star$.

Assumption 10.3 (Unique Optimal Action)**.** *For the ground truth model $M^\star \in \mathcal{M}$, the optimal action $\pi_{M^\star}$ is unique.*

Note that the latter assumption is standard in the literature on instance-optimality. We measure performance in terms of regret, which is given by

$$\mathbf{Reg}(T) := \sum_{t=1}^{T} \mathbb{E}_{\pi^t\sim p^t}\left[f^{M^\star}(\pi_{M^\star}) - f^{M^\star}(\pi^t)\right], \tag{10.1.1}$$

where p^t is the learner's randomization distribution for round t. In addition, we define $\Delta^M(\pi) = f^M(\pi_M) - f^M(\pi)$ as the *suboptimality gap* function for model M and decision π, and the *minimum suboptimality gap* as

$$\Delta^M_{\mathrm{min}} := \begin{cases} \inf_{\pi\in\Pi:\Delta^M(\pi)>0} \Delta^M(\pi), & \pi_M \text{ is unique,} \\ 0, & \text{otherwise.} \end{cases} \tag{10.1.2}$$

Since by assumption $\pi_{M^\star}$ is unique, we have $\Delta^{M^\star}_{\mathrm{min}} > 0$. Throughout, we replace dependence on $M^\star$ with "$\star$" when the meaning is clear from context, for example: $\Delta^\star_{\mathrm{min}} := \Delta^{M^\star}_{\mathrm{min}}$, $f^\star(\pi) := f^{M^\star}(\pi)$, or $\pi_\star = \pi_{M^\star}$.

Further Notation. We let $\mathcal{M}^+ = \{M : \Pi \to \triangle_{\mathcal{R}\times\mathcal{O}} \mid f^M(\pi) \in [0,1]\}$ denote the space of all possible models M with rewards in $\mathcal{R}$ and $f^M(\pi) \in [0,1]$. We use $\triangle_{\mathcal{X}}$ to refer to the set of probability distributions over any $\mathcal{X}$. Throughout, we often abbreviate $\mathbb{E}_{\pi\sim p}[\cdot]$ with $\mathbb{E}_p[\cdot]$.

10.1.2 Background: Asymptotic Instance-Optimality

Our aim is to develop algorithms that are *instance-optimal* in a strong sense: for *every model* $M^\star \in \mathcal{M}$, the regret of the algorithm under $M^\star$ is at least as good as that of any *consistent* algorithm; here, an algorithm is said to be "consistent" if it ensures that $\mathbb{E}^M[\mathbf{Reg}(T)] = o(T)$ for all $M \in \mathcal{M}$. Instance-optimality is a powerful notion of performance: no algorithm—even one designed specifically with $M^\star$ in mind—can achieve lower regret on $M^\star$ without giving up consistency. For multi-armed bandits, a long line of work initiated by Lai and Robbins [1985] characterizes the instance-optimal regret as a function of the instance $M^\star$, and provides efficient algorithms that attain instance-optimality in finite time [Garivier et al., 2016, Kaufmann et al., 2016, Lattimore, 2018, Garivier et al., 2019]. For the general decision making setting we consider, the forward-looking work of Graves and Lai [1997] (see Dong and Ma [2022] for a contemporary treatment) introduced a complexity measure we refer to as the *Graves-Lai Coefficient*, which asymptotically characterizes the instance-optimal performance as a function of the instance $M^\star$ and model class $\mathcal{M}$. Define the Kullback-Leibler divergence by

$$D_{\mathsf{KL}}(\mathbb{P} \,\|\, \mathbb{Q}) = \begin{cases} \int \log\left(\frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mathbb{Q}}\right)\mathrm{d}\mathbb{P}, & \mathbb{P} \ll \mathbb{Q}, \\ +\infty, & \text{otherwise.} \end{cases}$$

For any class $\mathcal{M}$ and model $M \in \mathcal{M}$, the Graves-Lai Coefficient is the value of the program

$$\mathsf{glc}(\mathcal{M}, M) := \inf_{\eta\in\mathbb{R}_+^\Pi} \left\{ \sum_{\pi\in\Pi} \eta_\pi \Delta^M(\pi) \mid \forall M' \in \mathcal{M}^{\mathrm{alt}}(M) : \sum_{\pi\in\Pi} \eta_\pi D_{\mathsf{KL}}(M(\pi) \,\|\, M'(\pi)) \geq 1 \right\}, \tag{10.1.3}$$

where, for M with unique optimal decision π_M, we define

$$\mathcal{M}^{\mathrm{alt}}(M) := \{M' \in \mathcal{M} \mid \pi_M \notin \boldsymbol{\pi}_{M'}\}, \tag{10.1.4}$$

the set of "alternative" models—the models $M' \in \mathcal{M}$ that disagree with M on the optimal decision[1]. When $\mathcal{M}$ is clear from context, we will abbreviate $\mathsf{g}^M := \mathsf{glc}(\mathcal{M}, M)$ (and $\mathsf{g}^\star := \mathsf{glc}(\mathcal{M}, M^\star)$). We also denote any solution to (10.1.3) by η^M—note that this is not in general unique. The characterization of [Graves and Lai, 1997] is as follows.

Proposition 10.1 ([Graves and Lai, 1997, Dong and Ma, 2022])**.** *For any model class* $\mathcal{M}$

[1]For $M \in \mathcal{M}$ such that π_M is not unique, we define $\mathcal{M}^{\mathrm{alt}}(M) := \{M' \in \mathcal{M} \mid \boldsymbol{\pi}_M \cap \boldsymbol{\pi}_{M'} = \varnothing\}$, and define $\mathsf{glc}(\mathcal{M}, M)$ as in (10.1.3), with respect to this $\mathcal{M}^{\mathrm{alt}}(M)$. We also define $\mathcal{M}^{\mathrm{alt}}(\pi) = \{M \in \mathcal{M} \mid \pi \notin \boldsymbol{\pi}_M\}$.

with $|\Pi| < \infty$*, any algorithm that is consistent with respect to* $\mathcal{M}$ *must have*

$$\mathbb{E}^{M^\star}[\mathbf{Reg}(T)] \geq \mathsf{glc}(\mathcal{M}, M^\star) \cdot \log(T) - o(\log(T)) \tag{10.1.5}$$

for any $M^\star \in \mathcal{M}$*, and there exists an algorithm which achieves*[2]

$$\mathbb{E}^{M^\star}[\mathbf{Reg}(T)] \leq \mathsf{glc}(\mathcal{M}, M^\star) \cdot \log(T) + o(\log(T)) \tag{10.1.6}$$

for any $M^\star \in \mathcal{M}$ *satisfying Assumption 10.3.*

The interpretation of the Graves-Lai Coefficient of $M^\star$ with respect to $\mathcal{M}$, $\mathsf{glc}(\mathcal{M}, M^\star)$, is simple. It asks, if $M^\star$ is known to the learner (to be clear, $M^\star$ is not known a-priori), what is the minimum regret that must be incurred to gather enough information to rule out all possible alternatives $M' \in \mathcal{M}$ which do not have $\pi_{M^\star}$ as an optimal decision (i.e., $\pi_{M^\star} \notin \boldsymbol{\pi}_{M'}$)? In other words, it aims to *certify* that $\pi_{M^\star}$ is indeed the optimal decision while incurring the minimum regret possible. This intuition is explicitly encoded in (10.1.3): the objective, $\sum_{\pi\in\Pi} \eta(\pi)\Delta^{M^\star}(\pi)$ denotes the regret that an allocation $\eta \in \mathbb{R}_+^\Pi$ will incur on the true instance $M^\star$, while the constraint $\sum_{\pi\in\Pi} \eta(\pi) D_{\mathsf{KL}}(M^\star(\pi) \,\|\, M'(\pi)) \geq 1$ ensures that $M^\star$ and M' are statistically distinguishable under the allocation η, for all $M' \in \mathcal{M}^{\mathrm{alt}}(M^\star)$.

As a simple example, for the finite-armed bandit problem where $\Pi = [A]$ and $\mathcal{M} = \{M(\pi) = \mathcal{N}(f(\pi), 1) \mid f \in \mathbb{R}^A\}$, a straightforward calculation shows that

$$\mathsf{glc}(\mathcal{M}, M^\star) \propto \sum_{\pi \neq \pi_{M^\star}} \frac{1}{\Delta^{M^\star}(\pi)}$$

and any optimal allocation has $\eta^{M^\star}(\pi) \propto \frac{1}{(\Delta^{M^\star}(\pi))^2}$ for $\pi \neq \pi_{M^\star}$. In this case, Proposition 10.1 recovers the well-known gap-dependent logarithmic regret bound

$$\mathbb{E}^{M^\star}[\mathbf{Reg}(T)] \lesssim \sum_{\pi \neq \pi_{M^\star}} \frac{\log(T)}{\Delta^{M^\star}(\pi)}$$

of Lai and Robbins [1985], and shows that it is instance-optimal. Note that this bound can offer significant improvement over the minimax rate $\Theta\big(\sqrt{AT}\big)$ when T is large.

The Graves-Lai Coefficient characterization is appealing in its simplicity, but the catch—at least when one moves beyond finite-armed bandits—is hiding in the lower-order terms, particularly for the upper bound (10.1.6). For general model classes, the best known finite-time regret bounds [Dong and Ma, 2022] take the form

$$\mathbb{E}^{M^\star}[\mathbf{Reg}(T)] \leq \mathsf{glc}(\mathcal{M}, M^\star) \cdot \log(T) + \mathrm{poly}(|\Pi|, (\Delta_{\min}^{M^\star})^{-1}) \cdot \log^{1-c}(T) \tag{10.1.7}$$

[2]To be precise, rather than scaling directly with $\mathsf{glc}(\mathcal{M}, M^\star)$, the upper bound given by Dong and Ma [2022] scales with a quantity $\mathsf{glc}_T(\mathcal{M}, M^\star)$ such that $\mathsf{glc}_T(\mathcal{M}, M^\star) \to_T \mathsf{glc}(\mathcal{M}, M^\star)$.

where $c > 0$ is a universal constant. While this indeed leads to instance-optimality as $T \to \infty$, the "lower-order" term in (10.1.7) scales with the size of the decision space, which is intractably large for most problems of interest. As an example, consider the problem of tabular reinforcement learning in an episodic MDP with S states, A actions, and horizon H. Here, we typically have $\mathsf{glc}(\mathcal{M}, M^\star) = \mathrm{poly}(S, A, H)$ (in particular, the Graves-Lai Coefficient can be bounded in terms of suboptimality gaps for the optimal value function and reachability for states of interest [Simchowitz and Jamieson, 2019, Dann et al., 2021, Wagenmaker et al., 2022c]), yet $|\Pi| = A^{HS}$. Consequently, the Graves-Lai Coefficient does not become the dominant term in (10.1.7) until $\geq \exp(\exp(S))$. That is, for realistic time horizons, asymptotic instance-optimality does not tell the full story.

Learning an optimal allocation. Given knowledge of an optimal Graves-Lai allocation $\eta^{M^\star}$ solving (10.1.3), a learner could simply take actions as specified by $\eta^{M^\star}$, and would achieve the instance-optimal rate given in Proposition 10.1. However, this is typically infeasible, as the optimal allocation itself depends strongly upon the ground truth model $M^\star$, and is therefore unknown to the learner. In light of this challenge, the approach taken by essentially all existing algorithms [Burnetas and Katehakis, 1996, Graves and Lai, 1997, Magureanu et al., 2014, Komiyama et al., 2015, Lattimore and Szepesvari, 2017, Combes et al., 2017, Hao et al., 2020, Van Parys and Golrezaei, 2023, Degenne et al., 2020b, Tirinzoni et al., 2020, Kirschner et al., 2021, Dong and Ma, 2022] is to first learn an estimate for a Graves-Lai allocation $\eta^{M^\star}$, and then take actions as specified by this estimate. In addition to being natural, this approach is *necessary* in a certain (weak) sense: for any algorithm that achieves instance-optimality, the expected decision frequencies must converge to an approximately optimal allocation as T grows (cf. Lemma 10.4.1).[3]

The presence of the lower-order term scaling with $|\Pi|$ in (10.1.6) (and in similar regret bounds from most existing work) reflects the sample complexity required to learn an optimal Graves-Lai allocation through uniform exploration. Specifically, one can estimate an allocation by uniformly exploring the decision space to gather data, and then solving an empirical approximation to the Graves-Lai program (10.1.3). Naive exploration of this type inevitably results in $\Omega(|\Pi|)$ sample complexity, and it is natural to ask whether a more deliberate exploration strategy, perhaps by exploiting the structure of the class $\mathcal{M}$, could lead to better finite-time regret bounds. For the setting of linear bandits, where $\Pi \subseteq \mathbb{R}^d$ and the mean reward function $\pi \mapsto f^M(\pi)$ is linear, this is indeed the case: a recent line of work [Tirinzoni et al., 2020, Kirschner et al., 2021] provides regret bounds of the form

$$\mathbb{E}^{M^\star}[\mathbf{Reg}(T)] \leq \mathsf{glc}(\mathcal{M}, M^\star) \cdot \log(T) + \mathrm{poly}(d, (\Delta_{\min}^{M^\star})^{-1}) \cdot \log^{1-c}(T).$$

This bound replaces the size of the decision space in the lower-order term by the dimension d, reflecting the fact that there are only d "effective" directions in which exploration is required.

[3]The connection between instance-optimal regret and learning an optimal allocation has many subtleties; we refer ahead to Section 10.4 for extensive discussion.

While this is an encouraging start, the techniques used in these works are specialized to linear bandits, and it is unclear how to generalize them beyond this setting. Recently, some progress has been made in understanding more general structured bandit problems [Jun and Zhang, 2020], yet a complete understanding of such problems remains lacking, and even less is known in complex decision making problems such as reinforcement learning.

10.1.3 A Motivating Example

As discussed in the prequel, for finite-armed bandits and linear bandits, it is possible to achieve instance-optimal regret bounds where the lower-order terms scale with reasonable problem-dependent quantities of interest: the number of actions A for bandits, and the dimension d for linear bandits. Informally, these quantities reflect the amount of exploration required to learn a near-optimal Graves-Lai allocation for the underlying model $M^\star$. One might be tempted to ask whether this phenomenon is universal. That is, can we always learn a near-optimal allocation with sample complexity no larger than, say, the minimax rate for $\mathcal{M}$? The starting point for our work is to recognize that in general, the answer is no: existing notions of statistical complexity—including those proposed in the minimax framework [Foster et al., 2021, 2022, 2023] and the Graves-Lai Coefficient itself—are insufficient to capture the complexity of learning the Graves-Lai allocation in finite time. To highlight this, consider the following simple example.

Example 10.1.1 (Searching for an informative arm)**.** Let $A, N \geq 2$ and $\beta \in (0,1)$ be parameters, and consider the class $\mathcal{M}$ of all models defined as follows. First, $\Pi = [A] \cup \{\pi_i^\circ\}_{i\in[N]}$; decisions in $[A]$ are "bandit arms", and decisions in $\{\pi_i^\circ\}_{i\in[N]}$ are "informative" (or, revealing) arms. Each model M has a unique optimal decision π_{M}, and the following structure, with $\mathcal{O} = [A] \cup \{\perp\}$.

- For each bandit arm $k \in [A]$ we have $r \sim \mathcal{N}(f^M(k), 1)$ for $f^M \in [0,1]$. There are no observations, i.e. $o = \perp$ almost surely.
- All informative arms π_k° give 0 reward almost surely. There exists a unique informative arm $\pi_{M}^\circ \in \{\pi_i^\circ\}_{i\in[N]}$ associated with M, so that if we play any π_k°, we receive an observation

$$o \sim \begin{cases} \mathrm{Unif}([A]), & \pi_k^\circ \neq \pi_{M}^\circ, \\ \beta \mathbb{I}_{\pi_{M}} + (1-\beta)\mathrm{Unif}([A]), & \pi_k^\circ = \pi_{M}^\circ. \end{cases}$$

We take $\mathcal{M}$ to consist of all possible models with this structure. The interpretation here is as follows. Suppose for concreteness that $\beta = 9/10$. If one were to ignore the revealing arms $\{\pi_i^\circ\}_{i\in[N]}$, this would be a standard finite-armed bandit problem. In particular, if we were to consider a model $M^\star$ with $f^{M^\star}(\pi) = \frac{1}{2} + \Delta\mathbb{I}\{\pi = i\}$ for $i \in [A]$, a standard calculation would

yield

$$\mathsf{g}^{M^\star} \propto \frac{A}{\Delta}. \tag{10.1.8}$$

However, the presence of the informative arms makes the problem substantially easier. With $\beta = 9/10$ (for concreteness), one can see that for the model M, pulling the informative arm $\pi^\circ_{M^\star}$ will give $o = \pi_{M^\star}$ with probability at least $9/10$, meaning that we can identify that $\pi_{M^\star}$ is optimal with high probability by pulling $\pi^\circ_{M^\star}$ a constant number of times. It follows that the optimal allocation is to ignore the bandit arms and set $\eta^{M^\star}(\pi) \propto \mathbb{I}\{\pi = \pi^\circ_{M^\star}\}$. This gives $\mathsf{g}^{M^\star} \leq O(1)$, which is substantially better than $\mathsf{g}^{M^\star} \propto \frac{A}{\Delta}$ if Δ is small or A is large.

If one is only concerned with asymptotic rates, this is the end of the story, but for non-asymptotic rates, we need to consider the amount of exploration required to *learn the optimal allocation*. In particular, in order to identify the informative arm $\pi^\circ_{M^\star}$, which is necessary to learn the optimal allocation, it is clear that in the worst case, any algorithm needs to try all of the revealing arms, leading to

$$\mathbb{E}[\mathbf{Reg}(T)] = \Omega(N).$$

This shows that while the complexity of learning the optimal allocation is washed away by an asymptotic analysis:

$$\lim_{T\to\infty} \frac{O(1) \cdot \log(T) + N}{\log(T)} = O(1),$$

it cannot be ignored for finite T. In addition, the $\Omega(N)$ factor cannot be explained away by standard complexity measures:

- By the argument above, $\sup_{M\in\mathcal{M}} \mathsf{g}^M = O(1)$, so the Graves-Lai Coefficient—even in a worst-case sense—does not reflect the complexity of learning an optimal allocation $\eta^{M^\star}$.
- The minimax rate for this problem is always bounded by $O(\sqrt{AT})$ (since we can treat it as a multi-armed bandit instance), which does not scale with N. Yet, $\Omega(A)$ sample complexity does not suffice to learn an optimal allocation. As a result, existing complexity measures such as the Decision-Estimation Coefficient [Foster et al., 2021], Eluder dimension [Russo and Van Roy, 2013], and information ratio [Russo and Van Roy, 2018], which are tailored to the minimax setting, cannot explain away the $\Omega(N)$ factor above.

Example 10.1.1 shows that if we want to achieve instance-optimality in finite time, new notions of problem complexity for the class $\mathcal{M}$ are required. This motivates the following central questions:

1. Can we develop algorithms for general model classes $\mathcal{M}$ that achieve non-asymptotic instance-optimal regret bounds of the form

$$\mathbb{E}^{M^\star}[\mathbf{Reg}(T)] \leq \mathsf{glc}(\mathcal{M}, M^\star) \cdot \log(T) + \mathrm{comp}(\mathcal{M}) \cdot \log^{1-c}(T), \tag{10.1.9}$$

where comp($\mathcal{M}$) is a complexity measure that reflects the intrinsic difficulty of exploration for $\mathcal{M}$—in particular, the difficulty of exploring to learn a Graves-Lai optimal allocation for? Ideally, such a complexity measure should be small for standard classes of interest, generalizing what is already known for finite-armed and linear bandits.

2. Can we understand when the presence of such lower-order terms is *necessary*?

10.1.4 The Allocation-Estimation Coefficient

To answer these questions and capture the statistical complexity of learning an optimal Graves-Lai allocation in finite time, we provide a new complexity measure, the *Allocation-Estimation Coefficient* (AEC). To describe the AEC, let us introduce some additional notation. For a model $M \in \mathcal{M}$ and parameter $\varepsilon \in [0, 1]$, we define

$$\Upsilon(M;\varepsilon) = \Bigg\{\lambda \in \triangle_\Pi \,|\exists \mathsf{n} \in \mathbb{R}_+ \text{ s.t. } \mathbb{E}_{\pi\sim\lambda}[\Delta^M(\pi)] \le \frac{(1+\varepsilon)\mathsf{g}^M}{\mathsf{n}}, \\ \inf_{M'\in\mathcal{M}^{\mathrm{alt}}(M)} \mathbb{E}_{\pi\sim\lambda}[D_{\mathsf{KL}}(M(\pi)\,\|\,M'(\pi))] \ge \frac{1-\varepsilon}{\mathsf{n}}\Bigg\} \tag{10.1.10}$$

the set of (normalized) allocations $\lambda \in \triangle_\Pi$ which are ε-optimal for the Graves-Lai program $\mathsf{glc}(\mathcal{M}, M)$ in (10.1.3)—both in terms of achieving the optimal objective value and satisfying the information constraint. In addition, for a distribution $\lambda \in \triangle_\Pi$, we define

$$\mathcal{M}^{\mathrm{gl}}_\varepsilon(\lambda) = \{M \in \mathcal{M} \mid \lambda \in \Upsilon(M;\varepsilon)\}. \tag{10.1.11}$$

Informally, $\mathcal{M}^{\mathrm{gl}}_\varepsilon(\lambda)$ represents the set of models for which the (normalized allocation) $\lambda \in \triangle_\Pi$ is ε-optimal for the Graves-Lai program $\mathsf{glc}(\mathcal{M}, M)$.

For a *reference model* $\overline{M} : \Pi \to \triangle_{\mathcal{R}\times\mathcal{O}}$ (not necessarily in $\mathcal{M}$) and parameter $\varepsilon > 0$, the Allocation-Estimation Coefficient is given by

$$\mathsf{aec}_\varepsilon(\mathcal{M}, \overline{M}) = \inf_{\lambda,\omega\in\triangle_\Pi} \sup_{M\in\mathcal{M}\setminus\mathcal{M}^{\mathrm{gl}}_\varepsilon(\lambda)} \left\{\frac{1}{\mathbb{E}_{\pi\sim\omega}\big[D_{\mathsf{KL}}\big(\overline{M}(\pi)\,\|\,M(\pi)\big)\big]}\right\}, \tag{10.1.12}$$

where we adopt the convention that the value is 0 if $\mathcal{M}^{\mathrm{gl}}_\varepsilon(\lambda) = \mathcal{M}$. In addition, letting co($\mathcal{M}$) denote the convex hull for $\mathcal{M}$, we define $\mathsf{aec}_\varepsilon(\mathcal{M}) := \sup_{\overline{M}\in\mathrm{co}(\mathcal{M})} \mathsf{aec}_\varepsilon(\mathcal{M}, \overline{M})$.

The Allocation-Estimation Coefficient is a game between a min-player choosing $\lambda, \omega \in \triangle_\Pi$ and a max-player choosing a model $M \in \mathcal{M}$ (with the restriction that $M \notin \mathcal{M}^{\mathrm{gl}}_\varepsilon(\lambda)$). Let us interpret the value. First:

- The distribution $\lambda \in \triangle_\Pi$ represents a normalized Graves-Lai allocation, while $\omega \in \triangle_\Pi$ is an *exploration* distribution used to gather information.

- The reference model $\overline{M}$ should be interpreted as a guess for the underlying model $M^\star \in \mathcal{M}$.

When $\lambda \in \triangle_\Pi$ is fixed, the value

$$\inf_{\omega \in \triangle_\Pi} \sup_{M \in \mathcal{M} \setminus \mathcal{M}_\varepsilon^{\mathrm{gl}}(\lambda)} \left\{ \frac{1}{\mathbb{E}_{\pi \sim \omega}\left[D_{\mathsf{KL}}\left(\overline{M}(\pi) \,\|\, M(\pi) \right) \right]} \right\}$$

represents the time required to gather enough information to distinguish between the reference model $\overline{M}$ and all alternative models $M \notin \mathcal{M}_\varepsilon^{\mathrm{gl}}(\lambda)$ for which λ is not an ε-optimal Graves-Lai allocation—provided that we explore optimally by minimizing over $\omega \in \triangle_\Pi$. For intuition, consider the case when $\overline{M} \in \mathcal{M}$. In this case, λ *must* be chosen so that $\overline{M} \in \mathcal{M}_\varepsilon^{\mathrm{gl}}(\lambda)$ (i.e., λ must be a Graves-Lai optimal allocation for $\overline{M}$), as otherwise the value of the AEC will be infinite, since $\mathbb{E}_{\pi \sim \omega}\left[D_{\mathsf{KL}}\left(\overline{M}(\pi) \,\|\, \overline{M}(\pi) \right) \right] = 0$ for all ω. Therefore, in such cases, the AEC reflects the *difficulty of distinguishing* $\overline{M}$ *from models that have different Graves-Lai optimal allocations*. In particular, such models might have the *same optimal decision as* $\overline{M}$ but, if our goal is to play a Graves-Lai optimal allocation for $\overline{M}$, we must still distinguish $\overline{M}$ from such models. The value $\mathsf{aec}_\varepsilon(\mathcal{M}, \overline{M})$ then reflects the *least possible time required* to distinguish such models, which is achieved by minimizing over the best possible (normalized) Graves-Lai allocation, $\lambda \in \triangle_\Pi$.

The Allocation-Estimation Coefficient plays a natural role for deriving both upper and lower bounds on the time required to learn an optimal allocation. For lower bounds, the significance of the AEC is somewhat immediate: it precisely quantifies the time required to acquire enough information to learn an ε-optimal allocation for the *best possible exploration strategy*, and thus leads to a lower bound on time required to learn such an allocation for any algorithm. Notably, the AEC serves as a lower bound for *all possible model classes* $\mathcal{M}$, and hence may be thought of as an intrinsic structural property of the class $\mathcal{M}$.

For upper bounds, the Allocation-Estimation Coefficient acts as a mechanism to drive exploration. Given an estimate $\widehat{M}^t$ for $M^\star$, if we select the decision $\pi^t \in \Pi$ using an appropriate mixture of the distributions (λ, ω) that attain the value of $\mathsf{aec}_\varepsilon(\mathcal{M}, \widehat{M}^t)$, we are guaranteed that one of two good outcomes occurs. Either:

1. the normalized allocation $\lambda \in \triangle_\Pi$ is ε-optimal for $\mathsf{glc}(\mathcal{M}, M^\star)$, so that by playing λ the learner matches the performance of the optimal allocation (up to a tolerance ε), thereby incurring the minimum amount of regret needed to gain information that allows it to distinguish $M^\star$ from $M \in \mathcal{M}^{\mathrm{alt}}(M^\star)$, or

2. $M^\star \notin \mathcal{M}_\varepsilon^{\mathrm{gl}}(\lambda)$ (that is, λ is not an ε-optimal Graves-Lai allocation), in which case—from the definition of the Allocation-Estimation Coefficient—playing $\omega \in \triangle_\Pi$ will allow us to better distinguish $M^\star$ from $\widehat{M}^t$.

As long as we can perform estimation in a consistent, online fashion, this reasoning will allow us to argue that $M^\star \in \mathcal{M}_\varepsilon^{\mathrm{gl}}(\lambda)$ the majority of the time, and that the total regret incurred through the learning process will scale with $\mathsf{aec}_\varepsilon(\mathcal{M})$.

Generalized Allocation-Estimation Coefficient. For certain results, we make use of the following, slightly more general variant of the AEC. For a reference model $\overline{M} : \Pi \to \triangle_{\mathcal{R}\times\mathcal{O}}$ and *subset of models* $\mathcal{M}_0 \subseteq \mathcal{M}$, we define

$$\mathsf{aec}_\varepsilon^{\mathcal{M}}(\mathcal{M}_0, \overline{M}) = \inf_{\lambda,\omega\in\triangle_\Pi} \sup_{M\in\mathcal{M}_0\setminus\mathcal{M}_\varepsilon^{\mathrm{gl}}(\lambda)} \left\{ \frac{1}{\mathbb{E}_{\pi\sim\omega}\left[D_{\mathsf{KL}}\left(\overline{M}(\pi) \,\|\, M(\pi)\right)\right]} \right\}, \tag{10.1.13}$$

where we adopt the convention that the value is 0 if $\mathcal{M}_0\setminus\mathcal{M}_\varepsilon^{\mathrm{gl}}(\lambda) = \varnothing$. Here $\mathcal{M}_0$ denotes the set we take the supremum over, while $\mathcal{M}$ denotes the set that $\mathcal{M}_\varepsilon^{\mathrm{gl}}(\lambda)$ is defined with respect to (i.e., the set with respect to which the Graves-Lai allocation is defined). When $\mathcal{M}_0 = \mathcal{M}$, we recover the AEC as defined in (10.1.12): $\mathsf{aec}_\varepsilon(\mathcal{M}, \overline{M}) = \mathsf{aec}_\varepsilon^{\mathcal{M}}(\mathcal{M}, \overline{M})$.

10.2 Non-Asymptotic Upper Bounds on Instance-Optimal Regret

This section presents our main algorithm and regret bounds. We begin by introducing the most basic variant of our algorithm, AE^2, and using it to provide instance-optimal regret bounds for simple settings (Sections 10.2.2 and 10.2.3); with preliminaries in Section 10.2.1. We then give a refined variant of the algorithm, $\mathsf{AE}^2_\star$, which adapts to the minimum gap $\Delta^\star_{\min}$ and leads to regret bounds under relaxed regularity conditions (Section 10.2.4 and Section 10.2.5). We conclude with an overview of our analysis in Section 10.2.6. For all results in this section, we assume that Assumptions 10.1 to 10.3 hold.

10.2.1 Regularity Conditions

To present our algorithm and results, we first introduce several regularity conditions for the model class $\mathcal{M}$.

Likelihood ratios. We next make two assumptions concerning smoothness of KL divergences and behavior of log-likelihood ratios.

Assumption 10.4 (Smooth KL)**.** *There exists $L_{\mathsf{KL}} > 0$ such that for all $M, M', M'' \in \mathcal{M}$ and $\pi \in \Pi$,*

$$|D_{\mathsf{KL}}(M(\pi) \,\|\, M''(\pi)) - D_{\mathsf{KL}}(M'(\pi) \,\|\, M''(\pi))| \le L_{\mathsf{KL}} \sqrt{D_{\mathsf{KL}}(M(\pi) \,\|\, M'(\pi))}.$$

Assumption 10.5 (Sub-Gaussian Log-Likelihood)**.** *There exists $V_{\mathcal{M}} > 0$ such that for all*

$M, M', M'' \in \mathcal{M}$ and $\pi \in \Pi$,

$$\mathbb{P}_{(r,o)\sim M(\pi)}\left[\left|\log \frac{\mathbb{P}^{M',\pi}(r,o)}{\mathbb{P}^{M'',\pi}(r,o)} - \mathbb{E}_{(r',o')\sim M(\pi)}\left[\log \frac{\mathbb{P}^{M',\pi}(r',o')}{\mathbb{P}^{M'',\pi}(r',o')}\right]\right| \geq x\right] \leq 2\exp(-x^2/V_{\mathcal{M}}^2)$$

for all $x \geq 0$.

Assumption 10.4 and Assumption 10.5 facilitate finite-sample estimation guarantees with respect to the KL divergence. Both assumptions are met by standard problem classes, including general structured bandit problems with Gaussian noise. Existing works that consider general model classes make similar assumptions [Dong and Ma, 2022].

Estimation. To provide estimation guarantees that accommodate infinite classes $\mathcal{M}$, we assume certain covering properties. We will consider the following notion of a cover.

Definition 10.2.1 ((ρ,μ)-Cover)**.** We say that a set $\mathcal{M}_{\text{cov}} \subseteq \mathcal{M}$ is a (ρ,μ)-cover of $\mathcal{M}$ if there exists some event $\mathcal{E}$ such that:[4]

1. $\sup_{M\in\mathcal{M}} \sup_{\pi\in\Pi} \mathbb{P}^{M,\pi}(\mathcal{E}^c) \leq \mu$.
2. For each $M \in \mathcal{M}$, there exists some $M' \in \mathcal{M}_{\text{cov}}$ such that

$$\left|\log \mathbb{P}^{M,\pi}(r,o) - \log \mathbb{P}^{M',\pi}(r,o)\right| \leq \rho$$

   for all $(r,o) \in \mathcal{R}\times\mathcal{O}$ with $\sup_{M''\in\mathcal{M}} \mathbb{P}^{M'',\pi}(r,o \mid \mathcal{E}) > 0$.

We denote the size of the smallest such cover by $\mathsf{N}_{\text{cov}}(\mathcal{M},\rho,\mu)$.

Definition 10.2.1 states that the log-likelihoods are "covered" under some good event $\mathcal{E}$ which occurs with high probability: for any model in the class, we can find some model in the cover with log-likelihoods that are "close" on $\mathcal{E}$. We assume that the covering number for the model class $\mathcal{M}$ is bounded, and has reasonable ("parametric") growth.

Assumption 10.6 (Bounded Covering Number)**.** *For some parameters $d_{\mathsf{cov}} \geq 1, C_{\mathsf{cov}} \geq 1$, we have*

$$\log \mathsf{N}_{\text{cov}}(\mathcal{M},\rho,\mu) \leq d_{\mathsf{cov}} \cdot \log\left(\frac{C_{\mathsf{cov}}}{\rho\mu}\right).$$

Note that the rate of growth of the covering number required by Assumption 10.6 is the standard rate of growth for parametric (e.g., linear) classes. Our results easily extend to accommodate general growth rates, but we adopt Assumption 10.6 because it suffices for all of the examples we will consider, and simplifies presentation.

[4]Note that we require that $\mathcal{M}_{\text{cov}} \subseteq \mathcal{M}$, i.e. that $\mathcal{M}_{\text{cov}}$ is a *proper* cover.

Information content of optimal decisions. As noted in the introduction, the Graves-Lai Coefficient $\mathsf{g}^\star = \mathsf{glc}(\mathcal{M}, M^\star)$ can be thought of as the minimal regret needed to distinguish $M^\star$ from all possible models with different optimal decisions. As playing the optimal decision, $\pi_\star$, incurs no regret, any allocation η which is optimal for the Graves-Lai program, (10.1.3), will still be optimal if we increase the number of plays of $\pi_\star$ arbitrarily. As we are interested in finite-time behavior in this work, it is undesirable to consider allocations for which the number of pulls of optimal decisions are arbitrarily large. Instead, we would like to consider allocations which play optimal decisions only as long as they still provides useful information about models in the alternate set. The following definition gives a formal quantification of this.

Definition 10.2.2 (Information Content of Optimal Decision)**.** Fix $\epsilon \in (0, 1/2]$. For a model $M \in \mathcal{M}$, we define $\mathsf{n}_\varepsilon^M > 0$ as the minimum value such that, for any allocation $\eta \in \mathbb{R}_+^\Pi$ satisfying

$$(1+\varepsilon)\mathsf{g}^M \geq \sum_{\pi\in\Pi} \eta(\pi)\Delta^M(\pi) \quad \text{and} \quad \inf_{M'\in\mathcal{M}^{\mathrm{alt}}(M)} \sum_{\pi\in\Pi} \eta(\pi) D_{\mathsf{KL}}(M(\pi) \,\|\, M'(\pi)) \geq 1-\varepsilon,$$

we have

$$\inf_{M'\in\mathcal{M}^{\mathrm{alt}}(M)} \sum_{\pi\in\Pi, \pi\notin\boldsymbol{\pi}_M} \eta(\pi) D_{\mathsf{KL}}(M(\pi) \,\|\, M'(\pi)) + \sum_{\pi\in\boldsymbol{\pi}_M} \mathsf{n}_\varepsilon^M D_{\mathsf{KL}}(M(\pi) \,\|\, M'(\pi)) \geq 1-2\varepsilon.$$

We denote $\mathsf{n}_\varepsilon^{\mathcal{M}} := \sup_{M\in\mathcal{M}} \mathsf{n}_\varepsilon^M$.

Intuitively, any allocation which is ε-optimal for the Graves-Lai program (10.1.3) need not play any optimal decision $\pi_M \in \boldsymbol{\pi}_M$ more than n_ε^M times. Therefore, for model M, n_ε^M can be thought of as a quantification of the extent to which playing optimal decisions provides useful information—no additional useful information can be acquired on models in the alternate set $\mathcal{M}^{\mathrm{alt}}(M)$ by playing optimal decisions more than n_ε^M times. As we will see, n_ε^M is bounded polynomially in problem parameters for many classes of interest.

Uniformly regular classes. We refer to a class as *uniformly regular* if $\mathsf{n}_\varepsilon^{\mathcal{M}} < \infty$, and the following assumption on the minimum gaps holds.

Assumption 10.7 (Lower-Bounded Minimum Gap)**.** *We have* $\inf_{M\in\mathcal{M}} \Delta_{\min}^M > 0$. *We denote by* $\Delta_{\min} > 0$ *a (known) lower bound on* $\inf_{M\in\mathcal{M}} \Delta_{\min}^M$.

Note that Assumption 10.7 implies that for all $M \in \mathcal{M}$, π_M is unique. For the results concerning the most basic version of our algorithm, AE^2 (Sections 10.2.2 and 10.2.3), we assume for expositional purposes that the class $\mathcal{M}$ is uniformly regular. Our more general algorithm, $\mathsf{AE}_\star^2$ (Section 10.2.5), achieves guarantees similar to those of AE^2, but without uniform regularity. In particular, $\mathsf{AE}_\star^2$ replaces dependence on $\Delta_{\min}$ with the minimum gap $\Delta_{\min}^\star := \Delta_{\min}^{M^\star}$ for the true model, and replaces dependence on $\mathsf{n}_\varepsilon^{\mathcal{M}}$ with $\mathsf{n}_\varepsilon^\star := \mathsf{n}_\varepsilon^{M^\star}$. We note,

however, that in cases where a lower bound Δ on the minimum gap $\Delta^\star_{\min}$ of the true model is known a-priori, Assumption 10.7 can be satisfied by restricting the model class to models with minimum gap at least Δ.

10.2.2 The AE^2 Algorithm

We now present the most basic variant of our main algorithm, AE^2 (Algorithm 10.1). This will serve as the starting point for the most general version of our algorithm, $\mathsf{AE}^2_\star$ (Section 10.2.5). To describe the algorithm, we first introduce the primitive of an *online estimation oracle* [Foster and Rakhlin, 2020, Foster et al., 2021].

Estimation oracles. Algorithm 10.1 makes use of an *online estimation oracle*, denoted by $\mathbf{Alg}_{\mathsf{KL}}$, which is an algorithm that, given knowledge of the class $\mathcal{M}$, estimates the underlying model $M^\star \in \mathcal{M}$ from data in a sequential fashion. When invoked at step $s \in \mathbb{N}$ with the data $(\pi^1, r^1, o^1), \dots, (\pi^{s-1}, r^{s-1}, o^{s-1})$ observed so far, the estimation oracle builds an estimate

$$\widehat{M}^s = \mathbf{Alg}_{\mathsf{KL}}\Big(\big\{(\pi^i, r^i, o^i)\big\}_{i=1}^{s-1}\Big)$$

which aims to approximate the true model $M^\star$. Following [Foster et al., 2021, Chen et al., 2022, Foster et al., 2024], we make use of *randomized* estimation oracles that, at each step produces $\xi^s = \mathbf{Alg}_{\mathsf{KL}}\big(\{(\pi^i, r^i, o^i)\}_{i=1}^{s-1}\big)$, where $\xi^s \in \triangle_{\mathcal{M}}$ is a randomization distribution, and draw $\widehat{M} \sim \xi^s$. We measure the oracle's performance in terms of cumulative estimation error, defined as follows.

Definition 10.2.3 (Cumulative Estimation Error)**.** Consider the process where, for each round $i \in \mathbb{N}$, given $(\pi^1, r^1, o^1), \dots, (\pi^{i-1}, r^{i-1}, o^{i-1})$ with $\pi^i \sim p^i$ and $(r^i, o^i) \sim M^\star(\pi^i)$, the estimation oracle returns $\xi^i = \mathbf{Alg}_{\mathsf{KL}}((\pi^1, r^1, o^1), \dots, (\pi^{i-1}, r^{i-1}, o^{i-1}))$. For any $s \in \mathbb{N}$, we define the oracle's cumulative KL estimation error under this process as:

$$\mathbf{Est}_{\mathsf{KL}}(s) := \sum_{i=1}^{s} \mathbb{E}_{\widehat{M}\sim\xi^i}\Big[\mathbb{E}_{\pi\sim p^i}\Big[D_{\mathsf{KL}}\big(M^\star(\pi) \,\|\, \widehat{M}(\pi)\big)\Big]\Big].$$

Algorithm 10.1 can be invoked with any off-the-shelf algorithm for estimation, but our main results make use of the fact that under Assumption 10.5 and Assumption 10.6, there exists an estimation oracle $\mathbf{Alg}_{\mathsf{KL}}$ (Algorithm H.4 in Appendix H.3.3) which ensures that with probability at least $1-\delta$, for all $s \in \mathbb{N}$:

$$\mathbf{Est}_{\mathsf{KL}}(s) \lesssim V_{\mathcal{M}} \cdot d_{\mathsf{cov}} \cdot \log^{3/2}\left(\frac{C_{\mathsf{cov}} \cdot s}{\delta}\right). \tag{10.2.1}$$

That is, the estimation oracle ensures that the KL divergence between the true model $M^\star$ and the estimates returned scales at most poly-logarithmically in the exploration horizon. Note that on its own, this guarantee does not necessarily imply that $\widehat{M}^s = \mathbb{E}_{M\sim\xi^s}[M] \to M^\star$—low

Algorithm 10.1 Allocation Estimation via Adaptive Exploration (AE^2)

1: **input:** optimality tolerance ε, model class $\mathcal{M}$.
2: Initialize $s \leftarrow 1$, $\mathsf{n}_{\max} \leftarrow \mathsf{n}_{\max}(\mathcal{M}, \varepsilon/6)$, and $q \leftarrow \frac{4\mathsf{n}_{\max}+\varepsilon \underline{\mathsf{g}}^{\mathcal{M}}}{4\mathsf{n}_{\max}+2\varepsilon \underline{\mathsf{g}}^{\mathcal{M}}}$ for $\underline{\mathsf{g}}^{\mathcal{M}} := \inf_{M\in\mathcal{M}:\mathsf{g}^M>0} \mathsf{g}^M$.
3: Compute $\xi^1 \leftarrow \mathbf{Alg}_{\mathsf{KL}}(\{\varnothing\})$ and $\widehat{M}^1 \leftarrow \mathbb{E}_{M\sim\xi^1}[M]$.
4: **for** $t = 1, 2, 3, \ldots$ **do**
5: **if** $\exists \pi_{\widehat{M}^s} \in \boldsymbol{\pi}_{\widehat{M}^s}$ s.t. $\forall M \in \mathcal{M}^{\mathrm{alt}}(\pi_{\widehat{M}^s})$, $\sum_{i=1}^{s-1} \mathbb{E}_{\widehat{M}\sim\xi^i}\left[\log \frac{\mathbb{P}^{\widehat{M},\pi^i}(r^i,o^i)}{\mathbb{P}^{M,\pi^i}(r^i,o^i)}\right] \geq \log(t\log t)$ **then**
6: Play $\pi_{\widehat{M}^s}$. // Exploit
7: **else** // Explore
8: Set $p^s \leftarrow q\lambda^s + (1-q)\omega^s$ for

$$\lambda^s, \omega^s \leftarrow \operatorname*{arg\,min}_{\lambda,\omega\in\triangle_\Pi} \sup_{M\in\mathcal{M}\setminus\mathcal{M}^{\mathrm{gl}}_{\varepsilon/6}(\lambda;\mathsf{n}_{\max})} \frac{1}{\mathbb{E}_{\widehat{M}\sim\xi^s}\left[\mathbb{E}_{\pi\sim\omega}\left[D_{\mathsf{KL}}\left(\widehat{M}(\pi) \,\|\, M(\pi)\right)\right]\right]}. \tag{10.2.2}$$

9: Draw $\pi^s \sim p^s$ and observe reward r^s and observation o^s.
10: Compute estimate $\xi^{s+1} \leftarrow \mathbf{Alg}_{\mathsf{KL}}(\{(\pi^i, r^i, o^i)\}_{i=1}^s)$ and $\widehat{M}^{s+1} = \mathbb{E}_{\widehat{M}\sim\xi^{s+1}}[\widehat{M}]$.
11: $s \leftarrow s+1$.

online estimation error only requires that $\widehat{M}^s$ is on average close to $M^\star$ on the decisions we have actually played.

Algorithm overview. We are now ready to give an overview of Algorithm 10.1. The algorithm alternates between *exploit* steps and *explore* steps, tracking the number of explore steps that have been performed with a counter $s \in \mathbb{N}$. For each step $t \in \mathbb{N}$, the algorithm makes use of an estimator $\widehat{M}^s = \mathbb{E}_{\widehat{M}\sim\xi^s}[\widehat{M}]$, where $\xi^s = \mathbf{Alg}_{\mathsf{KL}}\left(\left\{(\pi^i, r^i, o^i)\right\}_{i=1}^{s-1}\right)$ is computed by calling the estimation oracle with data gathered at previous explore steps. Given the estimator, AE^2 performs a test based on likelihood ratios (Line 5) to check whether it has collected enough information to rule out all models for which $\pi_{\widehat{M}^s} \in \boldsymbol{\pi}_{\widehat{M}^s}$ is not an optimal decision. If so, it *exploits*, and plays $\pi_{\widehat{M}^s}$ (Line 6), as in this case $\pi_{\widehat{M}^s} = \pi_\star$ with high probability. If the test fails, the algorithm must gather more information to eliminate alternatives, and it *explores* (Line 8). The key component of the explore phase is the choice of the exploration distribution in (10.2.2), which is based on the Allocation-Estimation Coefficient program, but incorporates some small modifications: 1) First, $\widehat{M}$ is randomized according to the distribution ξ^s, 2) Second, the set $M^\star \in \mathcal{M}^{\mathrm{gl}}_\varepsilon(\lambda)$ is replaced with a smaller set $M^\star \in \mathcal{M}^{\mathrm{gl}}_\varepsilon(\lambda; \mathsf{n}_{\max})$, which requires that λ obeys certain normalization constraints; this is detailed below. Using the distributions λ^s (representing a normalized allocation) and ω^s (representing an exploration distribution) returned in (10.2.2), the algorithm computes a mixture $p^s = q\lambda^s + (1-q)\omega^s$, where $q \in (0,1)$ is a carefully chosen parameter, and plays $\pi^s \sim p^s$ (Line 9). The reward and observation (r^s, o^s) that result from playing π^s are then used to update the estimation oracle for subsequent rounds (Line 10).

To understand the intuition behind the explore phase and why the Allocation-Estimation Coefficient plays a useful role here, we can consider two cases. In the first case, if λ^s is an ε-optimal Graves-Lai allocation for $M^\star$ (that is, $M^\star \in \mathcal{M}^{\mathrm{gl}}_{\varepsilon/6}(\lambda^s; \mathsf{n}_{\max})$), then playing λ^s will optimize the tradeoff between minimizing regret on $M^\star$ and collecting information that allows one to distinguish $M^\star$ from $M \in \mathcal{M}^{\mathrm{alt}}(M^\star)$, and will therefore match the optimal performance prescribed by the Graves-Lai Coefficient, incurring regret scaling as $\mathsf{g}^\star$.

In the second case, if λ^s is not an ε-optimal Graves-Lai allocation for $M^\star$, we have $M^\star \notin \mathcal{M}^{\mathrm{gl}}_{\varepsilon/6}(\lambda^s; \mathsf{n}_{\max})$, so by the definition of the AEC (10.2.2), ω^s will place mass on actions that ensure $\mathbb{E}_{\widehat{M}\sim\xi^s}[\mathbb{E}_{\pi\sim\omega}[D_{\mathsf{KL}}(\widehat{M}(\pi) \,\|\, M^\star(\pi))]]$ is large; exactly how large this quantity is will be quantified by the value of the AEC. Since p^s plays ω^s with constant probability, the quantity

$$\mathbb{E}_{M\sim\xi^s}[\mathbb{E}_{\pi\sim p^s}[D_{\mathsf{KL}}(M^\star(\pi) \,\|\, M(\pi))]]$$

will also be large, but if the estimation oracle is consistent in the sense of Definition 10.2.3, this can only happen a small number of times. In particular, if (10.2.1) holds, the number of times in which we encounter this second case is at most *logarithmic* in the number of exploration rounds. As such, we can show that λ^s must be a near-optimal Graves-Lai allocation for $M^\star$ on all but a logarithmic number of exploration rounds, and that AE^2 achieves the optimal rate on such rounds.

Critically, rather than exploring in a naive fashion (e.g., by sampling decisions uniformly), AE^2 explores only to the extent necessary to learn a Graves-Lai allocation for $M^\star$. There may exist instances $M \neq M^\star$ which differ significantly from $M^\star$ but have a similar Graves-Lai allocations—AE^2 will make no effort to distinguish such instances since, as long as it knows that one of these instances is correct, it can simply play their shared Graves-Lai allocation. This notion of exploration, which is targeted toward distinguishing instances that have different Graves-Lai allocations, is precisely the notion captured by the Allocation-Estimation Coefficient.

Normalization factor for allocations. As noted in Section 10.2.1, while an optimal Graves-Lai allocation may place an arbitrarily large number of pulls on an optimal decision, for finite-time guarantees it is useful to restrict to allocations which place only finite mass on optimal decisions. To this end, AE^2 restricts the optimization problem based on the AEC in (10.2.2) to only consider normalized allocations λ for which the *normalization factor* is at most

$$\mathsf{n}_{\max}(\mathcal{M}, \varepsilon) := \frac{64}{\Delta^2_{\min}} \cdot \left(\frac{1}{\varepsilon} + V_{\mathcal{M}} \mathsf{n}^{\mathcal{M}}_{\varepsilon}\right) \cdot \max_{M\in\mathcal{M}} \mathsf{g}^M. \tag{10.2.3}$$

where the normalization factor refers to the value n in the definition of $\Upsilon(M;\varepsilon)$ (see (10.1.10)). In particular, to enforce this restriction, the optimization problem in (10.2.2) restricts the max-player to $M \in \mathcal{M}\backslash\mathcal{M}^{\mathrm{gl}}_{\varepsilon/6}(\lambda; \mathsf{n}_{\max})$, where $\mathsf{n}_{\max} := \mathsf{n}_{\max}(\mathcal{M}, \varepsilon/6)$ and $\mathcal{M}^{\mathrm{gl}}_{\varepsilon/6}(\lambda; \mathsf{n}_{\max})$ is

defined identically to $\mathcal{M}^{\mathrm{gl}}_{\varepsilon/6}(\lambda)$ in (10.1.11), but with n restricted to $\mathsf{n} \leq \mathsf{n}_{\max}$. As we show in Lemma H.2.4, for $\mathsf{n}_{\max}(\mathcal{M}, \varepsilon)$ defined as in (10.2.3), the optimal value in (10.2.2) can be bounded by the AEC.

Computational efficiency. The primary computational burden in AE^2 lies in solving the optimization problem (10.2.2) to compute the exploration distributions. In general there is little hope of solving this efficiently (i.e., in time sublinear in $|\Pi|$ and $|\mathcal{M}|$)—indeed, in some cases it may be that to even determine whether $M \in \mathcal{M}^{\mathrm{gl}}_{\varepsilon}(\lambda)$ will require enumerating the model class $\mathcal{M}$. However, for nicely structured problems, we anticipate that this program can be solved, or at least approximated, efficiently. As the focus of our work is primarily statistical, we leave further exploration as to when the algorithm can be implemented efficiently to future work.

Simplicity. We emphasize the simplicity of AE^2. Most existing algorithms which achieve instance-optimality are quite complex, even in specialized settings such as linear bandits. In contrast, AE^2 is very simple and intuitive, and relies only on three basic components: an explore-exploit test, an estimation oracle, and a single optimization to compute the exploration distributions. Despite its simplicity, as we show, AE^2 obtains comparable or better performance over existing approaches.

Relation to existing approaches. At a very high level, AE^2 bears some similarity to the $\mathsf{E2D}$ algorithm of Foster et al. [2021], which achieves the *minimax* optimal rate for general classes $\mathcal{M}$ in the DMSO framework. Both algorithms rely on online estimation algorithms, and both solve min-max programs based on the output of the estimator to determine which allocations to play. However, the algorithm design and analysis principles for the two algorithms, and in particular the motivation for the min-max programs they solve, differ significantly.

10.2.3 AE^2 Algorithm: Regret Bound for Uniformly Regular Classes

We present upper bounds for AE^2 in the setting where our class $\mathcal{M}$ is uniformly regular: Assumption 10.7 holds and $\mathsf{n}^{\mathcal{M}}_{\varepsilon} < \infty$; these assumptions are relaxed by the $\mathsf{AE}^2_{\star}$ algorithm in the sequel. To state the regret bound for AE^2 in the tightest form possible, we introduce the following variant of the Allocation-Estimation Coefficient, which incorporates randomized estimators $\xi \in \triangle_{\mathcal{M}}$:

$$\overline{\mathsf{aec}}^{\mathcal{M}}_{\varepsilon}(\mathcal{M}_0, \xi) := \inf_{\lambda, \omega \in \triangle_{\Pi}} \sup_{M \in \mathcal{M} \setminus \mathcal{M}^{\mathrm{gl}}_{\varepsilon}(\lambda)} \frac{1}{\mathbb{E}_{\overline{M} \sim \xi}[\mathbb{E}_{\pi \sim \omega}[D_{\mathsf{KL}}\big(\overline{M}(\pi) \,\|\, M(\pi)\big)]]}, \tag{10.2.4}$$

with $\overline{\mathsf{aec}}_{\varepsilon}(\mathcal{M}, \xi) := \overline{\mathsf{aec}}^{\mathcal{M}}_{\varepsilon}(\mathcal{M}, \xi)$ and $\overline{\mathsf{aec}}_{\varepsilon}(\mathcal{M}) := \sup_{\xi \in \triangle_{\mathcal{M}}} \overline{\mathsf{aec}}_{\varepsilon}(\mathcal{M}, \xi)$. Note that one can always bound $\overline{\mathsf{aec}}_{\varepsilon}(\mathcal{M}) \leq \mathsf{aec}_{\varepsilon}(\mathcal{M})$ due to the convexity of the KL divergence. In fact, these definitions are equivalent up to dependence on problem-dependent parameters in Section 10.2.1 (indeed, our lower bounds in Section 10.4 scale with the latter quantity), but the former can

be simpler to bound for some of the examples we consider.

Our main theorem concerning the performance of AE^2 is as follows.

Theorem 10.1 (Regret Bound for AE^2)**.** *For any $\varepsilon \in (0, 1/2]$, there exists a choice for the estimation oracle $\mathbf{Alg}_{\mathsf{KL}}$ such that for all $T \in \mathbb{N}$, under Assumptions 10.4 to 10.7 and if $\mathsf{g}^\star > 0$, the expected regret of AE^2 is bounded by*

$$\mathbb{E}^{M^\star}[\mathbf{Reg}(T)] \le (1+\varepsilon)\mathsf{g}^\star \cdot \log(T) + \overline{\mathsf{aec}}_{\varepsilon/12}(\mathcal{M}) \cdot C_{\mathsf{aec}} \cdot \log^{3/2}(\log T) + C_{\mathsf{low}} \cdot \log^{1/2}(T), \tag{10.2.5}$$

where

$$C_{\mathsf{aec}} := c \cdot \frac{V_{\mathcal{M}}^2 d_{\mathsf{cov}} \log(C_{\mathsf{cov}}) \cdot \max_{M \in \mathcal{M}} \mathsf{g}^M}{\varepsilon \Delta_{\min}^3} \cdot \left(\varepsilon^{-1} + V_{\mathcal{M}} \mathsf{n}_{\varepsilon/6}^{\mathcal{M}}\right) \cdot \log(C_{\mathsf{low}}),$$

for a universal numerical constant $c > 0$, and C_{low} is a lower-order constant given by

$$C_{\mathsf{low}} := \mathrm{lin}\left(\max_{M \in \mathcal{M}} \mathsf{g}^M, \mathsf{aec}_{\varepsilon/12}^{1/2}(\mathcal{M}), \tfrac{1}{\varepsilon^2}, \tfrac{1}{\Delta_{\min}^3}, \mathsf{n}_{\varepsilon/6}^{\mathcal{M}}, L_{\mathsf{KL}}^2, V_{\mathcal{M}}^{13/2}, d_{\mathsf{cov}}, \log(C_{\mathsf{cov}}), \log\log T\right),$$

where $\mathrm{lin}(\cdot)$ *denotes a function multi-linear and poly-logarithmic in its arguments.*

We prove Theorem 10.1 in Appendix H.3.1, and give a proof sketch in Section 10.2.6. Theorem 10.1 shows that AE^2 achieves the asymptotically optimal Graves-Lai rate for $M^\star$, as given in Proposition 10.1, up to a $(1+\varepsilon)$ approximation factor. In more detail, if we label the terms in (10.2.5) as

$$\mathbb{E}^{M^\star}[\mathbf{Reg}(T)] \le \underbrace{(1+\varepsilon)\mathsf{g}^\star \cdot \log(T)}_{\text{(I)}} + \underbrace{\overline{\mathsf{aec}}_{\varepsilon/12}(\mathcal{M}) \cdot C_{\mathsf{aec}} \cdot \log^{3/2}(\log T)}_{\text{(II)}} + \underbrace{C_{\mathsf{low}} \cdot \log^{1/2}(T)}_{\text{(III)}},$$

the regret bound can be seen to consist of:

- The *leading-order term* (I) $= (1+\varepsilon)\mathsf{g}^\star \cdot \log(T)$. This is the only term that scales linearly with $\log(T)$, as a consequence we have $\lim_{T\to\infty} \frac{\mathbb{E}^{M^\star}[\mathbf{Reg}(T)]}{\log(T)} \le (1+\varepsilon)\mathsf{g}^\star$, which matches the instance-optimal rate given in Proposition 10.1 up to a factor of $(1+\varepsilon)$.

- A lower-order term (II) $= \overline{\mathsf{aec}}_{\varepsilon/12}(\mathcal{M}) \cdot C_{\mathsf{aec}} \cdot \log^{3/2}(\log T)$, which is polylogarithmic in $\log(T)$, and scales with $\overline{\mathsf{aec}}_{\varepsilon/12}(\mathcal{M})$, as well as regularity parameters from Section 10.2.1.

- A second lower-order term (III) $= C_{\mathsf{low}} \cdot \log^{1/2} T$. This term scales with $\log^{1/2}(T) = o(\log(T))$ and, like the term (II), scales with the AEC and regularity parameters from Section 10.2.1. Compared to (II), this term has worse dependence on $\log(T)$, but enjoys sublinear $\overline{\mathsf{aec}}_{\varepsilon/12}^{1/2}(\mathcal{M})$ scaling with the AEC.

Critically, both of the $o(\log T)$ lower-order terms above do not scale with (often exponentially large) terms such as $|\Pi|$ or $|\mathcal{M}|$ found in prior work, and instead scale principally with $\overline{\mathsf{aec}}_\varepsilon(\mathcal{M})$, which, as we will show in Section 10.4, is unavoidable in a certain sense. In particular, note that once T is large enough that

$$\log(T) \geq \widetilde{\Omega}^+(\overline{\mathsf{aec}}_{\varepsilon/12}(\mathcal{M})),$$

the leading-order term $(\mathrm{I}) = (1+\varepsilon)\mathsf{g}^\star \cdot \log(T)$ term in Theorem 10.1 will dominate the regret. This is precisely the time horizon given by the lower bound in Theorem 10.10, which is necessary for an algorithm to learn a near-optimal allocation for the Graves-Lai program. We offer a more thorough comparison of Theorem 10.1 with our lower bounds in Section 10.4.4. Below, we discuss the lower-order terms and asymptotic performance in greater detail.

Remark 10.2.1 (Additional Lower-Order Terms)**.** The lower-order terms in Theorem 10.1 depend on the model class $\mathcal{M}$ through the regularity, covering, and smoothness assumptions, as well as the minimum gap (Section 10.2.1). For many of the examples we consider, the Allocation-Estimation Coefficient will dominate these other terms, yet there may exist classes where this is not the case. Resolving the optimal dependence on these problem-dependent parameters in the lower-order terms, as well as understanding when these parameters are necessary, remains an interesting direction for future work.

In addition, let us mention that while both lower-order terms scale with $o(\log T)$ (note that the scaling is no larger than $O(\sqrt{\log T} \cdot \operatorname{polylog} \log T)$), it is not clear what the optimal dependence on T should be for the lower-order terms. For example, one might hope to replace the dependence on $\log^{1/2}(T)$ with $\log^a(T)$ for some constant $a < 1/2$, or even with $\operatorname{polylog}(\log(T))$. Precisely characterizing the optimal $\log(T)$ scaling for lower-order terms remains an interesting open question. To this end, we remark that Jun and Zhang [2020] show that in some cases, an $\Omega(\log \log T)$ term is indeed necessary.

Remark 10.2.2 (Asymptotic Performance)**.** Asymptotically, as $T \to \infty$, the regret of AE^2 scales with $(1+\varepsilon)\mathsf{g}^\star \cdot \log(T)$, which is a factor of $(1+\varepsilon)$ off from the asymptotic lower bound in Proposition 10.1. For any fixed $T \in \mathbb{N}$ of interest, as long as $\mathsf{aec}_\varepsilon(\mathcal{M}) = \operatorname{poly}(\varepsilon^{-1})$, one can obtain an asymptotic constant of 1 by choosing $\varepsilon = 1/\log^a(T)$ for a sufficiently small constant $a > 0$. For example, when $|\Pi| < \infty$, it is always possible to bound $\mathsf{aec}_\varepsilon(\mathcal{M}) \lesssim \operatorname{poly}(|\Pi|)/\varepsilon^4$ (see Proposition 10.2 below), so choosing ε as above ensures that the lower-order terms scale $o(\log T)$, while the leading-order term scales as $\mathsf{g}^\star \cdot \log(T)$ asymptotically.

10.2.3.1 Example: Searching for an Informative Arm

We next provide an example of a uniformly regular class in order to illustrate a case where Theorem 10.1 holds. In particular, we revisit the *informative arm* setting described in the introduction (Example 10.1.1). Recall that we exhibited a model class for which the complexity of learning the Graves-Lai allocation is not governed by existing complexity measures, and can be larger than the minimax optimal rate for learning with $\mathcal{M}$. In what

follows, we show that on this example the Allocation-Estimation Coefficient correctly adapts to the complexity of this model class. We emphasize that the main applications of our results, which take advantage of the more general $\mathsf{AE}^2_\star$ algorithm, will be given in Section 10.3.1 and Section 10.3.2, and we also present additional examples of uniformly regular classes in Section H.4.2.1.

Example 10.2.1 (Searching for an Informative Arm (revisited))**.** Let $\mathcal{M}$ denote the model class constructed in Example 10.1.1, with parameters $A, N \geq 5$ and $\beta \in [4/A, 9/10]$. We additionally discretize the space so that, for each $M \in \mathcal{M}$ and $\pi \in [A]$, we have $f^M(\pi) \in \{0, \Delta_{\min}, 2\Delta_{\min}, \dots, \lfloor \frac{1}{\Delta_{\min}} \rfloor \Delta_{\min}\}$,[5] and furthermore restrict $\mathcal{M}$ so that it does not include instances with multiple optimal arms. For this class, one can show that Assumptions 10.4 to 10.7 hold with $L_{\mathsf{KL}}, V_{\mathcal{M}} \leq O(\log A)$ and $d_{\mathsf{cov}} = O(A), C_{\mathsf{cov}} = O(N)$ (see Appendix H.4.4). Furthermore, we can bound $\mathsf{n}^{\mathcal{M}}_\varepsilon \leq \frac{2}{\Delta^2_{\min}}$, and

$$\mathsf{aec}_\varepsilon(\mathcal{M}) \leq \frac{64N}{\beta^2} + \frac{16A}{\Delta^2_{\min}}.$$

As a result, for this class, AE^2 has expected regret bounded as

$$\mathbb{E}^{M^\star}[\mathbf{Reg}(T)] \leq (1+\varepsilon)\mathsf{g}^\star \cdot \log(T) + N \cdot \mathrm{poly}(A, \tfrac{1}{\varepsilon}, \tfrac{1}{\Delta_{\min}}, \log N, \log\log T) \cdot \log^{1/2}(T).$$

Note that here the only term that scales linearly with N is the Allocation-Estimation Coefficient—every other class-dependent term appearing in the regret bound scales at most logarithmically in N. We are particularly interested in situations where the cost of finding the correct informative arm is much larger than any existing complexity measures for the problem: that is, when β is constant and $N \gg A, \frac{1}{\Delta_{\min}}$. In this case, we have $\mathsf{aec}_\varepsilon(\mathcal{M}) \leq O(N)$, and the Allocation-Estimation Coefficient correctly captures the intuitive complexity of learning the optimal allocation. In particular, the dependence on N reflects the fact that we need to test each informative arm at least once. Furthermore, as we show in Example 10.4.1, we can lower bound $\mathsf{aec}_\varepsilon(\mathcal{M}) \geq \Omega(N)$ as well, so in the regime where $N \gg A, \frac{1}{\Delta_{\min}}$, the AEC is the dominant lower-order term.

See Appendix H.4.2 for the proof of this example.

10.2.4 The $\mathsf{AE}^2_\star$ Algorithm

While it may be reasonable to assume that the minimum gap of $M^\star$, $\Delta^\star_{\min}$, is bounded away from 0, and that the amount of useful information playing $\pi_\star$ provides is also bounded

[5] The discretization is required to satisfy the uniform regularity assumption—we include it here to provide a concrete example of a uniformly regular class. However, it can be shown that, without this discretization assumption, Theorem 10.2 applies to the original formulation given in Example 10.1.1 with the AEC again scaling as $O(N)$.

Algorithm 10.2 Adaptive Exploration for Allocation Estimation for classes without uniform regularity ($\mathsf{AE}^2_\star$)

input: Optimality tolerance ε, estimation oracle $\mathbf{Alg}_{\mathsf{KL}}$, growth parameters α_q, α_n, $\alpha_{\mathcal{M}} \geq 0$.
$s \leftarrow 1$, $\ell \leftarrow 1$, $\delta \leftarrow \frac{\varepsilon}{4+2\varepsilon}$, $q^s \leftarrow 1 - s^{-\alpha_q}$, $\mathsf{n}^s \leftarrow s^{\alpha_{\mathsf{n}}}$.
Compute $\xi^1 \leftarrow \mathbf{Alg}_{\mathsf{KL}}(\{\varnothing\})$ and $\widehat{M}^1 \leftarrow \mathbb{E}_{M\sim\xi^1}[M]$.
for $t = 1, 2, 3, \ldots$ **do**
  if $s \geq 2^\ell$ **then** // Form active set and cover
    $\ell \leftarrow \ell + 1$.
    $\Delta^\ell \leftarrow \arg\min_{\Delta \geq 0} \Delta \quad \text{s.t.} \quad \mathsf{aec}^{\mathcal{M}}_{\delta/2}(\mathcal{M}_{\Delta,\frac{1}{\Delta}}) \leq s^{\alpha_{\mathcal{M}}}$.
    $\mathcal{M}^\ell \leftarrow \mathcal{M}_{\Delta^\ell,\frac{1}{\Delta^\ell}} \cap \left\{M \in \mathcal{M} \ : \ \mathsf{n}^M_\varepsilon + \frac{1}{\Delta^M_{\min}} + \frac{4\mathsf{g}^M}{\Delta^M_{\min}} + \frac{2\mathsf{n}^M_\varepsilon}{\mathsf{g}^M} + \frac{4}{\Delta^M_{\min}\delta} \leq \sqrt{\mathsf{n}^s}\right\}$.
    $\mathcal{M}^\ell_{\mathrm{cov}} \leftarrow (\rho_\ell, \mu_\ell)$-cover of $\mathcal{M}^\ell$ for $\rho_\ell \leftarrow 2^{-\ell}, \mu_\ell \leftarrow 2^{-5\ell}$, $\mathfrak{D}^\ell \leftarrow \varnothing$.
  if $\exists \pi_{\widehat{M}^s} \in \boldsymbol{\pi}_{\widehat{M}^s}$ s.t. $\forall M \in \mathcal{M}^{\mathrm{alt}}(\pi_{\widehat{M}^s})$, $\sum_{i=1}^{s-1} \mathbb{E}_{\widehat{M}\sim\xi^i}\left[\log \frac{\mathbb{P}^{\widehat{M},\pi^i}(r^i,o^i)}{\mathbb{P}^{M,\pi^i}(r^i,o^i)}\right] \geq \log(t\log t)$ **then**
    Play $\pi_{\widehat{M}^s}$. // Exploit
  else // Explore
    Set $p^s \leftarrow q^s\lambda^s + (1-q^s)\omega^s$ for

$$\lambda^s, \omega^s \leftarrow \arg\min_{\lambda,\omega\in\triangle_\Pi} \sup_{M\in\mathcal{M}^\ell\setminus\mathcal{M}^{\mathrm{gl}}_\varepsilon(\lambda;\mathsf{n}^s)} \frac{1}{\mathbb{E}_{\widehat{M}\sim\xi^s}[\mathbb{E}_{\pi\sim\omega}[D_{\mathsf{KL}}(\widehat{M}(\pi) \,\|\, M(\pi))]]}. \tag{10.2.6}$$

    Draw $\pi^s \sim p^s$, observe r^s, o^s, set $\mathfrak{D}^\ell \leftarrow \mathfrak{D}^\ell \cup \{(\pi^s, r^s, o^s)\}$.
    Compute estimate $\xi^{s+1} \leftarrow \mathbf{Alg}_{\mathsf{KL}}(\mathfrak{D}^\ell, \mathcal{M}^\ell_{\mathrm{cov}})$ and $\widehat{M}^{s+1} = \mathbb{E}_{M\sim\xi^{s+1}}[M]$.
    $s \leftarrow s + 1$.

on $M^\star$, assuming that this is true for every model in the model class (as in the prequel) is a significantly stronger assumption. For example, if we let $\mathcal{M}$ denote the space of all multi-armed bandits with means in $[0,1]$, the only possible value of $\Delta_{\min}$ is 0, as we can always find some instance with minimum gap arbitrarily close to 0. In this section, we dispense with the uniform regularity assumption: we relax Assumption 10.7, and additionally prove that it suffices if only $\mathsf{n}^\star_\varepsilon := \mathsf{n}^{M^\star}_\varepsilon$ (as opposed to $\mathsf{n}^{\mathcal{M}}_\varepsilon$) is bounded.

Our main algorithm for this section, $\mathsf{AE}^2_\star$, is given in Algorithm 10.2. It is very similar to AE^2 but to remove the requirement of uniform regularity, the algorithm avoids solving (10.2.2) over the entire model class $\mathcal{M}$, and instead solves it over a carefully restricted model class. For $x, y > 0$, define

$$\mathcal{M}_{x,y} := \{M \in \mathcal{M} \ : \ \Delta^M_{\min} \geq x, \mathsf{n}^M_\varepsilon \leq y\}. \tag{10.2.7}$$

$\mathsf{AE}^2_\star$ breaks its explore rounds into doubling epochs. For each epoch ℓ, (10.2.6) in Algorithm 10.2 solves an AEC-like optimization problem over a restricted class $\mathcal{M}^\ell \subseteq \mathcal{M}_{\Delta^\ell,\frac{1}{\Delta^\ell}}$,

which is chosen in Line 9 to explicitly ensure that the value of the optimization problem in (10.2.6), is bounded; this is guaranteed by the definition of Δ^ℓ in Line 8. Similar to AE^2, the value of the optimization in (10.2.6) quantifies how much information we are gaining about the Graves-Lai allocation of $M^\star$, and the regret of the explore phase can be bounded in terms of the value of this optimization. By restricting $\mathcal{M}^\ell$ so that the value of (10.2.6) is always bounded, we can therefore ensure that the regret during the exploration phase is bounded.

Intuitively, this restriction of $\mathcal{M}^\ell$ reduces the space of models we must distinguish $M^\star$ from in order to identify its Graves-Lai allocation: rather than distinguishing $M^\star$ from all models in $\mathcal{M}$, we must only distinguish it from models in $\mathcal{M}^\ell$, which could be significantly easier. The caveat is that, since we do not know the value of $\mathsf{n}^\star_\varepsilon$ or $\Delta^\star_{\min}$, $M^\star$ may not always be in $\mathcal{M}^\ell$. In such cases, little can be said about the exploration phase—we are not able to provide any meaningful guarantees on how much information λ^s and ω^s acquire about $M^\star$. To mitigate this, as s increases we gradually relax the criteria for inclusion in $\mathcal{M}^\ell$, ensuring that for large enough s, $M^\star$ will be in $\mathcal{M}^\ell$. In particular, one can show that the number of exploration rounds needed to guarantee $M^\star \in \mathcal{M}^\ell$ scales with $\mathsf{aec}^{\mathcal{M}}_\varepsilon(\mathcal{M}^\star)$, for

$$\mathcal{M}^\star := \left\{M \in \mathcal{M} \;:\; \Delta^M_{\min} \geq \Delta_\star, \mathsf{n}^M_{\varepsilon/36} \leq 1/\Delta_\star\right\} \quad \text{for} \quad \Delta_\star := \min\{\Delta^\star_{\min}, 1/\mathsf{n}^\star_{\varepsilon/36}\}. \tag{10.2.8}$$

That is, $\mathcal{M}^\star$ is the restriction of $\mathcal{M}$ to models with gap at least $\min\{\Delta^\star_{\min}, 1/\mathsf{n}^\star_{\varepsilon/36}\}$ (implying all models in $M^\star$ have a unique optimal decision), and for which the information content of the optimal decision is at most $\max\{1/\Delta^\star_{\min}, \mathsf{n}^\star_{\varepsilon/36}\}$.

Estimation oracle. While AE^2 simply requires that the estimation oracle $\mathbf{Alg}_{\mathsf{KL}}$ returns randomized estimators supported on $\triangle_{\mathcal{M}}$, for $\mathsf{AE}^2_\star$, we wish to ensure that the estimators produced are instead only supported on $\mathcal{M}^\ell$. To this end, we restrict the estimator to $\mathcal{M}^\ell_{\mathrm{cov}}$, the (ρ_ℓ, μ_ℓ)-cover of $\mathcal{M}^\ell$. We denote the resulting estimation oracle with $\mathbf{Alg}_{\mathsf{KL}}(\mathfrak{D}^\ell, \mathcal{M}^\ell_{\mathrm{cov}})$, where the first argument represents the set of available observations, and the second argument the set over which the estimation oracle must return an estimate.

Computational efficiency of $\mathsf{AE}^2_\star$. Similar to AE^2, it is not clear how to solve the main optimization required by $\mathsf{AE}^2_\star$, (10.2.6), in general. In addition, unlike AE^2, $\mathsf{AE}^2_\star$ maintains a version space of models, $\mathcal{M}^\ell$, which could increase the computational burden further. We emphasize that the focus of this work is primarily statistical, and leave addressing the computational challenges for future work.

10.2.5 $\mathsf{AE}^2_\star$ Algorithm: Regret Bound without Uniform Regularity

The following theorem provides the main guarantee for $\mathsf{AE}^2_\star$.

Theorem 10.2 (Regret Bound for $\mathsf{AE}^2_\star$)**.** *For any $\varepsilon \in (0, 1/2]$, if Assumptions 10.4 to 10.6 hold and $\mathsf{g}^\star > 0$, $\mathsf{AE}^2_\star$ (Algorithm 10.2) ensures that for all $T \in \mathbb{N}$, the expected regret is*

bounded as

$$\mathbb{E}^{M^\star}[\mathbf{Reg}(T)] \le (1+\varepsilon)\mathsf{g}^\star \cdot \log(T) + \left(\overline{\mathsf{aec}}^{\mathcal{M}}_{\varepsilon/12}(\mathcal{M}^\star)\right)^3 \cdot C_{\mathsf{aec}} \cdot \log^{3/2}(\log T) + C_{\mathsf{low}} \cdot \log^{6/7}(T),$$

where

$$C_{\mathsf{aec}} := \widetilde{O}\left(\frac{V_{\mathcal{M}}^3(V_{\mathcal{M}} + L_{\mathsf{KL}}) \cdot d_{\mathsf{cov}} \log(C_{\mathsf{cov}})}{\varepsilon \Delta^\star_{\min}}\right),$$

and C_{low} is a lower-order constant given by

$$C_{\mathsf{low}} := \mathrm{poly}\left(\mathsf{g}^\star, \tfrac{1}{\Delta^\star_{\min}}, \mathsf{n}^\star_{\varepsilon/6}, \tfrac{1}{\varepsilon}, V_{\mathcal{M}}, L_{\mathsf{KL}}, d_{\mathsf{cov}}, \log C_{\mathsf{cov}}, \log\log T\right).$$

The proof of Theorem 10.2 is given in Appendix H.3.2. As Theorem 10.2 illustrates, at the expense of a slightly larger polynomial dependence on the Allocation-Estimation Coefficient, and slightly larger lower-order terms, we can obtain near instance-optimal regret—matching the instance-optimal lower bound given in Proposition 10.1 up to a factor of $(1+\varepsilon)$—without requiring any assumption on the minimum gap, or boundedness of $\mathsf{n}^{\mathcal{M}}_\varepsilon$. Rather than scaling with the minimum gap for the entire class, $\Delta_{\min}$, Theorem 10.2 scales only with the minimum gap of the ground truth model, $\Delta^\star_{\min}$, which could be substantially larger than $\Delta_{\min}$. An additional advantage of Theorem 10.2 is that it scales with $\mathsf{aec}^{\mathcal{M}}_\varepsilon(\mathcal{M}^\star)$ as opposed to $\mathsf{aec}_\varepsilon(\mathcal{M})$; for the examples we consider in the sequel, the former quantity enjoys better dependence on problem-dependent parameters. For example, we show in Section 10.4 that for standard classes, $\mathsf{aec}_\varepsilon(\mathcal{M})$ can scale with the minimum gap amongst all models in the class. On the other hand $\mathsf{aec}^{\mathcal{M}}_\varepsilon(\mathcal{M}^\star)$ typically scales with $\Delta^\star_{\min}$.

We emphasize that $\mathsf{AE}^2_\star$ *requires no prior knowledge of* $\Delta^\star_{\min}$ *or* $\mathsf{n}^\star_\varepsilon$—it is able to adapt to the minimum gap and regularity of the underlying model.

Remark 10.2.3 (Dependence on $\mathsf{n}^\star_\varepsilon$)**.** As we show in the following examples, $\mathsf{n}^\star_\varepsilon$ is typically bounded polynomially in standard problem parameters, though in practice this needs to be verified for each problem instance. We remark that some scaling in terms of $\mathsf{n}^\star_\varepsilon$ seems unavoidable—if there is a significant amount of information to be gained playing the optimal decision, any algorithm which is nearly instance-optimal will play the optimal decision at least $\mathsf{n}^\star_\varepsilon$ times, and therefore the "effective horizon" to eliminate alternate instances scales with $\mathsf{n}^\star_\varepsilon$. As we are the first to formalize this notion of how informative the optimal decision is, we believe more research in this direction is required.

10.2.6 Overview of Analysis

To close this section we briefly sketch the proof of the regret bound for AE^2 (Theorem 10.1); the proof of the regret bound for $\mathsf{AE}^2_\star$ (Theorem 10.2) builds on these ideas, but is slightly more involved. See Appendix H.3 for full proofs.

Let us refer to the *exploit phase* as the subset of rounds t in which Line 6 of AE^2 is reached, and refer to the *explore phase* as the subset of rounds in which Line 8 is reached. We focus on bounding the regret in the explore phase—it can be shown (Lemma H.3.1) that in the exploit phase, where the if statement on Line 6 is true, $\pi_{\widehat{M}^s} = \pi_\star$ for all but $O(\log\log T)$ rounds, so that the regret incurred in this phase is at most $O(\log\log T)$.

Let s_T denote the total number of rounds in the explore phase up to time T. Fix an explore round $s \in [s_T]$. We bound the regret $\Delta^\star(p^s)$ by considering three cases.

Case 1: $M^\star \in \mathcal{M}\backslash\mathcal{M}^{\mathrm{gl}}_{\varepsilon/6}(\lambda^s; \mathsf{n}_{\max})$**.** In this case, λ^s is not an optimal (normalized) allocation for $M^\star$, but we can use the AEC to argue that the information gained by the algorithm is large. In particular, since p^s plays ω^s with probability at least $1-q$, we can bound

$$\begin{aligned}
&\frac{1}{\mathbb{E}_{\widehat{M}\sim\xi^s}[\mathbb{E}_{\pi\sim p^s}[D_{\mathsf{KL}}\big(\widehat{M}(\pi)\,\|\,M^\star(\pi)\big)]]} \\
&\le \frac{1}{1-q}\cdot\frac{1}{\mathbb{E}_{\widehat{M}\sim\xi^s}[\mathbb{E}_{\pi\sim\omega^s}[D_{\mathsf{KL}}\big(\widehat{M}(\pi)\,\|\,M^\star(\pi)\big)]]} \\
&\overset{(a)}{\le} \frac{1}{1-q}\cdot\min_{\lambda,\omega\in\triangle_\Pi}\sup_{M\in\mathcal{M}\backslash\mathcal{M}^{\mathrm{gl}}_{\varepsilon/6}(\lambda;\mathsf{n}_{\max})}\frac{1}{\mathbb{E}_{\widehat{M}\sim\xi^s}[\mathbb{E}_{\pi\sim\omega^s}[D_{\mathsf{KL}}\big(\widehat{M}(\pi)\,\|\,M(\pi)\big)]]} \\
&\lesssim \frac{1}{1-q}\cdot\overline{\mathsf{aec}}_{\varepsilon/12}(\mathcal{M})
\end{aligned}$$

as long as $\mathsf{n}_{\max}$ is chosen appropriately; here, (a) follows because $M^\star \in \mathcal{M}\backslash\mathcal{M}^{\mathrm{gl}}_{\varepsilon}(\lambda^s;\mathsf{n}_{\max})$ by assumption in this case, and by the choice of λ^s and ω^s given in (10.2.2). Rearranging this gives

$$1 \lesssim \frac{1}{1-q}\cdot\overline{\mathsf{aec}}_{\varepsilon/12}(\mathcal{M})\cdot\mathbb{E}_{\widehat{M}\sim\xi^s}[\mathbb{E}_{\pi\sim p^s}[D_{\mathsf{KL}}\big(\widehat{M}(\pi)\,\|\,M^\star(\pi)\big)]].$$

This reflects that, when $M^\star \in \mathcal{M}\backslash\mathcal{M}^{\mathrm{gl}}_{\varepsilon/6}(\lambda^s;\mathsf{n}_{\max})$, our choice of p^s ensures that $\widehat{M}\sim\xi^s$ and $M^\star$ can be distinguished, with the amount of information gained lower bounded by $O(\overline{\mathsf{aec}}_{\varepsilon/12}(\mathcal{M})^{-1})$

Adding and subtracting $\frac{2}{1-q}\cdot\overline{\mathsf{aec}}_{\varepsilon/12}(\mathcal{M})\cdot\mathbb{E}_{\widehat{M}\sim\xi^s}[\mathbb{E}_{\pi\sim p^s}[D_{\mathsf{KL}}\big(\widehat{M}(\pi)\,\|\,M^\star(\pi)\big)]]$ to $\Delta^\star(p^s)$, and using that $\Delta^\star(p^s)\le 1$ always, we then have that the instantaneous regret in this case is bounded by

$$\begin{aligned}
\Delta^\star(p^s) &= \Delta^\star(p^s) - \frac{2}{1-q}\cdot\overline{\mathsf{aec}}_{\varepsilon/12}(\mathcal{M})\cdot\mathbb{E}_{\widehat{M}\sim\xi^s}[\mathbb{E}_{\pi\sim p^s}[D_{\mathsf{KL}}\big(\widehat{M}(\pi)\,\|\,M^\star(\pi)\big)]] \\
&\quad + \frac{2}{1-q}\cdot\overline{\mathsf{aec}}_{\varepsilon/12}(\mathcal{M})\cdot\mathbb{E}_{\widehat{M}\sim\xi^s}[\mathbb{E}_{\pi\sim p^s}[D_{\mathsf{KL}}\big(\widehat{M}(\pi)\,\|\,M^\star(\pi)\big)]] \\
&\le -1 + \frac{2}{1-q}\cdot\overline{\mathsf{aec}}_{\varepsilon/12}(\mathcal{M})\cdot\mathbb{E}_{\widehat{M}\sim\xi^s}[\mathbb{E}_{\pi\sim p^s}[D_{\mathsf{KL}}\big(\widehat{M}(\pi)\,\|\,M^\star(\pi)\big)]].
\end{aligned}$$

Summing over s, it follows that the total regret in this case is bounded by

$$\sum_{s=1}^{s_T} \Delta^\star(p^s) \cdot \mathbb{I}\{s \text{ in Case 1}\} \le \frac{2}{1-q} \cdot \overline{\mathsf{aec}}_{\varepsilon/12}(\mathcal{M}) \cdot \mathbf{Est}_{\mathsf{KL}}(s_T) - \sum_{s=1}^{s_T} \mathbb{I}\{s \text{ in Case 1}\}.$$

Furthermore, since regret is always non-negative—that is, $\Delta^\star(p) \ge 0$ for all p—rearranging this inequality leads to a bound on the total number of times this case can occur:

$$\sum_{s=1}^{s_T} \mathbb{I}\{s \text{ in Case 1}\} \le \frac{2}{1-q} \cdot \overline{\mathsf{aec}}_{\varepsilon/12}(\mathcal{M}) \cdot \mathbf{Est}_{\mathsf{KL}}(s_T).$$

Critically, as given in (10.2.1), $\mathbf{Est}_{\mathsf{KL}}(s_T)$ scales at most poly-logarithmically in s_T. Thus, as long as s_T is at most $O(\log T)$, the total regret incurred in this case (as well as the total number of times this case can occur), will be at most $O(\overline{\mathsf{aec}}_{\varepsilon/12}(\mathcal{M}) \cdot \log\log T)$.

Case 2: $M^\star \in \mathcal{M}^{\mathrm{gl}}_{\varepsilon/6}(\lambda^s; \mathsf{n}_{\max})$ and $\pi_\star \in \boldsymbol{\pi}_{\widehat{M}^s}$. In this case, we have that λ^s is a Graves-Lai optimal allocation for $M^\star$. Thus, it follows that

$$\Delta^\star(\lambda^s) \le (1+\varepsilon/6)\mathsf{g}^\star/\mathsf{n}^\star \quad \text{and} \quad \inf_{M \in \mathcal{M}^{\mathrm{alt}}(M^\star)} \mathbb{E}_{\pi\sim\lambda^s}[D_{\mathsf{KL}}(M^\star(\pi) \,\|\, M(\pi))] \ge (1-\varepsilon/6)/\mathsf{n}^\star$$

for some $\mathsf{n}^\star \le \mathsf{n}_{\max}$. This implies that, for any $M \in \mathcal{M}^{\mathrm{alt}}(M^\star)$, we can bound

$$\begin{aligned}
\Delta^\star(p^s) &= \Delta^\star(p^s) - (1+\varepsilon)\mathsf{g}^\star \mathbb{E}_{\pi\sim p^s}[D_{\mathsf{KL}}(M^\star(\pi) \,\|\, M(\pi))] + (1+\varepsilon)\mathsf{g}^\star \mathbb{E}_{\pi\sim p^s}[D_{\mathsf{KL}}(M^\star(\pi) \,\|\, M(\pi))] \\
&\lesssim (1+\varepsilon/6)\mathsf{g}^\star/\mathsf{n}^\star - (1+\varepsilon)(1-\varepsilon/6)\mathsf{g}^\star/\mathsf{n}^\star + (1+\varepsilon)\mathsf{g}^\star \mathbb{E}_{\pi\sim p^s}[D_{\mathsf{KL}}(M^\star(\pi) \,\|\, M(\pi))] \\
&\lesssim -\varepsilon \mathsf{g}^\star/\mathsf{n}^\star + (1+\varepsilon)\mathsf{g}^\star \mathbb{E}_{\pi\sim p^s}[D_{\mathsf{KL}}(M^\star(\pi) \,\|\, M(\pi))] \\
&\le -\varepsilon \mathsf{g}^\star/\mathsf{n}_{\max} + (1+\varepsilon)\mathsf{g}^\star \mathbb{E}_{\pi\sim p^s}[D_{\mathsf{KL}}(M^\star(\pi) \,\|\, M(\pi))]
\end{aligned}$$

Since this bound holds uniformly for all $M \in \mathcal{M}^{\mathrm{alt}}(M^\star)$, it follows that the total regret in this case can be bounded as

$$\begin{aligned}
\sum_{s=1}^{s_T} \Delta^\star(p^s) \cdot \mathbb{I}\{s \text{ in Case 2}\} \lesssim (1+\varepsilon)\mathsf{g}^\star \cdot \sum_{s=1}^{s_T} \inf_{M \in \mathcal{M}^{\mathrm{alt}}(M^\star)} \mathbb{E}_{\pi\sim p^s}[D_{\mathsf{KL}}(M^\star(\pi) \,\|\, M(\pi))] \cdot \mathbb{I}\{\pi_\star \in \boldsymbol{\pi}_{\widehat{M}^s}\} \\
- \frac{\varepsilon \mathsf{g}^\star}{\mathsf{n}_{\max}} \sum_{s=1}^{s_T} \mathbb{I}\{s \text{ in Case 2}\}.
\end{aligned}$$

To bound this, the key observation is that, if we explore at round s, then it must be the case that, for all $\pi_{\widehat{M}^s} \in \boldsymbol{\pi}_{\widehat{M}^s}$, there exists some $M \in \mathcal{M}^{\mathrm{alt}}(\pi_{\widehat{M}^s})$ such that $\sum_{i=1}^{s-1} \mathbb{E}_{\widehat{M}\sim\xi^i}[\log \frac{\mathbb{P}^{\widehat{M},\pi^i}(r^i,o^i)}{\mathbb{P}^{M,\pi^i}(r^i,o^i)}] \le \log(T\log T)$. Using Assumption 10.4 and Assumption 10.5 to move from $\widehat{M} \sim \xi^i$ to $M^\star$ and to relate the observed log-likelihood ratios to the KL divergence, we can furthermore show

that

$$\inf_{M\in\mathcal{M}^{\mathrm{alt}}(M^\star)}\sum_{s=1}^{s_T}\mathbb{E}_{\pi\sim p^s}[D_{\mathsf{KL}}(M^\star(\pi)\,\|\,M(\pi))]\cdot\mathbb{I}\{\pi_\star\in\boldsymbol{\pi}_{\widehat{M}^s}\}$$
$$\lesssim\inf_{M\in\mathcal{M}^{\mathrm{alt}}(M^\star)}\sum_{s=1}^{s_T}\mathbb{E}_{\widehat{M}\sim\xi^s}\Big[\log\frac{\mathbb{P}^{\widehat{M},\pi^s}(r^s,o^s)}{\mathbb{P}^{M,\pi^s}(r^s,o^s)}\Big]+\sqrt{s_T}\cdot\mathbf{Est}_{\mathsf{KL}}(s_T)\lesssim\log(T\log T)+\sqrt{s_T}\cdot\mathbf{Est}_{\mathsf{KL}}(s_T).$$

This allows us to bound

$$\begin{aligned}(1+\varepsilon)\mathsf{g}^\star\cdot\sum_{s=1}^{s_T}&\inf_{M\in\mathcal{M}^{\mathrm{alt}}(M^\star)}\mathbb{E}_{\pi\sim p^s}[D_{\mathsf{KL}}(M^\star(\pi)\,\|\,M(\pi))]\cdot\mathbb{I}\{\pi_\star\in\boldsymbol{\pi}_{\widehat{M}^s}\}\\&\lesssim(1+\varepsilon)\mathsf{g}^\star\cdot\log T+\mathsf{g}^\star\sqrt{s_T}\cdot\mathbf{Est}_{\mathsf{KL}}(s_T).\end{aligned}$$

Thus, as long as $s_T=O(\log T)$, we can bound the total regret incurred in Case by $(1+\varepsilon)\mathsf{g}^\star\cdot\log T+o(\log T)$. Using that regret is always lower bounded by 0 in the same fashion as Case 1, we can further use this to bound the total number of times that Case 2 occurs by

$$\sum_{s=1}^{s_T}\mathbb{I}\{s\text{ in Case 2}\}\le\frac{\mathsf{n}_{\max}}{\varepsilon\mathsf{g}^\star}\Big(\mathsf{g}^\star\cdot\log T+\mathsf{g}^\star\sqrt{s_T}\cdot\mathbf{Est}_{\mathsf{KL}}(s_T)\Big).$$

The intuition for this case is that, since we are playing a Graves-Lai allocation for $M^\star$, the regret will scale with $\mathsf{g}^\star$, the instance-optimal rate, and, furthermore, the allocation will allow us to distinguish $M^\star$ from alternatives $M\in\mathcal{M}^{\mathrm{alt}}(M^\star)$. Using that the total estimation error is bounded, and that we only enter the explore phase if there exists some $M\in\mathcal{M}^{\mathrm{alt}}(M^\star)$ that we cannot distinguish from $M^\star$, this ultimately implies that the total number of times this phase occurs, and therefore the total regret incurred by this phase, is bounded.

Case 3: $M^\star\in\mathcal{M}^{\mathrm{gl}}_{\varepsilon/6}(\lambda^s;\mathsf{n}_{\max})$ and $\pi_\star\notin\boldsymbol{\pi}_{\widehat{M}^s}$. In this case, we bound $\Delta^\star(p^s)$ by adding and subtracting $\mathbb{E}_{\widehat{M}\sim\xi^s}[\mathbb{E}_{\pi\sim p^s}[D_{\mathsf{KL}}\big(\widehat{M}(\pi)\,\|\,M^\star(\pi)\big)]]$ in the same fashion as Case 1. Since $\pi_\star\notin\boldsymbol{\pi}_{\widehat{M}^s}$, it can be shown that ξ^s must place $\Omega(\Delta_{\min})$ probability mass on $M\in\mathcal{M}^{\mathrm{alt}}(M^\star)$, allowing us to lower bound $\mathbb{E}_{\widehat{M}\sim\xi^s}[\mathbb{E}_{\pi\sim p^s}[D_{\mathsf{KL}}\big(\widehat{M}(\pi)\,\|\,M^\star(\pi)\big)]]$ using the same reasoning as in Case 2. In total, we can show that the regret in this case is bounded by $O(\frac{\mathsf{g}^\star}{\Delta_{\min}}\cdot\mathbf{Est}_{\mathsf{KL}}(s_T))$, and the number of times this case can occur is at most $O(\frac{\mathsf{n}_{\max}}{\varepsilon\Delta_{\min}}\cdot\mathbf{Est}_{\mathsf{KL}}(s_T))$.

Concluding the Proof. Combining all three cases, we have shown that the regret of the explore phase is bounded by

$$(1+\varepsilon)\mathsf{g}^\star\cdot\log T+\widetilde{O}\bigg(\frac{1}{1-q}\cdot\overline{\mathsf{aec}}_{\varepsilon/12}(\mathcal{M})\cdot\mathbf{Est}_{\mathsf{KL}}(s_T)+\mathsf{g}^\star\sqrt{s_T}\cdot\mathbf{Est}_{\mathsf{KL}}(s_T)+\frac{\mathsf{g}^\star}{\Delta_{\min}}\cdot\mathbf{Est}_{\mathsf{KL}}(s_T)\bigg).$$

Furthermore, using our bounds on the number of times each case can occur, one can show that $s_T = O(\log T)$. Since $\mathbf{Est}_{\mathsf{KL}}(s_T)$ is at most polylogarithmic in s_T, it follows that the regret is bounded as

$$(1+\varepsilon)\mathsf{g}^\star \cdot \log T + \widetilde{O}\Big(\overline{\mathsf{aec}}_{\varepsilon/12}(\mathcal{M}) + \overline{\mathsf{aec}}^{1/2}_{\varepsilon/12}(\mathcal{M}) \cdot \log^{1/2} T\Big),$$

as stated in Theorem 10.1.

10.3 Examples: Non-Asymptotic Instance-Optimal Regret

We now instantiate Theorem 10.2 to give regret bounds for $\mathsf{AE}^2_\star$ in standard settings of interest, bounding the Allocation-Estimation Coefficient for each setting. We begin by focusing on structured bandit settings and contextual bandits (Section 10.3.1), then turn to tabular reinforcement learning in the sequel (Section 10.3.2).

10.3.1 Application: Structured and Contextual Bandits

We recall that, to map bandit problems to the DMSO framework, we take the decision space Π to be the set of "arms", the observation space $\mathcal{O} = \{\varnothing\}$, and the reward space $\mathcal{R}$ to be the rewards from the bandit (while we do not explicitly include the rewards in the observation space, we assume they are observed). We defer proofs for all examples to Appendix H.4.

10.3.1.1 The Uniform Exploration Coefficient

For the main examples we consider, we proceed by first bounding the Allocation-Estimation Coefficient in terms of another, somewhat simpler parameter we refer to as the *uniform exploration coefficient*.

Definition 10.3.1 (Uniform Exploration Coefficient)**.** For a randomized estimator $\xi \in \triangle_{\mathcal{M}}$, we define the uniform exploration coefficient with respect to ξ at scale $\varepsilon > 0$ as the value of the following program:

$$C^{\xi}_{\mathrm{exp}}(\varepsilon) := \min_{C \in \mathbb{R}_+, p \in \triangle_\Pi} \Big\{ C \;\Big|\; \forall M, M' \in \mathcal{M} \;:\; \begin{array}{r} \max_{M'' \in \{M, M'\}} \mathbb{E}_{\overline{M} \sim \xi}[\mathbb{E}_{\pi \sim p}[D_{\mathsf{KL}}\big(\overline{M}(\pi) \,\|\, M''(\pi)\big)]] \le \frac{1}{C} \\ \implies \max_{p' \in \triangle_\Pi} \mathbb{E}_{\pi \sim p'}[D_{\mathsf{KL}}(M(\pi) \,\|\, M'(\pi))] \le \varepsilon \end{array} \Big\}.$$

We define $p^{\xi}_{\mathrm{exp}}(\varepsilon)$ as any minimizing distribution for this program, and let

$$C_{\mathrm{exp}}(\mathcal{M}, \varepsilon) := \sup_{\xi \in \triangle_{\mathcal{M}}} C^{\xi}_{\mathrm{exp}}(\varepsilon)$$

denote the uniform exploration constant for class $\mathcal{M}$.

Intuitively, the uniform exploration coefficient characterizes the extent to which it is possible to explore by uniformly covering the decision space. In particular, one can always

choose p to be uniform over Π, which gives $C_{\mathrm{exp}}(\mathcal{M}, \varepsilon) \lesssim |\Pi|/\varepsilon$, but in cases where information is shared between actions, the parameter is significantly smaller, as we will show for familiar examples below. For example, in the case of linear bandits with dimension d, we have $C_{\mathrm{exp}}(\mathcal{M}, \varepsilon) \leq \widetilde{O}(\frac{d \cdot \log 1/\varepsilon}{\varepsilon})$.

The following result shows that the Allocation-Estimation Coefficient can be bounded in terms of the uniform exploration coefficient.

Proposition 10.2 (Informal)**.** *For $\mathcal{M}_0 \subseteq \mathcal{M}$, we can bound $\overline{\mathsf{aec}}^{\mathcal{M}}_{\varepsilon}(\mathcal{M}_0) \leq C_{\mathrm{exp}}(\mathcal{M}_0, \delta)$ for any*

$$\sqrt{\delta} \leq \min_{M \in \mathcal{M}_0} \min \left\{ \min \left\{ \frac{1}{81 L_{\mathsf{KL}}}, \frac{\Delta^M_{\mathrm{min}}}{34 V_{\mathcal{M}}} \right\} \cdot \frac{\varepsilon}{2\mathsf{g}^M / \Delta^M_{\mathrm{min}} + \mathsf{n}^M_{\varepsilon/36}}, \frac{\Delta^M_{\mathrm{min}}}{3} \right\}.$$

The full statement of Proposition 10.2 is given in Lemma H.2.6. Using Proposition 10.2, we obtain guarantees for $\mathsf{AE}^2_\star$ on several familiar classes, beginning with several bandit settings. We remark that Proposition 10.2 is not in general tight—it simply shows that the Allocation-Estimation Coefficient is bounded by a simple, general, and interpretable notion of how easily a class can be explored. As, such the bounds on the Allocation-Estimation Coefficient in the following examples can almost certainly be improved using more specialized tools.

10.3.1.2 Finite-Armed Bandits

We first consider the simplest bandit setting: multi-armed bandits with finite arms.

Example 10.3.1 (Finite-Armed Bandit)**.** Fix $A > 0$, and consider the class of finite-armed bandits with A arms and unit-variance Gaussian noise:

$$\mathcal{M} = \left\{ M(\pi) = \mathcal{N}(f^M(\pi), 1) \mid f^M \in [0,1]^A \right\}.$$

It is straightforward to verify that Assumptions 10.4 to 10.6 hold with $L_{\mathsf{KL}}, V_{\mathcal{M}} \leq 4$ and $d_{\mathsf{cov}} = O(A), C_{\mathsf{cov}} = O(1)$, and it can also be shown that, as long as $f^{M^\star}(\pi_\star) < 1$, we can bound $\mathsf{n}^\star_\varepsilon \leq c \cdot \frac{A^2}{\varepsilon(\Delta^\star_{\mathrm{min}})^4}$. In addition, we can bound $C_{\mathrm{exp}}(\mathcal{M}^\star, \varepsilon) \leq 4A/\varepsilon$, so Proposition 10.2 gives the following result.

Proposition 10.3. *For the finite-armed bandit problem with A actions, there exists a universal constant $c > 0$ such that*

$$\overline{\mathsf{aec}}^{\mathcal{M}}_{\varepsilon}(\mathcal{M}^\star) \leq c \cdot \frac{A^{15}}{\varepsilon^8 (\Delta^\star_{\mathrm{min}})^{24}}. \tag{10.3.1}$$

We immediately obtain the following corollary to Theorem 10.2.

Corollary 10.3. *For finite-armed bandits with Gaussian noise, as long as* $f^{M^\star}(\pi_\star) < 1$, $\mathsf{AE}^2_\star$ *has regret bounded by*

$$\mathbb{E}^{M^\star}[\mathbf{Reg}(T)] \le (1+\varepsilon)\mathsf{g}^\star \cdot \log(T) + \mathrm{poly}(A, \tfrac{1}{\varepsilon}, \tfrac{1}{\Delta^\star_{\min}}, \log\log T) \cdot \log^{6/7}(T).$$

With more refined analyses, various works have achieved instance-optimal regret bounds for finite-armed bandits with tighter lower-order terms than Corollary 10.3 [Garivier et al., 2016, Kaufmann et al., 2016, Lattimore, 2018, Garivier et al., 2019]. We emphasize that Corollary 10.3 is a special case of a much more general result. In particular, we proved the bound on the Allocation-Estimation Coefficient, (10.3.1), using tools which hold for general classes (e.g. Proposition 10.2). An analysis of $\mathsf{AE}^2_\star$ specialized to finite-armed bandits would likely yield a tighter result.

10.3.1.3 Structured Bandits

Many bandit problems exhibit richer structure than the multi-armed bandit setting, and the study of these settings has been the focus of much of the recent work on instance-optimal learning. We next consider one such setting, that of structured bandits with bounded *eluder dimension* [Russo and Van Roy, 2013].

Example 10.3.2 (Structured Bandits with Bounded Eluder Dimension)**.** Consider a bandit problem with unit-variance Gaussian noise but where the means are now given by a *general function class* $\mathcal{F}$ mapping from Π to $[0,1]$:

$$\mathcal{M} = \{M(\pi) = \mathcal{N}(f(\pi), 1) \mid f \in \mathcal{F}\}. \tag{10.3.2}$$

For such general settings, we might hope to capture the complexity of learning in terms of generalized notions of dimension for $\mathcal{F}$. We consider one such notion here: the eluder dimension.

Definition 10.3.2 (Eluder Dimension [Russo and Van Roy, 2013])**.** Let $\check{d}_{\mathrm{E}}(\mathcal{F}, \varepsilon')$ denote the length of the longest sequence of actions $\{\pi_1, \dots, \pi_d\}$ such that, for each $n \le d$, there exist functions $f, f' \in \mathcal{F}$ with $\sqrt{\sum_{i=1}^{n-1}(f(\pi_i) - f'(\pi_i))^2} \le \varepsilon'$ but $f(\pi_n) - f'(\pi_n) > \varepsilon'$. The *eluder dimension* of function class $\mathcal{F}$ at scale ε is then defined as $d_{\mathrm{E}}(\mathcal{F}, \varepsilon) = \sup_{\varepsilon' \ge \varepsilon} \check{d}_{\mathrm{E}}(\mathcal{F}, \varepsilon') \vee 1$.

The eluder dimension can be thought of as quantifying how easily a function class can be "explored": evaluating a pair of functions on $d_{\mathrm{E}}(\mathcal{F}, \varepsilon)$ points allows one to determine whether they are nearly identical over the entire space. It is known to be bounded for many standard classes—for example, for linear function classes with dimension d, $d_{\mathrm{E}}(\mathcal{F}, \varepsilon) \le \widetilde{O}(d \cdot \log 1/\varepsilon)$—and is also closely related to the *disagreement coefficient* [Foster et al., 2020b]. Furthermore, it can be shown to be a sufficient condition for bounded (worst-case) regret in general bandit problems [Russo and Van Roy, 2013]. The following result shows that the eluder dimension bounds the Allocation-Estimation Coefficient.

Proposition 10.4. *For the structured bandit class $\mathcal{M}$ considered in (10.3.2), we have*

$$C_{\mathrm{exp}}(\mathcal{M}^\star, \delta) \le \frac{16 d_{\mathrm{E}}(\mathcal{F}, \sqrt{\delta}/2)}{\delta}.$$

This implies that

$$\overline{\mathsf{aec}}_\varepsilon^{\mathcal{M}}(\mathcal{M}^\star) \le \frac{16 d_{\mathrm{E}}(\mathcal{F}, \sqrt{\delta}/2)}{\delta} \text{ for scale } \delta = c \cdot \frac{\varepsilon^2 \Delta_\star^8}{d_{\mathrm{E}}(\mathcal{F}, \frac{1}{2}\Delta_\star)^2} \text{ and } \Delta_\star := \min\{\Delta^\star_{\min}, 1/\mathsf{n}^\star_{\varepsilon/36}\},$$

where $c > 0$ is a universal constant.

Proposition 10.4 highlights the ability of the Allocation-Estimation Coefficient to adapt to the inherent complexity of "exploring" for the model class under consideration. We henceforth abbreviate $d_{\mathrm{E}} := d_{\mathrm{E}}(\mathcal{F}, \sqrt{\delta}/2)$.

It is straightforward to show that Assumptions 10.4 and 10.5 are met in this setting with $L_{\mathsf{KL}}, V_{\mathcal{M}} \le 4$ (see Appendix H.4.2). Furthermore, Assumption 10.6 can be shown to hold with d_{cov} scaling with the covering number of $\mathcal{F}$ in the distance $d(f, f') = \sup_{\pi \in \Pi} |f(\pi) - f'(\pi)|$ and $C_{\mathsf{cov}} = O(1)$. In general, it must be shown that $\mathsf{n}^\star_\varepsilon$ is bounded for each $M^\star$ and class of interest; as we show in the following examples, it is bounded for standard structured bandit settings such as linear bandits. We have the following corollary to Theorem 10.2.

Corollary 10.4. *In the structured bandit setting with bounded eluder dimension considered above, $\mathsf{AE}^2_\star$ has regret bounded as*

$$\mathbb{E}^{M^\star}[\mathbf{Reg}(T)] \le (1+\varepsilon)\mathsf{g}^\star \cdot \log(T) + \mathrm{poly}(d_{\mathrm{E}}, d_{\mathsf{cov}}, \tfrac{1}{\varepsilon}, \tfrac{1}{\Delta^\star_{\min}}, \mathsf{n}^\star_{\varepsilon/36}, \log\log T) \cdot \log^{6/7}(T).$$

To the best of our knowledge, Corollary 10.4 is the first result to show that it is possible to obtain the instance-optimal rate in classes with bounded eluder dimension, with lower-order terms scaling only polynomially in the eluder dimension. More generally, Corollary 10.4 illustrates that $\mathsf{AE}^2_\star$ can adapt to the structural properties of the given model class, and achieve regret scaling with existing notions of intrinsic dimension.

We next consider two examples of structured bandits where it is known that the eluder dimension is bounded: linear bandits and generalized linear models. While these results are immediate given Corollary 10.4, the additional structure present in these settings allows us to obtain somewhat more explicit results.

Example 10.3.3 (Linear Bandits)**.** Consider the class of linear bandits with unit-variance Gaussian noise defined as

$$\mathcal{M} = \{M(\pi) = \mathcal{N}(\langle \theta, x_\pi \rangle, 1) \mid \theta \in \Theta\},$$

where $\Theta \subseteq \mathbb{R}^d$ is some convex set with ℓ_2 diameter $O(1)$ and $\mathcal{X} := \{x_\pi : \pi \in \Pi\} \subseteq \mathbb{R}^d$ is the arm set, which we assume has $\|x_\pi\|_2 \le 1$ for all $\pi \in \Pi$. As in Example 10.3.2, Assumptions 10.4

and 10.5 are met in this setting with $L_{\mathsf{KL}}, V_{\mathcal{M}} \leq 4$; furthermore, Assumption 10.6 is also met with $d_{\mathsf{cov}} = O(d)$ and $C_{\mathsf{cov}} = O(1)$. We then have the following bound on the AEC.

Proposition 10.5. *For the linear bandit class $\mathcal{M}$ defined above, we have*

$$\overline{\mathsf{aec}}_{\varepsilon}^{\mathcal{M}}(\mathcal{M}^{\star}) \leq c \cdot \frac{d^3}{\varepsilon^2} \cdot \left(\frac{1}{\Delta_{\min}^{\star}} + \mathsf{n}_{\varepsilon/36}^{\star} \right)^8 \cdot \mathrm{polylog}\left(d, \tfrac{1}{\varepsilon}, \tfrac{1}{\Delta_{\min}^{\star}}, \mathsf{n}_{\varepsilon/36}^{\star} \right)$$

for a universal constant $c > 0$.

Using Proposition 10.5, we obtain the following corollary to Theorem 10.2.

Corollary 10.5. *In the linear bandit setting defined above, $\mathsf{AE}_{\star}^2$ has regret bounded as*

$$\mathbb{E}^{M^{\star}}[\mathbf{Reg}(T)] \leq (1+\varepsilon)\mathsf{g}^{\star} \cdot \log(T) + \mathrm{poly}(d, \tfrac{1}{\varepsilon}, \tfrac{1}{\Delta_{\min}^{\star}}, \mathsf{n}_{\varepsilon/36}^{\star}, \log\log T) \cdot \log^{6/7}(T).$$

Corollary 10.6 has lower-order terms scaling similarly to the best known lower-order terms in the linear bandit setting [Tirinzoni et al., 2020, Kirschner et al., 2021]. These works, however, develop algorithms which are specialized to the linear bandit setting, while Corollary 10.6 is the instantiation of a more general result designed for arbitrary general decision making settings.

The following result shows that under general conditions, we can bound the parameter $\mathsf{n}_{\varepsilon}^{\star}$ for linear bandits.

Proposition 10.6 (Informal)**.** *For linear bandits satisfying certain regularity conditions, $\mathsf{n}_{\varepsilon}^{\star}$ is bounded by a polynomial function of problem parameters and a geometry-dependent term scaling with the structure of $\mathcal{X}$ and Θ.*

The full statement of Proposition 10.6 is given in Proposition H.4. The regularity condition for Proposition 10.6 requires primarily that $\theta^{\star}$ lies sufficiently far within the interior of Θ (see Section H.4.2.4 for further details). We remark that the guarantees given in both Tirinzoni et al. [2020] and Kirschner et al. [2021] scale with geometric parameters very similar to $\mathsf{n}_{\varepsilon}^{\star}$.

Example 10.3.4 (Generalized Linear Models)**.** In the generalized linear model setting, we take the model class to be

$$\mathcal{M} = \{M(\pi) = \mathcal{N}(g(\langle \theta, x_{\pi} \rangle), 1) \mid \theta \in \Theta\},$$

where Θ and $\mathcal{X}$ are as in Example 10.3.3, and $g(\cdot)$ is a known link function which is increasing and Lipschitz, but potentially nonlinear. Let $g_{\max}$ and $g_{\min}$ denote upper and lower bounds on the derivative of g, respectively:

$$g_{\max} := \max_{\theta \in \Theta, x \in \mathrm{co}(\mathcal{X})} g'(\langle \theta, x \rangle) \quad \text{and} \quad g_{\min} := \min_{\theta \in \Theta, x \in \mathrm{co}(\mathcal{X})} g'(\langle \theta, x \rangle).$$

As in the linear bandit setting, we can show that Assumptions 10.4 and 10.5 are both met with $L_{\mathsf{KL}}, V_{\mathcal{M}} \leq 4$, and that Assumption 10.6 is also met with $d_{\mathsf{cov}} = O(d)$ and $C_{\mathsf{cov}} = O(g_{\max})$. Furthermore, under the same conditions as for linear bandits, $\mathsf{n}^\star_\varepsilon$ can be bounded for generalized linear models exactly as for linear bandits, but with an additional scaling of $(\frac{g_{\max}}{g_{\min}})^2$. We then have the following.

Proposition 10.7. *For the generalized linear model class $\mathcal{M}$ defined above, we have*

$$\overline{\mathsf{aec}}^{\mathcal{M}}_\varepsilon(\mathcal{M}^\star) \leq c \cdot \frac{d^3 g^3_{\max}}{\varepsilon^2 g^3_{\min}} \cdot \left(\frac{1}{\Delta^\star_{\min}} + \mathsf{n}^\star_{\varepsilon/36}\right)^8 \cdot \mathrm{polylog}\left(d, \tfrac{1}{\varepsilon}, \tfrac{1}{\Delta^\star_{\min}}, \mathsf{n}^\star_{\varepsilon/36}\right)$$

for a universal constant $c > 0$.

Using Proposition 10.5, we obtain the following corollary to Theorem 10.2.

Corollary 10.6. *In the generalized linear model setting defined above, $\mathsf{AE}^2_\star$ has regret bounded as*

$$\mathbb{E}^{M^\star}[\mathbf{Reg}(T)] \leq (1+\varepsilon)\mathsf{g}^\star \cdot \log(T) + \mathrm{poly}(d, \tfrac{g_{\max}}{g_{\min}}, \tfrac{1}{\varepsilon}, \tfrac{1}{\Delta^\star_{\min}}, \mathsf{n}^\star_{\varepsilon/36}, \log\log T) \cdot \log^{6/7}(T).$$

To the best of our knowledge, this is the first result to obtain finite-time instance-optimality for generalized linear models with lower-order terms polynomial in problem parameters.

10.3.1.4 Contextual Bandits

The previous examples illustrate that $\mathsf{AE}^2_\star$ is able to learn efficiently in a variety of structured bandit settings. We now show that it leads to new guarantees for finite-action contextual bandits with general function approximation.

Example 10.3.5 (Contextual Bandits with Finitely Many Arms)**.** Consider the contextual bandit setting with context set $\mathcal{X}$ (which could be arbitrarily large) and action set $\mathcal{A}$ such that $A := |\mathcal{A}| < \infty$. Let $p_{\mathcal{X}}$ denote the context distribution, which we assume is known to the learner. The learning protocol is then, for step $t = 1, 2, 3, \ldots$:

1. Environment samples context $x^t \sim p_{\mathcal{X}}$.

2. Learner chooses action $a^t \in \mathcal{A}$, receives reward r_t.

We assume that $r^t = f^\star(x^t, a^t) + w^t$ for $w^t \sim \mathcal{N}(0,1)$, for some $f^\star : \mathcal{X} \times \mathcal{A} \to [0,1]$. We assume as well that the learner is given access to a set of functions $\mathcal{F}$ such that $f^\star \in \mathcal{F}$.

To view this setting as a special case of the DMSO framework, we take the decision space to be the set $\Pi = (\mathcal{X} \to \mathcal{A})$ of all policies mapping from $\mathcal{X}$ to $\mathcal{A}$, and take $\mathcal{O} = \mathcal{X}$ as the observation space. The learner's decision at round t is a policy π^t, and they receive a

reward-observation pair $(r^t, o^t) = (r^t, x^t)$ under the process $x^t \sim p_{\mathcal{X}}$, $r^t \sim \mathcal{N}(f^\star(x^t, \pi^t(x^t)), 1)$. The model class $\mathcal{M}$ is the set of all instances of this form for $f^\star \in \mathcal{F}$.

The following result shows that the Allocation-Estimation Coefficient is be bounded by the number of actions A, and is *independent* of the size of the context space. See Appendix H.4.3 for a proof.

Proposition 10.8. *For the contextual bandit setting, we can bound*

$$C_{\mathrm{exp}}(\mathcal{M}^\star, \varepsilon) \leq \frac{4A}{\varepsilon},$$

which implies that

$$\overline{\mathsf{aec}}^{\mathcal{M}}_{\varepsilon}(\mathcal{M}^\star) \leq c \cdot \frac{A^3}{\varepsilon^2} \cdot \left(\frac{1}{\Delta^\star_{\mathrm{min}}} + \mathsf{n}^\star_{\varepsilon/36} \right)^8,$$

for a universal constant $c > 0$.

As in the cases of bandits with bounded eluder dimension, $\mathsf{n}^\star_\varepsilon$ must be bounded for each $M^\star$ and class $\mathcal{F}$ of interest. It is straightforward to show, however, that Assumptions 10.4 and 10.5 are met in this setting with $L_{\mathsf{KL}}, V_{\mathcal{M}} \leq 4$, and, furthermore, that Assumption 10.6 is also met with d_{cov} scaling as the covering number of $\mathcal{F}$ in the distance $d(f, f') = \sup_{x \in \mathcal{X}, a \in \mathcal{A}} |f(x, a) - f'(x, a)|$, and $C_{\mathsf{cov}} = O(1)$. We then have the following corollary.

Corollary 10.7. *In the finit-action contextual bandit setting considered above, $\mathsf{AE}^2_\star$ has regret bounded as*

$$\mathbb{E}^{M^\star}[\mathbf{Reg}(T)] \leq (1 + \varepsilon)\mathsf{g}^\star \cdot \log(T) + \mathrm{poly}(A, d_{\mathsf{cov}}, \tfrac{1}{\varepsilon}, \tfrac{1}{\Delta^\star_{\mathrm{min}}}, \mathsf{n}^\star_{\varepsilon/36}, \log\log T) \cdot \log^{6/7}(T).$$

To the best of our knowledge, Corollary 10.7 is the first instance-optimal guarantee in the contextual bandit setting with general function approximation that obtains lower-order term scaling polynomially in problem parameters. Notably, the lower-order term scales independently of the size of the context space, $|\mathcal{X}|$. We anticipate that extending this result to contextual bandit settings that have large action spaces, but which exhibit additional structure allowing for efficient exploration (e.g., linearity), will be straightforward.

10.3.2 Application: Tabular Reinforcement Learning

As a final application of our results, we turn to the setting of episodic tabular reinforcement learning.

Episodic Markov decision processes. Recall that episodic reinforcement learning is a special case of the DMSO framework in which each model $M \in \mathcal{M}$ is an episodic Markov Decision Process (MDP) given by the tuple $M = (\mathcal{S}, \mathcal{A}, H, \{P^M_h\}_{h=1}^H, \{R^M_h\}_{h=1}^H, s_1)$. Here $\mathcal{S}$

is a set of states, $\mathcal{A}$ a set of actions, H the horizon, $P_h^M : \mathcal{S} \times \mathcal{A} \to \triangle_{\mathcal{S}}$ the probability transition kernel at step h, $R_h^M : \mathcal{S} \times \mathcal{A} \to \triangle_{\mathbb{R}}$ the reward distribution at step h, and s_1 a deterministic initial state, which we take to be fixed across models. We assume that $R_h^M(s,a)$ is unit-variance Gaussian, and that $\mathbb{E}_{r_h \sim R_h^M(s,a)}[r_h] \in [0, 1/H]$.[6]

The decision space Π consists of non-stationary policies $\pi = (\pi_1, \dots, \pi_H)$, where $\pi_h : \mathcal{S} \to \mathcal{A}$. For a fixed policy π, an episode proceeds in an MDP M proceeds as follows. First, beginning from the initial state s_1, we take action $a_1 \sim \pi_1(s_1)$, receive reward $r_1 \sim R_1^M(s_1, a_1)$, and transition to $s_2 \sim P_1^M(\cdot \mid s_1, a_1)$. This continues for H steps at which point the episode terminates and the process repeats. We define $f^M(\pi) := \mathbb{E}^{M,\pi}\left[\sum_{h=1}^H r_h\right]$ as the expected reward achieved over the entire episode under this process.

Each round $t \in [T]$ in the DMSO framework corresponds to an episode in the underlying MDP $M^\star$. At each round, the learner selects a policy π^t, and receives reward $r^t = \sum_{h=1}^H r_h^t$ and $o^t = (s_1^t, a_1^t, r_1^t, \dots, s_H^t, a_H^t, r_H^t)$, where $(s_1^t, a_1^t, r_1^t, \dots, s_H^t, a_H^t, r_H^t)$ is the trajectory that results from executing π^t in $M^\star$ for a single episode.

Tabular model class. In the tabular RL setting, it is assumed that $S := |\mathcal{S}|$ and $A := |\mathcal{A}|$ are both finite, and we take Π to be the set of all deterministic policies. In addition to assuming that $\mathcal{M}$ consists of tabular MDPs, we restrict to the following subclass:

$$\mathcal{M}_{\mathrm{tab}}(P_{\min}) := \left\{ M = (\mathcal{S}, \mathcal{A}, H, \{P_h^M\}_{h=1}^H, \{R_h^M\}_{h=1}^H, s_1) \ : \ \min_{s,a,s',h} P_h^M(s' \mid s, a) \geq P_{\min} \right\}. \tag{10.3.3}$$

While the assumption that $\min_{s,a,s',h} P_h^M(s' \mid s, a) \geq P_{\min}$ may be seen as restrictive, the guarantees we provide scale only with $\log \frac{1}{P_{\min}}$, so $P_{\min}$ can be taken to be extremely small without affecting the result significantly.

Note that when our results are specialized to reinforcement learning, $\Delta_{\min}^\star$ denotes the gap between the performance of the optimal policy, and the next-best *deterministic* policy. This quantity can be lower bounded in terms of other standard quantities including gaps in the rewards at each state and the transition probabilities.

Toward instantiating Theorem 10.2 in this tabular RL setting, we first provide a bound on the Allocation-Estimation Coefficient, which we establish by first bounding the Uniform Exploration Coefficient.

---

[6] This assumption only serves to ensure that $f^M(\pi) \in [0, 1]$, in line with the convention for the rest of the paper. Our results continue to hold up to $\mathrm{poly}(H)$ factors if $\mathbb{E}_{r_h \sim R_h^M(s,a)}[r_h] \in [0, 1]$.

Proposition 10.9. *For $\mathcal{M} \leftarrow \mathcal{M}_{\mathrm{tab}}(P_{\min})$, we can bound*[7]

$$C^{\mathsf{H}}_{\mathrm{exp}}(\mathcal{M}^\star, \varepsilon) \leq c \cdot \frac{SAH^2 \cdot \log^2 H}{\varepsilon^2}$$

for a universal constant $c > 0$, which implies that

$$\overline{\mathsf{aec}}^{\mathcal{M}}_{\varepsilon}(\mathcal{M}^\star) \leq c \cdot \frac{S^5A^5H^{14} \cdot \log^{10} H}{\varepsilon^4} \cdot \left(\frac{1}{\Delta^\star_{\min}} + \mathsf{n}^\star_{\varepsilon/36}\right)^{24} \cdot \log^4 \frac{1}{P_{\min}}$$

for a universal constant $c > 0$.

Next, it can be shown that Assumptions 10.4 to 10.6 hold for $\mathcal{M} \leftarrow \mathcal{M}_{\mathrm{tab}}(P_{\min})$ with constants (see Appendix H.4.5):

$$L_{\mathsf{KL}} = V_{\mathcal{M}} = O(H \cdot \log 1/P_{\min}), \quad d_{\mathsf{cov}} = O(S^2AH), \quad C_{\mathsf{cov}} = O(H/P_{\min}).$$

We then obtain the following corollary to Theorem 10.2.

Corollary 10.8. *For $\mathcal{M} \leftarrow \mathcal{M}_{\mathrm{tab}}(P_{\min})$ and ε, $\mathsf{AE}^2_\star$ has regret bounded by*

$$\mathbb{E}^{M^\star}[\mathbf{Reg}(T)] \leq (1+\varepsilon)\mathsf{g}^\star \cdot \log(T) + \mathrm{poly}(S, A, H, \tfrac{1}{\varepsilon}, \tfrac{1}{\Delta^\star_{\min}}, \log \tfrac{1}{P_{\min}}, \mathsf{n}^\star_{\varepsilon/36}, \log\log(T)) \cdot \log^{6/7}(T).$$

To our knowledge, Corollary 10.8 is the first guarantee for tabular RL that achieves the instance-optimal rate while obtaining lower-order terms that scale only polynomially in problem parameters. As noted previously, existing approaches to instance-optimal regret in tabular RL [Ok et al., 2018, Dong and Ma, 2022] have lower-order terms that scale exponentially in problem parameters, and as a result are truly asymptotic in nature.

While Corollary 10.8 is stated in terms of $\mathsf{n}^\star_\varepsilon$ for the sake of generality, as we show in Appendix H.4.5, if $M^\star$ has rewards that are sufficiently small (in particular, if it satisfies $\mathbb{E}_{r_h \sim R_h^{M^\star}(s,a)}[r_h] < 1/H^2$ for all (s, a, h)), then we can bound

$$\mathsf{n}^\star_\varepsilon \leq c \cdot \frac{\mathsf{g}^\star}{\Delta^\star_{\min}} \cdot \left(1 + \frac{\mathsf{g}^\star}{\varepsilon(\Delta^\star_{\min})^2}\right).$$

In this case, Corollary 10.8 scales polynomially in all standard problem parameters.

Let us remark that the prior work of Dong and Ma [2022] does not require that $P_h^M(s' \mid s, a) \geq P_{\min}$ as we do (the work of Ok et al. [2018] only holds for ergodic MDPs, itself a very strong assumption). However, the lower-order term obtained in Dong and Ma [2022] scales polynomially in the inverse probability of observing the trajectory that occurs with

[7] Here $C^{\mathsf{H}}_{\mathrm{exp}}$ denote the uniform exploration coefficient as defined in Definition 10.3.1, but with $D_{\mathsf{KL}}(\cdot \,\|\, \cdot)$ replaced with $D^2_{\mathsf{H}}(\cdot, \cdot)$. To prove Proposition 10.9, we show that a variant of Proposition 10.2 still holds with this alternate definition of $C_{\mathrm{exp}}(\mathcal{M}, \varepsilon)$.

minimum (non-zero) probability. In general, this will scale exponentially in H, and inversely with the probability of the transition with minimum (non-zero) probability occurring, that is $\min_{s,a,s',h:P_h^M(s'|s,a)>0} P_h^M(s'|s,a)$. Thus, while we must impose the stronger condition that all transitions occur with some probability $P_{\min}$, our bounds only scale logarithmically in this quantity, and polynomially in S, A, and H, a significant improvement over Dong and Ma [2022]. Understanding whether it is possible to remove the additional restrictions we impose while still obtaining reasonable finite-time performance is an interesting direction for future work.

As far as we are aware, there is no prior work on instance-optimal algorithms for RL settings with general function approximation. While we have only instantiated Theorem 10.2 and $\mathsf{AE}^2_\star$ in the tabular RL setting, the tools we have developed can also be applied to RL with general model classes. Exploring the application of $\mathsf{AE}^2_\star$ to, for example, bilinear classes [Du et al., 2021] is an exciting avenue for future work.

10.4 Lower Bounds for Learning the Optimal Exploration Allocation

The AE^2 algorithm achieves instance-optimal regret by explicitly learning an ε-optimal Graves-Lai allocation for the underlying model $M^\star$. In this section, we introduce an abstract formulation for the problem of learning an optimal allocation (Section 10.4.1), and provide lower bounds which show that the Allocation-Estimation Coefficient is a fundamental limit for this task (Section 10.4.2). We then present several examples illustrating lower bounds on the Allocation-Estimation Coefficient (Section 10.4.3), and discuss how our lower bounds relate to the problem of minimizing regret (Section 10.4.4).

Additional Notation. Throughout this section we will also make use of the following definition:

$$\Upsilon(M;\varepsilon,\mathsf{n}_{\max}) := \Big\{\lambda \in \triangle_\Pi \;:\; \exists \mathsf{n} \in (0,\mathsf{n}_{\max}] \text{ s.t. } \Delta^M(\lambda) \le \frac{(1+\varepsilon)\mathsf{g}^M}{\mathsf{n}}, \inf_{M' \in \mathcal{M}^{\mathrm{alt}}(M)} \mathbb{E}_{\pi\sim\lambda}[D_{\mathsf{KL}}(M(\pi) \,\|\, M'(\pi))] \ge \frac{1-\varepsilon}{\mathsf{n}}\Big\}. \tag{10.4.1}$$

That is, $\Upsilon(M;\varepsilon,\mathsf{n}_{\max})$ denotes the set of normalized allocations that are Graves-Lai optimal for M with tolerance ε, and have normalization factor at most $\mathsf{n}_{\max}$. Unless otherwise stated, the results in this section do not make use of Assumption 10.3 or Assumption 10.7.

10.4.1 Learning the Optimal Allocation: Minimax Formulation

We consider the following protocol, which captures the task of learning an optimal Graves-Lai allocation for an unknown model $M^\star$.

- For $t = 1, \dots, T$, sample $\pi^t \sim p^t$ and observe (r^t, o^t).
- Based on the entire history $\mathcal{H}^T = (\pi^1, r^1, o^1), \dots, (\pi^T, r^T, o^T)$, output a normalized allocation $\widehat{\lambda} \in \triangle_\Pi$. The allocation may be randomized according to a distribution $q \in \triangle_{\triangle_\Pi}$.

We formalize an *algorithm* for this task as a pair $\mathbb{A} = (p, q)$, where $q(\cdot \mid \mathcal{H}^T)$ is the distribution over $\widehat{\lambda}$ given the history, and $p = \{p^t\}_{t \in [T]}$ is a sequence of exploration distributions of the form $p^t(\cdot \mid \mathcal{H}^{t-1})$. We let $\mathbb{P}^{M, \mathbb{A}}(\cdot)$ denote the law of $\mathcal{H}^T$ when M is the underlying model and $\mathbb{A}$ is the algorithm, and let $\mathbb{E}^{M, \mathbb{A}}[\cdot]$ denote the corresponding expectation. The goal of the algorithm is to ensure that $\widehat{\lambda}$ is an ε-optimal allocation for $M^\star$ with high probability, i.e.

$$\mathbb{P}^{M^\star, \mathbb{A}}\Big(\widehat{\lambda} \in \Upsilon(M^\star; \varepsilon, \mathsf{n}_{\max})\Big) \geq 1 - \delta$$

for some failure probability $\delta > 0$ and normalization factor $\mathsf{n}_{\max} > 0$

Intuitively, learning an optimal Graves-Lai allocation is closely related to achieving instance-optimal regret, but there are some subtle technical differences which we discuss in detail in the sequel. We study the former task because we find it to be more amenable to non-asymptotic lower bounds, and because it captures the behavior of "natural" algorithms such as AE^2 and essentially every existing asymptotically optimal algorithm we are aware of.

Minimax framework. To provide lower bounds on the complexity of learning an optimal Graves-Lai allocation, we consider a minimax framework. Our main quantity of interest will be:

$$T^{\mathrm{gl}}(\mathcal{M}; \varepsilon, \mathsf{n}_{\max}, \delta) = \inf_{\mathbb{A}} \inf\Big\{T \in \mathbb{N} \mid \mathbb{P}^{M, \mathbb{A}}\Big(\widehat{\lambda} \in \Upsilon(M; \varepsilon, \mathsf{n}_{\max})\Big) \geq 1 - \delta,\ \forall M \in \mathcal{M}\Big\}. \tag{10.4.2}$$

This represents the earliest time $T \in \mathbb{N}$ for which there exists an algorithm that learns an ε-optimal allocation with probability at least $1 - \delta$, and does so uniformly for all $M \in \mathcal{M}$. Recall that for our upper bounds (Theorem 10.1), the Allocation-Estimation Coefficient gives a bound on the lower-order terms in the regret (reflecting the time required to learn an ε-optimal allocation) that holds *uniformly* for all models in the class $\mathcal{M}$. To understand the optimality of uniform bounds of this type, a minimax framework is natural. This framework also naturally complements recent non-asymptotic algorithms for linear models such as [Tirinzoni et al., 2020, Kirschner et al., 2021], where the complexity of exploration is captured by problem-dependent quantities such as the feature dimension, which are bounded uniformly for all models in the class. Nonetheless, exploring other notions of optimality (for example, instance-dependent complexity) for learning the Graves-Lai allocation is an interesting direction for future research.

Note that the quantity (10.4.2) does not place any constraint on the regret of the

algorithm under consideration. It will also be useful to consider the notion

$$T^{\mathrm{gl}}(\mathcal{M};\varepsilon,\mathsf{n}_{\max},\delta,R)=\inf_{\mathbb{A}}\inf\Big\{T\in\mathbb{N}\mid\mathbb{P}^{M,\mathbb{A}}\Big(\widehat{\lambda}\in\Upsilon(M;\varepsilon,\mathsf{n}_{\max})\Big)\geq 1-\delta,\\ \mathbb{E}^{M,\mathbb{A}}[\mathbf{Reg}(T)]\leq R\cdot\log(T),\ \ \forall M\in\mathcal{M}\Big\},\quad(10.4.3)$$

which captures the minimax complexity of learning the Graves-Lai allocation, subject to the constraint that the algorithm achieves logarithmic regret throughout the learning process. This notion is particularly well suited to complement the upper bounds achieved by algorithms such as AE^2.

We remark that the restriction to allocations with normalization factor no more than $\mathsf{n}_{\max}$ in the definitions above is natural for several reasons. First, without such a restriction, the returned allocation can place an arbitrarily small amount of mass on informative actions, and an arbitrarily large amount of mass on the optimal action, since any Graves-Lai optimal allocation is still Graves-Lai optimal if the amount of mass on the optimal action is increased arbitrarily. While technically Graves-Lai optimal, such allocations do not reflect an allocation one could play over a finite time horizon in order to certify the optimal decision $\pi_\star$—as any algorithm with finite-time guarantees must do—and therefore do not reflect the cost such an algorithm must pay to learn an allocation. Second, given knowledge of the optimal action, any Graves-Lai optimal allocation which has a normalization factor larger than $\mathsf{n}_{\max}$ can be transformed into a Graves-Lai optimal allocation with normalization factor $\mathsf{n}_{\max}$ by adjusting the mass on the optimal action, assuming $\mathsf{n}_{\max}$ is taken to be sufficiently large (in particular, as large as $\mathsf{n}_{\max}(\mathcal{M},\varepsilon)$; cf. (10.2.3)). Therefore, if we restrict our attention to the task to learning the Graves-Lai allocation for a subclass of models which agree on the optimal action (as we do in Theorem 10.10), any lower bound for learning an allocation with normalization $\mathsf{n}_{\max}$ also applies to learning an unrestricted allocation, for large enough $\mathsf{n}_{\max}$. Finally, AE^2 itself plays allocations with bounded normalization, which as we note in Section 10.2.2, does not affect the optimality of its performance—allocations with bounded normalization are always sufficient.

10.4.2 Main Result

We state two lower bounds. The first scales with the version of the Allocation-Estimation Coefficient appearing in our upper bound (Theorem 10.1), but leads to a lower bound on T^{gl}, while the second lower bound scales with the AEC for a restriction $\mathcal{M}$, but is exponentially stronger in the sense that it provides a similar lower bound on $\log(T^{\mathrm{gl}})$. We remark briefly that the regularity conditions of Section 10.2.1 are not required to hold here, unless otherwise stated.

Theorem 10.9 (Main lower bound—weak variant)**.** *Let $\varepsilon>0$, $\mathsf{n}_{\max}>0$, and $\mathcal{M}_0\subseteq\mathcal{M}$ be*

given, and set $\delta := \frac{\varepsilon}{2} \cdot \min\{1, \inf_{M \in \mathcal{M}_0} \mathsf{g}^M / \mathsf{n}_{\max}\}$. *Unless*

$$T > \frac{\delta}{8} \cdot \sup_{\overline{M} \in \mathcal{M}^+} \mathsf{aec}^{\mathcal{M}}_{2\varepsilon}(\mathcal{M}_0, \overline{M}),$$

any algorithm must have, for some $M \in \mathcal{M}_0$*:*

$$\mathbb{P}^{M, \mathbb{A}}\left[\widehat{\lambda} \notin \Upsilon(M; \varepsilon, \mathsf{n}_{\max})\right] \geq \frac{\delta}{6}.$$

Stated equivalently, Theorem 10.9 implies that for any $\varepsilon > 0$, if we set $\delta = c \cdot \varepsilon \cdot \inf_{M \in \mathcal{M}_0} \mathsf{g}^M / \mathsf{n}_{\max}$ for sufficiently small numerical constant c, then

$$T^{\mathrm{gl}}(\mathcal{M}; \varepsilon, \mathsf{n}_{\max}, \delta) \geq \delta \cdot \sup_{\overline{M} \in \mathcal{M}^+} \mathsf{aec}_{2\varepsilon}(\mathcal{M}_0, \overline{M}).$$

Remark 10.4.1 (Choice of $\mathcal{M}_0$)**.** Theorem 10.9 is stated with respect to an arbitrary subset, $\mathcal{M}_0$, of $\mathcal{M}$. While one could simply choose $\mathcal{M}_0 \leftarrow \mathcal{M}$, in some cases it is advantageous to choose $\mathcal{M}_0 \subsetneq \mathcal{M}$. In particular, note that our lower bound scale with $\inf_{M \in \mathcal{M}_0} \mathsf{g}^M$. For classes $\mathcal{M}$ where $\inf_{M \in \mathcal{M}} \mathsf{g}^M = 0$—which will be the case, for example, if $\mathcal{M}$ corresponds to a model class where reward means form a compact set—it is advantageous to restrict $\mathcal{M}_0$ so that it corresponds only to instances with $\mathsf{g}^M > 0$.

We remark as well that the restriction of the lower bound to $\mathcal{M}_0$ is somewhat analogous to our upper bound Theorem 10.2, which provides guarantees in terms of the AEC of a restriction of $\mathcal{M}$, $\mathcal{M}^\star$. We may therefore choose $\mathcal{M}_0 \leftarrow \mathcal{M}^\star$ to obtain lower bounds matching the scaling of Theorem 10.2. In the following section we provide several examples of how $\mathcal{M}_0$ can be chosen to yield intuitive lower bounds.

Lastly, let us mention that $\mathcal{M}_0$ may also be chosen to yield lower bounds that have a more instance-dependent flavor. For example, given some instance $M^\star$, we could choose $\mathcal{M}_0$ to correspond to all instances identical to $M^\star$ up to a permutation of the decisions, in which case Theorem 10.9 is will yield lower bounds on the performance of any algorithm on a permutation of $M^\star$, rather than over the entire class.

Remark 10.4.2. The lower bound in Theorem 10.9, for $\mathcal{M}_0 = \mathcal{M}$, scales with the quantity

$$\sup_{\overline{M} \in \mathcal{M}^+} \mathsf{aec}_\varepsilon(\mathcal{M}, \overline{M}) \geq \sup_{\overline{M} \in \mathrm{co}(\mathcal{M})} \mathsf{aec}_\varepsilon(\mathcal{M}, \overline{M}) \geq \sup_{\xi \in \triangle_{\mathcal{M}}} \overline{\mathsf{aec}}_\varepsilon(\mathcal{M}, \xi) = \overline{\mathsf{aec}}_\varepsilon(\mathcal{M}),$$

which at first glance might appear to be larger than the version of the AEC appearing in our upper bounds. However, there is no contradiction, because these quantities can be shown to be equivalent (up to problem-dependent parameters) under the assumptions with which Theorem 10.1 is proven.

Our second lower bound yields a lower bound on $\log(T^{\mathrm{gl}})$ as opposed to T^{gl}—a significantly

stronger result—but scales with the AEC for a restricted class. To state the result, for $\overline{M} \in \mathcal{M}^+$ and $\mathcal{M}_0 \subseteq \mathcal{M}$, define

$$\mathcal{M}_0^{\mathrm{opt}}(\overline{M}) = \left\{ M \in \mathcal{M}_0 \mid \boldsymbol{\pi}_M \subseteq \boldsymbol{\pi}_{\overline{M}}, \;\; D_{\mathsf{KL}}\big(\overline{M}(\pi) \,\|\, M(\pi)\big) = 0 \;\; \forall \pi \in \boldsymbol{\pi}_{\overline{M}} \right\}.$$

This represents the set of models M where 1) the optimal decisions for M are also optimal for $\overline{M}$ and 2) playing the an optimal decision reveals no information that can distinguish M and $\overline{M}$. Our second lower bound scales with the AEC for $\mathcal{M}_0^{\mathrm{opt}}(\overline{M})$, and is restricted to algorithms with low regret.

Theorem 10.10 (Main lower bound—strong variant)**.** *Let $\varepsilon > 0$, $\mathsf{n}_{\max} > 0$, and $\mathcal{M}_0 \subseteq \mathcal{M}$ be given, and define $\delta = \frac{\varepsilon}{2} \cdot \min\{1, \inf_{M \in \mathcal{M}_0} \mathsf{g}^M / \mathsf{n}_{\max}\}$. Unless*

$$\sup_{M \in \mathcal{M}_0} \frac{\mathsf{g}^M}{\Delta_{\min}^M} \cdot \log(T) \geq \Omega(\delta^2) \cdot \sup_{\overline{M} \in \mathcal{M}^+} \mathsf{aec}_{2\varepsilon}^{\mathcal{M}}(\mathcal{M}_0^{\mathrm{opt}}(\overline{M}), \overline{M}),$$

there is no algorithm that simultaneously ensures that

1. $\mathbb{E}^{M, \mathbb{A}}[\mathbf{Reg}(T)] \leq 2 \cdot \mathsf{g}^M \log(T), \;\; \forall M \in \mathcal{M}_0.$
2. $\mathbb{P}^{M, \mathbb{A}}\left[\widehat{\lambda} \notin \Upsilon(M; \varepsilon, \mathsf{n}_{\max})\right] \leq \frac{\delta}{12}, \;\; \forall M \in \mathcal{M}_0.$

Equivalently, Theorem 10.10 implies that for any $\varepsilon > 0$, if we set $\delta = c \cdot \varepsilon \cdot \inf_{M \in \mathcal{M}_0} \mathsf{g}^M / \mathsf{n}_{\max}$ for a sufficiently small numerical constant c, then for $R := 2 \sup_{M \in \mathcal{M}_0} \mathsf{g}^M$,

$$\log(T^{\mathrm{gl}}(\mathcal{M}; \varepsilon, \mathsf{n}_{\max}, \delta, R)) \geq \frac{\delta^2 \cdot \inf_{M \in \mathcal{M}_0} \Delta_{\min}^M}{R} \cdot \sup_{\overline{M} \in \mathcal{M}^+} \mathsf{aec}_{2\varepsilon}^{\mathcal{M}}(\mathcal{M}_0^{\mathrm{opt}}(\overline{M}), \overline{M}).$$

10.4.3 Examples

As we discuss in the sequel, the dependence on the Allocation-Estimation Coefficient in our lower bounds (particularly Theorem 10.10) qualitatively matches the dependence on the AEC in our main upper bounds, Theorems 10.1 and 10.2. We defer a detailed discussion comparing our upper and lower bounds for a moment, and make matters concrete by considering three examples: the informative arm example from the introduction (Example 10.1.1), multi-armed bandits, and tabular reinforcement learning. We defer proofs for all examples to Appendix H.5.4.

Example 10.4.1 (Searching for an Informative Arm (revisited))**.** Let $\mathcal{M}$ be the class constructed in Example 10.1.1 with parameters $N, A \in \mathbb{N}$ and $\beta \in (0, 1/2)$. Let $\mathcal{M}_0$ denote the restriction of $\mathcal{M}$ to models with $\Delta_{\min}^M \geq \Delta$ for some $\Delta \in (0, 1/6)$ (so that π_M is unique),

and $f^M(\pi_M) < 1$. As long as $N \geq 16$ and $\beta \log(1 + \beta A) \geq 2\Delta/(A-1)$, we have

$$\sup_{\overline{M} \in \mathcal{M}^+} \mathsf{aec}_\varepsilon^{\mathcal{M}}(\mathcal{M}_0^{\mathrm{opt}}(\overline{M}), \overline{M}) \geq \frac{N}{4\beta}. \tag{10.4.4}$$

For this construction, we have $\Delta_{\min}^M \geq \Delta$ and $\Omega(1) \leq \mathsf{g}^M \leq O(\beta^{-1})$ for all $M \in \mathcal{M}_0$. Thus, Theorem 10.10 implies that any algorithm with $\mathbb{E}^{M,\mathbb{A}}[\mathbf{Reg}(T)] \leq 2\mathsf{g}^M \log(T)$ for all $M \in \mathcal{M}_0$ must fail to learn an ε-optimal allocation with probability $\delta = \Omega(\varepsilon/\mathsf{n}_{\max})$ unless

$$\sup_{M \in \mathcal{M}} \frac{\mathsf{g}^M}{\Delta_{\min}^M} \cdot \log(T) \gtrsim \frac{\varepsilon^2}{\mathsf{n}_{\max}^2} \cdot \frac{N}{\beta},$$

which implies that $\log(T^{\mathrm{gl}}(\mathcal{M}_0; \varepsilon, \mathsf{n}_{\max}, \delta, R)) \gtrsim \varepsilon^2/\mathsf{n}_{\max}^2 \cdot N$ for $R = 2\sup_{M \in \mathcal{M}} \mathsf{g}^M = O(\beta^{-1})$. Any Graves-Lai allocation need only take at most $O(\beta^{-1})$ pulls to eliminate alternative instances, so an appropriate choice of $\mathsf{n}_{\max}$ is $O(\beta^{-1})$, yielding $\log(T^{\mathrm{gl}}(\mathcal{M}_0; \varepsilon, \mathsf{n}_{\max}, \delta, R)) \gtrsim \beta^2\varepsilon^2 \cdot N$. While the dependence on the parameter $\mathsf{n}_{\max}, \varepsilon, \Delta > 0$ here is certainly loose, this formalizes the intuition sketched in the introduction: any algorithm that learns an optimal allocation must explore $\Omega(N)$ times, yet any algorithm that achieves near-instance-optimal regret $\mathbb{E}^{M,\mathbb{A}}[\mathbf{Reg}(T)] \leq 2\mathsf{g}^M \log(T) \lesssim \beta^{-1} \log(T)$ can play a sub-optimal decision no more than roughly $\beta^{-1} \log(T)/\Delta$ times, leading to the constraint that

$$\log(T) \gtrsim N.$$

We can also apply Theorem 10.9 to show that any algorithm must fail to learn an ε-optimal allocation with probability $\delta = \Omega(\varepsilon/\mathsf{n}_{\max})$ unless $T \gtrsim \varepsilon/\mathsf{n}_{\max} \cdot \frac{N}{\beta}$.

Example 10.4.2 (Finite-Armed Bandit)**.** Let $A \geq 6$ and $\Delta \in (0, 1/2)$ be given and set $\Pi = [A]$. Let $\mathcal{M}$ be the set of all multi-armed bandit instances with Gaussian noise:

$$\mathcal{M} = \left\{ M(\pi) = \mathcal{N}(f^M(\pi), 1/2) \mid f^M \in [0,1]^A \right\} \tag{10.4.5}$$

and let $\mathcal{M}_0 = \{M \in \mathcal{M} \ : \ |\boldsymbol{\pi}_M| = 1, \Delta^M(\pi) \in [\Delta, 2\Delta] \text{ for } \pi \neq \pi_M, f^M(\pi_M) < 1\}$ denote the set of all bandit instances where suboptimal arms have gaps on order Δ. Then for all $\varepsilon \in (0, 1/32)$,

$$\sup_{\overline{M} \in \mathcal{M}} \mathsf{aec}_\varepsilon^{\mathcal{M}}(\mathcal{M}_0^{\mathrm{opt}}(\overline{M}), \overline{M}) \geq c \cdot \varepsilon^{-2} \cdot \frac{A}{\Delta^2}, \tag{10.4.6}$$

where $c > 0$ is an absolute constant.

It can be shown that, by the construction of $\mathcal{M}_0$, $\mathsf{g}^M = \Theta(A/\Delta)$ for all $M \in \mathcal{M}_0$, so $\delta \propto \varepsilon \cdot \min\{1, \frac{A}{\Delta \mathsf{n}_{\max}}\}$. Theorem 10.9 then implies that any algorithm must fail to learn an

ε-optimal allocation with probability $\delta = \Omega(\varepsilon \cdot \min\{1, \frac{A}{\Delta \mathsf{n}_{\max}}\})$ unless

$$T \gtrsim \min\left\{1, \frac{A}{\Delta \mathsf{n}_{\max}}\right\} \cdot \frac{A}{\varepsilon \Delta^2}.$$

Any Graves-Lai allocation need only take $O(\frac{A}{\Delta^2})$ pulls to eliminate all alternate instances, so a reasonable choice of $\mathsf{n}_{\max}$ is therefore $O(\frac{A}{\Delta^2})$. With this choice of $\mathsf{n}_{\max}$, we have $T^{\mathrm{gl}}(\mathcal{M}; \varepsilon, \mathsf{n}_{\max}, \delta) \gtrsim \frac{A}{\varepsilon \Delta}$. We remark that the scaling on all parameters here is natural. Intuitively, we would expect that we need to pull each arm at least once to learn a near-optimal allocation, yielding an $\Omega(A)$ scaling. In addition, note that for multi-armed bandits, the optimal allocation places mass $\propto \frac{1}{\Delta^2}$ on arms with gap Δ. To correctly estimate this proportion requires an accurate estimate of Δ, which becomes increasingly difficult as Δ becomes smaller, yielding an $\Omega(\frac{1}{\Delta})$ scaling. Finally, as we decrease ε, we require that the returned allocation becomes closer to a truly optimal allocation, and we would therefore expect an $\Omega(\frac{1}{\varepsilon})$ scaling.

Note that the result derived by applying Theorem 10.9 above only gives a lower bound on T. To obtain a lower bound on $\log(T)$, we combine Theorem 10.10 and (10.4.6) with $\mathsf{n}_{\max} = O(\frac{A}{\Delta^2})$ as above, which implies that any algorithm with $\mathbb{E}^{M, \mathbb{A}}[\mathbf{Reg}(T)] \leq 2\mathsf{g}^M \log(T)$ for all $M \in \mathcal{M}_0$ must fail to learn an ε-optimal allocation with probability $\delta = \Omega(\varepsilon \cdot \min\{1, \frac{A}{\Delta \mathsf{n}_{\max}}\})$ unless

$$\sup_{M \in \mathcal{M}_0} \frac{\mathsf{g}^M}{\Delta^M_{\min}} \cdot \log(T) \gtrsim A,$$

or equivalently $\log(T^{\mathrm{gl}}(\mathcal{M}; \varepsilon, \mathsf{n}_{\max}, \delta, R)) \gtrsim \frac{A}{\sup_{M \in \mathcal{M}_0} \mathsf{g}^M / \Delta^M_{\min}}$ for $R = 2 \sup_{M \in \mathcal{M}} \mathsf{g}^M$. To see why such scaling is natural, note that any algorithm which has $\mathbb{E}^{M, \mathbb{A}}[\mathbf{Reg}(T)] \leq 2\mathsf{g}^M \log(T)$ for each instance M can afford to explore (that is, play a suboptimal decision) at most $2\frac{\mathsf{g}^M}{\Delta^M_{\min}} \log(T)$ times, or their regret could exceed $2\mathsf{g}^M \log(T)$. However, no algorithm has any hope of learning an optimal allocation unless they play every arm at least once, so taking at least A pulls of suboptimal arms seems unavoidable, and we therefore would expect that we must have $2\frac{\mathsf{g}^M}{\Delta^M_{\min}} \log(T) \gtrsim A$, which is precisely the necessary scaling shown here.

We make two remarks on this $\log(T)$ lower bound. First, note that due to the presence of the δ^2 term in (10.4.6) (which we believe to be loose), the lower bound we derive by applying Theorem 10.10 does not scale with the parameter ε^{-1}, as one might hope. Second, as noted, for $M \in \mathcal{M}_0$ for $\mathcal{M}_0$ chosen as in Example 10.4.2, we have $\mathsf{g}^M = \Omega(A/\Delta)$, in which cases the dependence on A cancels, and the lower bound becomes trivial. Note that this is somewhat to be expected. Theorem 10.10 will give a trivial lower bound whenever $\sup_{M \in \mathcal{M}_0} \mathsf{g}^M$ is much larger than the AEC. Recall that in our upper bound, Theorem 10.2, the leading order $\mathsf{g}^\star \cdot \log T$ term will dominate the lower-order terms once

$$\mathsf{g}^\star \cdot \log T \gtrsim \Omega(\mathsf{aec}_\varepsilon(\mathcal{M})).$$

Therefore, if $\mathsf{g}^\star$ is much larger than the AEC, Theorem 10.2 simply gives an upper bound scaling as approximately $\mathsf{g}^\star \cdot \log T$, as long as $\log T = \Omega(1)$. The interpretation in this setting is that the complexity of learning the Graves-Lai allocation is dominated by the regret incurred by *playing* the Graves-Lai allocation, which we know is necessary from Proposition 10.1, and therefore we would not expect lower-order terms to be significant components of the regret, as reflected by Theorem 10.2. However, as we have already shown, In settings such as Example 10.4.1 where this is not the case and the AEC is much larger than $\mathsf{g}^\star$, Theorem 10.10 will give a non-trivial lower bound which reflects the difficulty of learning the optimal allocation.

Example 10.4.3 (Tabular Reinforcement Learning)**.** Let $S, A, H \in \mathbb{N}$ and $\Delta \in (0, 1/2)$ be given, and assume that $SA \geq 24$ and $H \geq \log_2(S/2)$. Let $\mathcal{M}$ be the set of all tabular MDPs with 1) $|\mathcal{S}| = S$, $|\mathcal{A}| = A$ and horizon H, and 2) Gaussian rewards with variance $\sigma^2 = 1/2$ (cf. Section 10.3.2). Let $\mathcal{M}_0$ be the result of restricting $\mathcal{M}$ in the same fashion as Example 10.4.2. Then for all $\varepsilon \in (0, 1/32)$,

$$\sup_{\overline{M} \in \mathcal{M}} \mathsf{aec}_\varepsilon^{\mathcal{M}}(\mathcal{M}_0^{\mathrm{opt}}(\overline{M}), \overline{M}) \geq c \cdot \varepsilon^{-2} \cdot \frac{SA}{\Delta^2}, \tag{10.4.7}$$

where $c > 0$ is an absolute constant.

It can be shown that, by the construction of $\mathcal{M}_0$, $\mathsf{g}^M \geq \Omega(SA/\Delta)$ for all $M \in \mathcal{M}_0$. Thus, analogous to the multi-armed bandit example, Theorem 10.9 and (10.4.7) imply that any algorithm must fail to learn an ε-optimal allocation with probability $\delta = \Omega(\varepsilon \cdot \min\{1, \frac{SA}{\Delta \mathsf{n}_{\max}}\})$ unless

$$T \gtrsim \min\left\{1, \frac{SA}{\Delta \mathsf{n}_{\max}}\right\} \cdot \frac{SA}{\varepsilon \Delta^2}.$$

Choosing $\mathsf{n}_{\max} = O(\frac{SA}{\Delta^2})$ gives $T^{\mathrm{gl}}(\mathcal{M}; \varepsilon, \mathsf{n}_{\max}, \delta) \gtrsim \frac{SA}{\varepsilon \Delta}$. With this same choice of $\mathsf{n}_{\max}$, Theorem 10.10 implies that any algorithm with $\mathbb{E}^{M, \mathbb{A}}[\mathbf{Reg}(T)] \leq 2\mathsf{g}^M \log(T)$ for all $M \in \mathcal{M}_0$ must fail to learn an ε-optimal allocation with probability $\delta = \Omega(\varepsilon \Delta)$ unless

$$\sup_{M \in \mathcal{M}_0} \frac{\mathsf{g}^M}{\Delta_{\min}^M} \cdot \log(T) \gtrsim SA,$$

or equivalently $\log(T^{\mathrm{gl}}(\mathcal{M}; \varepsilon, \mathsf{n}_{\max}, \delta, R)) \gtrsim \frac{SA}{\sup_{M \in \mathcal{M}_0} \mathsf{g}^M / \Delta_{\min}^M}$ for $R = 2 \sup_{M \in \mathcal{M}_0} \mathsf{g}^M$.

10.4.4 Discussion and Interpretation

Our lower bounds show that the Allocation-Estimation Coefficient serves as a fundamental limit on the sample complexity required to learn an approximate Graves-Lai allocation. In particular, they capture phenomena such as the necessity of searching for an informative arm in Example 10.1.1 that are missed by purely asymptotic analyses. To the best of

our knowledge, our lower bounds represent the first attempt to systematically understand the sample complexity of learning the Graves-Lai allocation in a general decision making framework. As such, they are somewhat coarse (in particular, the dependence on parameters such as ε, $\mathsf{n}_{\max}$, and $\Delta^M_{\min}$ is almost certainly loose), and they are best thought of as a starting point for further research. In what follows, we provide additional interpretation of the results, and highlight some of the most interesting remaining questions.

Regret versus learning the optimal allocation. Theorems 10.9 and 10.10 lower bound the sample complexity required to learn an ε-optimal Graves-Lai allocation. Intuitively, this task is closely related to achieving instance-optimal regret. Our analysis of AE^2 shows that it is *sufficient*, and many prior works aim to directly estimate the optimal allocation as well. However, it is unclear to what extent learning the optimal allocation is *necessary* to achieve instance-optimal regret.

In more detail, it is quite straightforward to show that if an algorithm achieves instance-optimal regret, its empirical frequencies act as an optimal allocation *in expectation*.

Lemma 10.4.1. *Let $\varepsilon \in (0,2)$, and suppose that Assumption 10.7 holds. Fix $T \in \mathbb{N}$ and consider an algorithm $\mathbb{A}$ such that for all $M \in \mathcal{M}$,*

$$\mathbb{E}^{M,\mathbb{A}}[\mathbf{Reg}(T)] \le (1+\varepsilon)\mathsf{g}^M \cdot \log(T).$$

For each $M \in \mathcal{M}$, define $\eta^M \in \mathbb{R}^{\Pi}_+$ via $\eta^M(\pi) = \mathbb{E}^{M,\mathbb{A}}\left[\frac{T(\pi)}{\log(T)}\right]$, where $T(\pi)$ denotes the number of pulls of decision π, and define $\lambda^M = \eta^M / \|\eta^M\|_1$. Then if

$$\log(T) \ge \frac{6}{\varepsilon}\log\left(\sup_{M\in\mathcal{M}} \frac{2\mathsf{g}^M}{\Delta^M_{\min}} \cdot \log(T)\right),$$

we have that for all $M \in \mathcal{M}$,

$$\lambda^M \in \Upsilon(M;\varepsilon). \tag{10.4.8}$$

This result gives a guarantee on the expected frequencies of any instance-optimal algorithm, but does not give any guarantee for the realized frequencies. As such, without further assumptions on the algorithm under consideration, it is unclear whether instance-optimal regret implies that it is possible to learn an optimal allocation with high or even *constant* probability. We cannot currently rule out the existence of pathological algorithms for which η^M is optimal in expectation, yet the empirical arm frequencies deviate from the mean with moderate probability. Nonetheless, if one is willing to make stronger assumptions on the algorithm under consideration—in particular, that the *second moment* of regret is controlled—then it is possible to derive lower bounds on regret directly.

Theorem 10.11 (Simplified version of Theorem H.6)**.** *Let the time horizon $T \in \mathbb{N}$ and $\varepsilon \in (0, 1/2)$ be given, and suppose that Assumptions 10.5 and 10.7 hold. Suppose there exists an algorithm $\mathbb{A}$ with the property that for all $M \in \mathcal{M}$: 1) $\mathbb{E}^{M,\mathbb{A}}[\mathbf{Reg}(T)] \le (1+\varepsilon)\mathsf{g}^M \log(T)$, 2)*

$\sqrt{\mathbb{E}^{M,\mathbb{A}}[(\mathbf{Reg}(T))^2]} \le 2\mathsf{g}^M \log(T)$, *and 3) for all* $\pi \in \Pi$, *if* $\mathbb{E}^{M,\mathbb{A}}[T(\pi)] \neq 0$, *then* $\mathbb{E}^{M,\mathbb{A}}[T(\pi)] \ge 1$. *Then if we define* $\delta = \varepsilon \cdot \min\{1, \inf_{M \in \mathcal{M}_0} \frac{\mathsf{g}^M}{3\mathsf{g}^M/\Delta^M_{\min} + \mathsf{n}^{\mathcal{M}}_{\varepsilon}}\}$, *it must be the case that*

$$\log^3(T) \ge \frac{\delta^2}{C} \cdot \sup_{\overline{M} \in \mathcal{M}^+} \mathsf{aec}^{\mathcal{M}}_{4\varepsilon}(\mathcal{M}^{\mathrm{opt}}(\overline{M}), \overline{M}).$$

for $C \le O\Big((\sup_{M \in \mathcal{M}} \frac{\mathsf{g}^M}{\Delta^M_{\min}})^4 \cdot \frac{V^2_{\mathcal{M}} \log(\delta^{-1})}{\varepsilon^2}\Big)$.

See Appendix H.5.5 for a full statement and details. The idea behind the proof is to 1) show (via robust mean estimation) that any instance-optimal algorithm with well-behaved tails can be used to estimate the optimal allocation with high probability (with a small blowup in time horizon), and then 2) appeal to Theorem 10.10. More work is required to understand whether 1) we can prove lower bounds on regret directly, and 2) whether it is possible to show that low regret and learning the optimal allocation are equivalent in a stronger sense.

Comparing upper and lower bounds. Keeping the differences between regret minimization and learning the optimal allocation in mind, let us highlight that the lower bound on T provided by Theorem 10.10 seems to qualitatively match the upper bound from Theorems 10.1 and 10.2. In particular, ignoring problem-dependent parameters and polylogarithmic factors, the upper bound Theorem 10.1 scales, for every model $M \in \mathcal{M}$, as

$$\mathbb{E}^M[\mathbf{Reg}(T)] \le (1+\varepsilon)\mathsf{g}^M \log(T) + \widetilde{O}^+(\mathsf{aec}_{\varepsilon}(\mathcal{M})).$$

In order for this bound to simplify to, say,

$$\mathbb{E}^M[\mathbf{Reg}(T)] \le (1+2\varepsilon)\mathsf{g}^M \log(T),$$

we need

$$\mathsf{g}^M \cdot \log(T) \ge \widetilde{\Omega}^+(1) \cdot \frac{\mathsf{aec}_{\varepsilon}(\mathcal{M})}{\varepsilon},$$

which has similar scaling to the lower bound

$$\log(T) \gtrsim \widetilde{\Omega}^+(1) \cdot \sup_{\overline{M} \in \mathcal{M}^+} \mathsf{aec}^{\mathcal{M}}_{\varepsilon}(\mathcal{M}^{\mathrm{opt}}_0(\overline{M}), \overline{M})$$

from Theorem 10.10. As discussed in the prequel, the former result is concerned with regret, while the latter considers the task of learning the optimal allocation, but the scaling $\log(T) \gtrsim \mathsf{aec}_{\varepsilon}(\mathcal{M})$ seems to be fundamental for both. Of course, beyond the gap between regret and learning the allocation, there is still much room to improve the dependence on problem-dependent parameters in both results.

Comparing Theorem 10.9 and Theorem 10.10. Theorem 10.9 and Theorem 10.10 exhibit an interesting dichotomy: Theorem 10.9 places no constraints on the regret of the algorithm under consideration, and gives a lower of the form $T \gtrsim \widetilde{\Omega}^{+}(1) \cdot \sup_{\overline{M} \in \mathcal{M}^{+}} \mathsf{aec}_{\varepsilon}^{\mathcal{M}}(\mathcal{M}_0, \overline{M})$, while Theorem 10.10 gives a lower bound of the form $\log(T) \gtrsim \widetilde{\Omega}^{+}(1) \cdot \sup_{\overline{M} \in \mathcal{M}^{+}} \mathsf{aec}_{\varepsilon}^{\mathcal{M}}(\mathcal{M}_0^{\mathrm{opt}}(\overline{M}), \overline{M})$, or equivalently $T \gtrsim \exp(\widetilde{\Omega}^{+}(1) \cdot \sup_{\overline{M} \in \mathcal{M}^{+}} \mathsf{aec}_{\varepsilon}(\mathcal{M}_0^{\mathrm{opt}}(\overline{M}), \overline{M}))$; the latter lower bound is exponentially stronger, with the caveat that 1) the class $\mathcal{M}_0$ is replaced with the subclass $\mathcal{M}_0^{\mathrm{opt}}(\overline{M})$ and 2) Theorem 10.10 assumes that the algorithm achieve nearly-instance optimal regret for every model in $\mathcal{M}_0$ ($\mathbb{E}^{M, \mathbb{A}}[\mathbf{Reg}] \leq 2 \cdot \mathsf{g}^{M} \log(T)$). In what follows, we argue that this tradeoff is fundamental.

- First, let us consider the role of the assumption $\mathbb{E}^{M, \mathbb{A}}[\mathbf{Reg}] \leq 2 \cdot \mathsf{g}^{M} \log(T)$. Without this assumption, the lower bound from Theorem 10.9 is qualitatively tight: if the algorithm explores optimally for every round $t \in [T]$, it gains roughly $(\sup_{\overline{M} \in \mathcal{M}^{+}} \mathsf{aec}_{\varepsilon}^{\mathcal{M}}(\mathcal{M}_0, \overline{M}))^{-1}$ units of information per round, which is sufficient to identify an optimal allocation as soon as $T \gtrsim \sup_{\overline{M} \in \mathcal{M}^{+}} \mathsf{aec}_{\varepsilon}^{\mathcal{M}}(\mathcal{M}_0, \overline{M})$.

- On the other hand, if the require that $\mathbb{E}^{M, \mathbb{A}}[\mathbf{Reg}] \leq 2 \cdot \mathsf{g}^{M} \log(T)$, then for each $M \in \mathcal{M}_0$, the algorithm can afford to explore (i.e., play a non-optimal action) at most $2 \cdot \frac{\mathsf{g}^{M}}{\Delta_{\min}^{M}} \log(T)$ times. This changes the "effective" time horizon for exploration to $T' = 2 \cdot \frac{\mathsf{g}^{M}}{\Delta_{\min}^{M}} \log(T)$, but only if we restrict to models for which playing an optimal decision gives no information. This is precisely what the subclass $\mathcal{M}_0^{\mathrm{opt}}(\overline{M})$ captures: models $M \in \mathcal{M}_0$ for which decisions that are optimal for $\overline{M}$ lead to no information. Combining these insights leads to the lower bound $T' \approx \frac{\mathsf{g}^{M}}{\Delta_{\min}^{M}} \log(T) \gtrsim \Omega(1) \cdot \sup_{\overline{M} \in \mathcal{M}^{+}} \mathsf{aec}_{\varepsilon}^{\mathcal{M}}(\mathcal{M}_0^{\mathrm{opt}}(\overline{M}), \overline{M})$ in Theorem 10.10.

We remark in passing that the definition of $\mathcal{M}_0^{\mathrm{opt}}(\overline{M})$, which places the constraint that $D_{\mathsf{KL}}\big(\overline{M}(\pi) \,\|\, M(\pi)\big) = 0, \;\; \forall \pi \in \boldsymbol{\pi}_{\overline{M}}$ is somewhat coarse. We expect that both Theorem 10.10 and Theorem 10.1/Theorem 10.2 can be improved to scale with

$$\mathcal{M}_0^{\mathrm{opt}}(\overline{M}; \alpha) = \big\{ M \in \mathcal{M}_0 \mid \boldsymbol{\pi}_M \subseteq \boldsymbol{\pi}_{\overline{M}}, \;\; D_{\mathsf{KL}}\big(\overline{M}(\pi) \,\|\, M(\pi)\big) \leq \alpha^2 \;\; \forall \pi \in \boldsymbol{\pi}_{\overline{M}} \big\}$$

for $\alpha \approx 1/\sqrt{T}$; the intuition is that we get $\Omega(T)$ rounds worth of information on $\boldsymbol{\pi}_M$ for "free", which facilitates accurate estimation.

Minimax versus instance-dependent lower bounds. As mentioned in the prequel, our lower bounds have a (constrained) minimax flavor. Specifically, Theorem 10.10 shows that if T is not sufficiently large, then for any algorithm, there must exist a "worst-case" model $M \in \mathcal{M}$ for which the algorithm either 1) fails to achieve (approximately) instance-optimal regret or 2) fails to learn an ε-optimal Graves-Lai allocation. While this is quite different from a classical minimax analysis, and certainly is closely connected to instance-optimality, an interesting direction for future work is to develop a fully instance-dependent understanding of the complexity of learning Graves-Lai allocations.

BIBLIOGRAPHY

Yasin Abbasi-Yadkori and Csaba Szepesvári. Regret bounds for the adaptive control of linear quadratic systems. In *Proceedings of the 24th Annual Conference on Learning Theory*, pages 1–26. JMLR Workshop and Conference Proceedings, 2011.

Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. *Advances in neural information processing systems*, 24:2312–2320, 2011.

Marc Abeille and Alessandro Lazaric. Efficient optimistic exploration in linear-quadratic regulators via lagrangian relaxation. In *International Conference on Machine Learning*, pages 23–31. PMLR, 2020.

Naman Agarwal, Brian Bullins, Elad Hazan, Sham Kakade, and Karan Singh. Online control with adversarial disturbances. In *International Conference on Machine Learning*, pages 111–119. PMLR, 2019.

Naman Agarwal, Syomantak Chaudhuri, Prateek Jain, Dheeraj Nagaraj, and Praneeth Netrapalli. Online target q-learning with reverse experience replay: Efficiently finding the optimal policy for linear mdps. *arXiv preprint arXiv:2110.08440*, 2021.

Aymen Al Marjani, Aurélien Garivier, and Alexandre Proutiere. Navigating to the best policy in markov decision processes. *Advances in Neural Information Processing Systems*, 34: 25852–25864, 2021.

Ery Arias-Castro, Emmanuel J Candes, and Mark A Davenport. On the fundamental limits of adaptive sensing. *IEEE Transactions on Information Theory*, 59(1):472–481, 2012.

Karl J Åström and Björn Wittenmark. *Adaptive control*. Courier Corporation, 2013.

Jean-Yves Audibert and Sébastien Bubeck. Minimax policies for adversarial and stochastic bandits. In *COLT*, volume 7, pages 1–122, 2009.

Peter Auer and Ronald Ortner. Logarithmic online regret bounds for undiscounted reinforcement learning. *Advances in neural information processing systems*, 19, 2006.

Peter Auer, Nicolo Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47(2-3):235–256, 2002.

Peter Auer, Thomas Jaksch, and Ronald Ortner. Near-optimal regret bounds for reinforcement learning. *Advances in neural information processing systems*, 21, 2008.

Alex Ayoub, Zeyu Jia, Csaba Szepesvari, Mengdi Wang, and Lin Yang. Model-based reinforcement learning with value-targeted regression. In *International Conference on Machine Learning*, pages 463–474. PMLR, 2020.

Mohammad Gheshlaghi Azar, Ian Osband, and Rémi Munos. Minimax regret bounds for reinforcement learning. In *International Conference on Machine Learning*, pages 263–272. PMLR, 2017.

Leemon Baird. Residual algorithms: Reinforcement learning with function approximation. In *Machine Learning Proceedings 1995*, pages 30–37. Elsevier, 1995.

Märta Barenthin, Henrik Jansson, and Håkan Hjalmarsson. Applications of mixed h2 and hinfin; input design in identification. *IFAC Proceedings Volumes*, 38(1):458–463, 2005.

Nicholas M Boffi, Stephen Tu, and Jean-Jacques E Slotine. Regret bounds for adaptive nonlinear control. In *Learning for Dynamics and Control*, pages 471–483. PMLR, 2021.

Steven J Bradtke and Andrew G Barto. Linear least-squares algorithms for temporal difference learning. *Machine learning*, 22(1):33–57, 1996.

Ronen I Brafman and Moshe Tennenholtz. R-max-a general polynomial time algorithm for near-optimal reinforcement learning. *Journal of Machine Learning Research*, 3(Oct): 213–231, 2002.

Douglas A Bristow, Marina Tharayil, and Andrew G Alleyne. A survey of iterative learning control. *IEEE control systems magazine*, 26(3):96–114, 2006.

Lukas Brunke, Melissa Greeff, Adam W Hall, Zhaocong Yuan, Siqi Zhou, Jacopo Panerati, and Angela P Schoellig. Safe learning in robotics: From learning-based control to safe reinforcement learning. *Annual Review of Control, Robotics, and Autonomous Systems*, 5: 411–444, 2022.

Apostolos N Burnetas and Michael N Katehakis. Optimal adaptive policies for sequential allocation problems. *Advances in Applied Mathematics*, 17(2):122–142, 1996.

Kamalika Chaudhuri, Sham M Kakade, Praneeth Netrapalli, and Sujay Sanghavi. Convergence rates of active learning for maximum likelihood estimation. *Advances in Neural Information Processing Systems*, 28, 2015.

Probal Chaudhuri and Per A Mykland. Nonlinear experiments: Optimal design and inference based on likelihood. *Journal of the American Statistical Association*, 88(422):538–546, 1993.

Fan Chen, Song Mei, and Yu Bai. Unified algorithms for RL with decision-estimation coefficients: No-regret, PAC, and reward-free learning. *arXiv preprint arXiv:2209.11745*, 2022.

Lijie Chen, Jian Li, and Mingda Qiao. Nearly instance optimal sample complexity bounds for top-k arm selection. In *Artificial Intelligence and Statistics*, pages 101–110. PMLR, 2017.

Alon Cohen, Tomer Koren, and Yishay Mansour. Learning linear-quadratic regulators efficiently with only $\sqrt{T}$ regret. In *International Conference on Machine Learning*, pages 1300–1309. PMLR, 2019.

Richard Combes, Stefan Magureanu, and Alexandre Proutiere. Minimal exploration in structured stochastic bandits. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 1761–1769, 2017.

Varsha Dani, Thomas P Hayes, and Sham M Kakade. Stochastic linear optimization under bandit feedback. In *Conference on Learning Theory (COLT)*, 2008.

Christoph Dann and Emma Brunskill. Sample complexity of episodic fixed-horizon reinforcement learning. *Advances in Neural Information Processing Systems*, 28, 2015.

Christoph Dann, Tor Lattimore, and Emma Brunskill. Unifying pac and regret: Uniform pac bounds for episodic reinforcement learning. *Advances in Neural Information Processing Systems*, 30, 2017.

Christoph Dann, Lihong Li, Wei Wei, and Emma Brunskill. Policy certificates: Towards accountable reinforcement learning. In *International Conference on Machine Learning*, pages 1507–1516. PMLR, 2019.

Christoph Dann, Teodor Vanislavov Marinov, Mehryar Mohri, and Julian Zimmert. Beyond value-function gaps: Improved instance-dependent regret bounds for episodic reinforcement learning. *Advances in Neural Information Processing Systems*, 34, 2021.

Sarah Dean, Horia Mania, Nikolai Matni, Benjamin Recht, and Stephen Tu. Regret bounds for robust adaptive control of the linear quadratic regulator. In *Advances in Neural Information Processing Systems*, pages 4188–4197, 2018.

Sarah Dean, Stephen Tu, Nikolai Matni, and Benjamin Recht. Safely learning to control the constrained linear quadratic regulator. In *2019 American Control Conference (ACC)*, pages 5582–5588. IEEE, 2019.

Sarah Dean, Horia Mania, Nikolai Matni, Benjamin Recht, and Stephen Tu. On the sample complexity of the linear quadratic regulator. *Foundations of Computational Mathematics*, 20(4):633–679, 2020.

Rémy Degenne and Wouter M Koolen. Pure exploration with multiple correct answers. *Advances in Neural Information Processing Systems*, 32, 2019.

Rémy Degenne, Wouter M Koolen, and Pierre Ménard. Non-asymptotic pure exploration by solving games. *Advances in Neural Information Processing Systems*, 32, 2019.

Rémy Degenne, Pierre Ménard, Xuedong Shang, and Michal Valko. Gamification of pure exploration for linear bandits. In *International Conference on Machine Learning*, pages 2432–2442. PMLR, 2020a.

Rémy Degenne, Han Shao, and Wouter Koolen. Structure adaptive algorithms for stochastic bandits. In *International Conference on Machine Learning*, pages 2443–2452. PMLR, 2020b.

Omar Darwiche Domingues, Pierre Ménard, Emilie Kaufmann, and Michal Valko. Episodic reinforcement learning in finite mdps: Minimax lower bounds revisited. In *Algorithmic Learning Theory*, pages 578–598. PMLR, 2021.

Kefan Dong and Tengyu Ma. Asymptotic instance-optimal algorithms for interactive decision making. *arXiv preprint arXiv:2206.02326*, 2022.

Simon Du, Sham Kakade, Jason Lee, Shachar Lovett, Gaurav Mahajan, Wen Sun, and Ruosong Wang. Bilinear classes: A structural framework for provable generalization in rl. In *International Conference on Machine Learning*, pages 2826–2836. PMLR, 2021.

Simon S Du, Sham M Kakade, Ruosong Wang, and Lin F Yang. Is a good representation sufficient for sample efficient reinforcement learning? *arXiv preprint arXiv:1910.03016*, 2019.

Alessandro Epasto, Mohammad Mahdian, Vahab Mirrokni, and Emmanouil Zampetakis. Optimal approximation-smoothness tradeoffs for soft-max functions. *Advances in Neural Information Processing Systems*, 33:2651–2660, 2020.

Eyal Even-Dar, Shie Mannor, Yishay Mansour, and Sridhar Mahadevan. Action elimination and stopping conditions for the multi-armed bandit and reinforcement learning problems. *Journal of machine learning research*, 7(6), 2006.

Mohamad Kazem Shirani Faradonbeh, Ambuj Tewari, and George Michailidis. Finite time identification in unstable linear systems. *Automatica*, 96:342–353, 2018.

Aleksandr Aronovich Feldbaum. Dual control theory. i. *Avtomatika i Telemekhanika*, 21(9): 1240–1249, 1960.

Tanner Fiez, Lalit Jain, Kevin G Jamieson, and Lillian Ratliff. Sequential experimental design for transductive linear bandits. *Advances in neural information processing systems*, 32, 2019.

Dylan Foster, Tuhin Sarkar, and Alexander Rakhlin. Learning nonlinear dynamical systems from a single trajectory. In *Learning for Dynamics and Control*, pages 851–861. PMLR, 2020a.

Dylan J Foster and Alexander Rakhlin. Beyond UCB: Optimal and efficient contextual bandits with regression oracles. *International Conference on Machine Learning (ICML)*, 2020.

Dylan J Foster, Alexander Rakhlin, David Simchi-Levi, and Yunzong Xu. Instance-dependent complexity of contextual bandits and reinforcement learning: A disagreement-based perspective. *Conference on Learning Theory (COLT)*, 2020b.

Dylan J Foster, Sham M Kakade, Jian Qian, and Alexander Rakhlin. The statistical complexity of interactive decision making. *arXiv preprint arXiv:2112.13487*, 2021.

Dylan J Foster, Alexander Rakhlin, Ayush Sekhari, and Karthik Sridharan. On the complexity of adversarial decision making. *Advances in Neural Information Processing Systems*, 35: 35404–35417, 2022.

Dylan J Foster, Noah Golowich, and Yanjun Han. Tight guarantees for interactive decision making with the decision-estimation coefficient. In *The Thirty Sixth Annual Conference on Learning Theory*, pages 3969–4043. PMLR, 2023.

Dylan J Foster, Noah Golowich, Jian Qian, Alexander Rakhlin, and Ayush Sekhari. Model-free reinforcement learning with the decision-estimation coefficient. *Advances in Neural Information Processing Systems*, 36, 2024.

Marguerite Frank and Philip Wolfe. An algorithm for quadratic programming. *Naval research logistics quarterly*, 3(1-2):95–110, 1956.

David A Freedman. On tail probabilities for martingales. *The Annals of Probability*, pages 100–118, 1975.

Aurélien Garivier and Emilie Kaufmann. Optimal best arm identification with fixed confidence. In *Conference on Learning Theory*, pages 998–1027, 2016.

Aurélien Garivier, Tor Lattimore, and Emilie Kaufmann. On explore-then-commit strategies. In *Advances in Neural Information Processing Systems*, pages 784–792, 2016.

Aurélien Garivier, Pierre Ménard, and Gilles Stoltz. Explore first, exploit next: The true shape of regret in bandit problems. *Mathematics of Operations Research*, 44(2):377–399, 2019.

László Gerencsér and Håkan Hjalmarsson. Adaptive input design in system identification. In *Proceedings of the 44th IEEE Conference on Decision and Control*, pages 4988–4993. IEEE, 2005.

László Gerencsér, Jonas Mårtensson, and Håkan Hjalmarsson. Adaptive input design for arx systems. In *2007 European Control Conference (ECC)*, pages 5707–5714. IEEE, 2007.

László Gerencsér, Håkan Hjalmarsson, and Jonas Mårtensson. Identification of arx systems with non-stationary inputs—asymptotic analysis with application to adaptive input design. *Automatica*, 45(3):623–633, 2009.

Michel Gevers, Alexandre S Bazanella, Xavier Bombois, and Ljubisa Miskovic. Identification and the information matrix: how to get just sufficiently rich? *IEEE Transactions on Automatic Control*, 54(ARTICLE):2828–2840, 2009.

Richard D Gill, Boris Y Levit, et al. Applications of the van trees inequality: a bayesian cramér-rao bound. *Bernoulli*, 1(1-2):59–79, 1995.

Graham Clifford Goodwin and Robert L Payne. *Dynamic system identification: experiment design and data analysis*. Academic press, 1977.

Todd L Graves and Tze Leung Lai. Asymptotically efficient adaptive choice of control laws in controlled Markov chains. *SIAM journal on control and optimization*, 35(3):715–743, 1997.

Per Hägg, Christian A Larsson, and Håkan Hjalmarsson. Robust and adaptive excitation signal generation for input and output constrained systems. In *2013 European Control Conference (ECC)*, pages 1416–1421. IEEE, 2013.

Botao Hao, Tor Lattimore, and Csaba Szepesvari. Adaptive exploration in linear contextual bandit. In *International Conference on Artificial Intelligence and Statistics*, pages 3536–3545. PMLR, 2020.

Botao Hao, Tor Lattimore, Csaba Szepesvári, and Mengdi Wang. Online sparse reinforcement learning. In *International Conference on Artificial Intelligence and Statistics*, pages 316–324. PMLR, 2021.

Moritz Hardt, Tengyu Ma, and Benjamin Recht. Gradient descent learns linear dynamical systems. *The Journal of Machine Learning Research*, 19(1):1025–1068, 2018.

Elad Hazan, Holden Lee, Karan Singh, Cyril Zhang, and Yi Zhang. Spectral filtering for general linear dynamical systems. In *Advances in Neural Information Processing Systems*, pages 4634–4643, 2018.

Elad Hazan, Sham Kakade, Karan Singh, and Abby Van Soest. Provably efficient maximum entropy exploration. In *International Conference on Machine Learning*, pages 2681–2691. PMLR, 2019.

Jiafan He, Dongruo Zhou, and Quanquan Gu. Logarithmic regret for reinforcement learning with linear function approximation. pages 4171–4180, 2021.

Roland Hildebrand and Michel Gevers. Identification for control: optimal input design with respect to a worst-case ν-gap cost function. *SIAM Journal on Control and optimization*, 41 (5):1586–1608, 2002.

Håkan Hjalmarsson, Michel Gevers, and Franky De Bruyne. For model-based control design, closed-loop identification gives better performance. *Automatica*, 32(12):1659–1673, 1996.

Daniel Hsu, Sham Kakade, Tong Zhang, et al. A tail inequality for quadratic forms of subgaussian random vectors. *Electronic Communications in Probability*, 17, 2012.

Kevin Jamieson, Matthew Malloy, Robert Nowak, and Sébastien Bubeck. lil'ucb: An optimal exploration algorithm for multi-armed bandits. In *Conference on Learning Theory*, pages 423–439. PMLR, 2014.

Henrik Jansson and Håkan Hjalmarsson. Input design via lmis admitting frequency-wise model specifications in confidence regions. *IEEE transactions on Automatic Control*, 50 (10):1534–1549, 2005.

Yassir Jedra and Alexandre Proutiere. Sample complexity lower bounds for linear system identification. In *2019 IEEE 58th Conference on Decision and Control (CDC)*, pages 2676–2681. IEEE, 2019.

Zeyu Jia, Lin Yang, Csaba Szepesvari, and Mengdi Wang. Model-based reinforcement learning with value-targeted regression. In *Learning for Dynamics and Control*, pages 666–686. PMLR, 2020.

Nan Jiang, Akshay Krishnamurthy, Alekh Agarwal, John Langford, and Robert E Schapire. Contextual decision processes with low bellman rank are pac-learnable. In *International Conference on Machine Learning*, pages 1704–1713. PMLR, 2017.

Chi Jin, Zeyuan Allen-Zhu, Sebastien Bubeck, and Michael I Jordan. Is q-learning provably efficient? In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pages 4868–4878, 2018.

Chi Jin, Akshay Krishnamurthy, Max Simchowitz, and Tiancheng Yu. Reward-free exploration for reinforcement learning. In *International Conference on Machine Learning*, pages 4870–4879. PMLR, 2020a.

Chi Jin, Zhuoran Yang, Zhaoran Wang, and Michael I Jordan. Provably efficient reinforcement learning with linear function approximation. In *Conference on Learning Theory*, pages 2137–2143. PMLR, 2020b.

Chi Jin, Qinghua Liu, and Sobhan Miryoosefi. Bellman eluder dimension: New rich classes of rl problems, and sample-efficient algorithms. *Advances in neural information processing systems*, 34:13406–13418, 2021.

Anders Jonsson, Emilie Kaufmann, Pierre Ménard, Omar Darwiche Domingues, Edouard Leurent, and Michal Valko. Planning in markov decision processes with gap-dependent sample complexity. *Advances in Neural Information Processing Systems*, 33:1253–1263, 2020.

Kwang-Sung Jun and Chicheng Zhang. Crush optimism with pessimism: Structured bandits beyond asymptotic optimality. *Advances in Neural Information Processing Systems*, 33: 6366–6376, 2020.

Sham Kakade, Akshay Krishnamurthy, Kendall Lowrey, Motoya Ohnishi, and Wen Sun. Information theoretic regret bounds for online nonlinear control. *Advances in Neural Information Processing Systems*, 33:15312–15325, 2020.

Sham Machandranath Kakade. *On the sample complexity of reinforcement learning*. PhD thesis, UCL (University College London), 2003.

Dimitrios Katselis, Cristian R Rojas, Håkan Hjalmarsson, and Mats Bengtsson. Application-oriented finite sample experiment design: A semidefinite relaxation approach. *IFAC Proceedings Volumes*, 45(16):1635–1640, 2012.

Julian Katz-Samuels, Lalit Jain, Kevin G Jamieson, et al. An empirical process approach to the union bound: Practical algorithms for combinatorial and linear bandits. *Advances in Neural Information Processing Systems*, 33:10371–10382, 2020.

Emilie Kaufmann, Olivier Cappé, and Aurélien Garivier. On the complexity of best-arm identification in multi-armed bandit models. *The Journal of Machine Learning Research*, 17(1):1–42, 2016.

Michael Kearns and Satinder Singh. Finite-sample convergence rates for q-learning and indirect algorithms. *Advances in neural information processing systems*, 11, 1998.

Michael Kearns and Satinder Singh. Near-optimal reinforcement learning in polynomial time. *Machine learning*, 49(2):209–232, 2002.

Michael Kearns, Yishay Mansour, and Andrew Ng. Approximate planning in large pomdps via reusable trajectories. *Advances in Neural Information Processing Systems*, 12, 1999.

Johannes Kirschner, Tor Lattimore, Claire Vernade, and Csaba Szepesvári. Asymptotically optimal information-directed sampling. In *Conference on Learning Theory*, pages 2777–2821. PMLR, 2021.

Junpei Komiyama, Junya Honda, and Hiroshi Nakagawa. Regret lower bound and optimal algorithm in finite stochastic partial monitoring. *Advances in Neural Information Processing Systems*, 28, 2015.

Tze Leung Lai and Herbert Robbins. Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, 6(1):4–22, 1985.

Christian Larsson, Egon Geerardyn, and Johan Schoukens. Robust input design for resonant systems under limited a priori information. *IFAC Proceedings Volumes*, 45(16):1611–1616, 2012.

Tor Lattimore. Refining the confidence level for optimistic bandit strategies. *The Journal of Machine Learning Research*, 19(1):765–796, 2018.

Tor Lattimore and Marcus Hutter. Pac bounds for discounted mdps. In *International Conference on Algorithmic Learning Theory*, pages 320–334. Springer, 2012.

Tor Lattimore and Csaba Szepesvari. The end of optimism? an asymptotic analysis of finite-armed linear bandits. In *Artificial Intelligence and Statistics*, pages 728–737. PMLR, 2017.

Tor Lattimore and Csaba Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020.

Zhaoqi Li, Lillian Ratliff, Kevin G Jamieson, Lalit Jain, et al. Instance-optimal pac algorithms for contextual bandits. *Advances in Neural Information Processing Systems*, 35:37590–37603, 2022.

Kristian Lindqvist and Håkan Hjalmarsson. Identification for control: Adaptive input design using convex optimization. In *Proceedings of the 40th IEEE Conference on Decision and Control (Cat. No. 01CH37228)*, volume 5, pages 4326–4331. IEEE, 2001.

Lennart Ljung. *System identification*. Springer, 1998.

Gábor Lugosi and Shahar Mendelson. Mean estimation and regression under heavy-tailed distributions: A survey. *Foundations of Computational Mathematics*, 19(5):1145–1190, 2019.

Stefan Magureanu, Richard Combes, and Alexandre Proutiere. Lipschitz bandits: Regret lower bound and optimal algorithms. In *Conference on Learning Theory*, pages 975–999. PMLR, 2014.

Ian R Manchester. Input design for system identification via convex relaxation. In *49th IEEE Conference on Decision and Control (CDC)*, pages 2041–2046. IEEE, 2010.

Horia Mania, Stephen Tu, and Benjamin Recht. Certainty equivalence is efficient for linear quadratic control. *Advances in Neural Information Processing Systems*, 32, 2019.

Horia Mania, Michael I Jordan, and Benjamin Recht. Active learning for nonlinear system identification with guarantees. *J. Mach. Learn. Res.*, 23:32–1, 2022.

Teodor Vanislavov Marinov, Mehryar Mohri, and Julian Zimmert. Open problem: Finite-time instance dependent optimality for stochastic online learning with feedback graphs. In *Conference on Learning Theory*, pages 5644–5649. PMLR, 2022a.

Teodor Vanislavov Marinov, Mehryar Mohri, and Julian Zimmert. Stochastic online learning with feedback graphs: Finite-time and asymptotic optimality. *Advances in Neural Information Processing Systems*, 35:24947–24959, 2022b.

Aymen Al Marjani and Alexandre Proutiere. Best policy identification in discounted mdps: Problem-specific sample complexity. *arXiv preprint arXiv:2009.13405*, 2020.

Frank McSherry and Kunal Talwar. Mechanism design via differential privacy. In *48th Annual IEEE Symposium on Foundations of Computer Science (FOCS'07)*, pages 94–103. IEEE, 2007.

Raman Mehra. Optimal input signals for parameter estimation in dynamic systems–survey and new results. *IEEE Transactions on Automatic Control*, 19(6):753–768, 1974.

Raman K Mehra. Synthesis of optimal inputs for multiinput-multioutput (mimo) systems with process noise part i: Frequenc y-domain synthesis part ii: Time-domain synthesis. In *Mathematics in Science and Engineering*, volume 126, pages 211–249. Elsevier, 1976.

Francisco S Melo and M Isabel Ribeiro. Q-learning with linear function approximation. In *International Conference on Computational Learning Theory*, pages 308–322. Springer, 2007.

Pierre Ménard, Omar Darwiche Domingues, Anders Jonsson, Emilie Kaufmann, Edouard Leurent, and Michal Valko. Fast active learning for pure exploration in reinforcement learning. In *International Conference on Machine Learning*, pages 7599–7608. PMLR, 2021.

Ali Mesbah. Stochastic model predictive control with active uncertainty learning: A survey on dual control. *Annual Reviews in Control*, 45:107–117, 2018.

Rémi Munos and Csaba Szepesvári. Finite-time bounds for fitted value iteration. *Journal of Machine Learning Research*, 9(5), 2008.

Mojmir Mutny, Tadeusz Janik, and Andreas Krause. Active exploration via experiment design in markov chains. In *International Conference on Artificial Intelligence and Statistics*, pages 7349–7374. PMLR, 2023.

Duy Nguyen-Tuong and Jan Peters. Model learning for robot control: a survey. *Cognitive processing*, 12:319–340, 2011.

Jungseul Ok, Alexandre Proutiere, and Damianos Tranos. Exploration in structured reinforcement learning. *Advances in Neural Information Processing Systems*, 31, 2018.

Ian Osband and Benjamin Van Roy. On lower bounds for regret in reinforcement learning. *arXiv preprint arXiv:1608.02732*, 2016.

Samet Oymak. Stochastic gradient descent learns state equations with nonlinear activations. In *Conference on Learning Theory*, pages 2551–2579. PMLR, 2019.

Samet Oymak and Necmiye Ozay. Non-asymptotic identification of lti systems from a single trajectory. In *2019 American Control Conference (ACC)*, pages 5655–5661. IEEE, 2019.

Michael O'Connell, Guanya Shi, Xichen Shi, Kamyar Azizzadenesheli, Anima Anandkumar, Yisong Yue, and Soon-Jo Chung. Neural-fly enables rapid learning for agile flight in strong winds. *Science Robotics*, 7(66):eabm6597, 2022.

Luc Pronzato and Andrej Pázman. Design of experiments in nonlinear models. *Lecture notes in statistics*, 212, 2013.

Friedrich Pukelsheim. *Optimal design of experiments*. SIAM, 2006.

Ali Rahimi and Benjamin Recht. Uniform approximation of functions with random bases. In *2008 46th annual allerton conference on communication, control, and computing*, pages 555–561. IEEE, 2008.

Spencer M Richards, Navid Azizan, Jean-Jacques Slotine, and Marco Pavone. Adaptive-control-oriented meta-learning for nonlinear systems. *arXiv preprint arXiv:2103.04490*, 2021.

Cristian R Rojas, James S Welsh, Graham C Goodwin, and Arie Feuer. Robust optimal experiment design for system identification. *Automatica*, 43(6):993–1008, 2007.

Cristian R Rojas, Juan-Carlos Aguero, James S Welsh, Graham C Goodwin, and Arie Feuer. Robustness in experiment design. *IEEE Transactions on Automatic Control*, 57(4):860–874, 2011.

Daniel Russo. Simple bayesian algorithms for best arm identification. In *Conference on Learning Theory*, pages 1417–1418, 2016.

Daniel Russo and Benjamin Van Roy. Eluder dimension and the sample complexity of optimistic exploration. In *Advances in Neural Information Processing Systems*, pages 2256–2264, 2013.

Daniel Russo and Benjamin Van Roy. Learning to optimize via information-directed sampling. *Operations Research*, 66(1):230–252, 2018.

Tuhin Sarkar and Alexander Rakhlin. Near optimal finite time identification of arbitrary linear dynamical systems. In *International Conference on Machine Learning*, pages 5610–5618. PMLR, 2019.

Tuhin Sarkar, Alexander Rakhlin, and Munther A Dahleh. Nonparametric finite time lti system identification. *arXiv preprint arXiv:1902.01848*, 2019.

Yahya Sattar and Samet Oymak. Non-asymptotic and accurate learning of nonlinear dynamical systems. *The Journal of Machine Learning Research*, 23(1):6248–6296, 2022.

Ohad Shamir. A variant of azuma's inequality for martingales with subgaussian tails. *arXiv preprint arXiv:1110.2392*, 2011.

Ohad Shamir. On the complexity of bandit and derivative-free stochastic convex optimization. In *Conference on Learning Theory*, pages 3–24. PMLR, 2013.

Guanya Shi, Xichen Shi, Michael O'Connell, Rose Yu, Kamyar Azizzadenesheli, Animashree Anandkumar, Yisong Yue, and Soon-Jo Chung. Neural lander: Stable drone landing control using learned dynamics. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 9784–9790. IEEE, 2019.

Guanya Shi, Kamyar Azizzadenesheli, Michael O'Connell, Soon-Jo Chung, and Yisong Yue. Meta-adaptive nonlinear control: Theory and algorithms. *Advances in Neural Information Processing Systems*, 34:10013–10025, 2021a.

Guanya Shi, Wolfgang Hönig, Xichen Shi, Yisong Yue, and Soon-Jo Chung. Neural-swarm2: Planning and control of heterogeneous multirotor swarms using learned interactions. *IEEE Transactions on Robotics*, 38(2):1063–1079, 2021b.

Max Simchowitz and Dylan Foster. Naive exploration is optimal for online lqr. In *International Conference on Machine Learning*, pages 8937–8948. PMLR, 2020.

Max Simchowitz and Kevin G Jamieson. Non-asymptotic gap-dependent regret bounds for tabular mdps. *Advances in Neural Information Processing Systems*, 32, 2019.

Max Simchowitz, Kevin Jamieson, and Benjamin Recht. The simulator: Understanding adaptive sampling in the moderate-confidence regime. In *Conference on Learning Theory*, pages 1794–1834. PMLR, 2017.

Max Simchowitz, Horia Mania, Stephen Tu, Michael I Jordan, and Benjamin Recht. Learning without mixing: Towards a sharp analysis of linear system identification. In *Conference On Learning Theory*, pages 439–473. PMLR, 2018.

Max Simchowitz, Ross Boczar, and Benjamin Recht. Learning linear dynamical systems with semi-parametric least squares. In *Conference on Learning Theory*, pages 2714–2802. PMLR, 2019.

Max Simchowitz, Karan Singh, and Elad Hazan. Improper learning for non-stochastic control. In *Conference on Learning Theory*, pages 3320–3436. PMLR, 2020.

Herbert A Simon. Dynamic programming under uncertainty with a quadratic criterion function. *Econometrica, Journal of the Econometric Society*, pages 74–81, 1956.

Jean-Jacques E Slotine, Weiping Li, et al. *Applied nonlinear control*, volume 199. Prentice hall Englewood Cliffs, NJ, 1991.

Marta Soare, Alessandro Lazaric, and Rémi Munos. Best-arm identification in linear bandits. *Advances in Neural Information Processing Systems*, 27, 2014.

Yuda Song and Wen Sun. Pc-mlp: Model-based reinforcement learning with policy cover guided exploration. In *International Conference on Machine Learning*, pages 9801–9811. PMLR, 2021.

Alexander L Strehl, Lihong Li, Eric Wiewiora, John Langford, and Michael L Littman. Pac model-free reinforcement learning. In *Proceedings of the 23rd international conference on Machine learning*, pages 881–888, 2006.

Alexander L Strehl, Lihong Li, and Michael L Littman. Reinforcement learning in finite mdps: Pac analysis. *Journal of Machine Learning Research*, 10(11), 2009.

Richard S Sutton, David McAllester, Satinder Singh, and Yishay Mansour. Policy gradient methods for reinforcement learning with function approximation. *Advances in neural information processing systems*, 12, 1999.

Henri Theil. A note on certainty equivalence in dynamic planning. *Econometrica: Journal of the Econometric Society*, pages 346–349, 1957.

Andrea Tirinzoni, Matteo Pirotta, Marcello Restelli, and Alessandro Lazaric. An asymptotically optimal primal-dual incremental algorithm for contextual linear bandits. *Advances in Neural Information Processing Systems*, 33:1417–1427, 2020.

Andrea Tirinzoni, Matteo Pirotta, and Alessandro Lazaric. A fully problem-dependent regret lower bound for finite-horizon mdps. *arXiv preprint arXiv:2106.13013*, 2021.

Andrea Tirinzoni, Aymen Al Marjani, and Emilie Kaufmann. Near instance-optimal pac reinforcement learning for deterministic mdps. *Advances in neural information processing systems*, 35:8785–8798, 2022.

Anastasios Tsiamis and George J Pappas. Finite sample analysis of stochastic system identification. In *2019 IEEE 58th Conference on Decision and Control (CDC)*, pages 3648–3654. IEEE, 2019.

Alexandre B Tsybakov. Introduction to nonparametric estimation., 2009.

Stephen Tu, Ross Boczar, Andrew Packard, and Benjamin Recht. Non-asymptotic analysis of robust control from coarse-grained identification. *arXiv preprint arXiv:1707.04791*, 2017.

Bart Van Parys and Negin Golrezaei. Optimal learning for structured bandits. *Management Science*, 2023.

Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*, 2010.

Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

Andrew Wagenmaker and Dylan J Foster. Instance-optimality in interactive decision making: Toward a non-asymptotic theory. In *The Thirty Sixth Annual Conference on Learning Theory*, pages 1322–1472. PMLR, 2023.

Andrew Wagenmaker and Kevin Jamieson. Active learning for identification of linear dynamical systems. In *Conference on Learning Theory*, pages 3487–3582. PMLR, 2020.

Andrew Wagenmaker and Kevin G Jamieson. Instance-dependent near-optimal policy identification in linear mdps via online experiment design. *Advances in Neural Information Processing Systems*, 35:5968–5981, 2022.

Andrew Wagenmaker, Max Simchowitz, and Kevin Jamieson. Task-optimal exploration in linear dynamical systems. In *International Conference on Machine Learning*, pages 10641–10652. PMLR, 2021.

Andrew Wagenmaker, Yifang Chen, Max Simchowitz, Simon Du, and Kevin Jamieson. First-order regret in reinforcement learning with linear function approximation: A robust estimation approach. In *International Conference on Machine Learning*, pages 22384–22429. PMLR, 2022a.

Andrew Wagenmaker, Yifang Chen, Max Simchowitz, Simon Du, and Kevin Jamieson. Reward-free rl is no harder than reward-aware rl in linear markov decision processes. In *International Conference on Machine Learning*, pages 22430–22456. PMLR, 2022b.

Andrew Wagenmaker, Max Simchowitz, and Kevin Jamieson. Beyond no regret: Instance-dependent pac reinforcement learning. In *Conference on Learning Theory*, pages 358–418. PMLR, 2022c.

Andrew Wagenmaker, Guanya Shi, and Kevin G Jamieson. Optimal exploration for model-based rl in nonlinear systems. *Advances in Neural Information Processing Systems*, 36, 2024.

Ruosong Wang, Simon S Du, Lin Yang, and Russ R Salakhutdinov. On reward-free reinforcement learning with linear function approximation. *Advances in neural information processing systems*, 33:17816–17826, 2020.

Yining Wang, Ruosong Wang, Simon S Du, and Akshay Krishnamurthy. Optimism in reinforcement learning with generalized linear function approximation. *arXiv preprint arXiv:1912.04136*, 2019.

Yuanhao Wang, Ruosong Wang, and Sham Kakade. An exponential lower bound for linearly realizable mdp with constant suboptimality gap. *Advances in Neural Information Processing Systems*, 34:9521–9533, 2021.

Gellért Weisz, Philip Amortila, and Csaba Szepesvári. Exponential lower bounds for planning in mdps with linearly-realizable optimal action-value functions. In *Algorithmic Learning Theory*, pages 1237–1264. PMLR, 2021.

Grady Williams, Nolan Wagener, Brian Goldfain, Paul Drews, James M Rehg, Byron Boots, and Evangelos A Theodorou. Information theoretic mpc for model-based reinforcement learning. In *2017 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1714–1721. IEEE, 2017.

Jingfeng Wu, Vladimir Braverman, and Lin Yang. Gap-dependent unsupervised exploration for reinforcement learning. In *International Conference on Artificial Intelligence and Statistics*, pages 4109–4131. PMLR, 2022.

Haike Xu, Tengyu Ma, and Simon Du. Fine-grained gap-dependent bounds for tabular mdps via adaptive multi-step bootstrap. In *Conference on Learning Theory*, pages 4438–4472. PMLR, 2021.

Lin Yang and Mengdi Wang. Sample-optimal parametric q-learning using linearly additive features. In *International Conference on Machine Learning*, pages 6995–7004. PMLR, 2019.

Yuhong Yang and Andrew R Barron. An asymptotic property of model selection criteria. *IEEE Transactions on Information Theory*, 44(1):95–116, 1998.

Chenkai Yu, Guanya Shi, Soon-Jo Chung, Yisong Yue, and Adam Wierman. The power of predictions in online control. *Advances in Neural Information Processing Systems*, 33: 1994–2004, 2020.

Tom Zahavy, Brendan O'Donoghue, Guillaume Desjardins, and Satinder Singh. Reward is enough for convex mdps. *Advances in Neural Information Processing Systems*, 34: 25746–25759, 2021.

Andrea Zanette and Emma Brunskill. Tighter problem-dependent regret bounds in reinforcement learning without domain knowledge using value function bounds. In *International Conference on Machine Learning*, pages 7304–7312. PMLR, 2019.

Andrea Zanette, Mykel J Kochenderfer, and Emma Brunskill. Almost horizon-free structure-aware best policy identification with a generative model. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

Andrea Zanette, David Brandfonbrener, Emma Brunskill, Matteo Pirotta, and Alessandro Lazaric. Frequentist regret bounds for randomized least-squares value iteration. In *International Conference on Artificial Intelligence and Statistics*, pages 1954–1964. PMLR, 2020a.

Andrea Zanette, Alessandro Lazaric, Mykel Kochenderfer, and Emma Brunskill. Learning near optimal policies with low inherent bellman error. In *International Conference on Machine Learning*, pages 10978–10989. PMLR, 2020b.

Andrea Zanette, Alessandro Lazaric, Mykel J Kochenderfer, and Emma Brunskill. Provably efficient reward-agnostic navigation with linear value iteration. *Advances in Neural Information Processing Systems*, 33:11756–11766, 2020c.

Weitong Zhang, Dongruo Zhou, and Quanquan Gu. Reward-free model-based reinforcement learning with linear function approximation. *Advances in Neural Information Processing Systems*, 34, 2021a.

Zihan Zhang, Simon S Du, and Xiangyang Ji. Nearly minimax optimal reward-free reinforcement learning. *arXiv preprint arXiv:2010.05901*, 2020.

Zihan Zhang, Xiangyang Ji, and Simon Du. Is reinforcement learning more difficult than bandits? a near-optimal algorithm escaping the curse of horizon. In *Conference on Learning Theory*, pages 4528–4531. PMLR, 2021b.

Zihan Zhang, Jiaqi Yang, Xiangyang Ji, and Simon S Du. Variance-aware confidence set: Variance-dependent bound for linear bandits and horizon-free bound for linear mixture mdp. *arXiv preprint arXiv:2101.12745*, 2021c.

Dongruo Zhou, Quanquan Gu, and Csaba Szepesvari. Nearly minimax optimal reinforcement learning for linear mixture markov decision processes. In *Conference on Learning Theory*, pages 4532–4576. PMLR, 2021a.

Dongruo Zhou, Jiafan He, and Quanquan Gu. Provably efficient reinforcement learning for discounted mdps with feature mapping. In *International Conference on Machine Learning*, pages 12793–12802. PMLR, 2021b.

Ingvar Ziemann and Henrik Sandberg. On uninformative optimal policies in adaptive lqr with unknown b-matrix. In *Learning for Dynamics and Control*, pages 213–226. PMLR, 2021.

DEFERRED PROOFS

Appendix A

Deferred Proofs from Chapter 3

A.1 Deferred Proofs from Section 3.4

Lemma A.1.1. *If $T' \geq T_{\mathrm{ss}}(\zeta, k, \|x_0\|_2)$ for $T_{\mathrm{ss}}(\zeta, k, \|x_0\|_2)$ as defined in (3.4.1), we have that:*

$$\frac{1}{k}\left|\sum_{t=T'}^{T'+k-1}(v^\top x_t^{\boldsymbol{u}})^2 - v^\top \boldsymbol{\Lambda}_k^{\mathrm{freq}}(A_\star, \boldsymbol{u})v\right| \leq \zeta.$$

Proof. First note that, for any $t_0 \geq 0$, $v^\top \boldsymbol{\Lambda}_k^{\mathrm{freq}}(A_\star, \boldsymbol{u})v = \sum_{t=t_0}^{t_0+k-1}(v^\top x_t^{\boldsymbol{u},\mathrm{ss}})^2$. Thus:

$$\begin{aligned}
&\left|\sum_{t=T'}^{T'+k-1}(v^\top x_t^{\boldsymbol{u}})^2 - v^\top \boldsymbol{\Lambda}_k^{\mathrm{freq}}(A_\star, \boldsymbol{u})v\right| \\
&= \left|\sum_{t=T'}^{T'+k-1}\left(v^\top x_t^{\boldsymbol{u},\mathrm{ss}} + v^\top A_\star^t(x_0 - x_0^{\boldsymbol{u},\mathrm{ss}})\right)^2 - v^\top \boldsymbol{\Lambda}_k^{\mathrm{freq}}(A_\star, \boldsymbol{u})v\right| \\
&= \left|\sum_{t=T'}^{T'+k-1}\left(v^\top A_\star^t(x_0 - x_0^{\boldsymbol{u},\mathrm{ss}})\right)^2 + 2\sum_{t=T'}^{T'+k-1}\left(v^\top x_t^{\boldsymbol{u},\mathrm{ss}}\right)\left(v^\top A_\star^t(x_0 - x_0^{\boldsymbol{u},\mathrm{ss}})\right)\right| \\
&\leq \sum_{t=T'}^{T'+k-1}\left(v^\top A_\star^t(x_0 - x_0^{\boldsymbol{u},\mathrm{ss}})\right)^2 + 2\sqrt{\sum_{t=T'}^{T'+k-1}(v^\top x_t^{\boldsymbol{u},\mathrm{ss}})^2}\sqrt{\sum_{t=T'}^{T'+k-1}(v^\top A_\star^t(x_0 - x_0^{\boldsymbol{u},\mathrm{ss}}))^2} \\
&= \sum_{t=T'}^{T'+k-1}\left(v^\top A_\star^t(x_0 - x_0^{\boldsymbol{u},\mathrm{ss}})\right)^2 + 2\sqrt{v^\top \boldsymbol{\Lambda}_k^{\mathrm{freq}}(A_\star, \boldsymbol{u})v}\sqrt{\sum_{t=T'}^{T'+k-1}(v^\top A_\star^t(x_0 - x_0^{\boldsymbol{u},\mathrm{ss}}))^2} \\
&\leq \|x_0 - x_0^{\boldsymbol{u},\mathrm{ss}}\|_2^2\sum_{t=T'}^{T'+k-1}\|A_\star^t\|_2^2 + 2\|x_0 - x_0^{\boldsymbol{u},\mathrm{ss}}\|_2\sqrt{v^\top \boldsymbol{\Lambda}_k^{\mathrm{freq}}(A_\star, \boldsymbol{u})v}\sqrt{\sum_{t=T'}^{T'+k-1}\|A_\star^t\|_2^2} \\
&\leq \|x_0 - x_0^{\boldsymbol{u},\mathrm{ss}}\|_2^2\tau_\star^2\rho_\star^{2T'}\sum_{t=0}^{k-1}\rho_\star^{2t} + 2\|x_0 - x_0^{\boldsymbol{u},\mathrm{ss}}\|_2\tau_\star\rho_\star^{T'}\sqrt{v^\top \boldsymbol{\Lambda}_k^{\mathrm{freq}}(A_\star, \boldsymbol{u})v}\sqrt{\sum_{t=0}^{k-1}\rho_\star^{2t}}
\end{aligned}$$

$$\leq \frac{\|x_0 - x_0^{\boldsymbol{u},\mathrm{ss}}\|_2^2 \tau_\star^2 \rho_\star^{2T'}}{1-\rho_\star^2} + \frac{2\|x_0 - x_0^{\boldsymbol{u},\mathrm{ss}}\|_2 \tau_\star \sqrt{v^\top \boldsymbol{\Lambda}_k^{\mathrm{freq}}(A_\star, \boldsymbol{u}) v} \rho_\star^{T'}}{\sqrt{1-\rho_\star^2}}.$$

If we can ensure that T' is large enough that this is bounded as $k\zeta$, the result will follow. By Lemma A.1.2, we can upper bound

$$\|x_0^{\boldsymbol{u},\mathrm{ss}}\|_2 \leq \frac{2\tau_\star \|B_\star\|_2 k\gamma}{1-\rho_\star^k} \leq \frac{2\tau_\star \|B_\star\|_2 k\gamma}{1-\rho_\star}$$

so it follows that

$$\frac{\|x_0 - x_0^{\boldsymbol{u},\mathrm{ss}}\|_2^2 \tau_\star^2 \rho_\star^{2T'}}{1-\rho_\star^2} \leq \frac{\tau_\star^2}{1-\rho_\star}\left(\|x_0\|_2 + \frac{2\tau_\star \|B_\star\|_2 k\gamma}{1-\rho_\star}\right)^2 \cdot \rho_\star^{2T'}$$

and so we can bound this all by $k\zeta/2$ as long as

$$T' \geq \frac{1}{2\log\frac{1}{\rho_\star}} \cdot \log\left[\frac{2}{k\zeta} \cdot \frac{\tau_\star^2}{1-\rho_\star}\left(\|x_0\|_2 + \frac{2\tau_\star \|B_\star\|_2 k\gamma}{1-\rho_\star}\right)^2\right]$$

Next, we can bound:

$$\begin{aligned}
v^\top \boldsymbol{\Lambda}_k^{\mathrm{freq}}(A_\star, \boldsymbol{u}) v \leq \left\|\boldsymbol{\Lambda}_k^{\mathrm{freq}}(A_\star, \boldsymbol{u})\right\|_{\mathrm{op}} &= \frac{1}{k}\left\|\sum_{\ell=1}^k (e^{\iota\frac{2\pi\ell}{k}} I - A_\star)^{-1} B_\star \check{u}_\ell \check{u}_\ell^{\mathsf{H}} B_\star^{\mathsf{H}} (e^{\iota\frac{2\pi\ell}{k}} I - A_\star)^{-\mathsf{H}}\right\|_{\mathrm{op}} \\
&\leq \frac{1}{k}\left(\max_{\ell=1,\ldots,k} \|(e^{\iota\frac{2\pi\ell}{k}} I - A_\star)^{-1} B_\star\|_2^2\right) \sum_{\ell=1}^k \check{u}_\ell^{\mathsf{H}} \check{u}_\ell \\
&\leq k\gamma^2 \left(\max_{\ell=1,\ldots,k} \|(e^{\iota\frac{2\pi\ell}{k}} I - A_\star)^{-1} B_\star\|_2^2\right) \\
&\leq k\gamma^2 \|A_\star\|_{\mathcal{H}_\infty}^2 \|B_\star\|_{\mathrm{op}}^2
\end{aligned}$$

where we have used that, by our assumption that $\sum_{t=0}^{k-1} u_t^\top u_t \leq k\gamma^2$ and Parseval's Theorem, $\sum_{\ell=1}^k \check{u}_\ell^{\mathsf{H}} \check{u}_\ell \leq k^2\gamma^2$. It follows then that

$$\frac{2\|x_0 - x_0^{\boldsymbol{u},\mathrm{ss}}\|_2 \tau_\star \sqrt{v^\top \boldsymbol{\Lambda}_k^{\mathrm{freq}}(A_\star, \boldsymbol{u}) v} \rho_\star^{T'}}{\sqrt{1-\rho_\star^2}} \leq \frac{2\tau_\star \sqrt{k}\gamma \|A_\star\|_{\mathcal{H}_\infty} \|B_\star\|_{\mathrm{op}}}{\sqrt{1-\rho_\star}} \cdot \left(\|x_0\|_2 + \frac{2\tau_\star \|B_\star\|_2 k\gamma}{1-\rho_\star}\right) \cdot \rho_\star^{T'}.$$

We can bound all of this by $k\zeta/2$ as long as

$$T' \geq \frac{1}{\log\frac{1}{\rho_\star}} \cdot \log\left[\frac{2}{k\zeta} \cdot \frac{2\tau_\star \sqrt{k}\gamma \|A_\star\|_{\mathcal{H}_\infty} \|B_\star\|_{\mathrm{op}}}{\sqrt{1-\rho_\star}} \cdot \left(\|x_0\|_2 + \frac{2\tau_\star \|B_\star\|_2 k\gamma}{1-\rho_\star}\right)\right].$$

Note that, for $t \in [0,1]$, we have $\log(1-t) \leq -t$, which implies that $\log \frac{1}{\rho_\star} \geq 1 - \rho_\star$. Using this and some algebra, we see that each of these conditions on T' hold as long as $T' \geq T_{\text{ss}}(\zeta, k, \|x_0\|)$ for T_{ss} as defined the statement of result. □

Lemma A.1.2. *For all $t \geq 0$, if $x_0 = 0$: $\|x_t^{\boldsymbol{u}}\|_2 \leq \frac{2\tau_\star \|B_\star\|_2 k\gamma}{1-\rho_\star^k}$.*

Proof. We have:

$$\begin{aligned}
\|x_t^{\boldsymbol{u}}\|_2 &= \left\| \sum_{s=0}^{t-1} A_\star^{t-s-1} B_\star u_s \right\|_2 \leq \tau_\star \|B_\star\|_{\text{op}} \sum_{s=0}^{t-1} \rho_\star^{t-s-1} \|u_s\|_2 \\
&\leq \tau_\star \|B_\star\|_{\text{op}} \sum_{\ell=0}^{\lceil t/k \rceil - 2} \rho_\star^{t-k(\ell+1)-1} \sum_{s=k\ell}^{k(\ell+1)-1} \|u_s\|_2 + \tau_\star \|B_\star\|_{\text{op}} \sum_{s=(\lceil t/k \rceil - 1)k}^{t} \|u_s\|_2
\end{aligned}$$

By construction, we will have that $\sum_{s=k\ell}^{k(\ell+1)-1} \|u_s\|_2^2 \leq k\gamma^2$ so long as ℓ is large enough that $k\ell$ is in epoch i. However, since k is doubled at each epoch, this sum will contain an integer multiple of the period of the input regardless what the value of ℓ is, so we see that this inequality will hold for all values of ℓ. This implies that for all ℓ, $\sum_{s=k\ell}^{k(\ell+1)-1} \|u_s\|_2 \leq \sqrt{k}\sqrt{\sum_{s=k\ell}^{k(\ell+1)-1} \|u_s\|_2^2} \leq k\gamma$. So:

$$\begin{aligned}
&\tau_\star \|B_\star\|_{\text{op}} \sum_{\ell=0}^{\lceil t/k \rceil - 2} \rho_\star^{t-k(\ell+1)-1} \sum_{s=k\ell}^{k(\ell+1)-1} \|u_s\|_2 + \tau_\star \|B_\star\|_{\text{op}} \sum_{s=(\lceil t/k \rceil - 1)k}^{t} \|u_s\|_2 \\
&\leq \tau_\star \|B_\star\|_{\text{op}} k\gamma \sum_{\ell=0}^{\lceil t/k \rceil - 2} \rho_\star^{t-k(\ell+1)-1} + \tau_\star \|B_\star\|_{\text{op}} k\gamma \\
&= \tau_\star \|B_\star\|_{\text{op}} k\gamma \frac{\rho_\star^{t-1}}{\rho_\star^k} \frac{\frac{1}{\rho_\star^{k(\lceil t/k \rceil - 2)}} - \rho_\star^k}{1-\rho_\star^k} + \tau_\star \|B_\star\|_{\text{op}} k\gamma \\
&= \tau_\star \|B_\star\|_{\text{op}} k\gamma \frac{\frac{\rho_\star^k}{\rho_\star^{k\lceil t/k \rceil - t + 1}} - \rho_\star^{t-1}}{1-\rho_\star^k} + \tau_\star \|B_\star\|_{\text{op}} k\gamma \\
&\leq \frac{\tau_\star \|B_\star\|_{\text{op}} k\gamma}{1-\rho_\star^k} + \tau_\star \|B_\star\|_{\text{op}} k\gamma \\
&\leq \frac{2\tau_\star \|B_\star\|_{\text{op}} k\gamma}{1-\rho_\star^k}
\end{aligned}$$

where the last inequality holds since if t is divisible by k, $k\lceil t/k \rceil - t + 1 = 1$ so $\frac{\rho_\star^k}{\rho_\star^{k\lceil t/k \rceil - t+1}} \leq 1$, and if t is not divisible by k, $k\lceil t/k \rceil - t + 1 < k(t/k+1) - t + 1 = k+1$, and since $k\lceil t/k \rceil - t + 1$ is an integer, it follows that $\frac{\rho_\star^k}{\rho_\star^{k\lceil t/k \rceil - t+1}} \leq 1$.

□

A.2 Deferred Proofs from Section 3.5

Proof of Lemma 3.5.1. This follows directly by the formal definition of Algorithm 3.1. In particular, we see that at each epoch i, Algorithm 3.1 plays open-loop inputs $\widetilde{u}_t$ that are $\mathcal{F}_{\bar{t}_i}$ measurable. Furthermore, $\bar{t}_i$ and $\Gamma_{u,i}$ are deterministically specified at the start of the algorithm, $\mathrm{tr}(\Gamma_{u,i}) \le \gamma^2$, $\lambda_{\min}(\Gamma_{u,i}) \ge \gamma^2/(2d_u)$, and Lemma 3.5.3 gives $\sum_{t=0}^{T-1} \widetilde{u}_t^\top \widetilde{u}_t \le \gamma^2$ deterministically. The fact that $n = \mathcal{O}(\log T)$ follows since we increase the epoch length exponentially. Finally, the low-switching condition follows since the length of the epochs increase exponentially—once T is large enough that $T \ge T_{\mathrm{se}}(\pi_{\mathrm{exp}})$, we will have that at least half the initial interval is contained in the final epoch. Then for subsequent epochs, any interval of length $T_{\mathrm{se}}(\pi_{\mathrm{exp}})$ will contain at most one epoch boundary. □

Proof of Lemma 3.5.2. **Optimality of Inputs.** We apply Lemmas A.2.1 and A.2.3 which give that

$$\left|\lambda_{\min}\left(\frac{1}{k}\mathbf{\Lambda}_k^{\mathrm{freq}}(A_\star, \boldsymbol{U}^\star) + \mathbf{\Lambda}_k^{\mathrm{noise}}(A_\star, \sigma_u)\right) - \lambda_{\min}\left(\frac{1}{k}\mathbf{\Lambda}_k^{\mathrm{freq}}(A_\star, \widehat{\boldsymbol{U}}) + \mathbf{\Lambda}_k^{\mathrm{noise}}(A_\star, \sigma_u)\right)\right|$$
$$\le 32\gamma^2 \|A_\star\|_{\mathcal{H}_\infty}^3 \|B_\star\|_{\mathrm{op}}^2 \cdot \epsilon + 2\|\mathbf{\Lambda}_k^{\mathrm{noise}}(A_\star, \sigma_u) - \mathbf{\Lambda}_k^{\mathrm{noise}}(\widehat{A}, \sigma_u)\|_{\mathrm{op}} + \frac{\|\theta_\star\|_{\mathcal{H}_\infty}\|A_\star\|_{\mathcal{H}_\infty}^2\|B_\star\|_{\mathrm{op}}\gamma^2}{k}.$$

Furthermore, by Lemma C.1.1, we can bound

$$\|\mathbf{\Lambda}_k^{\mathrm{noise}}(A_\star, \sigma_u) - \mathbf{\Lambda}_k^{\mathrm{noise}}(\widehat{A}, \sigma_u)\|_{\mathrm{op}} \le \left(\frac{8(\sigma_w^2 + \sigma_u^2\|B_\star\|_{\mathrm{op}}^2)\tau_\star^3}{(1-\rho_\star^2)^2} + \frac{4\sigma_u^2\|B_\star\|_{\mathrm{op}}\tau_\star^2}{1-\rho_\star^2}\right)\epsilon + \frac{2\sigma_u^2\tau_\star^2}{1-\rho_\star^2}\epsilon^2.$$

By our condition on ϵ and k, we therefore have that

$$32\epsilon\gamma^2\|A_\star\|_{\mathcal{H}_\infty}^3\|B_\star\|_{\mathrm{op}}^2 + 2\|\mathbf{\Lambda}_k^{\mathrm{noise}}(A_\star, \sigma_u) - \mathbf{\Lambda}_k^{\mathrm{noise}}(\widehat{A}, \sigma_u)\|_{\mathrm{op}} + \frac{\|\theta_\star\|_{\mathcal{H}_\infty}\|A_\star\|_{\mathcal{H}_\infty}^2\|B_\star\|_{\mathrm{op}}\gamma^2}{k} \le \frac{1}{2}\lambda_{\mathrm{noise}}(\sigma_u).$$

As $\lambda_{\min}\left(\frac{1}{k}\mathbf{\Lambda}_k^{\mathrm{freq}}(A_\star, \boldsymbol{U}) + \mathbf{\Lambda}_k^{\mathrm{noise}}(A_\star, \sigma_u)\right) \ge \lambda_{\mathrm{noise}}(\sigma_u)$ for any $\boldsymbol{U}$, we therefore have

$$\lambda_{\min}\left(\frac{1}{k}\mathbf{\Lambda}_k^{\mathrm{freq}}(A_\star, \widehat{\boldsymbol{U}}) + \mathbf{\Lambda}_k^{\mathrm{noise}}(A_\star, \sigma_u)\right) \ge \frac{1}{2} \cdot \max_{\boldsymbol{U} \in \mathcal{U}_{\gamma^2,k}} \lambda_{\min}\left(\lambda_{\min}\left(\frac{1}{k}\mathbf{\Lambda}_k^{\mathrm{freq}}(A_\star, \boldsymbol{U}) + \mathbf{\Lambda}_k^{\mathrm{noise}}(A_\star, \sigma_u)\right)\right).$$

Infinite-Horizon Approximation. By Lemma 5.3.8, as long as

$$k \ge \max\left\{\frac{16\pi\|\theta_\star\|_{\mathcal{H}_\infty}\gamma^2}{\lambda_{\mathrm{noise}}(\sigma_u)}, \frac{\pi}{2\|\theta_\star\|_{\mathcal{H}_\infty}}\right\}\left(\max_{\omega\in[0,2\pi]}\|(e^{\iota\omega}I - A_\star)^{-2}B_\star\|_{\mathrm{op}}\right),$$

then for any $\bar{k}$ and input $\boldsymbol{U} \in \mathcal{U}_{\gamma^2,\bar{k}}$, we have that there exists an input $\boldsymbol{U}' \in \mathcal{U}_{\gamma^2,k}$ such that

$$\frac{1}{k}\mathbf{\Lambda}_k^{\mathrm{freq}}(A_\star, \boldsymbol{U}') \succeq \frac{1}{\bar{k}}\mathbf{\Lambda}_{\bar{k}}^{\mathrm{freq}}(A_\star, \boldsymbol{U}) - \frac{1}{4}\lambda_{\mathrm{noise}}(\sigma_u) \cdot I.$$

Furthermore, by Lemma A.2.4, as long as

$$k \geq \frac{1}{2(1-\rho_\star)} \cdot \log\left(\frac{2(\sigma_w^2\tau_\star^2 + \sigma_u^2\tau_\star^2\|B_\star\|_{\mathrm{op}}^2)}{\lambda_{\mathrm{noise}}(\sigma_u)(1-\rho_\star^2)}\right),$$

we have that $\mathbf{\Lambda}_k^{\mathrm{noise}}(A_\star, \sigma_u) \succeq \frac{1}{2}\mathbf{\Lambda}_{\bar{k}}^{\mathrm{noise}}(A_\star, \sigma_u)$. This implies that, for any $\bar{k} \geq d_x$:

$$\begin{aligned}
&\max_{\boldsymbol{U}\in\mathcal{U}_{\gamma^2,k}} \lambda_{\min}\left(\frac{1}{k}\mathbf{\Lambda}_k^{\mathrm{freq}}(A_\star, \boldsymbol{U}) + \mathbf{\Lambda}_k^{\mathrm{noise}}(A_\star, \sigma_u)\right) \\
&\geq \max_{\boldsymbol{U}\in\mathcal{U}_{\gamma^2,\bar{k}}} \lambda_{\min}\left(\frac{1}{\bar{k}}\mathbf{\Lambda}_{\bar{k}}^{\mathrm{freq}}(A_\star, \boldsymbol{U}) - \frac{1}{4}\lambda_{\mathrm{noise}}(\sigma_u)\cdot I + \frac{1}{2}\mathbf{\Lambda}_{\bar{k}}^{\mathrm{noise}}(A_\star, \sigma_u)\right) \\
&\geq \max_{\boldsymbol{U}\in\mathcal{U}_{\gamma^2,\bar{k}}} \lambda_{\min}\left(\frac{1}{\bar{k}}\mathbf{\Lambda}_{\bar{k}}^{\mathrm{freq}}(A_\star, \boldsymbol{U}) + \frac{1}{4}\mathbf{\Lambda}_{\bar{k}}^{\mathrm{noise}}(A_\star, \sigma_u)\right) \\
&\geq \frac{1}{4}\cdot \max_{\boldsymbol{U}\in\mathcal{U}_{\gamma^2,\bar{k}}} \lambda_{\min}\left(\frac{1}{\bar{k}}\mathbf{\Lambda}_{\bar{k}}^{\mathrm{freq}}(A_\star, \boldsymbol{U}) + \mathbf{\Lambda}_{\bar{k}}^{\mathrm{noise}}(A_\star, \sigma_u)\right)
\end{aligned}$$

where the second inequality follows since $\mathbf{\Lambda}_{\bar{k}}^{\mathrm{noise}}(A_\star, \sigma_u) \succeq \lambda_{\mathrm{noise}}(\sigma_u)\cdot I$. The result then follows since $\mathbf{\Lambda}_{\bar{k}}^{\mathrm{noise}}(A_\star, \sigma_u) \succeq \mathbf{\Lambda}_{\bar{k}}^{\mathrm{noise}}(A_\star, 0)$, and bounding $\max_{\omega\in[0,2\pi]} \|(e^{\iota\omega}I - A_\star)^{-2}B_\star\|_{\mathrm{op}} \leq \|A_\star\|_{\mathcal{H}_\infty}^2\|B_\star\|_{\mathrm{op}}$. □

Lemma A.2.1. *Assuming that $\|A_\star - \widehat{A}\|_{\mathrm{op}} \leq \epsilon$ for ϵ satisfying $\epsilon \leq \frac{1}{2\|A_\star\|_{\mathcal{H}_\infty}}$, then we have:*

$$\left|\lambda_{\min}\left(\frac{1}{k}\mathbf{\Lambda}_k^{\mathrm{freq}}(A_\star, \boldsymbol{U}^\star) + M\right) - \lambda_{\min}\left(\frac{1}{k}\mathbf{\Lambda}_k^{\mathrm{freq}}(A_\star, \widehat{\boldsymbol{U}}) + M\right)\right| \leq 32\gamma^2\|A_\star\|_{\mathcal{H}_\infty}^3\|B_\star\|_{\mathrm{op}}^2\cdot\epsilon + 2\|M - \widehat{M}\|_{\mathrm{op}}$$

where $\boldsymbol{U}^\star = \arg\max_{\boldsymbol{U}\in\mathcal{U}_{\gamma^2,k}} \lambda_{\min}(\frac{1}{k}\mathbf{\Lambda}_k^{\mathrm{freq}}(A_\star, \boldsymbol{U}) + M)$ and $\widehat{\boldsymbol{U}} = \arg\max_{\boldsymbol{U}\in\mathcal{U}_{\gamma^2,k}} \lambda_{\min}(\frac{1}{k}\mathbf{\Lambda}_k^{\mathrm{freq}}(\widehat{A}, \boldsymbol{U}) + \widehat{M})$.

Proof. Consider some functions $f(x), g(x)$, and assume that $|f(x) - g(x)| \leq \epsilon$ for all x. Let $x_f = \arg\min_{x\in\mathcal{X}} f(x)$. Then

$$g(x_f) \geq \max_{x\in\mathcal{X}} f(x) - \epsilon \geq \max_{x\in\mathcal{X}} g(x) - 2\epsilon.$$

Since $g(x_f) \leq \max_{x\in\mathcal{X}} g(x)$ by definition, we have $|g(x_f) - \max_{x\in\mathcal{X}} g(x)| \leq 2\epsilon$. Now consider any matrices $X, Y \succeq 0$, and let v and u denote the minimum eigenvalues of X and Y, respectively. Assume without loss of generality that $\lambda_{\min}(X) \geq \lambda_{\min}(Y)$, and note that $u^\top X u \geq v^\top X v = \lambda_{\min}(X)$.

$$|\lambda_{\min}(X) - \lambda_{\min}(Y)| = v^\top X v - u^\top Y u \leq u^\top X u - u^\top Y u \leq \sup_{v\in\mathcal{S}^{d-1}} |v^\top(X-Y)v|.$$

We now instantiate the above with $X(\boldsymbol{U}) = \frac{1}{k}\mathbf{\Lambda}_k^{\mathrm{freq}}(A_\star, \boldsymbol{U}) + M$ and $Y(\boldsymbol{U}) = \frac{1}{k}\mathbf{\Lambda}_k^{\mathrm{freq}}(\widehat{A}, \boldsymbol{U}) +$

$\widehat{M}$, $g(\boldsymbol{U}) = \lambda_{\min}(X(\boldsymbol{U}))$, and $f(\boldsymbol{U}) = \lambda_{\min}(Y(\boldsymbol{U}))$. The above observations give that, for any $\boldsymbol{U} \in \mathcal{U}_{\gamma^2,k}$:

$$\begin{aligned}|g(\boldsymbol{U}) - f(\boldsymbol{U})| &\leq \sup_{v \in \mathcal{S}^{d-1}} \left| v^\top \left(\frac{1}{k} \boldsymbol{\Lambda}_k^{\text{freq}}(A_\star, \boldsymbol{U}) + M \right) v - v^\top \left(\frac{1}{k} \boldsymbol{\Lambda}_k^{\text{freq}}(\widehat{A}, \boldsymbol{U}) + \widehat{M} \right) v \right| \\ &\leq 16\gamma^2 \|A_\star\|_{\mathcal{H}_\infty}^3 \|B_\star\|_{\text{op}}^2 \cdot \epsilon + \|M - \widehat{M}\|_{\text{op}}\end{aligned}$$

where the last inequality follows from Lemma A.2.2. The result follows. □

Lemma A.2.2. *If $\|A_\star - \widehat{A}\|_{\text{op}} \leq \epsilon$ for ϵ satisfying $\epsilon \leq \frac{1}{2\|A_\star\|_{\mathcal{H}_\infty}}$, we have that for any $v \in \mathcal{S}^{d-1}$ and $\boldsymbol{U} \in \mathcal{U}_{\gamma^2,k}$:*

$$\left| v^\top \boldsymbol{\Lambda}_k^{\text{freq}}(A_\star, \boldsymbol{U}) v - v^\top \boldsymbol{\Lambda}_k^{\text{freq}}(\widehat{A}, \boldsymbol{U}) v \right| \leq 16k\gamma^2 \|A_\star\|_{\mathcal{H}_\infty}^3 \|B_\star\|_{\text{op}}^2 \cdot \epsilon.$$

Proof. To bound $\left| v^\top \boldsymbol{\Lambda}_k^{\text{freq}}(A_\star, \boldsymbol{U}) v - v^\top \boldsymbol{\Lambda}_k^{\text{freq}}(\widehat{A}, \boldsymbol{U}) v \right|$, we calculate the directional derivative of $v^\top \boldsymbol{\Lambda}_k^{\text{freq}}(A, \boldsymbol{U}) v$ with respect to A and use this to bound the Lipschitz constant of the function $v^\top \boldsymbol{\Lambda}_k^{\text{freq}}(A, \boldsymbol{U}) v$ at $A \leftarrow A_\star$. The directional derivative is given by, for $\Delta \in \mathbb{R}^{d_x \times d_x}$, $\|\Delta\|_{\text{op}} = 1$:

$$D[v^\top \boldsymbol{\Lambda}_k^{\text{freq}}(A_\star, \boldsymbol{U}) v][\Delta] = \lim_{\delta \to 0} \frac{v^\top \boldsymbol{\Lambda}_k^{\text{freq}}(A_\star + \delta\Delta, \boldsymbol{U}) v - v^\top \boldsymbol{\Lambda}_k^{\text{freq}}(A_\star, \boldsymbol{U}) v}{|\delta|}$$

Lemma A.4.1 gives that, for small enough δ:

$$(e^{\iota\omega} I - A_\star - \delta\Delta)^{-1} = \sum_{s=0}^{\infty} (e^{\iota\omega} I - A_\star)^{-1} (\delta\Delta (e^{\iota\omega} I - A_\star)^{-1})^s$$

so:

$$\begin{aligned}v^\top \boldsymbol{\Lambda}_k^{\text{freq}}(A_\star + \delta\Delta, \boldsymbol{U}) v &= \frac{1}{k} \cdot v^\top \left(\sum_{\ell=1}^{k} (e^{\iota \frac{2\pi\ell}{k}} I - A_\star - \delta\Delta)^{-1} B_\star U_\ell B_\star^{\mathsf{H}} (e^{\iota \frac{2\pi\ell}{k}} I - A_\star - \delta\Delta)^{-\mathsf{H}} \right) v \\ &= \frac{1}{k} \cdot v^\top \left(\sum_{\ell=1} (e^{\iota \frac{2\pi\ell}{k}} I - A_\star)^{-1} B_\star U_\ell B_\star^{\mathsf{H}} (e^{\iota \frac{2\pi\ell}{k}} I - A_\star)^{-\mathsf{H}} \right) v \\ &\quad + \frac{1}{k} \cdot 2\delta v^\top \left(\sum_{\ell=1}^{k} (e^{\iota \frac{2\pi\ell}{k}} I - A_\star)^{-1} \Delta (e^{\iota \frac{2\pi\ell}{k}} I - A_\star)^{-1} B_\star U_\ell B_\star^{\mathsf{H}} (e^{\iota \frac{2\pi\ell}{k}} I - A_\star)^{-\mathsf{H}} \right) v \\ &\quad + O(\delta^2)\end{aligned}$$

and thus:

$$\lim_{\delta\to 0}\frac{v^\top \mathbf{\Lambda}_k^{\text{freq}}(A_\star+\delta\Delta,\boldsymbol{U})v - v^\top \mathbf{\Lambda}_k^{\text{freq}}(A_\star,\boldsymbol{U})v}{|\delta|}$$
$$= \frac{1}{k}\cdot 2v^\top\left(\sum_{\ell=1}^{k}(e^{\iota\frac{2\pi\ell}{k}}I - A_\star)^{-1}\Delta(e^{\iota\frac{2\pi\ell}{k}}I - A_\star)^{-1}B_\star U_\ell B_\star^{\mathsf{H}}(e^{\iota\frac{2\pi\ell}{k}}I - A_\star)^{-\mathsf{H}}\right)v.$$

By the Mean Value Theorem, and since $\|A_\star - \widehat{A}\|_{\text{op}} \le \epsilon$ by assumption, we then have

$$\left|v^\top \mathbf{\Lambda}_k^{\text{freq}}(A_\star,\boldsymbol{U})v - v^\top \mathbf{\Lambda}_k^{\text{freq}}(\widehat{A},\boldsymbol{U})v\right|$$
$$\le \max_{\substack{\delta\in[0,\epsilon]\\ \Delta\in\mathbb{R}^{d_x\times d_x}\\ \|\Delta\|_{\text{op}}=1}} \frac{2\epsilon}{k}\left|v^\top\left(\sum_{\ell=1}^{k}(e^{\iota\frac{2\pi\ell}{k}}I - A_\star - \delta\Delta)^{-1}\Delta(e^{\iota\frac{2\pi\ell}{k}}I - A_\star - \delta\Delta)^{-1}B_\star U_\ell B_\star^{\mathsf{H}}(e^{\iota\frac{2\pi\ell}{k}}I - A_\star - \delta\Delta)^{-\mathsf{H}}\right)v\right|.$$

Bounding the Gradient Magnitude. For any $v\in\mathcal{S}^{d-1}$ and $\boldsymbol{U}\in\mathcal{U}_{\gamma^2,k}$, we can bound

$$\left|v^\top\left(\sum_{\ell=1}^{k}(e^{\iota\frac{2\pi\ell}{k}}I - A_\star - \delta\Delta)^{-1}\Delta(e^{\iota\frac{2\pi\ell}{k}}I - A_\star - \delta\Delta)^{-1}B_\star U_\ell B_\star^{\mathsf{H}}(e^{\iota\frac{2\pi\ell}{k}}I - A_\star - \delta\Delta)^{-\mathsf{H}}\right)v\right|$$
$$\le \sum_{\ell=1}^{k}\|(e^{\iota\frac{2\pi\ell}{k}}I - A_\star - \delta\Delta)^{-1}\|_2^3\|B_\star\|_2^2\|U_\ell\|_{\text{op}}$$
$$\le \left(\max_{\ell\in[k]}\|(e^{\iota\frac{2\pi\ell}{k}}I - A_\star - \delta\Delta)^{-1}\|_2^3\|B_\star\|_2^2\right)\cdot\sum_{\ell=1}^{k}\text{tr}(U_\ell)$$
$$\le \max_{\ell\in[k]}k^2\gamma^2\|(e^{\iota\frac{2\pi\ell}{k}}I - A_\star - \delta\Delta)^{-1}\|_2^3\|B_\star\|_2^2$$

By Lemma A.4.1 and our condition on ϵ we have that:

$$(e^{\iota\omega}I - A_\star - \delta\Delta)^{-1} = \sum_{s=0}^{\infty}(e^{\iota\omega}I - A_\star)^{-1}(\delta\Delta(e^{\iota\omega}I - A_\star)^{-1})^s.$$

Thus, for $\delta\in[0,\epsilon]$ and Δ with $\|\Delta\|_{\text{op}} = 1$, we have

$$\|(e^{\iota\omega}I - A_\star - \delta\Delta)^{-1}\|_2 \le \|(e^{\iota\omega}I - A_\star)^{-1}\|_{\text{op}}\cdot\sum_{s=0}^{\infty}\epsilon^s\|(e^{\iota\omega}I - A_\star)^{-1}\|_{\text{op}}^s$$
$$\le \|(e^{\iota\omega}I - A_\star)^{-1}\|_{\text{op}}\cdot\sum_{s=0}^{\infty}\frac{1}{2^s}$$
$$= 2\|(e^{\iota\omega}I - A_\star)^{-1}\|_{\text{op}}$$

where the second inequality follows from the assumption that $\epsilon \leq \frac{1}{2\|A_\star\|_{\mathcal{H}_\infty}}$. Combining this with the above yields the result.

□

Lemma A.2.3. *For any $\boldsymbol{U} \in \mathcal{U}_{\gamma^2,k}$ and $\theta = (A, B)$ with $\rho(A) < 1$, there exists some $\boldsymbol{U}' \in \mathcal{U}^0_{\gamma^2,k}$ such that*

$$\left\| \frac{1}{k}\boldsymbol{\Lambda}_k^{\text{freq}}(A, \boldsymbol{U}') - \frac{1}{k}\boldsymbol{\Lambda}_k^{\text{freq}}(A, \boldsymbol{U}) \right\|_{\text{op}} \leq \|\theta\|_{\mathcal{H}_\infty}\|A\|^2_{\mathcal{H}_\infty}\|B\|_{\text{op}}\gamma^2 \cdot \frac{1}{k}.$$

Proof. Let $\boldsymbol{U}'$ be defined as

$$\boldsymbol{U}'_1 = \boldsymbol{U}_1 + \frac{1}{2}\boldsymbol{U}_k, \quad \boldsymbol{U}'_{k-1} = \boldsymbol{U}_{k-1} + \frac{1}{2}\boldsymbol{U}_k, \quad \boldsymbol{U}'_\ell = \boldsymbol{U}_\ell, 1 < \ell < k-1.$$

Clearly, $\boldsymbol{U}' \in \mathcal{U}^0_{\gamma^2,k}$. We then have

$$\begin{aligned}\frac{1}{k}\boldsymbol{\Lambda}_k^{\text{freq}}(A, \boldsymbol{U}') - \frac{1}{k}\boldsymbol{\Lambda}_k^{\text{freq}}(A, \boldsymbol{U}) &= \frac{1}{k^2}\frac{1}{2}(e^{\iota\frac{2\pi(k+1)}{k}}I - A)^{-1}BU_kB^{\mathsf{H}}(e^{\iota\frac{2\pi(k+1)}{k}}I - A)^{-\mathsf{H}} \\ &+ \frac{1}{k^2}\frac{1}{2}(e^{\iota\frac{2\pi(k-1)}{k}}I - A)^{-1}BU_kB^{\mathsf{H}}(e^{\iota\frac{2\pi(k-1)}{k}}I - A)^{-\mathsf{H}} - \frac{1}{k^2}(e^{\iota 2\pi}I - A)^{-1}BU_kB^{\mathsf{H}}(e^{\iota 2\pi}I - A)^{-\mathsf{H}}\end{aligned}$$

where we have used that $e^{\iota\frac{2\pi(k+1)}{k}} = e^{\iota\frac{2\pi}{k}}$. We can bound the operator norm of this difference as

$$\begin{aligned}\Big(&\|(e^{\iota\frac{2\pi(k+1)}{k}}I - A)^{-1}B - (e^{\iota 2\pi}I - A)^{-1}B\|_{\text{op}} \\ &+ \|(e^{\iota\frac{2\pi(k-1)}{k}}I - A)^{-1}B - (e^{\iota 2\pi}I - A)^{-1}B\|_{\text{op}}\Big)\frac{\|A\|_{\mathcal{H}_\infty}\|B\|_{\text{op}}\|U_k\|_{\text{op}}}{k^2}.\end{aligned}$$

Since $\boldsymbol{U} \in \mathcal{U}_{\gamma^2,k}$, we have $\|U_k\|_{\text{op}} \leq k^2\gamma^2$. By Lemma A.4.2, we have

$$\begin{aligned}\|(e^{\iota\frac{2\pi(k+1)}{k}}I - A)^{-1}B - (e^{\iota 2\pi}I - A)^{-1}B\|_{\text{op}} &\leq \|A\|_{\mathcal{H}_\infty}\|\theta\|_{\mathcal{H}_\infty} \cdot \frac{1}{k}, \\ \|(e^{\iota\frac{2\pi(k-1)}{k}}I - A)^{-1}B - (e^{\iota 2\pi}I - A)^{-1}B\|_{\text{op}} &\leq \|A\|_{\mathcal{H}_\infty}\|\theta\|_{\mathcal{H}_\infty} \cdot \frac{1}{k}\end{aligned}$$

for $\theta = (A, B)$. Putting this together gives the result. □

Lemma A.2.4. *As long as $k \geq \frac{1}{2(1-\rho_\star)} \cdot \log(\frac{2(\sigma_w^2\tau_\star^2 + \sigma_u^2\tau_\star^2\|B_\star\|^2_{\text{op}})}{\lambda_{\text{noise}}(\sigma_u)(1-\rho_\star^2)})$, we have, for any $\bar{k} \geq 0$:*

$$\boldsymbol{\Lambda}_k^{\text{noise}}(A_\star, \sigma_u) \succeq \frac{1}{2} \cdot \boldsymbol{\Lambda}_{\bar{k}}^{\text{noise}}(A_\star, \sigma_u).$$

Proof. The result is trivial for $k \leq k$, so we assume $\bar{k} > k$. We have

$$\|\sum_{t=0}^{k-1}(A_\star^t)(A_\star^t)^\top - \sum_{t=0}^{\bar{k}-1}(A_\star^t)(A_\star^t)^\top\|_{\text{op}} = \|\sum_{t=k}^{\bar{k}-1}(A_\star^t)(A_\star^t)^\top\|_{\text{op}} \leq \sum_{t=k}^{\bar{k}-1}\|A_\star^t\|_{\text{op}}^2 \leq \sum_{t=k}^{\bar{k}-1}\tau_\star^2 \cdot \rho_\star^{2t} \leq \tau_\star^2 \cdot \frac{\rho_\star^{2k}}{1-\rho_\star^2}.$$

Similarly,

$$\|\sum_{t=0}^{k-1}(A_\star^t)B_\star B_\star^\top(A_\star^t)^\top - \sum_{t=0}^{\bar{k}-1}(A_\star^t)(A_\star^t)^\top\|_{\text{op}} \leq \tau_\star^2\|B_\star\|_{\text{op}}^2 \cdot \frac{\rho_\star^{2k}}{1-\rho_\star^2}.$$

We also have

$$\sum_{t=0}^{k-1}\left[\sigma_w^2(A_\star^t)(A_\star^t)^\top + \sigma_u^2(A_\star^t)B_\star B_\star^\top(A_\star^t)^\top\right] \succeq \lambda_{\text{noise}}(\sigma_u)\cdot I$$

by definition. It follows that if k is large enough that

$$\sigma_w^2\tau_\star^2 \cdot \frac{\rho_\star^{2k}}{1-\rho_\star^2} + \sigma_u^2\tau_\star^2\|B_\star\|_{\text{op}}^2 \cdot \frac{\rho_\star^{2k}}{1-\rho_\star^2} \leq \frac{1}{2}\lambda_{\text{noise}}(\sigma_u),$$

then the conclusion holds. The result follows by rearranging this and noting that $\log\frac{1}{\rho_\star} \geq 1-\rho_\star$. □

A.2.1 Proof of Corollary 3.4

Proof. **Calculation of $\lambda_{\min}^\star$.** The stated rate can be attained by the input $\boldsymbol{u}$ defined as:

$$u_t = \sum_{i=1}^{d_x} a_i v_i \cos\left(\frac{2\pi i}{k}t\right)$$

for $k \geq \mathcal{O}\left(\max_{i=1,\ldots,d_x}\frac{i}{1-\lambda_i}\right)$ and some a_i to be specified satisfying $\sum_{i=1}^d a_i^2 = \gamma^2$. To see this, note that with this input we have

$$\begin{aligned}\mathbf{\Lambda}_k^{\text{freq}}(A_\star, \boldsymbol{u}) &= \sum_{i=1}^{d_x} a_i^2(e^{\iota\frac{2\pi i}{k}}I - A_\star)^{-1}v_i v_i^\top(e^{\iota\frac{2\pi i}{k}}I - A_\star)^{-\mathsf{H}} \\ &= V\left[\sum_{i=1}^{d_x}\frac{a_i^2}{(e^{\iota\frac{2\pi i}{k}}-\lambda_i)(e^{-\iota\frac{2\pi i}{k}}-\lambda_i)}e_i e_i^\top\right]V^\top.\end{aligned}$$

Note that

$$\begin{aligned}(e^{\iota\frac{2\pi i}{k}}-\lambda_i)(e^{-\iota\frac{2\pi i}{k}}-\lambda_i) &= 1+\lambda_i^2-\lambda_i\left(e^{-\iota\frac{2\pi i}{k}}+e^{\iota\frac{2\pi i}{k}}\right)\\ &= 1+\lambda_i^2-2\lambda_i\cos\frac{2\pi i}{k} \approx 1+\lambda_i^2-2\lambda_i\left(1-\frac{2\pi^2 i^2}{k^2}\right)\\ &= (1-\lambda_i)^2+\frac{4\lambda_i\pi^2 i^2}{k^2} = \mathcal{O}\left((1-\lambda_i)^2\right)\end{aligned}$$

where the last equality will hold as long as $\frac{2\sqrt{\lambda_i}\pi i}{1-\lambda_i} \leq k$. Assume that k satisfies this and choose $a_i^2 = \frac{\gamma^2}{\|\mathbf{1}-\lambda\|_2^2}(1-\lambda_i)^2$, then

$$\mathbf{\Lambda}_k^{\text{freq}}(A_\star, \boldsymbol{u}) = \mathcal{O}\left(\frac{\gamma^2}{\|\mathbf{1}-\lambda\|_2^2} V\left[\sum_{i=1}^d e_i e_i^\top\right] V^\top\right) = \mathcal{O}\left(\frac{\gamma^2}{\|\mathbf{1}-\lambda\|_2^2} I\right),$$

and the energy constraint will be satisfied since $\sum_{i=1}^d a_i^2 = \frac{\gamma^2}{\|\mathbf{1}-\lambda\|_2^2}\sum_{i=1}^d (1-\lambda_i)^2 = \gamma^2$. Thus, we have that $\lambda_{\min}(\mathbf{\Lambda}_k^{\text{freq}}(A_\star, \boldsymbol{u})) = \mathcal{O}(\frac{\gamma^2}{\|\mathbf{1}-\lambda\|_2^2})$. Since we have constructed a feasible input with minimum eigenvalue $\mathcal{O}(\frac{\gamma^2}{\|\mathbf{1}-\lambda\|_2^2})$, it follows that $\lambda_{\min}^\star \geq \mathcal{O}(\frac{\gamma^2}{\|\mathbf{1}-\lambda\|_2^2})$.

Calculation of Minimum Eigenvalue with Isotropic Noise. It is straightforward to see that, when playing isotropic noise, $u_t \sim \mathcal{N}(0, \frac{\gamma^2}{d_x} I)$, we have $\lambda_{\min}(\mathbf{\Lambda}_k^{\text{noise}}(A_\star, \frac{\gamma^2}{d_x})) = \mathcal{O}(\frac{\gamma^2/d_x}{1-\lambda_d})$.

Calculation of Minimum Eigenvalue for Optimal Noise. Consider now playing input $u_t \sim \mathcal{N}(0, \Sigma)$, for some Σ. First, note that satisfying the power constraint in this setting is equivalent to $\text{tr}(\Sigma) \leq \gamma^2$. Under this constraint, the optimal noise covariance can be obtained by solving:

$$\max_{\Sigma \succeq 0} \ \lambda_{\min}\left(\sigma^2 \textstyle\sum_{t=0}^k A_\star^t (A_\star^t)^\top + \sum_{t=0}^k A_\star^t B_\star \Sigma B_\star^\top (A_\star^t)^\top\right) \quad \text{s.t.} \quad \text{tr}(\Sigma) \leq \gamma^2.$$

In our setting, with $\gamma^2 \gg \sigma^2$, solving this is approximately equivalent to solving

$$\max_{\widetilde{\Sigma} \succeq 0} \ \lambda_{\min}\left(\textstyle\sum_{t=0}^k \Gamma^t \widetilde{\Sigma} \Gamma^t\right) \quad \text{s.t.} \quad \text{tr}(\widetilde{\Sigma}) \leq \gamma^2$$

where $\widetilde{\Sigma} = V^\top \Sigma V$. Let $\widetilde{\Sigma}^\star$ be the optimal diagonal solution, and note that, in this case, we have

$$\sum_{t=0}^k \Gamma^t \widetilde{\Sigma}^\star \Gamma^t = \frac{\gamma^2}{\sum_{i=1}^d \frac{1-\lambda_i^2}{1-\lambda_i^{2k}}} I.$$

To see this, note that for any diagonal $\widetilde{\Sigma}$ with ith element γ_i^2:

$$\left[\sum_{t=0}^{k} \Gamma^t \widetilde{\Sigma} \Gamma^t\right]_{ii} = \frac{\gamma_i^2(1-\lambda_i^{2k})}{1-\lambda_i^2}.$$

The optimal solution will clearly be the solution that balances the energy in every diagonal element, that is

$$\frac{\gamma_i^2(1-\lambda_i^{2k})}{1-\lambda_i^2} = \frac{\gamma_j^2(1-\lambda_j^{2k})}{1-\lambda_j^2}.$$

for all $i, j \in [d]$, so combining this constraint with the trace constraint yields

$$\frac{\gamma_j^2(1-\lambda_j^{2k})}{1-\lambda_j^2} \sum_{i=1}^{d} \frac{1-\lambda_i^2}{1-\lambda_i^{2k}} = \gamma^2 \implies \gamma_j^2 = \frac{1-\lambda_j^2}{1-\lambda_j^{2k}} \frac{\gamma^2}{\sum_{i=1}^{d} \frac{1-\lambda_i^2}{1-\lambda_i^{2k}}},$$

and thus the jth diagonal element will be $\gamma^2/(\sum_{i=1}^{d} \frac{1-\lambda_i^2}{1-\lambda_i^{2k}})$. Consider now some other matrix Δ that is not necessarily diagonal. Note then that

$$\begin{aligned}
\lambda_{\min}\left(\textstyle\sum_{t=0}^{k} \Gamma^t(\widetilde{\Sigma}^\star + \Delta)\Gamma^t\right) &= \lambda_{\min}\left(\textstyle\sum_{t=0}^{k} \Gamma^t \widetilde{\Sigma}^\star \Gamma^t + \sum_{t=0}^{k} \Gamma^t \Delta \Gamma^t\right) \\
&= \lambda_{\min}\left(\gamma^2 \cdot (\textstyle\sum_{i=1}^{d} \frac{1-\lambda_i^2}{1-\lambda_i^{2k}})^{-1} \cdot I + \sum_{t=0}^{k} \Gamma^t \Delta \Gamma^t\right) \\
&= \gamma^2 \cdot (\textstyle\sum_{i=1}^{d} \frac{1-\lambda_i^2}{1-\lambda_i^{2k}})^{-1} + \lambda_{\min}\left(\textstyle\sum_{t=0}^{k} \Gamma^t \Delta \Gamma^t\right)
\end{aligned}$$

For $\widetilde{\Sigma}^\star + \Delta$ to be in the constraint set, we must have that $\mathrm{tr}(\widetilde{\Sigma}^\star + \Delta) = \gamma^2 + \mathrm{tr}(\Delta) \leq \gamma^2 \implies \mathrm{tr}(\Delta) \leq 0$. To have that

$$\frac{\gamma^2}{\sum_{i=1}^{d} \frac{1-\lambda_i^2}{1-\lambda_i^{2k}}} + \lambda_{\min}\left(\textstyle\sum_{t=0}^{k} \Gamma^t \Delta \Gamma^t\right) \geq \frac{\gamma^2}{\sum_{i=1}^{d} \frac{1-\lambda_i^2}{1-\lambda_i^{2k}}}$$

we must have that $\sum_{t=0}^{k} \Gamma^t \Delta \Gamma^t$ is positive definite. However, this is not possible since the diagonal elements of $\sum_{t=0}^{k} \Gamma^t \Delta \Gamma^t$ are the sum of non-negative scalings of the diagonal elements of Δ, and since Δ must have at least one non-positive element on the diagonal to meet the constraint $\mathrm{tr}(\Delta) \leq 0$, it follows that $\sum_{t=0}^{k} \Gamma^t \Delta \Gamma^t$ has at least one non-positive diagonal element. Since the diagonal elements of every positive definite matrix are positive, $\sum_{t=0}^{k} \Gamma^t \Delta \Gamma^t$ cannot be positive definite, so we cannot increase the value of $\lambda_{\min}\left(\sum_{t=0}^{k} \Gamma^t(\widetilde{\Sigma}^\star + \Delta)\Gamma^t\right)$ with such a Δ. By convexity of the constraint set, it follows that the directional derivative in the direction of any other point in our constraint set is negative. Since this is a concave function, it follows that $\widetilde{\Sigma}^\star$ is optimal.

Thus, the optimal noise will yield a covariance with minimum eigenvalue $\frac{\gamma^2}{\sum_{i=1}^d \frac{1-\lambda_i^2}{1-\lambda_i^{2k}}}$. For k sufficiently large, we have

$$\frac{\gamma^2}{\sum_{i=1}^d \frac{1-\lambda_i^2}{1-\lambda_i^{2k}}} = \Theta\left(\frac{\gamma^2}{\|\mathbf{1}-\lambda\|_1}\right).$$

Concluding the Proof. By our lower bound on $\lambda_{\min}^\star$, Theorem 3.2 immediately gives the rate Algorithm 3.1 is claimed to achieve. The lower bound on the performance of the other strategies follows as a consequence of Lemma A.3.1, some algebra, and their respective minimum eigenvalue calculations.

□

A.3 Proof of Theorem 3.3

The starting point of our lower bound is the following result, originally due to Jedra and Proutiere [2019].

Lemma A.3.1. *For any matrix $A_\star$, for all $\epsilon > 0, \delta \in (0,1)$, the sample complexity $T_{\epsilon\delta}$ of any (ϵ,δ)-locally-stable algorithm in $A_\star$ satisfies:*

$$\lambda_{\min}\left(\mathbb{E}^{A_\star}\left[\sum_{t=1}^{T_{\epsilon\delta}} x_t x_t^\top\right]\right) \geq \frac{\sigma_w^2}{2\epsilon^2}\log\frac{1}{2.4\delta}$$

Proof. The proof of this result is essentially identical to the proof of Theorem 1 in Jedra and Proutiere [2019] and we omit it here. □

By Lemma A.3.1 we have that

$$\lambda_{\min}\left(\mathbb{E}^{A_\star,\pi_{\text{exp}}}\left[\sum_{t=1}^{T_{\epsilon\delta}} x_t x_t^\top\right]\right) \geq \frac{\sigma_w^2}{2\epsilon^2}\cdot\log\frac{1}{2.4\delta} \tag{A.3.1}$$

for $T_{\epsilon\delta}$ the sample complexity of any (ϵ,δ)-locally-stable algorithm in $A_\star$. Note that

$$\sum_{t=1}^{T_{\epsilon\delta}} x_t x_t^\top \preceq 2\sum_{t=1}^{T_{\epsilon\delta}}[x_t^{\boldsymbol{u}}(x_t^{\boldsymbol{u}})^\top + x_t^{\boldsymbol{w}}(x_t^{\boldsymbol{w}})^\top].$$

For any input (u_t), we see that

$$\|x_t^{\boldsymbol{u}}\|_2 = \left\|\sum_{s=0}^{t-1} A_\star^{t-s-1}B_\star u_s\right\|_2 \leq \tau_\star\|B_\star\|_{\text{op}}\sum_{s=0}^{t-1}\rho_\star^{t-s-1}\|u_s\|_2 \leq \tau_\star\|B_\star\|_{\text{op}}\sum_{s=0}^{t-1}\|u_s\|_2.$$

This implies that

$$\|x_t^{\boldsymbol{u}}\|_2^2 \le \tau_\star^2 \|B_\star\|_{\mathrm{op}}^2 \cdot \left(\sum_{s=0}^{t-1} \|u_s\|_2\right)^2 \le \tau_\star^2 \|B_\star\|_{\mathrm{op}}^2 t \cdot \sum_{s=0}^{t-1} \|u_s\|_2^2.$$

Thus,

$$\mathbb{E}^{A_\star,\pi_{\mathrm{exp}}} \left\| \sum_{t=1}^{T_{\epsilon\delta}} x_t^{\boldsymbol{u}} (x_t^{\boldsymbol{u}})^\top \right\|_{\mathrm{op}} \le \tau_\star^2 \|B_\star\|_{\mathrm{op}}^2 \cdot \sum_{t=1}^{T_{\epsilon\delta}} t \cdot \mathbb{E}^{A_\star,\pi_{\mathrm{exp}}} \sum_{s=0}^{t-1} \|u_s\|_2^2 \le \tau_\star^2 \|B_\star\|_{\mathrm{op}}^2 \gamma^2 T_{\epsilon\delta}^3,$$

where we have used that $\pi_{\mathrm{exp}} \in \Pi_{\gamma^2}$. Note also that $\mathbb{E}^{A_\star,\pi_{\mathrm{exp}}}[x_t^{\boldsymbol{w}}(x_t^{\boldsymbol{w}})^\top] = \boldsymbol{\Lambda}_t^{\mathrm{noise}}(A_\star, 0)$, and we can bound

$$\|\boldsymbol{\Lambda}_t^{\mathrm{noise}}(A_\star, 0)\|_{\mathrm{op}} \le \frac{\sigma_w^2 \tau_\star^2}{1-\rho_\star}.$$

Altogether then, we must have

$$2\tau_\star^2 \|B_\star\|_{\mathrm{op}}^2 \gamma^2 T_{\epsilon\delta}^3 + \frac{2\sigma_w^2 \tau_\star^2 T_{\epsilon\delta}}{1-\rho_\star} \ge \frac{\sigma_w^2}{2\epsilon^2} \cdot \log \frac{1}{2.4\delta}.$$

It follows then that, if ϵ and δ are such that $\frac{1}{\epsilon^2} \cdot \log \frac{1}{\delta} \ge \mathrm{poly}(d_x, d_u, \|A_\star\|_{\mathcal{H}_\infty}, \|B_\star\|_{\mathrm{op}}, \frac{1}{\lambda_{\mathrm{min}}^\star}, \gamma^2, \sigma_w^2)$, then we have that $T_{\epsilon\delta} \ge \mathrm{poly}(d_x, d_u, \|A_\star\|_{\mathcal{H}_\infty}, \|B_\star\|_{\mathrm{op}}, \frac{1}{\lambda_{\mathrm{min}}^\star}, \gamma^2, \sigma_w^2)$. The following result shows that, under this assumption, we can upper bound (A.3.1) by the best possible covariates.

Lemma A.3.2. *For $T \ge \mathrm{poly}\left(d_x, d_u, \|A_\star\|_{\mathcal{H}_\infty}, \|B_\star\|_{\mathrm{op}}, \frac{1}{\lambda_{\mathrm{min}}^\star}, \gamma^2, \sigma_w^2\right)$, we have*

$$\sup_{\pi_{\mathrm{exp}} \in \Pi_{\gamma^2}} \lambda_{\mathrm{min}} \left(\mathbb{E}^{A_\star,\pi_{\mathrm{exp}}} \left[\sum_{t=1}^{T} x_t x_t^\top \right] \right) \le 16T \cdot \limsup_{T' \to \infty} \sup_{\boldsymbol{U} \in \mathcal{U}_{\gamma^2,T'}} \lambda_{\mathrm{min}} \left(\boldsymbol{\Lambda}_{T'}^{\mathrm{ss}}(A_\star, \boldsymbol{U}, 0) \right) = 16T \cdot \lambda_{\mathrm{min}}^\star.$$

Proof. The proof of this result follows by the same sequence of steps as are used to prove Theorem 5.1. As such, we omit it for the sake of brevity. □

By Lemma A.3.2, if $T_{\epsilon\delta} \ge \mathrm{poly}(d_x, d_u, \|A_\star\|_{\mathcal{H}_\infty}, \|B_\star\|_{\mathrm{op}}, \frac{1}{\lambda_{\mathrm{min}}^\star}, \gamma^2, \sigma_w^2, \log T)$, we can bound

$$\lambda_{\mathrm{min}} \left(\mathbb{E}^{A_\star,\pi_{\mathrm{exp}}} \left[\sum_{t=1}^{T_{\epsilon\delta}} x_t x_t^\top \right] \right) \le 16 T_{\epsilon\delta} \cdot \lambda_{\mathrm{min}}^\star.$$

The result follows by rearranging.

A.4 Technical Results

Lemma A.4.1. *For $\delta < \frac{1}{\|(e^{\iota\omega}I-A)^{-1}\|_{\text{op}}}$ and $\Delta \in \mathbb{R}^{d_x \times d_x}$ with $\|\Delta\|_{\text{op}} = 1$, we have*

$$(e^{\iota\omega}I - A - \delta\Delta)^{-1} = \sum_{s=0}^{\infty}(e^{\iota\omega}I - A)^{-1}(\delta\Delta(e^{\iota\omega}I - A)^{-1})^s$$

Proof. To see that this is true, we can simply multiply the right hand side above by $(e^{\iota\omega}I - A - \delta\Delta)$ and observe that the result is I. In particular, we wish to show that

$$\left(\sum_{s=0}^{\infty}(e^{\iota\omega}I - A)^{-1}(\delta\Delta(e^{\iota\omega}I - A)^{-1})^s\right)(e^{\iota\omega}I - A - \delta\Delta) = I.$$

Consider, for fixed n:

$$\begin{aligned}
&\left\|\left(\sum_{s=0}^{n}(e^{\iota\omega}I - A)^{-1}(\delta\Delta(e^{\iota\omega}I - A)^{-1})^s\right)(e^{\iota\omega}I - A - \delta\Delta) - I\right\|_{\text{op}} \\
&= \left\|I - \delta(e^{\iota\omega}I - A)^{-1}\Delta + \sum_{s=1}^{n}(e^{\iota\omega}I - A)^{-1}\left(\delta\Delta(e^{\iota\omega}I - A)^{-1}\right)^{s-1}\left(\delta\Delta - \delta^2\Delta(e^{\iota\omega}I - A)^{-1}\Delta\right) - I\right\|_{\text{op}} \\
&= \left\|I - \delta(e^{\iota\omega}I - A)^{-1}\Delta + \sum_{s=1}^{n}\delta(e^{\iota\omega}I - A)^{-1}\Big(\left(\delta\Delta(e^{\iota\omega}I - A)^{-1}\right)^{s-1} - \left(\delta\Delta(e^{\iota\omega}I - A)^{-1}\right)^{s}\Big)\Delta - I\right\|_{\text{op}} \\
&= \left\|I - \delta(e^{\iota\omega}I - A)^{-1}\left(\delta\Delta(e^{\iota\omega}I - A)^{-1}\right)^{n}\Delta - I\right\|_{\text{op}} \\
&\leq \delta^{n+1}\|(e^{\iota\omega}I - A)^{-1}\|_{\text{op}}^{n+1}.
\end{aligned}$$

Since $\delta < \frac{1}{\|(e^{\iota\omega}I-A)^{-1}\|_{\text{op}}}$ by assumption, we can make $\delta^{n+1}\|(e^{\iota\omega}I - A)^{-1}\|_{\text{op}}^{n+1}$ arbitrarily small by making n large. Thus, for any $\epsilon > 0$, we can find an N such that for all $n \geq N$:

$$\left\|\left(\sum_{s=0}^{n}(e^{\iota\omega}I - A)^{-1}(\delta\Delta(e^{\iota\omega}I - A)^{-1})^s\right)(e^{\iota\omega}I - A - \delta\Delta) - I\right\|_2 \leq \epsilon.$$

This implies that

$$\begin{aligned}
&\lim_{n\to\infty}\left(\sum_{s=0}^{n}(e^{\iota\omega}I - A)^{-1}(\delta\Delta(e^{\iota\omega}I - A)^{-1})^s\right)(e^{\iota\omega}I - A - \delta\Delta) \\
&\quad = \left(\sum_{s=0}^{\infty}(e^{\iota\omega}I - A)^{-1}(\delta\Delta(e^{\iota\omega}I - A)^{-1})^s\right)(e^{\iota\omega}I - A - \delta\Delta) = I.
\end{aligned}$$

□

Lemma A.4.2. *Assume that $\rho(A) < 1$. Then for any ω_1, ω_2 and B, we have that*

$$\|(e^{\iota\omega_1} I - A)^{-1} B - (e^{\iota\omega_2} I - A)^{-1} B\|_2 \leq \left(\max_{\omega \in [0,2\pi]} \|(e^{\iota\omega} I - A)^{-2} B\|_2 \right) |\omega_1 - \omega_2|.$$

Proof. Noting that, since we assume $\rho(A) < 1$, using the identity that $(I + A)^{-1} = I - A + A^2 - A^3 + ...$, we have

$$(e^{\iota\omega} I - A)^{-1} = \left(e^{-\iota\omega} I + e^{-j2\theta} A + e^{-j3\theta} A^2 + ... \right).$$

Thus,

$$\frac{\mathrm{d}}{\mathrm{d}\omega} (e^{\iota\omega} I - A)^{-1} B = \sum_{\ell=0}^{\infty} -\iota(\ell+1) e^{-\iota(\ell+1)\omega} A^{\ell} \cdot B.$$

For any matrix A with $\rho(A) < 1$ we have

$$\begin{aligned} & (I + 2A + 3A^2 + 4A^3 + ...)(I - A)^2 = (I + A + A^2 + A^3 + ...)(I - A) = I \\ \implies & (I + 2A + 3A^2 + 4A^3 + ...)^{-1} = (I - A)^{-2}, \end{aligned}$$

which implies

$$\sum_{\ell=0}^{\infty} -\iota(\ell+1) e^{-\iota(\ell+1)\omega} A^{\ell} B = -\iota e^{-\iota\omega} \sum_{\ell=0}^{\infty} (\ell+1)(e^{-\iota\omega} A)^{\ell} B = -j e^{-\iota\omega} (I - e^{-\iota\omega} A)^{-2} B.$$

The result follows from the Mean Value Theorem and since

$$\max_{\omega \in [0,2\pi]} \| - j e^{-\iota\omega} (I - e^{-\iota\omega} A)^{-2} B\|_{\mathrm{op}} \leq \max_{\omega \in [0,2\pi]} \|(e^{\iota\omega} I - A)^{-2} B\|_{\mathrm{op}}.$$

□

Proof of Proposition 3.4. Fix some m and $j \in [d_u]$ and consider the segment of $\boldsymbol{u}_m$, $\boldsymbol{u}_m^j := (u_t)_{t=(j-1)km+1}^{jkm}$. By construction, this is a signal with period k. Assume we play this input starting from some state x_0 not necessarily equal to 0. Let $x_t^{\boldsymbol{u}_m^j}$ denote the response generated on the noiseless system. By Parseval's Theorem and Proposition 3.2, it follows that

$$\frac{1}{k} \boldsymbol{\Lambda}_k^{\mathrm{freq}}(\theta, \boldsymbol{u}_1^j) = \lim_{m \to \infty} \frac{1}{km} \sum_{t=0}^{km} x_t^{\boldsymbol{u}_m^j} (x_t^{\boldsymbol{u}_m^j})^{\top}$$

Furthermore, by the construction of $\boldsymbol{u}_1^j$ given in `ConstructTimeInput`, we have

$$\frac{1}{k}\boldsymbol{\Lambda}_k^{\text{freq}}(\theta, \boldsymbol{u}_1^j) = \frac{d_u}{k^2}\sum_{\ell=1}^{k} \lambda_{\ell,j}(e^{\iota\frac{2\pi\ell}{k}}I - A)^{-1}Bv_{\ell,j}v_{\ell,j}^{\mathsf{H}}B^{\mathsf{H}}(e^{\iota\frac{2\pi\ell}{k}}I - A)^{-\mathsf{H}}$$

Now note that, if we play the entire sequence of inputs $\boldsymbol{u}_m$, we will have

$$\sum_{t=0}^{d_u km} x_t^{\boldsymbol{u}_m}(x_t^{\boldsymbol{u}_m})^\top = \sum_{j=1}^{d_u}\sum_{t=(j-1)km+1}^{jkm} x_t^{\boldsymbol{u}_m^j}(x_t^{\boldsymbol{u}_m^j})^\top$$

where the starting state, $x_{(j-1)km+1}^{\boldsymbol{u}_m^j}$, is equal to the final state produced when playing the previous input, $x_{(j-1)km}^{\boldsymbol{u}_m^{j-1}}$. Note that, as we assume the system is stable and the input has bounded energy and is of period k, the norm of $x_{(j-1)km}^{\boldsymbol{u}_m^{j-1}}$ will scale sublinearly m (see Appendix C.2.3). It follows that,

$$\begin{aligned}
\lim_{m\to\infty}\frac{1}{d_u mk}\sum_{t=0}^{d_u mk} x_t^{\boldsymbol{u}_m}(x_t^{\boldsymbol{u}_m})^\top &= \lim_{m\to\infty}\frac{1}{d_u mk}\sum_{j=1}^{d_u}\sum_{t=(j-1)km+1}^{jkm} x_t^{\boldsymbol{u}_m^j}(x_t^{\boldsymbol{u}_m^j})^\top \\
&= \frac{1}{d_u}\sum_{j=1}^{d_u}\lim_{m\to\infty}\frac{1}{mk}\sum_{t=(j-1)km+1}^{jkm} x_t^{\boldsymbol{u}_m^j}(x_t^{\boldsymbol{u}_m^j})^\top \\
&= \frac{1}{d_u}\sum_{j=1}^{d_u}\frac{1}{k}\boldsymbol{\Lambda}_k^{\text{freq}}(\theta, \boldsymbol{u}_1^j) \\
&= \frac{1}{k^2}\sum_{\ell=1}^{k}\sum_{j=1}^{d_u}(e^{\iota\frac{2\pi\ell}{k}}I - A)^{-1}B(\lambda_{\ell,j}v_{\ell,j}v_{\ell,j}^{\mathsf{H}})B^{\mathsf{H}}(e^{\iota\frac{2\pi\ell}{k}}I - A)^{-\mathsf{H}} \\
&= \frac{1}{k^2}\sum_{\ell=1}^{k}(e^{\iota\frac{2\pi\ell}{k}}I - A)^{-1}BU_\ell B^{\mathsf{H}}(e^{\iota\frac{2\pi\ell}{k}}I - A)^{-\mathsf{H}} \\
&= \frac{1}{k}\boldsymbol{\Lambda}_k^{\text{freq}}(\theta, \boldsymbol{U})
\end{aligned}$$

where the second to last inequality follows by the definition of $\lambda_{\ell,j}, v_{\ell,j}$ given in `ConstructTimeInput`. To see that the power constraint holds, note that, by Parseval's Theorem and the construction of the input,

$$\sum_{t=1}^{d_u mk} u_t^\top u_t = \sum_{j=1}^{d_u}\sum_{t=(j-1)km+1}^{jkm} u_t^\top u_t = \sum_{j=1}^{d_u}\frac{m}{k}\sum_{\ell=1}^{k} d_u\lambda_{j,\ell}v_{j,\ell}^\top v_{j,\ell} = \sum_{j=1}^{d_u}\frac{m}{k}\sum_{\ell=1}^{k} d_u\text{tr}(\lambda_{j,\ell}v_{j,\ell}v_{j,\ell}^\top)$$

$$= \frac{md_u}{k} \sum_{\ell=1}^{k} \mathrm{tr}(U_\ell) \leq md_u k\gamma^2$$

where the final inequality holds since $\boldsymbol{U} \in \mathcal{U}_{\gamma^2,k}$. □

Proof of Proposition 3.2. 1. Follows by Parseval's Theorem and simple manipulations. For 2., take some ℓ such that $\frac{\ell}{k'} = \frac{n}{k}$ for some integer n. Then,

$$\check{u}'_\ell = \sum_{s=1}^{k'} u_s e^{-\iota \frac{2\pi \ell s}{k'}} = \sum_{s=1}^{k'} u_s e^{-\iota \frac{2\pi n s}{k}} = \sum_{s=1}^{k'} \widetilde{u}_{\mathrm{mod}(s,k)} e^{-\iota \frac{2\pi n \mathrm{mod}(s,k)}{k}} = \frac{k'}{k} \sum_{s=1}^{k} \widetilde{u}_s e^{-\iota \frac{2\pi n s}{k}} = \frac{k'}{k} \check{u}_n$$

Furthermore, if $\frac{\ell}{k'} \neq \frac{n}{k}$ for all integers n, we will have

$$\check{u}'_\ell = \sum_{s=1}^{k'} u_s e^{-\iota \frac{2\pi \ell s}{k'}} = \sum_{r=1}^{k} \widetilde{u}_r \sum_{s=0}^{k'/k-1} e^{-\iota \frac{2\pi \ell (ks+r)}{k'}} = \sum_{r=1}^{k} \widetilde{u}_r e^{-\iota \frac{2\pi \ell r}{k'}} \sum_{s=0}^{k'/k-1} e^{-\iota \frac{2\pi \ell k s}{k'}} = 0$$

Plugging this into the expression for $\boldsymbol{\Lambda}^{\mathrm{freq}}_{k'}(\theta, \boldsymbol{u}')$, the conclusion follows. □

Appendix B

Deferred Proofs from Chapter 4

B.1 Deferred Proofs from Section 4.1.1

Here we provide the formal proofs of the results stated in Section 4.1.1. For the remainder of Appendix B.1, unless otherwise stated we assume the expectation is taken with respect to θ and some fixed exploration policy π_{exp}. Hence, we write $\mathbb{E}[\cdot]$ in place of $\mathbb{E}_{\theta,\pi_{\text{exp}}}[\cdot]$.

B.1.1 Proof of Lemma 4.1.1

Our strategy is to show that an action $\mathfrak{a}$ with low excess risk can be used to produce an estimate of a parameter θ with low error in the task hessian norm $\|\cdot\|^2_{\mathcal{H}(\theta)}$. Specifically, we define the perturbation term

$$\begin{aligned}\delta_\star(\widehat{\mathfrak{a}}) &= \arg\min_{\delta} \|M_{\mathfrak{a}}(\theta_\star)^{1/2}\left((\widehat{\mathfrak{a}} - \mathfrak{a}_{\text{opt}}(\theta_\star)) - G_{\mathfrak{a}}(\theta_\star)\delta\right)\|_2 \\ &= (M_{\mathfrak{a}}(\theta_\star)^{1/2} G_{\mathfrak{a}}(\theta_\star))^\dagger M_{\mathfrak{a}}(\theta_\star)^{1/2}(\widehat{\mathfrak{a}} - \mathfrak{a}_{\text{opt}}(\theta_\star))\end{aligned}$$

and define the induced estimate

$$\widehat{\theta}(\widehat{\mathfrak{a}}) = \theta_\star + \delta_\star(\widehat{\mathfrak{a}})$$

Ensuring $\widehat{\mathfrak{a}}$ close $\mathfrak{a}_{\text{opt}}(\theta_\star)$: We first want to restrict the lower bound to being only over $\widehat{\mathfrak{a}}$ close $\mathfrak{a}_{\text{opt}}(\theta_\star)$. To this end, note that

$$\begin{aligned}\min_{\widehat{\mathfrak{a}}} \max_{\theta:\|\theta-\theta_\star\|_2^2\le r(T)} \mathbb{E}[\mathcal{R}(\widehat{\mathfrak{a}};\theta)] = \min\Bigg\{ &\min_{\widehat{\mathfrak{a}}:\|\widehat{\mathfrak{a}}-\mathfrak{a}_{\text{opt}}(\theta_\star)\|_2^2\le r_{\mathfrak{a}}^2} \max_{\theta:\|\theta-\theta_\star\|_2^2\le r^2} \mathbb{E}[\mathcal{R}(\widehat{\mathfrak{a}};\theta)], \\ &\min_{\widehat{\mathfrak{a}}:\|\widehat{\mathfrak{a}}-\mathfrak{a}_{\text{opt}}(\theta_\star)\|_2^2> r_{\mathfrak{a}}^2} \max_{\theta:\|\theta-\theta_\star\|_2^2\le r^2} \mathbb{E}[\mathcal{R}(\widehat{\mathfrak{a}};\theta)]\Bigg\}\end{aligned}$$

where we are free to choose $r_{\mathfrak{a}}$ as we wish but our choice will satisfy $r_{\mathfrak{a}} \ge L_{\mathfrak{a}1} r$. Now:

$$\|\widehat{\mathfrak{a}} - \mathfrak{a}_{\text{opt}}(\theta_\star)\|_2 \le \|\widehat{\mathfrak{a}} - \mathfrak{a}_{\text{opt}}(\theta)\|_2 + \|\mathfrak{a}_{\text{opt}}(\theta) - \mathfrak{a}_{\text{opt}}(\theta_\star)\|_2$$

$$\leq \|\widehat{\mathfrak{a}} - \mathfrak{a}_{\text{opt}}(\theta)\|_2 + L_{\mathfrak{a}1}\|\theta - \theta_\star\|_2$$
$$\leq \|\widehat{\mathfrak{a}} - \mathfrak{a}_{\text{opt}}(\theta)\|_2 + L_{\mathfrak{a}1} r$$

By the argument above, $\|\widehat{\mathfrak{a}} - \mathfrak{a}_{\text{opt}}(\theta_\star)\|_2^2 > r_{\mathfrak{a}}(T)$ then implies that:

$$\mathcal{R}(\widehat{\mathfrak{a}}; \theta) \geq \frac{\mu}{2}\|\widehat{\mathfrak{a}} - \mathfrak{a}_{\text{opt}}(\theta)\|_2^2 \geq \frac{\mu}{2}\left(\|\widehat{\mathfrak{a}} - \mathfrak{a}_{\text{opt}}(\theta_\star)\|_2 - L_{\mathfrak{a}1} r\right)^2 \geq \frac{\mu}{2}(r_{\mathfrak{a}} - L_{\mathfrak{a}1} r)^2$$

Choosing $r_{\mathfrak{a}} = (1 + \sqrt{2})L_{\mathfrak{a}1} r$,

$$\min_{\widehat{\mathfrak{a}}: \|\widehat{\mathfrak{a}} - \mathfrak{a}_{\text{opt}}(\theta_\star)\|_2^2 > r_{\mathfrak{a}}^2} \max_{\theta: \|\theta - \theta_\star\|_2^2 \leq r^2} \mathbb{E}[\mathcal{R}(\widehat{\mathfrak{a}}; \theta)] \geq \mu L_{\mathfrak{a}1}^2 r^2$$

Thus, definining the constant

$$C_{\mathfrak{a}} := (1 + \sqrt{2})L_{\mathfrak{a}1},$$

So ultimately we have:

$$\min_{\widehat{\mathfrak{a}}} \max_{\theta: \|\theta - \theta_\star\|_2^2 \leq r(T)} \mathbb{E}[\mathcal{R}(\widehat{\mathfrak{a}}; \theta)] \geq \min\left\{ \min_{\widehat{\mathfrak{a}}: \|\widehat{\mathfrak{a}} - \mathfrak{a}_{\text{opt}}(\theta_\star)\|_2^2 \leq C_{\mathfrak{a}}^2 r^2} \max_{\theta: \|\theta - \theta_\star\|_2^2 \leq r^2} \mathbb{E}[\mathcal{R}(\widehat{\mathfrak{a}}; \theta)], \mu L_{\mathfrak{a}1}^2 r^2 \right\} \tag{B.1.1}$$

We now proceed to lower bound the first term in the above expression. In particular, throughout we assume that

$$\|\widehat{\mathfrak{a}} - \mathfrak{a}_{\text{opt}}(\theta_\star)\|_2 \leq C_{\mathfrak{a}} r, \quad C_{\mathfrak{a}} := (1 + \sqrt{2})L_{\mathfrak{a}1}, \tag{B.1.2}$$

Taylor expansions: Fix a $t \in [0, 1]$, parameter θ, estimated action $\widehat{\mathfrak{a}}$, and define the interpolations

$$\widehat{\mathfrak{a}}_t = t\widehat{\mathfrak{a}} + (1 - t)\mathfrak{a}_{\text{opt}}(\theta), \quad \theta_t := t\theta + (1 - t)\theta_t$$

Throughout, we will let $\widehat{\mathfrak{a}}'$ and θ' denote certain values of $\widehat{\mathfrak{a}}_t$ and θ_t for some interpolation parameters $t', t'' \in [0, 1]$ chosen so as to satisfy the application of Taylor's theorem to follow.

First, by Taylor's theorem,

$$\begin{aligned}
\mathcal{R}(\widehat{\mathfrak{a}}; \theta) &= \mathcal{R}(\mathfrak{a}_{\text{opt}}(\theta); \theta) + \frac{d}{dt}\mathcal{R}(\widehat{\mathfrak{a}}_t; \theta)|_{t=0} + \frac{1}{2}\frac{d^2}{dt^2}\mathcal{R}(\widehat{\mathfrak{a}}_t; \theta)|_{t=0} + \frac{1}{6}\frac{d^3}{dt^3}\mathcal{R}(\widehat{\mathfrak{a}}_t; \theta)|_{t=t'} \\
&= \mathcal{R}(\mathfrak{a}_{\text{opt}}(\theta); \theta) + (\nabla_{\mathfrak{a}}\mathcal{R}(\mathfrak{a}; \theta)|_{\mathfrak{a}=\mathfrak{a}_{\text{opt}}(\theta)})^\top(\widehat{\mathfrak{a}} - \mathfrak{a}_{\text{opt}}(\theta)) \\
&\quad + \frac{1}{2}(\mathfrak{a}_{\text{opt}}(\theta) - \widehat{\mathfrak{a}})^\top(\nabla_{\mathfrak{a}}^2\mathcal{R}(\mathfrak{a}; \theta)|_{\mathfrak{a}=\mathfrak{a}_{\text{opt}}(\theta)})(\mathfrak{a}_{\text{opt}}(\theta) - \widehat{\mathfrak{a}}) \\
&\quad + \frac{1}{6}\nabla_{\mathfrak{a}}^3\mathcal{R}(\mathfrak{a}; \theta)|_{\mathfrak{a}=\mathfrak{a}'}[\widehat{\mathfrak{a}} - \mathfrak{a}_{\text{opt}}(\theta), \widehat{\mathfrak{a}} - \mathfrak{a}_{\text{opt}}(\theta), \widehat{\mathfrak{a}} - \mathfrak{a}_{\text{opt}}(\theta)]
\end{aligned}$$

where $t' \in [0,1], \mathfrak{a}' = \widehat{\mathfrak{a}}_{t'}$. The second equality follows by the chain rule and since $\frac{d}{dt}\widehat{\mathfrak{a}}_t = \widehat{\mathfrak{a}} - \mathfrak{a}_{\text{opt}}(\theta)$. Since $\mathfrak{a}_{\text{opt}}(\theta)$ minimizes the excess risk, we have

$$\mathcal{R}(\mathfrak{a}_{\text{opt}}(\theta);\theta) = 0, \quad \nabla_{\mathfrak{a}}\mathcal{R}(\mathfrak{a};\theta)|_{\mathfrak{a}=\mathfrak{a}_{\text{opt}}(\theta)} = 0$$

Thus, we may simplify

$$\begin{aligned}\mathcal{R}(\widehat{\mathfrak{a}};\theta) = \frac{1}{2}(\mathfrak{a}_{\text{opt}}(\theta) - \widehat{\mathfrak{a}})^\top(\nabla^2_{\mathfrak{a}}\mathcal{R}(\mathfrak{a};\theta)|_{\mathfrak{a}=\mathfrak{a}_{\text{opt}}(\theta)})(\mathfrak{a}_{\text{opt}}(\theta) - \widehat{\mathfrak{a}}) \\ + \frac{1}{6}\nabla^3_{\mathfrak{a}}\mathcal{R}(\mathfrak{a};\theta)|_{\mathfrak{a}=\mathfrak{a}'}[\widehat{\mathfrak{a}} - \mathfrak{a}_{\text{opt}}(\theta), \widehat{\mathfrak{a}} - \mathfrak{a}_{\text{opt}}(\theta), \widehat{\mathfrak{a}} - \mathfrak{a}_{\text{opt}}(\theta)]\end{aligned}$$

We can similarly Taylor expand $\mathfrak{a}_{\text{opt}}$ to get:

$$\mathfrak{a}_{\text{opt}}(\theta) = \mathfrak{a}_{\text{opt}}(\theta_\star) + G_{\mathfrak{a}}(\theta_\star)(\theta - \theta_\star) + \nabla^2_\theta \mathfrak{a}_{\text{opt}}(\theta)|_{\theta=\theta'}[\theta - \theta_\star, \theta - \theta_\star],$$

where again we set $\theta' = t''\theta + (1-t'')\theta_\star$ for some $t'' \in [0,1]$. Recall the definitions

$$\begin{aligned}\delta_\star &= \arg\min_\delta \| M_{\mathfrak{a}}(\theta_\star)^{1/2}\left((\widehat{\mathfrak{a}} - \mathfrak{a}_{\text{opt}}(\theta_\star)) - G_{\mathfrak{a}}(\theta_\star)\delta\right)\|_2 \\ &= (M_{\mathfrak{a}}(\theta_\star)^{1/2}G_{\mathfrak{a}}(\theta_\star))^\dagger M_{\mathfrak{a}}(\theta_\star)^{1/2}(\widehat{\mathfrak{a}} - \mathfrak{a}_{\text{opt}}(\theta_\star)) \\ \widehat{\theta}(\widehat{\mathfrak{a}}) &= \theta_\star + \delta_\star\end{aligned}$$

Writing $\widehat{\mathfrak{a}} = \mathfrak{a}_{\text{opt}}(\theta_\star) + G_{\mathfrak{a}}(\theta_\star)\delta_\star + \delta_{\widehat{\mathfrak{a}}}$ for some $\delta_{\widehat{\mathfrak{a}}}$ and denoting $\delta_\theta = \theta - \theta_\star$, we then have:

$$\begin{aligned}\mathcal{R}(\widehat{\mathfrak{a}};\theta) = \frac{1}{2}(\delta_\theta - \delta_\star)^\top G_{\mathfrak{a}}(\theta_\star)^\top M_{\mathfrak{a}}(\theta_\star)G_{\mathfrak{a}}(\theta_\star)(\delta_\theta - \delta_\star) \\ + \underbrace{\frac{1}{2}\delta_{\widehat{\mathfrak{a}}}^\top M_{\mathfrak{a}}(\theta_\star)\delta_{\widehat{\mathfrak{a}}}}_{(a1)} - \underbrace{\delta_{\widehat{\mathfrak{a}}}^\top M_{\mathfrak{a}}(\theta_\star)(\nabla^2_\theta \mathfrak{a}_{\text{opt}}(\theta)|_{\theta=\theta'}[\delta_\theta, \delta_\theta])}_{(a2)} \\ + \underbrace{\frac{1}{2}(\nabla^2_\theta \mathfrak{a}_{\text{opt}}(\theta)|_{\theta=\theta'}[\delta_\theta, \delta_\theta])^\top M_{\mathfrak{a}}(\theta_\star)(\nabla^2_\theta \mathfrak{a}_{\text{opt}}(\theta)|_{\theta=\theta'}[\delta_\theta, \delta_\theta])}_{(a3)} \\ + \underbrace{(\delta_\theta - \delta_\star)^\top G_{\mathfrak{a}}(\theta_\star)^\top M_{\mathfrak{a}}(\theta_\star)(\nabla^2_\theta \mathfrak{a}_{\text{opt}}(\theta)|_{\theta=\theta'}[\delta_\theta, \delta_\theta])}_{(a4)} - \underbrace{(\delta_\theta - \delta_\star)^\top G_{\mathfrak{a}}(\theta_\star)^\top M_{\mathfrak{a}}(\theta_\star)\delta_{\widehat{\mathfrak{a}}}}_{(a5)} \\ + \underbrace{\frac{1}{2}(\mathfrak{a}_{\text{opt}}(\theta) - \widehat{\mathfrak{a}})^\top\left(\nabla^2_{\mathfrak{a}}\mathcal{R}(\mathfrak{a};\theta)|_{\mathfrak{a}=\mathfrak{a}_{\text{opt}}(\theta)} - M_{\mathfrak{a}}(\theta_\star)\right)(\mathfrak{a}_{\text{opt}}(\theta) - \widehat{\mathfrak{a}})}_{(a6)} \\ + \underbrace{\frac{1}{6}\nabla^3_{\mathfrak{a}}\mathcal{R}(\mathfrak{a};\theta)|_{\mathfrak{a}=\mathfrak{a}'}[\widehat{\mathfrak{a}} - \mathfrak{a}_{\text{opt}}(\theta), \widehat{\mathfrak{a}} - \mathfrak{a}_{\text{opt}}(\theta), \widehat{\mathfrak{a}} - \mathfrak{a}_{\text{opt}}(\theta)]}_{(a7)}\end{aligned}$$

Controlling the Taylor Expansion through norm bounds: We verify thatwe are in the regime where Assumption 4.1 holds.

Claim 2. *For* $\mathfrak{a}$ *satisfying (B.1.2), it holds that*

$$\|\theta' - \theta_\star\|_2 \le \|\theta - \theta_\star\|_2 \le r_{\text{quad}}(\theta_\star) \tag{B.1.3}$$

$$\max\{\|\widehat{\mathfrak{a}} - \mathfrak{a}_{\text{opt}}(\theta_\star)\|, \|\mathfrak{a}' - \mathfrak{a}_{\text{opt}}(\theta_\star)\|_2, \|\mathfrak{a}_{\text{opt}}(\theta) - \mathfrak{a}_{\text{opt}}(\theta_\star)\|\} \le L_{\mathfrak{a}1} r_{\text{quad}}(\theta_\star) \tag{B.1.4}$$

Proof of Claim 2. By assumption, we have,

$$r \le \frac{1}{4} r_{\text{quad}}(\theta_\star) \tag{B.1.5}$$

Now recall that $\theta' = t'\theta + (1-t')\theta_\star$, for some $t' \in [0,1]$, so

$$\|\theta' - \theta_\star\|_2 \le \|\theta - \theta_\star\|_2 \le r$$

From this and trivial manipulations of $\|\theta - \theta_\star\|_{\text{op}}, \|\theta_0 - \theta_\star\|_{\text{op}}$, it follows that (B.1.5) implies (B.1.3).

To verify (B.1.4), recall that $\mathfrak{a}' = t''\widehat{\mathfrak{a}} + (1-t'')\mathfrak{a}_{\text{opt}}(\theta)$ for some $t'' \in [0,1]$. Hence,

$$\begin{aligned}&\max\{\|\widehat{\mathfrak{a}} - \mathfrak{a}_{\text{opt}}(\theta_\star)\|, \|\mathfrak{a}' - \mathfrak{a}_{\text{opt}}(\theta_\star)\|_2, \|\mathfrak{a}_{\text{opt}}(\theta) - \mathfrak{a}_{\text{opt}}(\theta_\star)\|\} \\ &\quad \le \|\widehat{\mathfrak{a}} - \mathfrak{a}_{\text{opt}}(\theta_\star)\| + \|\mathfrak{a}_{\text{opt}}(\theta) - \mathfrak{a}_{\text{opt}}(\theta_\star)\|\end{aligned}$$

From (B.1.2), it holds that $\|\widehat{\mathfrak{a}} - \mathfrak{a}_{\text{opt}}(\theta_\star)\| \le (1+\sqrt{2}) r L_{\mathfrak{a}1}$; moreover, since $\|\theta - \theta_\star\| \le r_{\text{quad}}(\theta_\star)$, the smoothness condition, Assumption 4.1, implies that that $\|\mathfrak{a}_{\text{opt}}(\theta_\star) - \mathfrak{a}_{\text{opt}}(\theta)\| \le L_{\mathfrak{a}1}\|\theta_\star - \theta\| \le r L_{\mathfrak{a}1}$. Hence,

$$\begin{aligned}&\max\{\|\widehat{\mathfrak{a}} - \mathfrak{a}_{\text{opt}}(\theta_\star)\|, \|\mathfrak{a}' - \mathfrak{a}_{\text{opt}}(\theta_\star)\|_2, \|\mathfrak{a}_{\text{opt}}(\theta) - \mathfrak{a}_{\text{opt}}(\theta_\star)\|\} \\ &\quad \le (2+\sqrt{2}) L_{\mathfrak{a}1} r \le 4 L_{\mathfrak{a}1} r\end{aligned}$$

Thus, (B.1.5) implies (B.1.4) holds.

□

The following bounds will be useful.

- By assumption: $\|\widehat{\mathfrak{a}} - \mathfrak{a}_{\text{opt}}(\theta_\star)\|_2^2 \le C_{\mathfrak{a}} r$,

$$\|\delta_\theta\|_2^2 = \|\theta - \theta_\star\|_2^2$$

- We have that

$$\|M_{\mathfrak{a}}(\theta_\star)^{1/2} G_{\mathfrak{a}}(\theta_\star)\delta_\star\|_2 \le \|M_{\mathfrak{a}}(\theta_\star)^{1/2}(\widehat{\mathfrak{a}} - \mathfrak{a}_{\text{opt}}(\theta_\star))\|_2 \le \|M_{\mathfrak{a}}(\theta_\star)^{1/2}\|_{\text{op}} C_{\mathfrak{a}} r$$

This follows since, recalling the definition of $\delta_\star$ and letting $U\Sigma V^\top = M_{\mathfrak{a}}(\theta_\star)^{1/2} G_{\mathfrak{a}}(\theta_\star)$, we have $\|M_{\mathfrak{a}}(\theta_\star)^{1/2} G_{\mathfrak{a}}(\theta_\star)\delta_\star\|_2 = \|\Sigma\Sigma^\dagger U^\top M_{\mathfrak{a}}(\theta_\star)^{1/2}(\widehat{\mathfrak{a}} - \mathfrak{a}_{\mathrm{opt}}(\theta_\star))\|_2$ and since $\|\Sigma\Sigma^\dagger\|_{\mathrm{op}} \le 1$.

- $\|M_{\mathfrak{a}}(\theta_\star)^{1/2}\delta_{\widehat{\mathfrak{a}}}\|_2 \le \|M_{\mathfrak{a}}(\theta_\star)^{1/2}(\widehat{\mathfrak{a}} - \mathfrak{a}_{\mathrm{opt}}(\theta_\star))\|_2 + \|M_{\mathfrak{a}}(\theta_\star)^{1/2} G_{\mathfrak{a}}(\theta_\star)\delta_\star\|_2 \le 2\|M_{\mathfrak{a}}(\theta_\star)^{1/2}(\widehat{\mathfrak{a}} - \mathfrak{a}_{\mathrm{opt}}(\theta_\star))\|_2$
- By Assumption 4.1, so long as (B.1.3) holds: $\|G_{\mathfrak{a}}(\theta_\star)\|_{\mathrm{op}} \le L_{\mathfrak{a}1}$, $\|\nabla^2_\theta \mathfrak{a}_{\mathrm{opt}}(\theta)|_{\theta=\theta'}\|_{\mathrm{op}} \le L_{\mathfrak{a}2}$.
- By Assumption 4.1, so long as (B.1.4) holds: $\|\nabla^3_{\mathfrak{a}} \mathcal{R}(\mathfrak{a};\theta)|_{\mathfrak{a}=\mathfrak{a}'}\|_{\mathrm{op}} \le L_{\mathcal{R}3}$.
- $\|M_{\mathfrak{a}}(\theta_\star)^{1/2}\|_{\mathrm{op}} = \sqrt{\|M_{\mathfrak{a}}(\theta_\star)\|_{\mathrm{op}}} = \sqrt{\|\nabla^2_{\mathfrak{a}} \mathcal{R}(\mathfrak{a};\theta_\star)|_{\mathfrak{a}=\mathfrak{a}_{\mathrm{opt}}(\theta_\star)}\|_{\mathrm{op}}} \le \sqrt{L_{\mathcal{R}2}}$. To see why the first equality holds, note that for any PSD $M = U\Sigma U^\top$, $\|M^{1/2}\|_{\mathrm{op}} = \|\Sigma^{1/2}\|_{\mathrm{op}} = \max_i \sqrt{\sigma_i} = \sqrt{\max_i \sigma_i} = \sqrt{\|M\|_{\mathrm{op}}}$.

Lower bounding the excess risk $\mathcal{R}(\widehat{\mathfrak{a}};\theta)$: Throughout the remainder of the proof, we let c denote a universal numerical constant which may change from line to line. From the above observations

$$(a2) = cL_{\mathfrak{a}2} L_{\mathcal{R}2} C_{\mathfrak{a}} r^3$$

By the bounds given above:

$$(a4) = cL_{\mathfrak{a}2}(L_{\mathfrak{a}1} + C_{\mathfrak{a}}) L_{\mathcal{R}2} r^3$$

To bound $(a6)$, we can apply Proposition 4.1 to get that, when (B.1.4) holds,

$$\|\nabla^2_{\mathfrak{a}} \mathcal{R}(\mathfrak{a};\theta)|_{\mathfrak{a}=\mathfrak{a}_{\mathrm{opt}}(\theta)} - \nabla^2_{\mathfrak{a}} \mathcal{R}(\mathfrak{a};\theta)|_{\mathfrak{a}=\mathfrak{a}_{\mathrm{opt}}(\theta_\star)}\|_{\mathrm{op}} \le L_{\mathcal{R}3}\|\mathfrak{a}_{\mathrm{opt}}(\theta) - \mathfrak{a}_{\mathrm{opt}}(\theta_\star)\|_2 \le L_{\mathcal{R}3} L_{\mathfrak{a}1} r$$

Using that $\|\mathfrak{a}_{\mathrm{opt}}(\theta) - \widehat{\mathfrak{a}}\|_2 \le \|\mathfrak{a}_{\mathrm{opt}}(\theta) - \mathfrak{a}_{\mathrm{opt}}(\theta_\star)\|_2 + \|\mathfrak{a}_{\mathrm{opt}}(\theta_\star) - \widehat{\mathfrak{a}}\|_2 \le (L_{\mathfrak{a}1} + C_{\mathfrak{a}} r)$, we have:

$$(a6) = c(L_{\mathfrak{a}1} + C_{\mathfrak{a}})^2 (L_{\mathcal{R}3} L_{\mathfrak{a}1} + L_{\mathrm{hess}}) r^3$$

This same bound on $\|\mathfrak{a}_{\mathrm{opt}}(\theta) - \widehat{\mathfrak{a}}\|_2$ gives:

$$(a7) = cL_{\mathcal{R}3}(L_{\mathfrak{a}1} + C_{\mathfrak{a}})^3 r^{3/2}$$

It remains to bound $(a5)$. Recall that

$$\delta_\star = \arg\min_{\delta} \|M_{\mathfrak{a}}(\theta_\star)^{1/2}(\widehat{\mathfrak{a}} - \mathfrak{a}_{\mathrm{opt}}(\theta_\star) - G_{\mathfrak{a}}(\theta_\star)\delta)\|_2$$

so $\delta_\star$ is the projection of $M_{\mathfrak{a}}(\theta_\star)^{1/2}(\widehat{\mathfrak{a}} - \mathfrak{a}_{\text{opt}}(\theta_\star))$ onto the image of $M_{\mathfrak{a}}(\theta_\star)^{1/2}G_{\mathfrak{a}}(\theta_\star)$. It follows that:

$$M_{\mathfrak{a}}(\theta_\star)^{1/2}(\widehat{\mathfrak{a}} - \mathfrak{a}_{\text{opt}}(\theta_\star) - G_{\mathfrak{a}}(\theta_\star)\delta_\star) = M_{\mathfrak{a}}(\theta_\star)^{1/2}\delta_{\widehat{\mathfrak{a}}} \quad \perp \quad \text{image}(M_{\mathfrak{a}}(\theta_\star)^{1/2}G_{\mathfrak{a}}(\theta_\star))$$

which implies

$$(a5) = -(\delta_\theta - \delta_\star)^\top G_{\mathfrak{a}}(\theta_\star)^\top M_{\mathfrak{a}}(\theta_\star)\delta_{\widehat{\mathfrak{a}}} = 0$$

Combining everything, we've shown that:

$$\mathcal{R}(\widehat{\mathfrak{a}};\theta) \geq \frac{1}{2}(\delta_\theta - \delta_\star)^\top G_{\mathfrak{a}}(\theta_\star)^\top M_{\mathfrak{a}}(\theta_\star)G_{\mathfrak{a}}(\theta_\star)(\delta_\theta - \delta_\star) + (a1) + (a3) - \mathcal{O}\Big(C_1 r^3\Big)$$

for

$$C_1 = 2L_{\mathfrak{a}1}L_{\mathfrak{a}2}L_{\mathcal{R}2} + 8L_{\mathfrak{a}1}^3 L_{\mathcal{R}3} + 4L_{\text{hess}}$$

However, $M_{\mathfrak{a}}(\theta_\star)$ is PSD so $(a1), (a3) \geq 0$, giving:

$$\mathcal{R}(\widehat{\mathfrak{a}};\theta) \geq \frac{1}{2}(\delta_\theta - \delta_\star)^\top G_{\mathfrak{a}}(\theta_\star)^\top M_{\mathfrak{a}}(\theta_\star)G_{\mathfrak{a}}(\theta_\star)(\delta_\theta - \delta_\star) - cC_1 r^3$$

Completing the proof: By definition, $\delta_\theta - \delta_\star = \theta - \widehat{\theta}(\widehat{\mathfrak{a}})$ and $G_{\mathfrak{a}}(\theta_\star)^\top M_{\mathfrak{a}}(\theta_\star)G_{\mathfrak{a}}(\theta_\star) = \mathcal{H}(\theta_\star)$, so

$$(\delta_\theta - \delta_\star)^\top G_{\mathfrak{a}}(\theta_\star)^\top M_{\mathfrak{a}}(\theta_\star)G_{\mathfrak{a}}(\theta_\star)(\delta_\theta - \delta_\star) = \|\theta - \widehat{\theta}(\widehat{\mathfrak{a}})\|^2_{\mathcal{H}(\theta_\star)}$$

Putting things together, we then have that

$$\min_{\widehat{\mathfrak{a}}:\|\widehat{\mathfrak{a}} - \mathfrak{a}_{\text{opt}}(\theta_\star)\|_2 \leq C_{\mathfrak{a}} r} \max_{\theta:\|\theta-\theta_\star\|_2 \leq r} \mathbb{E}[\mathcal{R}(\widehat{\mathfrak{a}};\theta)]$$
$$\geq \min_{\widehat{\mathfrak{a}}:\|\widehat{\mathfrak{a}} - \mathfrak{a}_{\text{opt}}(\theta_\star)\|_2 \leq C_{\mathfrak{a}} r} \max_{\theta:\|\theta-\theta_\star\|_2 \leq r} \mathbb{E}\left[\frac{1}{2}\|\theta - \widehat{\theta}(\widehat{\mathfrak{a}})\|^2_{\mathcal{H}(\theta_\star)}\right] - cC_1 r^3$$

Given knowledge of $\theta_\star$, $\widehat{\theta}(\widehat{\mathfrak{a}})$ is simply an estimator of θ, so it follows that from (B.1.1) that

$$\min_{\widehat{\mathfrak{a}}:\|\widehat{\mathfrak{a}} - \mathfrak{a}_{\text{opt}}(\theta_\star)\|_2^2 \leq C_{\mathfrak{a}} r(T)} \max_{\theta:\|\theta-\theta_\star\|_2^2 \leq r(T)} \mathbb{E}\left[\frac{1}{2}\|\theta - \widehat{\theta}(\widehat{\mathfrak{a}})\|^2_{\mathcal{H}(\theta_\star)}\right] \geq \min_{\widehat{\theta}} \max_{\theta:\|\theta-\theta_\star\|_2^2 \leq r(T)} \mathbb{E}\left[\frac{1}{2}\|\theta - \widehat{\theta}\|^2_{\mathcal{H}(\theta_\star)}\right]$$

This concludes the proof. □

B.1.2 Proof of Proposition 4.1 and Proposition 4.2

Proof of Proposition 4.1. We prove this for a generic function $f : \mathbb{R}^n \to \mathbb{R}^m$. Fix some $x, y \in \mathbb{R}^n$ and let $x_t = tx + (1-t)y$. Then, by Taylor's Theorem,

$$f(x) = f(y) + \frac{d}{dt} f(x_t)|_{t=t'}$$

for some $t' \in [0,1]$. By the chain rule, $\frac{d}{dt} f(x_t) = \nabla_x f(x)|_{x=x_t} \cdot \frac{d}{dt} x_t = \nabla_x f(x)|_{x=x_t} \cdot (x-y)$. So:

$$\|f(x) - f(y)\|_{\text{op}} \leq \|\nabla_x f(x)|_{x=x_{t'}}\|_{\text{op}} \cdot \|x - y\|_{\text{op}}$$

The result follows in our setting using the norm bounds given in Assumption 4.1. □

Proof of Proposition 4.2. Let $\theta_t = t\widehat{\theta} + (1-t)\theta_\star$. Note that for any t, by Proposition 4.1,

$$\|\mathfrak{a}_{\text{opt}}(\theta_t) - \mathfrak{a}_{\text{opt}}(\theta_\star)\|_2 \leq L_{\mathfrak{a}1}\|\theta_t - \theta_\star\|_2 \leq L_{\mathfrak{a}1}\|\widehat{\theta} - \theta_\star\|_2 \leq L_{\mathfrak{a}1} r_{\text{quad}}(\theta_\star)$$

where the last inequality follows by Assumption 4.1. We are therefore in the regime where the norm bounds given in Assumption 4.1 hold, which we will make use of throughout the proof. By Taylor's Theorem:

$$\mathcal{R}(\mathfrak{a}_{\text{opt}}(\widehat{\theta}); \theta_\star) = \mathcal{R}(\mathfrak{a}_{\text{opt}}(\theta_1); \theta_\star) = \mathcal{R}(\mathfrak{a}_{\text{opt}}(\theta_0); \theta_\star) + \frac{d}{dt}\mathcal{R}(\mathfrak{a}_{\text{opt}}(\theta_t); \theta_\star)|_{t=0} + \frac{1}{2}\frac{d^2}{dt^2}\mathcal{R}(\mathfrak{a}_{\text{opt}}(\theta_t); \theta_\star)|_{t=0} + \frac{1}{6}\frac{d^3}{dt^3}\mathcal{R}(\mathfrak{a}_{\text{opt}}(\theta_t); \theta_\star)|_{t=t'}$$

where $t' \in [0,1]$. Assumption 4.1 gives that $\mathcal{R}(\mathfrak{a}_{\text{opt}}(\theta_0); \theta_\star) = \mathcal{R}(\mathfrak{a}_{\text{opt}}(\theta_\star); \theta_\star) = 0$. Furthermore, $\frac{d}{dt}\mathcal{R}(\mathfrak{a}_{\text{opt}}(\theta_t); \theta_\star)|_{t=0} = \nabla_{\mathfrak{a}}\mathcal{R}(\mathfrak{a}; \theta_\star)|_{\mathfrak{a}=\mathfrak{a}_{\text{opt}}(\theta_0)} \cdot \nabla_\theta \mathfrak{a}_{\text{opt}}(\theta)|_{\theta=\theta_0} \cdot \frac{d}{dt}\theta_t|_{t=0}$, but by Assumption 4.1, $\nabla_{\mathfrak{a}}\mathcal{R}(\mathfrak{a}; \theta_\star)|_{\mathfrak{a}=\mathfrak{a}_{\text{opt}}(\theta_0)} = 0$. Finally, by the chain rule and since $\frac{d}{dt}\theta_t = \widehat{\theta} - \theta_\star$:

$$\frac{d^2}{dt^2}\mathcal{R}(\mathfrak{a}_{\text{opt}}(\theta_t); \theta_\star)|_{t=0} = (\widehat{\theta} - \theta_\star)^\top \nabla^2_\theta \mathcal{R}(\mathfrak{a}_{\text{opt}}(\theta); \theta_\star)|_{\theta=\theta_0}(\widehat{\theta} - \theta_\star) = (\widehat{\theta} - \theta_\star)^\top \mathcal{H}(\theta_\star)(\widehat{\theta} - \theta_\star)$$

$$\frac{d^3}{dt^3}\mathcal{R}(\mathfrak{a}_{\text{opt}}(\theta_t); \theta_\star) = \nabla^3_\theta \mathcal{R}(\mathfrak{a}_{\text{opt}}(\theta); \theta_\star)|_{\theta=\theta_{t'}}[\widehat{\theta} - \theta_\star, \widehat{\theta} - \theta_\star, \widehat{\theta} - \theta_\star]$$

It then follows that,

$$\left|\mathcal{R}(\mathfrak{a}_{\text{opt}}(\widehat{\theta}); \theta_\star) - \frac{1}{2}\|\widehat{\theta} - \theta_\star\|^2_{\mathcal{H}(\theta_\star)}\right| \leq \frac{1}{6}\|\nabla^3_\theta \mathcal{R}(\mathfrak{a}_{\text{opt}}(\theta); \theta_\star)|_{\theta=\theta_{t'}}[\widehat{\theta} - \theta_\star, \widehat{\theta} - \theta_\star, \widehat{\theta} - \theta_\star]\|_{\text{op}}$$

The chain rule gives,

$$\nabla^3_\theta \mathcal{R}(\mathfrak{a}_{\text{opt}}(\theta); \theta_\star)[\widehat{\theta} - \theta_\star, \widehat{\theta} - \theta_\star, \widehat{\theta} - \theta_\star]$$

$$
\begin{aligned}
&= \nabla_{\mathfrak{a}}^3 \mathcal{R}(\mathfrak{a}_{\text{opt}}(\theta); \theta_\star)[\nabla_\theta \mathfrak{a}_{\text{opt}}(\theta)[\widehat{\theta} - \theta_\star], \nabla_\theta \mathfrak{a}_{\text{opt}}(\theta)[\widehat{\theta} - \theta_\star], \nabla_\theta \mathfrak{a}_{\text{opt}}(\theta)[\widehat{\theta} - \theta_\star]] \\
&\quad + 3\nabla_{\mathfrak{a}}^2 \mathcal{R}(\mathfrak{a}_{\text{opt}}(\theta); \theta_\star)[\nabla_\theta^2 \mathfrak{a}_{\text{opt}}(\theta)[\widehat{\theta} - \theta_\star, \widehat{\theta} - \theta_\star], \nabla_\theta \mathfrak{a}_{\text{opt}}(\theta)[\widehat{\theta} - \theta_\star]] \\
&\quad + \nabla_{\mathfrak{a}} \mathcal{R}(\mathfrak{a}_{\text{opt}}(\theta); \theta_\star)[\nabla_\theta^3 \mathfrak{a}_{\text{opt}}(\theta)[\widehat{\theta} - \theta_\star, \widehat{\theta} - \theta_\star, \widehat{\theta} - \theta_\star]]
\end{aligned}
$$

so,

$$
\|\nabla_\theta^3 \mathcal{R}(\mathfrak{a}_{\text{opt}}(\theta); \theta_\star)|_{\theta=\theta_{t'}}[\widehat{\theta} - \theta_\star, \widehat{\theta} - \theta_\star, \widehat{\theta} - \theta_\star]\|_{\text{op}} \leq (L_{\mathcal{R}3} L_{\mathfrak{a}1}^3 + 3L_{\mathcal{R}2} L_{\mathfrak{a}2} L_{\mathfrak{a}1} + L_{\mathcal{R}1} L_{\mathfrak{a}3}) \|\widehat{\theta} - \theta_\star\|_{\text{op}}^3
$$

which proves the first inequality. For the second inequality, recall that by definition,

$$
\mathcal{H}(\theta_\star) = \nabla_\theta^2 \mathcal{R}(\mathfrak{a}_{\text{opt}}(\theta); \theta_\star)|_{\theta=\theta_\star}
$$

so, by Taylor's Theorem,

$$
\mathcal{H}(\theta_\star) = \nabla_\theta^2 \mathcal{R}(\mathfrak{a}_{\text{opt}}(\theta); \theta_\star)\big|_{\theta=\widehat{\theta}} + \frac{d}{dt} \nabla_\theta^2 \mathcal{R}(\mathfrak{a}_{\text{opt}}(\theta_t); \theta_\star)|_{t=t'}
$$

for $t' \in [0, 1]$. However,

$$
\frac{d}{dt} \nabla_\theta^2 \mathcal{R}(\mathfrak{a}_{\text{opt}}(\theta_t); \theta_\star)|_{t=t'} = \nabla_\theta^3 \mathcal{R}(\mathfrak{a}_{\text{opt}}(\theta); \theta_\star)|_{\theta=\theta_{t'}} \cdot \frac{d}{dt} \theta_t
$$

Thus,

$$
\|\mathcal{H}(\theta_\star) - \nabla_\theta^2 \mathcal{R}(\mathfrak{a}_{\text{opt}}(\theta); \theta_\star)|_{\theta=\widehat{\theta}}\|_{\text{op}} \leq (L_{\mathcal{R}3} L_{\mathfrak{a}1}^3 + 3L_{\mathcal{R}2} L_{\mathfrak{a}2} L_{\mathfrak{a}1} + L_{\mathcal{R}1} L_{\mathfrak{a}3}) \|\widehat{\theta} - \theta_\star\|_2
$$

By the chain rule,

$$
\nabla_\theta^2 \mathcal{R}(\mathfrak{a}_{\text{opt}}(\theta); \theta') = \nabla_{\mathfrak{a}}^2 \mathcal{R}(\mathfrak{a}; \theta')|_{\mathfrak{a}=\mathfrak{a}_{\text{opt}}(\theta)}[\nabla_\theta \mathfrak{a}_{\text{opt}}(\theta), \nabla_\theta \mathfrak{a}_{\text{opt}}(\theta)] + \nabla_{\mathfrak{a}} \mathcal{R}(\mathfrak{a}; \theta')|_{\mathfrak{a}=\mathfrak{a}_{\text{opt}}(\theta)} \cdot \nabla_\theta^2 \mathfrak{a}_{\text{opt}}(\theta)
$$

So, by Definition 4.1.1, since $\nabla_{\mathfrak{a}} \mathcal{R}(\mathfrak{a}; \theta')|_{\mathfrak{a}=\mathfrak{a}_{\text{opt}}(\theta')} = 0$, we have:

$$
\begin{aligned}
&\|\nabla_\theta^2 \mathcal{R}(\mathfrak{a}_{\text{opt}}(\theta); \theta_\star)|_{\theta=\widehat{\theta}} - \nabla_\theta^2 \mathcal{R}(\mathfrak{a}_{\text{opt}}(\theta); \widehat{\theta})|_{\theta=\widehat{\theta}}\|_{\text{op}} \\
&= \|\nabla_{\mathfrak{a}}^2 \mathcal{R}(\mathfrak{a}; \theta_\star)|_{\mathfrak{a}=\mathfrak{a}_{\text{opt}}(\widehat{\theta})}[\nabla_\theta \mathfrak{a}_{\text{opt}}(\theta)|_{\theta=\widehat{\theta}}, \nabla_\theta \mathfrak{a}_{\text{opt}}(\theta)|_{\theta=\widehat{\theta}}] + \nabla_{\mathfrak{a}} \mathcal{R}(\mathfrak{a}; \theta_\star)|_{\mathfrak{a}=\mathfrak{a}_{\text{opt}}(\widehat{\theta})} \cdot \nabla_\theta^2 \mathfrak{a}_{\text{opt}}(\theta)|_{\theta=\widehat{\theta}} \\
&\qquad - \nabla_{\mathfrak{a}}^2 \mathcal{R}(\mathfrak{a}; \widehat{\theta})|_{\mathfrak{a}=\mathfrak{a}_{\text{opt}}(\widehat{\theta})}[\nabla_\theta \mathfrak{a}_{\text{opt}}(\theta)|_{\theta=\widehat{\theta}}, \nabla_\theta \mathfrak{a}_{\text{opt}}(\theta)|_{\theta=\widehat{\theta}}]\|_{\text{op}} \\
&\leq L_{\text{hess}} \|\nabla_\theta \mathfrak{a}_{\text{opt}}(\theta)|_{\theta=\widehat{\theta}}\|_{\text{op}}^2 \|\widehat{\theta} - \theta_\star\|_2 + \|\nabla_{\mathfrak{a}} \mathcal{R}(\mathfrak{a}; \theta_\star)|_{\mathfrak{a}=\mathfrak{a}_{\text{opt}}(\widehat{\theta})} \cdot \nabla_\theta^2 \mathfrak{a}_{\text{opt}}(\theta)|_{\theta=\widehat{\theta}}\|_{\text{op}} \\
&\leq L_{\text{hess}} L_{\mathfrak{a}1}^2 \|\widehat{\theta} - \theta_\star\|_2 + L_{\mathfrak{a}2} \|\nabla_{\mathfrak{a}} \mathcal{R}(\mathfrak{a}; \theta_\star)|_{\mathfrak{a}=\mathfrak{a}_{\text{opt}}(\widehat{\theta})}\|_{\text{op}}
\end{aligned}
$$

However, $\nabla_{\mathfrak{a}} \mathcal{R}(\mathfrak{a}; \theta_\star)|_{\mathfrak{a}=\mathfrak{a}_{\text{opt}}(\widehat{\theta})}$ is Lipschitz continuous so, since $\nabla_{\mathfrak{a}} \mathcal{R}(\mathfrak{a}; \theta_\star)|_{\mathfrak{a}=\mathfrak{a}_{\text{opt}}(\theta_\star)} = 0$,

$$
\|\nabla_{\mathfrak{a}} \mathcal{R}(\mathfrak{a}; \theta_\star)|_{\mathfrak{a}=\mathfrak{a}_{\text{opt}}(\widehat{\theta})}\|_{\text{op}} = \|\nabla_{\mathfrak{a}} \mathcal{R}(\mathfrak{a}; \theta_\star)|_{\mathfrak{a}=\mathfrak{a}_{\text{opt}}(\widehat{\theta})} - \nabla_{\mathfrak{a}} \mathcal{R}(\mathfrak{a}; \theta_\star)|_{\mathfrak{a}=\mathfrak{a}_{\text{opt}}(\theta_\star)}\|_{\text{op}}
$$

$$\leq L_{\mathcal{R}2}\|\mathfrak{a}_{\mathrm{opt}}(\widehat{\theta}) - \mathfrak{a}_{\mathrm{opt}}(\theta_\star)\|_2$$
$$\leq L_{\mathcal{R}2}L_{\mathfrak{a}1}\|\widehat{\theta} - \theta_\star\|_2$$

Since $\mathcal{H}(\widehat{\theta}) = \nabla^2_\theta \mathcal{R}(\mathfrak{a}_{\mathrm{opt}}(\theta); \widehat{\theta})|_{\theta=\widehat{\theta}}$, we've shown that:

$$\|\mathcal{H}(\theta_\star) - \mathcal{H}(\widehat{\theta})\|_{\mathrm{op}} \leq (L_{\mathcal{R}3}L^3_{\mathfrak{a}1} + 3L_{\mathcal{R}2}L_{\mathfrak{a}2}L_{\mathfrak{a}1} + L_{\mathcal{R}2}L_{\mathfrak{a}1} + L_{\mathcal{R}1}L_{\mathfrak{a}3} + L_{\mathrm{hess}}L^2_{\mathfrak{a}1})\|\theta_\star - \widehat{\theta}\|_2$$

which proves the second inequality. □

B.2 Deferred Proofs from Section 4.2

Lemma B.2.1. *Assume our data is generated according to (4.0.1) and let*

$$\widehat{\theta}_{\mathrm{ls}} := \min_\theta \sum_{t=1}^T \|y_t - \theta^\top z_t\|_2^2$$

Then on the event

$$\mathcal{E} := \left\{\lambda_{\min}(\mathbf{\Sigma}_T) \geq \underline{\lambda} T, \mathbf{\Sigma}_T \preceq T\bar{\mathbf{\Lambda}}_T\right\}$$

with probability at least $1-\delta$:

$$\|\widehat{\theta}_{\mathrm{ls}} - \theta_\star\|_2 \leq C\sqrt{\frac{\log(1/\delta) + d_\theta + \log\det(\bar{\mathbf{\Lambda}}_T/\underline{\lambda} + I)}{\underline{\lambda} T}}.$$

Proof. Define the following events:

$$\mathcal{A} = \left\{\|\widehat{\theta}_i - \theta_\star\|_2 \leq C\sqrt{\frac{\log(1/\delta) + d_\theta + \log\det(\bar{\mathbf{\Lambda}}_T/\underline{\lambda} + I)}{\underline{\lambda} T}}\right\}$$
$$\mathcal{E}_1 = \left\{\left\|\left(\sum_{t=1}^T z_t z_t^\top\right)^{-1/2}\sum_{t=1}^T z_t w_t^\top\right\|_{\mathrm{op}} \leq c_2\sigma_w\sqrt{\log\frac{1}{\delta} + d_\theta + \log\det(\bar{\mathbf{\Lambda}}_T/\underline{\lambda} + I)}\right\}$$

Our goal is to show that $\mathbb{P}[\mathcal{A}^c \cap \mathcal{E}] \leq \delta$. The following is trivial.

$$\mathbb{P}[\mathcal{A}^c \cap \mathcal{E}] \leq \mathbb{P}[\mathcal{A}^c \cap \mathcal{E} \cap \mathcal{E}_1] + \mathbb{P}[\mathcal{E} \cap \mathcal{E}_1^c]$$

As $\widehat{\theta}_{\mathrm{ls}}$ is the least squares estimate, we will have that $\widehat{\theta}_{\mathrm{ls}} = (\sum_{t=1}^T z_t z_t^\top)^{-1}\sum_{t=1}^T z_t y_t =$

$\theta_\star + (\sum_{t=T-T_i}^{T} z_t z_t^\top)^{-1} \sum_{t=1}^{T} z_t w_t^\top$. Given this, the error can be decomposed as:

$$\|\widehat{\theta}_{\text{ls}} - \theta_\star\|_2 = \left\| \left(\sum_{t=1}^{T} z_t z_t^\top \right)^{-1} \sum_{t=1}^{T} z_t w_t^\top \right\|_2 \leq \left\| \left(\sum_{t=1}^{T} z_t z_t^\top \right)^{-1/2} \right\|_{\text{op}} \left\| \left(\sum_{t=1}^{T} z_t z_t^\top \right)^{-1/2} \sum_{t=1}^{T} z_t w_t^\top \right\|_2$$
$$= \left\| \left(\sum_{t=1}^{T} z_t z_t^\top \right)^{-1/2} \sum_{t=1}^{T} z_t w_t^\top \right\|_2 \Big/ \sqrt{\lambda_{\min} \left(\sum_{t=1}^{T} z_t z_t^\top \right)}$$

It follows that, on the event $\mathcal{E} \cap \mathcal{E}_1$, the error bound given in $\mathcal{A}$ holds. Thus, $\mathbb{P}[\mathcal{A}^c \cap \mathcal{E} \cap \mathcal{E}_1] = 0$. Proposition 3.1 implies that $\mathbb{P}[\mathcal{E} \cap \mathcal{E}_1^c] \leq \delta$, so $\mathbb{P}[\mathcal{A}^c \cap \mathcal{E}] \leq \delta$. □

B.3 Deferred Proofs from Section 4.3

Proof of Lemma 4.3.1. By Fubini's theorem and Bayes' rule,

$$\begin{aligned}
\mathbb{E}_{X \sim \mathbb{P}_X} \mathbb{E}_{Y \sim \mathbb{P}_{Y|X}(\cdot|X)}[f(X,Y)] &= \int \int f(x,y) \mathbb{P}_{Y|X}(y \mid x) \mathbb{P}_X(x) dy dx \\
&= \int \int \int f(x,y) \mathbb{P}_{X'|Y}(x' \mid y) \mathbb{P}_{Y|X}(y \mid x) \mathbb{P}_X(x) dx' dy dx \\
&= \int \int \int f(x,y) \frac{\mathbb{P}_{Y|X'}(y \mid x') \mathbb{P}_{X'}(x')}{\mathbb{P}_Y(y)} \mathbb{P}_{Y|X}(y \mid x) \mathbb{P}_X(x) dx' dy dx \\
&= \int \int \int f(x,y) \mathbb{P}_{X|Y}(x|y) \mathbb{P}_{Y|X}(y \mid x) \mathbb{P}_{X'}(x') dx' dy dx \\
&= \int \int \int f(x,y) \mathbb{P}_{X|Y}(x|y) \mathbb{P}_{Y|X}(y \mid x) \mathbb{P}_{X'}(x') dx dy dx' \\
&= \mathbb{E}_{X' \sim \mathbb{P}_{X'}} \mathbb{E}_{Y \sim \mathbb{P}_{Y|X}(\cdot|X')} \mathbb{E}_{X \sim \mathbb{P}_{X|Y}(\cdot|Y)}[f(X,Y)]
\end{aligned}$$

Relabeling gives the result. □

Proof of Lemma 4.3.2 .

$$\begin{aligned}
\mathbb{E}\left[\|Z - \mathbb{E}[Z \mid \mathcal{E}]\|_M^2 \mid \mathcal{E} \right] &\geq \mathbb{E}\left[\mathbb{I}\{\mathcal{E}\} \cdot \|Z - \mathbb{E}[Z \mid \mathcal{E}]\|_M^2 \right] \\
&= \underbrace{\mathbb{E}[\|Z - \mathbb{E}[Z \mid \mathcal{E}]\|_M^2]}_{(i)} - \underbrace{\mathbb{E}\left[\mathbb{I}\{\mathcal{E}^c\} \cdot \|Z - \mathbb{E}[Z \mid \mathcal{E}]\|_M^2 \right]}_{(ii)}
\end{aligned}$$

Next, we lower bound $(i) = \mathbb{E}[\|Z - \mathbb{E}[Z \mid \mathcal{E}]\|_M^2] \geq \mathbb{E}[\|Z - \mathbb{E}[Z]\|_M^2]$. Thus, it remains to upper bound (ii):

$$(ii) = \mathbb{E}\left[\mathbb{I}\{\mathcal{E}^c\} \cdot \|Z - \mathbb{E}[Z \mid \mathcal{E}]\|_M^2 \right] \leq 2\mathbb{E}\left[\mathbb{I}\{\mathcal{E}^c\} \cdot \left(\|Z - \mu\|_M^2 + \|\mu - \mathbb{E}[Z \mid \mathcal{E}]\|_M^2 \right) \right]$$

$$\leq 2\mathbb{E}\left[\mathbb{I}\{\mathcal{E}^c\}\cdot\left(\|Z-\mu\|_M^2 + r^2\|M\|_{\mathrm{op}}\right)\right]$$
$$\leq 2\mathbb{E}\left[\mathbb{I}\{\mathcal{E}^c\}\cdot\left(\|Z-\mu\|_2^2\|M\|_{\mathrm{op}} + r^2\|M\|_{\mathrm{op}}\right)\right],$$

where in the second line, we use that, under $\mathcal{E}$, $\|Z-\mu\|_2 \leq r$. Moreover, under $\mathcal{E}^c$, $\|Z-\mu\|_2^2 \geq r$, so that $\|Z-\mu\|_2^2 + r^2 \leq 2\|Z-\mu\|_2^2$. Hence, $(ii) \leq 4\|M\|_{\mathrm{op}}\mathbb{E}[\mathbb{I}\{\mathcal{E}^c\}\|Z-\mu\|_2^2]$. Thus,

$$\mathbb{E}[\|Z-\mathbb{E}[Z\mid\mathcal{E}]\|_M^2 \mid \mathcal{E}] \geq \mathbb{E}\,\|Z-\mathbb{E}[Z]\|_M^2 - 4\|M\|_{\mathrm{op}}\mathbb{E}[\mathbb{I}\{\mathcal{E}^c\}\|Z-\mu\|_2^2],$$

as needed. □

Proof of Lemma 4.3.3. Due to the fact that we have gaussian likelihoods, we we

$$\mathrm{d}\mathbb{P}(\mathfrak{D}_T \mid \theta) \propto \exp(-\frac{1}{2\sigma_w^2}\sum_{t=1}^T (y_t - \langle\theta, z_t\rangle)^2) \propto \exp(-\frac{1}{2\sigma_w^2}\theta^\top\mathbf{\Sigma}_T\theta + \frac{1}{\sigma_w^2}\theta^\top\sum_{t=1}^T z_t y_t^\top)$$

On the other hand, for any given θ, $\mathrm{d}\mathbb{P}(\theta) \propto \exp(-\frac{1}{2}\theta^\top\Gamma^{-1}\theta)$. Hence,

$$\mathrm{d}\mathbb{P}(\theta \mid \mathfrak{D}_T) \propto \exp(-\frac{1}{2\sigma_w^2}\theta^\top(\mathbf{\Sigma}_T + \Gamma^{-1})\theta + \frac{1}{\sigma_w^2}\theta^\top\sum_{t=1}^T z_t y_t^\top)$$

Thus, $\theta \mid \mathfrak{D}_T$ is conditionally Gaussian with covariance $\sigma_w^2(\mathbf{\Sigma}_T + \Gamma^{-1})^{-1}$. It follows that:

$$\begin{aligned}
&\mathbb{E}_{\theta'\sim\mathcal{D}_{\mathrm{full}}(\mathfrak{D}_T)}\left[\|\theta' - \mathbb{E}_{\theta''\sim\mathcal{D}_{\mathrm{full}}(\mathfrak{D}_T)}[\theta'']\|_M^2\right] \\
&= \mathrm{tr}\left(M^{1/2}\mathbb{E}_{\theta'\sim\mathcal{D}_{\mathrm{full}}(\mathfrak{D}_T)}\left[(\theta' - \mathbb{E}_{\theta''\sim\mathcal{D}_{\mathrm{full}}(\mathfrak{D}_T)}[\theta''])(\theta' - \mathbb{E}_{\theta''\sim\mathcal{D}_{\mathrm{full}}(\mathfrak{D}_T)}[\theta''])^\top\right]M^{1/2}\right) \\
&= \sigma_w^2\mathrm{tr}\left(M^{1/2}(\mathbf{\Sigma}_T + \Gamma^{-1})^{-1}M^{1/2}\right).
\end{aligned}$$

□

Proof of Lemma 4.3.4 . From Proposition 1.1. in Hsu et al. [2012], we have $t > 0$, it holds that $\mathbb{P}[\mathbf{g}^\top\Gamma\mathbf{g} > \mathrm{tr}(\Gamma) + 2\sqrt{t}\|\Gamma\|_{\mathrm{F}} + 2\|\Gamma\|_{\mathrm{op}}t] \leq e^{-t}$. In particular, if $\sqrt{t}\|\Gamma\|_{\mathrm{op}} \geq \|\Gamma\|_{\mathrm{F}}$, then $\mathbb{P}[\mathbf{g}^\top\Gamma\mathbf{g} > \mathrm{tr}(\Gamma) + 4t\|\Gamma\|_{\mathrm{op}}] \leq e^{-t}$. Reparametrizing $u = 4t\|\Gamma\|_{\mathrm{op}}$, we have that if $\sqrt{u\|\Gamma\|_{\mathrm{op}}}/2 \geq \|\Gamma\|_{\mathrm{F}}$, then $\mathbb{P}[\mathbf{g}^\top\Gamma\mathbf{g} > \mathrm{tr}(\Gamma) + u] \leq e^{-u/4\|\Gamma\|_{\mathrm{op}}}$. Lastly, the condition $\sqrt{u\|\Gamma\|_{\mathrm{op}}}/2 \leq \|\Gamma\|_{\mathrm{F}}$ is equivalent to $u \geq 4\|\Gamma\|_{\mathrm{F}}^2/\|\Gamma\|_{\mathrm{op}}$. Since $\|\Gamma\|_{\mathrm{F}}^2 \leq \mathrm{tr}(\Gamma)\|\Gamma\|_{\mathrm{op}}$, it suffices that $u \geq 4\mathrm{tr}(\Gamma)$. This concludes the proof. □

Proof of Lemma 4.3.5. Clearly, $AMA = BMB + CMC + BMC + CMB$. Since $CMC, BMB \succeq 0$:

$$\begin{aligned}
BMC + CMB &\succeq BMC + CMB - 4CMC - BMB/4 \\
&= -(B/2 - 2C)M(B/2 - 2C) \\
&\succeq -2(BMB/4 + 4CMC)
\end{aligned}$$

$$= -BMB/2 - 8CMC$$

Thus,

$$AMA \succeq BMB + CMC - BMB/2 - 8CMC = BMB/2 - 7CMC.$$

□

Appendix C

Deferred Proofs from Chapter 5

C.1 Deferred Proofs from Section 5.3

C.1.1 Linear Dynamical Systems Notation Proofs

Proof of Lemma 3.4.1 and Lemma 5.3.1. We have that

$$\begin{aligned}\|\widetilde{\theta}_\star\|_{\mathcal{H}_\infty} &= \max_{\omega\in[0,2\pi]} \left\| \left(e^{\iota\omega} I - \begin{bmatrix} A_\star & B_\star \\ 0 & 0 \end{bmatrix} \right)^{-1} \begin{bmatrix} 0 \\ I \end{bmatrix} \right\|_{\mathrm{op}} \\ &\le \max_{\omega\in[0,2\pi]} \left\| \left(e^{\iota\omega} I - \begin{bmatrix} A_\star & B_\star \\ 0 & 0 \end{bmatrix} \right)^{-1} \right\|_{\mathrm{op}}\end{aligned}$$

For each ω, set $A(\omega) := (e^{\iota\omega} I - A_\star)^{-1}$. Then, using the block matrix inverse formula,

$$\left(e^{\iota\omega} I - \begin{bmatrix} A_\star & B_\star \\ 0 & 0 \end{bmatrix} \right)^{-1} = \begin{bmatrix} A(\omega) & -A(\omega) e^{-\iota\omega} B_\star \\ 0 & e^{-\iota\omega} I \end{bmatrix}$$

Thus,

$$\begin{aligned}\max_{\omega\in[0,2\pi]} \left\| \left(e^{\iota\omega} I - \begin{bmatrix} A_\star & B_\star \\ 0 & 0 \end{bmatrix} \right)^{-1} \right\|_{\mathrm{op}} &\le \max_{\omega\in[0,2\pi]} \left\| \begin{bmatrix} A(\omega) & -A(\omega) e^{-\iota\omega} B_\star \\ 0 & e^{-\iota\omega} I \end{bmatrix} \right\| \\ &\le 1 + (1 + \|B_\star\|_{\mathrm{op}}) \max_{\omega\in[0,2\pi]} \|A(\omega)\|_{\mathrm{op}} \\ &= 1 + (1 + \|B_\star\|_{\mathrm{op}}) \|A_\star\|_{\mathcal{H}_\infty}.\end{aligned}$$

In the case of scalar $A_\star$, Lemma 4.1 Tu et al. [2017] shows that $\|A_\star^k\|_{\mathrm{op}} \le \|\frac{1}{\rho} A_\star\|_{\mathcal{H}_\infty} \rho^k$. In the case when $d_x > 1$, we can apply their proof to the sequence $u^\top A_\star^k v$ for some u, v with $\|u\|_2 = \|v\|_2 = 1$. Doing so, we obtain

$$u^\top A_\star^k v \le \|\tfrac{1}{\rho} A_\star\|_{\mathcal{H}_\infty} \rho^k$$

As this holds for all u and v, we have $\|A_\star^k\|_{\text{op}} \leq \|\frac{1}{\rho}A_\star\|_{\mathcal{H}_\infty}\rho^k$. As $\tau(A_\star, \rho)$ is the smallest value satisfying $\|A_\star^k\|_{\text{op}} \leq \tau(A_\star, \rho)\rho^k$ for all k, it follows that $\tau(A_\star, \rho) \leq \|\frac{1}{\rho}A_\star\|_{\mathcal{H}_\infty}$. We next wish to upper bound $\|\frac{1}{\rho}A_\star\|_{\mathcal{H}_\infty}$ by $\|A_\star\|_{\mathcal{H}_\infty}$ for some choice of ρ. Lemma F.9 of Wagenmaker and Jamieson [2020] gives that

$$\|A_\star - \tfrac{1}{\rho}A_\star\|_{\text{op}} \leq \frac{1}{2\|A_\star\|_{\mathcal{H}_\infty}} \quad \Longrightarrow \quad \|\tfrac{1}{\rho}A_\star\|_{\text{op}} \leq 2\|A_\star\|_{\mathcal{H}_\infty}$$

A sufficient condition to meet this is

$$\rho \geq \frac{2\|A_\star\|_{\mathcal{H}_\infty}\|A_\star\|_{\text{op}}}{1 + 2\|A_\star\|_{\mathcal{H}_\infty}\|A_\star\|_{\text{op}}}$$

As $\rho_\star$ satisfies this, it follows that $\tau(A_\star, \rho_\star) \leq \|\frac{1}{\rho_\star}A_\star\|_{\mathcal{H}_\infty} \leq 2\|A_\star\|_{\mathcal{H}_\infty}$. Combining this with Lemma 5.3.2, we conclude that

$$\tau_\star \leq (1 + \rho_\star^{-1}\|B_\star\|_{\text{op}})\tau(A_\star, \rho_\star) \leq 2(1 + \rho_\star^{-1}\|B_\star\|_{\text{op}})\|A_\star\|_{\mathcal{H}_\infty} \leq 2(1 + 2\|B_\star\|_{\text{op}})\|A_\star\|_{\mathcal{H}_\infty}$$

Finally, by definition of $\rho_\star$ it follows

$$\frac{1}{1 - \rho_\star} \leq \max\{1 + 2\|A_\star\|_{\mathcal{H}_\infty}\|A_\star\|_{\text{op}}, 2\} \leq 2 + 2\|A_\star\|_{\mathcal{H}_\infty}\|A_\star\|_{\text{op}}$$

We then upper bound $\|A_\star\|_{\text{op}} \leq \|A_\star\|_{\mathcal{H}_\infty}$ to obtain the final result. □

Proof of Lemma 5.3.2. Note that:

$$\widetilde{A}^k = \begin{bmatrix} A^k & A^{k-1}B \\ 0 & 0 \end{bmatrix}$$

Thus,

$$\|\widetilde{A}^k\|_{\text{op}} = \sup_{v \in \mathcal{S}^{d+p-1}} \left\| \begin{bmatrix} A^k & A^{k-1}B \\ 0 & 0 \end{bmatrix} v \right\|_{\text{op}} = \sup_{v \in \mathcal{S}^{d+p-1}} \left\| A^k v_1 + A^{k-1}Bv_2 \right\|_{\text{op}} \leq \|A^k\|_{\text{op}} + \|A^{k-1}B\|_{\text{op}}$$

so, for any $\rho > 0$,

$$\begin{aligned} \|\widetilde{A}^k\|_{\text{op}}\rho^{-k} &\leq \|A^k\|_{\text{op}}\rho^{-k} + \|B\|_{\text{op}}\|A^{k-1}\|_{\text{op}}\rho^{-k} \leq \tau(A, \rho) + \rho^{-1}\|B\|_{\text{op}}\|A^{k-1}\|_{\text{op}}\rho^{-(k-1)} \\ &\leq (1 + \rho^{-1}\|B\|_{\text{op}})\tau(A, \rho). \end{aligned}$$

□

Proof of Lemma 5.3.3. Fix some T. In the proof of Theorem 5.1 we showed that, for some

θ_0 satisfying $\|\theta_\star - \theta_0\|_F^2 \le 5(d_x^2 + d_x d_u)/(\lambda^\star_{\min,\infty} T^{5/6})$,

$$\min_{\pi \in \Pi_{\gamma^2}} \operatorname{tr}\left(\mathcal{H}(\theta_0)\left(\mathbb{E}_{\theta_\star,\pi}[\mathbf{\Sigma}_T] + \lambda^\star_{\min,\infty} T \cdot I\right)^{-1}\right) \ge \frac{1}{16T}\Phi^{\mathrm{ss}}_{\mathrm{opt}}(\gamma^2;\theta_\star) - \mathcal{O}\left(\frac{1}{T^{17/12}}\right)$$

Following the proof of Theorem 5.1, we can use Proposition 4.2 to show that

$$\begin{aligned}
&\min_{\pi \in \Pi_{\gamma^2}} \operatorname{tr}\left(\mathcal{H}(\theta_0)\left(\mathbb{E}_{\theta_\star,\pi}[\mathbf{\Sigma}_T] + \lambda^\star_{\min,\infty} T \cdot I\right)^{-1}\right) \\
&\quad \le \min_{\pi \in \Pi_{\gamma^2}} \operatorname{tr}\left(\mathcal{H}(\theta_\star)\left(\mathbb{E}_{\theta_\star,\pi}[\mathbf{\Sigma}_T] + \lambda^\star_{\min,\infty} T \cdot I\right)^{-1}\right) + \mathcal{O}\left(\frac{1}{T^{17/12}}\right) \\
&\quad \le \min_{\pi \in \Pi_{\gamma^2}} \operatorname{tr}\left(\mathcal{H}(\theta_\star)\left(\mathbb{E}_{\theta_\star,\pi}[\mathbf{\Sigma}_T]\right)^{-1}\right) + \mathcal{O}\left(\frac{1}{T^{17/12}}\right)
\end{aligned}$$

Renormalizing by T, it follows that for any T

$$\inf_{\pi \in \Pi_{\gamma^2}} \Phi_T(\theta_\star;\pi) = \min_{\pi \in \Pi_{\gamma^2}} \operatorname{tr}\left(\mathcal{H}(\theta_\star)\left(\mathbb{E}_{\theta_\star,\pi}[\mathbf{\Sigma}_T/T]\right)^{-1}\right) \ge \frac{1}{16}\Phi^{\mathrm{ss}}_{\mathrm{opt}}(\gamma^2;\theta_\star) - \mathcal{O}\left(\frac{1}{T^{5/12}}\right)$$

Taking $\liminf_{T\to\infty}$ of both sides proves the first inequality.

For the second inequality, some trivial manipulations of (C.3.1) in the proof of Lemma 5.4.2 shows that, for sufficiently large T,

$$\min_{\pi \in \Pi_{\gamma^2}} \operatorname{tr}\left(\mathcal{H}(\theta_0)\left(\mathbb{E}_{\theta_\star,\pi}[\mathbf{\Sigma}_T]\right)^{-1}\right) \le \frac{4}{T} \min_{u \in \mathcal{U}_{\gamma^2,T}} \operatorname{tr}\left(\mathcal{H}(\theta_\star)\mathbf{\Lambda}^{\mathrm{ss}}_{T,T}(\widetilde{\theta}_\star, u, 0)^{-1}\right)$$

Renormalizing by T and taking $\liminf_{T\to\infty}$ of both sides gives the result. □

Proof of Lemma 5.3.4. Fix T and consider playing the input $u_t \sim \mathcal{N}(0, \sigma_u^2 \cdot I)$. By definition,

$$\sum_{t=1}^{T} \mathbf{\Lambda}^{\mathrm{noise}}_t(\widetilde{\theta}_\star, \sigma_u) = \sum_{t=1}^{T} \mathbf{\Lambda}^{\mathrm{noise}}_t(\widetilde{\theta}_\star, 0) + \mathbb{E}\left[\sum_{t=1}^{T} x_t^{\boldsymbol{u}}(x_t^{\boldsymbol{u}})^\top\right]$$

By Lemma 5.3.5, it follows that there exists some input $\widetilde{\boldsymbol{u}} = (\widetilde{u}_t)_{t=0}^{k-1}$ with average expected power bounded by γ^2 such that

$$\mathbb{E}\left[\sum_{t=1}^{T} x_t^{\boldsymbol{u}}(x_t^{\boldsymbol{u}})^\top\right] \preceq \mathbb{E}\left[\sum_{t=1}^{2T+k} x_t^{\widetilde{\boldsymbol{u}}}(x_t^{\widetilde{\boldsymbol{u}}})^\top\right] + 5I$$

where $k := T_1 = 2H((d_x^2 + d_x)/2 + 1)$ and

$$H = \left\lceil \log\left(\frac{1-\rho^2}{8\tau(\widetilde{A}_\star, \rho)^3 \gamma^2 T^2}\right) / \log \rho \right\rceil = \mathcal{O}(\log T)$$

By Lemma 5.3.6,

$$\mathbb{E}\left[\sum_{t=1}^{2T+k} x_t^{\widetilde{\boldsymbol{u}}} (x_t^{\widetilde{\boldsymbol{u}}})^\top\right] \preceq \mathbb{E}\boldsymbol{\Lambda}_{2T+k}^{\mathrm{freq}}(\widetilde{\theta}_\star, \widetilde{\boldsymbol{u}}) + \left(\frac{\tau(\widetilde{A}_\star, \rho)\|\widetilde{\theta}_\star\|_{\mathcal{H}_\infty}^2 \sqrt{2T+k+1} k\gamma^2}{1-\rho^k} + \frac{\tau(\widetilde{A}_\star, \rho)^2 \|\widetilde{\theta}_\star\|_{\mathcal{H}_\infty}^2 k^2 \gamma^2}{(1-\rho^k)^2}\right) \cdot I$$

By definition of $\boldsymbol{\Lambda}^{\mathrm{freq}}$ and for any $\boldsymbol{U} \in \widetilde{\mathcal{U}}_{\gamma^2,k}$,

$$\begin{aligned}
\mathbb{E}\boldsymbol{\Lambda}_{2T+k}^{\mathrm{freq}}(\widetilde{\theta}, \boldsymbol{U}) &= \mathbb{E}\frac{4T}{k}\frac{1}{k}\sum_{t=1}^{k} (e^{\iota\frac{2\pi t}{k}} I - \widetilde{A})^{-1} \widetilde{B} \check{u}_t \check{u}_t^{\mathsf{H}} \widetilde{B}^{\mathsf{H}} (e^{\iota\frac{2\pi t}{k}} I - \widetilde{A})^{-\mathsf{H}} \\
&= \frac{4T}{k}\frac{1}{k}\sum_{t=1}^{k} (e^{\iota\frac{2\pi t}{k}} I - \widetilde{A})^{-1} \widetilde{B} \mathbb{E}[\check{u}_t \check{u}_t^{\mathsf{H}}] \widetilde{B}^{\mathsf{H}} (e^{\iota\frac{2\pi t}{k}} I - \widetilde{A})^{-\mathsf{H}}
\end{aligned}$$

Define $U_t := \mathbb{E}[\check{u}_t \check{u}_t^{\mathsf{H}}]$. By Parseval's Theorem, and the power constraint on $\widetilde{u}$, we have

$$\textstyle\sum_{t=1}^{k} \mathrm{tr}(U_t) = \mathbb{E}[\sum_{t=1}^{k} \check{u}_t^{\mathsf{H}} \check{u}_t] = \mathbb{E}[k \sum_{t=0}^{k-1} u_t^\top u_t] \le k^2\gamma^2$$

It follows that there exists some $\widetilde{\boldsymbol{U}} \in \mathcal{U}_{\gamma^2,k}$ such that $\mathbb{E}\boldsymbol{\Lambda}_{2T+k}^{\mathrm{freq}}(\widetilde{\theta}_\star, \widetilde{\boldsymbol{u}}) = \boldsymbol{\Lambda}_{2T+k}^{\mathrm{freq}}(\widetilde{\theta}_\star, \widetilde{\boldsymbol{U}})$. Putting this together, we have that

$$\begin{aligned}
\lambda_{\min}\left(\sum_{t=1}^{T} \boldsymbol{\Lambda}_t^{\mathrm{noise}}(\widetilde{\theta}_\star, \gamma/\sqrt{d_u})\right) &\le \lambda_{\min}\left((2T+k)\boldsymbol{\Lambda}_{2T+k}^{\mathrm{ss}}(\widetilde{\theta}_\star, \widetilde{\boldsymbol{U}})\right) + \mathcal{O}(\sqrt{T}\log T + \mathrm{poly}\log T) \\
&\le \sup_{\boldsymbol{U} \in \mathcal{U}_{\gamma^2,2T+k}} \lambda_{\min}\left((2T+k)\boldsymbol{\Lambda}_{2T+k}^{\mathrm{ss}}(\widetilde{\theta}_\star, \boldsymbol{U})\right) + \mathcal{O}(\sqrt{T}\log T + \mathrm{poly}\log T)
\end{aligned}$$

Dividing through by T and taking the $\limsup_{T\to\infty}$, we have

$$\limsup_{T\to\infty} \frac{1}{T}\lambda_{\min}\left(\sum_{t=1}^{T} \boldsymbol{\Lambda}_t^{\mathrm{noise}}(\widetilde{\theta}_\star, \sigma_u)\right) \le \limsup_{T\to\infty} \sup_{\boldsymbol{U} \in \mathcal{U}_{\gamma^2,T}} 2\lambda_{\min}(\boldsymbol{\Lambda}_T^{\mathrm{ss}}(\widetilde{\theta}_\star, \boldsymbol{U}))$$

Finally, we see that by definition and some algebra that

$$\limsup_{T\to\infty} \frac{1}{T}\lambda_{\min}\left(\sum_{t=1}^{T} \boldsymbol{\Lambda}_t^{\mathrm{noise}}(\widetilde{\theta}_\star, \sigma_u)\right) = \lambda_{\min}\left(\sigma_w^2 \sum_{t=0}^{\infty} \widetilde{A}_\star^t (\widetilde{A}_\star^t)^\top + \sigma_u^2 \sum_{t=0}^{\infty} \widetilde{A}_\star^t \widetilde{B}_\star \widetilde{B}_\star^\top (\widetilde{A}_\star^t)^\top\right)$$

$$= \min\left\{\lambda_{\min}\left(\sigma_w^2\sum_{t=0}^{\infty}A_\star^t(A_\star^t)^\top + \sigma_u^2\sum_{t=0}^{\infty}A_\star^tB_\star B_\star^\top(A_\star^t)^\top\right), \sigma_u^2\right\}$$

Noting that

$$\lambda_{\min}\left(\sigma_w^2\sum_{t=0}^{\infty}A_\star^t(A_\star^t)^\top + \sigma_u^2\sum_{t=0}^{\infty}A_\star^tB_\star B_\star^\top(A_\star^t)^\top\right) \geq \lambda_{\min}\left(\sigma_w^2\sum_{t=0}^{d_x}A_\star^t(A_\star^t)^\top + \sigma_u^2\sum_{t=0}^{d_x}A_\star^tB_\star B_\star^\top(A_\star^t)^\top\right)$$

completes the proof. □

C.1.2 Periodicity of Optimal Inputs

Proof of Lemma 5.3.5. In what follows, we regard $\epsilon > 0$ as fixed, and write $H \leftarrow H_\epsilon$.

We consider the response on the system with no process noise starting from $x_0^{\boldsymbol{u}} = 0$. Given some input $\{u_t\}_{t=0}^{T-1} \in \mathcal{U}_{\gamma^2}$, the state evolves as

$$x_t^{\boldsymbol{u}} = \sum_{s=0}^{t-1}A^{t-s-1}Bu_s$$

We will use G_t to denote the Markov parameters for this system:

$$G_t := [B, AB, \ldots, A^{t-2}B, A^{t-1}B]$$

and will define the extended input, $\mathfrak{u}_t$, and truncated extended input, $\mathfrak{u}_{t;H}$ as:

$$\mathfrak{u}_t := [u_t; u_{t-1}; \ldots; u_1; u_0], \quad \mathfrak{u}_{t;H} := [u_t; u_{t-1}; \ldots; u_{t-H+2}; u_{t-H+1}]$$

If $t < H-1$, we define $u_{-s} = 0$ for all $s > 0$. Then the state can be written as $x_t^{\boldsymbol{u}} = G_t\mathfrak{u}_{t-1}$. We can approximate the state using the last H inputs as $x_{t;H}^{\boldsymbol{u}} = G_H\mathfrak{u}_{t-1;H}$. The following result bounds the error in such an approximation.

Claim 3. *Fix some input u_t with $\mathbb{E}[\sum_{t=0}^{T-1}u_t^\top u_t] \leq T\gamma^2$ and assume we start from $x_0^{\boldsymbol{u}} = 0$. Then, under our choice of $H \geq \log(\frac{\epsilon(1-\rho^2)}{2\|B\|_{\mathrm{op}}^2\tau(A,\rho)^3\gamma^2T})/\log\rho$, we will have*

$$\mathbb{E}\|x_t^{\boldsymbol{u}}(x_t^{\boldsymbol{u}})^\top - x_{t;H}^{\boldsymbol{u}}(x_{t;H}^{\boldsymbol{u}})^\top\|_{\mathrm{op}} \leq \epsilon$$

Proof of Claim 3. We first bound the state difference:

$$\|x_{t;H}^{\boldsymbol{u}} - x_t^{\boldsymbol{u}}\|_2 = \left\|\sum_{s=0}^{t-H-1}A^{t-s-1}Bu_s\right\|_2 \leq \|A^H\|_{\mathrm{op}}\left\|\sum_{s=0}^{t-H-1}A^{t-H-s-1}Bu_s\right\|_2 \leq \tau(A,\rho)\rho^H\|x_{t-H}^{\boldsymbol{u}}\|_2$$

By Jensen's inequality, we can bound $\mathbb{E}\|x_t^{\boldsymbol{u}}\|_2$ as

$$\begin{aligned}\mathbb{E}\|x_t^{\boldsymbol{u}}\|_2 = \mathbb{E}\left\|\sum_{s=0}^{t-1} A^{t-s-1}Bu_s\right\|_2 &\le \|B\|_{\mathrm{op}}\tau(A,\rho)\mathbb{E}\left(\sum_{s=0}^{t-1}\rho^{t-s-1}\|u_s\|_2\right)\\ &\le \|B\|_{\mathrm{op}}\tau(A,\rho)\sqrt{\sum_{s=0}^{t-1}\rho^{2(t-s-1)}}\mathbb{E}\sqrt{\sum_{s=0}^{t-1}\|u_s\|_2^2}\\ &\le \|B\|_{\mathrm{op}}\tau(A,\rho)\sqrt{\sum_{s=0}^{t-1}\rho^{2(t-s-1)}}\sqrt{\mathbb{E}\sum_{s=0}^{T-1}\|u_s\|_2^2}\\ &\le \frac{\|B\|_{\mathrm{op}}\tau(A,\rho)\gamma\sqrt{T}}{\sqrt{1-\rho^2}}\end{aligned}$$

Thus, $\mathbb{E}\|x_{t;H}^{\boldsymbol{u}} - x_t^{\boldsymbol{u}}\|_2 \le \frac{\|B\|_{\mathrm{op}}\tau(A,\rho)^2\gamma\sqrt{2T}}{\sqrt{1-\rho^2}}\rho^H$. Thus, by the triangle inequality and what we have just shown,

$$\mathbb{E}\|x_t^{\boldsymbol{u}}(x_t^{\boldsymbol{u}})^\top - x_{t;H}^{\boldsymbol{u}}(x_{t;H}^{\boldsymbol{u}})^\top\|_{\mathrm{op}} \le \mathbb{E}[(\|x_t^{\boldsymbol{u}}\|_2 + \|x_{t;H}^{\boldsymbol{u}}\|_2)\|x_t^{\boldsymbol{u}} - x_{t;H}^{\boldsymbol{u}}\|_2] \le \frac{2\|B\|_{\mathrm{op}}^2\tau(A,\rho)^3\gamma^2 T}{1-\rho^2}\rho^H$$

where the last inequality follows by noting that our above argument also applies to bounding $\mathbb{E}\|x_{t;H}^{\boldsymbol{u}}\|_2$. The conclusion follows by some algebra. □

Fix $H = \lceil \log(\frac{\epsilon(1-\rho^2)}{8\|B\|_{\mathrm{op}}^2\tau(A,\rho)^3\gamma^2T^2})/\log\rho\rceil$, then, by Claim 3,

$$\mathbb{E}\sum_{t=1}^{T} x_t^{\boldsymbol{u}}(x_t^{\boldsymbol{u}})^\top \preceq \mathbb{E}\sum_{j=1}^{\lceil T/H\rceil}\sum_{t=H(j-1)+1}^{Hj} x_t^{\boldsymbol{u}}(x_t^{\boldsymbol{u}})^\top \preceq \mathbb{E}\sum_{j=1}^{\lceil T/H\rceil}\sum_{t=H(j-1)+1}^{Hj} x_{t;H}^{\boldsymbol{u}}(x_{t;H}^{\boldsymbol{u}})^\top + \epsilon I$$

Now consider a realization of (u_t) in our probability space.

Defining

$$\mathbf{U}_{j;H} := \sum_{t=H(j-1)+1}^{Hj} \mathfrak{u}_{t-1;H}\mathfrak{u}_{t-1;H}^\top \tag{C.1.1}$$

we can rewrite the covariates in terms of the Markov parameters as

$$\sum_{j=1}^{\lceil T/H\rceil}\sum_{t=H(j-1)+1}^{Hj} x_{t;H}^{\boldsymbol{u}}(x_{t;H}^{\boldsymbol{u}})^\top = \sum_{j=1}^{\lceil T/H\rceil} G_H\left(\sum_{t=H(j-1)+1}^{Hj}\mathfrak{u}_{t-1;H}\mathfrak{u}_{t-1;H}^\top\right)G_H^\top = \sum_{j=1}^{\lceil T/H\rceil} G_H\mathbf{U}_{j;H}G_H^\top$$

We will define the set of normalized covariance matrices as

$$\mathcal{M}_H := \left\{ G_H \mathbf{U} G_H^\top \ : \ \mathbf{U} \text{ of form (C.1.1) for input } \{\bar{u}_t\}_{t=1}^{2H}, \sum_{t=1}^{2H} \bar{u}_t^\top \bar{u}_t = 1 \right\}$$

The following result will allow us to express this in a more convenient form.

Claim 4. *Consider any n and some $\{\mathbf{U}'_{j;H}\}_{j=1}^n$, $\mathbf{U}'_{j;H} \in \mathcal{M}_H$. Let $p_j \in [0,1], \sum_{j=1}^n p_j = 1$. Then there exists some set of matrix inputs $\{\mathbf{U}''_{j;H}\}_{j=1}^{(d_x^2+d_x)/2+1}$, $\mathbf{U}''_{j;H} \in \mathcal{M}_H$, and some set of weights $q_j \in [0,1], \sum_{j=1}^{(d_x^2+d_x)/2+1} q_j = 1$ such that*

$$\sum_{j=1}^{n} p_j G_H \mathbf{U}'_{j;H} G_H^\top = \sum_{j=1}^{d_u(d_u+1)/2+1} q_j G_H \mathbf{U}''_{j;H} G_H^\top.$$

Proof. This is a direct consequence of Caratheodory's Theorem. By definition, $\mathcal{M} \subseteq \mathcal{S}_+^{d_x}$. The dimension of $\mathcal{S}_+^{d_x}$ is $(d_x^2 + d_x)/2$ so the points in $\mathcal{M}$ can be thought of as living in a $(d_x^2 + d_x)/2$-dimensional space. Caratheodory's Theorem then gives that, for any point, x, that is a convex combination of elements of $\mathcal{M}$, x can also be written as a convex combination of at most $\dim(\mathcal{M}) + 1$ points in $\mathcal{M}$. Taking $x = \sum_{j=1}^n p_j G_H \mathbf{U}'_{j;H} G_H^\top$, it follows that there exists $(d_x^2 + d_x)/2 + 1$ points $\mathbf{U}''_{j;H} \in \mathcal{M}$ and set of weights $q_j \in [0,1], \sum_{j=1}^m q_j = 1$ such that $x = \sum_{j=1}^{(d_x^2+d_x)/2+1} q_j G_H \mathbf{U}''_{j;H} G_H^\top$. □

We shall use the following definition going forward:

Definition C.1.1. For a given $\mathbf{U}_{j;H}$, let $\gamma^2[\mathbf{U}_{j;H}]$ denote the power of the input corresponding to $\mathbf{U}_{j;H}$. That is, if $\mathbf{U}_{j;H}$ is formed according to (C.1.1),

$$\gamma^2[\mathbf{U}_{j;H}] := \sum_{t=H(j-2)+1}^{Hj-1} u_t^\top u_t \tag{C.1.2}$$

Note then that $\mathbf{U}_{j;H} = \gamma^2[\mathbf{U}_{j;H}] \cdot \mathbf{U}'_{j;H}$ for some $\mathbf{U}'_{j;H} \in \mathcal{M}_H$.

Instantiating Claim 4 with

$$\mathbf{U}'_{j;H} \leftarrow \mathbf{U}_{j;H} / \gamma^2[\mathbf{U}_{j;H}], \quad \text{and} \quad p_j = \gamma^2[\mathbf{U}_{j;H}] / \left(\sum_{i=1}^{\lceil T/H \rceil} \gamma^2[\mathbf{U}_{i;H}] \right),$$

we have that there exists some set of matrices $\{\tilde{\mathbf{U}}_{j;H}\}_{j=1}^{(d_x^2+d_x)/2+1} \subseteq \mathcal{M}_H$ and some set of

weights q_j such that

$$\sum_{j=1}^{\lceil T/H \rceil} G_H \mathbf{U}_{j;H} G_H^\top = \left(\sum_{i=1}^{\lceil T/H \rceil} \gamma^2 [\mathbf{U}_{i;H}] \right) \sum_{j=1}^{(d_x^2+d_x)/2+1} q_j G_H \tilde{\mathbf{U}}_{j;H} G_H^\top \tag{C.1.3}$$

For future reference, we denote

$$\widetilde{\gamma}^2 := \sum_{i=1}^{\lceil T/H \rceil} \gamma^2 [\mathbf{U}_{i;H}]$$

Note that while $\tilde{\mathbf{U}}_{j;H} \in \mathcal{M}_H$, the associate covariates may not be realizable in only $H((d_x^2 + d_x)/2 + 1)$ steps, because $\bar{u}_t$ for a given t will be present in both blocks $\mathbf{U}_{j;H}$ and $\mathbf{U}_{j+1;H}$— these blocks cannot be chosen independently. However, this response can be realized in $2H((d_x^2 + d_x)/2 + 1)$ steps, which we recall is precisely our definition of T_ϵ.

For a given $j \in \{1, \dots, (d_x^2 + d_x)/2 + 1\}$, let $\{\widetilde{u}_{t;j}\}_{t=0}^{2H-1}$ be the set of inputs for which (C.1.1) is satisfied for $\tilde{\mathbf{U}}_{j;H}$, and such that $\sum_{t=0}^{2H-1} \widetilde{u}_{t;j}^\top \widetilde{u}_{t;j} = 1$. Let $\{\widetilde{u}_t\}_{t=0}^{T_\epsilon}$ denote the sequence of inputs formed by concatenating $\{\sqrt{\frac{1}{T}\widetilde{\gamma}^2 q_j T_\epsilon}\widetilde{u}_{t;j}\}_{t=0}^{2H-1}$ for all j. That is, set

$$\widetilde{u}_t = \sqrt{\tfrac{1}{T}\widetilde{\gamma}^2 q_j T_\epsilon}\widetilde{u}_{t',j'} \quad \text{where} \quad j' = \lfloor t/j \rfloor + 1, t' = t - (j'-1)2H$$

Finally, extend $\widetilde{u}_t$ to all t via $\widetilde{u}_t = \widetilde{u}_{\mathrm{mod}(t,T_\epsilon)}$, where the equality holds almost surely. Then,

$$\frac{1}{T}\frac{\widetilde{\gamma}^2 q_j T_\epsilon}{2} G_H \tilde{\mathbf{U}}_{j;H} G_H^\top = \sum_{t=2Hj-H+1}^{2Hj} x_{t;H}^{\tilde{\boldsymbol{u}}} (x_{t;H}^{\tilde{\boldsymbol{u}}})^\top \tag{C.1.4}$$

so $\sum_{j=1}^{(d_x^2+d_x)/2+1} \frac{T_\epsilon}{2T}\widetilde{\gamma}^2 q_j G_H \tilde{\mathbf{U}}_{j;H} G_H^\top$ corresponds to half of the input response, and thus

$$\begin{aligned}
\widetilde{\gamma}^2 \sum_{j=1}^{(d_x^2+d_x)/2+1} q_j G_H \tilde{\mathbf{U}}_{j;H} G_H^\top &= \frac{T}{T_\epsilon/2} \sum_{j=1}^{(d_x^2+d_x)/2+1} \sum_{t=2Hj-H+1}^{2Hj} x_{t;H}^{\tilde{\boldsymbol{u}}} (x_{t;H}^{\tilde{\boldsymbol{u}}})^\top \\
&\preceq \left\lceil \frac{T}{T_\epsilon/2} \right\rceil \sum_{j=1}^{(d_x^2+d_x)/2+1} \sum_{t=2Hj-H+1}^{2Hj} x_{t;H}^{\tilde{\boldsymbol{u}}} (x_{t;H}^{\tilde{\boldsymbol{u}}})^\top \\
&\overset{(a)}{=} \sum_{j=1}^{\lceil \frac{T}{T_\epsilon/2} \rceil ((d_x^2+d_x)/2+1)} \sum_{t=2Hj-H+1}^{2Hj} x_{t;H}^{\tilde{\boldsymbol{u}}} (x_{t;H}^{\tilde{\boldsymbol{u}}})^\top
\end{aligned}$$

$$\overset{(b)}{\preceq} \sum_{t=1}^{2H\left\lceil \frac{T}{T_\epsilon/2}\right\rceil((d_x^2+d_x)/2+1)} x_{t;H}^{\widetilde{\boldsymbol{u}}}(x_{t;H}^{\widetilde{\boldsymbol{u}}})^\top$$
$$\preceq \sum_{t=1}^{2T+T_\epsilon} x_{t;H}^{\widetilde{\boldsymbol{u}}}(x_{t;H}^{\widetilde{\boldsymbol{u}}})^\top$$

where (a) holds because, by construction, we will have $x_{t;H}^{\widetilde{\boldsymbol{u}}} = x_{t+j2H((d_x^2+d_x)/2+1);H}^{\widetilde{\boldsymbol{u}}}$ for any j, since the input is T_ϵ-periodic, and (b) follows as we are simply including more PSD terms in the sum. As this holds pointwise in our probability space, it follows that

$$\mathbb{E}\sum_{j=1}^{\lceil T/H\rceil}\sum_{t=H(j-1)+1}^{Hj} x_{t;H}^{\boldsymbol{u}}(x_{t;H}^{\boldsymbol{u}})^\top = \mathbb{E}\widetilde{\gamma}^2\sum_{j=1}^{(d_x^2+d_x)/2+1} q_j G_H \tilde{\mathbf{U}}_{j;H} G_H^\top \preceq \mathbb{E}\sum_{t=1}^{2T+T_\epsilon} x_{t;H}^{\widetilde{\boldsymbol{u}}}(x_{t;H}^{\widetilde{\boldsymbol{u}}})^\top$$

The input sequence $\{\widetilde{u}_t\}_{t=0}^{T_\epsilon}$ satisfies

$$\mathbb{E}\sum_{t=0}^{T_\epsilon-1}\widetilde{u}_t^\top\widetilde{u}_t = \mathbb{E}\frac{T_\epsilon/2}{T}\cdot\widetilde{\gamma}^2\cdot\sum_{j=1}^{(d_x^2+d_x)/2+1} q_j\sum_{t=0}^{2H-1}\widetilde{u}_{t;j}^\top\widetilde{u}_{t;j}$$
$$\overset{(a)}{=} \mathbb{E}\frac{T_\epsilon/2}{T}\cdot\widetilde{\gamma}^2\cdot\sum_{j=1}^{(d_x^2+d_x)/2+1} q_j$$
$$= \frac{T_\epsilon/2}{T}\mathbb{E}\widetilde{\gamma}^2$$
$$\overset{(b)}{\leq} T_\epsilon\gamma^2$$

where (a) follows since $\sum_{t=0}^{2H-1}\widetilde{u}_{t;j}^\top\widetilde{u}_{t;j} = 1$ almost surely, and (b) follows since

$$\mathbb{E}\widetilde{\gamma}^2 = \mathbb{E}\sum_{i=1}^{\lceil T/H\rceil}\gamma^2(\mathbf{U}_{i;H}) = \mathbb{E}\sum_{i=1}^{\lceil T/H\rceil}\sum_{t=H(i-2)+1}^{Hi-1} u_t^\top u_t \leq 2T\gamma^2$$

Thus, $\widetilde{u}_t$ satisfies the power constraint $\mathbb{E}\sum_{t=0}^{T_\epsilon-1}\widetilde{u}_t^\top\widetilde{u}_t \leq T_\epsilon\gamma^2$, which implies it also satisfies the constraint $\mathbb{E}\sum_{t=1}^{T}\widetilde{u}_t^\top\widetilde{u}_t \leq T\gamma^2$. By Claim 3, given our choice of H and this power constraint,

$$\mathbb{E}\sum_{t=1}^{2T+T_\epsilon} x_{t;H}^{\widetilde{\boldsymbol{u}}}(x_{t;H}^{\widetilde{\boldsymbol{u}}})^\top \preceq \mathbb{E}\sum_{t=1}^{2T+T_\epsilon} x_t^{\widetilde{\boldsymbol{u}}}(x_t^{\widetilde{\boldsymbol{u}}})^\top + 4\epsilon I$$

Finally, note that for any s, we can bound

$$\mathbb{E}\sum_{t=s}^{T_\epsilon+s-1}\widetilde{u}_t^\top\widetilde{u}_t \le 2T_\epsilon\gamma^2$$

since the sum can be contained by at most two periods of the input. The conclusion follows. □

C.1.3 Frequency Domain Approximation

Proof of Lemma 5.3.6. Define $G(e^{\iota\omega}) := (e^{\iota\omega}I - A)^{-1}B$ and let $(\check{x}_t)_{t=1}^T = \mathfrak{F}^{-1}((x_t)_{t=1}^T)$ denote the T point DFT of x_t. Then, by Parseval's Theorem,

$$\begin{aligned}\left\|\sum_{t=0}^{T}x_tx_t^\top - \frac{1}{T}\sum_{t=1}^{T}G(e^{\iota\frac{2\pi t}{T}})\check{u}_t\check{u}_t^{\mathsf{H}}G(e^{\iota\frac{2\pi t}{T}})^{\mathsf{H}}\right\|_{\mathrm{op}} &= \left\|\frac{1}{T}\sum_{t=1}^{T}\check{x}_t\check{x}_t^{\mathsf{H}} - \frac{1}{T}\sum_{t=1}^{T}G(e^{\iota\frac{2\pi t}{T}})\check{u}_t\check{u}_t^{\mathsf{H}}G(e^{\iota\frac{2\pi t}{T}})^{\mathsf{H}}\right\|_{\mathrm{op}}\\ &\le \frac{1}{T}\sum_{t=1}^{T}\|G(e^{\iota\frac{2\pi t}{T}})\check{u}_t - \check{x}_t\|_2(\|\check{x}_t\|_2 + \|G(e^{\iota\frac{2\pi t}{T}})\check{u}_t\|_2)\end{aligned}$$

By Taylor expanding,

$$G(e^{\iota\omega}) = \sum_{s=0}^{\infty}e^{-\iota\omega(s+1)}A^sB$$

By definition of a DFT, and since $x_0 = 0$,

$$\check{x}_\ell = \sum_{t=0}^{T-1}e^{-\iota\frac{2\pi\ell}{T}t}x_t = \sum_{t=1}^{T-1}\sum_{s=0}^{t-1}e^{-\iota\frac{2\pi\ell}{T}t}A^{t-s-1}Bu_s = \sum_{s=0}^{T-1}\left(\sum_{t=0}^{T-s-2}e^{-\iota\frac{2\pi\ell}{T}(t+1)}A^t\right)e^{-\iota\frac{2\pi\ell}{T}s}Bu_s$$

$$\check{u}_s = \sum_{t=0}^{T-1}e^{-\iota\frac{2\pi\ell}{T}t}u_t$$

Therefore,

$$\begin{aligned}G(e^{\iota\frac{2\pi\ell}{T}})\check{u}_\ell - \check{x}_\ell &= \sum_{s=0}^{T-1}\left(\sum_{t=0}^{\infty}e^{-\iota\frac{2\pi\ell}{T}(t+1)}A^t\right)e^{-\iota\frac{2\pi\ell}{T}s}Bu_s - \sum_{s=0}^{T-1}\left(\sum_{t=0}^{T-s-2}e^{-\iota\frac{2\pi\ell}{T}(t+1)}A^t\right)e^{-\iota\frac{2\pi\ell}{T}s}Bu_s\\ &= \sum_{s=0}^{T-1}\left(\sum_{t=T-s-1}^{\infty}e^{-\iota\frac{2\pi\ell}{T}(t+1)}A^t\right)e^{-\iota\frac{2\pi\ell}{T}s}Bu_s\\ &= \sum_{s=0}^{T-1}e^{-\iota\frac{2\pi\ell}{T}(T-s-1)}A^{T-s-1}\left(\sum_{t=0}^{\infty}e^{-\iota\frac{2\pi\ell}{T}(t+1)}A^t\right)e^{-\iota\frac{2\pi\ell}{T}s}Bu_s\end{aligned}$$

$$= e^{-\iota\frac{2\pi\ell}{T}(T-1)}\sum_{s=0}^{T-1} A^{T-s-1}G(e^{\iota\frac{2\pi\ell}{T}})Bu_s$$

Thus, since $u_s = u_{s+k}$ by assumption,

$$\begin{aligned}
\|G(e^{\iota\frac{2\pi\ell}{T}})\check{u}_\ell - \check{x}_\ell\|_2 &\le \tau(A,\rho)\|\theta\|_{\mathcal{H}_\infty}\|B\|_{\text{op}}\sum_{s=0}^{T-1}\rho^{T-s-1}\|u_s\|_2 \\
&\le \tau(A,\rho)\|\theta\|_{\mathcal{H}_\infty}\|B\|_{\text{op}}\sum_{j=0}^{\lceil T/k\rceil}\rho^{kj}\sum_{s=0}^{k-1}\|u_s\|_2 \\
&\le \tau(A,\rho)\|\theta\|_{\mathcal{H}_\infty}\|B\|_{\text{op}}\sqrt{k}\sqrt{\sum_{s=0}^{k-1}\|u_s\|_2^2}\sum_{j=0}^{\lceil T/k\rceil}\rho^{kj} \\
&\le \frac{\tau(A,\rho)\|\theta\|_{\mathcal{H}_\infty}\|B\|_{\text{op}}\sqrt{k}}{1-\rho^k}\sqrt{\sum_{s=0}^{k-1}\|u_s\|_2^2}
\end{aligned}$$

By Parseval's Theorem, and again since $u_s = u_{s+k}$,

$$\sum_{t=1}^{T}\|\check{u}_t\|_2^2 = T\sum_{t=0}^{T-1}\|u_t\|_2^2 \le T\lceil T/k\rceil\sum_{t=0}^{k-1}\|u_t\|_2^2$$

So,

$$\begin{aligned}
\frac{1}{T}\sum_{t=1}^{T}\|G(e^{\iota\frac{2\pi t}{T}})\check{u}_t - \check{x}_t\|_2\|G(e^{\iota\frac{2\pi t}{T}})\check{u}_t\|_2 &\le \frac{\|\theta\|_{\mathcal{H}_\infty}}{T}\sqrt{\sum_{t=1}^{T}\|G(e^{\iota\frac{2\pi t}{T}})\check{u}_t - \check{x}_t\|_2^2}\sqrt{\sum_{t=1}^{T}\|\check{u}_t\|_2^2} \\
&\le \frac{\|\theta\|_{\mathcal{H}_\infty}}{T}\sqrt{T\frac{\tau(A,\rho)^2\|\theta\|_{\mathcal{H}_\infty}^2\|B\|_{\text{op}}^2 k}{(1-\rho^k)^2}\sum_{s=0}^{k-1}\|u_s\|_2^2}\sqrt{T\lceil T/k\rceil\sum_{t=0}^{k-1}\|u_t\|_2^2} \\
&= \frac{\tau(A,\rho)\|\theta\|_{\mathcal{H}_\infty}^2\|B\|_{\text{op}}\sqrt{k\lceil T/k\rceil}}{1-\rho^k}\sum_{t=0}^{k-1}\|u_t\|_2^2
\end{aligned}$$

Again by Parseval's theorem, and by the same calculation as was performed above,

$$\begin{aligned}
\sum_{t=1}^{T}\|\check{x}_t\|_2^2 = T\sum_{t=1}^{T-1}\|x_t\|_2^2 = T\sum_{t=1}^{T-1}\left\|\sum_{s=0}^{t-1}A^{t-s-1}Bu_s\right\|_2^2 &\le T\sum_{t=1}^{T-1}\tau(A,\rho)^2\|B\|_{\text{op}}^2\left(\sum_{s=0}^{t-1}\rho^{t-s-1}\|u_s\|_2\right)^2 \\
&\le \frac{T\tau(A,\rho)^2\|B\|_{\text{op}}^2 k}{(1-\rho^k)^2}\sum_{s=0}^{k-1}\|u_s\|_2^2
\end{aligned}$$

So,

$$\begin{aligned}\frac{1}{T}\sum_{t=1}^{T}\|G(e^{\iota\frac{2\pi t}{T}})\check{u}_t-\check{x}_t\|_2\|X(e^{\iota\frac{2\pi t}{T}})\|_2 &\le \frac{1}{T}\sqrt{\sum_{t=1}^{T}\|G(e^{\iota\frac{2\pi t}{T}})\check{u}_t-\check{x}_t\|_2^2}\sqrt{\sum_{t=1}^{T}\|\check{x}_t\|_2^2}\\ &\le \frac{\|\theta\|_{\mathcal{H}_\infty}}{T}\sqrt{T\frac{\tau(A,\rho)^2\|\theta\|_{\mathcal{H}_\infty}^2\|B\|_{\mathrm{op}}^2 k}{(1-\rho^k)^2}\sum_{s=0}^{k-1}\|u_s\|_2^2}\sqrt{\frac{T\tau(A,\rho)^2\|B\|_{\mathrm{op}}^2 k}{(1-\rho^k)^2}\sum_{s=0}^{k-1}\|u_s\|_2^2}\\ &= \frac{\tau(A,\rho)^2\|\theta\|_{\mathcal{H}_\infty}^2\|B\|_{\mathrm{op}}^2 k}{(1-\rho^k)^2}\sum_{t=0}^{k-1}\|u_t\|_2^2\end{aligned}$$

It follows that

$$\begin{aligned}&\mathbb{E}\left\|\sum_{t=0}^{T}x_t x_t^\top-\frac{1}{T}\sum_{t=1}^{T}G(e^{\iota\frac{2\pi t}{T}})\check{u}_t\check{u}_t^{\mathsf{H}}G(e^{\iota\frac{2\pi t}{T}})^{\mathsf{H}}\right\|_{\mathrm{op}}\\ &\quad\le \frac{\tau(A,\rho)\|\theta\|_{\mathcal{H}_\infty}^2\|B\|_{\mathrm{op}}\sqrt{k\lceil T/k\rceil}}{1-\rho^k}\mathbb{E}\sum_{t=0}^{k-1}\|u_t\|_2^2+\frac{\tau(A,\rho)^2\|\theta\|_{\mathcal{H}_\infty}^2\|B\|_{\mathrm{op}}^2 k}{(1-\rho^k)^2}\mathbb{E}\sum_{t=0}^{k-1}\|u_t\|_2^2\\ &\quad\le \frac{\tau(A,\rho)\|\theta\|_{\mathcal{H}_\infty}^2\|B\|_{\mathrm{op}}\sqrt{T+1}k\gamma^2}{1-\rho^k}+\frac{\tau(A,\rho)^2\|\theta\|_{\mathcal{H}_\infty}^2\|B\|_{\mathrm{op}}^2 k^2\gamma^2}{(1-\rho^k)^2}\end{aligned}$$

and the conclusion follows. □

C.1.4 Smoothness of Covariates

Proof of Lemma 5.3.7. For convenience denote $\epsilon=\|\theta-\theta_\star\|_{\mathrm{op}}$. As $\check{\mathbf{\Lambda}}^{\mathrm{ss}}_{T_1,T_2}(\theta,\boldsymbol{U},\sigma_u)=I_{d_x}\otimes\mathbf{\Lambda}^{\mathrm{ss}}_{T_1,T_2}(\theta,\boldsymbol{U},\sigma_u)$,

$$\|\check{\mathbf{\Lambda}}^{\mathrm{ss}}_{T_1,T_2}(\theta,\boldsymbol{U},\sigma_u)-\check{\mathbf{\Lambda}}^{\mathrm{ss}}_{T_1,T_2}(\theta_\star,\boldsymbol{U},\sigma_u)\|_{\mathrm{op}}=\|\mathbf{\Lambda}^{\mathrm{ss}}_{T_1,T_2}(\theta,\boldsymbol{U},\sigma_u)-\mathbf{\Lambda}^{\mathrm{ss}}_{T_1,T_2}(\theta_\star,\boldsymbol{U},\sigma_u)\|_{\mathrm{op}}$$

By definition, when T_2 is divisible by k,

$$\mathbf{\Lambda}^{\mathrm{ss}}_{T_1,T_2}(\theta,\boldsymbol{U},\sigma_u)=\frac{1}{T_2}\mathbf{\Lambda}^{\mathrm{freq}}_{T_2}(\theta,\boldsymbol{U})+\frac{1}{T_1}\sum_{t=1}^{T_1}\mathbf{\Lambda}^{\mathrm{noise}}_t(\theta,\sigma_u)$$

$$\mathbf{\Lambda}^{\mathrm{freq}}_{T_2}(\theta,\boldsymbol{U})=\frac{T_2}{k}\frac{1}{k}\sum_{\ell=1}^{k}(e^{\iota\frac{2\pi\ell}{k}}I-A)^{-1}BU_\ell B^{\mathsf{H}}(e^{\iota\frac{2\pi\ell}{k}}I-A)^{-\mathsf{H}}$$

$$\mathbf{\Lambda}_t^{\text{noise}}(\theta, \sigma_u) = \sigma_w^2 \sum_{s=0}^{t-1} A^s (A^s)^\top + \sigma_u^2 \sum_{s=0}^{t-1} A^s B B^\top (A^s)^\top$$

Thus,

$$\begin{aligned}
&\|\mathbf{\Lambda}_{T_1,T_2}^{\text{ss}}(\theta, \boldsymbol{U}, \sigma_u) - \mathbf{\Lambda}_{T_1,T_2}^{\text{ss}}(\theta_\star, \boldsymbol{U}, \sigma_u)\|_{\text{op}} \\
&\qquad \le \underbrace{\frac{1}{T_2}\|\mathbf{\Lambda}_{T_2}^{\text{freq}}(\theta, \boldsymbol{U}) - \mathbf{\Lambda}_{T_2}^{\text{freq}}(\theta_\star, \boldsymbol{U})\|_{\text{op}}}_{(a)} + \underbrace{\frac{1}{T_1}\sum_{t=1}^{T_1} \|\mathbf{\Lambda}_t^{\text{noise}}(\theta, \sigma_u) - \mathbf{\Lambda}_t^{\text{noise}}(\theta_\star, \sigma_u)\|_{\text{op}}}_{(b)}
\end{aligned}$$

Lemma C.1.2 gives that, when (5.3.5) holds,

$$\begin{aligned}
(a) &\le \Big(\max_{\omega\in[0,2\pi]} \gamma^2 \|(e^{\iota\omega} I - A_\star)^{-1}\|_{\text{op}}^2 \|B_\star\|_{\text{op}} \left(16 \|(e^{\iota\omega} I - A_\star)^{-1}\|_{\text{op}} \|B_\star\|_{\text{op}} + 2\right) \Big)\epsilon \\
&\qquad + \Big(\max_{\omega\in[0,2\pi]} \gamma^2 \|(e^{\iota\omega} I - A_\star)^{-1}\|_{\text{op}}^2 (32 \|(e^{\iota\omega} I - A_\star)^{-1}\|_{\text{op}} \|B_\star\|_{\text{op}} + 2)\Big)\epsilon^2 \\
&\qquad + \Big(\max_{\omega\in[0,2\pi]} 16\gamma^2 \|(e^{\iota\omega} I - A_\star)^{-1}\|_{\text{op}}^3 \Big)\epsilon^3 \\
&\le \Big(\max_{\omega\in[0,2\pi]} 34\gamma^2 \|(e^{\iota\omega} I - A_\star)^{-1}\|_{\text{op}}^3 (\|B_\star\|_{\text{op}} + 1)^2 \Big)\epsilon
\end{aligned}$$

while Lemma C.1.1 gives that, when (5.3.5) holds,

$$\begin{aligned}
(b) &\le \left(\frac{8(\sigma_w^2 + \sigma_u^2 \|B_\star\|_{\text{op}}^2)\tau(A_\star, \rho)^3}{(1-\rho^2)^2} + \frac{4\sigma_u^2 \|B_\star\|_{\text{op}} \tau(A_\star, \rho)^2}{1-\rho^2} \right)\epsilon + \frac{2\sigma_u^2 \tau(A_\star, \rho)^2}{1-\rho^2}\epsilon^2 \\
&\le \left(\frac{8(\sigma_w^2 + \sigma_u^2 \|B_\star\|_{\text{op}}^2)\tau(A_\star, \rho)^3}{(1-\rho^2)^2} + \frac{4\sigma_u^2 (\|B_\star\|_{\text{op}} + 1)\tau(A_\star, \rho)^2}{1-\rho^2} \right)\epsilon
\end{aligned}$$

The result follows. □

Lemma C.1.1. *Assume that* $\|[A, B] - [\widehat{A}, \widehat{B}]\|_{\text{op}} \le \epsilon$ *and that* $\epsilon \le \frac{1-\rho}{2\tau(A,\rho)}$*, then*

$$\begin{aligned}
&\left\| \sum_{k=0}^{t} \left(\sigma_w^2 A^k (A^k)^\top + \sigma_u^2 A^k B B^\top (A^k)^\top\right) - \sum_{k=0}^{t} \left(\sigma_w^2 \widehat{A}^k (\widehat{A}^k)^\top + \sigma_u^2 \widehat{A}^k \widehat{B}\widehat{B}^\top (\widehat{A}^k)^\top\right) \right\|_{\text{op}} \\
&\qquad \le \left(\frac{8(\sigma_w^2 + \sigma_u^2 \|B\|_{\text{op}}^2)\tau(A, \rho)^3}{(1-\rho^2)^2} + \frac{4\sigma_u^2 \|B\|_{\text{op}} \tau(A, \rho)^2}{1-\rho^2} \right)\epsilon + \frac{2\sigma_u^2 \tau(A, \rho)^2}{1-\rho^2}\epsilon^2.
\end{aligned}$$

Proof. First note that $\|[A, B] - [\widehat{A}, \widehat{B}]\|_{\text{op}} \le \epsilon$ implies $\|A - \widehat{A}\|_{\text{op}} \le \epsilon, \|B - \widehat{B}\|_{\text{op}} \le \epsilon$. We

will denote $\widehat{A} = A + \Delta_A$, where $\|\Delta_A\|_{\text{op}} \leq \epsilon$. We can upper bound

$$\begin{aligned}
&\left\| \sum_{k=0}^{t} \left(\sigma_w^2 A^k (A^k)^\top + \sigma_u^2 A^k B B^\top (A^k)^\top \right) - \sum_{k=0}^{t} \left(\sigma_w^2 \widehat{A}^k (\widehat{A}^k)^\top + \sigma_u^2 \widehat{A}^k \widehat{B} \widehat{B}^\top (\widehat{A}^k)^\top \right) \right\|_{\text{op}} \\
&\leq \sigma_w^2 \sum_{k=0}^{t} \|A^k (A^k)^\top - (A + \Delta_A)^k ((A + \Delta_A)^k)^\top\|_{\text{op}} \\
&\quad + \sigma_u^2 \sum_{k=0}^{t} \|A^k B B^\top (A^k)^\top - (A + \Delta_A)^k \widehat{B} \widehat{B}^\top ((A + \Delta_A)^k)^\top\|_{\text{op}}
\end{aligned}$$

By the triangle inequality,

$$\begin{aligned}
&\|A^k (A^k)^\top - (A + \Delta_A)^k ((A + \Delta_A)^k)^\top\|_{\text{op}} \\
&\quad \leq \|A^k (A^k)^\top - A^k ((A + \Delta_A)^k)^\top + A^k ((A + \Delta_A)^k)^\top - (A + \Delta_A)^k ((A + \Delta_A)^k)^\top\|_{\text{op}} \\
&\quad \leq (\|A^k\|_{\text{op}} + \|(A + \Delta_A)^k\|_{\text{op}}) \|A^k - (A + \Delta_A)^k\|_{\text{op}}
\end{aligned}$$

By Proposition C.1,

$$\|(A + \Delta_A)^k\|_{\text{op}} \leq \tau(A, \rho)(\rho + \tau(A, \rho)\epsilon)^k$$

and

$$\|(A + \Delta_A)^k - A^k\|_{\text{op}} \leq k\tau(A, \rho)^2 (\rho + \tau(A, \rho)\epsilon)^{k-1} \epsilon$$

Combining all of this we have

$$\|A^k (A^k)^\top - (A + \Delta_A)^k ((A + \Delta_A)^k)^\top\|_{\text{op}} \leq k\tau(A, \rho)^3 (\rho^k + (\rho + \tau(A, \rho)\epsilon)^k)(\rho + \tau(A, \rho)\epsilon)^{k-1} \epsilon$$

Denote $\rho_2 := \rho + \tau(A, \rho)\epsilon$. Since we have assumed that $\epsilon \leq \frac{1-\rho}{2\tau(A,\rho)}$, $\rho_2 \leq \frac{1}{2} + \frac{1}{2}\rho$. Then it follows:

$$\begin{aligned}
\sigma_w^2 \sum_{k=0}^{t} \|A^k (A^k)^\top - (A + \Delta_A)^k ((A + \Delta_A)^k)^\top\|_{\text{op}} &\leq \frac{\sigma_w^2 \tau(A, \rho)^3 \epsilon}{\rho_2} \sum_{k=0}^{t} k((\rho\rho_2)^k + \rho_2^{2k}) \\
&\leq \sigma_w^2 \tau(A, \rho)^3 \epsilon \left(\frac{\rho}{(1 - \rho\rho_2)^2} + \frac{\rho_2}{(1 - \rho_2^2)^2} \right) \\
&\leq \frac{2\sigma_w^2 \tau(A, \rho)^3 \epsilon}{(1 - \rho_2^2)^2} \\
&\leq \frac{8\sigma_w^2 \tau(A, \rho)^3 \epsilon}{(1 - \rho^2)^2}
\end{aligned}$$

where the last inequality follows since $1-\rho_2^2 \geq \frac{1}{2}(1-\rho^2)$. Denoting $\widehat{B} = B + \Delta_B$, and using what we have already shown, we have

$$
\begin{aligned}
&\|A^k BB^\top (A^k)^\top - (A+\Delta_A)^k \widehat{B}\widehat{B}^\top ((A+\Delta_A)^k)^\top\|_{\mathrm{op}} \\
&= \|A^k BB^\top (A^k)^\top - (A+\Delta_A)^k BB^\top ((A+\Delta_A)^k)^\top - (A+\Delta_A)^k \Delta_B B^\top ((A+\Delta_A)^k)^\top \\
&\qquad - (A+\Delta_A)^k B\Delta_B{}^\top ((A+\Delta_A)^k)^\top - (A+\Delta_A)^k \Delta_B \Delta_B{}^\top ((A+\Delta_A)^k)^\top\|_{\mathrm{op}} \\
&\leq \|A^k BB^\top (A^k)^\top - A^k BB^\top ((A+\Delta_A)^k)^\top + A^k BB^\top ((A+\Delta_A)^k)^\top - (A+\Delta_A)^k BB^\top ((A+\Delta_A)^k)^\top\|_{\mathrm{op}} \\
&\qquad 2\|B\|_{\mathrm{op}} \tau(A,\rho)^2 (\rho + \tau(A,\rho)\epsilon)^{2k}\epsilon + \tau(A,\rho)^2(\rho+\tau(A,\rho)\epsilon)^{2k}\epsilon^2 \\
&\leq \|B\|_{\mathrm{op}}^2 (\|A^k\|_{\mathrm{op}} + \|(A+\Delta_A)^k\|_{\mathrm{op}}) \|A^k - (A+\Delta_A)^k\|_{\mathrm{op}} + 2\|B\|_{\mathrm{op}}\tau(A,\rho)^2(\rho+\tau(A,\rho)\epsilon)^{2k}\epsilon \\
&\qquad + \tau(A,\rho)^2(\rho+\tau(A,\rho)\epsilon)^{2k}\epsilon^2 \\
&\leq \|B\|_{\mathrm{op}}^2 k\tau(A,\rho)^3(\rho^k + \rho_2^k)\rho_2^{k-1}\epsilon + 2\|B\|_{\mathrm{op}}\tau(A,\rho)^2\rho_2^{2k}\epsilon + \tau(A,\rho)^2\rho_2^{2k}\epsilon^2
\end{aligned}
$$

Thus,

$$
\begin{aligned}
&\sigma_u^2 \sum_{k=0}^t \|A^k BB^\top (A^k)^\top - (A+\Delta_A)^k \widehat{B}\widehat{B}^\top ((A+\Delta_A)^k)^\top\|_{\mathrm{op}} \\
&\qquad \leq \sigma_u^2 \|B\|_{\mathrm{op}}^2 \tau(A,\rho)^3 \epsilon \sum_{k=0}^t k(\rho^k + \rho_2^k)\rho_2^{k-1} + 2\sigma_u^2\|B\|_{\mathrm{op}}\tau(A,\rho)^2\epsilon \sum_{k=0}^t \rho_2^{2k} + \sigma_u^2\tau(A,\rho)^2\epsilon^2 \sum_{k=0}^t \rho_2^{2k} \\
&\qquad \leq \frac{8\sigma_u^2\|B\|_{\mathrm{op}}^2\tau(A,\rho)^3\epsilon}{(1-\rho^2)^2} + \frac{4\sigma_u^2\|B\|_{\mathrm{op}}\tau(A,\rho)^2\epsilon}{1-\rho^2} + \frac{2\sigma_u^2\tau(A,\rho)^2\epsilon^2}{1-\rho^2}
\end{aligned}
$$

The conclusion follows. □

Lemma C.1.2. *Assume that* $\|[\widehat{A},\widehat{B}] - [A,B]\|_{\mathrm{op}} \leq \epsilon$, $\epsilon \leq (\max_{\omega\in[0,2\pi]} 2\|(e^{\iota\omega}I - A)^{-1}\|_{\mathrm{op}})^{-1}$, *and* $U_\ell \succeq 0$, $\sum_{\ell=1}^k \mathrm{tr}(U_\ell) \leq k^2\gamma^2$, *then:*

$$
\begin{aligned}
&\frac{1}{k}\left\|\sum_{\ell=1}^k (e^{\iota\omega_\ell}I - \widehat{A})^{-1}\widehat{B}U_\ell\widehat{B}^{\mathsf{H}}(e^{\iota\omega_\ell}I-\widehat{A})^{-\mathsf{H}} - \sum_{\ell=1}^k (e^{\iota\omega_\ell}I - A)^{-1}BU_\ell B^{\mathsf{H}}(e^{\iota\omega_\ell}I-A)^{-\mathsf{H}}\right\|_{\mathrm{op}} \\
&\qquad \leq \Big(\max_{\omega\in[0,2\pi]} k\gamma^2\|(e^{\iota\omega}I-A)^{-1}\|_{\mathrm{op}}^2\|B\|_{\mathrm{op}}\left(16\|(e^{\iota\omega}I-A)^{-1}\|_{\mathrm{op}}\|B\|_{\mathrm{op}}+2\right)\Big)\epsilon \\
&\qquad\quad + \Big(\max_{\omega\in[0,2\pi]} k\gamma^2\|(e^{\iota\omega}I-A)^{-1}\|_{\mathrm{op}}^2(32\|(e^{\iota\omega}I-A)^{-1}\|_{\mathrm{op}}\|B\|_{\mathrm{op}}+2)\Big)\epsilon^2 \\
&\qquad\quad + \Big(\max_{\omega\in[0,2\pi]} 16k\gamma^2\|(e^{\iota\omega}I-A)^{-1}\|_{\mathrm{op}}^3\Big)\epsilon^3
\end{aligned}
$$

Proof. Note first that $\|[\widehat{A},\widehat{B}]-[A,B]\|_{\mathrm{op}} \le \epsilon$ implies $\|\widehat{A}-A\|_{\mathrm{op}} \le \epsilon, \|\widehat{B}-B\|_{\mathrm{op}} \le \epsilon$ since:

$$\|[\widehat{A},\widehat{B}]-[A,B]\|_{\mathrm{op}} = \max_{u\in\mathcal{S}^{2d},v\in\mathcal{S}^{d+p}} u^\top([\widehat{A},\widehat{B}]-[A,B])v \ge \max_{\substack{u\in\mathcal{S}^{2d},u_{d+1:2d}=0\\ v\in\mathcal{S}^{d+p},v_{d+1:d+p}=0}} u^\top([\widehat{A},\widehat{B}]-[A,B])v = \|\widehat{A}-A\|_{\mathrm{op}}$$

If we denote $\widehat{A} = A+\Delta_A, \widehat{B} = B+\Delta_B$, then:

$$\begin{aligned}
&\left\|\sum_{\ell=1}^k (e^{\iota\omega_\ell}I-\widehat{A})^{-1}\widehat{B}U_\ell\widehat{B}^{\mathsf{H}}(e^{\iota\omega_\ell}I-\widehat{A})^{-\mathsf{H}} - \sum_{\ell=1}^k (e^{\iota\omega_\ell}I-A)^{-1}BU_\ell B^{\mathsf{H}}(e^{\iota\omega_\ell}I-A)^{-\mathsf{H}}\right\|_{\mathrm{op}}\\
&\le \underbrace{\left\|\sum_{\ell=1}^k (e^{\iota\omega_\ell}I-\widehat{A})^{-1}\widehat{B}U_\ell\widehat{B}^{\mathsf{H}}(e^{\iota\omega_\ell}I-\widehat{A})^{-\mathsf{H}} - \sum_{\ell=1}^k (e^{\iota\omega_\ell}I-A)^{-1}\widehat{B}U_\ell\widehat{B}^{\mathsf{H}}(e^{\iota\omega_\ell}I-A)^{-\mathsf{H}}\right\|_{\mathrm{op}}}_{(i)}\\
&\quad + \underbrace{\left\|\sum_{\ell=1}^k (e^{\iota\omega_\ell}I-A)^{-1}\widehat{B}U_\ell\widehat{B}^{\mathsf{H}}(e^{\iota\omega_\ell}I-A)^{-\mathsf{H}} - \sum_{\ell=1}^k (e^{\iota\omega_\ell}I-A)^{-1}BU_\ell B^{\mathsf{H}}(e^{\iota\omega_\ell}I-A)^{-\mathsf{H}}\right\|_{\mathrm{op}}}_{(ii)}
\end{aligned}$$

By Lemma A.2.2:

$$\begin{aligned}
(i) &\le \max_{\omega\in[0,2\pi]} 16k^2\gamma^2\|(e^{\iota\omega}I-A)^{-1}\|_{\mathrm{op}}\|(e^{\iota\omega}I-A)^{-1}\widehat{B}\|_{\mathrm{op}}^2\epsilon\\
&\le \max_{\omega\in[0,2\pi]} 16k^2\gamma^2\|(e^{\iota\omega}I-A)^{-1}\|_{\mathrm{op}}^3(\|B\|_{\mathrm{op}}+\epsilon)^2\epsilon
\end{aligned}$$

Note that these lemmas assume that U_ℓ are rank 1, but a trivial modification extends the results to arbitrary $U_\ell \succeq 0$ satisfying $\sum_{\ell=1}^k \mathrm{tr}(U_\ell) \le k^2\gamma^2$. Further:

$$\begin{aligned}
(ii) &\le 2\left\|\sum_{\ell=1}^k (e^{\iota\omega_\ell}I-A)^{-1}\Delta_B U_\ell B^{\mathsf{H}}(e^{\iota\omega_\ell}I-A)^{-\mathsf{H}}\right\|_{\mathrm{op}} + \left\|\sum_{\ell=1}^k (e^{\iota\omega_\ell}I-A)^{-1}\Delta_B U_\ell \Delta_B{}^{\mathsf{H}}(e^{\iota\omega_\ell}I-A)^{-\mathsf{H}}\right\|_{\mathrm{op}}\\
&\le (2\|\Delta_B\|_{\mathrm{op}}\|B\|_{\mathrm{op}} + \|\Delta_B\|_{\mathrm{op}}^2)\left(\max_{\omega\in[0,2\pi]}\|(e^{\iota\omega}I-A)^{-1}\|_{\mathrm{op}}^2\right)\sum_{\ell=1}^k\|U_\ell\|_{\mathrm{op}}\\
&\le k^2\gamma^2(2\epsilon\|B\|_{\mathrm{op}}+\epsilon^2)\left(\max_{\omega\in[0,2\pi]}\|(e^{\iota\omega}I-A)^{-1}\|_{\mathrm{op}}^2\right)
\end{aligned}$$

The result is then immediate. □

Proposition C.1. *Assume that $\|A^k\|_{\mathrm{op}} \le \tau\rho^k$ for all $k \ge 0$ and that $\|\Delta\|_{\mathrm{op}} \le \epsilon$. Then, for $k \ge 1$,*

$$\|(A+\Delta)^k\|_{\mathrm{op}} \le \tau(\rho+\tau\epsilon)^k, \quad \|(A+\Delta)^k - A^k\| \le \tau(\rho+\tau\epsilon)^k - \tau\rho^k \le k\tau^2(\rho+\tau\epsilon)^{k-1}\epsilon$$

Proof. If A and Δ were scalars, the Binomial Theorem would give:

$$(A+\Delta)^k = \sum_{s=0}^{k} \binom{k}{s} A^{k-s}\Delta^s$$

As matrix multiplication does not commute, we cannot simply apply the Binomial Theorem. However, we note that, for a fixed s, we will have $\binom{k}{s}$ terms of the form

$$A^{n_1}\Delta^{m_1}A^{n_2}\Delta^{m_2}\dots A^{n_s}\Delta^{m_s}A^{n_{s+1}}$$

where $\sum_{i=1}^{s+1} n_i = k-s, \sum_{i=1}^{s} m_i = s$. Critically, there will be at most s Δ terms in this product. Then, using our assumption on $\|A^k\|_{\text{op}}$, we have

$$\begin{aligned}
\|A^{n_1}\Delta^{m_1}A^{n_2}\Delta^{m_2}\dots A^{n_s}\Delta^{m_s}A^{n_{s+1}}\|_{\text{op}} &\le \left(\prod_{i=1}^{s+1}\|A^{n_i}\|_{\text{op}}\right)\left(\prod_{i=1}^{s}\|\Delta^{m_i}\|_{\text{op}}\right) \\
&\le \tau^{s+1}\left(\prod_{i=1}^{s+1}\rho^{n_i}\right)\left(\prod_{i=1}^{s}\epsilon^{m_i}\right) \\
&= \tau^{s+1}\rho^{k-s}\epsilon^s
\end{aligned}$$

As this bound does not depend on the specific values of n_i, m_i, we will have

$$\|(A+\Delta)^k\|_{\text{op}} \le \sum_{s=0}^{k}\binom{k}{s}\tau^{s+1}\rho^{k-s}\epsilon^s = \tau(\rho+\tau\epsilon)^k$$

where the final equality holds by the Binomial Theorem. Similarly, note that $\|(A+\Delta)^k - A^k\|_{\text{op}}$ will behave identically, except that the A^k term will be removed from the expansion of $(A+\Delta)^k$. Thus,

$$\|(A+\Delta)^k - A^k\|_{\text{op}} \le \sum_{s=1}^{k}\binom{k}{s}\tau^{s+1}\rho^{k-s}\epsilon^s = \tau(\rho+\tau\epsilon)^k - \tau\rho^k$$

To show the final conclusion, note that, for $k \ge 1$, the derivative of $(a+x)^k$ is $\frac{d}{dx}(a+x)^k = k(a+x)^{k-1}$. By the Mean Value Theorem,

$$|(a+x)^k - (a+y)^k| \le \left(\max_{z\in[x,y]} k|(a+z)^{k-1}|\right)|x-y|$$

Applying this observation in our setting gives that

$$\tau(\rho+\tau\epsilon)^k - \tau\rho^k \le \tau\left(\max_{z\in[0,\tau\epsilon]} k(\rho+z)^{k-1}\right)\tau\epsilon \le k\tau^2(\rho+\tau\epsilon)^{k-1}\epsilon$$

□

C.1.5 Infinite-Horizon Approximation

Proof of Lemma 5.3.8. For simplicity denote $G_{k,\ell} = (e^{\iota 2\pi \ell/k} I - A)^{-1} B$. Consider some k and, given $\ell \in [1, \bar{k}]$, let $\ell_k(\ell) \in [1, k]$ be the index such that $|\ell_k(\ell)/k - \ell/\bar{k}|$ is minimized. Let $\ell_k^{-1}(\ell) : \{1, \dots, k\} \to 2^{\{1,\dots,\bar{k}\}}$ return the set of indices that map to ℓ. Then:

$$\begin{aligned}\frac{1}{\bar{k}}\mathbf{\Lambda}_{\bar{k}}^{\text{freq}}(\theta, \boldsymbol{U}^\star) &= \frac{1}{\bar{k}^2}\sum_{\ell=1}^{\bar{k}} G_{\bar{k},\ell} U_{\bar{k},\ell}^\star G_{\bar{k},\ell}^{\mathsf{H}} = \frac{1}{\bar{k}^2}\sum_{\ell=1}^{\bar{k}} (G_{k,\ell_k(\ell)} + \Delta_\ell) U_{\bar{k},\ell}^\star (G_{k,\ell_k(\ell)} + \Delta_\ell)^{\mathsf{H}} \\ &= \sum_{\ell=1}^{k} G_{k,\ell}\left(\frac{1}{\bar{k}^2}\sum_{\ell' \in \ell_k^{-1}(\ell)} U_{\bar{k},\ell'}^\star\right) G_{k,\ell}^{\mathsf{H}} + \frac{1}{\bar{k}^2}\sum_{\ell=1}^{\bar{k}}\left(G_{k,\ell_k(\ell)} U_{\bar{k},\ell}^\star \Delta_\ell^{\mathsf{H}} + \Delta_\ell U_{\bar{k},\ell}^\star G_{k,\ell_k(\ell)}^{\mathsf{H}} + \Delta_\ell U_{\bar{k},\ell}^\star \Delta_\ell^{\mathsf{H}}\right)\end{aligned}$$

Set $U_{k,\ell} = \frac{1}{\bar{k}^2}\sum_{\ell' \in \ell_k^{-1}(\ell)} U_{\bar{k},\ell'}^\star$, denote $\delta := \max_{\ell \in [1,\dots,\bar{k}]} \|\Delta_\ell\|_2$, and note that $\|\theta\|_{\mathcal{H}_\infty} \geq \|G_{k,\ell}$ for all k, ℓ. Then:

$$\begin{aligned}\left\|\frac{1}{k}\mathbf{\Lambda}_k^{\text{freq}}(\theta, \boldsymbol{U}) - \frac{1}{\bar{k}}\mathbf{\Lambda}_{\bar{k}}^{\text{freq}}(\theta, \boldsymbol{U}^\star)\right\|_{\text{op}} &\leq \left\|\frac{1}{\bar{k}^2}\sum_{\ell=1}^{\bar{k}}\left(G_{k,\ell_k(\ell)} U_{\bar{k},\ell}^\star \Delta_\ell^{\mathsf{H}} + \Delta_\ell U_{\bar{k},\ell}^\star G_{k,\ell_k(\ell)}^{\mathsf{H}} + \Delta_\ell U_{\bar{k},\ell}^\star \Delta_\ell^{\mathsf{H}}\right)\right\|_{\text{op}} \\ &\leq \frac{2\delta\|\theta\|_{\mathcal{H}_\infty} + \delta^2}{\bar{k}^2}\sum_{\ell=1}^{\bar{k}} \|U_{\bar{k},\ell}^\star\|_{\text{op}} \\ &\leq \frac{2\delta\|\theta\|_{\mathcal{H}_\infty} + \delta^2}{\bar{k}^2}\sum_{\ell=1}^{\bar{k}} \text{tr}(U_{\bar{k},\ell}^\star) \\ &\leq (2\delta\|\theta\|_{\mathcal{H}_\infty} + \delta^2)\gamma^2\end{aligned}$$

where the final equality holds since $\widetilde{U}_{\bar{k},\ell}^\star$ is a feasible input and so must meet the power constraint. By Lemma A.4.2:

$$\|G_{k,\ell} - G_{k',\ell'}\|_2 \leq 2\pi|\ell/k - \ell'/k'|\left(\max_{\omega \in [0,2\pi]} \|(e^{\iota\omega} I - A)^{-2} B\|_{\text{op}}\right)$$

Using this we can bound:

$$\|G_{\bar{k},\ell} - G_{k,\ell_k(\ell)}\|_2 \leq 2\pi|\ell/\bar{k} - \ell_k(\ell)/k|\left(\max_{\omega \in [0,2\pi]} \|(e^{\iota\omega} I - A)^{-2} B\|_{\text{op}}\right) \leq \pi/k\left(\max_{\omega \in [0,2\pi]} \|(e^{\iota\omega} I - A)^{-2} B\|_{\text{op}}\right)$$

since any $x \in [0, 1]$ is at most $1/(2k)$ from the nearest fraction i/k. This implies

$$\delta \leq \pi/k\left(\max_{\omega \in [0,2\pi]} \|(e^{\iota\omega} I - A)^{-2} B\|_{\text{op}}\right)$$

so:

$$\begin{aligned}\left\|\frac{1}{k}\mathbf{\Lambda}_k^{\text{freq}}(\theta,\boldsymbol{U})-\frac{1}{\bar{k}}\mathbf{\Lambda}_{\bar{k}}^{\text{freq}}(\theta,\boldsymbol{U}^\star)\right\|_{\text{op}} &\leq \frac{2\pi\|\theta\|_{\mathcal{H}_\infty}\gamma^2}{k}\left(\max_{\omega\in[0,2\pi]}\|(e^{\iota\omega}I-A)^{-2}B\|_{\text{op}}\right)\\ &\quad+\frac{\pi^2\gamma^2}{k^2}\left(\max_{\omega\in[0,2\pi]}\|(e^{\iota\omega}I-A)^{-2}B\|_{\text{op}}\right)^2\\ &\leq \frac{4\pi\|\theta\|_{\mathcal{H}_\infty}\gamma^2}{k}\left(\max_{\omega\in[0,2\pi]}\|(e^{\iota\omega}I-A)^{-2}B\|_{\text{op}}\right)\end{aligned}$$

where the last inequality holds so long as $k \geq \frac{\pi}{2\|\theta\|_{\mathcal{H}_\infty}}\left(\max_{\omega\in[0,2\pi]}\|(e^{\iota\omega}I-A)^{-2}B\|_{\text{op}}\right)$. To make this less than ϵ, we must choose:

$$k \geq \frac{4\pi\|\theta\|_{\mathcal{H}_\infty}\gamma^2}{\epsilon}\left(\max_{\omega\in[0,2\pi]}\|(e^{\iota\omega}I-A)^{-2}B\|_{\text{op}}\right)$$

Finally, note that the input $\widetilde{U}_{k,\ell}$ is feasible since:

$$\sum_{\ell=1}^{k}\text{tr}(U_{k,\ell}) = \frac{1}{\bar{k}^2}\sum_{\ell=1}^{k}\sum_{\ell'\in\ell_k^{-1}(\ell)}\text{tr}(U^\star_{\bar{k},\ell'}) \leq \frac{1}{\bar{k}^2}\sum_{\ell=1}^{\bar{k}}\text{tr}(U^\star_{\bar{k},\ell}) \leq \gamma^2$$

The conclusion follows immediately. □

Proof of Lemma 5.3.9. By definition of $\mathbf{\Lambda}_t^{\text{noise}}$,

$$\begin{aligned}&\left\|\frac{1}{T}\sum_{t=1}^{T}\mathbf{\Lambda}_t^{\text{noise}}(\theta,\sigma_u)-\frac{1}{T'}\sum_{t=1}^{T'}\mathbf{\Lambda}_t^{\text{noise}}(\theta,\sigma_u)\right\|_{\text{op}}\\ &\quad=\left\|\frac{1}{T}\sum_{t=1}^{T}\sum_{s=0}^{t-1}\left(\sigma_w^2A^s(A^s)^\top+\sigma_u^2A^sBB^\top(A^s)^\top\right)-\frac{1}{T'}\sum_{t=1}^{T'}\sum_{s=0}^{t-1}\left(\sigma_w^2A^s(A^s)^\top+\sigma_u^2A^sBB^\top(A^s)^\top\right)\right\|_{\text{op}}\\ &\quad=\left\|\frac{1}{T}\sum_{s=0}^{T-1}\left(\sigma_w^2(T-s)A^s(A^s)^\top+\sigma_u^2(T-s)A^sBB^\top(A^s)^\top\right)\right.\\ &\qquad\qquad\left.-\frac{1}{T'}\sum_{s=0}^{T'-1}\left(\sigma_w^2(T'-s)A^s(A^s)^\top+\sigma_u^2(T'-s)A^sBB^\top(A^s)^\top\right)\right\|_{\text{op}}\\ &\quad=\left\|\sum_{s=0}^{T-1}\left(\sigma_w^2(1-\frac{s}{T})A^s(A^s)^\top+\sigma_u^2(1-\frac{s}{T})A^sBB^\top(A^s)^\top\right)\right.\\ &\qquad\qquad\left.-\sum_{s=0}^{T'-1}\left(\sigma_w^2(1-\frac{s}{T'})A^s(A^s)^\top+\sigma_u^2(1-\frac{s}{T'})A^sBB^\top(A^s)^\top\right)\right\|_{\text{op}}\end{aligned}$$

$$
\begin{aligned}
&= \Bigg\| \sum_{s=0}^{T-1} \left(\left(\frac{s}{T'} - \frac{s}{T} \right) (\sigma_w^2 A^s (A^s)^\top + \sigma_u^2 A^s B B^\top (A^s)^\top) \right) \\
&\qquad\qquad - \sum_{s=T}^{T'-1} \left(\sigma_w^2 (1 - \frac{s}{T'}) A^s (A^s)^\top + \sigma_u^2 (1 - \frac{s}{T'}) A^s B B^\top (A^s)^\top \right) \Bigg\|_{\mathrm{op}} \\
&\leq |T'^{-1} - T^{-1}| \tau(A,\rho)^2 (\sigma_w^2 + \sigma_u^2 \|B\|_{\mathrm{op}}^2) \sum_{s=0}^{T-1} s\rho^{2s} + (\sigma_w^2 + \sigma_u^2 \|B\|_{\mathrm{op}}^2) \tau(A,\rho)^2 \sum_{s=T}^{T'-1} (1 - s/T') \rho^{2s} \\
&\leq |T'^{-1} - T^{-1}| \tau(A,\rho)^2 (\sigma_w^2 + \sigma_u^2 \|B\|_{\mathrm{op}}^2) \frac{\rho^2 + \rho^{2T+2} T}{(1-\rho^2)^2} \\
&\qquad + \tau(A,\rho)^2 (\sigma_w^2 + \sigma_u^2 \|B\|_{\mathrm{op}}^2) \frac{\rho^{2T'+2} + \rho^{2T+2} T + \rho^{2T} T'}{(1-\rho^2)^2 T'} \\
&\leq \frac{4\tau(A,\rho)^2 (\sigma_w^2 + \sigma_u^2 \|B\|_{\mathrm{op}}^2)}{(1-\rho^2)^2} \left(T^{-1} + \rho^{2T} \right)
\end{aligned}
$$

If we set T large enough such that

$$
\frac{4\tau(A,\rho)^2 (\sigma_w^2 + \sigma_u^2 \|B\|_{\mathrm{op}}^2)}{(1-\rho^2)^2} T^{-1} \leq \epsilon/2, \qquad \frac{4\tau(A,\rho)^2 (\sigma_w^2 + \sigma_u^2 \|B\|_{\mathrm{op}}^2)}{(1-\rho^2)^2} \rho^{2T} \leq \epsilon/2
$$

the desired bound will hold. Rearranging these gives the result. □

C.2 Deferred Proofs from Section 5.4.1

C.2.1 Sequential Open-Loop Policies Satisfy Assumption 4.4

Proof of Lemma 5.4.1. Let:

$$
\Sigma_i := \sum_{t=\bar{t}_{i-1}}^{\bar{t}_i - 1} z_t z_t^\top, i = 1, \ldots, n, \qquad \Sigma_{t:t'} := \sum_{s=t}^{t'} z_s z_s^\top
$$

Sufficient Excitation: We first show that the sufficient excitation condition is met by π_{exp}. First, note that by Lemma C.2.2 and some algebra, for any time t, $\Sigma_t \preceq T\beta_\infty I$ with probability at least $1 - \delta/n$. Fix a time $t \geq T_{\mathrm{se}}(\pi_{\mathrm{exp}})$ where $T_{\mathrm{se}}(\pi_{\mathrm{exp}})$ is defined as above. Since $\pi_{\mathrm{exp}} \in \Pi_{\gamma^2}^{\mathrm{sol}}$, the low-switching condition implies that there exists some set of epochs $\{i_1, \ldots, i_m\} \subseteq [n]$, such that for $j \in [m]$, $\bar{t}_{i_j+1} - \bar{t}_{i_j} \geq \frac{1}{2} T_{\mathrm{se}}(\pi_{\mathrm{exp}})$, $\bar{t}_{i_j+1} \leq t$, and

$$
\sum_{j=1}^{m} (\bar{t}_{i_j+1} - \bar{t}_{i_j}) \geq \frac{1}{2} t
$$

This follows directly from the fact that, for any t_0, there exists some epoch $i \in [n]$ such that $|\{t_0, t_0 + T_{\mathrm{se}}(\pi_{\mathrm{exp}}) - 1\} \cap \{\bar{t}_i, \dots, \bar{t}_{i+1} - 1\}| \geq \frac{1}{2} T_{\mathrm{se}}(\pi_{\mathrm{exp}})$. Now consider some $j \in [m]$. By Lemma C.2.1, if

$$\bar{t}_{i_j+1} - \bar{t}_{i_j} \geq c_1 d_x \Big((d_x + d_u) \log(\beta_\infty / \lambda_{\min}(\mathbf{\Lambda}^{\mathrm{noise}}_{d_x}(\widetilde{\theta}_\star, \Gamma_{u,i_j}) + 1) + \log \frac{n}{\delta} \Big) \tag{C.2.1}$$

we will have that, with probability at least $1 - \delta/n$, for some c_2,

$$\Sigma_{\bar{t}_{i_j}:\bar{t}_{i_j+1}} \succeq c_2 (\bar{t}_{i_j+1} - \bar{t}_{i_j}) \lambda_{\min}(\mathbf{\Lambda}^{\mathrm{noise}}_{d_x}(\widetilde{\theta}_\star, \Gamma_{u,i_j})))$$

However, by definition of $\mathbf{\Lambda}^{\mathrm{noise}}$ and since by assumption $\lambda_{\min}(\Gamma_{u,i_j}) \geq \sigma_u^2$, we will have $\lambda_{\min}(\mathbf{\Lambda}^{\mathrm{noise}}_{d_x}(\widetilde{\theta}_\star, \Gamma_{u,i_j}) \geq \lambda_{\mathrm{noise}}(\sigma_u)$, which implies that

$$\log(\beta_\infty / \lambda_{\mathrm{noise}}(\sigma_u) + 1) \geq \log(\beta_\infty / \lambda_{\min}(\mathbf{\Lambda}^{\mathrm{noise}}_{d_x}(\widetilde{\theta}_\star, \Gamma_{u,i_j})) + 1).$$

As we know that $t_{i_j+1} - t_{i_j} \geq \frac{1}{2} T_{\mathrm{se}}(\pi_{\mathrm{exp}})$, it follows that (C.2.1) is met, so we conclude that with probability at least $1 - \delta/n$,

$$\Sigma_{\bar{t}_{i_j}:\bar{t}_{i_j+1}} \succeq c_2 (\bar{t}_{i_j+1} - \bar{t}_{i_j}) \lambda_{\mathrm{noise}}(\sigma_u)$$

Union bounding over this event holding for each $j \in [m]$, it follows that with probability at least $1 - \delta$,

$$\Sigma_t \succeq \sum_{j=1}^{m} \Sigma_{\bar{t}_{i_j}:\bar{t}_{i_j+1}} \succeq \sum_{j=1}^{m} c_2 (\bar{t}_{i_j+1} - \bar{t}_{i_j}) \lambda_{\mathrm{noise}}(\sigma_u) \geq \frac{c_2}{2} t \lambda_{\mathrm{noise}}(\sigma_u)$$

By proper choice of constants, we will then have that the sufficient excitation condition of Assumption 4.4 is met with

$$T_{\mathrm{se}}(\pi_{\mathrm{exp}}) = c_1 d_x \Big((d_x + d_u) \log(\beta_\infty / \lambda_{\min}(\mathbf{\Lambda}^{\mathrm{noise}}_{d_x}(\widetilde{\theta}_\star, \sigma_u)) + 1) + \log \frac{n}{\delta} \Big)$$
$$\underline{\lambda} = c_2 \lambda_{\min}(\mathbf{\Lambda}^{\mathrm{noise}}_{d_x}(\widetilde{\theta}_\star, \sigma_u)).$$

Concentration of Covariates: We define the following events for some ϵ_i to be specified.

$$\mathcal{E} = \{\|\mathbf{\Sigma}_T - \mathbb{E}_{\theta_\star, \pi_{\mathrm{exp}}}[\mathbf{\Sigma}_T]\|_{\mathrm{op}} \leq \tfrac{C_{\mathrm{con}}}{T^\alpha} \lambda_{\min}(\mathbb{E}_{\theta_\star, \pi_{\mathrm{exp}}}[\mathbf{\Sigma}_T])\} \qquad \text{(Good event)}$$
$$\mathcal{E}_1 = \{\lambda_{\min}(\mathbf{\Sigma}_T) \geq c_2 \lambda_{\min}(\mathbf{\Lambda}^{\mathrm{noise}}_{d_x}(\widetilde{\theta}_\star, \sigma_u)) T\} \qquad \text{(Sufficient excitation)}$$
$$\mathcal{E}_{2,i} = \{\|\mathbf{\Sigma}_i - \mathbb{E}_{\theta_\star, \pi_{\mathrm{exp}}}[\mathbf{\Sigma}_i]\|_{\mathrm{op}} \leq \epsilon_i\} \qquad \text{(Concentration of covariates)}$$
$$\mathcal{E}_{3,i} = \left\{ \|z_{\bar{t}_{i-1}}\|_2 \leq \sqrt{\tfrac{c\tau_\star^2 \gamma^2 T}{1 - \rho_\star^2}} \right\} \qquad \text{(Bounded states)}$$

We would like to show that $\mathcal{E}$ holds with high probability. The following is trivial.

$$\mathbb{P}[\mathcal{E}^c] \le \mathbb{P}[\mathcal{E}^c \cap \mathcal{E}_1 \cap (\cap_{i=1}^n \mathcal{E}_{2,i}) \cap (\cap_{i=1}^n \mathcal{E}_{3,i})] + \mathbb{P}[\mathcal{E}_1^c] + \sum_{i=1}^n \mathbb{P}[\mathcal{E}_{3,i} \cap \mathcal{E}_{2,i}^c] + \sum_{i=1}^n \mathbb{P}[\mathcal{E}_{3,i}^c]$$

We first show that $\mathcal{E}_1, \mathcal{E}_{2,i}, \mathcal{E}_{3,i}$ hold with high probability.

Note first that, by what we have just shown, $\mathbb{P}[\mathcal{E}_1^c] \le \delta$ as long as $T \ge T_{\mathrm{se}}(\pi_{\mathrm{exp}})$. By Lemma C.2.7, since $\pi_{\mathrm{exp}} \in \Pi_{\gamma^2}^{\mathrm{sol}}$ and Definition 5.4.1, we will have that $\mathbb{P}[\mathcal{E}_{3,i}^c] \le \delta$ as long as

$$T \ge (n + \sqrt{d_x}\sigma_w^2/\gamma^2) \log \frac{2(n+1)}{\delta} \tag{C.2.2}$$

Next, we show that $\mathbb{P}[\mathcal{E}_{3,i} \cap \mathcal{E}_{2,i}^c] \le \delta$. Setting

$$\epsilon_i = \left(\frac{c_3 \tau_\star \|\mathbf{\Lambda}_T^{\mathrm{noise}}(\widetilde{\theta}_\star, \Gamma_{u,i})\|_{\mathrm{op}}}{1-\rho_\star} \sqrt{\bar{t}_i - \bar{t}_{i-1}} + \frac{c_4 \tau_\star^2 (\sqrt{T}\gamma + \sqrt{(c\tau_\star^2\gamma^2 T)(1-\rho^2)})}{(1-\rho_\star)^2} \max\{\sigma_w, \sqrt{\|\Gamma_u\|_{\mathrm{op}}}\} \right)$$
$$\cdot \sqrt{\log\frac{1}{\delta} + d_x + d_u} + \frac{c_5 \max\{\sigma_w^2, \|\Gamma_u\|_{\mathrm{op}}\}\tau_\star^2}{(1-\rho_\star)^2} (\log\frac{1}{\delta} + d_x + d_u)$$

Lemma C.2.3 with $\rho = \rho_\star$ implies directly that $\mathbb{P}[\mathcal{E}_{3,i} \cap \mathcal{E}_{2,i}^c] \le \delta$. For future convenience, we can upper bound ϵ_i by

$$\frac{c_3 \tau_\star^3 (\sigma_w^2 + \gamma^2)\sqrt{T}}{(1-\rho_\star)^{5/2}} \sqrt{\log\frac{1}{\delta} + d_x + d_u}$$

as long as

$$T \ge \log\frac{1}{\delta} + d_x + d_u \tag{C.2.3}$$

On the event $\mathcal{E}_1 \cap (\cap_{i=1}^n \mathcal{E}_{2,i}) \cap (\cap_{i=1}^n \mathcal{E}_{3,i})$, we have

$$\|\mathbf{\Sigma}_T - \mathbb{E}_{\theta_\star, \pi_{\mathrm{exp}}}[\mathbf{\Sigma}_T]\|_{\mathrm{op}} = \Big\| \sum_{i=1}^n \mathbf{\Sigma}_i - \sum_{i=1}^n \mathbb{E}_{\theta_\star, \pi_{\mathrm{exp}}}[\mathbf{\Sigma}_i] \Big\|_{\mathrm{op}} \le \sum_{i=1}^n \|\mathbf{\Sigma}_i - \mathbb{E}_{\theta_\star, \pi_{\mathrm{exp}}}[\mathbf{\Sigma}_i]\|_{\mathrm{op}} \le \sum_{i=1}^n \epsilon_i$$
$$\le n \frac{c_3 \tau_\star^3 (\sigma_w^2 + \gamma^2)\sqrt{T}}{(1-\rho_\star)^{5/2}} \sqrt{\log\frac{1}{\delta} + d_x + d_u}$$

and

$$\mathbb{E}_{\theta_\star, \pi_{\mathrm{exp}}}[\mathbf{\Sigma}_T] \succeq \mathbb{E}_{\theta_\star, \pi_{\mathrm{exp}}}[\mathbb{I}\{\mathcal{E}_1\}\mathbf{\Sigma}_T] \ge \mathbb{P}[\mathcal{E}_1] c_2 \lambda_{\min}(\mathbf{\Lambda}_{d_x}^{\mathrm{noise}}(\widetilde{\theta}_\star, \sigma_u)) T \cdot I \ge \frac{1}{2}\lambda_{\min}(\mathbf{\Lambda}_{d_x}^{\mathrm{noise}}(\widetilde{\theta}_\star, \sigma_u)) T \cdot I$$

Thus, $\mathbb{P}[\mathcal{E}^c \cap \mathcal{E}_1 \cap (\cap_{i=1}^n \mathcal{E}_{2,i}) \cap (\cap_{i=1}^n \mathcal{E}_{3,i})] = 0$ with $\alpha = 1/2$ and

$$C_{\text{con}} := n \frac{c_3 \tau_\star^3 (\sigma_w^2 + \gamma^2)}{(1-\rho_\star)^{5/2} \lambda_{\min}(\mathbf{\Lambda}_{d_x}^{\text{noise}}(\widetilde{\theta}_\star, \sigma_u))} \sqrt{\log \frac{1}{\delta} + d_x + d_u}$$

Thus, $\mathbb{P}[\mathcal{E}^c] \leq 2n + 1$. Rescaling δ and setting $T_{\text{con}}(\pi_{\text{exp}})$ to guarantee (C.2.2) and (C.2.3) hold gives the result.

□

C.2.2 Concentration of Covariates

Lemma C.2.1 (Lemma E.3 of Wagenmaker and Jamieson [2020])**.** *Assume that our system θ is driven by some input $u_t = \widetilde{u}_t + u_t^w$ where $\widetilde{u}_t$ is deterministic and $u_t^w \sim \mathcal{N}(0, \Gamma_u)$. Then on the event that $\sum_{t=1}^T z_t z_t^\top \preceq T\bar{\mathbf{\Lambda}}_T$, for some $\bar{\mathbf{\Lambda}}_T$, choosing k so that:*

$$T \geq \frac{25600}{27} k \left(2(d_x + d_u) \log \frac{200}{3} + \log \det(\bar{\mathbf{\Lambda}}_T (\mathbf{\Lambda}_k^{\text{noise}})^{-1}) + \log \frac{1}{\delta} \right) \tag{C.2.4}$$

we will have with probability less than δ:

$$\sum_{t=1}^T z_t z_t^\top \not\preceq \frac{27}{25600} T \mathbf{\Lambda}_k^{\text{noise}}(\theta, \Gamma_u).$$

Lemma C.2.2. *Assume that we are playing a policy $\pi_{\text{exp}} \in \Pi_{\gamma^2}^{\text{sol}}$ and that $\delta \in (0, 1/3)$. Then with probability at least $1 - \delta$, and assuming we start from some state $x_0 = 0$,*

$$\begin{aligned}
\sum_{t=1}^T z_t z_t^\top &\preceq T \frac{c\tau(\widetilde{A}_\star, \rho)^2}{1-\rho} \left(\sqrt{d_x} \log \frac{T}{\delta} + \frac{\gamma^2}{1-\rho} \right) \cdot I \\
&\preceq cT \left((1 + \|B_\star\|_{\text{op}}^2) \|A_\star\|_{\mathcal{H}_\infty}^4 \right) \left(\sqrt{d_x} \log \frac{T}{\delta} + \gamma^2 \|A_\star\|_{\mathcal{H}_\infty}^2 \right) \cdot I.
\end{aligned}$$

Proof. Note that we can break the state into the portion driven by the conditionally non-random input, $\widetilde{u}_t$, and the process noise and random input. We denote these components as $z_t^{\boldsymbol{u}}$ and $z_t^{\boldsymbol{w}}$. Then

$$\left\| \sum_{t=1}^T z_t z_t^\top \right\|_{\text{op}} \leq \sum_{t=1}^T \|z_t\|_2^2 \leq 2 \sum_{t=1}^T (\|z_t^{\boldsymbol{u}}\|_2^2 + \|z_t^{\boldsymbol{w}}\|_2^2)$$

Note that the input $\widetilde{u}_t$ will almost surely satisfy $\sum_{t=0}^{T-1} \widetilde{u}_t^\top \widetilde{u}_t \leq T\gamma^2$. We can then apply

Lemma C.2.6 to get that

$$\sum_{t=1}^{T} \|z_t^{\boldsymbol{u}}\|_2^2 \le \frac{4\tau(\widetilde{A}_\star, \rho)^2 \gamma^2 T}{(1-\rho)^2}$$

To bound the component $\sum_{t=1}^{T} \|z_t^{\boldsymbol{w}}\|_2^2$, we apply Lemma C.2.7 with $\gamma^2 = 0$, union bounding over all T steps. We simplify the bound by upper bounding n by T and using that $\delta \in (0, 1/3)$ implies $\log T/\delta \ge 1$. The second bound follows by Lemma 5.3.1. □

Lemma C.2.3. *Consider the system*

$$x_{t+1} = Ax_t + Bu_t + \Gamma_w^{1/2} w_t$$

where $A \in \mathbb{R}^{d\times d}, w_t \sim \mathcal{N}(0, I)$ and u_t is deterministic and satisfies $\sum_{t=1}^{T} u_t^\top u_t \le T\gamma^2$. Assume that we start from some state x_0. Then we will have that, with probability at least $1-\delta$

$$\begin{aligned}\Big\| \sum_{t=0}^{T} x_t x_t^\top - \mathbb{E}\sum_{t=0}^{T} x_t x_t^\top \Big\|_{\mathrm{op}} &\le \frac{c_3 \|\Gamma_w\|_{\mathrm{op}} \tau(A,\rho)^2}{(1-\rho)^2}(\log\frac{1}{\delta} + d) \\ &+ \left(\frac{c_1 \tau(A,\rho) \|\boldsymbol{\Lambda}_T^{\mathrm{noise}}(\theta, 0)\|_{\mathrm{op}}}{1-\rho}\sqrt{T} + \frac{c_2 \tau(A,\rho)^2 (\sqrt{T}\gamma\|B\|_{\mathrm{op}} + \|x_0\|_2)}{(1-\rho)^2}\sqrt{\|\Gamma_w\|_{\mathrm{op}}} \right) \sqrt{\log\frac{1}{\delta} + d}\end{aligned}$$

for universal constants c_1, c_2, c_3.

Proof. Consider the systems

$$x_{t+1}^{\boldsymbol{u}} = Ax_t^{\boldsymbol{u}} + Bu_t, \quad x_{t+1}^{\boldsymbol{w}} = Ax_t^{\boldsymbol{w}} + \Gamma_w^{1/2} w_t$$

and note that $x_t = x_t^{\boldsymbol{u}} + x_t^{\boldsymbol{w}}$. Therefore,

$$\sum_{t=0}^{T} x_t x_t^\top = \sum_{t=0}^{T} x_t^{\boldsymbol{u}} x_t^{\boldsymbol{u}\top} + \sum_{t=0}^{T} \left(x_t^{\boldsymbol{u}} x_t^{\boldsymbol{w}\top} + x_t^{\boldsymbol{w}} x_t^{\boldsymbol{u}\top} \right) + \sum_{t=0}^{T} x_t^{\boldsymbol{w}} x_t^{\boldsymbol{w}\top}, \quad \mathbb{E}\sum_{t=0}^{T} x_t x_t^\top = \sum_{t=0}^{T} x_t^{\boldsymbol{u}} x_t^{\boldsymbol{u}\top} + \mathbb{E}\sum_{t=0}^{T} x_t^{\boldsymbol{w}} x_t^{\boldsymbol{w}\top}$$

The second equality is true as $x_t^{\boldsymbol{u}}$ is deterministic and $x_t^{\boldsymbol{w}}$ is mean 0. Fix some $v \in \mathcal{S}^{d-1}$. By Lemma C.2.4 and Lemma C.2.5, we'll have, simultaneously with probability $1-\delta$:

$$\left| \sum_{t=1}^{T} (v^\top x_t^{\boldsymbol{w}})^2 - \mathbb{E}\sum_{t=1}^{T} (v^\top x_t^{\boldsymbol{w}})^2 \right| \le \frac{2\tau(A,\rho)\|\boldsymbol{\Lambda}_T^{\mathrm{noise}}(\theta,0)\|_{\mathrm{op}}}{1-\rho^2}\sqrt{T\log\frac{4}{\delta}} + \frac{8\|\Gamma_w\|_{\mathrm{op}}\tau(A,\rho)^2}{(1-\rho)^2}\log\frac{4}{\delta}$$

$$\left| \sum_{t=1}^{T} v^\top x_t^{\boldsymbol{u}} x_t^{\boldsymbol{w}\top} v \right| \le \frac{\tau(A,\rho)^2(4\sqrt{T}\gamma\|B\|_{\mathrm{op}} + \|x_0^{\boldsymbol{u}}\|_2)}{(1-\rho)^2}\sqrt{2\|\Gamma_w\|_{\mathrm{op}}\log\frac{4}{\delta}}$$

Which implies that

$$\left|\sum_{t=0}^{T}(v^\top x_t)^2 - \mathbb{E}\sum_{t=0}^{T}(v^\top x_t)^2\right| \le \left|\sum_{t=1}^{T}(v^\top x_t^{\boldsymbol{w}})^2 - \mathbb{E}\sum_{t=1}^{T}(v^\top x_t^{\boldsymbol{w}})^2\right| + 2\left|\sum_{t=1}^{T} v^\top x_t^{\boldsymbol{u}} x_t^{\boldsymbol{w}\top} v\right|$$
$$\le \frac{2\tau(A,\rho)\|\mathbf{\Lambda}_T^{\text{noise}}(\theta,0)\|_{\text{op}}}{1-\rho^2}\sqrt{T\log\frac{4}{\delta}} + \frac{8\|\Gamma_w\|_{\text{op}}\tau(A,\rho)^2}{(1-\rho)^2}\log\frac{4}{\delta}$$
$$+ \frac{2\tau(A,\rho)^2(4\sqrt{T}\gamma\|B\|_{\text{op}} + \|x_0\|_2)}{(1-\rho)^2}\sqrt{2\|\Gamma_w\|_{\text{op}}\log\frac{4}{\delta}}$$

Note that if M is symmetric $\|M\|_{\text{op}} = \sup_{v\in\mathcal{S}^{d-1}}|v^\top M v|$. Fix v to be a vector for which this equality is attained. Let $\mathcal{T}$ be an ϵ-net of $\mathcal{S}^{d-1}$. Then we can then find some $v_0 \in \mathcal{T}$ such that $\|v - v_0\|_2 \le \epsilon$, and thus,

$$|v^\top M v - v_0^\top M v_0| \le |v^\top M v - v_0^\top M v| + |v_0^\top M v - v_0^\top M v_0| \le 2\|M\|_{\text{op}}\|v_0 - v\|_2 \le 2\epsilon\|M\|_{\text{op}}$$

Therefore,

$$|v_0^\top M v_0| \ge |v^\top M v| - |v^\top M v - v_0^\top M v_0| \le (1-2\epsilon)\|M\|_{\text{op}}$$

so $\|M\|_{\text{op}} \le \frac{1}{1-2\epsilon}\max_{v\in\mathcal{T}}|v^\top M v|$. Applying this in our setting and choosing $\epsilon = 1/2$, gives

$$\left\|\sum_{t=0}^{T} x_t^{\boldsymbol{w}} x_t^{\boldsymbol{w}\top} - \mathbb{E}\sum_{t=0}^{T} x_t^{\boldsymbol{w}} x_t^{\boldsymbol{w}\top}\right\|_{\text{op}} \le 2\max_{v\in\mathcal{T}}\left|\sum_{t=1}^{T}(v^\top x_t^{\boldsymbol{w}})^2 - \mathbb{E}\sum_{t=1}^{T}(v^\top x_t^{\boldsymbol{w}})^2\right|$$

By Corollary 4.2.13 of Vershynin [2018], we will have $|\mathcal{T}| \le 5^d$. Using our high probability bound on $\left|\sum_{t=1}^{T}(v^\top x_t^w)^2 - \mathbb{E}\sum_{t=1}^{T}(v^\top x_t^w)^2\right|$ given above, and union bounding over $\mathcal{T}$, we conclude that, with probability at least $1-\delta$

$$\left\|\sum_{t=0}^{T} x_t x_t^\top - \mathbb{E}\sum_{t=0}^{T} x_t x_t^\top\right\|_{\text{op}} \le \frac{2\tau(A,\rho)\|\mathbf{\Lambda}_T^{\text{noise}}(\theta,0)\|_{\text{op}}}{1-\rho^2}\sqrt{T(\log\frac{4}{\delta} + d\log 5)}$$
$$+ \frac{8\|\Gamma_w\|_{\text{op}}\tau(A,\rho)^2}{(1-\rho)^2}(\log\frac{4}{\delta} + d\log 5) + \frac{2\tau(A,\rho)^2(4\sqrt{T}\gamma\|B\|_{\text{op}} + \|x_0\|_2)}{(1-\rho)^2}\sqrt{2\|\Gamma_w\|_{\text{op}}(\log\frac{4}{\delta} + d\log 5)}.$$

□

Lemma C.2.4. *Consider the system*

$$x_{t+1} = Ax_t + \Gamma_w^{1/2} w_t$$

where $A \in \mathbb{R}^{d\times d}, w_t \sim \mathcal{N}(0, I)$, and assume $x_0 = 0$. Then, for any $v \in \mathcal{S}^{d-1}$, we'll have that,

with probability at least $1-\delta$

$$\left|\sum_{t=1}^{T}(v^\top x_t)^2 - \mathbb{E}\sum_{t=1}^{T}(v^\top x_t)^2\right| \le \frac{2\tau(A,\rho)\|\mathbf{\Lambda}_T^{\text{noise}}(\theta,0)\|_{\text{op}}}{1-\rho^2}\sqrt{T\log\frac{2}{\delta}} + \frac{8\|\Gamma_w\|_{\text{op}}\tau(A,\rho)^2}{(1-\rho)^2}\log\frac{2}{\delta}.$$

Proof. This is a direct consequence of the Hanson-Wright Inequality. Note that,

$$x_t = \sum_{s=0}^{t-1} A^{t-s-1}\Gamma_w^{1/2} w_s = G_t \mathfrak{w}_t$$

where we have defined $G_t := [A^{t-1}\Gamma_w^{1/2}, A^{t-2}\Gamma_w^{1/2}, \dots, A\Gamma_w^{1/2}, \Gamma_w^{1/2}, 0, \dots, 0] \in \mathbb{R}^{d\times dT}$, $\mathfrak{w}_T := [w_0^\top, w_1^\top, \dots, w_{T-2}^\top, w_{T-1}^\top]^\top$. So,

$$(v^\top x_t)^2 = \mathfrak{w}_T^\top G_t^\top v v^\top G_t \mathfrak{w}_T$$

and,

$$\sum_{t=1}^{T}(v^\top x_t)^2 = \sum_{t=1}^{T}\mathfrak{w}_T^\top G_t^\top v v^\top G_t \mathfrak{w}_T = \mathfrak{w}_T^\top \mathbf{G}_T \mathfrak{w}_T$$

where $\mathbf{G}_T := \sum_{t=1}^{T} G_t^\top v v^\top G_t$. The Hanson-Wright inequality then immediately gives that,

$$\mathbb{P}\left[\left|\sum_{t=1}^{T}(v^\top x_t)^2 - \mathbb{E}\sum_{t=1}^{T}(v^\top x_t)^2\right| \ge \epsilon\right] \le 2\exp\left(-c\min\left\{\frac{\epsilon^2}{\|\mathbf{G}_T\|_F^2}, \frac{\epsilon}{\|\mathbf{G}_T\|_{\text{op}}}\right\}\right)$$

For a fixed δ, rearranging gives

$$\mathbb{P}\left[\left|\sum_{t=1}^{T}(v^\top x_t)^2 - \mathbb{E}\sum_{t=1}^{T}(v^\top x_t)^2\right| \ge 2\sqrt{\|\mathbf{G}_T\|_F^2\log(2/\delta)} + 2\|\mathbf{G}_T\|_{\text{op}}\log(2/\delta)\right] \le \delta$$

We proceed to bound $\|\mathbf{G}_T\|_{\text{op}}$ and $\|\mathbf{G}_T\|_F^2$. Consider some $u \in \mathcal{S}^{dT-1}$ and note that, if we write $u = [u_0^\top, u_1^\top, \dots, u_{T-2}^\top, u_{T-1}^\top]^\top$, where $u_i \in \mathbb{R}^d$, using the definition of G_t given above, we have:

$$G_t u = \sum_{s=0}^{t-1} A^{t-s-1}\Gamma_w^{1/2} u_s = x_t^{\boldsymbol{u}}$$

where x_t^u is the state of the system with matrix A when the input u is played and there is no noise. Thus,

$$u^\top \mathbf{G}_T u = \sum_{t=1}^{T}(v^\top G_t u)^2 = \sum_{t=1}^{T}(v^\top x_t^{\boldsymbol{u}})^2 \le \lambda_{\max}\left(\sum_{t=1}^{T} x_t^{\boldsymbol{u}} x_t^{\boldsymbol{u}\top}\right)$$

Then, invoking Lemma C.2.6 with $\gamma^2 = \|\Gamma_w\|_{\text{op}}$ and $B = I$, we can bound,

$$\lambda_{\max}\left(\sum_{t=1}^{T} x_t^{\boldsymbol{u}} x_t^{\boldsymbol{u}\top}\right) \le \sum_{t=1}^{T} \|x_t^{\boldsymbol{u}}\|_2^2 \le \frac{4\tau(A,\rho)^2\|\Gamma_w\|_{\text{op}}}{(1-\rho)^2}$$

As this does not depend on u, it is a valid bound on $\|\mathbf{G}_T\|_{\text{op}}$:

$$\|\mathbf{G}_T\|_{\text{op}} \le \frac{4\tau(A,\rho)^2\|\Gamma_w\|_{\text{op}}}{(1-\rho)^2} \tag{C.2.5}$$

To bound $\|\mathbf{G}_T\|_F^2$, we can write,

$$\|\mathbf{G}_T\|_F^2 = \text{tr}(\mathbf{G}_T^\top \mathbf{G}_T) = \sum_{t=1}^{T}\sum_{s=1}^{T} \text{tr}(G_t^\top v v^\top G_t G_s^\top v v^\top G_s) = \sum_{t=1}^{T}\sum_{s=1}^{T} (v^\top G_t G_s^\top v)^2 \le \sum_{t=1}^{T}\sum_{s=1}^{T} \|G_t G_s^\top\|_{\text{op}}^2$$

From the definition of G, we have,

$$G_t G_s^\top = A^{\max\{t-s,0\}}\left(\sum_{k=0}^{\min\{t,s\}} A^k \Gamma_w A^{k\top}\right) A^{\max\{s-t,0\}^\top} = A^{\max\{t-s,0\}} \mathbf{\Lambda}^{\text{noise}}_{\min\{t,s\}}(\theta,0) A^{\max\{s-t,0\}^\top}$$

so,

$$\|G_t G_s^\top\|_{\text{op}} \le \|A^{\max\{t-s,0\}}\|_{\text{op}} \|A^{\max\{s-t,0\}}\|_{\text{op}} \|\mathbf{\Lambda}^{\text{noise}}_{\min\{t,s\}}(\theta,0)\|_{\text{op}} \le \tau(A,\rho)\rho^{|t-s|}\|\mathbf{\Lambda}^{\text{noise}}_{\min\{t,s\}}(\theta,0)\|_{\text{op}}$$

which implies

$$\begin{aligned}\sum_{t=1}^{T}\sum_{s=1}^{T} \|G_t G_s^\top\|_{\text{op}}^2 &\le \tau(A,\rho)^2 \|\mathbf{\Lambda}_T^{\text{noise}}(\theta,0)\|_{\text{op}}^2 \sum_{t=1}^{T}\sum_{s=1}^{T} \rho^{2|t-s|} \\ &= \tau(A,\rho)^2 \|\mathbf{\Lambda}_T^{\text{noise}}(\theta,0)\|_{\text{op}}^2 \frac{(1-\rho^4)T + 2\rho^{2(T+1)} - 2\rho^2}{(1-\rho^2)^2} \le \frac{\tau(A,\rho)^2 \|\mathbf{\Lambda}_T^{\text{noise}}(\theta,0)\|_{\text{op}}^2 T}{(1-\rho^2)^2}\end{aligned}$$

Combining everything, we have shown that, for any $v \in \mathcal{S}^{d-1}$,

$$\mathbb{P}\left[\left|\sum_{t=1}^{T}(v^\top x_t)^2 - \mathbb{E}\sum_{t=1}^{T}(v^\top x_t)^2\right| \ge \frac{2\tau(A,\rho)\|\mathbf{\Lambda}_T^{\text{noise}}(\theta,0)\|_{\text{op}}}{1-\rho^2}\sqrt{T\log\frac{2}{\delta}} + \frac{8\|\Gamma_w\|_{\text{op}}\tau(A,\rho)^2}{(1-\rho)^2}\log\frac{2}{\delta}\right] \le \delta$$

□

Lemma C.2.5. *Consider the systems*

$$x_{t+1}^{\boldsymbol{u}} = Ax_t^{\boldsymbol{u}} + Bu_t, \quad x_{t+1}^{\boldsymbol{w}} = Ax_t^{\boldsymbol{w}} + \Gamma_w^{1/2} w_t$$

where $A \in \mathbb{R}^{d\times d}, w_t \sim \mathcal{N}(0, I)$ and u_t a deterministic signal with $\sum_{t=1}^{T} u_t^\top u_t \leq T\gamma^2$. Assume that $x_0^{\boldsymbol{w}} = 0$. Then, for any $v \in \mathcal{S}^{d-1}$, we will have that, with probability at least $1-\delta$

$$\left|\sum_{t=1}^{T} v^\top x_t^{\boldsymbol{u}} x_t^{\boldsymbol{w}\top} v\right| \leq \frac{\tau(A,\rho)^2(4\sqrt{T}\gamma\|B\|_{\text{op}} + \|x_0^{\boldsymbol{u}}\|_2)}{(1-\rho)^2}\sqrt{2\|\Gamma_w\|_{\text{op}}\log\frac{2}{\delta}}.$$

Proof. We adopt the same notation as in the proof of Lemma C.2.4. Defining

$$G_t^u := [A^{t-1}B, A^{t-2}B, \ldots, AB, B, 0, \ldots, 0] \in \mathbb{R}^{d\times d_u T}, \quad \mathfrak{u}_T := [u_0^\top, u_1^\top, \ldots, u_{T-2}^\top, u_{T-1}^\top]^\top \in \mathbb{R}^{d_u T}$$

we have

$$x_t^{\boldsymbol{u}} = G_t^u \mathfrak{u}_T + A^t x_0^{\boldsymbol{u}}, \quad x_t^{\boldsymbol{w}} = G_t \mathfrak{w}_T$$

which implies

$$\sum_{t=1}^{T} v^\top x_t^{\boldsymbol{u}} x_t^{\boldsymbol{w}\top} v = \left(\mathfrak{u}_T^\top \sum_{t=1}^{T}(G_t^u)^\top v v^\top G_t + x_0^{u\top}\sum_{t=1}^{T} A^{t\top} v v^\top G_t\right)\mathfrak{w}_T =: \mathbf{g}^\top \mathfrak{w}_T \sim \mathcal{N}(0, \mathbf{g}^\top\mathbf{g})$$

By standard Gaussian concentration results, we then have that

$$\mathbb{P}\left[\left|\sum_{t=1}^{T} v^\top x_t^{\boldsymbol{u}} x_t^{\boldsymbol{w}\top} v\right| \geq \sqrt{2\mathbf{g}^\top\mathbf{g}\log\frac{2}{\delta}}\right] \leq \delta$$

It remains to bound $\mathbf{g}^\top\mathbf{g}$. To this end, note that

$$\mathbf{g}^\top\mathbf{g} \leq \left(\left\|\mathfrak{u}_T^\top \sum_{t=1}^{T}(G_t^u)^\top v v^\top G_t\right\|_2 + \left\|x_0^{\boldsymbol{u}\top}\sum_{t=1}^{T} A^{t\top} v v^\top G_t\right\|_2\right)^2$$

We can bound $\|A^t\|_{\text{op}} \leq \tau(A,\rho)\rho^t$ and,

$$\|G_t\|_{\text{op}} \leq \|\Gamma_w^{1/2}\|_{\text{op}}\sum_{s=0}^{t-1}\|A^s\|_{\text{op}} \leq \|\Gamma_w^{1/2}\|_{\text{op}}\tau(A,\rho)\sum_{s=0}^{t-1}\rho^s \leq \frac{\|\Gamma_w^{1/2}\|_{\text{op}}\tau(A,\rho)}{1-\rho}$$

so,

$$\left\|x_0^{\boldsymbol{u}\top}\sum_{t=1}^{T} A^{t\top} v v^\top G_t\right\|_2 \leq \frac{\|x_0^{\boldsymbol{u}}\|_2\|\Gamma_w^{1/2}\|_{\text{op}}\tau(A,\rho)^2}{1-\rho}\sum_{t=1}^{T}\rho^t \leq \frac{\|x_0^{\boldsymbol{u}}\|_2\|\Gamma_w^{1/2}\|_{\text{op}}\tau(A,\rho)^2}{(1-\rho)^2}$$

Furthermore, letting $G'_t = [A^{t-1}, A^{t-2}, \dots, A, I, 0, \dots, 0]$, we have

$$\left\| \mathfrak{u}_T^\top \sum_{t=1}^T (G_t^u)^\top v v^\top G_t \right\|_2 \le \|\mathfrak{u}_T\|_2 \|B\|_{\mathrm{op}} \|\Gamma_w^{1/2}\|_{\mathrm{op}} \left\| \sum_{t=1}^T (G'_t)^\top v v^\top G'_t \right\|_2 \le \|\mathfrak{u}_T\|_2 \|B\|_{\mathrm{op}} \|\Gamma_w^{1/2}\|_{\mathrm{op}} \|\mathbf{G}'_T\|_{\mathrm{op}}$$

where $\mathbf{G}'_T = \sum_{t=1}^T (G'_t)^\top v v^\top G'_t$. By (C.2.5), $\|\mathbf{G}'_T\|_{\mathrm{op}} \le 4\tau(A,\rho)^2/(1-\rho)^2$. Since we have assumed that $\sum_{t=1}^T u_t^\top u_t \le T\gamma^2$, we also have $\|\mathfrak{u}_T\|_2 \le \sqrt{T}\gamma$. Combining everything, we have shown that

$$\mathbf{g}^\top \mathbf{g} \le \|\Gamma_w\|_{\mathrm{op}} \left(\frac{4\sqrt{T}\gamma \|B\|_{\mathrm{op}} \tau(A,\rho)^2}{(1-\rho)^2} + \frac{\|x_0^{\boldsymbol{u}}\|_2 \tau(A,\rho)^2}{(1-\rho)^2} \right)^2$$

Thus,

$$\mathbb{P}\left[\left| \sum_{t=1}^T v^\top x_t^{\boldsymbol{u}} x_t^{\boldsymbol{w}\top} v \right| \ge \frac{\tau(A,\rho)^2 (4\sqrt{T}\gamma \|B\|_{\mathrm{op}} + \|x_0^{\boldsymbol{u}}\|_2)}{(1-\rho)^2} \sqrt{2\|\Gamma_w\|_{\mathrm{op}} \log \frac{2}{\delta}} \right] \le \delta$$

□

C.2.3 State Norm Bounds

Lemma C.2.6. *Consider the system*

$$x_{t+1} = Ax_t + Bu_t$$

and assume that we start at state $x_0 = 0$. Then if $\sum_{t=0}^{T-1} u_t^\top u_t \le T\gamma^2$, we will have

$$\sum_{t=1}^T \|x_t\|_2^2 \le \frac{4\tau(A,\rho)^2 \|B\|_{\mathrm{op}}^2 \gamma^2 T}{(1-\rho)^2}.$$

Proof. By definition $x_T = \sum_{s=0}^{T-1} A^{T-s-1} B u_s$, so

$$\sum_{t=1}^T \|x_t\|_2^2 \le \sum_{t=1}^T \left(\sum_{s=0}^{t-1} \|A^{t-s-1}\|_{\mathrm{op}} \|B\|_{\mathrm{op}} \|u_s\|_2 \right)^2 \le \tau(A,\rho)^2 \|B\|_{\mathrm{op}}^2 \sum_{t=1}^T \left(\sum_{s=0}^{t-1} \rho^{t-s-1} \|u_s\|_2 \right)^2$$

Letting $\rho_t = \rho^t$, $v_t = \|u_t\|_2$, we define $y_t = \sum_{s=0}^{t-1} \rho^{t-s-1} \|u_s\|_2 = (\rho * v)[t]$, where $*$ denotes convolution. By Parseval's Theorem,

$$\sum_{t=1}^T \left(\sum_{s=0}^{t-1} \rho^{t-s-1} \|u_s\|_2 \right)^2 = \sum_{t=1}^T y_t^2 = \frac{1}{T} \sum_{k=1}^T |Y_k|^2$$

where Y_k denotes the DFT of y_t. As convolution in the time domain is multiplication in the

frequency domain, we will have $Y_k = P_k V_k$ where P_k is the DFT of ρ_t and V_k is the DFT of v_t. We can explicitly calculate P_k as:

$$P_k = \sum_{t=0}^{T-1} \rho^t e^{-\iota \frac{2\pi k t}{T}} = \frac{1 - e^{-\iota 2\pi k} \rho^T}{1 - e^{-\iota \frac{2\pi k}{T}} \rho}$$

Thus,

$$\frac{1}{T} \sum_{k=1}^{T} |Y_k|^2 = \frac{1}{T} \sum_{k=1}^{T} \left| \frac{1 - e^{-\iota 2\pi k} \rho^T}{1 - e^{-\iota \frac{2\pi k}{T}} \rho} \right|^2 |V_k|^2$$

Note that, also by Parseval's Theorem, the constraint $\sum_{t=1}^{T} \|u_t\|_2^2 \leq \gamma^2$ translates to $\frac{1}{T} \sum_{k=1}^{T} |V_k|^2 \leq \gamma^2$. So,

$$\begin{aligned} \frac{1}{T} \sum_{k=1}^{T} \left| \frac{1 - e^{-\iota 2\pi k} \rho^T}{1 - e^{-\iota \frac{2\pi k}{T}} \rho} \right|^2 |V_k|^2 &\leq \max_{z : \|z\|_1 \leq T^2 \gamma^2} \frac{1}{T} \sum_{k=1}^{T} \left| \frac{1 - e^{-\iota 2\pi k} \rho^T}{1 - e^{-\iota \frac{2\pi k}{T}} \rho} \right|^2 z_k \\ &= \gamma^2 T \max_{k \in \{1, \ldots, T\}} \left| \frac{1 - e^{-\iota 2\pi k} \rho^T}{1 - e^{-\iota \frac{2\pi k}{T}} \rho} \right|^2 \leq \frac{4\gamma^2 T}{(1-\rho)^2} \end{aligned}$$

The conclusion follows. □

Lemma C.2.7. *Assume that we are playing a policy $\pi_{\text{exp}} \in \Pi^{\text{sol}}_{\gamma^2}$. Then with probability at least $1 - \delta$, assuming $z_0 = 0$,*

$$\|z_t\|_2^2 \leq \frac{c\tau(\widetilde{A}_\star, \rho)^2}{1 - \rho^2} \left(\gamma^2 T + (\sqrt{d_x}\sigma_w^2 + \sqrt{nT}\gamma^2) \sqrt{\log \frac{2(n+1)}{\delta}} + (\sigma_w^2 + n\gamma^2) \log \frac{2(n+1)}{\delta} \right)$$

and if

$$T \geq (n + \sqrt{d_x}\sigma_w^2/\gamma^2) \log \frac{2(n+1)}{\delta}$$

this bound can be simplified to

$$\|z_t\|_2^2 \leq \frac{c\tau(\widetilde{A}_\star, \rho)^2 \gamma^2 T}{1 - \rho^2}.$$

Proof. By Definition 5.4.1, for s in epoch j, we can always write the input u_s as $u_s = \widetilde{u}_s + u_s^w$, where $\widetilde{u}_s$ is $\mathcal{F}_{\bar{t}_j}$ measurable and $u_s^w \sim \mathcal{N}(0, \Gamma_{u,j})$. Given this, we break the state up into the component driven by $\widetilde{u}_s$, which we denote as $z_t^{\boldsymbol{u}}$, and the component driven by the process noise and u_s^w, which we denote as $z_t^{\boldsymbol{w}}$. By linearity, we will have that $z_t = z_t^{\boldsymbol{u}} + z_t^{\boldsymbol{w}}$, so

$$\|z_t\|_2^2 \leq 2\|z_t^{\boldsymbol{u}}\|_2^2 + 2\|z_t^{\boldsymbol{w}}\|_2^2$$

We can easily bound $\|z_t^{\boldsymbol{u}}\|_2^2$ as:

$$\|z_t^{\boldsymbol{u}}\|_2^2 = \Big\| \sum_{s=0}^{t-1} \widetilde{A}_\star^{t-s-1} \widetilde{B}_\star u_s \Big\|_2^2 \le \tau(\widetilde{A}_\star, \rho)^2 \Big(\sum_{s=0}^{t-1} \rho^{t-s-1} \|u_s\|_2 \Big)^2$$
$$\le \tau(\widetilde{A}_\star, \rho)^2 \Big(\sum_{s=0}^{t-1} \rho^{2(t-s-1)} \Big) \Big(\sum_{s=0}^{t-1} \|u_s\|_2^2 \Big) \le \frac{\tau(\widetilde{A}_\star, \rho)^2 \gamma^2 T}{1-\rho^2}$$

where the final inequality follows since, by assumption, $\sum_{s=0}^{t-1} \|u_s\|_2^2 \le T\gamma^2$ almost surely. We now bound $\|z_t^{\boldsymbol{w}}\|_2^2$. Note that due to the possible correlations between $\Gamma_{u,j}$ and previous epochs, we cannot naively apply Gaussian concentration. We first upper bound $\|z_t^{\boldsymbol{w}}\|_2^2$ as

$$\|z_t^{\boldsymbol{w}}\|_2^2 = \Big\| \sum_{s=0}^{t-1} \widetilde{A}_\star^{t-s-1} (\widetilde{B}_\star u_s^w + w_s) \Big\|_2^2 \le 2 \Big\| \sum_{s=0}^{t-1} \widetilde{A}_\star^{t-s-1} \widetilde{B}_\star u_s^w \Big\|_2^2 + 2 \Big\| \sum_{s=0}^{t-1} \widetilde{A}_\star^{t-s-1} w_s \Big\|_2^2$$
$$\le 2\tau(\widetilde{A}_\star, \rho)^2 \Big(\sum_{s=0}^{t-1} \rho^{t-s-1} \|u_s^w\|_2 \Big)^2 + 2 \Big\| \sum_{s=0}^{t-1} \widetilde{A}_\star^{t-s-1} w_s \Big\|_2^2$$
$$\le 2\tau(\widetilde{A}_\star, \rho)^2 \Big(\sum_{s=0}^{t-1} \rho^{2(t-s-1)} \Big) \Big(\sum_{s=0}^{t-1} \|u_s^w\|_2^2 \Big) + 2 \Big\| \sum_{s=0}^{t-1} \widetilde{A}_\star^{t-s-1} w_s \Big\|_2^2$$

We note that $\| \sum_{s=0}^{t-1} \widetilde{A}_\star^{t-s-1} w_s \|_2^2$ is simply the norm of the the state of a dynamical system driven by noise w_s. We can therefore apply Lemma C.2.8 to get that with probability at least $1-\delta$,

$$\Big\| \sum_{s=0}^{t-1} \widetilde{A}_\star^{t-s-1} w_s \Big\|_2^2 \le 4 \sqrt{\|\boldsymbol{\Lambda}_t^{\text{noise}}(\widetilde{\theta}_\star, 0)\|_F^2 \log \frac{2}{\delta}} + 4 \|\boldsymbol{\Lambda}_t^{\text{noise}}(\widetilde{\theta}_\star, 0)\|_{\text{op}} \log \frac{2}{\delta}$$

We can upper bound

$$\|\boldsymbol{\Lambda}_t^{\text{noise}}(\widetilde{\theta}_\star, 0)\|_{\text{op}} = \sigma_w^2 \Big\| \sum_{s=0}^{t-1} \widetilde{A}_\star^t (\widetilde{A}_\star^t)^\top \Big\|_{\text{op}} \le \sigma_w^2 \tau(\widetilde{A}_\star, \rho)^2 \sum_{s=0}^{t-1} \rho^{2t} \le \frac{\sigma_w^2 \tau(\widetilde{A}_\star, \rho)^2}{1-\rho^2}$$
$$\|\boldsymbol{\Lambda}_t^{\text{noise}}(\widetilde{\theta}_\star, 0)\|_F^2 \le d_x \|\boldsymbol{\Lambda}_t^{\text{noise}}(\widetilde{\theta}_\star, 0)\|_{\text{op}}^2$$

To bound $\sum_{s=0}^{t-1} \|u_s^w\|_2^2$ we can apply Hanson-Wright to some epoch j to get that

$$\sum_{s=\bar{t}_j}^{\bar{t}_{j+1}-1} \|u_s^w\|_2^2 \le \mathbb{E} \sum_{s=\bar{t}_j}^{\bar{t}_{j+1}-1} \|u_s^w\|_2^2 + 2\sqrt{(\bar{t}_{j+1} - \bar{t}_j) \|\Gamma_{u,j}\|_F^2 \log \frac{2}{\delta}} + 2\|\Gamma_{u,j}\|_{\text{op}} \log \frac{2}{\delta}$$

$$= \sum_{s=\bar{t}_j}^{\bar{t}_{j+1}-1} \operatorname{tr}(\Gamma_{u,j}) + 2\sqrt{(\bar{t}_{j+1}-\bar{t}_j)\|\Gamma_{u,j}\|_F^2 \log\frac{2}{\delta}} + 2\|\Gamma_{u,j}\|_{\mathrm{op}} \log\frac{2}{\delta}$$

$$\leq (\bar{t}_{j+1}-\bar{t}_j)\gamma^2 + 2\sqrt{(\bar{t}_{j+1}-\bar{t}_j)\gamma^4 \log\frac{2}{\delta}} + 2\gamma^2 \log\frac{2}{\delta}$$

Assume that t occurs in epoch i. Then if this bound holds for all epoch $j \leq i$, we can bound

$$\sum_{s=0}^{t-1} \|u_s^w\|_2^2 \leq t\gamma^2 + \sum_{j=0}^{i} \sqrt{(\bar{t}_{j+1}-\bar{t}_j)\gamma^4 \log\frac{2}{\delta}} + 2i\gamma^2 \log\frac{2}{\delta} \leq t\gamma^2 + \sqrt{it\gamma^4 \log\frac{2}{\delta}} + 2i\gamma^2 \log\frac{2}{\delta}$$

Union bounding over all n epochs and the bound on the process noise, we then have that with probability at least $1-\delta$,

$$\|z_t^{\boldsymbol{w}}\|_2^2 \leq \frac{c\tau(\widetilde{A}_\star,\rho)^2}{1-\rho^2}\Big(\gamma^2 t + (\sqrt{d_x}\sigma_w^2 + \sqrt{it}\gamma^2)\sqrt{\log\frac{2(n+1)}{\delta}} + (\sigma_w^2 + i\gamma^2)\log\frac{2(n+1)}{\delta}\Big)$$

The result the follows by combining this with our bound on $\|z_t^{\boldsymbol{u}}\|_2^2$, and upper bounding t by T and i by n. The simplified bound holds by noting that for large enough T, we can upper bound the two lower order terms in the bound on $\|z_t^{\boldsymbol{w}}\|_2^2$ by $(d_x\sigma_w^2+\gamma)T$. □

Lemma C.2.8. *Consider the system*

$$x_{t+1} = Ax_t + \Gamma_w^{1/2} w_t$$

where $w_t \sim \mathcal{N}(0, I)$ and assume that we start at state $x_0 = 0$. Then, with probability at least $1-\delta$

$$\|x_T\|_2^2 \leq 2\sqrt{\|\boldsymbol{\Lambda}_T^{\mathrm{noise}}(\theta,0)\|_F^2 \log\frac{2}{\delta}} + 2\|\boldsymbol{\Lambda}_T^{\mathrm{noise}}(\theta,0)\|_{\mathrm{op}} \log\frac{2}{\delta}.$$

Proof. Using the same notation as in the proof of Lemma C.2.4 and C.2.5, we will have that $x_T = G_T \mathfrak{w}_T$. Applying Hanson-Wright then gives that, with probability at least $1-\delta$,

$$\left|{x_T}^\top x_T - \operatorname{tr}(G_T^\top G_T)\right| \leq 2\sqrt{\|G_T^\top G_T\|_F^2 \log\frac{2}{\delta}} + 2\|G_T^\top G_T\|_{\mathrm{op}} \log\frac{2}{\delta}$$

By definition of G_T we have that

$$G_T^\top G_T = \sum_{s=0}^{T-1} \Gamma_w^{1/2} (A^s)^\top A^s \Gamma_w^{1/2}, \quad G_T G_T^\top = \sum_{s=0}^{T-1} A^s \Gamma_w (A^s)^\top = \boldsymbol{\Lambda}_T^{\mathrm{noise}}(\theta,0)$$

So,

$$\|G_T^\top G_T\|_{\mathrm{op}} = \|G_T G_T^\top\|_{\mathrm{op}} = \|\boldsymbol{\Lambda}_T^{\mathrm{noise}}(\theta,0)\|_{\mathrm{op}}, \quad \|G_T^\top G_T\|_F^2 = \operatorname{tr}(G_T G_T^\top G_T G_T^\top) = \|\boldsymbol{\Lambda}_T^{\mathrm{noise}}(\theta,0)\|_F^2$$

This concludes the proof. □

Theorem C.1 (Hanson-Wright Inequality, Vershynin [2018]). *Let $X \in \mathbb{R}^d$ be a random vector with independent, mean-zero, sub-Gaussian coordinates. Let $A \in \mathbb{R}^{d\times d}$. Then, for every $\epsilon \geq 0$, we have*

$$\mathbb{P}\left[|X^\top AX - \mathbb{E}X^\top AX| \geq \epsilon\right] \leq 2\exp\left(-c\min\left\{\frac{\epsilon^2}{K^4\|A\|_F^2}, \frac{\epsilon}{K^2\|A\|_{\text{op}}}\right\}\right)$$

where $K = \max_i \|X_i\|_{\psi_2}$.

Recall that, if X_i is gaussian with variance σ_w^2, $\|X_i\|_{\psi_2} \leq C\sigma_w$.

C.3 Deferred Proofs from Section 5.4.2

Lemma C.3.1 (Matrix Perturbation Bound). *Assume $A, B \in \mathbb{S}^d_{++}$, $\|A - B\|_{\text{op}} \leq \epsilon$, and $\epsilon < \lambda_{\min}(B)$. Then*

$$\|A^{-1} - B^{-1}\|_{\text{op}} \leq \frac{\epsilon}{\lambda_{\min}(B)(\lambda_{\min}(B) - \epsilon)}$$

Proof. Denote $\Delta = A - B$. By the matrix inversion lemma:

$$A^{-1} = (B + \Delta)^{-1} = B^{-1} - B^{-1}(B^{-1} + \Delta^{-1})^{-1}B^{-1}$$

so:

$$\begin{aligned}\|A^{-1} - B^{-1}\|_{\text{op}} &= \|B^{-1}(B^{-1} + \Delta^{-1})^{-1}B^{-1}\|_{\text{op}} \\ &\leq \|B^{-1}\|_{\text{op}}^2\|(B^{-1} + \Delta^{-1})^{-1}\|_{\text{op}} \\ &= \frac{1}{\lambda_{\min}(B)^2\sigma_d(B^{-1} + \Delta^{-1})}\end{aligned}$$

However:

$$\sigma_d(B^{-1} + \Delta^{-1}) \geq \sigma_d(\Delta^{-1}) - \sigma_1(B^{-1}) = \frac{1}{\|\Delta\|_{\text{op}}} - \frac{1}{\lambda_{\min}(B)} = \frac{\lambda_{\min}(B) - \|\Delta\|_{\text{op}}}{\lambda_{\min}(B)\|\Delta\|_{\text{op}}}$$

Since we have assumed $\epsilon < \lambda_{\min}(B)$ and since $\|\Delta\|_{\text{op}} \leq \epsilon$, this lower bound on $\sigma_d(B^{-1} + \Delta^{-1})$ will be positive, so:

$$\frac{1}{\lambda_{\min}(B)^2\sigma_d(B^{-1} + \Delta^{-1})} \leq \frac{\lambda_{\min}(B)\|\Delta\|_{\text{op}}}{\lambda_{\min}(B)^2(\lambda_{\min}(B) - \|\Delta\|_{\text{op}})} \leq \frac{\|\Delta\|_{\text{op}}}{\lambda_{\min}(B)(\lambda_{\min}(B) - \|\Delta\|_{\text{op}})}$$

The result follows since $\|\Delta\|_{\text{op}} \leq \epsilon$. □

C.3.1 Operator Norm Estimation

Lemma C.3.2. *Let*

$$\widehat{\theta}_{\text{ls}} = \min_{A,B} \sum_{t=1}^{T} \|x_{t+1} - Ax_t - Bu_t\|_2^2$$

Then on the event

$$\mathcal{E} := \left\{ \lambda_{\min}(\Sigma_T) \geq \underline{\lambda} T, \Sigma_T \preceq T\bar{\mathbf{\Lambda}}_T \right\}$$

with probability at least $1 - \delta$*:*

$$\|\widehat{\theta}_{\text{ls}} - \theta_\star\|_{\text{op}} \leq C\sqrt{\frac{\log(1/\delta) + d_x + \log\det(\bar{\mathbf{\Lambda}}_T/\underline{\lambda} + I)}{\underline{\lambda} T}}.$$

Proof. Define the following events:

$$\mathcal{A} = \left\{ \|\widehat{\theta}_i - \theta_\star\|_{\text{op}} \leq C\sqrt{\frac{\log(1/\delta) + d_x + \log\det(\bar{\mathbf{\Lambda}}_T/\underline{\lambda} + I)}{\underline{\lambda} T}} \right\}$$

$$\mathcal{E}_1 = \left\{ \left\| \left(\sum_{t=1}^{T} z_t z_t^\top \right)^{-1/2} \sum_{t=1}^{T} z_t w_t^\top \right\|_{\text{op}} \leq c_2 \sigma_w \sqrt{\log \frac{1}{\delta} + d_x + \log\det(\bar{\mathbf{\Lambda}}_T/\underline{\lambda} + I)} \right\}$$

Our goal is to show that $\mathbb{P}[\mathcal{A}^c \cap \mathcal{E}] \leq \delta$. The following is trivial.

$$\mathbb{P}[\mathcal{A}^c \cap \mathcal{E}] \leq \mathbb{P}[\mathcal{A}^c \cap \mathcal{E} \cap \mathcal{E}_1] + \mathbb{P}[\mathcal{E} \cap \mathcal{E}_1^c]$$

As $\widehat{\theta}_{\text{ls}}$ is the least squares estimate, we will have that $\widehat{\theta}_{\text{ls}}^\top = (\sum_{t=1}^{T} z_t z_t^\top)^{-1} \sum_{t=1}^{T} z_t x_{t+1}^\top = \theta_\star^\top + (\sum_{t=T-T_i}^{T} z_t z_t^\top)^{-1} \sum_{t=1}^{T} z_t w_t^\top$. Given this, the error can be decomposed as:

$$\begin{aligned}
\|\widehat{\theta}_{\text{ls}} - \theta_\star\|_{\text{op}} &= \left\| \left(\sum_{t=1}^{T} z_t z_t^\top \right)^{-1} \sum_{t=1}^{T} z_t w_t^\top \right\|_{\text{op}} \leq \left\| \left(\sum_{t=1}^{T} z_t z_t^\top \right)^{-1/2} \right\|_{\text{op}} \left\| \left(\sum_{t=1}^{T} z_t z_t^\top \right)^{-1/2} \sum_{t=1}^{T} z_t w_t^\top \right\|_{\text{op}} \\
&= \left\| \left(\sum_{t=1}^{T} z_t z_t^\top \right)^{-1/2} \sum_{t=1}^{T} z_t w_t^\top \right\|_{\text{op}} / \sqrt{\lambda_{\min}\left(\sum_{t=1}^{T} z_t z_t^\top \right)}
\end{aligned}$$

It follows that, on the event $\mathcal{E} \cap \mathcal{E}_1$, the error bound given in $\mathcal{A}$ holds. Thus, $\mathbb{P}[\mathcal{A}^c \cap \mathcal{E} \cap \mathcal{E}_1] = 0$. Proposition 3.1 implies that $\mathbb{P}[\mathcal{E} \cap \mathcal{E}_1^c] \leq \delta$, so $\mathbb{P}[\mathcal{A}^c \cap \mathcal{E}] \leq \delta$. □

C.3.2 Optimality of Inputs

Proof of Lemma 5.4.2. We will show that the following set of inequalities hold for large enough T and arbitrary $T' \geq T$:

$$\begin{aligned}
&\operatorname{tr}\Big(\mathcal{H}(\theta_\star)(\mathbb{E}_{\theta_\star,\boldsymbol{U}_i,T_i}[I_{d_x}\otimes\textstyle\sum_{t=T-T_i}^{T} z_t z_t^\top])^{-1}\Big) && \text{(C.3.1)}\\
&\overset{(a)}{\leq} \frac{1}{(1-\epsilon)^2 T_i}\operatorname{tr}\left(\mathcal{H}(\theta_\star)\check{\boldsymbol{\Lambda}}^{\mathrm{ss}}_{T_i,T_i/d_u}(\widetilde{\theta}_\star,\boldsymbol{U}_i,\gamma/\sqrt{2d_u})^{-1}\right) && \text{(Steady-state)}\\
&\overset{(b)}{\leq} \min_{\boldsymbol{U}\in\mathcal{U}_{\gamma^2/2,k_i}} \frac{1}{(1-\epsilon)^2 T_i}\operatorname{tr}\left(\mathcal{H}(\theta_\star)\check{\boldsymbol{\Lambda}}^{\mathrm{ss}}_{T_i,T_i/d_u}(\widetilde{\theta}_\star,\boldsymbol{U},\gamma/\sqrt{2d_u})^{-1}\right) + \frac{C_{\mathrm{exp}}}{(1-\epsilon)^2} && \text{(Optimal inputs)}\\
&\overset{(c)}{\leq} \min_{\boldsymbol{U}\in\mathcal{U}_{\gamma^2/2,T'}} \frac{1}{(1-\epsilon)^3 T_i}\operatorname{tr}\left(\mathcal{H}(\theta_\star)\check{\boldsymbol{\Lambda}}^{\mathrm{ss}}_{T',T'}(\widetilde{\theta}_\star,\boldsymbol{U},\gamma/\sqrt{2d_u})^{-1}\right) + \frac{C_{\mathrm{exp}}}{(1-\epsilon)^2} && \text{(Infinite horizon)}\\
&\overset{(d)}{\leq} \min_{\boldsymbol{U}\in\mathcal{U}_{\gamma^2,T'}} \frac{2}{(1-\epsilon)^3 T_i}\operatorname{tr}\left(\mathcal{H}(\theta_\star)\check{\boldsymbol{\Lambda}}^{\mathrm{ss}}_{T',T'}(\widetilde{\theta}_\star,\boldsymbol{U},0)^{-1}\right) + \frac{C_{\mathrm{exp}}}{(1-\epsilon)^2} && \text{(Noiseless inputs)}
\end{aligned}$$

Steady-state: By definition,

$$\mathbb{E}_{\theta_\star,\boldsymbol{U}_i,T_i}[I_{d_x}\otimes\textstyle\sum_{t=T-T_i}^{T} z_t z_t^\top] = T_i\check{\boldsymbol{\Lambda}}_{T_i}(\widetilde{\theta}_\star,\boldsymbol{U}_i,\gamma/\sqrt{2d_u},z_{T-T_i})$$

Note that the inputs played by Algorithm 5.1 are constructed as in Lemma C.3.3. It follows then that, by Lemma C.3.3, as long as,

$$\begin{aligned}
T_i \geq \max\Bigg\{ &\frac{2d_u\tau(\widetilde{A}_\star,\rho)^3(2k_i\gamma+\|z_{T-T_i}\|_2)((4+2\tau(\widetilde{A}_\star,\rho))k_i\gamma+\tau(\widetilde{A}_\star,\rho)\|z_{T-T_i}\|_2))}{(1-\rho^{k_i})^2(1-\rho)\lambda_{\min}(\boldsymbol{\Lambda}^{\mathrm{noise}}_{T_i/2}(\widetilde{\theta}_\star,\gamma/\sqrt{2d_u}))\epsilon},\\
&\left(\frac{\tau(\widetilde{A}_\star,\rho)^2k_i^2d_u\gamma^2}{(1-\rho^{k_i})^2}+\frac{\tau(\widetilde{A}_\star,\rho)k_id_u\gamma^2\sqrt{T_i/d_u}}{1-\rho^{k_i}}\right)\frac{16d_u\|\widetilde{\theta}_\star\|^2_{\mathcal{H}_\infty}}{\lambda_{\min}(\boldsymbol{\Lambda}^{\mathrm{noise}}_{T_i/2}(\theta_\star,\gamma/\sqrt{2d_u}))\epsilon}\Bigg\}
\end{aligned}$$

we will have

$$\boldsymbol{\Lambda}_{T_i}(\widetilde{\theta}_\star,\boldsymbol{U}_i,\gamma/\sqrt{2d_u},z_{T-T_i}) \succeq (1-\epsilon)^2\boldsymbol{\Lambda}^{\mathrm{ss}}_{T_i,T_i/d_u}(\widetilde{\theta}_\star,\boldsymbol{U}_i,\gamma/\sqrt{2d_u})$$

This implies that

$$\check{\boldsymbol{\Lambda}}_{T_i}(\widetilde{\theta}_\star,\boldsymbol{U}_i,\gamma/\sqrt{2d_u},z_{T-T_i}) \succeq (1-\epsilon)^2\check{\boldsymbol{\Lambda}}^{\mathrm{ss}}_{T_i,T_i/d_u}(\widetilde{\theta}_\star,\boldsymbol{U}_i,\gamma/\sqrt{2d_u})$$

from which (a) follows.

Optimal inputs: We next apply Lemma C.3.4, which bounds the suboptimality of certainty equivalent experiment design, to show (b). We instantiate Lemma C.3.4 with

$$\underline{\lambda} = \lambda_{\min}(\mathbf{\Lambda}_k^{\text{noise}}(\widetilde{\theta}_\star, \gamma/\sqrt{2d_u}))/2, \quad \mathbf{\Lambda}(\theta, \boldsymbol{U}) = \check{\mathbf{\Lambda}}^{\text{ss}}_{T_i, T_i/d_u}(\theta, \boldsymbol{U}, \gamma/\sqrt{2d_u})$$

Note that Lemma 5.3.7 gives that the smoothness condition (C.3.3) holds for $L_{\text{cov}}(\theta_\star, \gamma^2)$ as defined in (5.3.6). Furthermore, it is clear that, as long as $T_i \geq 2k$, we will have $\lambda_{\min}(\mathbf{\Lambda}(\widetilde{\theta}_\star, \boldsymbol{U})) \geq \underline{\lambda}$ for all $\boldsymbol{U}$. To apply Lemma C.3.4, we need

$$\|\widehat{\theta}_{i-1} - \theta_\star\|_{\text{op}} \leq \min\{r_{\text{cov}}(\theta_\star), \lambda_{\min}(\mathbf{\Lambda}_k^{\text{noise}}(\widetilde{\theta}_\star, \gamma/\sqrt{2d_u}))/(4L_{\text{cov}}(\theta_\star, \gamma^2))\}, \quad \|\widehat{\theta}_{i-1} - \theta_\star\|_F \leq r_{\text{quad}}(\theta_\star)$$

where we choose to instantiate Lemma C.3.4 in the operator norm, and since the matrix Frobenius norm coincides with the vector 2-norm. The condition on $\|\widehat{\theta}_{i-1} - \theta_\star\|_{\text{op}}$ will hold as long as our assumption on $\epsilon_{\text{op},i-1}$ holds. Since $\widehat{\theta}_i - \theta_\star$ is at most rank d_x, we have

$$\|\widehat{\theta}_{i-1} - \theta_\star\|_F \leq \sqrt{d_x}\|\widehat{\theta}_{i-1} - \theta_\star\|_{\text{op}} \leq r_{\text{quad}}(\theta_\star)$$

where the last inequality again holds so long as our assumption on $\epsilon_{\text{op},i-1}$ holds. Then, since we design the input $\boldsymbol{U}_i$ on the estimate $\widehat{\theta}_{i-1}$, the conditions of Lemma C.3.4 are met for $\boldsymbol{U}_i$, so

$$\begin{aligned}
&\frac{1}{T_i}\left|\text{tr}\left(\mathcal{H}(\theta_\star)\check{\mathbf{\Lambda}}^{\text{ss}}_{T_i, T_i/d_u}(\widetilde{\theta}_\star, \boldsymbol{U}_i, \gamma/\sqrt{2d_u})^{-1}\right) - \min_{\boldsymbol{U} \in \mathcal{U}_{\gamma^2/2, k_i}} \text{tr}\left(\mathcal{H}(\theta_\star)\check{\mathbf{\Lambda}}^{\text{ss}}_{T_i, T_i/d_u}(\widetilde{\theta}_\star, \boldsymbol{U}, \gamma/\sqrt{2d_u})^{-1}\right)\right| \\
&\quad \leq \left(\frac{4\sqrt{d_x}(d_x^2 + d_x d_u)L_{\text{hess}}\lambda_{\min}(\mathbf{\Lambda}_k^{\text{noise}}(\widetilde{\theta}_\star, \gamma/\sqrt{2d_u}) + 8L_{\text{cov}}(\theta_\star, \gamma^2)\text{tr}(\mathcal{H}(\theta_\star))}{\lambda_{\min}(\mathbf{\Lambda}_k^{\text{noise}}(\widetilde{\theta}_\star, \gamma/\sqrt{2d_u})^2}\right)\frac{\epsilon_{\text{op},i-1}}{T_i} \\
&\qquad\qquad + \frac{8(d_x^2 + d_x d_u)L_{\text{cov}}(\theta_\star, \gamma^2)L_{\text{hess}}}{\lambda_{\min}(\mathbf{\Lambda}_k^{\text{noise}}(\widetilde{\theta}_\star, \gamma/\sqrt{2d_u})^2}\frac{\epsilon_{\text{op},i-1}^2}{T_i} \\
&\quad =: C_{\text{exp}}
\end{aligned}$$

and thus (b) holds.

Infinite horizon: Now,

$$\mathbf{\Lambda}^{\text{ss}}_{T_i, T_i/d_u}(\widetilde{\theta}_\star, \boldsymbol{U}, \gamma/\sqrt{2d_u}) = \frac{d_u}{T_i}\mathbf{\Lambda}^{\text{freq}}_{T_i/d_u}(\widetilde{\theta}_\star, \boldsymbol{U}) + \frac{1}{T_i}\sum_{t=1}^{T_i}\mathbf{\Lambda}_t^{\text{noise}}(\widetilde{\theta}_\star, \gamma/\sqrt{2d_u})$$

If $\boldsymbol{U} \in \mathcal{U}_{\gamma^2/2, k_i}$, since $T_i/(d_u k_i)$ is an integer,

$$\mathbf{\Lambda}^{\text{freq}}_{T_i/d_u}(\widetilde{\theta}_\star, \boldsymbol{U}) = \frac{T_i}{d_u k_i^2}\sum_{\ell=1}^{k_i}(e^{\iota\frac{2\pi\ell}{k_i}}I - \widetilde{A}_\star)^{-1}\widetilde{B}_\star U_\ell \widetilde{B}_\star^\top (e^{\iota\frac{2\pi\ell}{k_i}}I - \widetilde{A}_\star)^{-\mathsf{H}}$$

Then, by Lemma 5.3.8, as long as

$$k_i \geq \max\left\{\frac{8\pi\|\widetilde{\theta}_\star\|_{\mathcal{H}_\infty}\gamma^2}{\lambda_{\min}(\mathbf{\Lambda}_k^{\mathrm{noise}}(\widetilde{\theta}_\star, \gamma/\sqrt{2d_u}))\epsilon}, \frac{\pi}{2\|\widetilde{\theta}_\star\|_{\mathcal{H}_\infty}}\right\}\left(\max_{\omega\in[0,2\pi]}\|(e^{\iota\omega}I - \widetilde{A}_\star)^{-2}\widetilde{B}\|_{\mathrm{op}}\right)$$

then for any $T' \geq k_i$ and $\boldsymbol{U}^\star \in \mathcal{U}_{\gamma^2/2,T'}$, there exists a feasible $\boldsymbol{U}' \in \mathcal{U}_{\gamma^2/2,k_i}$ such that

$$\left\|\frac{d_u}{T_i}\mathbf{\Lambda}^{\mathrm{freq}}_{T_i/d_u}(\widetilde{\theta}_\star, \boldsymbol{U}') - \frac{1}{T'}\mathbf{\Lambda}^{\mathrm{freq}}_{T'}(\widetilde{\theta}_\star, \boldsymbol{U}^\star)\right\|_{\mathrm{op}} \leq \frac{\epsilon}{2}\lambda_{\min}(\mathbf{\Lambda}_k^{\mathrm{noise}}(\widetilde{\theta}_\star, \gamma/\sqrt{2d_u}))$$

Furthermore, by Lemma 5.3.9, if

$$T_i \geq \max\left\{\frac{16\tau(\widetilde{A}_\star,\rho)^2(\sigma_w^2 + \gamma^2/(2d_u))}{(1-\rho^2)^2\lambda_{\min}(\mathbf{\Lambda}_k^{\mathrm{noise}}(\widetilde{\theta}_\star, \gamma/\sqrt{2d_u}))\epsilon}, \log\left(\frac{(1-\rho^2)^2\lambda_{\min}(\mathbf{\Lambda}_k^{\mathrm{noise}}(\widetilde{\theta}_\star, \gamma/\sqrt{2d_u}))\epsilon}{16\tau(\widetilde{A}_\star,\rho)^2(\sigma_w^2 + \gamma^2/(2d_u))}\right)\frac{1}{2\log\rho}\right\}$$

then, for any $T' \geq T_i$,

$$\left\|\frac{1}{T_i}\sum_{t=1}^{T_i}\mathbf{\Lambda}_t^{\mathrm{noise}}(\widetilde{\theta}_\star, \gamma/\sqrt{2d_u}) - \frac{1}{T'}\sum_{t=1}^{T'}\mathbf{\Lambda}_t^{\mathrm{noise}}(\widetilde{\theta}_\star, \gamma/\sqrt{2d_u})\right\|_{\mathrm{op}} \leq \frac{\epsilon}{2}\lambda_{\min}(\mathbf{\Lambda}_k^{\mathrm{noise}}(\widetilde{\theta}_\star, \gamma/\sqrt{2d_u}))$$

Now note that, by definition,

$$\min_{\boldsymbol{U}\in\mathcal{U}_{\gamma^2/2,k_i}} \mathrm{tr}\left(\mathcal{H}(\theta_\star)\check{\mathbf{\Lambda}}^{\mathrm{ss}}_{T_i,T_i/d_u}(\widetilde{\theta}_\star, \boldsymbol{U}, \gamma/\sqrt{2d_u})^{-1}\right) \leq \mathrm{tr}\left(\mathcal{H}(\theta_\star)\check{\mathbf{\Lambda}}^{\mathrm{ss}}_{T_i,T_i/d_u}(\widetilde{\theta}_\star, \boldsymbol{U}', \gamma/\sqrt{2d_u})^{-1}\right)$$

By what we've just shown, for any $T' \geq T_i$

$$\begin{aligned}
&\check{\mathbf{\Lambda}}^{\mathrm{ss}}_{T_i,T_i/d_u}(\widetilde{\theta}_\star, \boldsymbol{U}', \gamma/\sqrt{2d_u}) \\
&\succeq \check{\mathbf{\Lambda}}^{\mathrm{ss}}_{T',T'}(\widetilde{\theta}_\star, \boldsymbol{U}^\star, \gamma/\sqrt{2d_u}) - \frac{\epsilon}{2}\lambda_{\min}(\mathbf{\Lambda}_k^{\mathrm{noise}}(\widetilde{\theta}_\star, \gamma/\sqrt{2d_u}))\cdot I \\
&\succeq (1-\epsilon)\check{\mathbf{\Lambda}}^{\mathrm{ss}}_{T',T'}(\widetilde{\theta}_\star, \boldsymbol{U}^\star, \gamma/\sqrt{2d_u}) + \epsilon\frac{1}{2T'}\sum_{t=1}^{T'}\mathbf{\Lambda}_t^{\mathrm{noise}}(\widetilde{\theta}_\star, \gamma/\sqrt{2d_u}) - \frac{\epsilon}{2}\lambda_{\min}(\mathbf{\Lambda}_k^{\mathrm{noise}}(\widetilde{\theta}_\star, \gamma/\sqrt{2d_u}))\cdot I \\
&\succeq (1-\epsilon)\check{\mathbf{\Lambda}}^{\mathrm{ss}}_{T',T'}(\widetilde{\theta}_\star, \boldsymbol{U}^\star, \gamma/\sqrt{2d_u}) + \frac{\epsilon}{2}\lambda_{\min}(\mathbf{\Lambda}_{T'/2}^{\mathrm{noise}}(\widetilde{\theta}_\star, \gamma/\sqrt{2d_u}))\cdot I - \frac{\epsilon}{2}\lambda_{\min}(\mathbf{\Lambda}_k^{\mathrm{noise}}(\widetilde{\theta}_\star, \gamma/\sqrt{2d_u}))\cdot I \\
&\succeq (1-\epsilon)\check{\mathbf{\Lambda}}^{\mathrm{ss}}_{T',T'}(\widetilde{\theta}_\star, \boldsymbol{U}^\star, \gamma/\sqrt{2d_u})
\end{aligned}$$

From this (c) follows directly.

Noiseless inputs: Finally, since

$$\check{\mathbf{\Lambda}}^{\mathrm{ss}}_{T',T'}(\widetilde{\theta}_\star, \boldsymbol{U}^\star, \gamma/\sqrt{2d_u}) \succeq \check{\mathbf{\Lambda}}^{\mathrm{ss}}_{T',T'}(\widetilde{\theta}_\star, \boldsymbol{U}^\star, 0)$$

we can bound

$$\min_{\boldsymbol{U}\in\mathcal{U}_{\gamma^2/2,T'}} \operatorname{tr}\left(\mathcal{H}(\theta_\star)\check{\boldsymbol{\Lambda}}^{\mathrm{ss}}_{T',T'}(\widetilde{\theta}_\star,\boldsymbol{U},\gamma/\sqrt{2d_u})^{-1}\right) \le \min_{\boldsymbol{U}\in\mathcal{U}_{\gamma^2/2,T'}} \operatorname{tr}\left(\mathcal{H}(\theta_\star)\check{\boldsymbol{\Lambda}}^{\mathrm{ss}}_{T',T'}(\widetilde{\theta}_\star,\boldsymbol{U},0)^{-1}\right)$$
$$\le 2\min_{\boldsymbol{U}\in\mathcal{U}_{\gamma^2,T'}} \operatorname{tr}\left(\mathcal{H}(\theta_\star)\check{\boldsymbol{\Lambda}}^{\mathrm{ss}}_{T',T'}(\widetilde{\theta}_\star,\boldsymbol{U},0)^{-1}\right)$$

which proves (d). Finally, to simplify the bound, we note that $\lambda_{\min}(\boldsymbol{\Lambda}^{\mathrm{noise}}_k(\widetilde{\theta}_\star,\gamma/\sqrt{2d_u})) \ge \lambda_{\mathrm{noise}}$, and we choose $\rho=\rho_\star$. □

Lemma C.3.3. *Fix an input $\boldsymbol{U}\in\mathcal{U}_{\gamma^2,k}$ and consider a system $\theta=(A,B)$. If we start from some state x_0 and play the time domain input $u_t=$ `ConstructTimeInput`$(\boldsymbol{U},T/d_u,k)$ where* `ConstructTimeInput` *is defined in Algorithm 3.2 then, so long as T/d_u is divisible by k, and*

$$T \ge \max\left\{\frac{2d_u\tau(A,\rho)^3(2\|B\|_{\mathrm{op}}k\gamma+\|x_0\|_2)((4+2\tau(A,\rho))\|B\|_{\mathrm{op}}k\gamma+\tau(A,\rho)\|x_0\|_2))}{(1-\rho^k)^2(1-\rho)\lambda_{\min}(\boldsymbol{\Lambda}^{\mathrm{noise}}_{T/2}(\theta,\gamma/\sqrt{2d_u}))\epsilon},\right.$$
$$\left.\left(\frac{\tau(A,\rho)^2k^2d_u\gamma^2}{(1-\rho^k)^2}+\frac{\tau(A,\rho)kd_u\gamma^2\sqrt{T/d_u}}{1-\rho^k}\right)\frac{16d_u\|\theta\|^2_{\mathcal{H}_\infty}}{\lambda_{\min}(\boldsymbol{\Lambda}^{\mathrm{noise}}_{T/2}(\theta,\gamma/\sqrt{2d_u}))\epsilon}\right\}$$

we will have

$$\boldsymbol{\Lambda}_T(\theta,\boldsymbol{U},\gamma/\sqrt{2d_u},x_0) \succeq (1-\epsilon)^2\boldsymbol{\Lambda}^{\mathrm{ss}}_{T,T/d_u}(\theta,\boldsymbol{U},\gamma/\sqrt{2d_u}).$$

Proof. Let $\check{u}_{\ell,j}$ be defined as in Algorithm 3.2 for this $\boldsymbol{U}$, and denote $\boldsymbol{U}_j=(\check{u}_{\ell,j}\check{u}^{\mathsf{H}}_{\ell,j})_{\ell=1}^k$. Let $\widetilde{u}_{j,t}$ denote the time domain version of $\{\check{u}_{\ell,j}\}_{\ell=1}^k$, as is specified in Algorithm 3.2. Since $\boldsymbol{U}\in\mathcal{U}_{\gamma^2,k}$, some algebra shows that

$$\frac{1}{k}\sum_{t=1}^{k}\widetilde{u}^\top_{j,t+s}\widetilde{u}_{j,t+s} \le 2d_u\gamma^2$$

for any $s\ge 0$. We break up the sum of the response based on which input is being played:

$$\sum_{t=1}^{T}x^{\boldsymbol{u}}_t(x^{\boldsymbol{u}}_t)^\top = \sum_{j=1}^{d_u}\sum_{t=(j-1)T/d_u+1}^{jT/d_u}x^{\boldsymbol{u}}_t(x^{\boldsymbol{u}}_t)^\top$$

which gives:

$$\boldsymbol{\Lambda}_T(\theta,\boldsymbol{U},\gamma/\sqrt{2d_u},0) = \frac{1}{T}\sum_{j=1}^{d_u}\sum_{t=(j-1)T/d_u+1}^{jT/d_u}x^{\boldsymbol{u}}_t(x^{\boldsymbol{u}}_t)^\top + \frac{1}{T}\sum_{t=1}^{T}\boldsymbol{\Lambda}^{\mathrm{noise}}_t(\theta,\gamma/\sqrt{2d_u})$$
$$= \frac{1}{T}\sum_{j=1}^{d_u}\left(\boldsymbol{\Lambda}^{\mathrm{in}}_{T/d_u}(\theta,\boldsymbol{U}_j,x^{\boldsymbol{u}}_{(j-1)T/d_u}) + \frac{1}{d_u}\sum_{t=1}^{T}\boldsymbol{\Lambda}^{\mathrm{noise}}_t(\theta,\gamma/\sqrt{2d_u})\right)$$

If we start from some initial state x',

$$x_t^{\boldsymbol{u}} = A^t x' + \sum_{s=0}^{t-1} A^{t-s-1} B u_s =: x_t^{\boldsymbol{u},0} + \widetilde{x}_t^{\boldsymbol{u}}$$

then,

$$\boldsymbol{\Lambda}^{\text{in}}_{T/d_u}(\theta, \boldsymbol{U}_j, x^{\boldsymbol{u}}_{(j-1)T/d_u}) = \sum_{t=0}^{T} x_t^u (x_t^u)^\top = \sum_{t=0}^{T} \left[\widetilde{x}_t^{\boldsymbol{u}}(\widetilde{x}_t^{\boldsymbol{u}})^\top + \widetilde{x}_t^{\boldsymbol{u}}(x_t^{\boldsymbol{u},0})^\top + x_t^{\boldsymbol{u},0}(\widetilde{x}_t^{\boldsymbol{u}})^\top + x_t^{\boldsymbol{u},0}(x_t^{\boldsymbol{u},0})^\top\right]$$

Now,

$$\begin{aligned}\left\|\sum_{t=0}^{T} x_t^{\boldsymbol{u},0}(\widetilde{x}_t^{\boldsymbol{u}})^\top\right\|_{\text{op}} &\le \|x^{\boldsymbol{u}}_{(j-1)T/d_u}\|_2 \sum_{t=0}^{T} \|A^t\|_{\text{op}} \|\widetilde{x}_t^{\boldsymbol{u}}\|_2 \le \|x^{\boldsymbol{u}}_{(j-1)T/d_u}\|_2 \tau(A,\rho) \sum_{t=0}^{T} \rho^t \|\widetilde{x}_t^{\boldsymbol{u}}\|_2 \\ &\le \frac{2\tau(A,\rho)^2 \|x^{\boldsymbol{u}}_{(j-1)T/d_u}\|_2 \|B\|_{\text{op}} k\gamma}{(1-\rho^k)(1-\rho)}\end{aligned}$$

where the last inequality follows by Lemma A.1.2, which gives:

$$\|\widetilde{x}_t^{\boldsymbol{u}}\|_2 \le \frac{2\tau(A,\rho)\|B\|_{\text{op}} k\gamma}{1-\rho^k}$$

Furthermore,

$$\left\|\sum_{t=0}^{T} x_t^{\boldsymbol{u},0}(x_t^{\boldsymbol{u},0})^\top\right\|_{\text{op}} \le \|x^{\boldsymbol{u}}_{(j-1)T/d_u}\|_2^2 \sum_{t=0}^{T} \|A^t\|_{\text{op}}^2 \le \frac{\|x^{\boldsymbol{u}}_{(j-1)T/d_u}\|_2^2 \tau(A,\rho)^2}{1-\rho}$$

Therefore,

$$\begin{aligned}&\boldsymbol{\Lambda}^{\text{in}}_{T/d_u}(\theta, \boldsymbol{U}_j, x^{\boldsymbol{u}}_{(j-1)T/d_u}) + \frac{1}{d_u}\sum_{t=1}^{T} \boldsymbol{\Lambda}_t^{\text{noise}}(\theta, \gamma/\sqrt{2d_u}) \\ &\quad\succeq \boldsymbol{\Lambda}^{\text{in}}_{T/d_u}(\theta, \boldsymbol{U}_j, 0) + \frac{1}{d_u}\sum_{t=1}^{T} \boldsymbol{\Lambda}_t^{\text{noise}}(\theta, \gamma/\sqrt{2d_u}) \\ &\qquad\qquad - \frac{\tau(A,\rho)^2 \|x^{\boldsymbol{u}}_{(j-1)T/d_u}\|_2 (4\|B\|_{\text{op}} k\gamma + (1-\rho^k)\|x^{\boldsymbol{u}}_{(j-1)T/d_u}\|_2)}{(1-\rho^k)(1-\rho)} \\ &\quad\succeq (1-\epsilon)\boldsymbol{\Lambda}^{\text{in}}_{T/d_u}(\theta, \boldsymbol{U}_j, 0) + \frac{1-\epsilon}{d_u}\sum_{t=1}^{T} \boldsymbol{\Lambda}_t^{\text{noise}}(\theta, \gamma/\sqrt{2d_u}) + \frac{\epsilon T}{2d_u}\lambda_{\min}(\boldsymbol{\Lambda}^{\text{noise}}_{T/2}(\theta, \gamma/\sqrt{2d_u}))\cdot I\end{aligned}$$

$$- \frac{\tau(A,\rho)^2 \|x^{\boldsymbol{u}}_{(j-1)T/d_u}\|_2 (4\|B\|_{\mathrm{op}} k\gamma + (1-\rho^k)\|x^{\boldsymbol{u}}_{(j-1)T/d_u}\|_2)}{(1-\rho^k)(1-\rho)}$$

$$\succeq (1-\epsilon)\mathbf{\Lambda}^{\mathrm{in}}_{T/d_u}(\theta, \boldsymbol{U}_j, 0) + \frac{1-\epsilon}{d_u}\sum_{t=1}^{T} \mathbf{\Lambda}^{\mathrm{noise}}_t(\theta, \gamma/\sqrt{2d_u})$$

where the last inequality holds as long as

$$T \geq \frac{2d_u \tau(A,\rho)^2 \|x^{\boldsymbol{u}}_{(j-1)T/d_u}\|_2 (4\|B\|_{\mathrm{op}} k\gamma + (1-\rho^k)\|x^{\boldsymbol{u}}_{(j-1)T/d_u}\|_2)}{(1-\rho^k)(1-\rho)\lambda_{\min}(\mathbf{\Lambda}^{\mathrm{noise}}_{T/2}(\theta, \gamma/\sqrt{2d_u}))\epsilon} \tag{C.3.2}$$

By Proposition 3.2, since T/d_u is divisible by k, we will have that

$$\mathbf{\Lambda}^{\mathrm{freq}}_{T/d_u}(\theta, \boldsymbol{U}_j) = \frac{d_u}{T}\sum_{\ell=1}^{T} (e^{\iota\frac{2\pi\ell}{T}} I - A)^{-1} B U_{j,\ell} U_{j,\ell}^{\mathsf{H}} B^{\mathsf{H}} (e^{\iota\frac{2\pi\ell}{T}} I - A)^{-\mathsf{H}}$$

where $\{U_{j,\ell}\}_{\ell=1}^{T} = \mathfrak{F}(\widetilde{u}_{j,0}, \ldots, \widetilde{u}_{j,T/d_u})$. Then, by Lemma 5.3.6, we will have,

$$\begin{aligned}
&\|\mathbf{\Lambda}^{\mathrm{in}}_{T/d_u}(\theta, \boldsymbol{U}_j, 0) - \mathbf{\Lambda}^{\mathrm{freq}}_{T/d_u}(\theta, \boldsymbol{U}_j)\|_{\mathrm{op}} \\
&\qquad \leq \left(\frac{8\tau(A,\rho)^2 k^2 d_u \gamma^2}{(1-\rho^k)^2} + \frac{8\tau(A,\rho) k d_u \gamma^2 \sqrt{T/d_u}}{1-\rho^k}\right)\left(\max_{\omega\in[0,2\pi]} \|(e^{-\iota\omega} I - A)^{-1} B\|_{\mathrm{op}}^2\right) \\
&\qquad \leq \frac{\epsilon T}{2d_u}\lambda_{\min}(\mathbf{\Lambda}^{\mathrm{noise}}_{T/2}(\theta, \gamma/\sqrt{2d_u}))
\end{aligned}$$

where the last inequality is true so long as

$$T \geq \left(\frac{8\tau(A,\rho)^2 k^2 d_u \gamma^2}{(1-\rho^k)^2} + \frac{8\tau(A,\rho) k d_u \gamma^2 \sqrt{T/d_u}}{1-\rho^k}\right)\frac{2d_u \cdot \max_{\omega\in[0,2\pi]} \|(e^{\iota\omega} I - A)^{-1} B\|_{\mathrm{op}}^2}{\lambda_{\min}(\mathbf{\Lambda}^{\mathrm{noise}}_{T/2}(\theta, \gamma/\sqrt{2d_u}))\epsilon}$$

Thus,

$$\begin{aligned}
&\mathbf{\Lambda}^{\mathrm{in}}_{T/d_u}(\theta, \boldsymbol{U}_j, 0) + \frac{1}{d_u}\sum_{t=1}^{T} \mathbf{\Lambda}^{\mathrm{noise}}_t(\theta, \gamma/\sqrt{2d_u}) \\
&\qquad \succeq \mathbf{\Lambda}^{\mathrm{freq}}_{T/d_u}(\theta, \boldsymbol{U}_j) + \frac{1}{d_u}\sum_{t=1}^{T} \mathbf{\Lambda}^{\mathrm{noise}}_t(\theta, \gamma/\sqrt{2d_u}) - \frac{\epsilon T}{2d_u}\lambda_{\min}(\mathbf{\Lambda}^{\mathrm{noise}}_{T/2}(\theta, \gamma/\sqrt{2d_u}))\cdot I \\
&\qquad \succeq (1-\epsilon)\mathbf{\Lambda}^{\mathrm{freq}}_{T/d_u}(\theta, \boldsymbol{U}_j) + \frac{1-\epsilon}{d_u}\sum_{t=1}^{T} \mathbf{\Lambda}^{\mathrm{noise}}_t(\theta, \gamma/\sqrt{2d_u})
\end{aligned}$$

It follows that if (C.3.2) holds for each j,

$$\begin{aligned}\mathbf{\Lambda}_T(\theta, \boldsymbol{U}, \gamma/\sqrt{2d_u}, 0) &\succeq \frac{(1-\epsilon)^2}{T}\sum_{j=1}^{d_u} \mathbf{\Lambda}^{\mathrm{freq}}_{T/d_u}(\theta, \boldsymbol{U}_j) + \frac{(1-\epsilon)^2}{T}\sum_{t=1}^{T} \mathbf{\Lambda}^{\mathrm{noise}}_t(\theta, \gamma/\sqrt{2d_u}) \\ &= \frac{d_u(1-\epsilon)^2}{T}\mathbf{\Lambda}^{\mathrm{freq}}_{T/d_u}(\theta, \boldsymbol{U}) + \frac{(1-\epsilon)^2}{T}\sum_{t=1}^{T} \mathbf{\Lambda}^{\mathrm{noise}}_t(\theta, \gamma/\sqrt{2d_u}) \\ &= (1-\epsilon)^2 \mathbf{\Lambda}^{\mathrm{ss}}_{T,T/d_u}(\theta, \boldsymbol{U}, \gamma/\sqrt{2d_u})\end{aligned}$$

It remains to ensure that (C.3.2) holds by bounding $\|x^{\boldsymbol{u}}_{(j-1)T/d_u}\|_2$. Again by Lemma A.1.2, we have

$$\|x^{\boldsymbol{u}}_{(j-1)T/d_u}\|_2 \le \frac{2\tau(A,\rho)\|B\|_{\mathrm{op}} k\gamma}{1-\rho^k}\mathbb{I}\{j>1\} + \tau(A,\rho)\rho^{(j-1)T/d_u}\|x_0\|_2 \le \frac{\tau(A,\rho)(2\|B\|_{\mathrm{op}} k\gamma + \|x_0\|_2)}{1-\rho^k}$$

Some algebra gives the result. □

C.3.3 Certainty Equivalence Experiment Design

Lemma C.3.4. *Fix a nominal instance $\theta_\star$ and let $\widehat{\theta}$ be some instance such that $\|\theta_\star - \widehat{\theta}\|_\circ \le \epsilon_\circ$, for $\circ \in \{\mathrm{op}, 2\}$. Let $\mathbf{\Lambda}(\theta, \boldsymbol{U}) \in \mathcal{S}_+^{d_\theta}$ be a map that satisfies, for all $\boldsymbol{U} \in \mathcal{U}_{\gamma^2}$ and all θ with $\|\theta - \theta_\star\|_\circ \le r_{\mathrm{cov}}(\theta_\star)$,*

$$\|\mathbf{\Lambda}(\theta, \boldsymbol{U}) - \mathbf{\Lambda}(\theta_\star, \boldsymbol{U})\|_{\mathrm{op}} \le L_{\mathrm{cov}}(\theta_\star, \gamma^2) \cdot \|\theta - \theta_\star\|_\circ \tag{C.3.3}$$

Assume that for all $\boldsymbol{U} \in \mathcal{U}_{\gamma^2,k}$, $\lambda_{\min}(\mathbf{\Lambda}(\theta_\star, \boldsymbol{U})) \ge \underline{\lambda}$ and that $\epsilon_\circ < \min\{r_{\mathrm{cov}}(\theta_\star), \underline{\lambda}/(2L_{\mathrm{cov}}(\theta_\star, \gamma^2))\}$, $\epsilon_2 \le r_{\mathrm{quad}}(\theta_\star)$. Let:

$$\widehat{\boldsymbol{U}} = \operatorname*{arg\,min}_{\boldsymbol{U}\in\mathcal{U}_{\gamma^2,k}} \mathrm{tr}\left(\mathcal{H}(\widehat{\theta})\mathbf{\Lambda}(\widehat{\theta}, \boldsymbol{U})^{-1}\right), \quad \boldsymbol{U}^\star = \operatorname*{arg\,min}_{\boldsymbol{U}\in\mathcal{U}_{\gamma^2,k}} \mathrm{tr}\left(\mathcal{H}(\theta_\star)\mathbf{\Lambda}(\theta_\star, \boldsymbol{U})^{-1}\right)$$

Then, under Assumption 4.1, we have:

$$\begin{aligned}\left|\mathrm{tr}\left(\mathcal{H}(\theta_\star)\mathbf{\Lambda}(\theta_\star, \boldsymbol{U}^\star)^{-1}\right) - \mathrm{tr}\left(\mathcal{H}(\theta_\star)\mathbf{\Lambda}(\theta_\star, \widehat{\boldsymbol{U}})^{-1}\right)\right| &\le \frac{2d_\theta L_{\mathrm{hess}}}{\underline{\lambda}}\epsilon_2 + \frac{4L_{\mathrm{cov}}(\theta_\star, \gamma^2)\mathrm{tr}(\mathcal{H}(\theta_\star))}{\underline{\lambda}^2}\epsilon_\circ \\ &\quad + \frac{4d_\theta L_{\mathrm{cov}}(\theta_\star, \gamma^2)L_{\mathrm{hess}}}{\underline{\lambda}^2}\epsilon_\circ\epsilon_2.\end{aligned}$$

Proof. By Proposition 4.2, under Assumption 4.1 and since $\|\theta_\star - \widehat{\theta}\|_2 \le r_{\mathrm{quad}}(\theta_\star)$, we have

$$\|\mathcal{H}(\theta_\star) - \mathcal{H}(\widehat{\theta})\|_{\mathrm{op}} \le L_{\mathrm{hess}}\|\theta_\star - \widehat{\theta}\|_2$$

Furthermore, by Lemma C.3.1, (C.3.3), and since $\epsilon_\circ < \underline{\lambda}/(2L_{\text{cov}}(\theta_\star, \gamma^2))$,

$$\|\mathbf{\Lambda}(\widehat{\theta}, \boldsymbol{U})^{-1} - \mathbf{\Lambda}(\theta_\star, \boldsymbol{U})^{-1}\|_{\text{op}} \leq \frac{\|\mathbf{\Lambda}(\widehat{\theta}, \boldsymbol{U}) - \mathbf{\Lambda}(\theta_\star, \boldsymbol{U})\|_{\text{op}}}{\underline{\lambda}(\underline{\lambda} - L_{\text{cov}}(\theta_\star, \gamma^2)\epsilon_\circ)} \leq \frac{2L_{\text{cov}}(\theta_\star, \gamma^2)\epsilon_\circ}{\underline{\lambda}^2}$$

Thus, denoting $\Delta_{\mathcal{H}} = \mathcal{H}(\widehat{\theta}) - \mathcal{H}(\theta_\star)$ and $\Delta_{\Sigma^{-1}} = \mathbf{\Lambda}(\theta_\star, \boldsymbol{U})^{-1} - \mathbf{\Lambda}(\widehat{\theta}, \boldsymbol{U})^{-1}$, the above bounds and Von Neumann's trace inequality imply:

$$\begin{aligned}
&\left|\text{tr}\left(\mathcal{H}(\widehat{\theta})\mathbf{\Lambda}(\widehat{\theta}, \boldsymbol{U})^{-1}\right) - \text{tr}\left(\mathcal{H}(\theta_\star)\mathbf{\Lambda}(\theta_\star, \boldsymbol{U})^{-1}\right)\right| \\
&\leq |\text{tr}\left(\Delta_{\mathcal{H}}\mathbf{\Lambda}(\theta_\star, u)^{-1}\right)| + |\text{tr}(\mathcal{H}(\theta_\star)\Delta_{\Sigma^{-1}})| + |\text{tr}(\Delta_{\mathcal{H}}\Delta_{\Sigma^{-1}})| \\
&\leq L_{\text{hess}}\text{tr}\left(\mathbf{\Lambda}(\theta_\star, \boldsymbol{U})^{-1}\right)\epsilon_2 + \frac{2L_{\text{cov}}(\theta_\star, \gamma^2)\text{tr}(\mathcal{H}(\theta_\star))}{\underline{\lambda}^2}\epsilon_\circ + \frac{2d_\theta L_{\text{cov}}(\theta_\star, \gamma^2)L_{\text{hess}}}{\underline{\lambda}^2}\epsilon_\circ\epsilon_2 \\
&\leq \frac{d_\theta L_{\text{hess}}}{\underline{\lambda}}\epsilon_2 + \frac{2L_{\text{cov}}(\theta_\star, \gamma^2)\text{tr}(\mathcal{H}(\theta_\star))}{\underline{\lambda}^2}\epsilon_\circ + \frac{2d_\theta L_{\text{cov}}(\theta_\star, \gamma^2)L_{\text{hess}}}{\underline{\lambda}^2}\epsilon_\circ\epsilon_2 \\
&=: f(\epsilon)
\end{aligned}$$

Assume that $\text{tr}\left(\mathcal{H}(\widehat{\theta})\mathbf{\Lambda}(\widehat{\theta}, \widehat{\boldsymbol{U}})^{-1}\right) > \text{tr}\left(\mathcal{H}(\theta_\star)\mathbf{\Lambda}(\theta_\star, \boldsymbol{U}^\star)^{-1}\right)$, then:

$$\begin{aligned}
&\left|\text{tr}\left(\mathcal{H}(\theta_\star)\mathbf{\Lambda}(\theta_\star, \boldsymbol{U}^\star)^{-1}\right) - \text{tr}\left(\mathcal{H}(\theta_\star)\mathbf{\Lambda}(\theta_\star, \widehat{\boldsymbol{U}})^{-1}\right)\right| \\
&\leq \left|\text{tr}\left(\mathcal{H}(\theta_\star)\mathbf{\Lambda}(\theta_\star, \boldsymbol{U}^\star)^{-1}\right) - \text{tr}\left(\mathcal{H}(\widehat{\theta})\mathbf{\Lambda}(\widehat{\theta}, \widehat{\boldsymbol{U}})^{-1}\right)\right| + \left|\text{tr}\left(\mathcal{H}(\widehat{\theta})\mathbf{\Lambda}(\widehat{\theta}, \widehat{\boldsymbol{U}})^{-1}\right) - \text{tr}\left(\mathcal{H}(\theta_\star)\mathbf{\Lambda}(\theta_\star, \widehat{\boldsymbol{U}})^{-1}\right)\right| \\
&\leq \left|\text{tr}\left(\mathcal{H}(\theta_\star)\mathbf{\Lambda}(\theta_\star, \boldsymbol{U}^\star)^{-1}\right) - \text{tr}\left(\mathcal{H}(\widehat{\theta})\mathbf{\Lambda}(\widehat{\theta}, \boldsymbol{U}^\star)^{-1}\right)\right| + \left|\text{tr}\left(\mathcal{H}(\widehat{\theta})\mathbf{\Lambda}(\widehat{\theta}, \widehat{\boldsymbol{U}})^{-1}\right) - \text{tr}\left(\mathcal{H}(\theta_\star)\mathbf{\Lambda}(\theta_\star, \widehat{\boldsymbol{U}})^{-1}\right)\right| \\
&\leq 2f(\epsilon)
\end{aligned}$$

where the second inequality holds because $\widehat{\boldsymbol{U}}$ is the minimizer of $\text{tr}\left(\mathcal{H}(\widehat{\theta})\mathbf{\Lambda}(\widehat{\theta}, \boldsymbol{U})^{-1}\right)$. If instead $\text{tr}\left(\mathcal{H}(\widehat{\theta})\mathbf{\Lambda}(\widehat{\theta}, \widehat{\boldsymbol{U}})^{-1}\right) \leq \text{tr}\left(\mathcal{H}(\theta_\star)\mathbf{\Lambda}(\theta_\star, \boldsymbol{U}^\star)^{-1}\right)$, we can replace $\text{tr}\left(\mathcal{H}(\theta_\star)\mathbf{\Lambda}(\theta_\star, \boldsymbol{U}^\star)^{-1}\right)$ with $\text{tr}\left(\mathcal{H}(\theta_\star)\mathbf{\Lambda}(\theta_\star, \widehat{\boldsymbol{U}})^{-1}\right)$ in the above calculation to get the same result. The conclusion follows.

□

C.4 Proof of Corollary 5.3

The following result shows that the LQR problem satisfies all assumptions required by the LDDM setting, allowing us to state Corollary 5.3 as a direct corollary of Theorem 5.2.

Theorem C.2. *If $\theta_\star$ is stabilizable and $R_{\mathbf{u}}, R_{\mathbf{x}} \succeq I$, Assumption 4.1 is satisfied for $\mathcal{R}_{LQR}$ with:*

1. $\mu = 2$.
2. $r_{\text{quad}}(\theta_\star) = \min\left\{\frac{1}{150\Psi_{P_\star}^5}, \frac{1}{240\Psi_{B_\star}\Psi_{P_\star}^5}, \frac{\Psi_{B_\star}}{2}\right\}$.
3. $L_{\mathcal{R}1} = c_1 d_x \Psi_{B_\star}\sqrt{\Psi_{P_\star}} + \frac{c_2 d_x \Psi_{R_{\mathbf{u}}}(1+1/\Psi_{B_\star})}{\sqrt{\Psi_{P_\star}}}$, $\quad L_{\mathcal{R}2} = c_3 d_x\Big(\Psi_{B_\star}^2\Psi_{P_\star}^2 + \Psi_{R_{\mathbf{u}}}(1+\Psi_{B_\star})\Psi_{P_\star}\Big)$, $L_{\mathcal{R}3} = c_4 d_x\Big(\Psi_{B_\star}^3\Psi_{P_\star}^{7/2} + \Psi_{R_{\mathbf{u}}}\Psi_{B_\star}(1+\Psi_{B_\star})\Psi_{P_\star}^{5/2}\Big)$.
4. $L_{\mathfrak{a}1} = 8\Psi_{P_\star}^{7/2}$, $\quad L_{\mathfrak{a}2} = \text{poly}(\Psi_{P_\star})$, $\quad L_{\mathfrak{a}3} = \text{poly}(\Psi_{P_\star}, \Psi_{B_\star})$.
5. $L_{\text{hess}} = d_x\text{poly}(\Psi_{P_\star}, \Psi_{R_{\mathbf{u}}}, \Psi_{B_\star}, 1/\Psi_{B_\star}) + d_x^2\Big(\Psi_{B_\star}^2\Psi_{P_\star}^4 + (1+\Psi_{B_\star})\Psi_{R_{\mathbf{u}}}\Psi_{P_\star}^3 + \frac{\Psi_{R_{\mathbf{u}}}\Psi_{P_\star}}{\Psi_{B_\star}}\Big)$.

for universal constants c_1, c_2, c_3, c_4.

Proof. The proof of Theorem C.2 is quite technical, and somewhat tangential to the main focus of this thesis. As such, we omit it here and refer the reader to Section J and Theorem J.1 of Wagenmaker et al. [2021] for a complete proof. □

The proofs of Proposition 5.1, Proposition 5.2, and Proposition 5.3 are similarly rather technical, and we omit them from this thesis as well. For the proofs of Proposition 5.1 and Proposition 5.2, please refer to Section K of Wagenmaker et al. [2021], and for Proposition 5.3 see Section B.7 of Wagenmaker et al. [2021].

Appendix D

Deferred Proofs from Chapter 6

D.1 Deferred Proofs from Section 6.5

Lemma D.1.1. *Under Assumptions 6.1 and 6.3, the system (6.0.1) satisfies Assumptions 6.11 and D.1 with*

$$\boldsymbol{\psi}(\boldsymbol{\tau}) \leftarrow I_{d_x} \otimes \sum_{h=1}^{H} \boldsymbol{\phi}(x_h^{\boldsymbol{\tau}}, u_h^{\boldsymbol{\tau}}) \boldsymbol{\phi}(x_h^{\boldsymbol{\tau}}, u_h^{\boldsymbol{\tau}})^\top, \quad D = d_x H B_{\boldsymbol{\phi}}^2,$$

$d_{\boldsymbol{\psi}} \leftarrow d_{\boldsymbol{\phi}} d_x$, and where $x_h^{\boldsymbol{\tau}}$ (resp. $u_h^{\boldsymbol{\tau}}$) denotes the state (resp. input) at step h of trajectory $\boldsymbol{\tau}$. Furthermore, it satisfies Assumption 6.10 with $\mathbb{A}_{\mathcal{R}}$ instantiated with the LC^3 algorithm of Kakade et al. [2020] and

$$C_{\mathcal{R}} = C \cdot H \sqrt{d_{\boldsymbol{\phi}}(d_{\boldsymbol{\phi}} + d_x + B_A)} \cdot \log(1 + B_{\boldsymbol{\phi}} H / \sigma_w), \quad p_{\mathcal{R}} = 3/2, \quad \alpha = 1/2$$

for a universal constant C.

Proof. That Assumption D.1 is satisfied is immediate under Assumption 6.3. It is clear that $\boldsymbol{\psi}(\boldsymbol{\tau}) \in \mathcal{S}_+^{d_{\boldsymbol{\phi}} d_x}$. To obtain a bound on D, we only need to bound the trace of $\boldsymbol{\psi}(\boldsymbol{\tau})$:

$$\begin{aligned}
\operatorname{tr}(\boldsymbol{\psi}(\boldsymbol{\tau})) &= \sum_{h=1}^{H} \operatorname{tr}(I_{d_x}) \cdot \operatorname{tr}(\boldsymbol{\phi}(x_h^{\boldsymbol{\tau}}, u_h^{\boldsymbol{\tau}})) \boldsymbol{\phi}(x_h^{\boldsymbol{\tau}}, u_h^{\boldsymbol{\tau}}))^\top) \\
&= d_x \sum_{h=1}^{H} \|\boldsymbol{\phi}(x_h^{\boldsymbol{\tau}}, u_h^{\boldsymbol{\tau}}))\|_2^2 \\
&\leq d_x H B_{\boldsymbol{\phi}}^2
\end{aligned}$$

where the inequality holds under Assumption 6.1.

To show that Assumption 6.10 is satisfied in this setting, we have by Lemma D.1.3 that with probability at least $1 - \delta$, LC^3 has regret bounded as (using that $c_{\max} \leq 1$ in the setting

of Assumption 6.10):

$$\mathcal{R}_T \le C \cdot H\sqrt{d_{\boldsymbol{\phi}} \cdot (d_{\boldsymbol{\phi}} + d_x + B_A + \log\frac{1}{\delta}) \cdot T} \cdot \log\left(1 + B_{\boldsymbol{\phi}} HT/\sigma_w\right)$$

for C a universal constant. We can therefore take $\alpha = 1/2$, $p_{\mathcal{R}} = 3/2$, and

$$C_{\mathcal{R}} = C' \cdot H\sqrt{d_{\boldsymbol{\phi}}(d_{\boldsymbol{\phi}} + d_x + B_A)} \cdot \log(1 + B_{\boldsymbol{\phi}} H/\sigma_w).$$

□

Lemma D.1.2. *Let $\bar{T}_\ell$ denote the total number of episodes collected by Algorithm 6.1 at round ℓ. For*

$$T_\ell \ge \mathrm{poly}\left(d_{\boldsymbol{\phi}}, d_x, \frac{1}{\bar{\lambda}_{\min}}, B_A, B_{\boldsymbol{\phi}}, \log\frac{1}{\sigma_w}, H, \log\frac{T_\ell}{\delta}\right), \tag{D.1.1}$$

on the success event of Lemma 6.5.1, we have $2T_\ell \ge \bar{T}_\ell$ and

$$2T_\ell \ge \sum_{i=1}^{\ell-1} \bar{T}_i.$$

Proof. Recall that $T_\ell = 2^\ell$ and $N_\ell = \lceil T_\ell / T_\ell^{\mathrm{oed}} \rceil$ for $T_\ell^{\mathrm{oed}} = |\Pi_\ell|$. By Lemma 6.5.1, $\bar{T}_\ell$ can be bounded as

$$\bar{T}_\ell \le N_\ell T_\ell^{\mathrm{oed}} + (16 + 2\log T_\ell^{\mathrm{oed}}) T_\ell^{\mathrm{oed}}$$

and T_ℓ^{oed} can be bounded as

$$T_\ell^{\mathrm{oed}} \le \mathrm{poly}\left(d_{\boldsymbol{\phi}}, d_x, \frac{1}{\bar{\lambda}_{\min}}, B_A, B_{\boldsymbol{\phi}}, \log\frac{1}{\sigma_w}, H, \log\frac{T_\ell}{\delta}\right). \tag{D.1.2}$$

Note that we can bound

$$\begin{aligned}
\bar{T}_i &\le N_i T_i^{\mathrm{oed}} + (16 + 2\log T_i^{\mathrm{oed}}) T_i^{\mathrm{oed}} \\
&\le T_i + (17 + 2\log T_i^{\mathrm{oed}}) T_i^{\mathrm{oed}} \\
&\le T_i + \mathrm{poly}\left(d_{\boldsymbol{\phi}}, d_x, \frac{1}{\bar{\lambda}_{\min}}, B_A, B_{\boldsymbol{\phi}}, \log\frac{1}{\sigma_w}, H, \log\frac{T_\ell}{\delta}\right).
\end{aligned}$$

From this it is immediately obvious that $2T_\ell \ge \bar{T}_\ell$ as long as (D.1.1) is satisfied.

To show the second conclusion note that, by our choice of $T_i = 2^i$, we have that

$T_\ell \geq \sum_{i=1}^{\ell-1} T_i$, so it therefore remains to show that

$$T_\ell \geq \sum_{i=1}^{\ell-1} \text{poly}\left(d_{\boldsymbol{\phi}}, d_x, \frac{1}{\bar{\lambda}_{\min}}, B_A, B_{\boldsymbol{\phi}}, \log\frac{1}{\sigma_w}, H, \log\frac{T_i}{\delta}\right).$$

However, we can bound

$$\sum_{i=1}^{\ell-1} \text{poly}\left(d_{\boldsymbol{\phi}}, d_x, \tfrac{1}{\bar{\lambda}_{\min}}, B_A, B_{\boldsymbol{\phi}}, \log\tfrac{1}{\sigma_w}, H, \log\tfrac{T_i}{\delta}\right) \leq \text{poly}\left(d_{\boldsymbol{\phi}}, d_x, \tfrac{1}{\bar{\lambda}_{\min}}, B_A, B_{\boldsymbol{\phi}}, \log\tfrac{1}{\sigma_w}, H, \log\tfrac{T_\ell}{\delta}\right) \cdot \log T_\ell,$$

so a sufficient condition is

$$T_\ell \geq \text{poly}\left(d_{\boldsymbol{\phi}}, d_x, \frac{1}{\bar{\lambda}_{\min}}, B_A, B_{\boldsymbol{\phi}}, \log\frac{1}{\sigma_w}, H, \log\frac{T_\ell}{\delta}\right) \cdot \log T_\ell$$

which we see is met when (D.1.1) holds. □

Lemma D.1.3. *Consider running the LC³ algorithm of Kakade et al. [2020] modified so that the parameter β^t in the definition of $\textsc{Ball}^t$ is given by*

$$\beta^t := \sqrt{\lambda} B_A + \sqrt{8 d_x \log 5 + 8 \log(T \det(\boldsymbol{\Sigma}_t) \det(\boldsymbol{\Sigma}_0)^{-1}/\delta)}$$

for $\boldsymbol{\Sigma}_t$ the covariates obtained running LC³ for t episodes. Then with regret defined as

$$\mathcal{R}_T(\Pi) := \sum_{t=1}^{T} \mathcal{J}_{A_\star}(\pi^t) - T \cdot \min_{\pi \in \Pi} \mathcal{J}_{A_\star}(\pi),$$

and assuming $\text{cost}(\boldsymbol{\tau}) \leq c_{\max}$ for all trajectories $\boldsymbol{\tau}$, with any policy class Π, LC³ has regret bounded as, with probability at least $1-\delta$:

$$\mathcal{R}_T(\Pi) \leq C \cdot c_{\max} H \sqrt{d_{\boldsymbol{\phi}} \cdot (d_{\boldsymbol{\phi}} + d_x + B_A + \log\frac{1}{\delta}) \cdot T} \cdot \log\left(1 + B_{\boldsymbol{\phi}} H T/\sigma_w\right)$$

for a universal constant C.

Proof. The proof of this result is a small modification from the proof of the main LC³ regret bound given in Kakade et al. [2020]. We refer the reader to Theorem 6 of Wagenmaker et al. [2024] for a full proof. □

Proof of Lemma 6.5.1. By Lemma D.1.1, the assumptions of Lemma D.3.4 and Lemma D.3.5 are met, so we can therefore apply these results in our setting. By Lemma D.3.4, the event $\mathcal{E}_{\text{exp}}$ occurs with probability at least $1-\delta$. Throughout the remainder of the proof we union bound over the success event of Lemma D.3.5 and $\mathcal{E}_{\text{exp}}$, which together occur with probability at least $1-4\delta$.

Let

$$\widetilde{\mathbf{\Lambda}} := \sum_{t=1}^{NT_{\mathrm{out}}} \boldsymbol{\psi}(\boldsymbol{\tau}^t) = I_{d_x} \otimes \sum_{t=1}^{NT_{\mathrm{out}}} \sum_{h=1}^{H} \boldsymbol{\phi}(x_h^t, u_h^t)\boldsymbol{\phi}(x_h^t, u_h^t)^\top$$

denote the features returned by rerunning every policy in Π_{out} N times. By Lemma D.3.5, we then have that:

$$\left\|\widetilde{\mathbf{\Lambda}} - N \cdot \textstyle\sum_{\pi\in\Pi_{\mathrm{out}}} \check{\mathbf{\Lambda}}_\pi\right\|_{\mathrm{op}} \le \underbrace{\frac{\sqrt{T_{\mathrm{out}}} \cdot \sqrt{8d_{\boldsymbol{\phi}}d_x \log(1+8\sqrt{NT_{\mathrm{out}}}) + 8\log 1/\delta}}{\sqrt{N}\cdot 6272 d_{\boldsymbol{\phi}} d_x \log\frac{68\widetilde{N}}{\delta}}}_{=:\beta} \cdot \lambda_{\min}\left(N \cdot \textstyle\sum_{\pi\in\Pi_{\mathrm{out}}} \check{\mathbf{\Lambda}}_\pi\right).$$

Applying Theorem 4.2 with $\mathcal{E}$ the event that the above conclusion holds and $\mathbf{\Gamma} := N \cdot \sum_{\pi\in\Pi_{\mathrm{out}}} \check{\mathbf{\Lambda}}_\pi$, we obtain that, with probability at least $1-\delta$ (using the mapping to the martingale regression setting described in Section 6.1):

$$\|\mathrm{vec}(\widehat{A} - A_\star)\|_{\mathcal{H}}^2 \le 5(1+\zeta)\cdot\sigma_w^2 \log\frac{6d_x d_{\boldsymbol{\phi}}}{\delta} \cdot \mathrm{tr}(\mathcal{H}\mathbf{\Gamma}^{-1})$$

for $\zeta = 26\beta^2\lambda_{\max}(\mathbf{\Gamma})\mathrm{tr}(\mathbf{\Gamma}^{-1})$ and β as defined above. Since $\|\boldsymbol{\phi}(x,u)\|_2 \le B_{\boldsymbol{\phi}}$ under Assumption 6.1, we have $\|\check{\mathbf{\Lambda}}_\pi\|_2 \le B_{\boldsymbol{\phi}}^2$, so we can upper bound $\lambda_{\max}(\mathbf{\Gamma}) \le NT_{\mathrm{out}}B_{\boldsymbol{\phi}}^2$. By Lemma D.3.5, we can also bound (using that $D = d_x H B_{\boldsymbol{\phi}}^2$ by Lemma D.1.1):

$$\mathrm{tr}(\mathbf{\Gamma}^{-1}) \le \frac{1}{N\cdot 6272 d_x H B_{\boldsymbol{\phi}}^2 \log\frac{68\widetilde{N}}{\delta}}.$$

Combining these and using that

$$T_{\mathrm{out}} \le \mathrm{poly}\left(d_{\boldsymbol{\phi}}, d_x, \frac{1}{\bar{\lambda}_{\min}}, B_A, B_{\boldsymbol{\phi}}, \log\frac{1}{\sigma_w}, H, \log\frac{\widetilde{N}}{\delta}\right) \tag{D.1.3}$$

as shown in Lemma D.3.6 (and using our bounds on $C_{\mathcal{R}}$ and $p_{\mathcal{R}}$ in Lemma D.1.1), we can therefore bound

$$\zeta \le \mathrm{poly}\left(d_{\boldsymbol{\phi}}, d_x, \frac{1}{\bar{\lambda}_{\min}}, B_A, B_{\boldsymbol{\phi}}, \log\frac{1}{\sigma_w}, H, \log\frac{\widetilde{N}}{\delta}\right)\cdot\frac{1}{N}.$$

Using that $\mathrm{tr}(\mathcal{H}\mathbf{\Gamma}^{-1}) \le \|\mathcal{H}\|_{\mathrm{op}}\cdot\mathrm{tr}(\mathbf{\Gamma}^{-1})$, and the bound on $\mathrm{tr}(\mathbf{\Gamma}^{-1})$ given above, it follows that

$$\|\mathrm{vec}(\widehat{A} - A_\star)\|_{\mathcal{H}}^2 \le 5\sigma_w^2 \log\frac{6d_x d_{\boldsymbol{\phi}}}{\delta}\cdot\mathrm{tr}(\mathcal{H}\mathbf{\Gamma}^{-1}) + \mathrm{poly}\left(d_{\boldsymbol{\phi}}, d_x, \tfrac{1}{\bar{\lambda}_{\min}}, B_A, B_{\boldsymbol{\phi}}, \log\tfrac{1}{\sigma_w}, H, \|\mathcal{H}\|_{\mathrm{op}}, \log\tfrac{\widetilde{N}}{\delta}\right)\cdot\frac{1}{N^2}.$$

Finally, by Lemma D.3.5, we can bound

$$\operatorname{tr}(\mathcal{H}\boldsymbol{\Gamma}^{-1}) \le \frac{12}{NT_{\text{out}}} \cdot \min_{\boldsymbol{\Lambda}\in\boldsymbol{\Omega}} \operatorname{tr}(\mathcal{H}\check{\boldsymbol{\Lambda}}^{-1}).$$

By Lemma D.3.6, we can bound the total number of episodes collected by Algorithm D.2 by $(16 + 2\log(T_{\text{out}})) \cdot T_{\text{out}}$.

Bound on Frobenius Norm Error. By Lemma D.3.5, we can lower bound

$$\lambda_{\min}(\widetilde{\boldsymbol{\Lambda}}) \ge N \cdot 6272 d_x d_{\boldsymbol{\phi}} H B_{\boldsymbol{\phi}}^2 \log \frac{68\widetilde{N}}{\delta}.$$

Furthermore, since $\|\boldsymbol{\phi}(x,u)\|_2 \le B_{\boldsymbol{\phi}}$, we always have $\|\widetilde{\boldsymbol{\Lambda}}\|_{\text{op}} \le NT_{\text{out}}B_{\boldsymbol{\phi}}^2$, which implies $\widetilde{\boldsymbol{\Lambda}} \preceq NT_{\text{out}}B_{\boldsymbol{\phi}}^2 \cdot I$. By Lemma B.2.1, we then have that with probability at least $1-\delta$ (again using the mapping to the martingale regression setting described in Section 6.1):

$$\begin{aligned}
\|\widehat{A} - A_\star\|_{\text{F}} &\le C \cdot \sigma_w \sqrt{\frac{\log 1/\delta + d_x d_{\boldsymbol{\phi}} + \log\det(\frac{T_{\text{out}}}{6272 d_x d_{\boldsymbol{\phi}} H \log\frac{68\widetilde{N}}{\delta}} \cdot I + I)}{N \cdot 6272 d_x d_{\boldsymbol{\phi}} H B_{\boldsymbol{\phi}}^2 \log\frac{68\widetilde{N}}{\delta}}} \\
&\le \operatorname{poly}\left(d_{\boldsymbol{\phi}}, d_x, \frac{1}{\bar{\lambda}_{\min}}, B_A, B_{\boldsymbol{\phi}}, \sigma_w, H, \log\frac{\widetilde{N}}{\delta}\right) \cdot \frac{1}{\sqrt{N}}.
\end{aligned}$$

□

Proof of Lemma 6.5.2. We have

$$\begin{aligned}
\min_{\check{\boldsymbol{\Lambda}}\in\boldsymbol{\Omega}} \operatorname{tr}(\mathcal{H}\check{\boldsymbol{\Lambda}}^{-1}) &= \min_{\check{\boldsymbol{\Lambda}}\in\boldsymbol{\Omega}} \operatorname{tr}(\mathcal{H}'\check{\boldsymbol{\Lambda}}^{-1}) + \operatorname{tr}((\mathcal{H}-\mathcal{H}')\check{\boldsymbol{\Lambda}}^{-1}) \\
&\le \min_{\check{\boldsymbol{\Lambda}}\in\boldsymbol{\Omega}} \operatorname{tr}(\mathcal{H}'\check{\boldsymbol{\Lambda}}^{-1}) + \|\mathcal{H}-\mathcal{H}'\|_{\text{op}} \cdot \operatorname{tr}(\check{\boldsymbol{\Lambda}}^{-1})
\end{aligned}$$

Under Assumption 6.3, we know that there exists some $\check{\boldsymbol{\Lambda}}' \in \boldsymbol{\Omega}$ such that $\lambda_{\min}(\check{\boldsymbol{\Lambda}}') \ge \bar{\lambda}_{\min}$. We can then bound

$$\begin{aligned}
&\min_{\check{\boldsymbol{\Lambda}}\in\boldsymbol{\Omega}} \operatorname{tr}(\mathcal{H}'\check{\boldsymbol{\Lambda}}^{-1}) + \|\mathcal{H}-\mathcal{H}'\|_{\text{op}} \cdot \operatorname{tr}(\check{\boldsymbol{\Lambda}}^{-1}) \\
&\le \min_{\check{\boldsymbol{\Lambda}}\in\boldsymbol{\Omega}} \operatorname{tr}(\mathcal{H}'(\frac{1}{2}\check{\boldsymbol{\Lambda}} + \frac{1}{2}\check{\boldsymbol{\Lambda}}')^{-1}) + \|\mathcal{H}-\mathcal{H}'\|_{\text{op}} \cdot \operatorname{tr}((\frac{1}{2}\check{\boldsymbol{\Lambda}} + \frac{1}{2}\check{\boldsymbol{\Lambda}}')^{-1}) \\
&\le \min_{\check{\boldsymbol{\Lambda}}\in\boldsymbol{\Omega}} 2\operatorname{tr}(\mathcal{H}'\check{\boldsymbol{\Lambda}}^{-1}) + 2\|\mathcal{H}-\mathcal{H}'\|_{\text{op}} \cdot \frac{d_x d_{\boldsymbol{\phi}}}{\bar{\lambda}_{\min}}
\end{aligned}$$

which proves the result. □

D.2 Proof of Theorem 6.2

Proof of Theorem 6.2. This proof follows immediately from Lemma D.2.1, by lower bounding the right-hand side of (D.2.1) by the min over all policies in Π. □

Lemma D.2.1. *Under Assumptions 6.1, 6.2 and 6.4 to 6.8 and if $\mathfrak{a}_{\mathrm{opt}}$ is defined as in (6.1.1), as long as*

$$T \geq \mathrm{poly}(\|A_\star\|_{\mathrm{op}}, L_{\boldsymbol{\phi}}, L_{\mathfrak{a}}, L_{\pi_\star}, L_{\mathrm{cost}}, \sigma_w, \sigma_w^{-1}, B_{\boldsymbol{\phi}}, H, d_x, d_{\boldsymbol{\phi}}, \underline{\lambda}^{-1}, \mu^{-1}, r_{\mathrm{cost}}(A_\star)^{-1}, r_{\mathfrak{a}}(A_\star)^{-1}, r_\mu(A_\star)^{-1}),$$

for any $\omega_{\mathrm{exp}} \in \triangle_\Pi$, we have

$$\min_{\widehat{\mathfrak{a}}} \max_{A \in \mathcal{B}_T} \mathbb{E}_{\mathfrak{D}_T \sim A, \omega_{\mathrm{exp}}}[\mathcal{J}_A(\widehat{\mathfrak{a}}(\mathfrak{D}_T)) - \mathcal{J}_A(\mathfrak{a}_{\mathrm{opt}}(A))] \geq \frac{\sigma_w^2}{3T} \cdot \mathrm{tr}(\mathcal{H}(A_\star)\mathbb{E}_{\pi \sim \omega_{\mathrm{exp}}}[\check{\boldsymbol{\Lambda}}_\pi]^{-1}) - \frac{C_{\mathrm{lb}}}{T^{5/4}} \tag{D.2.1}$$

for $\mathcal{B}_T := \{A \ : \ \|A - A_\star\|_{\mathrm{F}}^2 \leq 5 d_x d_{\boldsymbol{\phi}} / (\underline{\lambda} T H)^{5/6}\}$, where $\mathbb{E}_{\mathfrak{D}_T \sim A, \omega_{\mathrm{exp}}}[\cdot] = \mathbb{E}_{\pi \sim \omega_{\mathrm{exp}}}[\mathbb{E}_{\mathfrak{D}_T \sim A, \pi}[\cdot]]$ denotes the expectation over trajectories generated by running policies π drawn according to ω_{exp} on system A for T episodes, and

$$C_{\mathrm{lb}} := \mathrm{poly}(\|A_\star\|_{\mathrm{op}}, L_{\boldsymbol{\phi}}, L_{\mathfrak{a}}, L_{\pi_\star}, L_{\mathrm{cost}}, \sigma_w, \sigma_w^{-1}, B_{\boldsymbol{\phi}}, H, d_x, d_{\boldsymbol{\phi}}, \underline{\lambda}^{-1}, \mu^{-1}, r_{\mathrm{cost}}(A_\star)^{-1}, r_{\mathfrak{a}}(A_\star)^{-1}, r_\mu(Ast)^{-1}).$$

Proof. This result is a direct consequence of Theorem 4.3—to obtain the result we must only verify that the assumptions of this result are met. We verify each assumption below.

Verifying Assumption 4.1. Part 1 of Assumption 4.1 is met by Assumption 6.7 within diameter $r_\mu(A_\star)$. Furthermore, under Assumptions 6.1, 6.2 and 6.4 to 6.6 and by Lemma D.2.7, the additional parts of Assumption 4.1 are also met with diameter $\min\{r_{\mathrm{cost}}(A_\star), r_{\mathfrak{a}}(A_\star)\}$ and smoothness constant $\mathrm{poly}(\|A_\star\|_{\mathrm{op}}, L_{\boldsymbol{\phi}}, L_{\mathfrak{a}}, L_{\pi_\star}, L_{\mathrm{cost}}, \sigma_w^{-1}, H, d_x)$.

Verifying Assumptions 4.2 and 4.3. Assumption 4.2 is immediately met by Assumption 6.8. Furthermore, Assumption 4.3 is met by Lemma D.2.2 with $c_{\mathrm{cov}} = 1, L_{\mathrm{cov}}(\theta_\star, \gamma^2) = \frac{H^2 B_{\boldsymbol{\phi}}^3}{\sigma_w^2} \cdot \mathrm{poly}(d_x)$, and $C_{\mathrm{cov}} = 0$.

Given that these assumptions are met, the result follows noting that, if we run for T episodes, then the effective horizon is $d_x T H$ (using the mapping from the setting of (6.0.1) to the martingale regression setting described in Section 6.1). Note that the final bound scales with $\frac{1}{T}$ instead of $\frac{1}{d_x T H}$ as we are able to bring the $d_x H$ factor into the $\check{\boldsymbol{\Lambda}}_{\pi_{\mathrm{exp}}}$ term, since $\boldsymbol{\Lambda}_{\pi_{\mathrm{exp}}}$ is not normalized by $d_x H$. □

Lemma D.2.2. *Under Assumption 6.1, for any policy distribution $\omega \in \triangle_{\Pi_{\mathrm{exp}}}$ and A, A', we have*

$$\mathbb{E}_{\pi \sim \omega}[\boldsymbol{\Lambda}_{A,\pi}] \preceq \mathbb{E}_{\pi \sim \omega}[\boldsymbol{\Lambda}_{A',\pi}] + \frac{H^2 B_{\boldsymbol{\phi}}^3}{\sigma_w^2} \cdot \mathrm{poly}(d_x) \cdot \|A - A'\|_{\mathrm{F}} \cdot I.$$

Proof. We will prove that the desired bound follows for a particular $\pi \in \Pi_{\text{exp}}$, which immediately implies that it holds for $\omega \in \triangle_{\Pi_{\text{exp}}}$. By definition we have

$$\mathbf{\Lambda}_{A,\pi} = \int \left(\sum_{h=1}^{H} \boldsymbol{\phi}(x_h^{\boldsymbol{\tau}}, u_h^{\boldsymbol{\tau}}) \boldsymbol{\phi}(x_h^{\boldsymbol{\tau}}, u_h^{\boldsymbol{\tau}})^\top \right) \cdot f_{A,\pi}(\boldsymbol{\tau}) \mathrm{d}\boldsymbol{\tau}$$

and

$$f_{A,\pi}(\boldsymbol{\tau}) = \prod_{h=1}^{H} f_A(x_{h+1}^{\boldsymbol{\tau}} \mid x_h^{\boldsymbol{\tau}}, u_h^{\boldsymbol{\tau}}) \pi_h(u_h^{\boldsymbol{\tau}} \mid x_h^{\boldsymbol{\tau}}).$$

Fix some $\boldsymbol{v} \in \mathcal{S}^{d_{\boldsymbol{\phi}}-1}$, and note that, given the expression above, we have

$$\boldsymbol{v}^\top \mathbf{\Lambda}_{A,\pi} \boldsymbol{v} = \int \sum_{h=1}^{H} (\boldsymbol{v}^\top \boldsymbol{\phi}(x_h^{\boldsymbol{\tau}}, u_h^{\boldsymbol{\tau}}))^2 \cdot f_{A,\pi}(\boldsymbol{\tau}) \mathrm{d}\boldsymbol{\tau}.$$

It follows that

$$\begin{aligned} \nabla_A \boldsymbol{v}^\top \mathbf{\Lambda}_{A,\pi} \boldsymbol{v} &= \int \sum_{h=1}^{H} (\boldsymbol{v}^\top \boldsymbol{\phi}(x_h^{\boldsymbol{\tau}}, u_h^{\boldsymbol{\tau}}))^2 \cdot \nabla_A f_{A,\pi}(\boldsymbol{\tau}) \mathrm{d}\boldsymbol{\tau} \\ &= \int \sum_{h=1}^{H} (\boldsymbol{v}^\top \boldsymbol{\phi}(x_h^{\boldsymbol{\tau}}, u_h^{\boldsymbol{\tau}}))^2 \cdot f_{A,\pi}(\boldsymbol{\tau}) \nabla_A \log f_{A,\pi}(\boldsymbol{\tau}) \mathrm{d}\boldsymbol{\tau}. \end{aligned}$$

We can show (see e.g. the proof of Lemma D.1 in Wagenmaker et al. [2024]), we have, for any Δ,

$$\nabla_A \log f_{A,\pi}(\boldsymbol{\tau})[\Delta] = \sum_{h=1}^{H} \frac{1}{\sigma_w^2} (x_{h+1}^{\boldsymbol{\tau}} - A\boldsymbol{\phi}(x_h^{\boldsymbol{\tau}}, u_h^{\boldsymbol{\tau}})) \cdot \Delta \boldsymbol{\phi}(x_h^{\boldsymbol{\tau}}, u_h^{\boldsymbol{\tau}}),$$

so we can bound

$$\|\nabla_A \log f_{A,\pi}(\boldsymbol{\tau})\|_{\text{op}} \le \frac{B_{\boldsymbol{\phi}}}{\sigma_w^2} \sum_{h=1}^{H} \|x_{h+1}^{\boldsymbol{\tau}} - A\boldsymbol{\phi}(x_h^{\boldsymbol{\tau}}, u_h^{\boldsymbol{\tau}})\|_2.$$

Furthermore, we can also bound $(\boldsymbol{v}^\top \boldsymbol{\phi}(x_h^{\boldsymbol{\tau}}, u_h^{\boldsymbol{\tau}}))^2 \le B_{\boldsymbol{\phi}}^2$. We therefore have

$$\|\nabla_A \boldsymbol{v}^\top \mathbf{\Lambda}_{A,\pi} \boldsymbol{v}\|_{\text{op}} \le H B_{\boldsymbol{\phi}}^3 \int \sum_{h=1}^{H} \|x_{h+1}^{\boldsymbol{\tau}} - A\boldsymbol{\phi}(x_h^{\boldsymbol{\tau}}, u_h^{\boldsymbol{\tau}})\|_2 \cdot f_{A,\pi}(\boldsymbol{\tau}) \mathrm{d}\boldsymbol{\tau}$$

$$\leq \frac{H^2 B_{\phi}^3}{\sigma_w^2} \cdot \mathrm{poly}(d_x)$$

where the last inequality follows from Lemma D.2.3. It follows from the Mean Value Theorem that

$$|\boldsymbol{v}^\top \boldsymbol{\Lambda}_{A,\pi} \boldsymbol{v} - \boldsymbol{v}^\top \boldsymbol{\Lambda}_{A',\pi} \boldsymbol{v}| \leq \frac{H^2 B_{\phi}^3}{\sigma_w^2} \cdot \mathrm{poly}(d_x) \cdot \|A - A'\|_{\mathrm{F}}.$$

As this holds for all $\boldsymbol{v} \in \mathcal{S}^{d-1}$, it follows that

$$\|\boldsymbol{\Lambda}_{A,\pi} - \boldsymbol{\Lambda}_{A',\pi}\|_{\mathrm{op}} \leq \frac{H^2 B_{\phi}^3}{\sigma_w^2} \cdot \mathrm{poly}(d_x) \cdot \|A - A'\|_{\mathrm{F}}.$$

□

Lemma D.2.3. *Let $w_i \sim \mathcal{N}(0, I_{d_x})$ for all $i \in \{1, 2, \ldots, n\}$. Then,*

$$\mathbb{E}\left[\left(\sum_{i=1}^{n} \|w_i\|_2\right)^c\right] \leq n^2 \cdot \mathrm{poly}(d_x).$$

for c an absolute constant.

Proof. We first bound

$$(\sum_{i=1}^{n} \|w_i\|_2)^c \leq n \cdot \max_i \|w_i\|_2^c \leq n \cdot \sum_i \|w_i\|_2^c.$$

The result then follows since we can bound the $\mathbb{E}[\|w_i\|_2^c] \leq \mathrm{poly}(d)$ for $w_i \sim \mathcal{N}(0, I)$ and c an absolute constant. □

D.2.1 Smooth Nonlinear Systems

The following result shows that under Assumption 6.6, $\mathfrak{a}_{\mathrm{opt}}$ is differentiable.

Proposition D.1. *Under, Assumptions 6.1, 6.2 and 6.4 to 6.6, there exists some $r_{\mathfrak{a}}(A_\star) > 0$ and $L_{\pi_\star} < \infty$ such that, for all $A \in \mathcal{B}_{\mathrm{F}}(A_\star; r_{\mathfrak{a}}(A_\star))$, $\mathfrak{a}_{\mathrm{opt}}(A)$ satisfies:*

- $\nabla_{\mathfrak{a}} \mathcal{J}_A(\pi^{\mathfrak{a}})|_{\mathfrak{a}=\mathfrak{a}_{\mathrm{opt}}(A)} = 0$,
- *$\mathfrak{a}_{\mathrm{opt}}(A)$ is three-times differentiable in A, and we can bound $\|\nabla_A^{(i)} \mathfrak{a}_{\mathrm{opt}}(A)\|_{\mathrm{op}} \leq L_{\pi_\star}$ for some $L_{\pi_\star} > 0$ and $i \in \{1, 2, 3\}$.*

We then have the following results, analogous to the results of Chapter 4.

Lemma D.2.4. *Under Assumptions 6.1, 6.2 and 6.4 to 6.6, for* $\widehat{A} \in \mathcal{B}_{\mathrm{F}}(A_\star; \min\{r_{\mathrm{cost}}(A_\star), r_{\mathfrak{a}}(A_\star)\})$, *we have*

$$\mathcal{J}_{A_\star}(\mathfrak{a}_{\mathrm{opt}}(\widehat{A})) - \mathcal{J}_{A_\star}(\mathfrak{a}_{\mathrm{opt}}(A_\star)) \le \mathrm{vec}(\widehat{A} - A_\star)^\top \mathcal{H}(A_\star)\mathrm{vec}(\widehat{A} - A_\star) + M[\widehat{A} - A_\star, \widehat{A} - A_\star, \widehat{A} - A_\star].$$

for some tensor M *such that*

$$\|M[\widehat{A} - A_\star, \widehat{A} - A_\star, \widehat{A} - A_\star]\|_{\mathrm{op}} \le \mathrm{poly}(L_{\pi_\star}, \|A_\star\|_{\mathrm{op}}, B_{\boldsymbol{\phi}}, L_{\boldsymbol{\phi}}, L_{\mathfrak{a}}, L_{\mathrm{cost}}, \sigma_w^{-1}, H, d_x) \cdot \|\widehat{A} - A_\star\|_{\mathrm{op}}^3.$$

Lemma D.2.5. *Under Assumptions 6.1, 6.2 and 6.4 to 6.6, and if* $\widehat{A} \in \mathcal{B}_{\mathrm{F}}(A_\star; \min\{r_{\mathrm{cost}}(A_\star), r_{\mathfrak{a}}(A_\star)\})$, *we can bound*

$$\|\mathcal{H}(A_\star) - \mathcal{H}(\widehat{A})\|_{\mathrm{op}} \le \mathrm{poly}(L_{\pi_\star}, \|A_\star\|_{\mathrm{op}}, B_{\boldsymbol{\phi}}, L_{\boldsymbol{\phi}}, L_{\mathfrak{a}}, L_{\mathrm{cost}}, \sigma_w^{-1}, H, d_x) \cdot \|\widehat{A} - A_\star\|_{\mathrm{op}}.$$

Lemma D.2.6. *Under Assumptions 6.1, 6.2 and 6.4 to 6.6, for all* $A \in \mathcal{B}_{\mathrm{F}}(A_\star; \min\{r_{\mathrm{cost}}(A_\star), r_{\mathfrak{a}}(A_\star)\})$, *we can bound*

$$\|\mathcal{H}(A)\|_{\mathrm{op}} \le \mathrm{poly}(\|A_\star\|_{\mathrm{op}}, B_{\boldsymbol{\phi}}, L_{\boldsymbol{\phi}}, L_{\mathfrak{a}}, L_{\mathrm{cost}}, L_{\pi_\star}, \sigma_w^{-1}, H, d_x)$$

Lemma D.2.7. *Under Assumptions 6.1, 6.2, 6.4 and 6.5, for any* A *satisfying* $A \in \mathcal{B}_{\mathrm{F}}(A_\star; r_{\mathfrak{a}}(A_\star))$, *the controller loss* $\mathcal{J}_A(\mathfrak{a})$ *is four-times differentiable in* $\mathfrak{a}$ *and* A. *Furthermore, we can bound*

$$\|\nabla_A^{(i)} \nabla_{\mathfrak{a}}^{(j)} \mathcal{J}_A(\mathfrak{a})\|_{\mathrm{op}} \le \mathrm{poly}(\|A\|_{\mathrm{op}}, B_{\boldsymbol{\phi}}, L_{\boldsymbol{\phi}}, L_{\mathfrak{a}}, L_{\mathrm{cost}}, \sigma_w^{-1}, H, d_x)$$

for $i, j \in \{0, 1, 2, 3, 4\}$ *satisfying* $1 \le i + j \le 3$.

The proofs of these results are somewhat technical and tangential to the main focus of this thesis. As such, we omit them and refer the reader to Appendix D of Wagenmaker et al. [2024].

D.3 Deferred Proofs from Section 6.6

D.3.1 Collecting Full-Rank Data

In this section, we consider the setting where $\boldsymbol{\psi}(\boldsymbol{\tau}) \in \mathcal{S}_+^{d_{\boldsymbol{\psi}}}$, and our goal is to collect $\{\boldsymbol{\tau}_t\}_{t=1}^T$ such that $\boldsymbol{\psi}(\frac{1}{T}\sum_{t=1}^T \boldsymbol{\psi}(\boldsymbol{\tau}_t)) > 0$. For this to be achievable, we need the following assumption, a generalization of Assumption 6.3.

Assumption D.1 (Full-Rank Data)**.** *Consider* $\boldsymbol{\psi}(\boldsymbol{\tau})$ *such that* $\boldsymbol{\psi}(\boldsymbol{\tau}) \in \mathbb{S}_+^{d_{\boldsymbol{\psi}}}$. *Then we have* $\sup_{\boldsymbol{\Gamma} \in \boldsymbol{\Omega}_{\boldsymbol{\psi}}} \lambda_{\min}(\boldsymbol{\Gamma}) \ge \bar{\lambda}_{\min}$ *for some* $\bar{\lambda}_{\min} > 0$.

Throughout this section we also assume that Assumption 6.10 is satisfied with $\alpha = 1/2$ (though all results generalize in a straightforward way for $\alpha \ne 1/2$). We have the following result.

Algorithm D.1 Minimum Eigenvalue Maximization (MinEig)

1: **input**: scale N, confidence δ, regret minimization algorithm $\mathbb{A}_{\mathcal{R}}$, exploration policies Π_{exp}
2: **for** $j = 1, 2, 3, \ldots$ **do**
3: $\quad N_j \leftarrow \lceil 2^{j/3} \rceil - 1, K_j \leftarrow \lceil 2^{2j/3} \rceil, T_j \leftarrow (N_j + 1)K_j, \lambda_j \leftarrow T_j^{-1/18}, \delta_j \leftarrow \frac{\delta}{4j^2}$
4: $\quad \mathbf{\Sigma}_j, \Pi_j \leftarrow \textsc{DynamicOED}(\Phi, N_j, K_j, \delta_j, \mathbb{A}_{\mathcal{R}}, \Pi_{\text{exp}})$ for $\Phi(\mathbf{\Gamma}) = \operatorname{tr}((\mathbf{\Gamma} + \lambda_j \cdot I)^{-1})$
5: $\quad$ **if** $\lambda_{\min}(\mathbf{\Sigma}_j) \geq 12544 D d_{\boldsymbol{\psi}} \log \frac{2N(2+32T_j)}{\delta}$ **then**
6: $\quad\quad$ **break**
7: **return** Π_j

Lemma D.3.1. *Under Assumptions 6.10, 6.11 and D.1, running Algorithm D.1 we have that with probability at least $1 - \delta$, it will terminate after collecting at most*

$$2|\Pi| \leq \text{poly}\left(d_{\boldsymbol{\psi}}, \frac{1}{\bar{\lambda}_{\min}}, D, C_{\mathcal{R}}, \log^{p_{\mathcal{R}}} \frac{N}{\delta}\right)$$

episodes, and return policy set Π such that

$$\lambda_{\min}\left(\textstyle\sum_{\pi \in \Pi} \mathbf{\Gamma}_\pi\right) \geq 6272 D d_{\boldsymbol{\psi}} \log \frac{68N}{\delta}.$$

Furthermore, if we rerun each policy in Π once, the resulting features $\mathbf{\Sigma}$ will satisfy, with probability at least $1 - \delta/N$:

$$\lambda_{\min}\left(\mathbf{\Sigma}\right) \geq 6272 D d_{\boldsymbol{\psi}} \log \frac{68N}{\delta}.$$

Proof. By Lemma D.3.2 and our choice of N_j and K_j in Algorithm D.1, we have that if $\lambda_j \leq \frac{\bar{\lambda}_{\min}}{4 d_{\boldsymbol{\psi}}}$ and

$$T_j^{1/3} \geq \widetilde{\Omega}\left(\left(D\lambda_j^{-2}(D^3\lambda_j^{-1} + C_{\mathcal{R}} \cdot \log^{p_{\mathcal{R}}} \frac{1}{\delta_j})\right) \cdot \frac{\bar{\lambda}_{\min}}{d_{\boldsymbol{\psi}}}\right), \tag{D.3.1}$$

then $\lambda_{\min}(\mathbf{\Sigma}_j) \geq \frac{\bar{\lambda}_{\min}}{4 d_{\boldsymbol{\psi}}} \cdot T_j$ with probability at least $1 - \delta_j$. It follows that, with probability at least $1 - \delta_j$, the if statement on Line 5 will be true once $\lambda_j \leq \frac{\bar{\lambda}_{\min}}{4 d_{\boldsymbol{\psi}}}$, (D.3.1) holds, and

$$\frac{\bar{\lambda}_{\min}}{4 d_{\boldsymbol{\psi}}} \cdot T_j \geq 12544 D d_{\boldsymbol{\psi}} \log \frac{2N(2+32T_j)}{\delta}. \tag{D.3.2}$$

By our choice of $\lambda_j = T_j^{-1/18}$, a sufficient condition to ensure $\lambda_j \leq \frac{\bar{\lambda}_{\min}}{4 d_{\boldsymbol{\psi}}}$, (D.3.1), and (D.3.2)

is

$$T_j \geq \widetilde{\Omega}\left(\max\left\{\left(\frac{d_{\boldsymbol{\psi}}}{\bar{\lambda}_{\min}}\right)^{18}, \left(\frac{D^4\bar{\lambda}_{\min}}{d_{\boldsymbol{\psi}}}\right)^{6}, \left(DC_{\mathcal{R}}\cdot\log^{p_{\mathcal{R}}}\frac{1}{\delta_j}\cdot\frac{\bar{\lambda}_{\min}}{d_{\boldsymbol{\psi}}}\right)^{9/2}, \frac{Dd_{\boldsymbol{\psi}}^2}{\bar{\lambda}_{\min}}\cdot\log\frac{NT_j}{\delta}\right\}\right).$$

Since $T_j = \lceil 2^{j/3}\rceil\lceil 2^{2j/3}\rceil \in [2^j, 4\cdot 2^j]$, it follows that the if statement on Line 5 will be met after running for at most

$$\widetilde{\mathcal{O}}\left(\max\left\{\left(\frac{d_{\boldsymbol{\psi}}}{\bar{\lambda}_{\min}}\right)^{18}, \left(\frac{D^4\bar{\lambda}_{\min}}{d_{\boldsymbol{\psi}}}\right)^{6}, \left(DC_{\mathcal{R}}\cdot\log^{p_{\mathcal{R}}}\frac{1}{\delta_j}\cdot\frac{\bar{\lambda}_{\min}}{d_{\boldsymbol{\psi}}}\right)^{9/2}, \frac{Dd_{\boldsymbol{\psi}}^2}{\bar{\lambda}_{\min}}\cdot\log\frac{N}{\delta}\right\}\right) \tag{D.3.3}$$

episodes.

By Lemma D.3.3, if $\lambda_{\min}(\boldsymbol{\Sigma}_j) \geq 12544Dd_{\boldsymbol{\psi}}\log\frac{2N(2+32T_j)}{\delta}$ and we rerun all policies in Π_j, then we will collect data $\widetilde{\boldsymbol{\Sigma}}$ such that $\lambda_{\min}(\widetilde{\boldsymbol{\Sigma}}) \geq \frac{1}{2}\lambda_{\min}(\boldsymbol{\Sigma}_j)$, with probability at least $1-\delta/2N$. As the if statement on Line 5 will only be true once this is met, it follows that, with probability at least $1-\delta/2N$, rerunning all policies in Π_j once, we will collect data $\boldsymbol{\Sigma}$ which satisfies

$$\lambda_{\min}(\boldsymbol{\Sigma}) \geq \frac{1}{2}\lambda_{\min}(\boldsymbol{\Sigma}_j) \geq 6272Dd_{\boldsymbol{\psi}}\log\frac{2N(2+32T_j)}{\delta} \geq 6272Dd_{\boldsymbol{\psi}}\log\frac{68N}{\delta}.$$

The lower bound on $\lambda_{\min}(\sum_{\pi\in\Pi}\boldsymbol{\Gamma}_\pi)$ follows analogously from Lemma D.3.3.

The result then follows noting that the failure probability of running DynamicOED is at most

$$\sum_{j=1}^{\infty}\frac{\delta}{4j^2} \leq \delta/2.$$

□

D.3.1.1 Supporting Lemmas

Lemma D.3.2. *Under Assumptions 6.10, 6.11 and D.1, consider running* DynamicOED *on the objective*

$$\Phi(\boldsymbol{\Gamma}) = \operatorname{tr}((\boldsymbol{\Gamma}+\lambda\cdot I)^{-1})$$

with $N = \lceil 2^{i/3}\rceil - 1$ *and* $K = \lceil 2^{2i/3}\rceil$, *for some* $\lambda > 0$ *and* i. *Let* $T := (N+1)K$. *Then if* $\lambda \leq \frac{\bar{\lambda}_{\min}}{4d_{\boldsymbol{\psi}}}$ *and*

$$T^{1/3} \geq \widetilde{\Omega}\left(\left(D\lambda^{-2}(D^3\lambda^{-1} + C_{\mathcal{R}}\cdot\log^{p_{\mathcal{R}}}\frac{1}{\delta})\right)\cdot\frac{\bar{\lambda}_{\min}}{d_{\boldsymbol{\psi}}}\right), \tag{D.3.4}$$

with probability at least $1-\delta$,

$$\lambda_{\min}\left(\sum_{t=1}^{T}\boldsymbol{\psi}(\boldsymbol{\tau}_t)\right) \geq \frac{\bar{\lambda}_{\min}}{4d_{\boldsymbol{\psi}}}\cdot T.$$

Proof. Applying Corollary 6.5 with $\mathcal{H}=I$ and $\boldsymbol{\Gamma}_0=\lambda\cdot I$, we have that, with probability at least $1-\delta$:

$$\begin{aligned}
&\operatorname{tr}\left(\left(\sum_{t=1}^{T}\boldsymbol{\psi}(\boldsymbol{\tau}_t)+T\lambda\cdot I\right)^{-1}\right)\\
&\leq \frac{1}{T}\cdot\min_{\boldsymbol{\Gamma}\in\boldsymbol{\Omega}_{\boldsymbol{\psi}}}\operatorname{tr}((\boldsymbol{\Gamma}+\lambda\cdot I)^{-1})+\frac{8D^4\lambda^{-3}}{T(N+1)}+\frac{8D\lambda^{-2}(\log^{1/2}\frac{T}{\delta}+C_{\mathcal{R}}\log^{p_{\mathcal{R}}}\frac{T}{\delta})}{T\sqrt{K}}\\
&\leq \frac{1}{T}\cdot\min_{\boldsymbol{\Gamma}\in\boldsymbol{\Omega}_{\boldsymbol{\psi}}}\operatorname{tr}((\boldsymbol{\Gamma}+\lambda\cdot I)^{-1})+\frac{24D^4\lambda^{-3}}{T^{4/3}}+\frac{24D\lambda^{-2}(\log^{1/2}\frac{T}{\delta}+C_{\mathcal{R}}\log^{p_{\mathcal{R}}}\frac{T}{\delta})}{T^{4/3}}
\end{aligned}$$

where the second inequality follows since

$$T=\lceil 2^{i/3}\rceil\lceil 2^{2i/3}\rceil\leq 4\cdot 2^{i}$$

which implies $N+1=\lceil 2^{i/3}\rceil\geq T^{1/3}/4^{1/3}$ and $K=\lceil 2^{2i/3}\rceil\geq T^{2/3}/4^{2/3}$. If T satisfies (D.3.4), then we can bound

$$\frac{24D^4\lambda^{-3}}{T^{4/3}}+\frac{24D\lambda^{-2}(\log^{1/2}\frac{T}{\delta}+C_{\mathcal{R}}\log^{p_{\mathcal{R}}}\frac{T}{\delta})}{T^{4/3}}\leq\frac{1}{T}\cdot\frac{d_{\boldsymbol{\psi}}}{\bar{\lambda}_{\min}}.$$

Furthermore, under Assumption 6.3 there exists some $\boldsymbol{\Gamma}\in\boldsymbol{\Omega}_{\boldsymbol{\psi}}$ such that $\boldsymbol{\Gamma}\succeq\bar{\lambda}_{\min}\cdot I$, so we can upper bound

$$\min_{\boldsymbol{\Gamma}\in\boldsymbol{\Omega}_{\boldsymbol{\psi}}}\operatorname{tr}((\boldsymbol{\Gamma}+\lambda\cdot I)^{-1})\leq\frac{d_{\boldsymbol{\psi}}}{\bar{\lambda}_{\min}}$$

and we can lower bound

$$\operatorname{tr}\left(\left(\sum_{t=1}^{T}\boldsymbol{\psi}(\boldsymbol{\tau}_t)+T\lambda\cdot I\right)^{-1}\right)\geq\frac{1}{\lambda_{\min}(\sum_{t=1}^{T}\boldsymbol{\psi}(\boldsymbol{\tau}_t))+T\lambda}.$$

Thus,

$$\frac{1}{\lambda_{\min}(\sum_{t=1}^{T}\boldsymbol{\psi}(\boldsymbol{\tau}_t))+T\lambda}\leq\frac{1}{T}\cdot\frac{2d_{\boldsymbol{\psi}}}{\bar{\lambda}_{\min}}\implies\lambda_{\min}\left(\sum_{t=1}^{T}\boldsymbol{\psi}(\boldsymbol{\tau}_t)\right)\geq\frac{T\bar{\lambda}_{\min}}{2d_{\boldsymbol{\psi}}}-T\lambda.$$

It follows that if $\lambda \le \frac{\bar{\lambda}_{\min}}{4d_{\psi}}$, then we have

$$\lambda_{\min}\left(\sum_{t=1}^{T} \boldsymbol{\psi}(\boldsymbol{\tau}_t)\right) \ge T \cdot \frac{\bar{\lambda}_{\min}}{4d_{\psi}}$$

which proves the result. □

Lemma D.3.3. *Consider running some policies $(\pi_\tau)_{\tau=1}^T$, for π_τ $\mathcal{F}_{\tau-1}$-measurable, and collecting covariance $\boldsymbol{\Sigma}_T = \sum_{t=1}^T \boldsymbol{\psi}(\boldsymbol{\tau}_t)$. Then under Assumption 6.11, as long as*

$$\lambda_{\min}(\boldsymbol{\Sigma}_T) \ge 12544 D d_{\psi} \log \frac{2+32T}{\delta}$$

with probability at least $1-\delta$, if we rerun each $(\pi_\tau)_{\tau=1}^T$, we will collect features $\widetilde{\boldsymbol{\Sigma}}_T$ such that

$$\lambda_{\min}(\widetilde{\boldsymbol{\Sigma}}_T) \ge \frac{1}{2}\lambda_{\min}(\boldsymbol{\Sigma}_T).$$

Furthermore,

$$\lambda_{\min}\left(\sum_{\tau=1}^{T} \boldsymbol{\Gamma}_{\pi_\tau}\right) \ge \frac{1}{2}\lambda_{\min}(\boldsymbol{\Sigma}_T).$$

Proof. This follows from applying Lemma D.7 of Wagenmaker and Jamieson [2022] to the matrix $\frac{1}{D}\boldsymbol{\Sigma}_T$. Note that while Wagenmaker and Jamieson [2022] considers the setting of linear MDPs, the proof of Lemma D.7 of Wagenmaker and Jamieson [2022] does not make use of the linear MDP assumption, and the proof therefore extends immediately to our setting. Furthermore, though it is not explicitly stated, the lower bound on $\lambda_{\min}(\sum_{\tau=1}^T \boldsymbol{\Gamma}_{\pi_\tau})$ is also proved in Lemma D.7 of Wagenmaker and Jamieson [2022]. □

D.3.2 Rerunning Policies (LearnExpΠ)

In this section, we build on the analysis of the DynamicOED algorithm to show that, not only do the features collected by DynamicOED approximately minimize Φ, but that, under certain conditions, if we rerun the policies that DynamicOED ran to collect this data, we will collect a new set of features which also approximately minimizes Φ.

In particular, we specialize this argument to objectives of the form $\Phi(\boldsymbol{\Gamma}) = \mathrm{tr}(\mathcal{H}\boldsymbol{\Gamma}^{-1})$. LearnExpΠ (Algorithm D.2) proceeds by first calling MinEig to collect full-rank data, using this data as a regularizer of $\Phi(\boldsymbol{\Gamma})$, and the running DynamicOED on this objective. After meeting a certain termination criteria, it terminates, and returns the policies it has run over its operation.

Lemma D.3.4. *Let $\mathcal{E}_{\mathrm{exp}}$ denote the event that, for all $i = 1, 2, 3, \ldots$, the success event of*

Algorithm D.2 Learn Minimizing Exploration Policies (LearnExpΠ)

input: $\mathcal{H}$, iterates bound $\widetilde{N}$, confidence δ, regret minimization algorithm $\mathbb{A}_{\mathcal{R}}$, exploration policies Π_{exp}
for $i = 1, 2, 3, \ldots$ **do**
  $N_i \leftarrow \lceil 2^{i/3} \rceil - 1, K_i \leftarrow \lceil 2^{2i/3} \rceil, T_i \leftarrow (N_i + 1)K_i, \delta_i \leftarrow \delta/4i^2$
  $\Pi^i_{\textsc{MinEig}} \leftarrow \textsc{MinEig}(\widetilde{N}T_i, \delta_i, \mathbb{A}_{\mathcal{R}}, \Pi_{\text{exp}})$
  Run policies in $\Pi^i_{\textsc{MinEig}}$ $\lceil T_i / |\Pi^i_{\textsc{MinEig}}| \rceil$ times, set $\mathbf{\Gamma}^i_0$ to collected features
  $\Phi(\mathbf{\Gamma}) \leftarrow \operatorname{tr}(\mathcal{H}(\mathbf{\Gamma} + T_i^{-1}\mathbf{\Gamma}^i_0)^{-1})$
  $\mathbf{\Gamma}^i_{\text{fw}}, \Pi^i_{\text{fw}} \leftarrow \textsc{DynamicOED}(\Phi, N_i, K_i, \delta, \mathbb{A}_{\mathcal{R}}, \Pi_{\text{exp}})$
  if

$$\max_{j=1,\ldots,i} |\Pi^j_{\textsc{MinEig}}| \le T_i \tag{D.3.5}$$

$$\frac{16D^4\|\mathcal{H}\|_{\text{op}}\|(T_i^{-1}\mathbf{\Gamma}_0)^{-1}\|^3_{\text{op}}}{T_i(N_i+1)} + \frac{16D\|\mathcal{H}\|_{\text{op}}\|(T_i^{-1}\mathbf{\Gamma}_0)^{-1}\|^2_{\text{op}}(\log^{1/2}\frac{4T_i}{\delta} + C_{\mathcal{R}}\log^{p_{\mathcal{R}}}\frac{2T_i}{\delta})}{T_i\sqrt{K_i}} \le \operatorname{tr}\left(\mathcal{H}\left(\mathbf{\Gamma}^i_{\text{fw}} + \mathbf{\Gamma}^i_0\right)^{-1}\right) \tag{D.3.6}$$

$$\operatorname{tr}(\mathcal{H}) \cdot D\sqrt{2T_i}\sqrt{8d_\psi \log(1 + 8\sqrt{2T_i}) + 8\log 1/\delta} \cdot \frac{2}{\lambda_{\min}\left(\mathbf{\Gamma}^i_{\text{fw}} + \mathbf{\Gamma}^i_0\right)^2} \le \operatorname{tr}\left(\mathcal{H}\left(\mathbf{\Gamma}^i_{\text{fw}} + \mathbf{\Gamma}^i_0\right)^{-1}\right) \tag{D.3.7}$$

$$D\sqrt{2T_i}\sqrt{8d_\psi \log(1 + 8\sqrt{2T_i}) + 8\log 1/\delta} \le \frac{1}{2}\lambda_{\min}\left(\mathbf{\Gamma}^i_{\text{fw}} + \mathbf{\Gamma}^i_0\right) \tag{D.3.8}$$

  then
    $\mathbf{\Gamma}_{\text{out}} \leftarrow \mathbf{\Gamma}^i_{\text{fw}} + \mathbf{\Gamma}^i_0, \Pi_{\text{out}} \leftarrow \Pi^i_{\textsc{MinEig}} \cup (\cup_{j=1}^{\lceil T_i/|\Pi_{\textsc{MinEig}}|\rceil} \Pi^i_{\text{fw}})$
    return $\mathbf{\Gamma}_{\text{out}}, \Pi_{\text{out}}$

MinEig (Lemma D.3.1) and DynamicOED occur, and

$$\lambda_{\min}(\mathbf{\Gamma}^i_0) \ge \lceil T_i/|\Pi^i_{\textit{MinEig}}| \rceil \cdot 6272 D d_\psi \log \frac{68\widetilde{N}}{\delta}.$$

Then if Assumptions 6.10, 6.11 and D.1 hold, $\mathbb{P}[\mathcal{E}_{\text{exp}}] \ge 1 - \delta$.

Lemma D.3.5. *Consider rerunning each policy in* Π_{out} $N \le \widetilde{N}$ *times, and let* $\widetilde{\mathbf{\Gamma}}$ *denote the obtained features. Then, if Assumptions 6.10, 6.11 and D.1 hold, with probability at least*

$1-3\delta$, *on the event* $\mathcal{E}_{\text{exp}}$*:*

$$\left\|\widetilde{\mathbf{\Gamma}} - N \cdot \textstyle\sum_{\pi\in\Pi_{\text{out}}} \mathbf{\Gamma}_\pi\right\|_{\text{op}} \le \frac{\sqrt{T_{\text{out}}} \cdot \sqrt{8d_{\boldsymbol{\psi}} \log(1+8\sqrt{NT_{\text{out}}}) + 8\log 1/\delta}}{\sqrt{N} \cdot 6272 d_{\boldsymbol{\psi}} \log \frac{68\widetilde{N}}{\delta}} \cdot \lambda_{\min}\left(N \cdot \textstyle\sum_{\pi\in\Pi_{\text{out}}} \mathbf{\Gamma}_\pi\right), \tag{D.3.9}$$

$$N \cdot 6272 D d_{\boldsymbol{\psi}} \log \frac{68\widetilde{N}}{\delta} \le \min\left\{\lambda_{\min}\left(N \cdot \textstyle\sum_{\pi\in\Pi_{\text{out}}} \mathbf{\Gamma}_\pi\right), \lambda_{\min}(\widetilde{\mathbf{\Gamma}})\right\}, \tag{D.3.10}$$

and

$$\operatorname{tr}\left(\mathcal{H}\left(\textstyle\sum_{\pi\in\Pi_{\text{out}}} \mathbf{\Gamma}_\pi\right)^{-1}\right) \le \frac{12}{T_{\text{out}}} \cdot \min_{\mathbf{\Gamma}\in\mathbf{\Omega}} \operatorname{tr}\left(\mathcal{H}\mathbf{\Gamma}^{-1}\right) \tag{D.3.11}$$

for $T_{\text{out}} := |\Pi_{\text{out}}|$.

Lemma D.3.6. *On the event* $\mathcal{E}_{\text{exp}}$*, under Assumptions 6.10, 6.11 and D.1, we can bound*

$$T_{\text{out}} \le \text{poly}\left(d_{\boldsymbol{\psi}}, \frac{1}{\bar{\lambda}_{\min}}, D, C_{\mathcal{R}}, \log^{p_{\mathcal{R}}} \frac{\widetilde{N}}{\delta}\right).$$

Furthermore, the total number of episodes collected by Algorithm D.2 is bounded by $(16 + 2\log(T_{\text{out}})) \cdot T_{\text{out}}$.

D.3.2.1 Supporting Lemmas and Proofs for LearnExpΠ

Lemma D.3.7. *Under Assumption 6.11, for any* $\mathbf{\Gamma} = \mathbb{E}_{\boldsymbol{\tau}\sim\omega}[\boldsymbol{\psi}(\boldsymbol{\tau})]$ *and* $\mathcal{H} \succeq 0$ *we can bound*

$$\operatorname{tr}(\mathcal{H}\mathbf{\Gamma}^{-1}) \ge D^{-1} \cdot \operatorname{tr}(\mathcal{H}).$$

Proof. By Von Neumann's Trace Inequality we can lower bound

$$\operatorname{tr}(\mathcal{H}\mathbf{\Gamma}^{-1}) \ge \lambda_{\min}(\mathbf{\Gamma}^{-1}) \cdot \operatorname{tr}(\mathcal{H}) = \|\mathbf{\Gamma}\|_{\text{op}}^{-1} \cdot \operatorname{tr}(\mathcal{H}).$$

By our assumption that $\operatorname{tr}(\boldsymbol{\psi}(\boldsymbol{\tau})) \le D$, we can bound $\|\mathbf{\Gamma}\|_{\text{op}} \le D$, which proves the result. □

Lemma D.3.8. *Assume* $\operatorname{tr}(\boldsymbol{\psi}(\boldsymbol{\tau})) \le D$ *for all* $\boldsymbol{\tau}$*. Let* $\mathbf{\Gamma}_K$ *denote the time-normalized features obtained by playing policies* $\{\pi_k\}_{k=1}^K$*, where* π_k *is* $\mathcal{F}_{k-1}$*-measurable. Then, with probability at least* $1-\delta$*,*

$$\left\|\frac{1}{K}\sum_{k=1}^{K} \mathbf{\Gamma}_{\pi_k} - \mathbf{\Gamma}_K\right\|_{\text{op}} \le D\sqrt{\frac{8d_{\boldsymbol{\psi}} \log(1+8\sqrt{K}) + 8\log 1/\delta}{K}}.$$

Proof. This follows from an argument identical to the proof of Lemma C.4 of Wagenmaker and Jamieson [2022]. While Wagenmaker and Jamieson [2022] considers the setting of linear MDPs, we note that the proof of Lemma C.4 of Wagenmaker and Jamieson [2022] nowhere relies on the linear MDP assumption. The result stated here then follows identically as Lemma C.4 of Wagenmaker and Jamieson [2022], after normalizing $\boldsymbol{\psi}(\boldsymbol{\tau})$ by D. □

Proof of Lemma D.3.4. By Lemma D.3.1, the failure probability of running MinEig at round i is $\delta_i = \delta/8i^2$, and by Corollary 6.5 the failure probability of DynamicOED at round i is also bounded by $\delta_i = \delta/8i^2$. It follows that the total failure probability of running MinEig and DynamicOED is bounded by

$$\sum_{i=1}^{\infty} \frac{2 \cdot \delta}{8i^2} \leq \frac{\delta}{2}.$$

Furthermore, by Lemma D.3.1, we have that rerunning all policies in $\Pi^i_{\textsc{MinEig}}$, we will obtain features $\boldsymbol{\Gamma}$ satisfying, with probability at least $1 - \delta_i/\widetilde{N}T_i$:

$$\lambda_{\min}(\boldsymbol{\Gamma}) \geq 6272 D d_{\boldsymbol{\psi}} \log \frac{68\widetilde{N}}{\delta_i}.$$

Repeating this $\lceil T_i/|\Pi^i_{\textsc{MinEig}}| \rceil$ times and union bounding, we have that

$$\lambda_{\min}(\boldsymbol{\Gamma}^i_0) \geq \lceil T_i/|\Pi^i_{\textsc{MinEig}}| \rceil \cdot 6272 D d_{\boldsymbol{\psi}} \log \frac{68\widetilde{N}}{\delta_i}$$

with probability at least $1 - \delta/\widetilde{N}T_i \cdot \lceil T_i/|\Pi^i_{\textsc{MinEig}}| \rceil \geq 1 - \delta/\widetilde{N}$. □

Proof of Lemma D.3.5. Let $\Pi_{\textsc{MinEig}}, \boldsymbol{\Gamma}_0, \Pi_{\mathrm{fw}}, \boldsymbol{\Gamma}_{\mathrm{fw}}$ denote the policies and features obtained on the round at which Algorithm D.2 terminates. Let $T_{\mathrm{fw}} = |\Pi_{\mathrm{fw}}|$ denote the number of episodes of DynamicOED on the terminating round, and $N_{\mathrm{fw}}, K_{\mathrm{fw}}$ the corresponding values of N_i and K_i. Throughout the proof we make use of the fact that at termination of Algorithm D.2, all of (D.3.5)-(D.3.8) are met.

Proof of (D.3.9) and (D.3.10). By Lemma D.3.8 we have that, with probability at least $1 - \delta$:

$$\left\| \widetilde{\boldsymbol{\Gamma}} - N \cdot \textstyle\sum_{\pi \in \Pi_{\mathrm{out}}} \boldsymbol{\Gamma}_\pi \right\|_{\mathrm{op}} \leq D\sqrt{NT_{\mathrm{out}}} \cdot \sqrt{8 d_{\boldsymbol{\psi}} \log(1 + 8\sqrt{NT_{\mathrm{out}}}) + 8 \log 1/\delta}.$$

On $\mathcal{E}_{\mathrm{exp}}$, by Lemma D.3.1, we can bound

$$|\Pi_{\textsc{MinEig}}| \leq \mathrm{poly}\left(d_{\boldsymbol{\psi}}, \frac{1}{\bar{\lambda}_{\min}}, D, C_{\mathcal{R}}, \log^{p_{\mathcal{R}}} \frac{\widetilde{N}T_{\mathrm{out}}}{\delta} \right)$$

and, furthermore, we can lower bound

$$\lambda_{\min}\left(\sum_{\pi\in\Pi_{\textsc{MinEig}}}\mathbf{\Gamma}_\pi\right) \geq 6272Dd_{\boldsymbol{\psi}}\log\frac{68\widetilde{N}}{\delta}.$$

Since $\Pi_{\textsc{MinEig}}\subseteq\Pi_{\text{out}}$, it follows that

$$\lambda_{\min}\left(N\cdot\sum_{\pi\in\Pi_{\text{out}}}\mathbf{\Gamma}_\pi\right) \geq N\cdot 6272Dd_{\boldsymbol{\psi}}\log\frac{68\widetilde{N}}{\delta}.$$

Combining these, we therefore have that, with probability at least $1-\delta$:

$$\left\|\widetilde{\mathbf{\Gamma}}-N\cdot\sum_{\pi\in\Pi_{\text{out}}}\mathbf{\Gamma}_\pi\right\|_{\text{op}} \leq \frac{\sqrt{NT_{\text{out}}}\cdot\sqrt{8d_{\boldsymbol{\psi}}\log(1+8\sqrt{NT_{\text{out}}})+8\log 1/\delta}}{N\cdot 6272d_{\boldsymbol{\psi}}\log\frac{68\widetilde{N}}{\delta}}\cdot\lambda_{\min}\left(N\cdot\sum_{\pi\in\Pi_{\text{out}}}\mathbf{\Gamma}_\pi\right).$$

In addition, also by Lemma D.3.1, we have that with probability at least $1-\delta/\widetilde{N}T_{\text{out}}\cdot N\geq 1-\delta$, that

$$\lambda_{\min}(\widetilde{\mathbf{\Gamma}}) \geq N\cdot 6272Dd_{\boldsymbol{\psi}}\log\frac{68\widetilde{N}}{\delta}.$$

Proof of (D.3.11)**.** By Corollary 6.5, on $\mathcal{E}_{\text{exp}}$ we have that:

$$\Phi(\mathbf{\Gamma}_{\text{fw}}) = \operatorname{tr}\left(\mathcal{H}\left(\mathbf{\Gamma}_{\text{fw}}+\mathbf{\Gamma}_0\right)^{-1}\right) \leq \frac{\min_{\mathbf{\Gamma}\in\mathbf{\Omega}}\operatorname{tr}\left(\mathcal{H}(\mathbf{\Gamma}+T_{\text{fw}}^{-1}\mathbf{\Gamma}_0)^{-1}\right)}{T_{\text{fw}}}$$
$$+\frac{8D^4\|\mathcal{H}\|_{\text{op}}\|(T_{\text{fw}}^{-1}\mathbf{\Gamma}_0)^{-1}\|_{\text{op}}^3}{T_{\text{fw}}(N_{\text{fw}}+1)}+\frac{8D\|\mathcal{H}\|_{\text{op}}\|(T_{\text{fw}}^{-1}\mathbf{\Gamma}_0)^{-1}\|_{\text{op}}^2(\log^{1/2}\frac{4T_{\text{fw}}}{\delta}+C_{\mathcal{R}}\log^{p_{\mathcal{R}}}\frac{2T_{\text{fw}}}{\delta})}{T_{\text{fw}}\sqrt{K_{\text{fw}}}}.$$

Since T_{fw} satisfies (D.3.6), we can bound

$$\frac{8D^4\|\mathcal{H}\|_{\text{op}}\|(T_{\text{fw}}^{-1}\mathbf{\Gamma}_0)^{-1}\|_{\text{op}}^3}{T_{\text{fw}}(N_{\text{fw}}+1)}+\frac{8D\|\mathcal{H}\|_{\text{op}}\|(T_{\text{fw}}^{-1}\mathbf{\Gamma}_0)^{-1}\|_{\text{op}}^2(\log^{1/2}\frac{4T_{\text{fw}}}{\delta}+C_{\mathcal{R}}\log^{p_{\mathcal{R}}}\frac{2T_{\text{fw}}}{\delta})}{T_{\text{fw}}\sqrt{K_{\text{fw}}}} \leq \frac{1}{2}\Phi(\mathbf{\Gamma}_{\text{fw}}).$$

It follows that

$$\Phi(\mathbf{\Gamma}_{\text{fw}}) \leq \frac{\min_{\mathbf{\Gamma}\in\mathbf{\Omega}}\operatorname{tr}\left(\mathcal{H}(\mathbf{\Gamma}+T_{\text{fw}}^{-1}\mathbf{\Gamma}_0)^{-1}\right)}{T_{\text{fw}}}+\frac{1}{2}\Phi(\mathbf{\Gamma}_{\text{fw}})$$
$$\implies \Phi(\mathbf{\Gamma}_{\text{fw}}) \leq 2\cdot\frac{\min_{\mathbf{\Gamma}\in\mathbf{\Omega}}\operatorname{tr}\left(\mathcal{H}(\mathbf{\Gamma}+T_{\text{fw}}^{-1}\mathbf{\Gamma}_0)^{-1}\right)}{T_{\text{fw}}}. \tag{D.3.12}$$

By Lemma D.3.8, we have that, with probability at least $1-\delta$:

$$\left\|\sum_{\pi\in\Pi_{\text{out}}}\mathbf{\Gamma}_\pi-(\mathbf{\Gamma}_{\text{fw}}+\mathbf{\Gamma}_0)\right\|_{\text{op}}\leq D\sqrt{T_{\text{out}}}\sqrt{8d_{\boldsymbol{\psi}}\log(1+8\sqrt{T_{\text{out}}})+8\log 1/\delta}.$$

Since (D.3.8) is satisfied and $|\Pi^i_{\textsc{MinEig}}|\leq T_i$, we have

$$D\sqrt{T_{\text{out}}}\sqrt{8d_{\boldsymbol{\psi}}\log(1+8\sqrt{T_{\text{out}}})+8\log 1/\delta}\leq\frac{1}{2}\lambda_{\min}\left(\mathbf{\Gamma}_{\text{fw}}+\mathbf{\Gamma}_0\right).$$

By Lemma C.3.1 it follows that

$$\left\|\left(\sum_{\pi\in\Pi_{\text{out}}}\mathbf{\Gamma}_\pi\right)^{-1}-\left(\mathbf{\Gamma}_{\text{fw}}+\mathbf{\Gamma}_0\right)^{-1}\right\|_{\text{op}}\leq D\sqrt{T_{\text{out}}}\sqrt{8d_{\boldsymbol{\psi}}\log(1+8\sqrt{T_{\text{out}}})+8\log 1/\delta}\cdot\frac{2}{\lambda_{\min}\left(\mathbf{\Gamma}_{\text{fw}}+\mathbf{\Gamma}_0\right)^2}.$$

This implies that

$$\begin{aligned}\text{tr}\left(\mathcal{H}\left(\sum_{\pi\in\Pi_{\text{out}}}\mathbf{\Gamma}_\pi\right)^{-1}\right)&\leq\text{tr}\left(\mathcal{H}\left(\mathbf{\Gamma}_{\text{fw}}+\mathbf{\Gamma}_0\right)^{-1}\right)\\&\quad+\text{tr}(\mathcal{H})\cdot D\sqrt{T_{\text{out}}}\sqrt{8d_{\boldsymbol{\psi}}\log(1+8\sqrt{T_{\text{out}}})+8\log 1/\delta}\cdot\frac{2}{\lambda_{\min}\left(\mathbf{\Gamma}_{\text{fw}}+\mathbf{\Gamma}_0\right)^2}.\end{aligned}$$

Now if

$$\text{tr}(\mathcal{H})\cdot D\sqrt{T_{\text{out}}}\sqrt{8d_{\boldsymbol{\psi}}\log(1+8\sqrt{T_{\text{out}}})+8\log 1/\delta}\cdot\frac{2}{\lambda_{\min}\left(\mathbf{\Gamma}_{\text{fw}}+\mathbf{\Gamma}_0\right)^2}\leq\text{tr}\left(\mathcal{H}\left(\mathbf{\Gamma}_{\text{fw}}+\mathbf{\Gamma}_0\right)^{-1}\right),\tag{D.3.13}$$

we can bound this all by

$$\leq 2\text{tr}\left(\mathcal{H}\left(\mathbf{\Gamma}_{\text{fw}}+\mathbf{\Gamma}_0\right)^{-1}\right)\leq\frac{4}{T_{\text{fw}}}\cdot\min_{\mathbf{\Gamma}\in\boldsymbol{\Omega}}\text{tr}\left(\mathcal{H}(\mathbf{\Gamma}+T_{\text{fw}}^{-1}\mathbf{\Gamma}_0)^{-1}\right)$$

where the last inequality follows from (D.3.12). However, note that (D.3.13) since (D.3.7) holds. Finally, note that

$$T_{\text{out}}=T_{\text{fw}}+\lceil T_{\text{fw}}/|\Pi_{\textsc{MinEig}}|\rceil|\Pi_{\textsc{MinEig}}|\leq 2T_{\text{fw}}+|\Pi_{\textsc{MinEig}}|\leq 3T_{\text{fw}}$$

where the last inequality follows since (D.3.5) holds. We can therefore upper bound $\frac{4}{T_{\text{fw}}}\leq\frac{12}{T_{\text{out}}}$. Putting this together proves the result.

□

Proof of Lemma D.3.6. To bound T_{out}, it suffices to show that (D.3.5)-(D.3.8) are satisfied for sufficiently large T_i.

On $\mathcal{E}_{\text{exp}}$, by Lemma D.3.1, we can bound

$$|\Pi^i_{\textsc{MinEig}}| \le \text{poly}\left(d_{\boldsymbol{\psi}}, \frac{1}{\bar{\lambda}_{\min}}, D, C_{\mathcal{R}}, \log^{p_{\mathcal{R}}} \frac{\widetilde{N} T_i}{\delta}\right),$$

so to ensure (D.3.5) is met it suffices that

$$T_i \ge \text{poly}\left(d_{\boldsymbol{\psi}}, \frac{1}{\bar{\lambda}_{\min}}, D, C_{\mathcal{R}}, \log^{p_{\mathcal{R}}} \frac{\widetilde{N}}{\delta}\right).$$

On $\mathcal{E}_{\text{exp}}$, we have

$$\lambda_{\min}(\boldsymbol{\Gamma}^i_{\text{fw}} + \boldsymbol{\Gamma}^i_0) \ge \lambda_{\min}(\boldsymbol{\Gamma}^i_0) \ge \lceil T_i / |\Pi^i_{\textsc{MinEig}}| \rceil \cdot 6272 D d_{\boldsymbol{\psi}} \log \frac{68\widetilde{N}}{\delta}.$$

Which also implies

$$\|(T_i^{-1}\boldsymbol{\Gamma}^i_0)^{-1}\|_{\text{op}} = \frac{T_i}{\lambda_{\min}(\boldsymbol{\Gamma}^i_0)} \le \frac{T_i}{\lceil T_i/|\Pi^i_{\textsc{MinEig}}|\rceil} \cdot \frac{1}{6272 D d_{\boldsymbol{\psi}} \log \frac{68\widetilde{N}}{\delta}} \le \frac{|\Pi^i_{\textsc{MinEig}}|}{6272 D d_{\boldsymbol{\psi}} \log \frac{68\widetilde{N}}{\delta}}$$

Furthermore, by Lemma D.3.7 we can lower bound

$$\text{tr}\left(\mathcal{H}\left(\boldsymbol{\Gamma}^i_{\text{fw}} + \boldsymbol{\Gamma}^i_0\right)^{-1}\right) \ge \frac{\text{tr}(\mathcal{H})}{D(T_i + \lceil T_i/|\Pi^i_{\textsc{MinEig}}|\rceil |\Pi^i_{\textsc{MinEig}}|)} \ge \frac{\text{tr}(\mathcal{H})}{3DT_i}.$$

Combining these and using that $N_i = \mathcal{O}(T_i^{1/3})$ and $K_i = \mathcal{O}(T_i^{2/3})$, it is easy to see that (D.3.6)-(D.3.8) will be met once

$$T_i \ge \text{poly}\left(d_{\boldsymbol{\psi}}, \frac{1}{\bar{\lambda}_{\min}}, D, C_{\mathcal{R}}, \log^{p_{\mathcal{R}}} \frac{\widetilde{N}}{\delta}\right).$$

The bound on T_{out} then follows since $T_i = \lceil 2^{i/3} \rceil \lceil 2^{2i/3} \rceil \in [2^i, 4 \cdot 2^i]$, so it can be at most a constant larger than the sufficient condition before terminating.

Let $i^\star$ denote the round that Algorithm D.2 terminates on. Note that at round i, MinEig runs for at most $|\Pi^i_{\textsc{MinEig}}|$, DynamicOED runs for at most T_i episodes, and we run for an additional $\lceil T_i/|\Pi^i_{\textsc{MinEig}}|\rceil \cdot |\Pi^i_{\textsc{MinEig}}|$ episodes on Line 5. In total, then, the number of episodes Algorithm D.2 runs for is bounded by

$$\sum_{i=1}^{i^\star} (T_i + |\Pi^i_{\textsc{MinEig}}| + \lceil T_i/|\Pi^i_{\textsc{MinEig}}|\rceil \cdot |\Pi^i_{\textsc{MinEig}}|) \le 2 \sum_{i=1}^{i^\star} (T_i + |\Pi^i_{\textsc{MinEig}}|)$$

$$\leq 16T_{i^\star} + 2\sum_{i=1}^{i^\star} |\Pi^i_{\textsc{MinEig}}|$$

where the last inequality follows since $T_i \in [2^i, 4 \cdot 2^i]$. Now note that, since Algorithm D.2 only terminates once (D.3.5) is met, we will have $\max_{j=1,\dots,i^\star} |\Pi^j_{\textsc{MinEig}}| \leq T_{i^\star}$. This implies that $2\sum_{i=1}^{i^\star} |\Pi^i_{\textsc{MinEig}}| \leq 2i^\star T_{i^\star} \leq 2\log(T_{i^\star}) \cdot T_{i^\star}$. Bounding $T_{i^\star} \leq T_{\text{out}}$ gives the result.

□

Appendix E

Deferred Proofs from Chapter 7

E.1 Deferred Proofs from Section 7.2.1 and Section 7.3

E.1.1 Interpreting the Gap-Visitation Complexity

Proposition E.1. *The gap-visitation complexity, $\mathcal{C}(M, \epsilon)$, satisfies*

$$\mathcal{C}(M,\epsilon) = \sum_{h=1}^{H} \inf_{\pi} \max_{s} \frac{1}{w_h^\pi(s)} \sum_{a} \min\left\{ \frac{1}{\widetilde{\Delta}_h(s,a)^2}, \frac{W_h(s)^2}{\epsilon^2} \right\} + \frac{H^2 |\mathrm{OPT}(\epsilon)|}{\epsilon^2}.$$

Furthermore, when M has unique optimal actions, the best-policy gap-visitation complexity, $\mathcal{C}^\star(M)$, satisfies

$$\mathcal{C}^\star(M) = \sum_{h=1}^{H} \inf_{\pi} \max_{s} \frac{1}{w_h^\pi(s)} \sum_{a:\Delta_h(s,a)>0} \frac{1}{\Delta_h(s,a)^2}.$$

Proof. Consider the optimization

$$\min_{\lambda \in \triangle(X)} \max_{x \in X} a_x/\lambda_x.$$

It is easy to see that

$$\sum_{x \in X} a_x = \min_{\lambda \in \triangle(X)} \max_{x \in X} a_x/\lambda_x$$

and the optimal λ is

$$\lambda_x^* = \frac{a_x}{\sum_{x' \in X} a_{x'}}.$$

For any policy π, we will have that $\sum_a \pi_h(a|s) = 1$, and $\pi_h(a|s)$ must be a valid distribution over a. This implies that $w_h^\pi(s,a) = w_h^\pi(s)\pi_h(a|s)$. Now fix π for steps $h' = 1, \ldots, h-1$, then

it follows that

$$\inf_{\pi_h} \max_{s,a} \min \left\{ \frac{1}{w_h^\pi(s,a)\widetilde{\Delta}_h(s,a)^2}, \frac{W_h(s)^2}{w_h^\pi(s,a)\epsilon^2} \right\} = \inf_{\pi_h} \max_s \frac{1}{w_h^\pi(s)} \max_a \frac{1}{\pi_h(a|s)} \min \left\{ \frac{1}{\widetilde{\Delta}_h(s,a)^2}, \frac{W_h(s)^2}{\epsilon^2} \right\}.$$

Now for a given s, we can use that $w_h^\pi(s)$ is independent of π_h and apply our above calculation to get that

$$\inf_{\pi_h} \frac{1}{w_h^\pi(s)} \max_a \frac{1}{\pi_h(a|s)} \min \left\{ \frac{1}{\widetilde{\Delta}_h(s,a)^2}, \frac{W_h(s)^2}{\epsilon^2} \right\} = \frac{1}{w_h^\pi(s)} \sum_a \min \left\{ \frac{1}{\widetilde{\Delta}_h(s,a)^2}, \frac{W_h(s)^2}{\epsilon^2} \right\}.$$

As the maximum over s is over a finite set and $\pi_h(\cdot|s)$ can be chosen independently of $\pi_h(\cdot|s')$ for any $s \neq s'$, we have that

$$\inf_{\pi_h} \max_s \frac{1}{w_h^\pi(s)} \max_a \frac{1}{\pi_h(a|s)} \min \left\{ \frac{1}{\widetilde{\Delta}_h(s,a)^2}, \frac{W_h(s)^2}{\epsilon^2} \right\} = \max_s \frac{1}{w_h^\pi(s)} \sum_a \min \left\{ \frac{1}{\widetilde{\Delta}_h(s,a)^2}, \frac{W_h(s)^2}{\epsilon^2} \right\}.$$

Since taking an inf over π is equivalent to taking an inf over $\{\pi_{h'}\}_{h'=1}^{h-1}$ and π_h, we can take the inf of this over $\{\pi_{h'}\}_{h'=1}^{h-1}$ to get

$$\inf_{\pi} \max_{s,a} \min \left\{ \frac{1}{w_h^\pi(s,a)\widetilde{\Delta}_h(s,a)^2}, \frac{W_h(s)^2}{w_h^\pi(s,a)\epsilon^2} \right\} = \inf_{\pi} \max_s \frac{1}{w_h^\pi(s)} \sum_a \min \left\{ \frac{1}{\widetilde{\Delta}_h(s,a)^2}, \frac{W_h(s)^2}{\epsilon^2} \right\}.$$

The same line of reasoning can be used to obtain the expression for $\mathcal{C}^\star(M)$. □

Proof of Proposition 7.2. Let π^{sh} denote the policy that achieves $w_h^{\pi^{sh}}(s) = W_h(s)$. Consider the state visitation distribution:

$$w'_h(s) = \frac{\sum_{s'} w_h^{\pi^{s'h}}(s) \cdot \sum_a \min \left\{ \frac{1}{W_h(s')\widetilde{\Delta}_h(s',a)^2}, \frac{W_h(s')}{\epsilon^2} \right\}}{\sum_{s',a} \min \left\{ \frac{1}{W_h(s')\widetilde{\Delta}_h(s',a)^2}, \frac{W_h(s')}{\epsilon^2} \right\}}.$$

Since the set of state visitations realizable on a given MDP is convex and for any realizable state distribution there exists a policy with that state distribution by Proposition E.4, and since w'_h is a convex combination of state visitation distributions, it follows that there exists some policy $\widetilde{\pi}$ such that $w'_h(s) = w_h^{\widetilde{\pi}}(s)$. Furthermore, by definition,

$$w_h^{\widetilde{\pi}}(s) \geq \frac{w_h^{\pi^{sh}}(s) \cdot \sum_a \min \left\{ \frac{1}{W_h(s)\widetilde{\Delta}_h(s,a)^2}, \frac{W_h(s)}{\epsilon^2} \right\}}{\sum_{s',a} \min \left\{ \frac{1}{W_h(s')\widetilde{\Delta}_h(s',a)^2}, \frac{W_h(s')}{\epsilon^2} \right\}} = W_h(s) \cdot \frac{\sum_a \min \left\{ \frac{1}{W_h(s)\widetilde{\Delta}_h(s,a)^2}, \frac{W_h(s)}{\epsilon^2} \right\}}{\sum_{s',a} \min \left\{ \frac{1}{W_h(s')\widetilde{\Delta}_h(s',a)^2}, \frac{W_h(s')}{\epsilon^2} \right\}}.$$

Thus, since $\widetilde{\pi}$ is a feasible policy, using the expression for $\mathcal{C}(M^\star, \epsilon)$ given in Proposition E.1,

it follows that

$$\mathcal{C}(M^\star, \epsilon) = \sum_{h=1}^{H} \inf_{\pi} \max_{s} \frac{1}{w_h^\pi(s)} \sum_a \min\left\{ \frac{1}{\widetilde{\Delta}_h(s,a)^2}, \frac{W_h(s)^2}{\epsilon^2} \right\} + \frac{H^2|\mathrm{OPT}(\epsilon)|}{\epsilon^2}$$
$$\leq \sum_{h=1}^{H} \sum_{s,a} \min\left\{ \frac{1}{W_h(s)\widetilde{\Delta}_h(s,a)^2}, \frac{W_h(s)}{\epsilon^2} \right\} + \frac{H^2|\mathrm{OPT}(\epsilon)|}{\epsilon^2}.$$

To obtain the first bound, we use the second bound to get

$$\mathcal{C}(M^\star, \epsilon) \leq \sum_{s,a,h} \frac{H^2 W_h(s)}{\epsilon^2} \leq \frac{H^3 SA}{\epsilon^2}$$

and use that $|\mathrm{OPT}(\epsilon_{\mathrm{tol}})| \leq SAH$. □

Proof of Proposition 7.3. This follows directly from Proposition E.1. □

E.1.2 Proof of Proposition 7.4

Instance Class E.1.1. Given gap parameters $\Delta_1, \Delta_2 > 0$ and transition probability $p \in (0, 1/2)$, consider an MDP with $H = S = A = 2$ which always starts in state s_0 and has rewards and transitions defined as (where we drop the horizon subscript for simplicity):

$$P(s_1|s_0,a_1) = 1-p, \quad P(s_2|s_0,a_1) = p, \quad P(s_1|s_0,a_2) = 0, \quad P(s_2|s_0,a_2) = 1$$
$$R(s_0,a_1) \sim \text{Bernoulli}(1), \quad R(s_0,a_2) \sim \text{Bernoulli}(0)$$
$$R(s_i,a_1) \sim \text{Bernoulli}(0.5+\Delta_i), \quad R(s_i,a_2) \sim \text{Bernoulli}(0.5), i \in \{1,2\}$$

At $h = 2$, we can then think of each state as simply a two-armed bandit with gap Δ_i. We assume that $p < 1/2$, so that $1-p$ can be thought of as a constant. This instance is illustrated in Figure 7.1.

Proposition E.2 (Formal Statement of Proposition 7.4)**.** *Given any MDP in Instance Class E.1.1, any learner executing Protocol 7.3.1 which computes an optimal policy with probability at least $1-\delta$ must collect at least*

$$K \geq \Omega\left(\frac{\log 1/\delta}{\Delta_1^2} + \frac{\log 1/\delta}{p\Delta_2^2} \right)$$

episodes, as long as $\frac{\log 1/\delta}{\Delta_2^2} \geq c \max\{C_2, C_1^{\frac{1}{1-\alpha}} p^{\frac{-\alpha}{1-\alpha}}\}$, for a universal constant c. However, on this instance,

$$\mathcal{C}^\star(M^\star) \leq \mathcal{O}\left(\frac{1}{\Delta_1^2} + \frac{1}{\Delta_2^2} \right)$$

and so, with probability $1-\delta$, MOCA *will terminate in at most* $K \leq \widetilde{\mathcal{O}}(\mathcal{C}^\star(M^\star)\cdot \log 1/\delta)$ *episodes and return the optimal policy.*

Proof of Proposition 7.4. To get the complexity bound of MOCA, we apply Theorem 7.1 and Proposition E.1. The stated complexity follows since $W_2(s_1) = 1-p \geq 1/2$ and $W_2(s_2) = 1$, from which the stated complexity follows directly.

Complexity of Low-Regret Algorithms. The expected regret of any algorithm is given by

$$N_1(s_0,a_2) + \Delta_1 N_2(s_1,a_2) + \Delta_2 N_2(s_2,a_2)$$

where we let $N_h(s_i,a_j)$ denote the expected number of times action a_j is taken in state s_i at timestep h. Our assumption on the regret implies that $N_1(s_1,a_2) \leq C_1 K^\alpha + C_2$.

From standard lower bounds on bandits (Theorem 4 of Kaufmann et al. [2016]), and using that for small Δ $\mathrm{KL}(\mathrm{Bernoulli}(0.5)||\mathrm{Bernoulli}(0.5+\Delta)) = \Theta(\Delta^2)$, to solve the bandit in s_1 with probability at least $1-\delta$, we must have that $N_2(s_1) \geq c\frac{\log 1/\delta}{\Delta_1^2}$, and similarly, to solve the bandit in s_2, we must have that $N_2(s_2) \geq c\frac{\log 1/\delta}{\Delta_2^2}$, for an absolute constant c.

Note that $N_2(s_1) = (1-p)N_1(s_0,a_1)$ and $N_2(s_2) = pN_1(s_0,a_1) + N_1(s_0,a_2)$, and that the total number of episodes run is $N_1(s_0,a_1) + N_1(s_0,a_2)$. This implies that we must have

$$N_1(s_0,a_1) \geq \frac{c\log 1/\delta}{(1-p)\Delta_1^2}, \quad pN_1(s_0,a_1) + N_1(s_0,a_2) \geq \frac{c\log 1/\delta}{\Delta_2^2}.$$

However, since $N_1(s_0,a_2) \leq C_1K^\alpha + C_2$, $N_1(s_0,a_1)$ must at least satisfy

$$pN_1(s_0,a_1) + C_1K^\alpha + C_2 \geq \frac{\log 1/\delta}{\Delta_2^2} \implies N_1(s_0,a_1) \geq \frac{1}{p}\left(\frac{c\log 1/\delta}{\Delta_2^2} - C_1K^\alpha - C_2\right).$$

Thus, we need

$$K = N_1(s_0,a_1) + N_1(s_0,a_2) \geq N_1(s_0,a_1) \geq \max\left\{\frac{c\log 1/\delta}{(1-p)\Delta_1^2}, \frac{1}{p}\left(\frac{c\log 1/\delta}{\Delta_2^2} - C_1K^\alpha - C_2\right)\right\}.$$

Assume that $\frac{c\log 1/\delta}{\Delta_2^2} \geq 2C_2$, then

$$K \geq \frac{1}{p}\left(\frac{c\log 1/\delta}{\Delta_2^2} - C_1K^\alpha - C_2\right)$$

implies

$$K \geq \frac{1}{p}\left(\frac{c\log 1/\delta}{2\Delta_2^2} - C_1K^\alpha\right) \implies 2\max\{pK, C_1K^\alpha\} \geq \frac{c\log 1/\delta}{2\Delta_2^2}.$$

The second expression is equivalent to

$$K \geq \min\left\{\frac{c\log 1/\delta}{4p\Delta_2^2}, (\frac{c\log 1/\delta}{4C_1\Delta_2^2})^{1/\alpha}\right\}$$

and we will have that the minimizer of this is $\frac{c\log 1/\delta}{4p\Delta_2^2}$ as long as $\frac{\log 1/\delta}{\Delta_2^2} \geq c' C_1^{\frac{1}{1-\alpha}} p^{\frac{-\alpha}{1-\alpha}}$. The result follows. □

E.1.3 Proof for Instance Class 7.3.1

Instance Class E.1.2 (Formal Definition of Instance Class 7.3.1)**.** Given a number of states $S \in \mathbb{N}$, consider MDP with horizon $H = 2$, S states, and $S+1$ actions. We assume we always start in state s_0 and define our transition kernel and reward function as follows:

$$P(s_i|s_0, a^\star) = \frac{2^{-i}}{1-2^{-S}}, \quad P(s_i|s_0, a_i) = 1, i \in [S]$$

$$R(s_0, a^\star) \sim \text{Bernoulli}(1), \quad R(s_0, a_i) \sim \text{Bernoulli}(0), i \in [S]$$
$$\forall i: \quad R(s_i, a^\star) \sim \text{Bernoulli}(0.9), \quad R(s_i, a_j) \sim \text{Bernoulli}(0.1), j \in [S].$$

Note that $a^\star$ is the optimal action in every state.

Proposition E.3 (Formal Statement of Proposition 7.5)**.** *For the MDP in Instance Class E.1.2 with S states, and any*

$$\epsilon \in [2^{-S}, c\min\{C_1^{-1/\alpha}(S\log 1/\delta)^{\frac{1-\alpha}{\alpha}}, C_2^{-1} S\log 1/\delta\}]$$

where c is an absolute constant, to find an ϵ-optimal policy with probability $1-\delta$ any learner executing Protocol 7.3.1 with a low-regret algorithm satisfying Definition 7.3.1 must collect at least

$$\Omega\left(\frac{S\log 1/\delta}{\epsilon}\right)$$

episodes. In contrast, on this example $\mathcal{C}^\star(M^\star) = \mathcal{O}(S^2)$ and $\epsilon^\star = 1/3$, so, for $\epsilon \in [2^{-S}, 1/3]$, with probability $1-\delta$, MOCA will terminate and output $\pi^\star$ in $\widetilde{\mathcal{O}}(C_{\text{LOT}}(1/3))$ episodes.

Randomized to deterministic policies. Assume we are given some randomized policy π which for every (s,h) choose action a with probability $\pi_h(a|s)$. Then we define the deterministic policy $\widetilde{\pi}$ given this randomized policy as

$$\widetilde{\pi}_h(s) = \arg\max_a \pi_h(a|s).$$

We will use this mapping in our lower bound.

Proof of Proposition E.3. The complexity for Algorithm 7.4 follows directly from Theorem 7.1 and Proposition E.1 and since in this example we will have $W_2(s) = 1$ for each s and so $\epsilon^\star = 1/3$. Furthermore, $\mathcal{C}^\star(M^\star) = \mathcal{O}(S^2)$. The stated complexity follows.

Complexity of Low-Regret Algorithms. Let $\Delta_{\text{KL}} := \text{KL}(\text{Bernoulli}(0.1)||\text{Bernoulli}(0.9)) \approx 1.76$ denote the KL divergence between the reward distributions of the optimal and suboptimal actions at any state for $h = 2$, and $\Delta := 0.9 - 0.1$ the suboptimality gap.

Assume that a policy π takes action $a^\star$ in s_0. Then, the total suboptimality of the policy is given by

$$\sum_{i=1}^{S} \frac{2^{-i}}{1 - 2^{-S}} \epsilon_2(s_i, \pi)$$

where $\epsilon_2(s_i, \pi)$ denotes the suboptimality of policy π in $s_i, i \in [S]$. In particular, for any i_ϵ, to guarantee our policy is ϵ-good we need

$$\frac{2^{-i_\epsilon}}{1 - 2^{-S}} \epsilon_2(s_{i_\epsilon}, \pi) \leq \epsilon.$$

By the structure of the reward in any state s_{i_ϵ}, the total suboptimality in this state will be

$$\epsilon_2(s_{i_\epsilon}, \pi) = (1 - \sum_{j=1}^{S} \pi_2(a_j|s_{i_\epsilon}))\Delta$$

It follows that if $\epsilon_2(s_{i_\epsilon}, \pi) < \Delta/4$, then we will have that $\widetilde{\pi}_2(s_{i_\epsilon}) = a^\star$, where $\widetilde{\pi}$ is the deterministic policy derived from π. Choose $i_\epsilon = \lfloor -\log_2(2\epsilon(1-2^{-S})/\Delta) - 1 \rfloor$. Then it follows that,

$$\frac{2^{-i_\epsilon}}{1 - 2^{-S}}(1 - \sum_{j=1}^{S} \pi_2(a_j|s_{i_\epsilon}))\Delta \geq 4\epsilon(1 - \sum_{j=1}^{S} \pi_2(a_j|s_{i_\epsilon}))$$

and thus, for the policy to be ϵ-optimal, we must have that $(1 - \sum_{j=1}^{S} \pi_2(a_j|s_{i_\epsilon})) \leq 1/4$. This implies that $\widetilde{\pi}_2(s_{i_\epsilon}) = a^\star$, so we have therefore derived a deterministic policy from our stochastic one that is optimal in $(s_{i_\epsilon}, 2)$. By Theorem 4 of Kaufmann et al. [2016], to identify the optimal action in state s_{i_ϵ} with probability $1 - \delta$ we must have that

$$N_2(s_{i_\epsilon}) \geq \frac{(S+1)}{\Delta_{\text{KL}}} \log \frac{1}{2.4\delta}$$

where $N_2(s_{i_\epsilon})$ is the expected number of samples collected in s_{i_ϵ} at $h = 2$. As we have deterministically derived $\widetilde{\pi}$ from π, and since $\widetilde{\pi}$ will play the optimal action in s_{i_ϵ} for any ϵ-optimal π, it follows that this lower bound on $N_2(s_{i_\epsilon})$ applies here.

If our low-regret algorithm has regret bounded as $C_1 K^\alpha + C_2$, then we must have that

$$\sum_{i=1}^{S} N_1(s_1, a_i) \leq C_1 K^\alpha + C_2$$

since every time action $a_i \neq a^\star$ is taken we will incur a loss of 1. This implies that

$$N_2(s_{i_\epsilon}) \leq C_1 K^\alpha + C_2 + \frac{2^{-i_\epsilon}}{1 - 2^{-S}} K$$

since if action $a^\star$ is taken in state s_1, we will only reach state s_{i_ϵ} with probability $\frac{2^{-i_\epsilon}}{1-2^{-S}}$. Combining these, to ensure that the optimal action is learned in s_{i_ϵ}, we will need that

$$\frac{(S+1)}{\Delta_{\mathrm{KL}}} \log \frac{1}{2.4\delta} \leq C_1 K^\alpha + C_2 + \frac{2^{-i_\epsilon}}{1 - 2^{-S}} K \leq C_1 K^\alpha + C_2 + \frac{4\epsilon}{\Delta} K$$

where the second inequality follows by our choice of i_ϵ. It follows that we need

$$\begin{aligned} K &\geq \frac{\Delta}{4\epsilon} \left(\frac{(S+1)}{\Delta_{\mathrm{KL}}} \log \frac{1}{2.4\delta} - C_1 K^\alpha - C_2 \right) \geq \frac{(S+1)\log 1/2.4\delta}{12\epsilon} - C_1 K^\alpha - C_2 \\ &\geq \frac{(S+1)\log 1/2.4\delta}{24\epsilon} - C_1 K^\alpha \end{aligned}$$

where the final inequality holds as long as $\frac{(S+1)\log 1/2.4\delta}{12\epsilon} \geq 2C_2$. This implies

$$2\max\{K, C_1 K^\alpha\} \geq \frac{(S+1)\log 1/2.4\delta}{24\epsilon}$$

which is equivalent to

$$K \geq \min\left\{ \frac{(S+1)\log 1/2.4\delta}{48\epsilon}, \left(\frac{(S+1)\log 1/2.4\delta}{48 C_1 \epsilon} \right)^{1/\alpha} \right\}.$$

For

$$\epsilon \leq \mathcal{O}\left(C_1^{-1/\alpha} (S \log 1/\delta)^{\frac{1-\alpha}{\alpha}} \right)$$

we will have that the minimizer is the first term, and

$$K \geq \Omega\left(\frac{S \log 1/\delta}{\epsilon} \right).$$

□

E.1.4 Lower Bounds on Best Policy Identification

Lemma E.1.1. *Consider MDPs M and M' with the same state space $\mathcal{S}$, actions space $\mathcal{A}$, horizon H, and initial state distribution P_0. Fix some $(s,h) \in \mathcal{S} \times [H]$, and for any $a \in \mathcal{A}$ let $\nu_h(s,a)$ denote the law of the joint distribution of (s', R) where $s' \sim P_M(\cdot|s,a)$ and $R \sim R_M(s,a)$. Define the law $\nu'_h(s,a)$ analogously with respect to M'. For any almost-sure stopping time τ with respect to $(\mathcal{F}_k)$,*

$$\sum_{s,a,h} \mathbb{E}_M[N_h^\tau(s,a)]\mathrm{KL}(\nu_h(s,a), \nu'_h(s,a)) \geq \sup_{\mathcal{E} \in \mathcal{F}_\tau} d(\mathbb{P}_M(\mathcal{E}), \mathbb{P}_{M'}(\mathcal{E}))$$

where $d(x,y) = x \log \frac{x}{y} + (1-x) \log \frac{1-x}{1-y}$ and $N_h^\tau(s,a)$ denotes the number of visits to (s,a,h) in the τ episodes.

Proof. This is the MDP analogue of Lemma 1 of Kaufmann et al. [2016] and its proof follows identically. □

Definition E.1.1. We say an algorithm is δ-correct if, for any MDP $M \in \mathfrak{M}$, we have that M terminates at some (possibly random) episode K_δ and outputs $\pi^\star$, with probability at least $1-\delta$.

E.1.4.1 Proof of Proposition 7.1

MDP Construction. Fix some $\bar{h} \in [H]$, gaps $\{\mathsf{gap}(s,a)\}_{s\in[S],a\in[A-1]} \subseteq (0,1/2)^{SA}$, and arbitrary transition kernels $\{P_h\}_{h=1}^{\bar{h}-1}$. For each s, fix a single a and set $\mathsf{gap}(s,a) = 0$. Let $M^\star$ denote the MDP with transitions $\{P_h\}_{h=1}^{\bar{h}-1}$, and for $h \geq \bar{h}$ define

$$P_h(s|s,a) = 1, \quad \forall a \in \mathcal{A}.$$

Then let the rewards be defined as follows. For all $h > \bar{h}$ and all s, choose any a' and set $R_h(s,a') = 1$, and $R_h(s,a) = 0$ for all $a \neq a'$. For $h = \bar{h}$, set

$$R_h(s,a) \sim \text{Bernoulli}(3/4 - \mathsf{gap}(s,a)).$$

For $h < \bar{h}$, let

$$\pi_h^\star(s) = \arg\max_a \sum_{s'} P_h(s'|s,a) V_{h+1}^\star(s')$$

where $V_{h+1}^\star(s')$ is the optimal value function at step $h+1$ (note that the MDP is now fully specified for $h' > h$ so this is well-defined). Then set $R_h(s, \pi_h^\star(s)) = 1$ and $R_h(s,a) = 0$ for $a \neq \pi_h^\star(s)$ (if $\pi_h^\star(s)$ is not unique, simply choose some $\pi^\star$ out of all $\pi_h^\star(s)$ arbitrarily, set $R_h(s,\pi^\star) = 1$, and all other $R_h(s,a) = 0$).

Note that we could have just as easily encoded the gaps in the transition function and set the rewards to be, for example, deterministic at level $\bar{h}$.

Lemma E.1.2. *The MDP constructed above has gaps which satisfy*

$$\Delta_{\bar{h}}(s,a) = \mathsf{gap}(s,a), \quad \forall s \in \mathcal{S}, a \in \mathcal{A}, a \neq \pi^\star_h(s)$$
$$\Delta_h(s,a) \geq 1, \quad \forall s \in \mathcal{S}, a \in \mathcal{A}, h \neq \bar{h}$$

Furthermore, for each s and $h > \bar{h}$, we have $W_h(s) = W_{\bar{h}}(s)$.

Proof. We begin with level $\bar{h}$. Since the action take at $(s, \bar{h})$ does not effect the outgoing transition, we have that, for $a \neq \pi^\star_h(s)$,

$$\Delta_{\bar{h}}(s,a) = \max_{a'} Q^\star_{\bar{h}}(s,a') - Q^\star_{\bar{h}}(s,a) = 3/4 - (3/4 - \mathsf{gap}(s,a)) = \mathsf{gap}(s,a).$$

For $h > \bar{h}$, we again have that the outgoing transition is not effected by the action taken, so it follows that the gap depends exclusively on the reward function at this state. Since the reward is set to 1 for a single action and 0 otherwise, it follows that the gaps are all 1.

For $h \leq \bar{h}$, we will have that

$$\begin{aligned}
\Delta_h(s,a) &= \max_{a'} Q^\star_h(s,a') - Q^\star_h(s,a) \\
&= 1 + \max_{a'} \sum_{s'} P_h(s'|s,a') V^\star_{h+1}(s') - \sum_{s'} P_h(s'|s,a) V^\star_{h+1}(s') \\
&\geq 1.
\end{aligned}$$

Finally, that $W_h(s) = W_{\bar{h}}(s)$ for all s and $h > \bar{h}$ follows since for all steps after $\bar{h}$, state s transitions to state s with probability 1. □

Lemma E.1.3. *On this example,*

$$\mathcal{C}^\star(M^\star) \leq \inf_\pi \max_{s,a} \frac{1}{w^\pi_{\bar{h}}(s,a)\Delta_h(s,a)^2} + \max_{s,h} \frac{SAH}{W_h(s)}.$$

Proof. By definition,

$$\mathcal{C}^\star(M^\star) = \sum_{h=1}^{H} \inf_\pi \max_{s,a} \frac{1}{w^\pi_h(s,a)\Delta_h(s,a)^2}.$$

By Lemma E.1.2, we can bound

$$\sum_{h \neq \bar{h}} \inf_\pi \max_{s,a} \frac{1}{w^\pi_h(s,a)\Delta_h(s,a)^2} \leq \sum_{h \neq \bar{h}} \inf_\pi \max_{s,a} \frac{1}{w^\pi_h(s,a)}.$$

Consider the policy π' which is the mixture over the policies π^{sh} where $w_h^{\pi^{sh}}(s) = W_h(s)$. Then,

$$\sum_{h \neq \bar{h}} \inf_{\pi} \max_{s,a} \frac{1}{w_h^{\pi}(s,a)} \leq \sum_{h \neq \bar{h}} \max_{s,a} \frac{1}{w_h^{\pi'}(s,a)} \leq \sum_{h \neq \bar{h}} \max_{s} \frac{SA}{W_h(s)} \leq \max_{s} \frac{SAH}{W_h(s)}.$$

□

Lemma E.1.4. *On the MDP constructed above, any δ-correct algorithm will have*

$$\begin{aligned}\mathbb{E}_{M^\star}[K_\delta] &\geq \inf_{\pi} \max_{s,a} \frac{1}{6 w_{\bar{h}}^{\pi}(s,a) \Delta_{\bar{h}}(s,a)^2} \cdot \log \frac{1}{2.4\delta} \\ &\gtrsim \mathcal{C}^\star(M^\star) \cdot \log \frac{1}{2.4\delta} - \max_{s,h} \frac{SAH}{W_h(s)}.\end{aligned}$$

Proof. We will apply Lemma E.1.1 on our MDP, $M^\star$, and MDP M' which is identical to $M^\star$ except that, for some (s,a), $a \neq \pi_{\bar{h}}^\star(s)$, we set $R_{\bar{h}}(s,a) \sim \text{Bernoulli}(3/4 + \alpha)$ for small α. Note that in this case we have that the optimal policy on $M^\star$ and M' differ at $(s, \bar{h})$. Since $M^\star$ and M' are identical at all points but this one, we have

$$\begin{aligned}&\sum_{s,a,h} \mathbb{E}_{M^\star}[N_h^\tau(s,a)] \text{KL}(\nu_h(s,a), \nu_h'(s,a)) \\ &\quad = \mathbb{E}_{M^\star}[N_{\bar{h}}^\tau(s,a)] \text{KL}\big(\text{Bernoulli}(3/4 - \mathsf{gap}(s,a)), \text{Bernoulli}(3/4 + \alpha)\big).\end{aligned}$$

Let $\pi^\star(M^\star)$ denote the optimal policy on $M^\star$, and $\widehat{\pi}$ denote the policy returned by our algorithm. Let $\mathcal{E} = \{\widehat{\pi} = \pi^\star(M^\star)\}$. Since we assume our algorithm is δ-correct, and since the optimal policies on $M^\star$ and M' differ, we have $\mathbb{P}_{M^\star}(\mathcal{E}) \geq 1 - \delta$ and $\mathbb{P}_{M'}(\mathcal{E}) \leq \delta$. By Kaufmann et al. [2016], we can then lower bound

$$d(\mathbb{P}_{M^\star}(\mathcal{E}), \mathbb{P}_{M'}(\mathcal{E})) \geq \log \frac{1}{2.4\delta}.$$

Thus, by Lemma E.1.1, we have shown that, for any (s,a), $a \neq \pi_{\bar{h}}^\star(s)$,

$$\mathbb{E}_{M^\star}[N_{\bar{h}}^\tau(s,a)] \geq \frac{1}{\text{KL}\big(\text{Bernoulli}(3/4 - \mathsf{gap}(s,a)), \text{Bernoulli}(3/4 + \alpha)\big)} \cdot \log \frac{1}{2.4\delta}.$$

For small α, we can bound (see e.g. Lemma 2.7 of Tsybakov [2009])

$$\text{KL}\big(\text{Bernoulli}(3/4 - \mathsf{gap}(s,a)), \text{Bernoulli}(3/4 + \alpha)\big) \leq 6(\mathsf{gap}(s,a) - \alpha)^2.$$

Taking $\alpha \to 0$, we have

$$\mathbb{E}_{M^\star}[N_h^\tau(s,a)] \geq \frac{1}{6\mathsf{gap}(s,a)^2} \cdot \log \frac{1}{2.4\delta}.$$

We can write $\mathbb{E}_{M^\star}[N_h^\tau(s,a)] = \mathbb{E}_{M^\star}[\sum_{k=1}^\tau w_h^{\pi_k}(s,a)]$ where π_k denotes the policy our algorithm played at episode k. Note that all state-visitation distributions lie in a convex set in $[0,1]^{SA}$ and that for any valid state-visitation distribution, there exists some policy that realizes it, by Proposition E.4. By Caratheodory's Theorem, it follows that there exists some set of policies Π with $|\Pi| \leq SA+1$ such that, for any π and all s,a, $w_h^\pi(s,a) = \sum_{\pi' \in \Pi} \lambda_{\pi'} w_h^{\pi'}(s,a)$, for some $\lambda \in \triangle_\Pi$. Letting λ^k denote this distribution satisfying the above inequality for π_k, it follows that

$$\begin{aligned}
\mathbb{E}_{M^\star}[\sum_{k=1}^\tau w_h^{\pi_k}(s,a)] &= \mathbb{E}_{M^\star}[\sum_{k=1}^\tau \sum_{\pi \in \Pi} \lambda_\pi^k w_h^\pi(s,a)] \\
&= \sum_{\pi \in \Pi} \mathbb{E}_{M^\star}[\sum_{k=1}^\tau \lambda_\pi^k] w_h^\pi(s,a) \\
&= \mathbb{E}_{M^\star}[\tau] \sum_{\pi \in \Pi} \frac{\mathbb{E}_{M^\star}[\sum_{k=1}^\tau \lambda_\pi^k]}{\mathbb{E}_{M^\star}[\tau]} w_h^\pi(s,a).
\end{aligned}$$

Note that $\sum_{\pi \in \Pi} \mathbb{E}_{M^\star}[\sum_{k=1}^\tau \lambda_\pi^k] = \mathbb{E}_{M^\star}[\sum_{k=1}^\tau \sum_{\pi \in \Pi} \lambda_\pi^k] = \mathbb{E}_{M^\star}[\tau]$ so it follows that $(\frac{\mathbb{E}_{M^\star}[\sum_{k=1}^\tau \lambda_\pi^k]}{\mathbb{E}_{M^\star}[\tau]})_{\pi \in \Pi} \in \triangle_\Pi$. Thus, a δ-correct algorithm must satisfy, for all s,a and some $\lambda \in \triangle_\Pi$,

$$\mathbb{E}_{M^\star}[\tau] \geq \frac{1}{6\mathsf{gap}(s,a)^2 \cdot \sum_{\pi \in \Pi} \lambda_\pi w_h^\pi(s,a)} \cdot \log \frac{1}{2.4\delta}.$$

Since the set of state visitation distributions is convex, and since for any state-visitation distribution we can find some policy realizing that distribution, for any $\lambda \in \triangle_\Pi$, it follows that there exists some π' such that, for all s,a, $\sum_{\pi \in \Pi} \lambda_\pi w_h^\pi(s,a) = w_h^{\pi'}(s,a)$. So, we need, for all s,a

$$\mathbb{E}_{M^\star}[\tau] \geq \frac{1}{6\mathsf{gap}(s,a)^2 \cdot w_h^{\pi'}(s,a)} \cdot \log \frac{1}{2.4\delta}.$$

It follows that *every* δ-correct algorithm must satisfy

$$\mathbb{E}_{M^\star}[\tau] \geq \inf_\pi \max_{s,a} \frac{1}{6\mathsf{gap}(s,a)^2 \cdot w_h^\pi(s,a)} \cdot \log \frac{1}{2.4\delta},$$

from which the first inequality follows. The second follows from Lemma E.1.3. □

E.2 Deferred Proofs from Section 7.4

In this section we provide the missing proofs from Section 7.4.

Notation. Throughout the proof, we let ϵ_{tol} denote the tolerance and δ_{tol} the confidence given as an input to MOCA, and $\epsilon = \epsilon_{\text{tol}(m)}$ and $\delta = \delta_{\text{tol}(m)}$ the tolerance and confidence given as an input to `Moca-SE` at epoch m of MOCA, respectively. For convenience, we will also define $\epsilon_0 = H$. For a single call of `Moca-SE`, we will use the following notation:

- For a given h, i, and ℓ, consider the call to `CollectSamples` on Line 13, and let $\{\mathcal{X}^{\ell}_{hij}\}_{j=1}^{\varsigma_\epsilon}$ denote the partition returned by calling `Learn2Explore` on Line 2 of `CollectSamples`. Similarly, let $\{\Pi^{\ell}_{hij}\}_{j=1}^{\varsigma_\epsilon}$ and $\{N^{\ell}_{hij}\}_{j=1}^{\varsigma_\epsilon}$ denote the policies and minimum number of samples returned by `Learn2Explore`, respectively.

- For a given h, consider the call to `CollectSamples` on Line 18, and let $\{\mathcal{X}^{\ell_\epsilon}_{hj}\}_{j=1}^{\varsigma_\epsilon}$ denote the partition returned by calling `Learn2Explore` on Line 2 of `CollectSamples`. As before, let $\{\Pi^{\ell_\epsilon+1}_{hj}\}_{j=1}^{\varsigma_\epsilon}$ and $\{N^{\ell_\epsilon+1}_{hj}\}_{j=1}^{\varsigma_\epsilon}$ denote the policies and minimum number of samples.

Good Events. We next define the good events, which we will assume hold throughout the remainder of the proof.

First, let $\mathcal{E}_{\text{exp}}$ be the event on which, for all calls to `Moca-SE` simultaneously:

- For every $h = 1, \dots, H$, $i = 1, \dots, \varsigma_\epsilon$, $\ell = 1, \dots, \ell_\epsilon$, we collect at least n^{ℓ}_{i1} samples from each $(s, a) \in \mathcal{Z}^{\ell}_{hi}$. Furthermore, $\cup_{j=1}^{\varsigma_\epsilon} \mathcal{X}^{\ell}_{hij} = \mathcal{Z}^{\ell}_{hi}$ and $\mathcal{X}^{\ell}_{hij}$ satisfy

$$\sup_{\pi} \sum_{(s,a)\in\mathcal{X}^{\ell}_{hij}} w^{\pi}_h(s, a) \le 2^{-j+1}.$$

- For every $h = 1, \dots, H$, if `Moca-SE` is run with `FinalRound` = `True`, then we collect at least $n^{\ell_\epsilon+1}_{j}$ samples from each $(s, a) \in \mathcal{X}^{\ell_\epsilon+1}_{hj}$. Furthermore, $\cup_{j=1}^{\varsigma_\epsilon} \mathcal{X}^{\ell_\epsilon+1}_{hj} = \mathcal{Z}^{\ell_\epsilon+1}_{h}$ and $\mathcal{X}^{\ell_\epsilon+1}_{hj}$ satisfies

$$\sup_{\pi} \sum_{(s,a)\in\mathcal{X}^{\ell_\epsilon+1}_{hj}} w^{\pi}_h(s, a) \le 2^{-j+1}.$$

- $\widehat{W}_h(s) \le W_h(s) \le 32\widehat{W}_h(s)$ for all $s \in \mathcal{Z}_h$.

- Following Line 7 of `Moca-SE`, $\mathcal{Z}_h$ satisfies, for all h,

$$\sup_{\pi} \max_{s\in\mathcal{Z}^c_h} w^{\pi}_h(s) \le \frac{\epsilon}{2H^2S}.$$

Next, let $\mathcal{E}_{\mathrm{est}}$ be the event on which, for all calls to `Moca-SE`,

$$|\widehat{Q}^{\widehat{\pi}}_{h,\ell}(s,a) - Q^{\widehat{\pi}}_h(s,a)| \le \sqrt{\frac{H^2\varsigma_\delta}{N_h^{hi\ell}(s,a)}}, \quad \forall (s,a) \in \mathcal{Z}^\ell_{hi}, \forall h \in [H], i \in [\varsigma_\epsilon], \ell \in [\ell_\epsilon]$$

$$|\widehat{Q}^{\widehat{\pi}}_{h,\ell_\epsilon+1}(s,a) - Q^{\widehat{\pi}}_h(s,a)| \le \sqrt{\frac{H^2\varsigma_\delta}{N_h^{h(\ell_\epsilon+1)}(s,a)}}, \quad \forall (s,a) \in \mathcal{Z}^{\ell_\epsilon}_h, \forall h \in [H]$$

where $\widehat{Q}^{\widehat{\pi}}_{h,\ell}(s,a)$ is the estimate of $Q^{\widehat{\pi}}_h(s,a)$ formed on Line 12 of `EliminateActions`, $N_h^{hi\ell}(s,a)$ is the number of samples collected from (s,a,h) at iteration (h,i,ℓ), and $\widehat{Q}^{\widehat{\pi}}_{h,\ell_\epsilon+1}(s,a)$ and $N_h^{h(\ell_\epsilon+1)}(s,a)$ are the analogous quantities for the sampling done if `FinalRound` = `True`.

E.2.1 Correctness of `Moca-SE`.

Proof of Lemma 7.4.2. We first claim that the optimal action with respect to $\widehat{\pi}$ must always be active.

Claim 5. *On the event $\mathcal{E}_{\mathrm{est}} \cap \mathcal{E}_{\mathrm{exp}}$, for any h, s, and $\ell \in [\ell_\epsilon+1]$, we will have that $\widehat{a}^\star_h(s) \in \mathcal{A}^\ell_h(s)$ where $\widehat{a}^\star_h(s) = \arg\max_a Q^{\widehat{\pi}}_h(s,a)$.*

We prove this claim in Appendix E.2.3. By construction, we will always have that $|\mathcal{A}^\ell_h(s)| \ge 1$. If $|\mathcal{A}^\ell_h(s)| = 1$, from Claim 5 it follows that $\mathcal{A}^\ell_h(s) = \{\widehat{a}^\star_h(s)\}$, and thus $\max_{a'} Q^{\widehat{\pi}}_h(s,a') - Q^{\widehat{\pi}}_h(s,a) = 0$.

Assume then that $|\mathcal{A}^\ell_h(s)| > 1$, $\ell \le \ell_\epsilon$, and $s \in \mathcal{Z}_{hi}$. The result is trivial when $\ell = 0$, since in this case $\epsilon_\ell = H$, and we will always have $\Delta_h(s,a) \le H, W_h(s) \le 1$. On the event $\mathcal{E}_{\mathrm{exp}}$, for all $i \in [\varsigma_\epsilon]$ we will collect at least $n^\ell_{i1} = 2^{18} \cdot 2^{-2i} H^2 \varsigma_\delta / \epsilon^2_\ell$ samples from (s,a) for each $a \in \mathcal{A}^\ell_h(s)$, and on $\mathcal{E}_{\mathrm{est}}$ we will then have that

$$|\widehat{Q}^{\widehat{\pi}}_{h,\ell}(s,a) - Q^{\widehat{\pi}}_h(s,a)| \le \sqrt{\frac{H^2\varsigma_\delta}{n^\ell_{i1}}} = 2^i \epsilon_\ell / 2^9.$$

Thus, for any $a \in \mathcal{A}^\ell_h(s)$, we have

$$\begin{aligned}\max_{a' \in \mathcal{A}^\ell_h(s)} \widehat{Q}^{\widehat{\pi}}_{h,\ell}(s,a') - \widehat{Q}^{\widehat{\pi}}_{h,\ell}(s,a) &\ge \max_{a' \in \mathcal{A}^\ell_h(s)} Q^{\widehat{\pi}}_h(s,a') - Q^{\widehat{\pi}}_h(s,a) - 2 \cdot 2^i \epsilon_\ell / 2^9 \\ &= \max_{a'} Q^{\widehat{\pi}}_h(s,a') - Q^{\widehat{\pi}}_h(s,a) - 2 \cdot 2^i \epsilon_\ell / 2^9\end{aligned}$$

where the equality follows since $\widehat{a}^\star_h(s) \in \mathcal{A}^\ell_h(s)$. It follows that if

$$\Delta^{\widehat{\pi}}_h(s,a) = \max_{a'} Q^{\widehat{\pi}}_h(s,a') - Q^{\widehat{\pi}}_h(s,a) \ge 4 \cdot 2^i \epsilon_\ell / 2^9$$

then

$$\max_{a' \in \mathcal{A}_h^\ell(s)} \widehat{Q}_{h,\ell}^{\widehat{\pi}}(s,a') - \widehat{Q}_{h,\ell}^{\widehat{\pi}}(s,a) \geq 2 \cdot 2^i \epsilon_\ell / 2^9.$$

so the exit condition on Line 16 for `EliminateActions` is met for our choice of $\gamma_{ij}^\ell = 2^i \epsilon_\ell / 2^8$ (note that in this case, since γ_{ij}^ℓ is the same for all ℓ, Line 15 has no effect), and therefore $a \notin \mathcal{A}_h^{\ell+1}(s)$. Thus, any $a \in \mathcal{A}_h^{\ell+1}(s)$ must satisfy

$$\Delta_h^{\widehat{\pi}}(s,a) \leq 2^i \epsilon_\ell / 2^7.$$

By construction, we will have that $\widehat{W}_h(s) \in [2^{-i}, 2^{-i+1}]$ and on $\mathcal{E}_{\text{exp}}$, $\widehat{W}_h(s) \leq W_h(s) \leq 32\widehat{W}_h(s)$. Thus, we can upper bound

$$\Delta_h^{\widehat{\pi}}(s,a) \leq 2^i \epsilon_\ell / 2^7 \leq \frac{2\epsilon_\ell}{\widehat{W}_h(s) 2^7} \leq \frac{32 \cdot 2\epsilon_\ell}{W_h(s) 2^7} = \frac{\epsilon_\ell}{2 W_h(s)}.$$

Finally, the following claim, proved in Appendix E.2.3, allows us to relate $\Delta_h^{\widehat{\pi}}(s,a)$ to $\Delta_h(s,a)$:

Claim 6. *On the event $\mathcal{E}_{\text{est}} \cap \mathcal{E}_{\text{exp}}$, for any (s,a,h), we will have $|\Delta_h^{\widehat{\pi}}(s,a) - \Delta_h(s,a)| \leq \epsilon / W_h(s)$.*

Applying Claim 6, we can lower bound $\Delta_h^{\widehat{\pi}}(s,a) \geq \Delta_h(s,a) - \epsilon / W_h(s) \geq \Delta_h(s,a) - \epsilon_\ell / W_h(s)$. Rearranging this gives the second conclusion.

The argument for the third conclusion is similar to the preceding argument. However, we now have the extra subtlety that for $a \neq a'$ with $a, a' \in \mathcal{A}_h^{\ell_\epsilon}(s)$, we may collect a different number of samples from (s,a) and (s,a') since it's possible that $(s,a) \in \mathcal{X}_{hj}^{\ell_\epsilon+1}$ and $(s,a') \in \mathcal{X}_{hj'}^{\ell_\epsilon+1}$ for $j \neq j'$. Denote

$$j(s) = \arg\max_j j \quad \text{s.t.} \quad \exists a, (s,a) \in \mathcal{X}_{hj}^{\ell_\epsilon+1}.$$

Note that, on $\mathcal{E}_{\text{exp}}$, we are guaranteed that there exists some $j \in [\varsigma_\epsilon]$ such that $(s,a) \in \mathcal{X}_{hj}^{\ell_\epsilon+1}$ so $j(s)$ is always well-defined. We can repeat the above argument, but now we can only guarantee that

$$|\widehat{Q}_{h,\ell_\epsilon+1}^{\widehat{\pi}}(s,a) - Q_h^{\widehat{\pi}}(s,a)| \leq \sqrt{\frac{H^2 \varsigma_\delta}{n_{j(s)}^{\ell_\epsilon+1}}} = \frac{\epsilon}{8H\varsigma_\epsilon 2^{-j(s)+1}}.$$

since we can only guarantee we collect $n_{j(s)}^{\ell_\epsilon+1}$ samples from each $(s,a), a \in \mathcal{A}_h^{\ell_\epsilon}(s)$. It again

follows that if

$$\Delta_h^{\widehat{\pi}}(s,a) \geq 4 \cdot \frac{\epsilon}{8H\varsigma_\epsilon 2^{-j(s)+1}}$$

then

$$\max_{a' \in \mathcal{A}_h^{\ell_\epsilon}(s)} \widehat{Q}_{h,\ell_\epsilon+1}^{\widehat{\pi}}(s,a') - \widehat{Q}_{h,\ell_\epsilon+1}^{\widehat{\pi}}(s,a) \geq 2 \cdot \frac{\epsilon}{8H\varsigma_\epsilon 2^{-j(s)+1}}.$$

As this is precisely the elimination criteria used in `EliminateActions`, it follows that a will be eliminated. Thus, all $a \in \mathcal{A}_h^{\ell_\epsilon+1}(s)$ must satisfy

$$\Delta_h^{\widehat{\pi}}(s,a) \leq 4 \cdot \frac{\epsilon}{8H\varsigma_\epsilon 2^{-j(s)+1}}$$

which gives the third conclusion.

□

Proof of Lemma 7.4.5. Proposition 7.6 gives that, if $\widehat{\pi}$ satisfies $\max_a Q_h^{\widehat{\pi}}(s,a) - Q_h^{\widehat{\pi}}(s,\widehat{\pi}_h(s)) \leq \epsilon_h(s)$ for all h and s, then $\widehat{\pi}$ is at most

$$\sum_{h=1}^{H} \sup_{\pi} \sum_{s} w_h^{\pi}(s)\epsilon_h(s) \tag{E.2.1}$$

suboptimal. When running Algorithm 7.3, for a particular h every state s can be classified in one of three ways:

- $s \notin \mathcal{Z}_h$: In this case, on $\mathcal{E}_{\text{exp}}$ we will have $\sup_\pi w_h^\pi(s) \leq \epsilon/(2H^2S)$ and $\epsilon_h(s) \leq H$.
- $s \in \mathcal{Z}_h$ and $|\mathcal{A}_h^{\ell_\epsilon+1}(s)| = 1$: In this case, by Lemma 7.4.2, since $\widehat{\pi}$ only takes actions that are in $\mathcal{A}_h^{\ell_\epsilon+1}(s)$, we will have $\epsilon_h(s) = \max_a Q_h^{\widehat{\pi}}(s,a) - Q_h^{\widehat{\pi}}(s,\widehat{\pi}_h(s)) = 0$.
- $s \in \mathcal{Z}_h$, $|\mathcal{A}_h^{\ell_\epsilon+1}(s)| > 1$: Then we can apply Lemma 7.4.2 to get

$$\epsilon_h(s) = \max_{a'} Q_h^{\widehat{\pi}}(s,a') - Q_h^{\widehat{\pi}}(s,\widehat{\pi}_h(s)) \leq \frac{\epsilon}{2H\varsigma_\epsilon \cdot 2^{-j(s)+1}}$$

Let $\widetilde{\mathcal{X}}_j = \{s \ : \ j(s) = j\}$ and note that $\{s \in \mathcal{Z}_h \ : \ |\mathcal{A}_h^{\ell_\epsilon+1}(s)| > 1\} \subseteq \cup_{j=1}^{\varsigma_\epsilon} \widetilde{\mathcal{X}}_j$ since, on $\mathcal{E}_{\text{exp}}$, for every s satisfying $s \in \mathcal{Z}_h, |\mathcal{A}_h^{\ell_\epsilon+1}(s)| > 1$, we will have $(s,a) \in \mathcal{Z}_h^{\ell_\epsilon+1}$ for some a, so we must have that $(s,a) \in \mathcal{X}_{hj}^{\ell_\epsilon+1}$ for some $j \in [\varsigma_\epsilon]$. Furthermore, by definition of $j(s)$, if $s \in \widetilde{\mathcal{X}}_j$,

then $(s,a) \in \mathcal{X}_{hj}^{\ell_\epsilon+1}$ for some a. Then, plugging all of this into (E.2.1), on $\mathcal{E}_{\text{exp}}$,

$$\begin{aligned}
\sum_{h=1}^{H} \sup_{\pi} \sum_{s} w_h^\pi(s)\epsilon_h(s) &\le \sum_{h=1}^{H} \sup_{\pi} \sum_{j=1}^{\varsigma_\epsilon} \sum_{s\in\widetilde{\mathcal{X}}_j} w_h^\pi(s)\epsilon_h(s) + H\sum_{h=1}^{H} \sup_{\pi} \sum_{s\in\mathcal{Z}_h^c} w_h^\pi(s) \\
&\le \frac{\epsilon}{2H\varsigma_\epsilon} \sum_{h=1}^{H} \sup_{\pi} \sum_{j=1}^{\varsigma_\epsilon} \sum_{s\in\widetilde{\mathcal{X}}_j} w_h^\pi(s) 2^{j(s)-1} + H\sum_{h=1}^{H} \sup_{\pi} \sum_{s\in\mathcal{Z}_h^c} w_h^\pi(s) \\
&\overset{(a)}{\le} \frac{\epsilon}{2H\varsigma_\epsilon} \sum_{h=1}^{H} \sum_{j=1}^{\varsigma_\epsilon} 2^{j-1} \sup_{\pi} \sum_{(s,a)\in\mathcal{X}_{hj}^{\ell_\epsilon+1}} w_h^\pi(s,a) + H\sum_{h=1}^{H} \sup_{\pi} \sum_{s\in\mathcal{Z}_h^c} w_h^\pi(s) \\
&\le \frac{\epsilon}{2H\varsigma_\epsilon} \sum_{h=1}^{H} \sum_{j=1}^{\varsigma_\epsilon} 2^{j-1}2^{-j+1} + H\sum_{h=1}^{H} \sum_{s\in\mathcal{Z}_h^c} \frac{\epsilon}{2H^2S} \\
&\le \epsilon
\end{aligned}$$

where (a) holds since for $s \in \widetilde{\mathcal{X}}_j$, $j(s) = j$, and since we can always choose π so that $\pi_h(s) = a$ so $w_h^\pi(s,a) = w_h^\pi(s)$. It follows that $\widehat{\pi}$ is at most ϵ-suboptimal. □

E.2.2 Sample Complexity

Lemma E.2.1. `CollectSamples`$(\mathcal{Z}_{hi}^\ell, \{n_{ij}^\ell\}_{j=1}^{\varsigma_\epsilon}, h, \widehat{\pi}, \frac{\delta}{H\varsigma_\epsilon\ell_\epsilon}, \frac{\epsilon_{\text{exp}}}{32})$ *terminates in at most*

$$\frac{cH^2\varsigma_\delta\varsigma_\epsilon}{\epsilon_\ell^2} \sum_{j=1}^{\varsigma_\epsilon} 2^j \sum_{(s,a)\in\mathcal{X}_{hij}^\ell} W_h(s)^2 + \frac{\text{poly}(S,A,H,\log 1/\delta,\log 1/\epsilon)}{\epsilon}$$

episodes and `CollectSamples`$(\mathcal{Z}_h^{\ell_\epsilon+1}, \{n_j^{\ell_\epsilon+1}\}_{j=1}^{\varsigma_\epsilon}, h, \widehat{\pi}, \frac{\delta}{H}, \frac{\epsilon_{\text{exp}}}{32})$ *terminates in at most*

$$\frac{cH^4\varsigma_\delta\varsigma_\epsilon^2}{\epsilon^2}|\mathcal{Z}_h^{\ell_\epsilon+1}| + \frac{\text{poly}(S,A,H,\log 1/\delta,\log 1/\epsilon)}{\epsilon}$$

episodes.

Proof. Recall that $\epsilon_{\text{exp}} = \frac{\epsilon}{2H^2S}$. The complexity of `CollectSamples`$(\mathcal{Z}_{hi}^\ell, n_i^\ell, h, \widehat{\pi}, \frac{\delta}{H\varsigma_\epsilon\ell_\epsilon}, \frac{\epsilon}{64H^2S})$ can be bounded by the sum of the complexity of calling `Learn2Explore` to learn a set of exploration policies, and the complexity of playing these policies to collect samples. By Theorem E.1, we can bound the complexity of calling `Learn2Explore` by

$$C_K(\tfrac{\delta}{H\varsigma_\epsilon\ell_\epsilon}, \delta_{\text{samp}}, \varsigma_\epsilon)\frac{256H^2S}{\epsilon}$$

where $\delta_{\text{samp}} = \frac{\delta}{H\varsigma_\epsilon \ell_\epsilon} \cdot \frac{1}{\varsigma_\epsilon \max_j n_{ij}^\ell} \leq \frac{\delta \epsilon_\ell^2}{2^{17} H^3 \varsigma_\delta \varsigma_\epsilon^2 \ell_\epsilon}$. As shown in Appendix E.3, $C_K(\frac{\delta}{H\varsigma_\epsilon \ell_\epsilon}, \delta_{\text{samp}}, \varsigma_\epsilon)$ is $\text{poly}(S, A, H, \log 1/\epsilon, \log 1/\delta)$, so this entire term is $\frac{\text{poly}(S,A,H,\log 1/\epsilon, \log 1/\delta)}{\epsilon}$.

Since rerunning the policies in Π_{hij}^ℓ yields at least $N_{hij}^\ell/2$ samples from each (s,a) in X_{hij}^ℓ, if we desire n_{ij}^ℓ samples from each (s,a), the complexity of running the policies returned by `Learn2Explore` in order to collect the desired samples is clearly given by

$$\sum_{j=1}^{\varsigma_\epsilon} |\Pi_{hij}^\ell| \lceil 2n_{ij}^\ell / N_{hij}^\ell \rceil.$$

By the construction of Π_{hij}^ℓ and definition of N_{hij}^ℓ given in `Learn2Explore`, we have that

$$|\Pi_{hij}^\ell| = 2^j C_K(\tfrac{\delta}{H\varsigma_\epsilon \ell_\epsilon}, \delta_{\text{samp}}, j), \quad N_{hij}^\ell = \frac{|\Pi_{hij}^\ell|}{4 M_{hij}^\ell 2^j}.$$

where $M_{hij}^\ell = \sum_{j'=j}^{\varsigma_\epsilon+1} |\mathcal{X}_{hij'}^\ell|$ and $\mathcal{X}_{hi(\varsigma_\epsilon+1)}^\ell = \mathcal{Z}_{hi}^\ell \setminus \cup_{j=1}^{\varsigma_\epsilon} \mathcal{X}_{hij}^\ell$. As we are on $\mathcal{E}_{\text{exp}}$, $\mathcal{Z}_{hi}^\ell = \cup_{j=1}^{\varsigma_\epsilon} \mathcal{X}_{hij}^\ell$, so $|\mathcal{X}_{hi(\varsigma_\epsilon+1)}^\ell| = 0$. It follows that the complexity can be upper bounded as

$$\begin{aligned}
\sum_{j=1}^{\varsigma_\epsilon} |\Pi_{hij}^\ell| \lceil 2n_{ij}^\ell / N_{hij}^\ell \rceil &\leq 8 \sum_{j=1}^{\varsigma_\epsilon} 2^j M_{hij}^\ell n_{ij}^\ell + \sum_{j=1}^{\varsigma_\epsilon} 2^j C_K(\tfrac{\delta}{H\varsigma_\epsilon \ell_\epsilon}, \delta_{\text{samp}}, j) \\
&\leq 8 \sum_{j=1}^{\varsigma_\epsilon} 2^j M_{hij}^\ell n_{ij}^\ell + 2^{\varsigma_\epsilon+1} C_K(\tfrac{\delta}{H\varsigma_\epsilon \ell_\epsilon}, \delta_{\text{samp}}, \varsigma_\epsilon) \\
&= 8 \frac{2^{17} H^2 \varsigma_\delta}{2^{2i} \epsilon_\ell^2} \sum_{j=1}^{\varsigma_\epsilon} 2^j M_{hij}^\ell + 2^{\varsigma_\epsilon+1} C_K(\tfrac{\delta}{H\varsigma_\epsilon \ell_\epsilon}, \delta_{\text{samp}}, \varsigma_\epsilon)
\end{aligned}$$

The term $2^{\varsigma_\epsilon+1} C_K(\frac{\delta}{H\varsigma_\epsilon \ell_\epsilon}, \delta_{\text{samp}}, \varsigma_\epsilon)$ is $\frac{\text{poly}(S,A,H,\log 1/\epsilon, \log 1/\delta)}{\epsilon}$ by definition of ς_ϵ and C_K. Furthermore,

$$\sum_{j=1}^{\varsigma_\epsilon} 2^j M_{hij}^\ell = \sum_{j=1}^{\varsigma_\epsilon} 2^j \sum_{j'=j}^{\varsigma_\epsilon} |\mathcal{X}_{hij'}^\ell| \leq \varsigma_\epsilon \sum_{j=1}^{\varsigma_\epsilon} 2^j |\mathcal{X}_{hij}^\ell|.$$

We can therefore bound

$$\frac{2^{17} H^2 \varsigma_\delta}{2^{2i} \epsilon_\ell^2} \sum_{j=1}^{\varsigma_\epsilon} 2^j M_{hij}^\ell \leq \frac{c H^2 \varsigma_\delta \varsigma_\epsilon}{\epsilon_\ell^2} \sum_{j=1}^{\varsigma_\epsilon} 2^{j-2i} |\mathcal{X}_{hij}^\ell|.$$

Finally, using that on $\mathcal{E}_{\text{exp}}$ $W_h(s) \geq \widehat{W}_h(s)$, and that all $(s,a) \in \mathcal{X}_{hij}^\ell$ have a value of $\widehat{W}_h(s)$ within a factor of 2 of every other, we can upper bound $2^{-i} \leq 4W_h(s)$ for any $(s,a) \in \mathcal{X}_{hij}^\ell$. This completes the proof of the first claim.

The second claim follows similarly. By the same argument as above, we can upper bound the sample complexity of calling `CollectSamples`$(\mathcal{Z}_h^{\ell_\epsilon+1}, \{n_j^{\ell_\epsilon+1}\}_{j=1}^{\varsigma_\epsilon}, h, \widehat{\pi}, \frac{\delta}{H}, \frac{\epsilon_{\exp}}{32})$ as

$$\begin{aligned}
&\sum_{j=1}^{\varsigma_\epsilon} |\Pi_{hj}^{\ell_\epsilon+1}| \lceil 2n_j^{\ell_\epsilon+1}/N_{hj}^{\ell_\epsilon+1} \rceil + \frac{\mathrm{poly}(S,A,H,\log 1/\epsilon, \log 1/\delta)}{\epsilon} \\
&\qquad \le 8 \sum_{j=1}^{\varsigma_\epsilon} 2^j M_{hj}^{\ell_\epsilon+1} n_j^{\ell_\epsilon+1} + \frac{\mathrm{poly}(S,A,H,\log 1/\epsilon, \log 1/\delta)}{\epsilon} \\
&\qquad \overset{(a)}{=} \tfrac{cH^4 \varsigma_\delta \varsigma_\epsilon^2}{\epsilon^2} \sum_{j=1}^{\varsigma_\epsilon} 2^{-j} M_{hj}^{\ell_\epsilon+1} + \frac{\mathrm{poly}(S,A,H,\log 1/\epsilon, \log 1/\delta)}{\epsilon} \\
&\qquad \overset{(b)}{\le} \tfrac{cH^4 \varsigma_\delta \varsigma_\epsilon^2}{\epsilon^2} |\mathcal{Z}_h^{\ell_\epsilon+1}| + \frac{\mathrm{poly}(S,A,H,\log 1/\epsilon, \log 1/\delta)}{\epsilon}
\end{aligned}$$

where (a) follows by our setting of $n_j^{\ell_\epsilon+1}$ and (b) follows since $M_{hj}^{\ell_\epsilon+1} \le |\mathcal{Z}_h^{\ell_\epsilon+1}|$. The second conclusion follows. □

Proof of Lemma 7.4.3. With `FinalRound` = `False` , the complexity of `Moca-SE` is given by the complexity incurred calling `Learn2Explore` on Line 4 and calling `CollectSamples` on Line 13. By Theorem E.1 and since we call `Learn2Explore` at most SH times, we can bound the complexity of calling `Learn2Explore` by

$$\frac{\mathrm{poly}(S,A,H,\log 1/\epsilon, \log 1/\delta)}{\epsilon}.$$

Next, we turn to upper bounding the sample complexity of `Learn2Explore`. We can lower bound

$$|\mathcal{X}_{hij}^{\ell}| \sup_\pi \min_{(s,a)\in\mathcal{X}_{hij}^{\ell}} w_h^\pi(s,a) \le \sup_\pi \sum_{(s,a)\in\mathcal{X}_{hij}^{\ell}} w_h^\pi(s,a).$$

so, on $\mathcal{E}_{\exp}$, $2^j \le 2(|\mathcal{X}_{hij}^{\ell}| \sup_\pi \min_{(s,a)\in\mathcal{X}_{hij}^{\ell}} w_h^\pi(s,a))^{-1}$. Plugging this into the bound given in Lemma E.2.1, we can bound the leading term in the sample complexity of a single call to `CollectSamples` as

$$\begin{aligned}
\frac{cH^2\varsigma_\delta\varsigma_\epsilon}{\epsilon_\ell^2} \sum_{j=1}^{\varsigma_\epsilon} 2^j \sum_{(s,a)\in\mathcal{X}_{hij}^{\ell}} W_h(s)^2 &\le \frac{cH^2\varsigma_\delta\varsigma_\epsilon}{\epsilon_\ell^2} \sum_{j=1}^{\varsigma_\epsilon} \frac{1}{|\mathcal{X}_{hij}^{\ell}| \sup_\pi \min_{(s,a)\in\mathcal{X}_{hij}^{\ell}} w_h^\pi(s,a)} \sum_{(s,a)\in\mathcal{X}_{hij}^{\ell}} W_h(s)^2 \\
&\overset{(a)}{\le} \frac{cH^2\varsigma_\delta\varsigma_\epsilon}{\epsilon_\ell^2} \sum_{j=1}^{\varsigma_\epsilon} \inf_\pi \max_{(s,a)\in\mathcal{X}_{hij}^{\ell}} \frac{W_h(s)^2}{w_h^\pi(s,a)} \\
&\le \frac{cH^2\varsigma_\delta\varsigma_\epsilon^2}{\epsilon_\ell^2} \inf_\pi \max_{j\in\{1,\dots,\varsigma_\epsilon\}} \max_{(s,a)\in\mathcal{X}_{hij}^{\ell}} \frac{W_h(s)^2}{w_h^\pi(s,a)}
\end{aligned}$$

where (a) holds since all $s \in \mathcal{X}_{hij}^{\ell}$ have values of $\widehat{W}_h(s)$ within a constant factor of each other, and since on $\mathcal{E}_{\text{exp}}$ $\widehat{W}_h(s) \leq W_h(s) \leq 32\widehat{W}_h(s)$, which together imply that

$$\max_{s \in \mathcal{X}_{hij}^{\ell}} W_h(s) \leq c \min_{s \in \mathcal{X}_{hij}^{\ell}} W_h(s).$$

If $(s,a) \in \mathcal{X}_{hij}^{\ell}$, then we must have that $(s,a) \in \mathcal{Z}_{hi}^{\ell}$ since $\mathcal{X}_{hij}^{\ell} \subseteq \mathcal{Z}_{hi}^{\ell}$, and, by the definition of $\mathcal{Z}_{hi}^{\ell}$, $a \in \mathcal{A}_h^{\ell-1}(s)$ and $|\mathcal{A}_h^{\ell-1}(s)| > 1$. Lemma 7.4.2 gives that any $a \in \mathcal{A}_h^{\ell-1}(s)$ satisfies $\Delta_h(s,a) \leq 3\epsilon_{\ell-1}/(2W_h(s))$. Since $|\mathcal{A}_h^{\ell-1}(s)| > 1$, it follows there exists a, a', $a \neq a'$, such that

$$\Delta_h(s,a) \leq 3\epsilon_{\ell-1}/(2W_h(s)) \quad \text{and} \quad \Delta_h(s,a') \leq 3\epsilon_{\ell-1}/(2W_h(s)).$$

Thus, if $(s,a) \in \mathcal{X}_{hij}^{\ell}$, $\frac{1}{4\epsilon_\ell^2} = \frac{1}{\epsilon_{\ell-1}^2} \leq \frac{9}{4W_h(s)^2\Delta_h(s,a)^2}$ and $\frac{1}{4\epsilon_\ell^2} = \frac{1}{\epsilon_{\ell-1}^2} \leq \frac{9}{4W_h(s)^2\Delta_h(s,a')^2}$, which implies $\frac{1}{4\epsilon_\ell^2} \leq \frac{9}{4W_h(s)^2 \max\{\Delta_h(s,a)^2, \Delta_h(s,a')^2\}}$. Note that $\max\{\Delta_h(s,a)^2, \Delta_h(s,a')^2\} \geq \widetilde{\Delta}_h(s,a)^2$ since if $\Delta_h(s,a) = 0$, we will have $\max\{\Delta_h(s,a)^2, \Delta_h(s,a')^2\} = \Delta_h(s,a')^2$, so either a is the unique optimal action at (s,h), in which case $\Delta_h(s,a') \geq \Delta_{\min}(s,h) = \widetilde{\Delta}_h(s,a)$, or there are multiple optimal actions, in which case $\Delta_h(s,a') \geq 0 = \widetilde{\Delta}_h(s,a)$. Thus,

$$\begin{aligned}
&\frac{cH^2\varsigma_\delta\varsigma_\epsilon^2}{\epsilon_\ell^2} \inf_\pi \max_{j \in \{1,\dots,\varsigma_\epsilon\}} \max_{(s,a) \in \mathcal{X}_{hij}^{\ell}} \frac{W_h(s)^2}{w_h^\pi(s,a)} \\
&\qquad \leq cH^2\varsigma_\delta\varsigma_\epsilon^2 \inf_\pi \max_{j \in \{1,\dots,\varsigma_\epsilon\}} \max_{(s,a) \in \mathcal{X}_{hij}^{\ell}} \min\left\{ \frac{1}{w_h^\pi(s,a)\widetilde{\Delta}_h(s,a)^2}, \frac{W_h(s)^2}{w_h^\pi(s,a)\epsilon_\ell^2} \right\} \\
&\qquad \leq cH^2\varsigma_\delta\varsigma_\epsilon^2 \inf_\pi \max_{(s,a) \in \mathcal{Z}_{hi}^{\ell}} \min\left\{ \frac{1}{w_h^\pi(s,a)\widetilde{\Delta}_h(s,a)^2}, \frac{W_h(s)^2}{w_h^\pi(s,a)\epsilon_\ell^2} \right\}.
\end{aligned}$$

Summing over ℓ and using that for all $(s,a) \in \mathcal{Z}_{hi}^{\ell}$, $s \in \mathcal{Z}_{hi}$, proves the first conclusion. Summing over i, and h gives

$$\begin{aligned}
&\sum_{h=1}^{H} \sum_{i=1}^{\varsigma_\epsilon} cH^2\varsigma_\delta\varsigma_\epsilon^2\ell_\epsilon \inf_\pi \max_{s \in \mathcal{Z}_{hi}, a} \min\left\{ \frac{1}{w_h^\pi(s,a)\widetilde{\Delta}_h(s,a)^2}, \frac{W_h(s)^2}{w_h^\pi(s,a)\epsilon^2} \right\} \\
&\qquad \leq cH^2\varsigma_\delta\varsigma_\epsilon^3\ell_\epsilon \sum_{h=1}^{H} \inf_\pi \max_{s,a} \min\left\{ \frac{1}{w_h^\pi(s,a)\widetilde{\Delta}_h(s,a)^2}, \frac{W_h(s)^2}{w_h^\pi(s,a)\epsilon^2} \right\}.
\end{aligned}$$

This proves the result. □

Proof of Lemma 7.4.4. The only additional samples taken when running `Moca-SE` with `FinalRound` = `True` as compared to running it with `FinalRound` = `False` is incurred by calling `CollectSamples` on Line 18 of `Moca-SE`. Thus, the total complexity can be bounded by adding the complexity bound from Lemma 7.4.3 to this additional cost.

In particular, by Lemma E.2.1, this additional call of `CollectSamples` will require at most

$$\frac{cH^4\varsigma_\delta\varsigma_\epsilon^2}{\epsilon^2}|\mathcal{Z}_h^{\ell_\epsilon+1}| + \frac{\mathrm{poly}(S, A, H, \log 1/\delta, \log 1/\epsilon)}{\epsilon}$$

episodes to terminate, from which the first conclusion follows. We can repeat the argument from the proof of Lemma 7.4.3 to get that $\mathcal{Z}_h^{\ell_\epsilon+1} \subseteq \mathcal{W}_h^{\ell_\epsilon+1}$, where we define $\mathcal{W}_h^{\ell_\epsilon} := \{(s,a) : s \in \mathcal{Z}_h, \exists a' \neq a, \max\{\Delta_h(s,a), \Delta_h(s,a')\} \leq 3\epsilon_{\ell_\epsilon-1}/(2W_h(s))\}$. However, note that $\epsilon_{\ell_\epsilon-1} \leq 2\epsilon$, and the condition $\exists a' \neq a, \max\{\Delta_h(s,a), \Delta_h(s,a')\} \leq 3\epsilon_{\ell_\epsilon-1}/(2W_h(s))$ implies $\widetilde{\Delta}_h(s,a) \leq 3\epsilon_{\ell_\epsilon-1}/(2W_h(s))$. It follows that

$$\mathcal{W}_h^{\ell_\epsilon+1} \subseteq \left\{(s,a) \;:\; \widetilde{\Delta}_h(s,a) \leq 3\epsilon/W_h(s)\right\} =: \mathrm{OPT}(\epsilon, h)$$

Summing over h gives the result.

□

E.2.3 Proofs of Additional Lemmas and Claims

Proof of Lemma 7.4.1. $\mathcal{E}_{\mathrm{est}}$ **holds.** That $\mathcal{E}_{\mathrm{est}}$ holds with probability $1 - \delta_{\mathrm{tol}}/2$ follows directly from Hoeffding's inequality and a union bound, since $\mathring{Q}_h^{\widehat{\pi},t}(s_h^t, a_h^t) \leq H$ almost surely. In particular, note that for any given call to `Moca-SE`, we will form at most $SAH\varsigma_\epsilon(\ell_\epsilon + 1)$ estimates of $Q_h^{\widehat{\pi}}(s,a)$. By Hoeffding's inequality and our choice of ς_δ, that each of these estimates concentrates as given on $\mathcal{E}_{\mathrm{est}}$ then holds with probability

$$1 - SAH\varsigma_\epsilon(\ell_\epsilon + 1) \cdot \frac{\delta}{SAH\varsigma_\epsilon(\ell_\epsilon + 1)} = 1 - \delta.$$

With our choice of $\delta_{\mathrm{tol}(m)} = \frac{\delta_{\mathrm{tol}}}{36m^2}$, union bounding over this holding for each call to `Moca-SE`, we then have that $\mathcal{E}_{\mathrm{est}}$ holds with probability at least

$$1 - \sum_{m=1}^{\lceil \log H/\epsilon \rceil} \frac{\delta_{\mathrm{tol}}}{36m^2} \geq 1 - \frac{\delta_{\mathrm{tol}}}{2},$$

which is the desired result.

$\mathcal{E}_{\mathrm{exp}}$ **holds.** We show that the desired events hold for a single call of `Moca-SE`, then union bound over all calls to `Moca-SE` to get the final result. Let $\mathcal{E}_{\mathrm{exp}}^m$ denote the event on which all conditions of $\mathcal{E}_{\mathrm{exp}}$ hold for the mth call to `Moca-SE`.

Assume that we run `Moca-SE` with tolerance $\epsilon_{\mathrm{tol}(m)}$ and confidence $\delta_{\mathrm{tol}(m)}$. Let $\mathcal{E}_{\mathsf{L2E}}^{sh}$ denote the success event of calling `Learn2Explore` on Line 4, $\mathcal{E}_{\mathsf{L2E}}^{hi\ell}$ denote the success event of calling `Learn2Explore` in the call to `CollectSamples` at iteration (h, i, ℓ) on Line 13, and $\mathcal{E}_{\mathsf{L2E}}^{h}$ the success event of calling `Learn2Explore` in the call to `CollectSamples` on Line 18.

By Theorem E.1 and the confidence with which we call `Learn2Explore`, we have that $\mathbb{P}[\mathcal{E}^{sh}_{\mathsf{L2E}}] \geq 1 - \delta_{\mathrm{tol}(m)}/SH$, $\mathbb{P}[\mathcal{E}^{hi\ell}_{\mathsf{L2E}}] \geq 1 - \delta_{\mathrm{tol}(m)}/(H\varsigma_\epsilon \ell_\epsilon)$, and $\mathbb{P}[\mathcal{E}^{h}_{\mathsf{L2E}}] \geq 1 - \delta_{\mathrm{tol}(m)}/H$. Union bounding over these events, and using that there are at most $H\varsigma_\epsilon\ell_\epsilon$ indices (h, i, ℓ), we get that the event

$$(\cap_{s,h}\mathcal{E}^{sh}_{\mathsf{L2E}}) \cap (\cap_{h=1}^{H} \cap_{i=1}^{\varsigma_\epsilon} \cap_{\ell=1}^{\ell_\epsilon} \mathcal{E}^{hi\ell}_{\mathsf{L2E}}) \cap (\cap_{h=1}^{H}\mathcal{E}^{h}_{\mathsf{L2E}})$$

holds with probability at least $1 - 3\delta_{\mathrm{tol}(m)}$.

That

$$\sup_{\pi} \sum_{(s,a)\in\mathcal{X}^{\ell}_{hij}} w^{\pi}_h(s,a) \leq 2^{-j+1}$$

for $j \in [\varsigma_\epsilon]$, is a direct consequence of $\mathcal{E}^{hi\ell}_{\mathsf{L2E}}$ holding, and similarly that

$$\sup_{\pi} \sum_{(s,a)\in\mathcal{X}^{\ell_\epsilon+1}_{hj}} w^{\pi}_h(s,a) \leq 2^{-j+1}$$

holds for $j \in [\varsigma_\epsilon]$, is a direct consequence of $\mathcal{E}^{h}_{\mathsf{L2E}}$. In addition, that

$$\sup_{\pi} \max_{s\in\mathcal{Z}^{c}_{h}} w^{\pi}_h(s) \leq \frac{\epsilon}{2H^2S}$$

holds for all h is immediate on $\cap_{s,h}\mathcal{E}^{sh}_{\mathsf{L2E}}$.

On the event $\mathcal{E}^{hi\ell}_{\mathsf{L2E}}$, if we run the policies returned by `Learn2Explore` for some $j \in \{1, \ldots, \varsigma_\epsilon\}$, Π^{ℓ}_{hij}, Theorem E.1 and our choice of δ_{samp} gives that we will collect at least $\frac{1}{2}N^{\ell}_{hij}$ samples from each $(s,a) \in \mathcal{X}^{\ell}_{hij}$ with probability at least $1 - \delta_{\mathrm{tol}(m)}/(H\varsigma^2_\epsilon\ell_\epsilon n^{\ell}_{i1})$. As `CollectSamples` runs each policy $\lceil 2n^{\ell}_{i1}/N^{\ell}_{hij}\rceil$ times, it follows that we will collect at least $\lceil 2n^{\ell}_{i1}/N^{\ell}_{hij}\rceil \cdot \frac{1}{2}N^{\ell}_{hij} \geq n^{\ell}_{i1}$ samples from each $(s,a) \in \mathcal{X}^{\ell}_{hij}$ with probability at least $1 - \delta_{\mathrm{tol}(m)}/(H\varsigma^2_\epsilon\ell_\epsilon n^{\ell}_{i1}) \cdot \lceil 2n^{\ell}_{1i}/N^{\ell}_{hij}\rceil \geq 1 - 3\delta_{\mathrm{tol}(m)}/(H\varsigma^2_\epsilon\ell_\epsilon)$. Union bounding over this for each h, i, ℓ and $j \in [\varsigma_\epsilon]$ gives that with probability at least $1 - 3\delta_{\mathrm{tol}(m)}$, we collect at least n^{ℓ}_{i1} samples from each $(s,a) \in \mathcal{X}^{\ell}_{hij}$. The same argument gives that with probability at least $1 - 3\delta_{\mathrm{tol}(m)}$ we collect at least $n^{\ell_\epsilon+1}_{j}$ samples from each $(s,a) \in \mathcal{X}^{\ell_\epsilon+1}_{hj}$, $j = 1, \ldots, \varsigma_\epsilon, h \in [H]$.

Relating $\widehat{W}_h(s)$ to $W_h(s)$. It remains to show that $\widehat{W}_h(s) \leq W_h(s) \leq 32\widehat{W}_h(s)$ for all $s \in \mathcal{Z}_h$, $\cup_{j=1}^{\varsigma_\epsilon}\mathcal{X}^{\ell}_{hij} = \mathcal{Z}^{\ell}_{hi}$, and $\cup_{j=1}^{\varsigma_\epsilon}\mathcal{X}^{\ell_\epsilon+1}_{hj} = \mathcal{Z}^{\ell_\epsilon+1}_{h}$.

We first show $\widehat{W}_h(s) \leq W_h(s) \leq 32\widehat{W}_h(s)$. Consider running `Learn2Explore` with $\mathcal{X} = \{(s,a)\}$ for arbitrary a and assume that $\mathcal{X}^{sh}_j$ is the returned partition containing (s,a). By Theorem E.1, on $\mathcal{E}^{sh}_{\mathsf{L2E}}$ we will have that

$$W_h(s) \leq 2^{-j+1}$$

and, furthermore, that with probability at least $1/2$, if we rerun all policies in Π_j^{sh} returned by `Learn2Explore`, we will obtain at least $N_j^{sh}/2 = |\Pi_j^{sh}|/(8|\mathcal{X}|2^j) = |\Pi_j^{sh}|/(8 \cdot 2^j)$ samples from (s, a, h).

Let X be a random variable which is the count of total samples collected in (s, a, h) when running $\pi_k \in \Pi_j^{sh}$. Then Markov's inequality and the above property of Π_j^{sh} gives

$$\frac{1}{2} \le \mathbb{P}[X \ge N_j^{sh}/2] \le \frac{2\mathbb{E}[X]}{N_j^{sh}} = \frac{2}{N_j^{sh}} \sum_{\pi \in \Pi_j^{sh}} w_h^\pi(s,a) \le \frac{2|\Pi_j^{sh}|}{N_j^{sh}} W_h(s) = 8 \cdot 2^j W_h(s).$$

Rearranging this and recalling that we set $\widehat{W}_h(s) = \frac{1}{16 \cdot 2^j}$, we have that $\widehat{W}_h(s) \le W_h(s)$. However, we also have

$$W_h(s) \le 2^{-j+1} = 32\widehat{W}_h(s).$$

This proves that $\widehat{W}_h(s) \le W_h(s) \le 32\widehat{W}_h(s)$.

Now note that any $s \in \mathcal{Z}_h$ has $\widehat{W}_h(s) \ge \frac{\epsilon_{\mathrm{tol}(m)}}{32H^2S}$, which, combined with the above, implies that $W_h(s) \ge \frac{\epsilon_{\mathrm{tol}(m)}}{32H^2S}$. Fix (h, i, ℓ), and note that the call to `Learn2Explore` in the call to `CollectSamples` for index (h, i, ℓ) uses input tolerance $\frac{\epsilon_{\mathrm{tol}(m)}}{64H^2S}$. Theorem E.1 then gives that, on $\mathcal{E}_{\mathsf{L2E}}^{hi\ell}$, we will have

$$\sup_\pi \sum_{(s,a) \in \mathcal{Z}_{hi}^\ell \setminus (\cup_{j=1}^{\varsigma_\epsilon} \mathcal{X}_{hij}^\ell)} w_h^\pi(s,a) \le \frac{\epsilon_{\mathrm{tol}(m)}}{64H^2S}.$$

However, as $W_h(s') \le \sup_\pi \sum_{(s,a) \in \mathcal{Z}_{hi}^\ell \setminus (\cup_{j=1}^{\varsigma_\epsilon} \mathcal{X}_{hij}^\ell)} w_h^\pi(s,a)$ for any $(s', a) \in \mathcal{Z}_{hi}^\ell \setminus (\cup_{j=1}^{\varsigma_\epsilon} \mathcal{X}_{hij}^\ell)$, we will have that any $(s, a) \in \mathcal{Z}_{hi}^\ell \setminus (\cup_{j=1}^{\varsigma_\epsilon} \mathcal{X}_{hij}^\ell)$ has $W_h(s) \le \frac{\epsilon_{\mathrm{tol}(m)}}{64H^2S}$. This is a contradiction since we know $W_h(s) \ge \frac{\epsilon_{\mathrm{tol}(m)}}{32H^2S}$ for any $(s, a) \in \mathcal{Z}_{hi}^\ell$. Thus, we must have that $\mathcal{Z}_{hi}^\ell \setminus (\cup_{j=1}^{\varsigma_\epsilon} \mathcal{X}_{hij}^\ell) = \emptyset$ so $\cup_{j=1}^{\varsigma_\epsilon} \mathcal{X}_{hij}^\ell = \mathcal{Z}_{hi}^\ell$. The same argument shows that $\cup_{j=1}^{\varsigma_\epsilon} \mathcal{X}_{hj}^{\ell_\epsilon+1} = \mathcal{Z}_h^{\ell_\epsilon+1}$.

Completing the proof. We have therefore shown that $\mathbb{P}[\mathcal{E}_{\mathrm{exp}}^m] \ge 1 - 9\delta_{\mathrm{tol}(m)}$. Union bounding over all m, by our choice of $\delta_{\mathrm{tol}(m)} = \frac{\delta_{\mathrm{tol}}}{36m^2}$, we have that

$$\mathbb{P}[\mathcal{E}_{\mathrm{exp}}] = \mathbb{P}[\cap_{m=1}^{\lceil \log H/\epsilon \rceil} \mathcal{E}_{\mathrm{exp}}^m] \ge 1 - \sum_{m=1}^{\lceil \log H/\epsilon \rceil} 9\frac{\delta_{\mathrm{tol}}}{36m^2} \ge 1 - \delta_{\mathrm{tol}}/2.$$

Union bounding over $\mathcal{E}_{\mathrm{exp}}$ and $\mathcal{E}_{\mathrm{est}}$ then gives the result. □

Proof of Claim 5. We proceed by induction. Consider some $s \in \mathcal{Z}_{hi}$. The base case is trivial as $\mathcal{A}_h^0(s) = \mathcal{A}$. Fix some $\ell \le \ell_\epsilon$ and assume that $\widehat{a}_h^\star(s) \in \mathcal{A}_h^{\ell-1}(s)$ and $|\mathcal{A}_h^{\ell-1}(s)| > 1$. Then, on $\mathcal{E}_{\mathrm{exp}}$, we can guarantee that we will collect at least $\frac{2^{18}H^2\varsigma_\delta}{2^{2i}\epsilon_\ell^2}$ samples from (s, a) for each

$a \in \mathcal{A}_h^{\ell-1}$. On the event $\mathcal{E}_{\text{est}}$, it then follows that for each $a \in \mathcal{A}_h^{\ell-1}(s)$,

$$|\widehat{Q}_{h,\ell}^{\widehat{\pi}}(s,a) - Q_h^{\widehat{\pi}}(s,a)| \le 2^i \epsilon_\ell / 2^9.$$

Thus, since by assumption $\widehat{a}_h^\star(s) \in \mathcal{A}_h^{\ell-1}(s)$,

$$\begin{aligned}\max_{a \in \mathcal{A}_h^{\ell-1}(s)} \widehat{Q}_{h,\ell}^{\widehat{\pi}}(s,a) - \widehat{Q}_{h,\ell}^{\widehat{\pi}}(s,\widehat{a}_h^\star(s)) &\le \max_{a \in \mathcal{A}_h^{\ell-1}(s)} Q_h^{\widehat{\pi}}(s,a) - Q_h^{\widehat{\pi}}(s,\widehat{a}_h^\star(s)) + 2 \cdot 2^i \epsilon_\ell / 2^9 \\ &\le 2 \cdot 2^i \epsilon_\ell / 2^9 \\ &= \gamma_{ij}^\ell\end{aligned}$$

for any j, so the exit condition on Line 16 of `EliminateActions` is not met for $\widehat{a}_h^\star(s)$, and thus $\widehat{a}_h^\star(s) \in \mathcal{A}_h^\ell(s)$. The result follows analogously if $\ell = \ell_\epsilon + 1$, in which case we simply use the different values of n and γ.

Now if $(s,a) \notin \mathcal{Z}_{hi}^\ell$ for all a, that means we will never remove arms from $\mathcal{A}_h^\ell(s)$ again. However, by the above inductive argument, if ℓ' is the last round such that $(s,a) \in \mathcal{Z}_{hi}^{\ell'}$ for some a, we will have that $\widehat{a}_h^\star(s) \in \mathcal{A}_h^{\ell'}(s)$, so it follows that $s \in \mathcal{A}_h^\ell(s)$.

Finally, if $s \notin \mathcal{Z}_h$, then we will never remove an arm from $\mathcal{A}_h^0(s)$, and since $\mathcal{A}_h^0(s) = \mathcal{A}$, the conclusion follows trivially. □

Proof of Claim 6. In Lemma 7.4.5, we showed that the local suboptimality bounds of $\widehat{\pi}$, $\epsilon_h(s)$, satisfy

$$\sum_{h=1}^H \sup_\pi \sum_s w_h^\pi(s) \epsilon_h(s) \le \epsilon.$$

By Lemma E.2.2, it follows that for any π' and any h,

$$\sum_s w_h^{\pi'}(s)(V_h^\star(s) - V_h^{\widehat{\pi}}(s)) \le \sum_{h'=h}^H \sup_\pi \sum_s w_{h'}^\pi(s) \epsilon_{h'}(s) \le \epsilon.$$

The result then follows from Lemma E.2.3. □

Lemma E.2.2. *Assume that for each h and s, $\widehat{\pi}$ plays an action which satisfies*

$$\max_a Q_h^{\widehat{\pi}}(s,a) - Q_h^{\widehat{\pi}}(s, \widehat{\pi}_h(s)) \le \epsilon_h(s). \tag{E.2.2}$$

Then for any h and π',

$$\sum_s w_h^{\pi'}(s)(V_h^\star(s) - V_h^{\widehat{\pi}}(s)) \le \sum_{h'=h}^H \sup_\pi \sum_s w_{h'}^\pi(s) \epsilon_{h'}(s).$$

Proof. We proceed by backwards induction. The base case, $h = H$, is trivial. Assume that at level h, for any π,

$$\sum_s w_h^\pi(s)(V_h^\star(s) - V_h^{\widehat{\pi}}(s)) \le \sum_{h'=h}^{H} \sup_{\pi'} \sum_{s'} w_{h'}^{\pi'}(s')\epsilon_{h'}(s')$$

and that at level $h-1$, for each s (E.2.2) holds. By definition,

$$\begin{aligned} V_{h-1}^\star(s) - V_{h-1}^{\widehat{\pi}}(s) &= Q_{h-1}^\star(s, \pi_{h-1}^\star(s)) - Q_{h-1}^{\widehat{\pi}}(s, \widehat{\pi}_{h-1}(s)) \\ &= Q_{h-1}^\star(s, \pi_{h-1}^\star(s)) - Q_{h-1}^{\widehat{\pi}}(s, \pi_{h-1}^\star(s)) + Q_{h-1}^{\widehat{\pi}}(s, \pi^\star(s)) - \max_a Q_{h-1}^{\widehat{\pi}}(s,a) \\ &\qquad + \max_a Q_{h-1}^{\widehat{\pi}}(s,a) - Q_{h-1}^{\widehat{\pi}}(s, \widehat{\pi}_{h-1}(s)). \end{aligned}$$

Clearly, $Q_{h-1}^{\widehat{\pi}}(s, \pi^\star(s)) - \max_a Q_{h-1}^{\widehat{\pi}}(s,a) \le 0$ and by assumption $\max_a Q_{h-1}^{\widehat{\pi}}(s,a) - Q_{h-1}^{\widehat{\pi}}(s, \widehat{\pi}_{h-1}(s)) \le \epsilon_{h-1}(s)$. Furthermore,

$$Q_{h-1}^\star(s, \pi_{h-1}^\star(s)) - Q_{h-1}^{\widehat{\pi}}(s, \pi_{h-1}^\star(s)) = \sum_{s'} P_{h-1}(s'|s, \pi_{h-1}^\star(s))(V_h^\star(s') - V_h^{\widehat{\pi}}(s')).$$

Then, for any π,

$$\begin{aligned} &\sum_s w_{h-1}^\pi(s)(V_{h-1}^\star(s) - V_{h-1}^{\widehat{\pi}}(s)) \\ &\le \sum_s w_{h-1}^\pi(s)\epsilon_{h-1}(s) + \sum_s \sum_{s'} w_{h-1}^\pi(s) P_{h-1}(s'|s, \pi_{h-1}^\star(s))(V_h^\star(s') - V_h^{\widehat{\pi}}(s')) \\ &= \sum_s w_{h-1}^\pi(s)\epsilon_{h-1}(s) + \sum_s w_h^{\pi'}(s)(V_h^\star(s) - V_h^{\widehat{\pi}}(s)) \\ &\le \sum_{h'=h-1}^{H} \sup_{\pi'} \sum_{s'} w_{h'}^{\pi'}(s')\epsilon_{h'}(s') \end{aligned}$$

where the last inequality follows by the inductive hypothesis and we have used that

$$\sum_s w_{h-1}^\pi(s) P_{h-1}(s'|s, \pi_{h-1}^\star(s)) = w_h^{\pi'}(s').$$

where $\pi'_{h'}(s) = \pi_{h'}(s)$ for all $h' \le h-2$ and $\pi'_{h'}(s) = \pi_{h'}^\star(s)$ for $h' \ge h-1$. The conclusion then follows. □

E.2.4 MDP Technical Results

Proof of Proposition 7.6. This result follows from the Performance-Difference Lemma. We give the full proof for completeness. The following is the standard proof of the Performance-

Difference Lemma:

$$\begin{aligned}
V_0^\star - V_0^\pi &= \mathbb{E}_{\pi^\star, s_0 \sim P_0}\left[\sum_{h=1}^H r_h(s_h, a_h)\right] - V_0^\pi \\
&= \mathbb{E}_{\pi^\star, s_0 \sim P_0}\left[\sum_{h=1}^H r_h(s_h, a_h) + V_h^\pi(s_h) - V_h^\pi(s_h)\right] - V_0^\pi \\
&= \mathbb{E}_{\pi^\star, s_0 \sim P_0}\left[\sum_{h=1}^H r_h(s_h, a_h) + V_{h+1}^\pi(s_{h+1}) - V_h^\pi(s_h)\right] \\
&= \mathbb{E}_{\pi^\star, s_0 \sim P_0}\left[\sum_{h=1}^H r_h(s_h, a_h) + \mathbb{E}[V_{h+1}^\pi(s')|s_h, a_h] - V_h^\pi(s_h)\right] \\
&= \mathbb{E}_{\pi^\star, s_0 \sim P_0}\left[\sum_{h=1}^H Q_h^\pi(s_h, a_h) - V_h^\pi(s_h)\right] \\
&= \sum_{h=1}^H \sum_{s,a} w_h^{\pi^\star}(s, a)(Q_h^\pi(s, a) - V_h^\pi(s)).
\end{aligned}$$

In the case when π is deterministic, we have $V_h^\pi(s) = Q_h^\pi(s, \pi_h(s))$. Furthermore, we can upper bound the above by

$$\sum_{h=1}^H \sum_{s,a} w_h^{\pi^\star}(s, a)(\max_{a'} Q_h^\pi(s, a') - Q_h^\pi(s, \pi_h(s))) = \sum_{h=1}^H \sum_{s} w_h^{\pi^\star}(s)\epsilon_h(s).$$

The result follows by upper bounding the visitation under $\pi^\star$ by the visitation under the worst-case policy. □

Lemma E.2.3. *Assume that*

$$\sup_\pi \sum_{s'} w_h^\pi(s')(V_h^\star(s') - V_h^{\widehat{\pi}}(s')) \le \epsilon \quad \textit{and} \quad \sup_\pi \sum_{s'} w_{h+1}^\pi(s')(V_{h+1}^\star(s') - V_{h+1}^{\widehat{\pi}}(s')) \le \epsilon.$$

Then, for any s,

$$|\Delta_h(s, a) - \Delta_h^{\widehat{\pi}}(s, a)| \le \epsilon / W_h(s).$$

Proof. By definition,

$$\begin{aligned}
|\Delta_h(s, a) - \Delta_h^{\widehat{\pi}}(s, a)| &= |V_h^\star(s) - Q_h^\star(s, a) - (\max_{a'} Q_h^{\widehat{\pi}}(s, a') - Q_h^{\widehat{\pi}}(s, a))| \\
&\le \max\{|V_h^\star(s) - \max_{a'} Q_h^{\widehat{\pi}}(s, a')|, |Q_h^{\widehat{\pi}}(s, a) - Q_h^\star(s, a)|\}.
\end{aligned}$$

where the last inequality follows since

$$V_h^\star(s) - Q_h^\star(s,a) - (\max_{a'} Q_h^{\widehat{\pi}}(s,a') - Q_h^{\widehat{\pi}}(s,a)) \leq V_h^\star(s) - \max_{a'} Q_h^{\widehat{\pi}}(s,a')$$

and

$$-(V_h^\star(s) - Q_h^\star(s,a) - (\max_{a'} Q_h^{\widehat{\pi}}(s,a') - Q_h^{\widehat{\pi}}(s,a))) \leq Q_h^\star(s,a) - Q_h^{\widehat{\pi}}(s,a).$$

Now,

$$\begin{aligned} V_h^\star(s) - \max_{a'} Q_h^{\widehat{\pi}}(s,a') &= V_h^\star(s) - Q_h^{\widehat{\pi}}(s,\widehat{\pi}_h(s)) + Q_h^{\widehat{\pi}}(s,\widehat{\pi}_h(s)) - \max_{a'} Q_h^{\widehat{\pi}}(s,a') \\ &\leq V_h^\star(s) - V_h^{\widehat{\pi}}(s) \end{aligned}$$

where the inequality follows since, by definition, $V_h^{\widehat{\pi}}(s) = Q_h^{\widehat{\pi}}(s,\widehat{\pi}_h(s))$ and $Q_h^{\widehat{\pi}}(s,\widehat{\pi}_h(s)) - \max_{a'} Q_h^{\widehat{\pi}}(s,a') \leq 0$. By assumption,

$$\sup_\pi \sum_{s'} w_h^\pi(s')(V_h^\star(s') - V_h^{\widehat{\pi}}(s')) \leq \epsilon$$

and furthermore, for any s,

$$\sup_\pi \sum_{s'} w_h^\pi(s')(V_h^\star(s') - V_h^{\widehat{\pi}}(s')) \geq W_h(s)(V_h^\star(s) - V_h^{\widehat{\pi}}(s))$$

so it follows that $|V_h^\star(s) - V_h^{\widehat{\pi}}(s)| \leq \epsilon / W_h(s)$. By definition,

$$Q_h^\star(s,a) - Q_h^{\widehat{\pi}}(s,a) = \sum_{s'} P_h(s'|s,a)(V_{h+1}^\star(s') - V_{h+1}^{\widehat{\pi}}(s'))$$

so

$$\begin{aligned} W_h(s)(Q_h^\star(s,a) - Q_h^{\widehat{\pi}}(s,a)) &= \sum_{s'} P_h(s'|s,a) W_h(s)(V_{h+1}^\star(s') - V_{h+1}^{\widehat{\pi}}(s')) \\ &\leq \sup_\pi \sum_{s'} w_{h+1}^\pi(s')(V_{h+1}^\star(s') - V_{h+1}^{\widehat{\pi}}(s')) \end{aligned}$$

where the inequality follows since $V_{h+1}^\star(s') \geq V_{h+1}^{\widehat{\pi}}(s')$, and since

$$P_h(s'|s,a)W_h(s) = \mathbb{P}[s_{h+1} = s'|s_h = s, a_h = a]\mathbb{P}_\pi[s_h = s] = \mathbb{P}_{\pi'}[s_{h+1} = s', s_h = s] \leq \mathbb{P}_{\pi'}[s_{h+1}]$$

where π denotes the policy achieving $\mathbb{P}_\pi[s_h = s] = W_h(s)$ and π' plays π up to h and then $\pi'_h(s) = a$. Thus, if $\sup_\pi \sum_{s'} w_{h+1}^\pi(s')(V_{h+1}^\star(s') - V_{h+1}^{\widehat{\pi}}(s')) \leq \epsilon$, rearranging the inequalities gives the result.

□

Proposition E.4. *Fix some MDP M. Then:*

1. *The set of valid state-action visitation distributions on M is convex.*
2. *For any valid state-action visitation distribution on M, there exists some policy which realizes it.*

Proof. The set of valid state-action visitation distributions, $\mathcal{W}$, is defined as

$$\mathcal{W} := \Big\{ w \in [0,1]^{SAH} \ : \ \exists \pi \in \Pi \text{ s.t. } w_h(s,a) = \pi_h(a|s) \cdot \sum_{s',a'} P_{h-1}(s|s',a') w_{h-1}(s',a'), \forall h \geq 1, \\ w_0(s,a) = \pi_0(a|s) P_0(s), \quad \sum_{s,a} w_h(s,a) = 1, \forall h \geq 0 \Big\}$$

where here $\Pi = \triangle(\mathcal{A})^{SH}$.

Fix some state-action visitation distributions $w, w' \in \mathcal{W}$, and let π and π' denote their correponding policies as above. Furthermore, denote $w_h(s) = \sum_a w_h(s,a)$ (and similarly for w'). Our goal is to show that for any $t \in [0,1]$, $\widetilde{w} = (1-t)w + tw' \in \mathcal{W}$. First, we show that there exists some policy $\widetilde{\pi}$ such that

$$(1-t)w_0(s,a) + tw'_0(s,a) = \widetilde{\pi}_0(a|s) P_0(s).$$

Note that we can take $\widetilde{\pi}_0(a|s) = (1-t)\pi_0(a|s) + t\pi'_0(a|s)$, since

$$((1-t)\pi_0(a|s) + t\pi'_0(a|s)) P_0(s) = (1-t)w_0(s,a) + tw'_0(s,a).$$

By construction, for any $h \geq 1$,

$$\widetilde{w}_h(s) = \sum_a \widetilde{w}_h(s,a) = (1-t) \sum_a w_h(s,a) + t \sum_a w'_h(s,a) = (1-t)w_h(s) + tw'_h(s).$$

Furthermore, since w is a valid state-action distribution,

$$w_h(s) = \sum_{s',a'} P_{h-1}(s|s',a') w_{h-1}(s',a')$$

and similarly for w'. Let $\widetilde{\pi}_h(a|s) = \widetilde{w}_h(s,a)/\widetilde{w}_h(s)$ (where we define $0/0 = 0$), and note that this is a valid distribution since by definition $\sum_a \widetilde{w}_h(s,a) = \widetilde{w}_h(s)$. Then,

$$\begin{aligned} \widetilde{w}_h(s,a) &= \widetilde{\pi}_h(a|s) \widetilde{w}_h(s) \\ &= \widetilde{\pi}_h(a|s)((1-t)w_h(s) + tw'_h(s)) \end{aligned}$$

$$= \widetilde{\pi}_h(a|s) \sum_{s',a'} P_{h-1}(s|s',a')((1-t)w_{h-1}(s',a') + tw'_{h-1}(s',a'))$$
$$= \widetilde{\pi}_h(a|s) \sum_{s',a'} P_{h-1}(s|s',a')\widetilde{w}_{h-1}(s',a')$$

where the last equality follows by the definition of $\widetilde{w}_{h-1}$. The other constraints are trivial to verity, so $\widetilde{w} \in \mathcal{W}$. This proves the first result.

For the second result, take some $w \in \mathcal{W}$, and let $\pi_h(a|s) = w_h(s,a)/w_h(s)$. By definition this is a valid distribution. Furthermore, it trivially holds that $w_0^\pi(s,a) = w_0(s,a)$. Assume that $w_{h-1}^\pi(s,a) = w_{h-1}(s,a)$ for all (s,a). By definition and the inductive hypothesis,

$$\begin{aligned} w_h^\pi(s,a) &= \pi_h(a|s) \sum_{s',a'} P_{h-1}(s|s',a') w_{h-1}^\pi(s',a') \\ &= \pi_h(a|s) \sum_{s',a'} P_{h-1}(s|s',a') w_{h-1}(s,a) \\ &= \pi_h(a|s) w_h(s) \\ &= w_h(s,a), \end{aligned}$$

which proves the second result. □

E.3 Proof of Theorem 7.2

Define the following value:

$$\begin{aligned} K_i(\delta, \delta_{\text{samp}}) = \Bigg\lceil 2^i \max \Bigg\{ & 288 c_{\text{eu}}^2 S^2 A^2 H (i+3) \log(576 c_{\text{eu}} SAH(i+3)), 288 c_{\text{eu}}^2 S^2 A^2 H \log \frac{2SAH}{\delta}, \\ & 2048 S^2 A^2 \log \frac{4SAH}{\delta_{\text{samp}}}, 256 c_{\text{eu}} S^3 A^2 H^4 (i+9)^3 \log^3 \left(512 c_{\text{eu}} SAH(i+9)\right), \\ & 128 c_{\text{eu}} S^3 A^2 H^4 \log^3 \frac{2SAH}{\delta} + 8H \log \frac{4}{\delta} \Bigg\} \Bigg\rceil \\ =: 2^i C_K(\delta, \delta_{\text{samp}}, i) \end{aligned} \tag{E.3.1}$$

and note that $C_K(\delta, \delta_{\text{samp}}, i) = \text{poly}(S, A, H, \log 1/\delta, \log 1/\delta_{\text{samp}}, i)$.

Theorem E.1 (Formal Statement of Theorem 7.2)**.** *Consider running* `Learn2Explore` *with tolerance* $\epsilon_{\mathsf{L2E}} \leftarrow \epsilon$ *and confidence* δ *and obtaining a partition* $\mathcal{X}_i \subseteq \mathcal{S} \times \mathcal{A}$ *and policies* Π_i*,* $i \in \{1, 2, \ldots, \lceil \log(1/\epsilon) \rceil\}$*. Let* $\mathcal{E}_{\mathsf{L2E}}$ *be the event on which, for all* i *simultaneously:*

1. *Sets $\mathcal{X}_i$ satisfy:*

$$\sup_{\pi} \sum_{(s,a)\in\mathcal{X}_i} w_h^\pi(s,a) \le 2^{-(i-1)}$$

2. *For any i, if all policies in Π_i are each rerun once, we will collect $\frac{1}{2}N_i$ samples from each $(s,a)\in\mathcal{X}_i$ with probability $1-\delta_{\mathrm{samp}}$, where we recall $N_i = K_i(\delta/\lceil\log(1/\epsilon)\rceil, \delta_{\mathrm{samp}})/(4\cdot 2^i|\mathcal{X}\setminus\cup_{i'=1}^{i-1}\mathcal{X}_{i'}|)$.*

3. *The remaining states, $\mathcal{X}\setminus(\cup_{i=1}^{\lceil\log(1/\epsilon)\rceil}\mathcal{X}_i)$ satisfy,*

$$\sup_{\pi} \sum_{(s,a)\in(\mathcal{X}\setminus(\cup_{i=1}^{\lceil\log(1/\epsilon)\rceil}\mathcal{X}_i))} w_h^\pi(s,a) \le \epsilon.$$

Then $\mathbb{P}[\mathcal{E}_{\mathsf{L2E}}] \ge 1-\delta$. Furthermore, Algorithm 7.1 takes at most

$$C_K\left(\frac{\delta}{\lceil\log 1/\epsilon\rceil}, \delta_{\mathrm{samp}}, \lceil\log 1/\epsilon\rceil\right)\frac{4}{\epsilon}$$

episodes to terminate.

Proof. This directly follows by induction and Lemma E.3.1. For $i=1$, it will clearly be the case that

$$\sup_{\pi} \sum_{(s,a)\in\mathcal{X}} w_h^\pi(s,a) \le 2^{-(i-1)} = 1$$

since $\sum_{s,a} w_h^\pi(s,a) = 1$ for any π and h. Now consider an epoch i and assume that

$$\sup_{\pi} \sum_{(s,a)\in\mathcal{X}} w_h^\pi(s,a) \le 2^{-(i-1)}.$$

By Lemma E.3.1, running `FindExplorableSets` will produce a set $\mathcal{X}_i$ and policies Π_i such that

$$\sup_{\pi} \sum_{(s,a)\in\mathcal{X}_i} w_h^\pi(s,a) \le 2^{-(i-1)}, \quad \sup_{\pi} \sum_{(s,a)\in\mathcal{X}\setminus\mathcal{X}_i} w_h^\pi(s,a) \le 2^{-i}$$

and rerunning every policy in Π_i at once will allow us to collect at least $\frac{1}{2}N_i$ samples from each $(s,a)\in\mathcal{X}_i$. As $\mathcal{X}\leftarrow\mathcal{X}\setminus\mathcal{X}_i$, the hypothesis will then be met at the next epoch, $i+1$.

Union bounding over epochs completes the first part of the proof. That

$$\sup_{\pi} \sum_{(s,a)\in(\mathcal{X}\setminus(\cup_{i=1}^{\lceil\log(1/\epsilon)\rceil}\mathcal{X}_i))} w_h^\pi(s,a) \le \epsilon$$

follows on this same event by Lemma E.3.1 and since we run until $i = \lceil\log(1/\epsilon)\rceil$ which implies $2^{-\lceil\log(1/\epsilon)\rceil} \le \epsilon$. Union bounding over each i gives the result.

The sample complexity bound follows by bounding

$$\begin{aligned}\sum_{i=1}^{\lceil\log(1/\epsilon)\rceil} K_i(\delta/\lceil\log 1/\epsilon\rceil, \delta_{\text{samp}}) &\le C_K\left(\frac{\delta}{\lceil\log 1/\epsilon\rceil}, \delta_{\text{samp}}, \lceil\log 1/\epsilon\rceil\right) \sum_{i=1}^{\lceil\log(1/\epsilon)\rceil} 2^i \\ &\le C_K\left(\frac{\delta}{\lceil\log 1/\epsilon\rceil}, \delta_{\text{samp}}, \lceil\log 1/\epsilon\rceil\right)\frac{4}{\epsilon}.\end{aligned}$$

□

Lemma E.3.1. *Assume that $\mathcal{X}$ satisfies*

$$\sup_{\pi} \sum_{(s,a)\in\mathcal{X}} w_h^\pi(s,a) \le 2^{-(i-1)}.$$

Then, if `FindExplorableSets`$(\mathcal{X}, h, \delta, K_i, N_i)$ *returns partition $\mathcal{X}_i$ and policies Π_i, with probability $1-\delta$ the returned partition $\mathcal{X}_i$ will satisfy*

$$\sup_{\pi} \sum_{(s,a)\in\mathcal{X}_i} w_h^\pi(s,a) \le 2^{-(i-1)}, \quad \sup_{\pi} \sum_{(s,a)\in\mathcal{X}\setminus\mathcal{X}_i} w_h^\pi(s,a) \le 2^{-i}.$$

Furthermore, if all policies in Π_i are each rerun once, we will collect $\frac{1}{2}N_i$ samples from each $(s,a,h) \in \mathcal{X}_i$ with probability $1-\delta_{\text{samp}}$.

Proof. The structure of this proof takes inspiration from the proof presented in Zhang et al. [2020]. The first conclusion is trivial since $\mathcal{X}_i \subseteq \mathcal{X}$ and by our assumption on $\mathcal{X}$.

We will simply denote $K_i := K_i(\delta, \delta_{\text{samp}})$ throughout the proof. In addition, we will let K_{ij} denote the total number of epochs taken for fixed j, and will let m_i denote the total number of times j is incremented. Therefore,

$$K_i = \sum_{j=1}^{m_i} K_{ij}.$$

Let $V_0^{\star,ij}$ denote the optimal value function on the reward function r_h^j at stage j of epoch i.

By our assumption on $\mathcal{X}$ and the definition of our reward function we can bound

$$V_0^{\star,ij} \le \sup_\pi \mathbb{E}_\pi[\mathbb{I}\{(s_h,a_h)\in\mathcal{X}\}] = \sup_\pi \sum_{(s,a)\in\mathcal{X}} w_h^\pi(s,a) \le 2^{-(i-1)}. \tag{E.3.2}$$

As `FindExplorableSets` runs EULER, by Lemma E.3.4 we will have, with probability at least $1-\delta$, for any fixed K and j,

$$\Big(\sum_{k=1}^K V_0^{\star,ij} - \sum_{k=1}^K V_0^{k,ij}\Big)|\mathcal{F}_{j-1} \le c_{\mathrm{eu}}\sqrt{SAHV_0^{\star,i1}K\log\frac{SAHK}{\delta}} + c_{\mathrm{eu}}S^2AH^4\log^3\frac{SAHK}{\delta} \tag{E.3.3}$$

where $\mathcal{F}_{j-1}$ denotes the filtration of up to iteration j, and we have used that $V_0^{\star,ij} \le V_0^{\star,i1}$ for all j since the reward function can only decrease as j increases. `FindExplorableSets` terminates and restarts EULER if the condition on Line 14 is met, but this is a *random* stopping condition. As such, to guarantee that (E.3.3) holds for any possible value of this stopping time, we union bound over all values. Since `FindExplorableSets` runs for at most K_i epochs, it suffices to union bound over K_i stopping times. We then have that

$$\Big(\sum_{k=1}^K V_0^{\star,ij} - \sum_{k=1}^K V_0^{k,ij}\Big)|\mathcal{F}_{j-1} \le 2c_{\mathrm{eu}}\sqrt{SAHV_0^{\star,i1}K\log\frac{2SAHK_i}{\delta}} + 8c_{\mathrm{eu}}S^2AH^4\log^3\frac{2SAHK_i}{\delta}$$

with probability at least $1-\frac{\delta}{2SA}$ for all $K\in[1,K_i]$ simultaneously. Since $m_i \le SA$, union bounding over all j we then have that, with probability at least $1-\delta/2$,

$$\begin{aligned}\sum_{j=1}^{m_i}\Big(\sum_{k=1}^{K_{ij}} V_0^{\star,ij} - \sum_{k=1}^{K_{ij}} V_0^{k,ij}\Big) &\le \sum_{j=1}^{m_i} 2c_{\mathrm{eu}}\sqrt{SAHV_0^{\star,i1}K_{ij}\log\frac{2SAHK_i}{\delta}} + 8c_{\mathrm{eu}}S^3A^2H^4\log^3\frac{2SAHK_i}{\delta}\\ &\le 2c_{\mathrm{eu}}\sqrt{S^2A^2HV_0^{\star,i1}K_i\log\frac{2SAHK_i}{\delta}} + 8c_{\mathrm{eu}}S^3A^2H^4\log^3\frac{2SAHK_i}{\delta}\end{aligned}$$

where the final inequality follows from Jensen's inequality. Using the same calculation as in the proof of Lemma E.3.4, we can bound

$$\mathbb{E}_{\pi_k}[(\sum_{h=1}^H R_h^j(s_h,a_h) - V_0^{k,ij})^2] \le 4V_0^{k,ij}$$

By (E.3.2), $4V_0^{k,ij} \le 4/2^{i-1}$, so we can apply Lemma E.3.5 with $\sigma_V^2 = 4/2^{i-1}$, to get that,

with probability at least $1-\delta/2$,

$$\left|\sum_{j=1}^{m_i}\sum_{k=1}^{K_{ij}}\sum_{h=1}^{H} R_h^j(s_h^{j,k},a_h^{j,k}) - \sum_{j=1}^{m_i}\sum_{k=1}^{K_{ij}} V_0^{k,ij}\right| \le \sqrt{32K_i 2^{-i}\log\frac{4}{\delta}} + 2H\log\frac{4}{\delta}.$$

Putting this together and union bounding over these events, we have that with probability at least $1-\delta$,

$$\sum_{j=1}^{m_i}\sum_{k=1}^{K_{ij}}\sum_{h=1}^{H} R_h^j(s_h^{j,k},a_h^{j,k}) \ge \sum_{j=1}^{m_i}\sum_{k=1}^{K_{ij}} V_0^{\star,ij} - \sqrt{64K_i 2^{-i}\log\frac{4}{\delta}} - 2c_{\text{eu}}\sqrt{S^2A^2HV_0^{\star,i1}K_i\log\frac{2SAHK_i}{\delta}} - C_{\mathcal{R}}$$

where we denote

$$C_{\mathcal{R}} := 8c_{\text{eu}}S^3A^2H^4\log^3\frac{2SAHK_i}{\delta} + 2H\log\frac{4}{\delta}.$$

Assume that $V_0^{\star,im_i} > 2^{-i}$. Using that the reward decreases monotonically so $V_0^{\star,im_i} \le V_0^{\star,ij}$ for any $j \le m_i$, we can lower bound the above as

$$\begin{aligned}
&\ge 2^{-i}K_i - \sqrt{64K_i 2^{-i}\log\frac{4}{\delta}} - 2c_{\text{eu}}\sqrt{S^2A^2HV_0^{\star,i1}K_i\log\frac{2SAHK_i}{\delta}} - C_{\mathcal{R}} \\
&\ge 2^{-i}K_i - 3c_{\text{eu}}\sqrt{S^2A^2H2^{-i}K_i\log\frac{2SAHK_i}{\delta}} - C_{\mathcal{R}}
\end{aligned}$$

where the second inequality follows by (E.3.2) and since $\sqrt{64K_i2^{-i}\log\frac{4}{\delta}}$ will then be dominated by the regret term, $c_{\text{eu}}\sqrt{S^2A^2HV_0^{\star,i1}K_i\log\frac{2SAHK_i}{\delta}}$. Lemma E.3.2 gives

$$K_i \ge 2^i \max\left\{4C_{\mathcal{R}}, 144c_{\text{eu}}^2S^2A^2H\log\frac{2SAHK_i}{\delta}\right\}$$

which implies

$$\frac{1}{4}2^{-i}K_i - C_{\mathcal{R}} \ge 0$$

and

$$\begin{aligned}
&\frac{1}{4}2^{-i}K_i - 3c_{\text{eu}}\sqrt{S^2A^2H2^{-i}K_i\log\frac{2SAHK_i}{\delta}} \\
&\ge \frac{2^i\cdot 144c_{\text{eu}}^2S^2A^2H\log\frac{2SAHK_i}{\delta}}{4\cdot 2^i} - 3c_{\text{eu}}\sqrt{S^2A^2H2^{-i}\log\frac{2SAHK_i}{\delta}\cdot 2^i 144c_{\text{eu}}^2S^2A^2H\log\frac{2SAHK_i}{\delta}}
\end{aligned}$$

$= 0.$

Thus, we can lower bound the above as

$$2^{-i}K_i - 3c_{\text{eu}}\sqrt{S^2A^2H2^{-i}K_i \log \frac{2SAHK_i}{\delta}} - C_{\mathcal{R}} \geq \frac{1}{2}2^{-i}K_i.$$

Note that we can collect a total reward of at most $|\mathcal{X}|N_i$. However, by our choice of $N_i = K_i/(4|\mathcal{X}| \cdot 2^i)$, we have that

$$|\mathcal{X}|N_i = \frac{1}{4 \cdot 2^i}K_i < \frac{1}{2 \cdot 2^i}K_i.$$

This is a contradiction. Thus, we must have that $V_0^{\star,im_i} \leq 1/2^i$. The second conclusion follows from this by definition of $V_0^{\star,im_i}$.

For the third conclusion, we can apply Lemma E.3.3. By construction, we will only add some (s, a, h) to $\mathcal{X}_i$ if we visit N_i times. It follows by Lemma E.3.3 that, with probability $1 - \delta_{\text{samp}}/(SAH)$, if we rerun all policies, we will collect at least

$$N_i - \sqrt{8K_i \max_k w_h^{\pi_k}(s, a) \log \frac{4SAH}{\delta_{\text{samp}}}} - \frac{4}{3}\log \frac{4SAH}{\delta_{\text{samp}}}$$

samples from (s, a, h). Note that $\max_k w_h^{\pi_k}(s, a) \leq 2^{-i}$ by our assumption on $\mathcal{X}$. Given our choice of N_i, we can then guarantee that we will collect at least

$$\frac{K_i}{4|\mathcal{X}|2^i} - \sqrt{\frac{8K_iH}{2^i} \log \frac{4SAH}{\delta_{\text{samp}}}} - \frac{4}{3}\log \frac{4SAH}{\delta_{\text{samp}}}$$

samples. Since $K_i \geq 2048S^2A^2 \log \frac{4SAH}{\delta_{\text{samp}}}$, and $|\mathcal{X}| \leq SA$, we will have that

$$\frac{K_i}{4|\mathcal{X}|2^i} - \sqrt{\frac{8K_iH}{2^i} \log \frac{4SAH}{\delta_{\text{samp}}}} - \frac{4}{3}\log \frac{4SAH}{\delta_{\text{samp}}} \geq \frac{K_i}{8|\mathcal{X}|2^i} = \frac{1}{2}N_i$$

The third conclusion follows by union bounding over every $(s, a, h) \in \mathcal{X}_i$. □

E.3.1 Technical Lemmas

Lemma E.3.2. *We will have that*

$$K_i(\delta, \delta_{\text{samp}}) \geq 2^i \max\left\{32c_{\text{eu}}S^3A^2H^4 \log^3 \frac{2SAHK_i(\delta, \delta_{\text{samp}})}{\delta} + 8H \log \frac{4}{\delta},\right.$$

$$144c_{\text{eu}}^2 S^2 A^2 H \log \frac{2SAHK_i(\delta)}{\delta} \Bigg\}.$$

Proof. Note that for any $i, j > 0$ and $C > 0$, if $x \geq C^i (i+3j)^j \log^j(C(i+3j))$, then $x \geq C^i \log^j x$ since

$$\begin{aligned} C^i \log^j x = C^i \log^j [C^i (i+3j)^j \log^j (C(i+3j))] &\leq C^i \log^j [C^{i+j}(i+3j)^{2j}] \\ &\leq C^i (i+3j)^j \log[C(i+3j)] \\ &= x \end{aligned}$$

and, furthermore, $\frac{d}{dy} y|_{y=C^{i+j}(\max\{i+j,2j\})^{2j}} = 1$, while

$$\frac{d}{dy} C^i \log^j y|_{y=C^{i+j}(\max\{i+j,2j\})^{2j}} = \frac{C^i \log^{j-1} y}{y}|_{y=C^{i+j}(\max\{i+j,2j\})^{2j}} \leq 1$$

and since the derivative of poly log functions decreases monotonically.

It follows that

$$K_i(\delta, \delta_{\text{samp}}) \geq 2^i \cdot 256 c_{\text{eu}} S^3 A^2 H^4 \log^3 K_i(\delta, \delta_{\text{samp}})$$

as long as

$$K_i(\delta, \delta_{\text{samp}}) \geq 2^i \cdot 256 c_{\text{eu}} S^3 A^2 H^4 (i+9)^3 \log^3 (512 c_{\text{eu}} SAH(i+9))$$

So

$$\begin{aligned} K_i(\delta, \delta_{\text{samp}}) &\geq 2 \cdot 2^i \max\{128 c_{\text{eu}} S^3 A^2 H^4 \log^3 K_i(\delta, \delta_{\text{samp}}), 128 c_{\text{eu}} S^3 A^2 H^4 \log^3 \frac{2SAH}{\delta} + 8H \log \frac{4}{\delta}\} \\ &\geq 2^i (32 c_{\text{eu}} S^3 A^2 H^4 \log^3 \frac{2SAHK_i(\delta, \delta_{\text{samp}})}{\delta} + 8H \log \frac{4}{\delta}) \end{aligned}$$

if

$$\begin{aligned} K_i(\delta, \delta_{\text{samp}}) \geq \max\{&2^i \cdot 256 c_{\text{eu}} S^3 A^2 H^4 (i+9)^3 \log^3 (512 c_{\text{eu}} SAH(i+9)), \\ &128 c_{\text{eu}} S^3 A^2 H^4 \log^3 \frac{2SAH}{\delta} + 8H \log \frac{4}{\delta}\}. \end{aligned}$$

Similarly,

$$K_i(\delta, \delta_{\text{samp}}) \geq 2^i \cdot 144 c_{\text{eu}}^2 S^2 A^2 H \log \frac{2SAHK_i}{\delta}$$

if

$$K_i(\delta, \delta_{\text{samp}}) \geq \max\{2^i \cdot 288 c_{\text{eu}}^2 S^2 A^2 H(i+3) \log(576 c_{\text{eu}} SAH(i+3)), 288 c_{\text{eu}}^2 S^2 A^2 H \log \frac{2SAH}{\delta}\}.$$

The result then follows recalling the definition of $K_i(\delta, \delta_{\text{samp}})$ given in (E.3.1). □

Lemma E.3.3. *Consider a set of policies $\{\pi_k\}_{k=1}^K$. Assume that running each of these policies once, we collect at least N samples from some (s, a, h). Then, if we rerun each of these policies once, we will collect, with probability $1-\delta$, at least*

$$N - \sqrt{8K \max_k w_h^{\pi_k}(s,a) \log 4/\delta} - 4/3 \log 4/\delta$$

samples from (s, a, h).

Proof. Note that when running π_k, the expected number of visits to (s, a, h) is $w_h^{\pi_k}(s, a)$. By Bernstein's inequality, and using that $\mathbb{I}\{(s_h^k, a_h^k) = (s, a)\} \sim \text{Bernoulli}(w_h^{\pi_k}(s, a))$, we then have that, with probability at least $1-\delta$,

$$\left| \sum_{k=1}^K w_h^{\pi_k}(s,a) - \sum_{k=1}^K \mathbb{I}\{(s_h^k, a_h^k) = (s,a)\} \right| \leq \sqrt{2K \max_k w_h^{\pi_k}(s,a) \log 2/\delta} + 2/3 \log 2/\delta$$

As our first draw from the policies yielded a value of at least N, we can apply Proposition E.5, which gives that, with probability at least $1-2\delta$,

$$\sum_{k=1}^K \mathbb{I}\{(s_h^k, a_h^k) = (s,a)\} \geq N - 2\sqrt{2K \max_k w_h^{\pi_k}(s,a) \log 2/\delta} - 4/3 \log 2/\delta$$

The result follows. □

Lemma E.3.4 (Lemma 3.4 of Jin et al. [2020a])**.** *If r_h^k is non-zero for at most one h per episode, the regret of* Euler *[Zanette and Brunskill, 2019] will be bounded, with probability at least $1-\delta$, as*

$$\sum_{k=1}^K V_0^\star - \sum_{k=1}^K V_0^{\pi_k} \leq c_{\text{eu}} \sqrt{SAHV_0^\star K \log \frac{SAHK}{\delta}} + c_{\text{eu}} S^2 AH^4 \log^3 \frac{SAHK}{\delta}$$

for some absolute constant c_{eu}.

Proof. The proof of this is identical to the proof of Lemma 3.4 in Jin et al. [2020a] but we include it for completeness. We therefore repeat their analysis, using an alternative upper

bound for equation (156) in Zanette and Brunskill [2019]:

$$\begin{aligned}\frac{1}{KH}\sum_{k=1}^{K}\mathbb{E}_{\pi_k}\left[(\sum_{h=1}^{H}r_h^k - V_0^{\pi_k})^2\right] &\le \frac{2}{KH}\sum_{k=1}^{K}\mathbb{E}_{\pi_k}\left[(\sum_{h=1}^{H}r_h^k)^2 + (V_0^{\pi_k})^2\right] \\ &\overset{(a)}{\le} \frac{2}{KH}\sum_{k=1}^{K}\mathbb{E}_{\pi_k}\left[\sum_{h=1}^{H}(r_h^k)^2 + V_0^{\pi_k}\right] \\ &\overset{(b)}{\le} \frac{2}{KH}\sum_{k=1}^{K}\mathbb{E}_{\pi_k}\left[\sum_{h=1}^{H}r_h^k + V_0^{\pi_k}\right] \\ &= \frac{4}{KH}\sum_{k=1}^{K}V_0^{\pi_k} \\ &\le 4V_0^\star/H\end{aligned}$$

where (a) follows since r_h^k is nonzero for at most one h and (b) follows since $r_h^k \le 1$. Thus, we can replace $\mathcal{G}^2$ in Theorem 1 of Zanette and Brunskill [2019] with $4V_0^\star$. As Zanette and Brunskill [2019] assume a stationary MDP while ours is non-stationary, we must replace S in their bound with SH. This gives the result.

□

Lemma E.3.5. *Consider some set of policies $\{\pi_k\}_{k=1}^K$ where π_k is $\mathcal{F}_{k-1}$ measurable. Let $\sum_{h=1}^H R_h^k$ denote the (random) reward obtained running π_k on the MDP M_k, and let V_0^k denote the value function of running π_k on M_k. Assume that*

$$\mathbb{E}_{\pi_k}[(\sum_{h=1}^{H}R_h^k - V_0^k)^2|\mathcal{F}_{k-1}] \le \sigma_V^2$$

for all k and constant σ_V^2 which is $\mathcal{F}_0$-measurable. Then, with probability at least $1-\delta$,

$$\left|\sum_{k=1}^{K}\sum_{h=1}^{H}R_h^k - \sum_{k=1}^{K}V_0^k\right| \le \sqrt{8K\sigma_V^2\log\frac{2}{\delta}} + 2H\log\frac{2}{\delta}.$$

Proof. By definition, $V_0^k = \mathbb{E}[\sum_{h=1}^H R_h^k|\mathcal{F}_{k-1}]$ and $|\sum_{h=1}^H R_h^k - V_0^k| \le H$ almost surely. The result then follows directly from Freedman's Inequality [Freedman, 1975].

□

Proposition E.5. *Consider some distribution $\mathbf{P}$ and assume that $\mathbb{P}_{x\sim\mathbf{P}}[x \in [\mu-c, \mu+c]] \ge 1-\delta$. Then $\mathbb{P}_{x,x'\overset{\text{i.i.d.}}{\sim}\mathbf{P}}[x \ge x' - 2c] \ge 1-2\delta$.*

Proof.

$$\begin{aligned}
\mathbb{P}_{x,x' \overset{\text{i.i.d.}}{\sim} \mathbf{P}}[x \geq x' - 2c] &= \mathbb{P}_{x,x' \overset{\text{i.i.d.}}{\sim} \mathbf{P}}[x' - \mu + \mu - x \leq 2c] \\
&\geq \mathbb{P}_{x,x' \overset{\text{i.i.d.}}{\sim} \mathbf{P}}[|x' - \mu| + |\mu - x| \leq 2c] \\
&\geq \mathbb{P}_{x \overset{\text{i.i.d.}}{\sim} \mathbf{P}}[|\mu - x| \leq c] \mathbb{P}_{x' \overset{\text{i.i.d.}}{\sim} \mathbf{P}}[|x' - \mu| \leq c] \\
&\geq (1 - \delta)^2 \\
&\geq 1 - 2\delta.
\end{aligned}$$

□

Appendix F

Deferred Proofs from Chapter 8

F.1 Properties of Linear MDPs

Lemma F.1.1. *For any linear MDP satisfying Definition 8.1.1, we must have that* $\|\boldsymbol{\phi}(s,a)\|_2 \geq 1/\sqrt{d}$ *for all* s *and* a*, and* $\|\boldsymbol{\phi}_{\pi,h}\|_2 \geq 1/\sqrt{d}$ *for all* π *and* h*.*

Proof. By Definition 8.1.1, we know that $P_h(\cdot|s,a) = \langle \boldsymbol{\phi}(s,a), \boldsymbol{\mu}_h(\cdot)\rangle$ forms a valid probability distribution, and that $\|\int_{\mathcal{S}} |\mathrm{d}\boldsymbol{\mu}_h(s)|\|_2 \leq \sqrt{d}$. It follows that

$$1 = \int_{\mathcal{S}} \langle \boldsymbol{\phi}(s,a), \mathrm{d}\boldsymbol{\mu}_h(s)\rangle \leq \|\boldsymbol{\phi}(s,a)\|_2 \| \int_{\mathcal{S}} |\mathrm{d}\boldsymbol{\mu}_h(s)|\|_2 \leq \sqrt{d}\|\boldsymbol{\phi}(s,a)\|_2$$

from which the first result follows.

For the second result, using that $1 = \int_{\mathcal{S}} \langle \boldsymbol{\phi}(s,a), \mathrm{d}\boldsymbol{\mu}_h(s)\rangle$, we get

$$\begin{aligned}
\int_{\mathcal{S}} \langle \boldsymbol{\phi}_{\pi,h}, \mathrm{d}\boldsymbol{\mu}_h(s)\rangle &= \int_{\mathcal{S}} \langle \mathbb{E}_\pi[\boldsymbol{\phi}_h], \mathrm{d}\boldsymbol{\mu}_h(s)\rangle \\
&= \mathbb{E}_\pi\left[\int_{\mathcal{S}} \langle \boldsymbol{\phi}_h, \mathrm{d}\boldsymbol{\mu}_h(s)\rangle\right] \\
&= \mathbb{E}_\pi[1] \\
&= 1
\end{aligned}$$

where we can exchange the order of integration by Fubini's Theorem since the integrand is absolutely integrable, by Definition 8.1.1. As above, we then have

$$1 = \int_{\mathcal{S}} \langle \boldsymbol{\phi}_{\pi,h}, \mathrm{d}\boldsymbol{\mu}_h(s)\rangle \leq \sqrt{d}\|\boldsymbol{\phi}_{\pi,h}\|_2$$

so the second result follows. □

F.1.1 Feature-Visitations in Linear MDPs

Lemma F.1.2. $\boldsymbol{\phi}_{\pi,h} = \mathcal{T}_{\pi,h}\boldsymbol{\phi}_{\pi,h-1} = \ldots = \mathcal{T}_{\pi,h}\ldots\mathcal{T}_{\pi,1}\boldsymbol{\phi}_{\pi,0}$.

Proof. By the linear MDP assumption, we have:

$$\begin{aligned}
\boldsymbol{\phi}_{\pi,h} &= \mathbb{E}_\pi[\boldsymbol{\phi}(s_h, a_h)] \\
&= \mathbb{E}_\pi[\mathbb{E}[\boldsymbol{\phi}(s_h, a_h)|\mathcal{F}_{h-1}]] \\
&= \mathbb{E}_\pi[\int \int \boldsymbol{\phi}(s,a)\mathrm{d}\pi_h(a|s)\mathrm{d}\boldsymbol{\mu}_{h-1}(s)^\top \boldsymbol{\phi}(s_{h-1}, a_{h-1})] \\
&= \mathbb{E}_\pi[\int \boldsymbol{\phi}_{\pi,h}(s)\mathrm{d}\boldsymbol{\mu}_{h-1}(s)^\top \boldsymbol{\phi}(s_{h-1}, a_{h-1})] \\
&= \int \boldsymbol{\phi}_{\pi,h}(s)\mathrm{d}\boldsymbol{\mu}_{h-1}(s)^\top \mathbb{E}_\pi[\boldsymbol{\phi}(s_{h-1}, a_{h-1})] \\
&= \mathcal{T}_{\pi,h}\boldsymbol{\phi}_{\pi,h-1}.
\end{aligned}$$

This yields the first equality. Repeating this calculation $h-1$ more times yields the final equality. □

Lemma F.1.3. *Fix some h and $i < h$, and consider the vector*

$$\boldsymbol{v} := \mathcal{T}_{\pi,i+1}^\top \mathcal{T}_{\pi,i+2}^\top \dots \mathcal{T}_{\pi,h-1}^\top \mathcal{T}_{\pi,h}^\top \boldsymbol{u}.$$

Assume that either $\boldsymbol{u} = \boldsymbol{\theta}_h$ for some $\boldsymbol{\theta}_h$ which is a valid reward vector as defined in Definition 8.1.1, or $\boldsymbol{u} \in \mathcal{S}^{d-1}$. In either case, we have that, for any s, a, $|\boldsymbol{v}^\top \boldsymbol{\phi}(s,a)| \le 1$, and $\|\boldsymbol{v}\|_2 \le \sqrt{d}$.

Proof. By the linear MDP structure (see Proposition 2.3 of Jin et al. [2020b]), for any j,

$$\begin{aligned}
Q_j^\pi(s,a) &= \langle \boldsymbol{\phi}(s,a), \boldsymbol{w}_j^\pi \rangle \\
&= \langle \boldsymbol{\phi}(s,a), \boldsymbol{\theta}_j \rangle + \int V_{j+1}^\pi(s')\mathrm{d}\boldsymbol{\mu}_j(s')^\top \boldsymbol{\phi}(s,a) \\
&= \langle \boldsymbol{\phi}(s,a), \boldsymbol{\theta}_j \rangle + \int \langle \boldsymbol{w}_{j+1}^\pi, \boldsymbol{\phi}_{j+1,\pi}(s') \rangle \mathrm{d}\boldsymbol{\mu}_j(s')^\top \boldsymbol{\phi}(s,a) \\
&= \langle \boldsymbol{\phi}(s,a), \boldsymbol{\theta}_j + \mathcal{T}_{\pi,j+1}^\top \boldsymbol{w}_{j+1}^\pi \rangle
\end{aligned}$$

so in general,

$$\boldsymbol{w}_i^\pi = \sum_{h'=i}^{H} \Big(\prod_{j=i+1}^{h'} \mathcal{T}_{\pi,j}^\top \Big) \boldsymbol{\theta}_{h'}$$

where we order the product $\prod_{j=i+1}^{h'} \mathcal{T}_{\pi,j}^\top = \mathcal{T}_{\pi,i+1}^\top \mathcal{T}_{\pi,i+1}^\top \dots \mathcal{T}_{\pi,h'}^\top$.

Case 1: $\boldsymbol{u} = \boldsymbol{\theta}_h$. We first consider the case where $\boldsymbol{u} = \boldsymbol{\theta}_h$ for some $\boldsymbol{\theta}_h$ which is a valid reward satisfying Definition 8.1.1. Assume that the reward in our MDP is set such that for

$h' \neq h$, $\boldsymbol{\theta}_{h'} = 0$. In this case, we then have that

$$\boldsymbol{w}_i^\pi = \mathcal{T}_{\pi,i+1}^\top \mathcal{T}_{\pi,i+2}^\top \dots \mathcal{T}_{\pi,h}^\top \boldsymbol{\theta}_h = \boldsymbol{v}.$$

In this case, we know that the trajectory rewards are always bounded by 1, so it follows that $Q_i^\pi(s,a) \leq 1$. Thus,

$$1 \geq Q_i^\pi(s,a) = \langle \boldsymbol{\phi}(s,a), \boldsymbol{w}_i^\pi \rangle = \langle \boldsymbol{\phi}(s,a), \boldsymbol{v} \rangle$$

and this holds for any s, a. Since Q-values are always positive, it also holds that $\langle \boldsymbol{\phi}(s,a), \boldsymbol{v} \rangle \geq 0$.

To bound the norm of $\boldsymbol{v}$, we note that by the Bellman equation and the calculation above,

$$\begin{aligned}
\|\boldsymbol{v}\|_2 = \|\boldsymbol{w}_i^\pi\|_2 &= \|\boldsymbol{\theta}_i + \int V_{i+1}^\pi(s') \mathrm{d}\boldsymbol{\mu}_i(s')\|_2 \\
&\leq \|\boldsymbol{\theta}_i\|_2 + \| \int |V_{i+1}^\pi(s')| \mathrm{d}\boldsymbol{\mu}_i(s')\|_2 \\
&\leq \| \int |\mathrm{d}\boldsymbol{\mu}_i(s')|\|_2 \\
&\leq \sqrt{d}
\end{aligned}$$

where we have used that $|V_{i+1}^\pi(s')| \leq 1$ since the total episode return is at most 1 on our augmented reward function, and the linear MDP assumption.

Case 2: $\boldsymbol{u} \in \mathcal{S}^{d-1}$. We can repeat the argument above in the case where we only assume $\boldsymbol{u} \in \mathcal{S}^{d-1}$. Since $\|\boldsymbol{\phi}(s,a)\|_2 \leq 1$, it follows that with the reward vector at level h set to $\boldsymbol{u}$, the reward will still be bounded in $[-1,1]$. Thus, essentially the same argument can be used, with the slight modification to handle Q-values that are negative. □

Lemma F.1.4. *The set $\boldsymbol{\Omega}_h$ is convex and compact.*

Proof. Take $\boldsymbol{\Lambda}_1, \boldsymbol{\Lambda}_2 \in \boldsymbol{\Omega}_h$. By definition, $\boldsymbol{\Lambda}_1 = \mathbb{E}_{\pi \sim \omega_1}[\boldsymbol{\Lambda}_{\pi,h}], \boldsymbol{\Lambda}_2 = \mathbb{E}_{\pi \sim \omega_2}[\boldsymbol{\Lambda}_{\pi,h}]$. It follows that, for any $t \in [0,1]$, $t\boldsymbol{\Lambda}_1 + (1-t)\boldsymbol{\Lambda}_2 = \mathbb{E}_{\pi \sim t\omega_1 + (1-t)\omega_2}[\boldsymbol{\Lambda}_{\pi,h}]$. For $t\omega_1 + (1-t)\omega_2$ the mixture of ω_1 and ω_2. As $t\omega_1 + (1-t)\omega_2$ is a valid mixture over policies, it follows that $t\boldsymbol{\Lambda}_1 + (1-t)\boldsymbol{\Lambda}_2 \in \boldsymbol{\Omega}_h$, which proves convexity.

Compactness follows since $\|\boldsymbol{\phi}(s,a)\|_2 \leq 1$ for all s, a, so $\|\boldsymbol{\Lambda}_{\pi,h}\|_{\mathrm{op}} \leq 1$, which implies $\|\boldsymbol{\Lambda}\|_{\mathrm{op}} \leq 1$ for any $\boldsymbol{\Lambda} \in \boldsymbol{\Omega}_h$. Furthermore, the set $\boldsymbol{\Omega}_h$ is clearly closed, which proves compactness.

□

F.1.2 Constructing the Policy Class

Lemma F.1.5 (Lemma B.1 of Jin et al. [2020b]). *Let $\boldsymbol{w}_h^\pi$ denote the set of weights such that $Q_h^\pi(s,a) = \langle \boldsymbol{\phi}(s,a), \boldsymbol{w}_h^\pi \rangle$. Then $\|\boldsymbol{w}_h^\pi\|_2 \le 2H\sqrt{d}$.*

Lemma F.1.6. *For any $\delta > 0$ there exists sets of actions $(\widetilde{\mathcal{A}}_s)_{s\in\mathcal{S}}$, $\widetilde{\mathcal{A}}_s \subseteq \mathcal{A}$, such that $|\widetilde{\mathcal{A}}_s| \le (1 + 8H\sqrt{d}/\delta)^d$ for all s and, for all $a \in \mathcal{A}$, s, h, and any π, there exists some $\widetilde{a} \in \widetilde{\mathcal{A}}_s$ such that*

$$|Q_h^\pi(s,a) - Q_h^\pi(s,\widetilde{a})| \le \delta, \quad |r_h(s,a) - r_h(s,\widetilde{a})| \le \delta.$$

Proof. Let $\mathcal{N}$ be a $\delta/(4H\sqrt{d})$ cover of the unit ball. By Lemma F.1.10 we can bound $|\mathcal{N}| \le (1 + 8H\sqrt{d}/\delta)^d$. Take any s and let $\widetilde{\mathcal{A}}_s = \emptyset$. Then for each $\boldsymbol{\phi} \in \mathcal{N}$, choose any a at random from the set $\{a \in \mathcal{A} \ : \ \|\boldsymbol{\phi}(s,a) - \boldsymbol{\phi}\|_2 \le \delta/2\}$ and set $\widetilde{\mathcal{A}}_s \leftarrow \widetilde{\mathcal{A}}_s \cup \{a\}$. With this construction, we claim that for all $a \in \mathcal{A}$, there exists some $\widetilde{a} \in \widetilde{\mathcal{A}}_s$ such that $\|\boldsymbol{\phi}(s,a) - \boldsymbol{\phi}(s,\widetilde{a})\|_2 \le \delta/(2H\sqrt{d})$. To see why this is, note that by construction of $\mathcal{N}$, there always exists some $\boldsymbol{\phi} \in \mathcal{N}$ such that $\|\boldsymbol{\phi}(s,a) - \boldsymbol{\phi}\|_2 \le \delta/(4H\sqrt{d})$. Since $\widetilde{\mathcal{A}}_s$ will contain some $\widetilde{a}$ such that $\|\boldsymbol{\phi}(s,\widetilde{a}) - \boldsymbol{\phi}\|_2 \le \delta/(4H\sqrt{d})$, the claim follows by the triangle inequality.

By Lemma F.1.5, we have that for any π, $\|\boldsymbol{w}_h^\pi\|_2 \le 2H\sqrt{d}$. Take $a \in \mathcal{A}$ and let $\widetilde{a} \in \widetilde{\mathcal{A}}_s$ be the action such that $\|\boldsymbol{\phi}(s,a) - \boldsymbol{\phi}(s,\widetilde{a})\|_2 \le \delta/(2H\sqrt{d})$. Then

$$|Q_h^\pi(s,a) - Q_h^\pi(s,\widetilde{a})| = |\langle \boldsymbol{\phi}(s,a) - \boldsymbol{\phi}(s,\widetilde{a}), w_h^\pi \rangle| \le 2H\sqrt{d}\|\boldsymbol{\phi}(s,a) - \boldsymbol{\phi}(s,\widetilde{a})\|_2 \le \delta.$$

The bound on $|r_h(s,a) - r_h(s,\widetilde{a})|$ follows analogously, since we assume our rewards are linear, and that $\|\boldsymbol{\theta}_h\|_2 \le \sqrt{d}$. □

Definition F.1.1 (Linear Softmax Policy). We say a policy is a *linear softmax policy* with parameters η and $\{\boldsymbol{w}_h\}_{h=1}^H$ if it can be written as

$$\pi_h(a|s) = \frac{e^{\eta\langle \boldsymbol{\phi}(s,a), \boldsymbol{w}_h \rangle}}{\sum_{a' \in \mathcal{A}} e^{\eta\langle \boldsymbol{\phi}(s,a'), \boldsymbol{w}_h \rangle}}$$

for some $\boldsymbol{w} = \{\boldsymbol{w}_h\}_{h=1}^H$. We will denote such a policy as $\pi^{\boldsymbol{w}}$.

Definition F.1.2 (Restricted-Action Linear Softmax Policy). We say a policy is a *restricted-action linear softmax policy* with parameters η, $\{\boldsymbol{w}_h\}_{h=1}^H$, and $(\widetilde{\mathcal{A}}_s)_{s\in\mathcal{S}}$ if it can be written as

$$\widetilde{\pi}_h(a|s) = \frac{e^{\eta\langle \boldsymbol{\phi}(s,a), \boldsymbol{w}_h \rangle} \cdot \mathbb{I}\{a \in \widetilde{\mathcal{A}}_s\}}{\sum_{a' \in \widetilde{\mathcal{A}}_s} e^{\eta\langle \boldsymbol{\phi}(s,a'), \boldsymbol{w}_h \rangle}}$$

for some $\boldsymbol{w} = \{\boldsymbol{w}_h\}_{h=1}^H$. We will denote such a policy as $\widetilde{\pi}^{\boldsymbol{w}}$.

Lemma F.1.7. *For any restricted-action linear softmax policies $\pi^{\boldsymbol{w}}$ and $\pi^{\boldsymbol{u}}$ with identical restricted sets $(\widetilde{\mathcal{A}}_s)_{s\in\mathcal{S}}$, we can bound*

$$|V_0^{\pi^{\boldsymbol{w}}}(s_1) - V_0^{\pi^{\boldsymbol{u}}}(s_1)| \leq 2dH\eta \sum_{h=1}^{H} \|\boldsymbol{w}_h - \boldsymbol{u}_h\|_2.$$

Proof. Note that for any policy π, the value of the policy can be expressed as

$$V_0^{\pi}(s_1) = \sum_{h=1}^{H} \langle \boldsymbol{\theta}_h, \boldsymbol{\phi}_{\pi,h} \rangle.$$

Thus,

$$|V_0^{\pi^{\boldsymbol{w}}}(s_1) - V_0^{\pi^{\boldsymbol{u}}}(s_1)| \leq \sum_{h=1}^{H} |\langle \boldsymbol{\theta}_h, \boldsymbol{\phi}_{\pi^{\boldsymbol{w}},h} - \boldsymbol{\phi}_{\pi^{\boldsymbol{u}},h} \rangle|.$$

So it suffices to bound $|\langle \boldsymbol{\theta}_h, \boldsymbol{\phi}_{\pi^{\boldsymbol{w}},h} - \boldsymbol{\phi}_{\pi^{\boldsymbol{u}},h} \rangle|$. Using the same decomposition as in the proof of Lemma 8.3.2, we have

$$\boldsymbol{\phi}_{\pi^{\boldsymbol{w}},h} - \boldsymbol{\phi}_{\pi^{\boldsymbol{u}},h} = \sum_{i=0}^{h-1} \left(\prod_{j=h-i+1}^{h} \mathcal{T}_{\pi^{\boldsymbol{w}},j} \right) (\mathcal{T}_{\pi^{\boldsymbol{w}},h-i} - \mathcal{T}_{\pi^{\boldsymbol{u}},h-i}) \boldsymbol{\phi}_{\pi^{\boldsymbol{u}},h-i-1}.$$

By definition,

$$\mathcal{T}_{\pi^{\boldsymbol{w}},h-i} - \mathcal{T}_{\pi^{\boldsymbol{u}},h-i} = \int (\boldsymbol{\phi}_{\pi^{\boldsymbol{w}},h-i}(s) - \boldsymbol{\phi}_{\pi^{\boldsymbol{u}},h-i}(s)) \mathrm{d}\boldsymbol{\mu}_{h-i-1}(s)^{\top}$$

where

$$\boldsymbol{\phi}_{\pi^{\boldsymbol{w}},h-i}(s) = \sum_{a\in\widetilde{\mathcal{A}}_s} \boldsymbol{\phi}(s,a) \pi_{h-i}^{\boldsymbol{w}}(a|s).$$

Our goal is to bound $\|\boldsymbol{\phi}_{\pi^{\boldsymbol{w}},h-i}(s) - \boldsymbol{\phi}_{\pi^{\boldsymbol{u}},h-i}(s)\|_2$. Let $\boldsymbol{w}^t = t\boldsymbol{w} + (1-t)\boldsymbol{u}$, and $\boldsymbol{\phi}_{t,h-i}(s) := \boldsymbol{\phi}_{\pi^{\boldsymbol{w}^t},h-i}(s)$. Then $\|\boldsymbol{\phi}_{\pi^{\boldsymbol{w}},h-i}(s) - \boldsymbol{\phi}_{\pi^{\boldsymbol{u}},h-i}(s)\|_2 = \|\boldsymbol{\phi}_{1,h-i}(s) - \boldsymbol{\phi}_{0,h-i}(s)\|_2$. By the Mean Value Theorem, we can bound

$$\|\boldsymbol{\phi}_{1,h-i}(s) - \boldsymbol{\phi}_{0,h-i}(s)\|_2 \leq \max_{t'\in[0,1]} \|\frac{\mathrm{d}}{\mathrm{d}t} \boldsymbol{\phi}_{t,h-i}(s)|_{t=t'}\|_2 \cdot |1-0|.$$

Some calculation shows that

$$\frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{\phi}_{t,h-i}(s) = \frac{1}{(\sum_{a'\in\widetilde{\mathcal{A}}_s} e^{\eta\langle\boldsymbol{\phi}(s,a'),\boldsymbol{w}_h^t\rangle})^2}\Bigg[\sum_{a\in\widetilde{\mathcal{A}}_s}\eta\boldsymbol{\phi}(s,a)\Big[e^{\eta\langle\boldsymbol{\phi}(s,a),\boldsymbol{w}_h^t\rangle}\langle\boldsymbol{\phi}(s,a),\boldsymbol{w}_{h-i}-\boldsymbol{u}_{h-i}\rangle\cdot\sum_{a'\in\widetilde{\mathcal{A}}_s}e^{\eta\langle\boldsymbol{\phi}(s,a'),\boldsymbol{w}_h^t\rangle}$$
$$-e^{\eta\langle\boldsymbol{\phi}(s,a),\boldsymbol{w}_h^t\rangle}\cdot\sum_{a'\in\widetilde{\mathcal{A}}_s}e^{\eta\langle\boldsymbol{\phi}(s,a'),\boldsymbol{w}_h^t\rangle}\langle\boldsymbol{\phi}(s,a'),\boldsymbol{w}_{h-i}-\boldsymbol{u}_{h-i}\rangle\Big]\Bigg].$$

Since $\|\boldsymbol{\phi}(s,a)\|_2 \le 1$, we can then bound

$$\begin{aligned}\|\frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{\phi}_{t,h-i}(s)\|_2 &\le \frac{2\eta}{(\sum_{a'\in\widetilde{\mathcal{A}}_s} e^{\eta\langle\boldsymbol{\phi}(s,a'),\boldsymbol{w}_h^t\rangle})^2}\Bigg[\sum_{a\in\widetilde{\mathcal{A}}_s}e^{\eta\langle\boldsymbol{\phi}(s,a),\boldsymbol{w}_h^t\rangle}\cdot\sum_{a'\in\widetilde{\mathcal{A}}_s}e^{\eta\langle\boldsymbol{\phi}(s,a'),\boldsymbol{w}_h^t\rangle}\Bigg]\cdot\|\boldsymbol{w}_{h-i}-\boldsymbol{u}_{h-i}\|_2\\ &= 2\eta\cdot\|\boldsymbol{w}_{h-i}-\boldsymbol{u}_{h-i}\|_2.\end{aligned}$$

It follows that

$$\|\boldsymbol{\phi}_{\pi^{\boldsymbol{w}},h-i}(s)-\boldsymbol{\phi}_{\pi^{\boldsymbol{u}},h-i}(s)\|_2 \le 2\eta\cdot\|\boldsymbol{w}_{h-i}-\boldsymbol{u}_{h-i}\|_2.$$

With Definition 8.1.1, this implies that

$$\|\mathcal{T}_{\pi^{\boldsymbol{w}},h-i}-\mathcal{T}_{\pi^{\boldsymbol{u}},h-i}\|_{\mathrm{op}} \le \int\|\boldsymbol{\phi}_{\pi^{\boldsymbol{w}},h-i}(s)-\boldsymbol{\phi}_{\pi^{\boldsymbol{u}},h-i}(s)\|_2\|\mathrm{d}\boldsymbol{\mu}_{h-i-1}(s)\|_2 \le 2\sqrt{d}\eta\|\boldsymbol{w}_{h-i}-\boldsymbol{u}_{h-i}\|_2.$$

By Lemma F.1.3, we can bound $\|\boldsymbol{\theta}_h^\top\left(\prod_{j=h-i+1}^{h}\mathcal{T}_{\pi^{\boldsymbol{w}},j}\right)\|_2$. Thus, returning to the error decomposition given above, we have

$$\begin{aligned}|V_0^{\pi^{\boldsymbol{w}}}(s_1)-V_0^{\pi^{\boldsymbol{u}}}(s_1)| &\le \sum_{h=1}^{H}\sum_{i=0}^{h-1}\left|\boldsymbol{\theta}_h^\top\left(\prod_{j=h-i+1}^{h}\mathcal{T}_{\pi^{\boldsymbol{w}},j}\right)(\mathcal{T}_{\pi^{\boldsymbol{w}},h-i}-\mathcal{T}_{\pi^{\boldsymbol{u}},h-i})\boldsymbol{\phi}_{\pi^{\boldsymbol{u}},h-i-1}\right|\\ &\le \sqrt{d}\sum_{h=1}^{H}\sum_{i=0}^{h-1}\|\mathcal{T}_{\pi^{\boldsymbol{w}},h-i}-\mathcal{T}_{\pi^{\boldsymbol{u}},h-i}\|_{\mathrm{op}}\|\boldsymbol{\phi}_{\pi^{\boldsymbol{u}},h-i-1}\|_2\\ &\le 2d\eta\sum_{h=1}^{H}\sum_{i=0}^{h-1}\|\boldsymbol{w}_{h-i}-\boldsymbol{u}_{h-i}\|_2\\ &\le 2dH\eta\sum_{h=1}^{H}\|\boldsymbol{w}_h-\boldsymbol{u}_h\|_2.\end{aligned}$$

□

Lemma F.1.8. *Let $\boldsymbol{w}^\star$ denote the weights such that $Q_h^\star(s,a)=\langle\boldsymbol{\phi}(s,a),\boldsymbol{w}_h^\star\rangle$, and $\pi^{\boldsymbol{w}^\star}$ the restricted-action linear softmax policy with action sets $(\widetilde{\mathcal{A}}_s)_{s\in\mathcal{S}}$ as defined in Lemma F.1.6*

with $\delta = \frac{\epsilon}{2H \cdot e^2}$. *Then*

$$|V_0^{\pi^{\boldsymbol{w}^\star}}(s_1) - V_0^\star(s_1)| \le \epsilon$$

as long as $\eta \ge 30dH \log(1 + 120Hd/\epsilon) \cdot \frac{1}{\epsilon}$.

Proof. We prove this by induction. Assume that at step h, for all s, we have $|V_h^\star(s) - V_h^{\pi^{\boldsymbol{w}^\star}}(s)| \le \delta_h$ for some δ_h. Then,

$$\begin{aligned}
|Q_{h-1}^{\pi^{\boldsymbol{w}^\star}}(s,a) - Q_{h-1}^\star(s,a)| &= \left| \int (V_h^{\pi^{\boldsymbol{w}^\star}}(s') - V_h^\star(s')) \mathrm{d}P_{h-1}(s' \mid s,a) \right| \\
&\le \int |V_h^{\pi^{\boldsymbol{w}^\star}}(s') - V_h^\star(s')| \mathrm{d}P_{h-1}(s' \mid s,a) \\
&\le \delta_h \cdot \int \mathrm{d}P_{h-1}(s' \mid s,a) \\
&\le \delta_h.
\end{aligned}$$

Thus,

$$\begin{aligned}
V_{h-1}^{\pi^{\boldsymbol{w}^\star}}(s) &= \frac{\sum_{a \in \widetilde{\mathcal{A}}_s} e^{\eta \langle \boldsymbol{\phi}(s,a), \boldsymbol{w}_{h-1}^\star \rangle} Q_{h-1}^{\pi^{\boldsymbol{w}^\star}}(s,a)}{\sum_{a \in \widetilde{\mathcal{A}}_s} e^{\eta \langle \boldsymbol{\phi}(s,a), \boldsymbol{w}_{h-1}^\star \rangle}} \\
&= \frac{\sum_{a \in \widetilde{\mathcal{A}}_s} e^{\eta Q_{h-1}^\star(s,a)} Q_{h-1}^{\pi^{\boldsymbol{w}^\star}}(s,a)}{\sum_{a \in \widetilde{\mathcal{A}}_s} e^{\eta Q_{h-1}^\star(s,a)}} \\
&\ge \frac{\sum_{a \in \widetilde{\mathcal{A}}_s} e^{\eta Q_{h-1}^\star(s,a)} Q_{h-1}^\star(s,a)}{\sum_{a \in \widetilde{\mathcal{A}}_s} e^{\eta Q_{h-1}^\star(s,a)}} - \delta_h.
\end{aligned}$$

By Lemma F.1.11, as long as $\eta \ge H \log |\widetilde{\mathcal{A}}_s| / \delta_h$, we can lower bound

$$\frac{\sum_{a \in \widetilde{\mathcal{A}}_s} e^{\eta Q_{h-1}^\star(s,a)} Q_{h-1}^\star(s,a)}{\sum_{a \in \widetilde{\mathcal{A}}_s} e^{\eta Q_{h-1}^\star(s,a)}} - \delta_h \ge \max_{a \in \widetilde{\mathcal{A}}_s} Q_{h-1}^\star(s,a) - (1 + 1/H)\delta_h.$$

Furthermore, by Lemma F.1.6 and our choice of $\widetilde{\mathcal{A}}_s$, we have

$$\begin{aligned}
\max_{a \in \widetilde{\mathcal{A}}_s} Q_{h-1}^\star(s,a) - (1 + 1/H)\delta_h &\ge \max_{a \in \mathcal{A}} Q_{h-1}^\star(s,a) - (1 + 1/H)\delta_h - \frac{\epsilon}{2H \cdot e^2} \\
&= V_{h-1}^\star(s) - (1 + 1/H)\delta_h - \frac{\epsilon}{2H \cdot e^2}.
\end{aligned}$$

Define recursively $\delta_{h-1} = (1+2/H)\delta_h$ and $\delta_H = \frac{\epsilon}{e^2}$. Then $\delta_h = (1+2/H)^{H-h}\frac{\epsilon}{e^2} \geq \frac{\epsilon}{e^2}$, so

$$V^\star_{h-1}(s) - (1+1/H)\delta_h - \frac{\epsilon}{2H\cdot e^2} \geq V^\star_{h-1}(s) - (1+1/H)\delta_h - \delta_h/H = V^\star_{h-1}(s) - \delta_{h-1}.$$

So $|V^\star_h(s) - V^{\pi^{\boldsymbol{w}^\star}}_h(s)| \leq \delta_{h-1}$ for all s, which proves the inductive step.

For the base case, we have

$$\begin{aligned} V^{\pi^{\boldsymbol{w}^\star}}_H(s) - V^\star_H(s) &= \frac{\sum_{a\in\widetilde{\mathcal{A}}_s} e^{\eta Q^\star_H(s,a)}\nu_H(s,a)}{\sum_{a\in\widetilde{\mathcal{A}}_s} e^{\eta Q^\star_H(s,a)}} - \max_a \nu_H(s,a) \\ &\geq \max_{a\in\widetilde{\mathcal{A}}_s} \nu_H(s,a) - \max_a \nu_H(s,a) - \delta_H/2 \\ &\geq -\delta_H \end{aligned}$$

where the first inequality holds by Lemma F.1.11 as long as $\eta \geq 2\log|\widetilde{\mathcal{A}}_s|/\delta_H$, and the second inequality holds by Lemma F.1.6 and our choice of $\widetilde{\mathcal{A}}_s$ and δ_H. This proves the base case, since $V^{\pi^{\boldsymbol{w}^\star}}_H(s) \leq V^\star_H(s)$.

Recursing this all the way back, we conclude that

$$V^{\pi^{\boldsymbol{w}^\star}}_0(s_1) \geq V^\star_0(s_1) - \delta_0$$

for $\delta_0 = (1+2/H)^H\delta_H \leq \epsilon$ since $(1+2/x)^x \leq e^2$ for all x.

For this argument to hold, we must choose $\eta \geq 2\log|\widetilde{\mathcal{A}}_s|/\delta_H$ and $\eta \geq H\log|\widetilde{\mathcal{A}}_s|/\delta_h$ for all s and h. By Lemma F.1.6 and our choice of $\widetilde{\mathcal{A}}_s$, we can bound

$$|\widetilde{\mathcal{A}}_s| \leq (1 + 16e^2H^2\sqrt{d}/\epsilon)^d \leq (1+120Hd/\epsilon)^{2d}$$

so, since $\delta_H = \epsilon/e^2 \leq \delta_h$, it suffices that we take $\eta \geq 30dH\log(1+120Hd/\epsilon)\cdot\frac{1}{\epsilon}$.

□

Lemma F.1.9. *Let $\eta = 30dH\log(1+120Hd/\epsilon)\cdot\frac{1}{\epsilon}$ and $\mathcal{W}$ an $\frac{\epsilon}{4dH^2\eta}$-net of $\mathcal{B}^d(2H\sqrt{d})$. Let Π denote the set of restricted-action linear softmax policy with vectors $\boldsymbol{w}\in\mathcal{W}^H$, parameter η, and action sets $(\widetilde{\mathcal{A}}_s)_{s\in\mathcal{S}}$ as defined in Lemma F.1.6 with $\delta = \frac{\epsilon}{2H\cdot e^2}$. Then for any MDP and reward function, there exists some $\pi\in\Pi$ such that $|V^\pi_0 - V^\star_0| \leq \epsilon$, and*

$$|\Pi| \leq \left(1 + \frac{480H^4d^{5/2}\log(1+120Hd/\epsilon)}{\epsilon^2}\right)^{dH}.$$

Proof. Consider some MDP and reward function, and let $\{\boldsymbol{w}^\star_h\}_{h=1}^H$ denote the optimal Q-function linear representation: $Q^\star_h(s,a) = \langle\boldsymbol{\phi}(s,a), \boldsymbol{w}^\star_h\rangle$. Let $\widetilde{\boldsymbol{w}}$ denote the vector in $\mathcal{W}^H$ such that $\sum_{h=1}^H \|\boldsymbol{w}^\star_h - \widetilde{\boldsymbol{w}}_h\|_2$ is minimized. Then by Lemma F.1.7 and Lemma F.1.8, as long as

$\eta \geq 30dH\log(1+120Hd/\epsilon)\cdot\frac{1}{\epsilon}$, we have

$$|V_0^{\pi^{\widetilde{\boldsymbol{w}}}}(s_1) - V_0^\star(s_1)| \leq |V_0^{\pi^{\widetilde{\boldsymbol{w}}}}(s_1) - V_0^{\pi^{\boldsymbol{w}^\star}}(s_1)| + |V_0^{\pi^{\boldsymbol{w}^\star}}(s_1) - V_0^\star(s_1)|$$
$$\leq 2dH\eta\sum_{h=1}^{H}\|\boldsymbol{w}_h^\star - \widetilde{\boldsymbol{w}}_h\|_2 + \epsilon/2.$$

The first conclusion then follows as long as we can find some $\widetilde{\boldsymbol{w}}$ such that

$$2dH\eta\sum_{h=1}^{H}\|\boldsymbol{w}_h^\star - \widetilde{\boldsymbol{w}}_h\|_2 \leq \epsilon/2.$$

However, by Lemma F.1.5, we can bound $\|\boldsymbol{w}_h^\star\|_2 \leq 2H\sqrt{d}$. Therefore, since $\mathcal{W}$ is a $\frac{\epsilon}{4dH^2\eta}$-net of $\mathcal{B}^d(2H\sqrt{d})$, for each h there will exist some $\widetilde{\boldsymbol{w}}_h \in \mathcal{W}$ such that $\|\boldsymbol{w}_h^\star - \widetilde{\boldsymbol{w}}_h\|_2 \leq \frac{\epsilon}{4dH^2\eta}$, which implies that we can find $\widetilde{\boldsymbol{w}} \in \mathcal{W}^d$ such that

$$2dH\eta\sum_{h=1}^{H}\|\boldsymbol{w}_h^\star - \widetilde{\boldsymbol{w}}_h\|_2 \leq \epsilon/2,$$

which gives the first conclusion.

To bound the size of Π, we apply Lemma F.1.10 and our choice of η to bound

$$|\mathcal{W}| \leq (1+\frac{16H^3d^{3/2}\eta}{\epsilon})^d \leq (1+\frac{480H^4d^{5/2}\log(1+120Hd/\epsilon)}{\epsilon^2})^d.$$

The bound on $|\Pi|$ follows since $|\Pi| = |\mathcal{W}|^H$.

□

F.1.3 Technical Results

Lemma F.1.10 (Vershynin [2010]). *For any $\epsilon > 0$, the ϵ-covering number of the Euclidean ball $\mathcal{B}^d(R) := \{\boldsymbol{x} \in \mathbb{R}^d : \|\boldsymbol{x}\|_2 \leq R\}$ with radius $R > 0$ in the Euclidean metric is upper bounded by $(1+2R/\epsilon)^d$.*

Lemma F.1.11 (McSherry and Talwar [2007], Epasto et al. [2020]). *Consider some $(x_i)_{i=1}^n$. Then if $\eta \geq \log(n)/\delta$, we have*

$$\frac{\sum_{i=1}^n e^{\eta x_i}x_i}{\sum_{i=1}^n e^{\eta x_i}} \geq \max_{i\in[n]} x_i - \delta.$$

Lemma F.1.12 (Azuma-Hoeffding). *et $\mathcal{F}_0 \subset \mathcal{F}_1 \subset \ldots \subset \mathcal{F}_T$ be a filtration and let $X_1, X_2, \ldots, X_T$ be real random variables such that X_t is $\mathcal{F}_t$-measurable, $\mathbb{E}[X_t|\mathcal{F}_{t-1}] = 0$,*

and $|X_t| \le b$ *almost surely. Then for any* $\delta \in (0,1)$*, we have with probability at least* $1-\delta$*,*

$$\left|\sum_{t=1}^{T} X_t\right| \le \sqrt{8b^2 \log 2/\delta}.$$

Lemma F.1.13 (Freedman's Inequality [Freedman, 1975])**.** *Let* $\mathcal{F}_0 \subset \mathcal{F}_1 \subset \ldots \subset \mathcal{F}_T$ *be a filtration and let* $X_1, X_2, \ldots, X_T$ *be real random variables such that* X_t *is* $\mathcal{F}_t$*-measurable,* $\mathbb{E}[X_t|\mathcal{F}_{t-1}] = 0$*,* $|X_t| \le b$ *almost surely, and* $\sum_{t=1}^{T} \mathbb{E}[X_t^2|\mathcal{F}_{t-1}] \le V$ *for some fixed* $V > 0$ *and* $b > 0$*. Then for any* $\delta \in (0,1)$*, we have with probability at least* $1-\delta$*,*

$$\sum_{t=1}^{T} X_t \le 2\sqrt{V \log 1/\delta} + b \log 1/\delta.$$

F.2 Deferred Proofs from Section 8.2.2

F.2.1 Linear Bandit Construction

In the linear bandit setting, at each time step t, the learner chooses some $\boldsymbol{z}_t \in \mathcal{Z}$, and observes y_t. We will consider the case when the noise is Bernoulli so that $y_t \sim \text{Bernoulli}(\langle \boldsymbol{\theta}_\star, \boldsymbol{z}_t \rangle + 1/2)$, and will set

$$\boldsymbol{\theta}_\star = \boldsymbol{e}_1, \quad \mathcal{Z} = \{\xi \boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_d, -\xi \boldsymbol{e}_1, -\boldsymbol{e}_2, \ldots, -\boldsymbol{e}_d, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_d\}, \quad \boldsymbol{x}_i = (\xi - \Delta)\boldsymbol{e}_1 + \gamma \boldsymbol{e}_i$$

for some ξ and Δ to be chosen. In this setting, the optimal arm is $\boldsymbol{z}^\star = \xi \boldsymbol{e}_1$, and $\Delta(\boldsymbol{e}_i) = \xi, i \ge 2$, $\Delta(\boldsymbol{x}_i) = \Delta$.

We will assume:

$$\frac{1}{52d} \ge \xi \ge \max\{\gamma/\sqrt{d}, \sqrt{\Delta}\}, \quad \max\left\{\zeta := \frac{\xi \mathcal{C}_1}{(d/200\Delta^2)^{1-\alpha}} + \frac{200\xi\mathcal{C}_2\Delta^2}{d}, \Delta\right\} \le \gamma^2. \quad \text{(F.2.1)}$$

We provide explicit values for ξ, Δ, and γ that satisfy this in Lemma F.2.2.

Definition F.2.1 ((ϵ,δ)-correct Stopping Rule)**.** We say a stopping rule τ is (ϵ,δ)-correct over Θ if for all $\boldsymbol{\theta} \in \Theta$, $\mathbb{P}_{\boldsymbol{\theta}}[\boldsymbol{\theta}^\top \mathbb{E}_{\boldsymbol{z}\sim\widehat{\omega}_\tau}[\boldsymbol{z}] \ge \max_{\boldsymbol{z}\in\mathcal{Z}} \boldsymbol{\theta}^\top \boldsymbol{z} - \epsilon] \ge 1-\delta$, where $\widehat{\omega}_\tau \in \triangle_{\mathcal{Z}}$ is a distribution over arms recommended at time τ.

We will say that some $\omega \in \triangle_{\mathcal{Z}}$ is ϵ-optimal for $\boldsymbol{\theta}$ if $\boldsymbol{\theta}^\top \mathbb{E}_{\boldsymbol{z}\sim\omega}[\boldsymbol{z}] \ge \max_{\boldsymbol{z}\in\mathcal{Z}} \boldsymbol{\theta}^\top \boldsymbol{z} - \epsilon$. We have the following.

Lemma F.2.1. *Consider running some low-regret algorithm satisfying Definition 8.2.1 on the linear bandit instance described above and let* τ *be some stopping rule. Then if* τ *is* (ϵ,δ)*-correct on the set of instances* $\Theta_{\text{lb}} := \{\boldsymbol{\theta} \in \mathbb{R}^d \ : \ \|\boldsymbol{\theta}\|_2 \le \frac{1}{4}(\sqrt{d}-2), |\langle \boldsymbol{\theta}, \boldsymbol{z}\rangle| \le 1/2, \forall \boldsymbol{z} \in \mathcal{Z}\}$ *for* $\delta \in (0, 1/10)$*,* $\epsilon < \min\{\Delta/2, \xi\}$*,* $d \ge 2116$*, and where* ξ, γ*, and* Δ *are chosen as in*

Lemma F.2.2, we must have that

$$\mathbb{E}_{\boldsymbol{\theta}_\star}[\tau] \geq \frac{d}{200\Delta^2} \cdot \log \frac{1}{2.4\delta}.$$

Proof. This proof follows closely the proof of Theorem 1 of Fiez et al. [2019] and relies on the Transportation Lemma of Kaufmann et al. [2016].

Bounding the number of pulls to $\{\pm\boldsymbol{e}_2, \ldots, \pm\boldsymbol{e}_d\}$. By assumption, we collect data with a low-regret algorithm satisfying Definition 8.2.1. Every time we pull $\pm\boldsymbol{e}_i, i \geq 2$, we incur a loss of ξ. Thus, we can lower bound

$$\mathbb{E}[V_0^\star - V_0^{\pi_k}] \geq \xi \sum_{i=2}^{d} \mathbb{E}[\mathbb{P}_{\pi_k}[\{\boldsymbol{z}_k = \boldsymbol{e}_i\} \cup \{\boldsymbol{z}_k = -\boldsymbol{e}_i\}]]$$

so, letting $T(\boldsymbol{e}_i)$ denote the total number of pulls to $\boldsymbol{e}_i$, we have

$$\mathcal{C}_1 K^\alpha + \mathcal{C}_2 \geq \sum_{k=1}^{K} \mathbb{E}[V_0^\star - V_0^{\pi_k}] \geq \xi \sum_{k=1}^{K} \sum_{i=2}^{d} \mathbb{E}[\mathbb{P}_{\pi_k}[\{\boldsymbol{z}_k = \boldsymbol{e}_i\} \cup \{\boldsymbol{z}_k = -\boldsymbol{e}_i\}]] = \xi \sum_{i=2}^{d} \mathbb{E}[T(\boldsymbol{e}_i) + T(-\boldsymbol{e}_i)]. \tag{F.2.2}$$

Applying the Transportation Lemma. Let $\Theta_{\text{alt}} \subseteq \Theta_{\text{lb}}$ denote the set of $\boldsymbol{\theta}$ vectors such that the set of ϵ-optimal distributions on $\boldsymbol{\theta}$ is disjoint from that on $\boldsymbol{\theta}_\star$. By the Transportation Lemma of Kaufmann et al. [2016] (Lemma 1 of Kaufmann et al. [2016]), for any $\boldsymbol{\theta} \in \Theta_{\text{alt}}$, assuming our stopping rule is (ϵ, δ)-correct, we then have

$$\sum_{\boldsymbol{z} \in \mathcal{Z}} \mathbb{E}[T(\boldsymbol{z})] \mathrm{KL}(\nu_{\boldsymbol{\theta}_\star, \boldsymbol{z}} || \nu_{\boldsymbol{\theta}, \boldsymbol{z}}) \geq \log \frac{1}{2.4\delta}.$$

Combining this with our constraint (F.2.2), it follows that $\sum_{\boldsymbol{z} \in \mathcal{Z}} \mathbb{E}[T(\boldsymbol{z})] \geq \sum_{\boldsymbol{z} \in \mathcal{Z}} t_{\boldsymbol{z}}$ for any $(t_{\boldsymbol{z}})_{\boldsymbol{z} \in \mathcal{Z}}$ that is a feasible solution to

$$\min \sum_{\boldsymbol{z} \in \mathcal{Z}} t_{\boldsymbol{z}} \quad \text{s.t.} \quad \min_{\boldsymbol{\theta} \in \Theta_{\text{alt}}} \sum_{\boldsymbol{z} \in \mathcal{Z}} t_{\boldsymbol{z}} \mathrm{KL}(\nu_{\boldsymbol{\theta}_\star, \boldsymbol{z}} || \nu_{\boldsymbol{\theta}, \boldsymbol{z}}) \geq \log \frac{1}{2.4\delta}, \mathcal{C}_1 (\sum_{\boldsymbol{z} \in \mathcal{Z}} t_{\boldsymbol{z}})^\alpha + \mathcal{C}_2 \geq \xi \sum_{i=2}^{d} (t_{\boldsymbol{e}_i} + t_{-\boldsymbol{e}_i}). \tag{F.2.3}$$

We can rearrange the second constraint to

$$\frac{\xi \mathcal{C}_1}{(\sum_{\boldsymbol{z} \in \mathcal{Z}} t_{\boldsymbol{z}})^{1-\alpha}} + \frac{\xi \mathcal{C}_2}{\sum_{\boldsymbol{z} \in \mathcal{Z}} t_{\boldsymbol{z}}} \geq \frac{\sum_{i=2}^{d} (t_{\boldsymbol{e}_i} + t_{-\boldsymbol{e}_i})}{\sum_{\boldsymbol{z} \in \mathcal{Z}} t_{\boldsymbol{z}}}.$$

Assume that the optimal solution to (F.2.3) satisfies $\sum_{\boldsymbol{z}\in\mathcal{Z}} t_{\boldsymbol{z}} \geq \frac{d}{200\Delta^2}$, then this constraint can be weakened to

$$\zeta := \frac{\xi \mathcal{C}_1}{(d/200\Delta^2)^{1-\alpha}} + \frac{200\xi\mathcal{C}_2\Delta^2}{d} \geq \frac{\sum_{i=2}^d t_{\boldsymbol{x}_i}}{\sum_{\boldsymbol{z}\in\mathcal{Z}} t_{\boldsymbol{z}}}.$$

It follows then that if the optimal value to

$$\min \sum_{\boldsymbol{z}\in\mathcal{Z}} t_{\boldsymbol{z}} \quad \text{s.t.} \quad \min_{\boldsymbol{\theta}\in\Theta_{\text{alt}}} \sum_{\boldsymbol{z}\in\mathcal{Z}} t_{\boldsymbol{z}} \mathrm{KL}(\nu_{\boldsymbol{\theta}_\star,\boldsymbol{z}} || \nu_{\boldsymbol{\theta},\boldsymbol{z}}) \geq \log\frac{1}{2.4\delta}, \zeta \geq \frac{\sum_{i=2}^d (t_{\boldsymbol{e}_i} + t_{-\boldsymbol{e}_i})}{\sum_{\boldsymbol{z}\in\mathcal{Z}} t_{\boldsymbol{z}}} \tag{F.2.4}$$

is at least $\frac{d}{200\Delta^2}$, then the optimal value to (F.2.3) is also at least $\frac{d}{200\Delta^2}$, so our assumption that $\sum_{\boldsymbol{z}\in\mathcal{Z}} t_{\boldsymbol{z}} \geq \frac{d}{200\Delta^2}$ will be justified.

For $\boldsymbol{z} \neq \boldsymbol{z}^\star$, let $\boldsymbol{\theta}_{\boldsymbol{z}}(\epsilon, t)$ denote the instance

$$\boldsymbol{\theta}_\star - \frac{(\boldsymbol{y}_{\boldsymbol{z}}^\top \boldsymbol{\theta}_\star + 2\epsilon)\widetilde{\mathbf{A}}(t)^{-1}\boldsymbol{y}_{\boldsymbol{z}}}{\boldsymbol{y}_{\boldsymbol{z}}^\top \widetilde{\mathbf{A}}(t)^{-1}\boldsymbol{y}_{\boldsymbol{z}}}$$

for $\boldsymbol{y}_{\boldsymbol{z}} = \boldsymbol{z}^\star - \boldsymbol{z}$, $\widetilde{\mathbf{A}}(t) = \sum_{\boldsymbol{z}\in\mathcal{Z}} \frac{t_{\boldsymbol{z}}}{\sum_{\boldsymbol{z}'\in\mathcal{Z}} t_{\boldsymbol{z}'}} \boldsymbol{z}\boldsymbol{z}^\top + \mathrm{diag}([\xi^2, \gamma^2/d, \ldots, \gamma^2/d])$. Note that $\boldsymbol{y}_{\boldsymbol{z}}^\top \boldsymbol{\theta}_{\boldsymbol{z}}(\epsilon, t) = -2\epsilon < 0$. Since $\epsilon < \Delta/2$ by assumption, any ϵ-optimal distribution ω for $\boldsymbol{\theta}_\star$ must place mass strictly more than $1/2$ on $\boldsymbol{z}^\star$. Since $\boldsymbol{y}_{\boldsymbol{z}}^\top \boldsymbol{\theta}_{\boldsymbol{z}}(\epsilon, t) = -2\epsilon$, any ϵ-optimal distribution on $\boldsymbol{\theta}_{\boldsymbol{z}}(\epsilon, t)$ must place mass less than or equal to $1/2$ on $\boldsymbol{z}^\star$. Thus, it follows that the ϵ-optimal distributions on $\boldsymbol{\theta}_\star$ and $\boldsymbol{\theta}_{\boldsymbol{z}}(\epsilon, t)$ are disjoint. Furthermore, we have:

Claim 7. *For all $\boldsymbol{z} \in \{\boldsymbol{x}_2, \ldots, \boldsymbol{x}_d\}$ and t, assuming that $\epsilon < \Delta$ and ξ and Δ are chosen as in Lemma F.2.2, we can bound $\|\boldsymbol{\theta}_{\boldsymbol{z}}(\epsilon, t)\|_2 \leq 11$. Furthermore, for any $\boldsymbol{v} \in \mathcal{Z}$, we have $|\langle \boldsymbol{\theta}_{\boldsymbol{z}}(\epsilon, t), \boldsymbol{v}\rangle| \leq 1/4$.*

Since we have chosen $\Theta_{\text{lb}} = \{\boldsymbol{\theta} \in \mathbb{R}^d \ : \ \|\boldsymbol{\theta}\|_2 \leq \frac{1}{4}(\sqrt{d} - 2), |\langle \boldsymbol{\theta}, \boldsymbol{z}\rangle| \leq 1/2, \forall \boldsymbol{z} \in \mathcal{Z}\}$ and assumed $d \geq 2116$, we have that $\boldsymbol{\theta}_{\boldsymbol{z}}(\epsilon, t) \in \Theta_{\text{lb}}$, so $\boldsymbol{\theta}_{\boldsymbol{z}}(\epsilon, t) \in \Theta_{\text{alt}}$. We can also bound:

Claim 8. *For all $\boldsymbol{z}, \boldsymbol{v} \in \mathcal{Z}$ and t,*

$$\mathrm{KL}(\nu_{\boldsymbol{\theta}_\star,\boldsymbol{v}} || \nu_{\boldsymbol{\theta}_{\boldsymbol{z}}(\epsilon,t),\boldsymbol{v}}) \leq 16(\boldsymbol{y}_{\boldsymbol{z}}^\top \boldsymbol{\theta}_\star + \epsilon)^2 \frac{\boldsymbol{y}_{\boldsymbol{z}}^\top \widetilde{\mathbf{A}}(t)^{-1}\boldsymbol{v}\boldsymbol{v}^\top \widetilde{\mathbf{A}}(t)^{-1}\boldsymbol{y}_{\boldsymbol{z}}}{(\boldsymbol{y}_{\boldsymbol{z}}^\top \widetilde{\mathbf{A}}(t)^{-1}\boldsymbol{y}_{\boldsymbol{z}})^2}.$$

This implies that, for any t,

$$\begin{aligned}
\sum_{\boldsymbol{v}\in\mathcal{Z}} t_{\boldsymbol{v}} \mathrm{KL}(\nu_{\boldsymbol{\theta}_\star,\boldsymbol{v}} || \nu_{\boldsymbol{\theta}_{\boldsymbol{z}}(\epsilon,t),\boldsymbol{v}}) &\leq 16 \sum_{\boldsymbol{v}\in\mathcal{Z}} t_{\boldsymbol{v}} (\boldsymbol{y}_{\boldsymbol{z}}^\top \boldsymbol{\theta}_\star + \epsilon)^2 \frac{\boldsymbol{y}_{\boldsymbol{z}}^\top \widetilde{\mathbf{A}}(t)^{-1}\boldsymbol{v}\boldsymbol{v}^\top \widetilde{\mathbf{A}}(t)^{-1}\boldsymbol{y}_{\boldsymbol{z}}}{(\boldsymbol{y}_{\boldsymbol{z}}^\top \widetilde{\mathbf{A}}(t)^{-1}\boldsymbol{y}_{\boldsymbol{z}})^2} \\
&= 16 \sum_{\boldsymbol{v}\in\mathcal{Z}} t_{\boldsymbol{v}} \cdot (\boldsymbol{y}_{\boldsymbol{z}}^\top \boldsymbol{\theta}_\star + \epsilon)^2 \frac{\boldsymbol{y}_{\boldsymbol{z}}^\top \widetilde{\mathbf{A}}(t)^{-1}(\sum_{\boldsymbol{v}\in\mathcal{Z}} \frac{t_{\boldsymbol{v}}}{\sum_{\boldsymbol{v}'\in\mathcal{Z}} t_{\boldsymbol{v}'}} \boldsymbol{v}\boldsymbol{v}^\top)\widetilde{\mathbf{A}}(t)^{-1}\boldsymbol{y}_{\boldsymbol{z}}}{(\boldsymbol{y}_{\boldsymbol{z}}^\top \widetilde{\mathbf{A}}(t)^{-1}\boldsymbol{y}_{\boldsymbol{z}})^2}
\end{aligned}$$

$$\leq 16 \sum_{\boldsymbol{v} \in \mathcal{Z}} t_{\boldsymbol{v}} \cdot (\boldsymbol{y}_{\boldsymbol{z}}^\top \boldsymbol{\theta}_\star + \epsilon)^2 \frac{\boldsymbol{y}_{\boldsymbol{z}}^\top \widetilde{\mathbf{A}}(t)^{-1} \widetilde{\mathbf{A}}(t) \widetilde{\mathbf{A}}(t)^{-1} \boldsymbol{y}_{\boldsymbol{z}}}{(\boldsymbol{y}_{\boldsymbol{z}}^\top \widetilde{\mathbf{A}}(t)^{-1} \boldsymbol{y}_{\boldsymbol{z}})^2}$$
$$= \sum_{\boldsymbol{v} \in \mathcal{Z}} t_{\boldsymbol{v}} \cdot \frac{16(\boldsymbol{y}_{\boldsymbol{z}}^\top \boldsymbol{\theta}_\star + \epsilon)^2}{\|\boldsymbol{y}_{\boldsymbol{z}}\|^2_{\widetilde{\mathbf{A}}(t)^{-1}}}$$

Thus:

$$\text{(F.2.4)} \geq \min \sum_{\boldsymbol{v} \in \mathcal{Z}} t_{\boldsymbol{v}} \quad \text{s.t.} \quad \min_{\boldsymbol{z} \in \{\boldsymbol{x}_1, \dots, \boldsymbol{x}_d\}} \sum_{\boldsymbol{v} \in \mathcal{Z}} t_{\boldsymbol{v}} \mathrm{KL}(\nu_{\boldsymbol{\theta}_\star, \boldsymbol{v}} || \nu_{\boldsymbol{\theta}_{\boldsymbol{z}}(\epsilon, t), \boldsymbol{v}}) \geq \log \frac{1}{2.4\delta}, \zeta \geq \frac{\sum_{i=2}^d (t_{\boldsymbol{e}_i} + t_{-\boldsymbol{e}_i})}{\sum_{\boldsymbol{v} \in \mathcal{Z}} t_{\boldsymbol{v}}}$$
$$\geq \min \sum_{\boldsymbol{v} \in \mathcal{Z}} t_{\boldsymbol{v}} \quad \text{s.t.} \quad \sum_{\boldsymbol{v} \in \mathcal{Z}} t_{\boldsymbol{v}} \geq \max_{\boldsymbol{z} \in \{\boldsymbol{x}_1, \dots, \boldsymbol{x}_d\}} \frac{\|\boldsymbol{y}_{\boldsymbol{z}}\|^2_{\widetilde{\mathbf{A}}(t)^{-1}}}{16(\boldsymbol{y}_{\boldsymbol{z}}^\top \boldsymbol{\theta}_\star + \epsilon)^2} \cdot \log \frac{1}{2.4\delta}, \zeta \geq \frac{\sum_{i=2}^d (t_{\boldsymbol{e}_i} + t_{-\boldsymbol{e}_i})}{\sum_{\boldsymbol{v} \in \mathcal{Z}} t_{\boldsymbol{v}}}$$
$$= \inf_{\lambda \in \widetilde{\triangle}} \max_{\boldsymbol{z} \in \{\boldsymbol{x}_1, \dots, \boldsymbol{x}_d\}} \frac{\|\boldsymbol{y}_{\boldsymbol{z}}\|^2_{\widetilde{\mathbf{A}}(\lambda)^{-1}}}{16(\boldsymbol{y}_{\boldsymbol{z}}^\top \boldsymbol{\theta}_\star + \epsilon)^2} \cdot \log \frac{1}{2.4\delta}$$

where $\widetilde{\mathbf{A}}(\lambda) = \sum_{\boldsymbol{z} \in \mathcal{Z}} \lambda_{\boldsymbol{z}} \boldsymbol{z}\boldsymbol{z}^\top$ and $\widetilde{\triangle} = \{\lambda \in \triangle_{\mathcal{Z}} \ : \ \zeta \geq \sum_{i=2}^d (\lambda_{\boldsymbol{e}_i} + \lambda_{-\boldsymbol{e}_i})\}$. We can further lower bound this by

$$\geq \inf_{\lambda \in \widetilde{\triangle}} \max_{i \geq 2} \frac{\|\boldsymbol{z}^\star - \boldsymbol{x}_i\|^2_{\widetilde{\mathbf{A}}(\lambda)^{-1}}}{16((\boldsymbol{z}^\star - \boldsymbol{x}_i)^\top \boldsymbol{\theta}_\star + \epsilon)^2} \cdot \log \frac{1}{2.4\delta}$$
$$= \inf_{\lambda \in \widetilde{\triangle}} \max_{i \geq 2} \frac{\|\Delta \boldsymbol{e}_1 - \gamma \boldsymbol{e}_i\|^2_{\widetilde{\mathbf{A}}(\lambda)^{-1}}}{16(\Delta + \epsilon)^2} \cdot \log \frac{1}{2.4\delta}.$$

By Lemma F.2.3, we have

$$\inf_{\lambda \in \widetilde{\triangle}} \max_{i \geq 2} \|\Delta \boldsymbol{e}_1 - \gamma \boldsymbol{e}_i\|^2_{\widetilde{\mathbf{A}}(\lambda)^{-1}}$$
$$\geq \inf_{\lambda \in \triangle_d} \max_{i \geq 2} (\Delta \boldsymbol{e}_1 - \gamma \boldsymbol{e}_i)^\top \left(2\xi^2 \boldsymbol{e}_1 \boldsymbol{e}_1^\top + 2\max\{\zeta, \gamma^2\} \lambda_i \boldsymbol{e}_i \boldsymbol{e}_i^\top + \mathrm{diag}([\xi^2, \gamma^2/d, \dots, \gamma^2/d]) \right)^{-1} (\Delta \boldsymbol{e}_1 - \gamma \boldsymbol{e}_i)$$
$$\geq \frac{\Delta^2}{3\xi^2} + \inf_{\lambda \in \triangle_d} \max_{i \geq 2} \frac{1}{2\lambda_i + 1/d}$$

where in the final equality we have used that $\zeta \leq \gamma^2$. However, this is clearly minimized by choosing $\lambda_i = 1/(d-1)$, which gives a lower bound of

$$\frac{1}{2/(d-1) + 1/d} \geq \frac{d-1}{3}.$$

Putting all of this together, we have shown that any feasibly solution $(t_{\boldsymbol{z}})_{\boldsymbol{z} \in \mathcal{Z}}$ to (F.2.3) must

satisfy

$$\sum_{\boldsymbol{z}\in\mathcal{Z}} t_{\boldsymbol{z}} \geq \frac{d-1}{48(\Delta+\epsilon)^2}\cdot\log\frac{1}{2.4\delta} \geq \frac{d}{200\Delta^2}\cdot\log\frac{1}{2.4\delta}$$

where the second inequality uses that $\epsilon \leq \Delta$ and $d \geq 2116$. Using that any feasible solution to (F.2.3) lower bounds $\sum_{\boldsymbol{z}\in\mathcal{Z}} \mathbb{E}[T(\boldsymbol{z})]$ and noting that our lower bound has justified our assumption that $\sum_{\boldsymbol{z}\in\mathcal{Z}} t_{\boldsymbol{z}} \geq d/200\Delta^2$, gives the result. □

Lemma F.2.2. *Take some $\Delta > 0$ satisfying:*

$$\Delta \leq \min\left\{\frac{1}{10000d^4}, \sqrt{\frac{1}{4\cdot 10816\mathcal{C}_2}}, \frac{1}{\sqrt{200}}\left(\frac{1}{10816d^{\alpha}\mathcal{C}_1}\right)^{\frac{1}{2(1-\alpha)}}\right\}$$

and set

$$\xi = \frac{1}{52d}, \qquad \gamma = \frac{1}{52\sqrt{d}}.$$

Then this choice of ξ, γ, δ satisfies (F.2.1) and, furthermore, $\|\boldsymbol{z}\|_2 \leq 1$ for all $\boldsymbol{z}\in\mathcal{Z}$.

Proof. We first fix $\xi = \frac{1}{52d}$. To satisfy (F.2.1), it then suffices to choose γ and Δ satisfying

$$\frac{\gamma}{\sqrt{d}} \leq \frac{1}{52d}, \quad \sqrt{\Delta} \leq \frac{1}{52d}, \quad \frac{\mathcal{C}_1}{52d(d/200\Delta^2)^{1-\alpha}} + \frac{4\mathcal{C}_2\Delta^2}{d^2} \leq \gamma^2, \quad \Delta \leq \gamma^2.$$

Thus, if

$$\Delta \leq \frac{1}{10000d^4}, \quad \Delta \leq \sqrt{\frac{1}{4\cdot 10816\mathcal{C}_2}}, \quad \Delta \leq \frac{1}{\sqrt{200}}\left(\frac{1}{10816d^{\alpha}\mathcal{C}_1}\right)^{\frac{1}{2(1-\alpha)}},$$

some algebra shows that it suffices that we take $\gamma = \frac{1}{52\sqrt{d}}$. The norm bound follows by our choice of ξ and since $\xi \geq \gamma/\sqrt{d} \geq \sqrt{\Delta}$. □

F.2.1.1 Additional Proofs

Proof of Claim 7. We first show that $|\langle\boldsymbol{\theta}_{\boldsymbol{z}}(\epsilon,t),\boldsymbol{v}\rangle| \leq 1/4$ for all $\boldsymbol{z}\in\{x_2,\ldots,x_d\}$ and $\boldsymbol{v}\in\mathcal{Z}$.

Let $\widetilde{\triangle} = \{\lambda\in\triangle_{\mathcal{Z}} \ : \ \zeta \geq \sum_{i=2}^{d}(\lambda_{\boldsymbol{e}_i}+\lambda_{-\boldsymbol{e}_i})\}$. By Lemma F.2.3,

$$\begin{aligned}\boldsymbol{y}_{\boldsymbol{z}}^{\top}\widetilde{\mathbf{A}}(t)^{-1}\boldsymbol{y}_{\boldsymbol{z}} &\geq \inf_{\lambda\in\widetilde{\triangle}} \boldsymbol{y}_{\boldsymbol{z}}^{\top}\Big(2\sum_{\boldsymbol{z}'\in\mathcal{Z}}\lambda_{\boldsymbol{z}'}\mathrm{diag}([(\boldsymbol{z}')^2]) + \mathrm{diag}([\xi^2,\gamma^2/d,\ldots,\gamma^2/d])\Big)^{-1}\boldsymbol{y}_{\boldsymbol{z}} \\ &\geq \boldsymbol{y}_{\boldsymbol{z}}^{\top}\Big(2\xi^2\boldsymbol{e}_1\boldsymbol{e}_1^{\top} + 2\max\{\zeta,\gamma^2\}\boldsymbol{e}_i\boldsymbol{e}_i^{\top} + \mathrm{diag}([\xi^2,\gamma^2/d,\ldots,\gamma^2/d])\Big)^{-1}\boldsymbol{y}_{\boldsymbol{z}}\end{aligned}$$

$$\geq \Delta^2 \frac{3}{\xi^2} + \frac{\gamma^2}{2\max\{\zeta, \gamma^2\} + \gamma^2/d}$$
$$\geq \frac{1}{3}$$

where the last inequality follows from our assumption that $\zeta \leq \gamma^2$. We can also upper bound

$$\boldsymbol{v}^\top \widetilde{\mathbf{A}}(t)^{-1} \boldsymbol{y}_{\boldsymbol{z}} \leq \boldsymbol{v}^\top \mathrm{diag}([\xi^2, \gamma^2/d, \ldots, \gamma^2/d])^{-1} \boldsymbol{y}_{\boldsymbol{z}} \leq \frac{\Delta}{\xi^2} + \frac{d\gamma}{\gamma^2} \leq \frac{2d}{\gamma}.$$

This gives an upper bound of

$$|\langle \boldsymbol{\theta}_{\boldsymbol{z}}(\epsilon, t), \boldsymbol{v} \rangle| \leq \xi + (\Delta + \epsilon)\frac{6d}{\gamma} \leq \xi + \frac{12d\Delta}{\sqrt{\Delta}} \leq 13d\xi.$$

By our assumption that $\xi \leq \frac{1}{52d}$, it follows that $|\langle \boldsymbol{\theta}_{\boldsymbol{z}}(\epsilon, t), \boldsymbol{v} \rangle| \leq 1/4$ for all $\boldsymbol{v} \in \mathcal{Z}$.

To prove the norm bound on $\boldsymbol{\theta}_{\boldsymbol{z}}(\epsilon, t)$, consider $\boldsymbol{v} = \boldsymbol{e}_i$, $i \neq 1$. Similar to the above calculation, we have

$$|\langle \boldsymbol{\theta}_{\boldsymbol{z}}(\epsilon, t), \boldsymbol{v} \rangle| = |[\boldsymbol{\theta}_{\boldsymbol{z}}(\epsilon, t)]_i| \leq (\Delta + \epsilon) \cdot \frac{6d}{\gamma} \leq 12d\sqrt{\Delta}$$

where the last inequality follows by taking $\epsilon < \Delta$, and using that by assumption $\gamma \geq \sqrt{\Delta}$. Now letting $\boldsymbol{v} = \xi \boldsymbol{e}_1 \in \mathcal{Z}$, we have

$$|\langle \boldsymbol{\theta}_{\boldsymbol{z}}(\epsilon, t), \boldsymbol{v} \rangle| = \xi |[\boldsymbol{\theta}_{\boldsymbol{z}}(\epsilon, t)]_1| \leq \xi + 12d\sqrt{\Delta}.$$

Thus, it follows that

$$\begin{aligned}
\|\boldsymbol{\theta}_{\boldsymbol{z}}(\epsilon, t)\|_2 &= \sqrt{\sum_{i=1}^{2} [\boldsymbol{\theta}_{\boldsymbol{z}}(\epsilon, t)]_i^2} \\
&\leq \sqrt{2 + 288d^2\Delta/\xi^2 + 144d^3\Delta)} \\
&\leq \sqrt{2} + \frac{d}{\xi}\sqrt{288\Delta} + d^{3/2}\sqrt{144\Delta}.
\end{aligned}$$

Now, if ξ and Δ are chosen as in Lemma F.2.2, we have

$$\begin{aligned}
\|\boldsymbol{\theta}_{\boldsymbol{z}}(\epsilon, t)\|_2 &\leq \sqrt{2} + 52d^2\sqrt{288\Delta} + d^{3/2}\sqrt{144\Delta} \\
&\leq \sqrt{2} + 52\sqrt{288/10000} + \sqrt{144/10000} \\
&\leq 11.
\end{aligned}$$

□

Proof of Claim 8. By Lemma 2.7 of Tsybakov [2009], as long as $\langle \boldsymbol{\theta}_{\boldsymbol{z}}(\epsilon,t), \boldsymbol{v}\rangle + 1/2 \in (0,1)$ and $\langle \boldsymbol{\theta}_\star, \boldsymbol{v}\rangle + 1/2 \in (0,1)$, which will be the case by the definition of $\boldsymbol{\theta}_\star$ and since $|\langle \boldsymbol{\theta}_{\boldsymbol{z}}(\epsilon,t), \boldsymbol{v}\rangle| \le 1/4$ as noted above, we have

$$\mathrm{KL}(\nu_{\boldsymbol{\theta}_\star,\boldsymbol{v}} || \nu_{\boldsymbol{\theta}_{\boldsymbol{z}}(\epsilon,t),\boldsymbol{v}}) \le \frac{\langle \boldsymbol{\theta}_{\boldsymbol{z}}(\epsilon,t) - \boldsymbol{\theta}_\star, \boldsymbol{v}\rangle^2}{(\langle \boldsymbol{\theta}_{\boldsymbol{z}}(\epsilon,t), \boldsymbol{v}\rangle + 1/2)(1/2 - \langle \boldsymbol{\theta}_{\boldsymbol{z}}(\epsilon,t), \boldsymbol{v}\rangle)}.$$

Using what we have just shown, we can upper bound this as

$$\begin{aligned}\frac{\langle \boldsymbol{\theta}_{\boldsymbol{z}}(\epsilon,t) - \boldsymbol{\theta}_\star, \boldsymbol{v}\rangle^2}{(\langle \boldsymbol{\theta}_{\boldsymbol{z}}(\epsilon,t), \boldsymbol{v}\rangle + 1/2)(1/2 - \langle \boldsymbol{\theta}_{\boldsymbol{z}}(\epsilon,t), \boldsymbol{v}\rangle)} &\le \frac{\langle \boldsymbol{\theta}_{\boldsymbol{z}}(\epsilon,t) - \boldsymbol{\theta}_\star, \boldsymbol{v}\rangle^2}{(-1/4 + 1/2)(1/2 - 1/4)} \\ &= 16\langle \boldsymbol{\theta}_{\boldsymbol{z}}(\epsilon,t) - \boldsymbol{\theta}_\star, \boldsymbol{v}\rangle^2.\end{aligned}$$

By our choice of $\boldsymbol{\theta}_{\boldsymbol{z}}(\epsilon,t)$, this is equal to:

$$16(\boldsymbol{y}_{\boldsymbol{z}}^\top \boldsymbol{\theta}_\star + \epsilon)^2 \frac{\boldsymbol{y}_{\boldsymbol{z}}^\top \widetilde{\mathbf{A}}(t)^{-1} \boldsymbol{v}\boldsymbol{v}^\top \widetilde{\mathbf{A}}(t)^{-1} \boldsymbol{y}_{\boldsymbol{z}}}{(\boldsymbol{y}_{\boldsymbol{z}}^\top \widetilde{\mathbf{A}}(t)^{-1} \boldsymbol{y}_{\boldsymbol{z}})^2}$$

which completes the proof. □

Lemma F.2.3.

$$\sum_{\boldsymbol{z}\in\mathcal{Z}} \lambda_{\boldsymbol{z}} \boldsymbol{z}\boldsymbol{z}^\top \preceq 2 \sum_{\boldsymbol{z}\in\mathcal{Z}} \lambda_{\boldsymbol{z}} \mathrm{diag}(\boldsymbol{z}^2).$$

Proof. This follows since every $\boldsymbol{z} \in \mathcal{Z}$ has at most two non-zero entries, and since $(a\boldsymbol{x} + b\boldsymbol{y})(a\boldsymbol{x} + b\boldsymbol{y})^\top \preceq 2a^2\boldsymbol{x}\boldsymbol{x}^\top + 2b^2\boldsymbol{y}\boldsymbol{y}^\top$. □

F.2.2 Mapping to Linear MDPs

We can map this linear bandit (with parameters chose as in Lemma F.2.2) to a linear MDP with state space $\mathcal{S} = \{s_0, s_1, \bar{s}_2, \ldots, \bar{s}_{d+1}\}$, action space $\mathcal{A} = \mathcal{Z} \cup \{\boldsymbol{e}_{d+1}/2\}$, parameters

$$\begin{aligned}\boldsymbol{\theta}_1 = \mathbf{0}, \quad \boldsymbol{\theta}_2 = \boldsymbol{e}_1 \\ \boldsymbol{\mu}_1(s_1) = [2\boldsymbol{\theta}, 1], \quad \boldsymbol{\mu}_1(\bar{s}_i) = \frac{1}{d}[-2\boldsymbol{\theta}, 1],\end{aligned}$$

for some $\boldsymbol{\theta}$ (in particular, we are interested in $\boldsymbol{\theta} = \boldsymbol{\theta}_\star$), and feature vectors

$$\begin{aligned}&\boldsymbol{\phi}(s_0, \boldsymbol{e}_{d+1}) = \boldsymbol{e}_{d+1}/2, \quad \boldsymbol{\phi}(s_0, \boldsymbol{z}) = [\boldsymbol{z}/2, 1/2], \quad \forall \boldsymbol{z} \in \mathcal{Z} \\ &\boldsymbol{\phi}(s_1, \boldsymbol{z}) = \boldsymbol{e}_1, \quad \boldsymbol{\phi}(\bar{s}_i, \boldsymbol{z}) = \boldsymbol{e}_i, i \ge 2, \quad \forall \boldsymbol{z} \in \mathcal{A}.\end{aligned}$$

Note that, if we take action $\boldsymbol{z}$ in state s_0, our expected episode reward is

$$P_1(s_1|s_0,\boldsymbol{z})\cdot 1+\sum_{i=2}^{d+1}P_1(\bar{s}_i|s_0,\boldsymbol{z})\cdot 0=\langle\boldsymbol{\theta},\boldsymbol{z}\rangle+1/2$$

since we always acquire a reward of 1 in any state s_1, and a reward of 0 in any state $\bar{s}_i$, and the reward distribution is Bernoulli.

Lemma F.2.4. *For any $\boldsymbol{\theta}\in\Theta_{\text{lb}}$, the MDP constructed above is a valid linear MDP as defined in Definition 8.1.1.*

Proof. For $\boldsymbol{z}\in\mathcal{Z}$ we have,

$$\begin{aligned}P_1(s_1|s_0,\boldsymbol{z})&=\langle\boldsymbol{\phi}(s_0,\boldsymbol{z}),\boldsymbol{\mu}_1(s_1)\rangle=\langle\boldsymbol{\theta},\boldsymbol{z}\rangle+1/2\geq 0\\P_1(\bar{s}_i|s_0,\boldsymbol{z})&=\langle\boldsymbol{\phi}(s_0,\boldsymbol{z}),\boldsymbol{\mu}_1(\bar{s}_i)\rangle=\frac{1}{d}(-\langle\boldsymbol{\theta},\boldsymbol{z}\rangle+1/2)\geq 0\end{aligned}$$

where the inequality follows since $|\langle\boldsymbol{\theta},\boldsymbol{z}\rangle|\leq 1/2$ for all $\boldsymbol{z}\in\mathcal{Z}$ by the definition of Θ_{lb}. In addition,

$$P_1(s_1|s_0,\boldsymbol{z})+\sum_{i=2}^{d+1}P_1(\bar{s}_i|s_0,\boldsymbol{z})=\langle\boldsymbol{\theta},\boldsymbol{z}\rangle+1/2+d\cdot\frac{1}{d}(-\langle\boldsymbol{\theta},\boldsymbol{z}\rangle+1/2)=1$$

Thus, $P_1(\cdot|s_0,\boldsymbol{z})$ is a valid probability distribution for $\boldsymbol{z}\in\mathcal{Z}$. A similar calculation shows the same for $\boldsymbol{z}=\boldsymbol{e}_{d+1}/2$.

It remains to check the normalization bounds. Clearly, by our construction of $\mathcal{Z}$, $\|\boldsymbol{\phi}(s,a)\|_2\leq 1$ for all s and a. It is also obvious that $\|\boldsymbol{\theta}_0\|_2\leq\sqrt{d}$ and $\|\boldsymbol{\theta}_1\|_2\leq\sqrt{d}$. Finally,

$$\||\boldsymbol{\mu}_1(\mathcal{S})|\|_2=\left\|\sum_{s\in\mathcal{S}\setminus s_0}|\boldsymbol{\mu}_1(s)|\right\|_2=\|[2\boldsymbol{\theta},1]+d\cdot\frac{1}{d}[2\boldsymbol{\theta},1]\|_2\leq 4\|\boldsymbol{\theta}\|_2+2\leq\sqrt{d}$$

where the last inequality follows from the definition of Θ_{lb}. Thus, all normalization bounds are met, so this is a valid linear MDP. □

Proof of Proposition 8.1. If we assume that the learner has prior access to the feature vectors, and also knows this is a linear MDP, then, even with no knowledge of the dynamics, we can guarantee an optimal policy is contained in the set of policies $\pi^{\boldsymbol{z},\boldsymbol{z}'}$ defined as:

$$\pi_1^{\boldsymbol{z},\boldsymbol{z}'}(s_0)=\boldsymbol{z},\pi_2^{\boldsymbol{z},\boldsymbol{z}'}(s_0)=\boldsymbol{z}',\pi_h^{\boldsymbol{z},\boldsymbol{z}'}(s_1)=\xi\boldsymbol{e}_1,\pi_h^{\boldsymbol{z},\boldsymbol{z}'}(\bar{s}_i)=\xi\boldsymbol{e}_1$$

This holds because in states s_1 and $\bar{s}_i$, the performance of each action is identical since the feature vectors are identical, so it doesn't matter which action we choose in these states. In this case, we can bound $|\Pi|\leq|\mathcal{Z}|^2\leq 4d^2$.

Now, for $\boldsymbol{z} \in \mathcal{A}$, $\boldsymbol{z} \neq \boldsymbol{e}_{d+1}/2$, we have

$$\begin{aligned}\boldsymbol{\phi}_{\pi^{\boldsymbol{z},\boldsymbol{z}'},1} &= [\boldsymbol{z}/2, 1/2] \\ \boldsymbol{\phi}_{\pi^{\boldsymbol{z},\boldsymbol{z}'},2} &= (\langle \boldsymbol{\theta}_\star, \boldsymbol{z}\rangle + 1/2)\boldsymbol{e}_1 + \frac{1}{d}(-\langle \boldsymbol{\theta}_\star, \boldsymbol{z}\rangle + 1/2)\sum_{i\geq 2}\boldsymbol{e}_i\end{aligned}$$

and if $\boldsymbol{z} = \boldsymbol{e}_{d+1}/2$, $\boldsymbol{\phi}_{\pi^{\boldsymbol{z},\boldsymbol{z}'},1} = \boldsymbol{e}_{d+1}/2$, $\boldsymbol{\phi}_{\pi^{\boldsymbol{z},\boldsymbol{z}'},2} = \boldsymbol{e}_1/2 + \frac{1}{2d}\sum_{i\geq 2}\boldsymbol{e}_i$. Let π_{exp} be the policy that plays action $\boldsymbol{e}_2$ in state s_0 at step $h = 1$. Then,

$$\boldsymbol{\Lambda}_{\pi_{\mathrm{exp}},2} = \frac{1}{2}\boldsymbol{e}_1\boldsymbol{e}_1 + \frac{1}{2d}\sum_{i\geq 3}\boldsymbol{e}_i\boldsymbol{e}_i^\top.$$

Since $\langle \boldsymbol{\theta}_\star, \boldsymbol{z}\rangle \leq \mathcal{O}(1/d)$ and $[\boldsymbol{z}]_1 \leq \mathcal{O}(1/d)$ for all $\boldsymbol{z}$ by construction, it follows that we can bound, for all $\boldsymbol{z}, \boldsymbol{z}'$,

$$\|\boldsymbol{\phi}_{\pi^{\boldsymbol{z},\boldsymbol{z}'},2}\|^2_{\boldsymbol{\Lambda}^{-1}_{\pi_{\mathrm{exp}},2}} = \mathcal{O}\left(1 + \sum_{i\geq 2}\frac{1}{d^2}\cdot d\right) = \mathcal{O}(1)$$

so

$$\inf_{\pi_{\mathrm{exp}}}\max_{\pi\in\Pi}\frac{\|\boldsymbol{\phi}_{\pi,2}\|^2_{\boldsymbol{\Sigma}^{-1}_{\pi_{\mathrm{exp}},2}}}{\max\{V_0^\star - V_0^\pi, \Delta^{\Pi}_{\min}, \epsilon\}^2} \leq \mathcal{O}(1/\epsilon^2).$$

Now let π_{exp} be the policy that, at step $h = 1$, plays $\xi\boldsymbol{e}_1$ with probability $1/4$, $\boldsymbol{e}_{d+1}$ with probability $1/4$, and plays $\boldsymbol{e}_i$ with probability $\frac{1}{4(d-1)}$ for $i \geq \{2, \ldots, d\}$. In this setting, we have

$$\boldsymbol{\Lambda}_{\pi_{\mathrm{exp}},1} = \frac{1}{4}\xi^2\boldsymbol{e}_1\boldsymbol{e}_1^\top + \frac{1}{4}\boldsymbol{e}_{d+1}\boldsymbol{e}_{d+1} + \frac{1}{4(d-1)}\sum_{i\in\{2,\ldots,d\}}\boldsymbol{e}_i\boldsymbol{e}_i^\top.$$

Note that $V_0^\star - V_0^{\pi^{\boldsymbol{z},\boldsymbol{z}'}} = \xi - \langle \boldsymbol{\theta}_\star, \boldsymbol{z}\rangle$, so for $\boldsymbol{z} = \boldsymbol{e}_2, \ldots, \boldsymbol{e}_{d+1}$, we have $V_0^\star - V_0^{\pi^{\boldsymbol{z},\boldsymbol{z}'}} = \xi = \mathcal{O}(1/d)$, while for $\boldsymbol{z} = \xi\boldsymbol{e}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_d$, we have $V_0^\star - V_0^{\pi^{\boldsymbol{z},\boldsymbol{z}'}} = \Delta$.

It's easy to see that for $\boldsymbol{z} = \boldsymbol{e}_2, \ldots, \boldsymbol{e}_{d+1}$, we have $\|\boldsymbol{\phi}_{\pi^{\boldsymbol{z},\boldsymbol{z}'},1}\|_{\boldsymbol{\Lambda}^{-1}_{\pi_{\mathrm{exp}},1}} \leq \mathcal{O}(d)$, and for $\boldsymbol{z} = \xi\boldsymbol{e}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_d$, $\|\boldsymbol{\phi}_{\pi^{\boldsymbol{z},\boldsymbol{z}'},1}\|_{\boldsymbol{\Lambda}^{-1}_{\pi_{\mathrm{exp}},1}} \leq \mathcal{O}(1 + d\gamma^2) = \mathcal{O}(1)$. Combining these bounds with the gap values, we conclude that

$$\inf_{\pi_{\mathrm{exp}}}\max_{\pi\in\Pi}\frac{\|\boldsymbol{\phi}_{\pi,1}\|^2_{\boldsymbol{\Sigma}^{-1}_{\pi_{\mathrm{exp}},1}}}{\max\{V_0^\star - V_0^\pi, \Delta^{\Pi}_{\min}, \epsilon\}^2} \leq \mathcal{O}(1/\epsilon^2 + \mathrm{poly}(d)).$$

The result then follows by Lemma 8.3.8 and Lemma F.2.4. □

Lower bounding the performance of low-regret algorithms. Assume that we have access to the linear bandit instance constructed in Appendix F.2.1 with parameters chosen as in Lemma F.2.2. That is, at every timestep t we can choose an arm $\boldsymbol{z}_t \in \mathcal{Z}$ and obtain and observe reward $y_t \sim \text{Bernoulli}(\langle \boldsymbol{\theta}_\star, \boldsymbol{z}_t \rangle + 1/2)$. Using the mapping up, we can use this bandit to simulate a linear MDP as follows:

1. Start in state s_0 and choose any action $\boldsymbol{z}_t \in \mathcal{A}$

2. Play action $\boldsymbol{z}_t$ in our linear bandit. If reward obtained is $y_t = 1$, then in MDP transition to any of the states s_1. If reward obtained is $y_t = 0$ transition to any of the states $\bar{s}_2, \ldots, \bar{s}_{d+1}$, each with probability $1/d$. If the chosen action was $\boldsymbol{z}_t = \boldsymbol{e}_{d+1}/2$, then play any action in the linear bandit and transition to state s_1 with probability $1/2$ and $\bar{s}_2, \ldots, \bar{s}_{d+1}$ with probability $1/2d$, regardless of y_t

3. Take any action in the state in which you end up, and receive reward of 1 if you are in s_1, and reward of 0 if you are in $\bar{s}_2, \ldots, \bar{s}_{d+1}$.

Note that this MDP has precisely the transition and reward structure as the MDP constructed above.

Lemma F.2.5. *Any algorithm that is (ϵ, δ)-PAC for linear MDPs satisfying Definition 8.1.1 is also (ϵ, δ)-PAC for linear bandits constructed as above, for any $\boldsymbol{\theta} \in \Theta_{\text{lb}}$.*

Proof. Consider some policy π that is ϵ-optimal on the linear MDP constructed as above. Note that, for all $\boldsymbol{\theta}$, we have $\max_{\boldsymbol{z} \in \mathcal{Z}} \langle \boldsymbol{z}, \boldsymbol{\theta} \rangle \geq 0$, so that the optimal action in the linear MDP is never $\boldsymbol{e}_{d+1}/2$. If π is ϵ-optimal in the linear MDP, this then implies that

$$\begin{aligned} &\sum_{\boldsymbol{z} \in \mathcal{Z}} \pi_1(\boldsymbol{z}|s_0)(\langle \boldsymbol{z}, \boldsymbol{\theta} \rangle + 1/2) + \pi_1(\boldsymbol{e}_{d+1}/2|s_0)/2 \geq \max_{\boldsymbol{z} \in \mathcal{Z}} \langle \boldsymbol{z}, \boldsymbol{\theta} \rangle + 1/2 - \epsilon \\ \implies &\langle \sum_{\boldsymbol{z} \in \mathcal{Z}} \pi_1(\boldsymbol{z}|s_0)\boldsymbol{z}, \boldsymbol{\theta} \rangle \geq \max_{\boldsymbol{z} \in \mathcal{Z}} \langle \boldsymbol{z}, \boldsymbol{\theta} \rangle - \epsilon. \end{aligned}$$

Thus, it follows that the distribution $\pi_1(\boldsymbol{z}|s_0)$ is ϵ-optimal for the linear bandit with parameter $\boldsymbol{\theta}$, implying that any policy ϵ-optimal on the linear MDP constructed above can be used to identify an ϵ-optimal solution to the linear bandit.

By Lemma F.2.4, for any $\boldsymbol{\theta} \in \Theta_{\text{lb}}$, the linear MDP constructed above with parameter $\boldsymbol{\theta}$ is a linear MDP satisfying Definition 8.1.1. Thus, it follows that any algorithm (ϵ, δ)-PAC on linear MDPs satisfying Definition 8.1.1 must also be (ϵ, δ)-PAC on our linear bandit.

□

Proof of Proposition 8.2. By Lemma F.2.5, we know that any algorithm which is (ϵ, δ)-PAC for linear MDPs satisfying Definition 8.1.1 is also (ϵ, δ)-PAC on the set of linear bandit

instances with arm set $\mathcal{Z}$ as constructed above, and parameter in Θ_{lb}. As the lower bound given in Lemma F.2.1 applies to any algorithm that is (ϵ, δ)-PAC on linear bandits with parameters in Θ_{lb}, it follows that it also applies to a (ϵ, δ)-PAC linear MDP algorithm, which proves the result. □

F.3 Deferred Proofs from Section 8.3

Lemma F.3.1. *With probability at least $1-\delta$:*

$$\|\widehat{\boldsymbol{\phi}}_{\pi,h} - \boldsymbol{\phi}_{\pi,h}\|_2 \leq d \sum_{h'=1}^{h-1} \left(2\sqrt{\log \frac{2Hd}{\delta}} + \frac{\log \frac{2Hd}{\delta}}{\sqrt{\lambda_{\min}(\boldsymbol{\Lambda}_{h'})}} + \sqrt{d\lambda} \right) \cdot \|\widehat{\boldsymbol{\phi}}_{\pi,h'}\|_{\boldsymbol{\Lambda}_{h'}^{-1}}.$$

Proof. We have:

$$\|\widehat{\boldsymbol{\phi}}_{\pi,h} - \boldsymbol{\phi}_{\pi,h}\|_2 \leq \|\widehat{\boldsymbol{\phi}}_{\pi,h} - \boldsymbol{\phi}_{\pi,h}\|_1 = \sum_{i=1}^{d} |[\widehat{\boldsymbol{\phi}}_{\pi,h}]_i - [\boldsymbol{\phi}_{\pi,h}]_i| = \sum_{i=1}^{d} |\langle \boldsymbol{e}_i, \widehat{\boldsymbol{\phi}}_{\pi,h} - \boldsymbol{\phi}_{\pi,h} \rangle|.$$

Since $\boldsymbol{e}_i \in \mathcal{S}^{d-1}$, we can apply Lemma 8.3.2 to bound, with probability $1-\delta/d$,

$$|\langle \boldsymbol{e}_i, \widehat{\boldsymbol{\phi}}_{\pi,h} - \boldsymbol{\phi}_{\pi,h} \rangle| \leq \sum_{h'=1}^{h-1} \left(2\sqrt{\log \frac{2Hd}{\delta}} + \frac{\log \frac{2Hd}{\delta}}{\sqrt{\lambda_{\min}(\boldsymbol{\Lambda}_{h'})}} + \sqrt{d\lambda} \right) \cdot \|\widehat{\boldsymbol{\phi}}_{\pi,h'}\|_{\boldsymbol{\Lambda}_{h'}^{-1}}.$$

Summing over i gives the result. □

Lemma F.3.2. *Assume we have collected data $\{\boldsymbol{\phi}(s_{h,\tau}, a_{h,\tau}), R_h(s_{h,\tau}, a_{h,\tau})\}_{\tau=1}^{K}$ and that for each τ', $R_h(s_{h,\tau'}, a_{h,\tau'})|(s_{h,\tau'}, a_{h,\tau'})$ is independent of $\{(s_{h,\tau}, a_{h,\tau})\}_{\tau \neq \tau'}$. Let*

$$\widehat{\boldsymbol{\theta}}_h = \arg\min_{\boldsymbol{\theta}} \sum_{\tau=1}^{K} (R_{h,\tau} - \langle \boldsymbol{\phi}_{h,\tau}, \boldsymbol{\theta} \rangle)^2 + \lambda \|\boldsymbol{\theta}\|_2^2$$

and fix $\boldsymbol{u} \in \mathbb{R}^d$ that is independent of $\{\boldsymbol{\phi}(s_{h,\tau}, a_{h,\tau}), R_h(s_{h,\tau}, a_{h,\tau})\}_{\tau=1}^{K}$. Then with probability at least $1-\delta$:

$$|\langle \boldsymbol{u}, \widehat{\boldsymbol{\theta}}_h - \boldsymbol{\theta}_h \rangle| \leq \left(\sqrt{\log 2/\delta} + \frac{\log 2/\delta}{\sqrt{\lambda_{\min}(\boldsymbol{\Lambda}_h)}} + \sqrt{d\lambda} \right) \cdot \|\boldsymbol{u}\|_{\boldsymbol{\Lambda}_h^{-1}}.$$

Proof. Let $\mathfrak{D} = \{(s_{h,\tau}, a_{h,\tau})\}_{\tau=1}^{K}$. Then by our assumption on the independence of $R_{h,\tau}$, we have that $R_{h,\tau}|(s_{h,\tau}, a_{h,\tau})$ has the same distribution as $r_{h,\tau}|\mathfrak{D}$. Conditioning on $\mathfrak{D}$, the $\boldsymbol{\phi}_{h,\tau}$ vectors are fixed, so $\boldsymbol{\Lambda}_h$ is also fixed.

By construction we have

$$\widehat{\boldsymbol{\theta}}_h = \boldsymbol{\Lambda}_h^{-1} \sum_{\tau=1}^{K} \boldsymbol{\phi}_{h,\tau} R_{h,\tau}.$$

Furthermore:

$$\boldsymbol{\theta}_h = \boldsymbol{\Lambda}_h^{-1} \boldsymbol{\Lambda}_h \boldsymbol{\theta}_h = \boldsymbol{\Lambda}_h^{-1} \sum_{\tau=1}^{K} \boldsymbol{\phi}_{h,\tau} \mathbb{E}[R_{h,\tau} | \mathcal{F}_{h-1,\tau}] + \lambda \boldsymbol{\Lambda}_h^{-1} \boldsymbol{\theta}_h.$$

Thus,

$$|\langle \boldsymbol{u}, \widehat{\boldsymbol{\theta}}_h - \boldsymbol{\theta}_h \rangle| \leq \underbrace{\left| \sum_{\tau=1}^{K} \boldsymbol{u}^\top \boldsymbol{\Lambda}_h^{-1} \boldsymbol{\phi}_{h,\tau} (R_{h,\tau} - \mathbb{E}[R_{h,\tau} | \mathcal{F}_{h-1,\tau}]) \right|}_{(a)} + \underbrace{\left| \lambda \boldsymbol{u}^\top \boldsymbol{\Lambda}_h^{-1} \boldsymbol{\theta}_h \right|}_{(b)}.$$

Since $R_{h,\tau} \in [0,1]$ almost surely, we can bound

$$|\boldsymbol{u}^\top \boldsymbol{\Lambda}_h^{-1} \boldsymbol{\phi}_{h,\tau} (R_{h,\tau} - \mathbb{E}[R_{h,\tau} | \mathcal{F}_{h-1,\tau}])| \leq \|\boldsymbol{u}\|_{\boldsymbol{\Lambda}_h^{-1}} \|\boldsymbol{\phi}_{h,\tau}\|_{\boldsymbol{\Lambda}_h^{-1}} \leq \|\boldsymbol{u}\|_{\boldsymbol{\Lambda}_h^{-1}} / \sqrt{\lambda_{\min}(\boldsymbol{\Lambda}_h)}.$$

Furthermore, we can bound

$$\begin{aligned}
&\operatorname{Var}\left[\boldsymbol{u}^\top \boldsymbol{\Lambda}_h^{-1} \boldsymbol{\phi}_{h,\tau} (R_{h,\tau} - \mathbb{E}[R_{h,\tau} | \mathcal{F}_{h-1,\tau}]) | \mathfrak{D} \right] \\
&\quad = \mathbb{E}\left[(\boldsymbol{u}^\top \boldsymbol{\Lambda}_h^{-1} \boldsymbol{\phi}_{h,\tau} (R_{h,\tau} - \mathbb{E}[R_{h,\tau} | \mathcal{F}_{h-1,\tau}]))^2 | \mathfrak{D} \right] \\
&\quad \leq \boldsymbol{u}^\top \boldsymbol{\Lambda}_h^{-1} \boldsymbol{\phi}_{h,\tau} \boldsymbol{\phi}_{h,\tau}^\top \boldsymbol{\Lambda}_h^{-1} \boldsymbol{u}.
\end{aligned}$$

By Bernstein's inequality, we then have, with probability at least $1 - \delta$ conditioned on $\mathfrak{D}$:

$$\begin{aligned}
(a) &\leq \sqrt{\sum_{\tau=1}^{K} \boldsymbol{u}^\top \boldsymbol{\Lambda}_h^{-1} \boldsymbol{\phi}_{h,\tau} \boldsymbol{\phi}_{h,\tau}^\top \boldsymbol{\Lambda}_h^{-1} \boldsymbol{u} \cdot \log 2/\delta} + \frac{\|\boldsymbol{u}\|_{\boldsymbol{\Lambda}_h^{-1}} \cdot \log 2/\delta}{\sqrt{\lambda_{\min}(\boldsymbol{\Lambda}_h)}} \\
&\leq (\sqrt{\log 2/\delta} + \frac{\log 2/\delta}{\sqrt{\lambda_{\min}(\boldsymbol{\Lambda}_h)}}) \cdot \|\boldsymbol{u}\|_{\boldsymbol{\Lambda}_h^{-1}}.
\end{aligned}$$

Applying the Law of Total Probability as in Lemma 8.3.1, we obtain

$$\mathbb{P}\left[(a) \geq (\sqrt{\log 2/\delta} + \frac{\log 2/\delta}{\sqrt{\lambda_{\min}(\boldsymbol{\Lambda}_h)}}) \cdot \|\boldsymbol{u}\|_{\boldsymbol{\Lambda}_h^{-1}} \right] \leq \delta.$$

By Definition 8.1.1, we can also bound

$$(b) \leq \sqrt{\lambda}\|\boldsymbol{u}\|_{\boldsymbol{\Lambda}_h^{-1}}\|\boldsymbol{\theta}_h\|_2 \leq \sqrt{d\lambda}\|\boldsymbol{u}\|_{\boldsymbol{\Lambda}_h^{-1}}.$$

Combining these proves the result. □

Proof of Corollary 8.2. By Lemma F.1.9, we can choose Π_ϵ to be the restricted-action linear softmax policy set constructed in Lemma F.1.9. Lemma F.1.9 shows that Π_ϵ will contain an ϵ-optimal policy for any MDP and reward function, and that

$$|\Pi_\epsilon| \leq \left(1 + \frac{480H^4d^{5/2}\log(1+120Hd/\epsilon)}{\epsilon^2}\right)^{dH}.$$

Combining this with the guarantee of Theorem 8.1 shows that $V_0^{\widehat{\pi}} \geq V_0^\star - 2\epsilon$ and that $V_0^\star(\Pi) - V_0^\pi$ is within a factor of ϵ of $V_0^\star - V_0^\pi$. To bound the complexity of this procedure, we apply the bound given in Theorem 8.1 with the bound on the cardinality of Π_ϵ given above. □

F.3.1 Interpreting the Complexity

Lemma F.3.3. *For any set of policies* Π*, we can bound*

$$\inf_{\boldsymbol{\Lambda}\in\boldsymbol{\Omega}_h}\sup_{\pi\in\Pi}\|\boldsymbol{\phi}_{\pi,h}\|^2_{\boldsymbol{\Lambda}^{-1}} \leq d.$$

Proof. By Jensen's inequality, for any $\boldsymbol{v} \in \mathbb{R}^d$, we have

$$\boldsymbol{v}^\top\boldsymbol{\Lambda}_{\pi,h}\boldsymbol{v} = \mathbb{E}_\pi[(\boldsymbol{v}^\top\boldsymbol{\phi}_h)^2] \geq (\mathbb{E}_\pi[\boldsymbol{v}^\top\boldsymbol{\phi}_h])^2 = (\boldsymbol{v}^\top\boldsymbol{\phi}_{\pi,h})^2.$$

It follows that, for any π,

$$\boldsymbol{\Lambda}_{\pi,h} \succeq \boldsymbol{\phi}_{\pi,h}\boldsymbol{\phi}_{\pi,h}^\top.$$

Take $\boldsymbol{\Lambda} \in \boldsymbol{\Omega}_h$. Then,

$$\boldsymbol{\Lambda} = \mathbb{E}_{\pi\sim\omega}[\boldsymbol{\Lambda}_{\pi,h}] \succeq \mathbb{E}_{\pi\sim\omega}[\boldsymbol{\phi}_{\pi,h}\boldsymbol{\phi}_{\pi,h}^\top].$$

It follows that we can upper bound

$$\inf_{\boldsymbol{\Lambda}\in\boldsymbol{\Omega}_h}\sup_{\pi\in\Pi}\|\boldsymbol{\phi}_{\pi,h}\|^2_{\boldsymbol{\Lambda}^{-1}} \leq \inf_{\lambda\in\triangle_\Pi}\sup_{\pi\in\Pi}\|\boldsymbol{\phi}_{\pi,h}\|^2_{A(\lambda)^{-1}}$$

where $A(\lambda) = \sum_\pi \lambda_\pi\boldsymbol{\phi}_{\pi,h}\boldsymbol{\phi}_{\pi,h}^\top$. By Kiefer-Wolfowitz [Lattimore and Szepesvári, 2020], this is upper bounded by d. □

Proof of Corollary 8.3. This follows directly from Lemma F.3.3 and Corollary 8.2, by upper

bounding:

$$\inf_{\mathbf{\Lambda}\in\mathbf{\Omega}_h}\max_{\pi\in\Pi_\epsilon}\frac{\|\boldsymbol{\phi}_{\pi,h}\|^2_{\mathbf{\Lambda}^{-1}}}{\max\{V_0^\star - V_0^\pi,\epsilon\}^2} \le \inf_{\mathbf{\Lambda}\in\mathbf{\Omega}_h}\max_{\pi\in\Pi_\epsilon}\frac{\|\boldsymbol{\phi}_{\pi,h}\|^2_{\mathbf{\Lambda}^{-1}}}{\epsilon^2} \le \frac{d}{\epsilon^2}.$$

□

F.3.1.1 Tabular MDPs

Proof of Proposition 8.3. We begin with an example where PEDEL has complexity smaller than the Gap-Visitation Complexity, and then turn to an example where the reverse is true.

PEDEL Improves on Gap-Visitation Complexity. Consider the tabular MDP with $|\mathcal{S}| = |\mathcal{A}| = N$, and where

$$\begin{aligned}
&P_h(s_1|s_1,a_1) = 1, \quad \nu_h(s_1,a_1) = 1, \forall h \in [H]\\
&P_h(s_1|s_1,a_j) = 0, \quad \nu_h(s_1,a_j) = 0, \forall h \in [H], j \neq 1\\
&P_h(s_1|s_i,a_j) = 0, \forall h \in [H], j \in [N], i \neq 1\\
&P_h(s_i|s_j,a_i) = 1, \forall h \in [H], j \in [N], i \neq 1\\
&R_h(s_i,a_1) = \epsilon, \forall h \in [H], i \neq 1, \quad R_h(s_i,a_j) = 0, \forall h \in [H], j \neq 1, i \neq 1.
\end{aligned}$$

In this MDP, the optimal policy simply plays action a_1 H times and is always in state s_1. The total reward it collects is H.

Note that, since the MDP is deterministic, we can choose Π as in Corollary 8.4, to be simply the set of policies which play a deterministic sequences of actions. It follows that any policy in Π that does not play a_1 H consecutive times has optimality gap of at least $1-\epsilon$. In this case, then, we have $\Delta_{\min}(\Pi) = 1-\epsilon$.

By Theorem 8.1, we can therefore upper bound the complexity of the leading-order term by $\widetilde{\mathcal{O}}(\mathrm{poly}(S,A,H,\log 1/\delta))$, so PEDEL will identify the optimal policy (since Π contains an optimal policy). Thus, the total complexity of PEDEL is $\mathcal{O}(\mathrm{poly}(S,A,H,\log 1/\delta))$.

On this example, in every state s_i, $i \neq 1$, action a_1 still collects a reward of ϵ. Thus, we have that $\Delta_h(s_i,a_j) = \epsilon$ for $j \neq 1$. The Gap-Visitation complexity is given by

$$\sum_{h=1}^{H}\inf_{\pi}\max_{s,a}\min\left\{\frac{1}{w_h^\pi(s,a)\Delta_h(s,a)^2}, \frac{W_h(s)^2}{\epsilon^2}\right\}.$$

Since $W_h(s) = 1$ for each s, we conclude that

$$\sum_{h=1}^{H}\inf_{\pi}\max_{s,a}\min\left\{\frac{1}{w_h^\pi(s,a)\Delta_h(s,a)^2}, \frac{W_h(s)^2}{\epsilon^2}\right\} \ge \sum_{h=1}^{H}\frac{1}{\epsilon^2}.$$

Thus, for small ϵ, the Gap-Visitation complexity can be arbitrarily worse than the complexity of PEDEL.

The Gap-Visitation Complexity Improves on PEDEL. To show that the Gap-Visitation Complexity improves on the complexity of PEDEL, we consider the example in Instance Class 7.3.1. As shown in Proposition 7.5 on this example, for any ϵ, the Gap-Visitation Complexity is $\widetilde{\mathcal{O}}(\text{poly}(S))$.

To bound the complexity of PEDEL on this example, we consider the complexity given in Lemma 8.3.8 with Π the set of all deterministic policies. Take $\epsilon \geq 2^{-S}$. Then, on this example, it follows that $\Delta_{\min}(\Pi) \leq \mathcal{O}(\epsilon)$, since we can find a policy π which is optimal on every state s_i at step $h = 2$ for $i = \mathcal{O}(\log 1/\epsilon)$, which will give it a policy gap of $\mathcal{O}(\epsilon)$. Furthermore, any near-optimal policy will have $[\boldsymbol{\phi}_{\pi,2}]_{s_1,a_1} = w_2^\pi(s_1, a_1) = \mathcal{O}(1)$, so we always have $\inf_{\boldsymbol{\Lambda} \in \boldsymbol{\Omega}_2} \max_{\pi \in \Pi(4\epsilon_\ell)} \|\boldsymbol{\phi}_{\pi,h}\|^2_{\boldsymbol{\Lambda}^{-1}} \geq \Omega(1)$. It follows that the complexity of PEDEL is lower bounded by $\Omega(1/\epsilon^2)$.

□

F.3.1.2 Deterministic, Tabular MDPs

Lemma F.3.4. *In the deterministic MDP setting,*

$$\inf_{\boldsymbol{\Lambda} \in \boldsymbol{\Omega}_h} \max_{\pi \in \Pi} \frac{\|\boldsymbol{\phi}_{\pi,h}\|^2_{\boldsymbol{\Lambda}^{-1}}}{(V_0^\star - V_0^\pi)^2 \vee \epsilon^2} \leq \sum_{s,a} \frac{1}{\bar{\Delta}_h(s,a)^2 \vee \epsilon^2}.$$

Proof. Note that $[\boldsymbol{\phi}_{\pi,h}]_{s_h^\pi, a_h^\pi} = 1$, and otherwise, for $(s,a) \neq (s_h^\pi, a_h^\pi)$, $[\boldsymbol{\phi}_{\pi,h}]_{s,a} = 0$. Furthermore, $\boldsymbol{\Lambda}_{\pi_{\text{exp}},h}$ will always be diagonal, with diagonal elements $w_h^\pi(s,a)$. We then have $\|\boldsymbol{\phi}_{\pi,h}\|^2_{\boldsymbol{\Lambda}^{-1}_{\pi_{\text{exp}},h}} = \frac{1}{w_h^{\pi_{\text{exp}}}(s_h^\pi, a_h^\pi)}$, so

$$\begin{aligned}
\inf_{\boldsymbol{\Lambda} \in \boldsymbol{\Omega}_h} \max_{\pi \in \Pi} \frac{\|\boldsymbol{\phi}_{\pi,h}\|^2_{\boldsymbol{\Lambda}^{-1}}}{(V_0^\star - V_0^\pi)^2 \vee \epsilon^2} &\leq \inf_{\pi_{\text{exp}}} \max_{\pi \in \Pi} \frac{\|\boldsymbol{\phi}_{\pi,h}\|^2_{\boldsymbol{\Lambda}^{-1}_{\pi_{\text{exp}},h}}}{(V_0^\star - V_0^\pi)^2 \vee \epsilon^2} \\
&= \inf_{\pi_{\text{exp}}} \max_{\pi \in \Pi} \frac{w_h^{\pi_{\text{exp}}}(s_h^\pi, a_h^\pi)^{-1}}{(V_0^\star - V_0^\pi)^2 \vee \epsilon^2} \\
&\overset{(a)}{=} \inf_{\pi_{\text{exp}}} \max_{s,a} \max_{\pi \in \Pi_{sah}} \frac{w_h^{\pi_{\text{exp}}}(s_h^\pi, a_h^\pi)^{-1}}{(V_0^\star - V_0^\pi)^2 \vee \epsilon^2} \\
&\overset{(b)}{=} \inf_{\pi_{\text{exp}}} \max_{s,a} \max_{\pi \in \Pi_{sah}} \frac{w_h^{\pi_{\text{exp}}}(s,a)^{-1}}{(V_0^\star - V_0^\pi)^2 \vee \epsilon^2} \\
&= \inf_{\pi_{\text{exp}}} \max_{s,a} \frac{w_h^{\pi_{\text{exp}}}(s,a)^{-1}}{(V_0^\star - \max_{\pi \in \Pi_{sah}} V_0^\pi)^2 \vee \epsilon^2} \\
&\overset{(c)}{=} \inf_{\pi_{\text{exp}}} \max_{s,a} \frac{w_h^{\pi_{\text{exp}}}(s,a)^{-1}}{\bar{\Delta}_h(s,a)^2 \vee \epsilon^2}
\end{aligned}$$

where (a) follows since $\Pi = \cup_{s,a}\Pi_{sah}$, (b) follows since by definition, for any $\pi \in \Pi_{sah}$, $(s_h^\pi, a_h^\pi) = (s, a)$, and (c) follows by the definition of $\bar{\Delta}_h(s, a)$.

Let π^{sa} denote any policy such that $(s_h^\pi, a_h^\pi) = (s, a)$. Set

$$\lambda_{\pi^{sa}} = \frac{\max\{\bar{\Delta}_h(s,a), \epsilon\}^{-2}}{\sum_{s',a'} \max\{\bar{\Delta}_h(s',a'), \epsilon\}^{-2}}.$$

Note that this is a valid distribution. Let $\pi_{\mathrm{exp}} = \sum_{s,a} \lambda_{\pi^{sa}} \pi^{sa}$, then $w_h^{\pi_{\mathrm{exp}}}(s,a) = \lambda_{\pi^{sa}}$, so

$$\begin{aligned}\inf_{\pi_{\mathrm{exp}}} \max_{s,a} \frac{w_h^{\pi_{\mathrm{exp}}}(s,a)^{-1}}{\bar{\Delta}_h(s,a)^2 \vee \epsilon^2} &\le \max_{s,a} \frac{\lambda_{\pi^{sa}}^{-1}}{\bar{\Delta}_h(s,a)^2 \vee \epsilon^2} \\ &\le \sum_{s,a} \frac{1}{\bar{\Delta}_h(s,a)^2 \vee \epsilon^2}\end{aligned}$$

which proves the result. □

Proof of Corollary 8.4. As in tabular MDPs, we can set Π to correspond to the set of all deterministic policies. However, since our MDP is also deterministic, at any given h, we only need to specify $\pi_h(s)$ for a single s—the state we will end up in at step h with probability 1. Thus, we can take Π to be a set of cardinality $|\Pi| = A^H$. The result then follows directly from Lemma F.3.4 and Theorem 8.1. □

F.4 Deferred Proofs from Section 8.4

Throughout this section, we will denote $\gamma_\Phi := \max_{\phi \in \Phi} \|\phi\|_2$ and let $f(\mathbf{\Lambda}) := \widetilde{\mathsf{XY}}_{\mathrm{opt}}(\mathbf{\Lambda})$.

F.4.1 Approximating Non-Smooth Optimal Design with Smooth Optimal Design

Lemma F.4.1.

$$|\mathsf{XY}_{\mathrm{opt}}(\mathbf{\Lambda}) - \widetilde{\mathsf{XY}}_{\mathrm{opt}}(\mathbf{\Lambda})| \le \frac{\log|\Phi|}{\eta}, \qquad \mathsf{XY}_{\mathrm{opt}}(\mathbf{\Lambda}) \le \widetilde{\mathsf{XY}}_{\mathrm{opt}}(\mathbf{\Lambda}).$$

Proof. This result is standard but we include the proof for completeness. We prove it for some generic sequence $(a_i)_{i=1}^n$. Take $\eta > 0$. Clearly,

$$\exp(\max_i \eta a_i) \le \sum_{i=1}^n \exp(\eta a_i) \le n \exp(\max_i \eta a_i)$$

so

$$\max_i \eta a_i \le \log\left(\sum_{i=1}^n \exp(\eta a_i)\right) \le \log n + \max_i \eta a_i.$$

The result follows by rearranging and dividing by η. □

Lemma F.4.2. *If $\eta \ge \widetilde{\eta} \ge 0$, then $\widetilde{\mathsf{XY}}_{\mathrm{opt}}(\mathbf{\Lambda};\eta) \le \widetilde{\mathsf{XY}}_{\mathrm{opt}}(\mathbf{\Lambda};\widetilde{\eta})$.*

Proof. We will prove this for some generic sequence $(a_i)_{i=1}^n$, $a_i \ge 0$. Note that,

$$\frac{\mathrm{d}}{\mathrm{d}\eta}\frac{1}{\eta}\log\left(\sum_i e^{\eta a_i}\right) = -\frac{1}{\eta^2}\log\left(\sum_i e^{\eta a_i}\right) + \frac{1}{\eta}\frac{1}{\sum_i e^{\eta a_i}}\cdot\sum_i a_i e^{\eta a_i}.$$

We are done if we can show this is non-positive. Note that,

$$\log\left(\sum_i e^{\eta a_i}\right) \ge \max_i \log\left(e^{\eta a_i}\right) = \max_i \eta a_i$$

so

$$\begin{aligned}
-\frac{1}{\eta^2}\log\left(\sum_i e^{\eta a_i}\right) + \frac{1}{\eta}\frac{1}{\sum_i e^{\eta a_i}}\cdot\sum_i a_i e^{\eta a_i} &\le -\frac{1}{\eta}\max_i a_i + \frac{1}{\eta}\frac{1}{\sum_i e^{\eta a_i}}\cdot\sum_i a_i e^{\eta a_i} \\
&\le -\frac{1}{\eta}\max_i a_i + \frac{1}{\eta}\max_i a_i \\
&= 0.
\end{aligned}$$

The result follows since $\widetilde{\mathsf{XY}}_{\mathrm{opt}}$ has this form. □

Lemma F.4.3. *We have,*

$$\inf_{\mathbf{\Lambda}\succeq 0, \|\mathbf{\Lambda}\|_{\mathrm{op}}\le 1} \mathsf{XY}_{\mathrm{opt}}(\mathbf{\Lambda}) \ge \frac{\gamma_\Phi}{1+\|\mathbf{\Lambda}_0\|_{\mathrm{op}}}.$$

Proof. Note that $\|\mathbf{A}(\mathbf{\Lambda})\|_{\mathrm{op}} \le 1 + \|\mathbf{\Lambda}_0\|_{\mathrm{op}}$, so

$$\inf_{\mathbf{\Lambda}\succeq 0, \|\mathbf{\Lambda}\|_{\mathrm{op}}\le 1} \max_{\boldsymbol{\phi}\in\Phi} \|\boldsymbol{\phi}\|^2_{\mathbf{A}(\mathbf{\Lambda})^{-1}} \ge \inf_{\mathbf{\Lambda}\succeq 0, \|\mathbf{\Lambda}\|_{\mathrm{op}}\le 1+\|\mathbf{\Lambda}_0\|_{\mathrm{op}}} \|\boldsymbol{\phi}\|^2_{\mathbf{A}(\mathbf{\Lambda})^{-1}} \ge \frac{\max_{\boldsymbol{\phi}\in\Phi}\|\boldsymbol{\phi}\|_2}{1+\|\mathbf{\Lambda}_0\|_{\mathrm{op}}}.$$

□

Lemma F.4.4. *Assume that we set $\eta \ge \frac{2}{\gamma_\Phi}(1+\|\mathbf{\Lambda}_0\|_{\mathrm{op}})\cdot\log|\Phi|$. Then*

$$\min_{\mathbf{\Lambda}\in\mathbf{\Omega}} \widetilde{\mathsf{XY}}_{\mathrm{opt}}(\mathbf{\Lambda})) \le 2\cdot\min_{\mathbf{\Lambda}\in\mathbf{\Omega}}\max_{\boldsymbol{\phi}\in\Phi}\|\boldsymbol{\phi}\|^2_{\mathbf{A}(\mathbf{\Lambda})^{-1}}.$$

Proof. Denote $f(\mathbf{\Lambda}) \leftarrow \text{LogSumExp}\left(\{e^{\eta\|\boldsymbol{\phi}\|^2_{\mathbf{A}(\mathbf{\Lambda})^{-1}}}\}_{\boldsymbol{\phi}\in\Phi}; \eta\right)$. By Lemma F.4.1 and Lemma F.4.3, we have

$$|\max_{\boldsymbol{\phi}\in\Phi}\|\boldsymbol{\phi}\|^2_{\mathbf{A}(\mathbf{\Lambda})^{-1}} - f(\mathbf{\Lambda})| \le \frac{\log|\Phi|}{\eta} \le \frac{\gamma_\Phi}{2(1+\|\mathbf{\Lambda}_0\|_{\text{op}})} \le \min_{\mathbf{\Lambda}\succeq 0, \|\mathbf{\Lambda}\|_{\text{op}}\le 1} \frac{1}{2} f(\mathbf{\Lambda})$$
$$\implies f(\mathbf{\Lambda}) \le 2\max_{\boldsymbol{\phi}\in\Phi}\|\boldsymbol{\phi}\|^2_{\mathbf{A}(\mathbf{\Lambda})^{-1}}.$$

The result follows. □

F.4.2 Bounding the Smoothness

Lemma F.4.5. *$f(\mathbf{\Lambda}) = \widetilde{\text{XY}}_{\text{opt}}(\mathbf{\Lambda})$ satisfies all conditions of Algorithm 8.2 with*

$$\beta = 2\|\mathbf{\Lambda}_0^{-1}\|^3_{\text{op}}(1+\eta\|\mathbf{\Lambda}_0^{-1}\|_{\text{op}}), \quad M = \|\mathbf{\Lambda}_0^{-1}\|^2_{\text{op}}$$

$$\nabla_{\mathbf{\Lambda}} f(\mathbf{\Lambda}) = \left(\sum_{\boldsymbol{\phi}\in\Phi} e^{\eta\|\boldsymbol{\phi}\|^2_{\mathbf{A}(\mathbf{\Lambda})^{-1}}}\right)^{-1} \cdot \sum_{\boldsymbol{\phi}\in\Phi} e^{\eta\|\boldsymbol{\phi}\|^2_{\mathbf{A}(\mathbf{\Lambda})^{-1}}} \mathbf{A}(\mathbf{\Lambda})^{-1}\boldsymbol{\phi}\boldsymbol{\phi}^\top \mathbf{A}(\mathbf{\Lambda})^{-1} =: \Xi_{\mathbf{\Lambda}}.$$

Proof. Using Lemma F.4.6, the gradient of $f(\mathbf{\Lambda})$ with respect to $\mathbf{\Lambda}_{ij}$ is

$$\nabla_{\mathbf{\Lambda}_{ij}} f(\mathbf{\Lambda}) = -\left(\sum_{\boldsymbol{\phi}\in\Phi} e^{\eta\|\boldsymbol{\phi}\|^2_{\mathbf{A}(\mathbf{\Lambda})^{-1}}}\right)^{-1} \cdot \sum_{\boldsymbol{\phi}\in\Phi} e^{\eta\|\boldsymbol{\phi}\|^2_{\mathbf{A}(\mathbf{\Lambda})^{-1}}} \boldsymbol{\phi}^\top \mathbf{A}(\mathbf{\Lambda})^{-1}\boldsymbol{e}_i\boldsymbol{e}_j^\top \mathbf{A}(\mathbf{\Lambda})^{-1}\boldsymbol{\phi}$$

from which the expression for $\nabla_{\mathbf{\Lambda}} f(\mathbf{\Lambda})$ follows directly.

To bound M, we see that

$$\sup_{\mathbf{\Lambda},\widetilde{\mathbf{\Lambda}}\succeq 0, \|\mathbf{\Lambda}\|_{\text{op}}\le 1, \|\widetilde{\mathbf{\Lambda}}\|_{\text{op}}\le 1} |\text{tr}(\nabla f(\mathbf{\Lambda})^\top \widetilde{\mathbf{\Lambda}})| \le \left(\sum_{\boldsymbol{\phi}\in\Phi} e^{\eta\|\boldsymbol{\phi}\|^2_{\mathbf{A}(\mathbf{\Lambda})^{-1}}}\right)^{-1} \cdot \sum_{\boldsymbol{\phi}\in\Phi} e^{\eta\|\boldsymbol{\phi}\|^2_{\mathbf{A}(\mathbf{\Lambda})^{-1}}} \|\mathbf{A}(\mathbf{\Lambda})^{-1}\|^2_{\text{op}} \|\widetilde{\mathbf{\Lambda}}\|_{\text{op}}$$
$$\le \|\mathbf{\Lambda}_0^{-1}\|^2_{\text{op}}$$

where the last inequality follows since $\mathbf{A}(\mathbf{\Lambda}) \succeq \mathbf{\Lambda}_0$ for all $\mathbf{\Lambda}$.

To bound the smoothness, again by the Mean Value Theorem it suffices to bound the operator norm of the Hessian. Standard calculus gives that, using $\nabla^2 f(\mathbf{\Lambda})[\widetilde{\mathbf{\Lambda}}, \bar{\mathbf{\Lambda}}]$ to denote the Hessian of f in direction $(\widetilde{\mathbf{\Lambda}}, \bar{\mathbf{\Lambda}})$:

$$\nabla^2 f(\mathbf{\Lambda})[\widetilde{\mathbf{\Lambda}}, \bar{\mathbf{\Lambda}}] = -\frac{d}{dt}\left(\sum_{\boldsymbol{\phi}\in\Phi} e^{\eta\|\boldsymbol{\phi}\|^2_{\mathbf{A}(\mathbf{\Lambda}+t\bar{\mathbf{\Lambda}})^{-1}}}\right)^{-1} \cdot \sum_{\boldsymbol{\phi}\in\Phi} e^{\eta\|\boldsymbol{\phi}\|^2_{\mathbf{A}(\mathbf{\Lambda}+t\bar{\mathbf{\Lambda}})^{-1}}} \boldsymbol{\phi}^\top \mathbf{A}(\mathbf{\Lambda}+t\bar{\mathbf{\Lambda}})^{-1}\widetilde{\mathbf{\Lambda}}\mathbf{A}(\mathbf{\Lambda}+t\bar{\mathbf{\Lambda}})^{-1}\boldsymbol{\phi}$$

$$
\begin{aligned}
&= -\eta \left(\sum_{\boldsymbol{\phi} \in \Phi} e^{\eta \|\boldsymbol{\phi}\|^2_{\mathbf{A}(\boldsymbol{\Lambda})^{-1}}} \right)^{-2} \left(\sum_{\boldsymbol{\phi} \in \Phi} e^{\eta \|\boldsymbol{\phi}\|^2_{\mathbf{A}(\boldsymbol{\Lambda})^{-1}}} \boldsymbol{\phi}^\top \mathbf{A}(\boldsymbol{\Lambda})^{-1} \bar{\boldsymbol{\Lambda}} \mathbf{A}(\boldsymbol{\Lambda})^{-1} \boldsymbol{\phi} \right) \\
&\qquad \cdot \left(\sum_{\boldsymbol{\phi} \in \Phi} e^{\eta \|\boldsymbol{\phi}\|^2_{\mathbf{A}(\boldsymbol{\Lambda})^{-1}}} \boldsymbol{\phi}^\top \mathbf{A}(\boldsymbol{\Lambda})^{-1} \widetilde{\boldsymbol{\Lambda}} \mathbf{A}(\boldsymbol{\Lambda})^{-1} \boldsymbol{\phi} \right) \\
&\quad + \eta \left(\sum_{\boldsymbol{\phi} \in \Phi} e^{\eta \|\boldsymbol{\phi}\|^2_{\mathbf{A}(\boldsymbol{\Lambda})^{-1}}} \right)^{-1} \sum_{\boldsymbol{\phi} \in \Phi} e^{\eta \|\boldsymbol{\phi}\|^2_{\mathbf{A}(\boldsymbol{\Lambda})^{-1}}} \left(\boldsymbol{\phi}^\top \mathbf{A}(\boldsymbol{\Lambda})^{-1} \bar{\boldsymbol{\Lambda}} \mathbf{A}(\boldsymbol{\Lambda})^{-1} \boldsymbol{\phi} \right) \left(\boldsymbol{\phi}^\top \mathbf{A}(\boldsymbol{\Lambda})^{-1} \widetilde{\boldsymbol{\Lambda}} \mathbf{A}(\boldsymbol{\Lambda})^{-1} \boldsymbol{\phi} \right) \\
&\quad + \left(\sum_{\boldsymbol{\phi} \in \Phi} e^{\eta \|\boldsymbol{\phi}\|^2_{\mathbf{A}(\boldsymbol{\Lambda})^{-1}}} \right)^{-1} \sum_{\boldsymbol{\phi} \in \Phi} e^{\eta \|\boldsymbol{\phi}\|^2_{\mathbf{A}(\boldsymbol{\Lambda})^{-1}}} \boldsymbol{\phi}^\top \mathbf{A}(\boldsymbol{\Lambda})^{-1} \bar{\boldsymbol{\Lambda}} \mathbf{A}(\boldsymbol{\Lambda})^{-1} \widetilde{\boldsymbol{\Lambda}} \mathbf{A}(\boldsymbol{\Lambda})^{-1} \boldsymbol{\phi} \\
&\quad + \left(\sum_{\boldsymbol{\phi} \in \Phi} e^{\eta \|\boldsymbol{\phi}\|^2_{\mathbf{A}(\boldsymbol{\Lambda})^{-1}}} \right)^{-1} \sum_{\boldsymbol{\phi} \in \Phi} e^{\eta \|\boldsymbol{\phi}\|^2_{\mathbf{A}(\boldsymbol{\Lambda})^{-1}}} \boldsymbol{\phi}^\top \mathbf{A}(\boldsymbol{\Lambda})^{-1} \widetilde{\boldsymbol{\Lambda}} \mathbf{A}(\boldsymbol{\Lambda})^{-1} \bar{\boldsymbol{\Lambda}} \mathbf{A}(\boldsymbol{\Lambda})^{-1} \boldsymbol{\phi}.
\end{aligned}
$$

We can bound this as

$$
\sup_{\boldsymbol{\Lambda}, \widetilde{\boldsymbol{\Lambda}}, \bar{\boldsymbol{\Lambda}} \succeq 0, \|\boldsymbol{\Lambda}\|_{\mathrm{op}} \le 1, \|\widetilde{\boldsymbol{\Lambda}}\|_{\mathrm{op}} \le 1, \|\bar{\boldsymbol{\Lambda}}\|_{\mathrm{op}} \le 1} |\nabla^2 f(\boldsymbol{\Lambda})[\widetilde{\boldsymbol{\Lambda}}, \bar{\boldsymbol{\Lambda}}]| \le 2\eta \|\boldsymbol{\Lambda}_0^{-1}\|_{\mathrm{op}}^4 + 2\|\boldsymbol{\Lambda}_0^{-1}\|_{\mathrm{op}}^3.
$$

Convexity of $f(\boldsymbol{\Lambda})$ follows since it is the composition of a convex function with a strictly increasing convex function, so it is itself convex. □

Lemma F.4.6. *For $\boldsymbol{\Lambda}$ invertible, $\frac{d}{dt}(\boldsymbol{\Lambda} + t\boldsymbol{e}_i\boldsymbol{e}_j^\top)^{-1} = -\boldsymbol{\Lambda}^{-1}\boldsymbol{e}_i\boldsymbol{e}_j^\top\boldsymbol{\Lambda}^{-1}$.*

Proof. We can compute the gradient as

$$
\frac{d}{dt}(\boldsymbol{\Lambda} + t\boldsymbol{e}_i\boldsymbol{e}_j^\top)^{-1} = \lim_{t \to 0} \frac{(\boldsymbol{\Lambda} + t\boldsymbol{e}_i\boldsymbol{e}_j^\top)^{-1} - \boldsymbol{\Lambda}^{-1}}{t}.
$$

By the Sherman-Morrison formula,

$$
(\boldsymbol{\Lambda} + t\boldsymbol{e}_i\boldsymbol{e}_j^\top)^{-1} = \boldsymbol{\Lambda}^{-1} - \frac{t\boldsymbol{\Lambda}^{-1}\boldsymbol{e}_i\boldsymbol{e}_j^\top\boldsymbol{\Lambda}^{-1}}{1 + t\boldsymbol{e}_j^\top\boldsymbol{\Lambda}^{-1}\boldsymbol{e}_i}
$$

so as $t \to 0$,

$$
(\boldsymbol{\Lambda} + t\boldsymbol{e}_i\boldsymbol{e}_j^\top)^{-1} \to \boldsymbol{\Lambda}^{-1} - t\boldsymbol{\Lambda}^{-1}\boldsymbol{e}_i\boldsymbol{e}_j^\top\boldsymbol{\Lambda}^{-1}
$$

Thus,

$$
\lim_{t \to 0} \frac{(\boldsymbol{\Lambda} + t\boldsymbol{e}_i\boldsymbol{e}_j^\top)^{-1} - \boldsymbol{\Lambda}^{-1}}{t} = \lim_{t \to 0} \frac{\boldsymbol{\Lambda}^{-1} - t\boldsymbol{\Lambda}^{-1}\boldsymbol{e}_i\boldsymbol{e}_j^\top\boldsymbol{\Lambda}^{-1} - \boldsymbol{\Lambda}^{-1}}{t} = -\boldsymbol{\Lambda}^{-1}\boldsymbol{e}_i\boldsymbol{e}_j^\top\boldsymbol{\Lambda}^{-1}.
$$

□

F.4.3 Data Collection via Online Frank-Wolfe

Proof of Lemma 8.4.1. Note that by Lemma 6.6.2, if $f_i(\widehat{\mathbf{\Lambda}}_i) \leq K_i N_i \epsilon$ then this implies that $f_i(\frac{1}{K_i N_i}\widehat{\mathbf{\Sigma}}_i) \leq K_i N_i \epsilon$, where $\widehat{\mathbf{\Sigma}}_i$ are the actual observed covariates at round i. Thus, upon termination the desired conclusion of Lemma 8.4.1 will be met. It then remains to find a setting of i large enough to guarantee that the termination criteria is met.

To show this, we first argue that the FORCE algorithm of Wagenmaker et al. [2022a] satisfies Assumption 6.10. By Theorem 5 of Wagenmaker et al. [2022a], we have that with probability at least $1-\delta$, FORCE has regret bounded after K episodes as:

$$c_1\sqrt{d^4H^4K \cdot \log^3(HK/\delta)} + c_2 d^4 H^3 \log^{7/2}(HK/\delta) \leq c_3 d^4 H^4 \log^{7/2}(K/\delta)\cdot\sqrt{K}.$$

It follows that FORCE satisfies Assumption 6.10 with $C_{\mathcal{R}} = c_3 d^4 H^4$, $p_{\mathcal{R}} = 7/2$, and $\alpha = 1/2$.

By Theorem 6.3, we then have with probability at least $1-\delta/2i^2$:

$$f_i(\widehat{\mathbf{\Lambda}}_i) \leq \min_{\mathbf{\Lambda}\in\mathbf{\Omega}} f_i(\mathbf{\Lambda}) + \frac{\beta_i R^2(\log N_i + 1)}{2(N_i+1)} + cMd^4H^4\log^{7/2}\frac{4i^2N_iK_i}{\delta}\cdot\sqrt{\frac{1}{K_i}} + M\sqrt{\frac{8\log(8i^2N_i/\delta)}{K_i}}.$$

By assumption, we have $\min_{\mathbf{\Lambda}\in\mathbf{\Omega}} f_i(\mathbf{\Lambda}) \geq f_{\min}$. By our choice $N_i = 2^i$ and $K_i = 2^{2i}$, we see that we can bound

$$\frac{\beta R^2(\log N_i + 1)}{2(N_i+1)} + cMd^4H^4\log^{7/2}\frac{4i^2N_iK_i}{\delta}\cdot\sqrt{\frac{1}{K_i}} + M\sqrt{\frac{8\log(8i^2N_i/\delta)}{K_i}} \leq f_{\min}$$

as long as

$$i \geq c\cdot\log\max\{\beta, M, d, H, f_{\min}^{-1}, \log\frac{1}{\delta}\}$$

for a universal constant c. Thus, for i satisfying this, we have $f_i(\widehat{\mathbf{\Lambda}}_i) \leq 2\min_{\mathbf{\Lambda}\in\mathbf{\Omega}} f_i(\mathbf{\Lambda})$. So to ensure that $f_i(\widehat{\mathbf{\Lambda}}_i) \leq K_iN_i\epsilon$ it suffices that $2\min_{\mathbf{\Lambda}\in\mathbf{\Omega}} f_i(\mathbf{\Lambda}) \leq K_iN_i\epsilon$. It then suffices that $\frac{2}{\epsilon}\cdot\min_{\mathbf{\Lambda}\in\mathbf{\Omega}} f_i(\mathbf{\Lambda}) \leq 2^{4i}$. Since we assume $f(\mathbf{\Lambda}) \geq f_i(\mathbf{\Lambda})$, this is implied by $\frac{2}{\epsilon}\cdot\min_{\mathbf{\Lambda}\in\mathbf{\Omega}} f(\mathbf{\Lambda}) \leq 2^{3i}$.

As the failure probability of the success event of Theorem 6.3 at round i is bounded by $\delta/2i^2$, the total failure probability is bounded by $\sum_{i=1}^{\infty}\delta/2i^2 \leq \delta$. On this success event, it suffices that we will terminate after round $\widehat{i}$ for any $\widehat{i}$ satisfying

$$\widehat{i} \geq c\cdot\log\max\{\beta, R, M, d, H, f_{\min}^{-1}, \log\frac{1}{\delta}\}, \quad 2^{3\widehat{i}} \geq \frac{2}{\epsilon}\cdot\min_{\mathbf{\Lambda}\in\mathbf{\Omega}} f(\mathbf{\Lambda}). \tag{F.4.1}$$

If we terminate at epoch $\widehat{i}$, the total complexity will be bounded by

$$\sum_{i=1} N_i K_i = \sum_{i=1}^{\widehat{i}} 2^{3i} \le \frac{8}{7} \cdot 2^{3\widehat{i}}.$$

Since we will have terminated as soon as $\widehat{i}$ satisfies (F.4.1), it follows that

$$\widehat{i} - 1 < c \cdot \log \max\{\beta, R, M, d, H, f_{\min}^{-1}, \log \frac{1}{\delta}\} \quad \text{or} \quad 2^{3(\widehat{i}-1)} < \frac{2}{\epsilon} \cdot \min_{\mathbf{\Lambda} \in \mathbf{\Omega}} f(\mathbf{\Lambda}).$$

This implies that

$$2^{3\widehat{i}} = 8 \cdot 2^{3(\widehat{i}-1)} \le \frac{16}{\epsilon} \cdot \min_{\mathbf{\Lambda} \in \mathbf{\Omega}} f(\mathbf{\Lambda}) + \mathrm{poly}\left(\beta, R, M, d, H, f_{\min}^{-1}, \log \frac{1}{\delta}\right).$$

The result follows. □

Lemma F.4.7. *There exists a procedure which takes as input parameter $N, \underline{\lambda}, \delta$ and with probability at least $1 - \delta$ terminates after at most*

$$N + \mathrm{poly}\left(d, \frac{1}{\bar{\lambda}_{\min}}, H, \log \frac{N}{\delta}, \underline{\lambda}\right)$$

episodes, and returns covariates $\mathbf{\Sigma}$ such that

$$\lambda_{\min}(\mathbf{\Sigma}) \ge \frac{N}{C_{\lambda_{\min}}} + \max\{d \log 1/\delta, \underline{\lambda}\}$$

for $C_{\lambda_{\min}} = \mathrm{poly}\left(d, \frac{1}{\bar{\lambda}_{\min}}, H, \log \frac{N+\underline{\lambda}}{\delta}\right)$, and

$$\|\mathbf{\Sigma}\|_{\mathrm{op}} \le N + \mathrm{poly}\left(d, \frac{1}{\bar{\lambda}_{\min}}, H, \log \frac{N}{\delta}, \underline{\lambda}\right).$$

Proof. Applying Lemma D.3.1, we have that the MinEig algorithm with scale $N + \underline{\lambda}$, confidence δ, and Force as the regret minimization algorithm, will return some set of policies Π such that, rerunning each policy in Π once, with probability at least $1 - \delta/(N + \underline{\lambda}/d)$, we will collect covariates $\mathbf{\Sigma}$ such that

$$\lambda_{\min}(\mathbf{\Sigma}) \ge 6272 d \log \frac{68(N + \underline{\lambda})}{\delta}.$$

Furthermore, MinEig will run for at most

$$2|\Pi| \le \mathrm{poly}\left(d, \frac{1}{\bar{\lambda}_{\min}}, H, \log \frac{N + \underline{\lambda}}{\delta}\right) =: C_{\lambda_{\min}}$$

episodes. This implies that, rerunning each policy $N/|\Pi| + \underline{\lambda}/d + 1$ times, we have that the covariates returned $\mathbf{\Sigma}$ satisfy, with probability at least $1 - \delta$:

$$\begin{aligned}\lambda_{\min}(\mathbf{\Sigma}) &\geq \frac{N}{|\Pi|} \cdot 6272 d \log \frac{68(N + \underline{\lambda})}{\delta} + \max\{d \log 1/\delta, \underline{\lambda}\} \\ &\geq \frac{N}{C_{\lambda_{\min}}} \cdot 6272 d \log \frac{68(N + \underline{\lambda})}{\delta} + \max\{d \log 1/\delta, \underline{\lambda}\}.\end{aligned}$$

Note that the total number of episodes collected by this procedure is bounded by

$$|\Pi| \cdot \frac{N}{|\Pi|} + |\Pi| \cdot \underline{\lambda}/d + |\Pi| \leq N + \text{poly}\left(d, \frac{1}{\bar{\lambda}_{\min}}, H, \log \frac{N + \underline{\lambda}}{\delta}, \underline{\lambda}\right).$$

The bound on the operator norm of $\mathbf{\Sigma}$ follows from this and since each $\boldsymbol{\phi}$ has norm at most 1. $\square$

Appendix G

Deferred Proofs from Chapter 9

G.1 Deferred Proofs from Section 9.3

Lemma G.1.1 (Elliptic Potential Lemma, Lemma 11 of Abbasi-Yadkori et al. [2011])**.** *Consider a sequence of vectors* $(\boldsymbol{x}_t)_{t=1}^T, \boldsymbol{x}_t \in \mathbb{R}^d$*, and assume that* $\|\boldsymbol{x}_t\|_2 \leq a$ *for all* t*. Let* $\mathbf{V}_t = \lambda I + \sum_{s=1}^t \boldsymbol{x}_s \boldsymbol{x}_s^\top$ *for some* $\lambda > 0$*. Then we will have that*

$$\sum_{t=1}^{T} \min\{1, \|\boldsymbol{x}_t\|^2_{\mathbf{V}_{t-1}^{-1}}\} \leq 2d \log(1 + a^2 T/(d\lambda)).$$

Furthermore, if $\lambda \geq \max\{1, a^2\}$*,*

$$\sum_{t=1}^{T} \|\boldsymbol{x}_t\|^2_{\mathbf{V}_{t-1}^{-1}} \leq 2d \log(1 + a^2 T/(d\lambda)).$$

Lemma G.1.2. *If* $x \geq C(2n)^n \log^n(2nCB)$ *for* $n, C, B \geq 1$*, then* $x \geq C \log^n(Bx)$*.*

Proof. If $x = C(2n)^n \log^n(2nCB)$, then

$$\begin{aligned} C \log^n(Bx) &= C \log^n \left[C(2n)^n \log^n(2nCB)\right] \\ &\leq C \log^n \left[C^{1+n}(2n)^{2n} B^n\right] \\ &\leq C \log^n \left[C^{2n}(2n)^{2n} B^{2n}\right] \\ &\leq C(2n)^n \log^n \left[2nCB\right] \\ &= x. \end{aligned}$$

The result then follows since x increases more quickly than $C \log^n(Bx)$. □

G.1.1 FORCE satisfies Definition 9.1.1

We invoke Theorem 8 of Wagenmaker et al. [2022a]. To do so, we need to control an appropriate covering number. Recall the ball $\mathcal{B}^d := \{\boldsymbol{\phi} \in \mathbb{R}^d : \|\boldsymbol{\phi}\| \leq 1\}$. Given a class of

functions $\mathscr{F}$ denote a set of functions $f : \mathcal{B}^d \to \mathbb{R}$, we define the distance on $f_1, f_2 \in \mathscr{F}$

$$\operatorname{dist}_\infty(f_1, f_2) := \sup_{\boldsymbol{\phi} \in \mathcal{B}^d} |f_1(\boldsymbol{\phi}) - f_2(\boldsymbol{\phi})|.$$

Lemma G.1.3. *Consider the class of functions*

$$\mathscr{R} := \left\{ r : \mathcal{B}^d \to \mathbb{R} \ : \ r(\boldsymbol{\phi}) = \begin{cases} 1 & \|\boldsymbol{\phi}\|^2_{\boldsymbol{\Lambda}^{-1}} > \boldsymbol{\gamma}^2, \boldsymbol{\phi} \in \mathcal{X} \\ \boldsymbol{\gamma}^{-2}\|\boldsymbol{\phi}\|^2_{\boldsymbol{\Lambda}^{-1}} & \|\boldsymbol{\phi}\|^2_{\boldsymbol{\Lambda}^{-1}} \le \boldsymbol{\gamma}^2, \boldsymbol{\phi} \in \mathcal{X}, \boldsymbol{\Lambda} \succeq I \\ 0 & \boldsymbol{\phi} \notin \mathcal{X} \end{cases} \right\}$$

for some $\boldsymbol{\gamma}^2 > 0$ and $\mathcal{X} \subseteq \mathbb{R}^d$. Then

$$\mathsf{N}(\mathscr{R}, \operatorname{dist}_\infty, \epsilon) \le d^2 \log\left(1 + \frac{2\sqrt{d}}{\boldsymbol{\gamma}^2 \epsilon}\right).$$

Proof. Consider $r_1, r_2 \in \mathscr{R}$ parameterized by $\boldsymbol{\Lambda}_1, \boldsymbol{\Lambda}_2$. Then,

$$\operatorname{dist}_\infty(r_1, r_2) = \sup_{\boldsymbol{\phi} \in \mathcal{B}^d} |r_1(\boldsymbol{\phi}) - r_2(\boldsymbol{\phi})| \le \sup_{\boldsymbol{\phi} \in \mathcal{B}^d} \frac{1}{\boldsymbol{\gamma}^2} \left| \|\boldsymbol{\phi}\|^2_{\boldsymbol{\Lambda}_1^{-1}} - \|\boldsymbol{\phi}\|^2_{\boldsymbol{\Lambda}_2^{-1}} \right|.$$

The inequality follows easily by noting that in every possible case, we can bound

$$|r_1(\boldsymbol{\phi}) - r_2(\boldsymbol{\phi})| \le \frac{1}{\boldsymbol{\gamma}^2} \left| \|\boldsymbol{\phi}\|^2_{\boldsymbol{\Lambda}_1^{-1}} - \|\boldsymbol{\phi}\|^2_{\boldsymbol{\Lambda}_2^{-1}} \right|.$$

Now note that

$$\left| \|\boldsymbol{\phi}\|^2_{\boldsymbol{\Lambda}_1^{-1}} - \|\boldsymbol{\phi}\|^2_{\boldsymbol{\Lambda}_2^{-1}} \right| = \left| \boldsymbol{\phi}^\top (\boldsymbol{\Lambda}_1^{-1} - \boldsymbol{\Lambda}_2^{-1}) \boldsymbol{\phi} \right| \le \|\boldsymbol{\Lambda}_1^{-1} - \boldsymbol{\Lambda}_2^{-1}\|_{\mathrm{op}} \le \|\boldsymbol{\Lambda}_1^{-1} - \boldsymbol{\Lambda}_2^{-1}\|_{\mathrm{F}}.$$

Let $\mathcal{N}$ be an $\boldsymbol{\gamma}^2\epsilon$ cover of $\{\mathbf{A} \in \mathbb{R}^{d \times d} \ : \ \|\mathbf{A}\|_{\mathrm{F}} \le \sqrt{d}\}$. Then by Lemma F.1.10, $\log|\mathcal{N}| \le d^2 \log(1 + 2\sqrt{d}/(\boldsymbol{\gamma}^2\epsilon))$. Furthermore, for any $\boldsymbol{\Lambda} \succeq I$, we can find some $\mathbf{A} \in \mathcal{N}$ such that $\|\boldsymbol{\Lambda}^{-1} - \mathbf{A}\|_{\mathrm{F}} \le \boldsymbol{\gamma}^2\epsilon$. Let

$$\mathscr{V} := \left\{ r(\cdot) \ : \ r(\boldsymbol{\phi}) = \begin{cases} 1 & \|\boldsymbol{\phi}\|^2_{\mathbf{A}^{-1}} > \boldsymbol{\gamma}^2, \boldsymbol{\phi} \in \mathcal{X} \\ \boldsymbol{\gamma}^{-2}\|\boldsymbol{\phi}\|^2_{\mathbf{A}} & \|\boldsymbol{\phi}\|^2_{\mathbf{A}} \le \boldsymbol{\gamma}^2, \boldsymbol{\phi} \in \mathcal{X} \\ 0 & \boldsymbol{\phi} \notin \mathcal{X} \end{cases}, \mathbf{A} \in \mathcal{N} \right\},$$

then it follows that $\mathscr{V}$ is an ϵ-net of $\mathscr{R}$ in the $\operatorname{dist}_\infty$ norm, and that $\log|\mathscr{V}| \le d^2 \log(1 + 2\sqrt{d}/(\boldsymbol{\gamma}^2\epsilon))$. The result follows. □

Lemma G.1.4. *Let r^k be as defined in* EGS. *Then* FORCE *satisfies Definition 9.1.1 with*

$$\mathcal{C}_1 = d^4 H^3 \log(e + \sqrt{d}/\boldsymbol{\gamma}^2), \quad p_1 = 3$$

$$\mathcal{C}_2 = d^4 H^3 \log^{3/2}(e + \sqrt{d}/\boldsymbol{\gamma}^2), \quad p_2 = 7/2$$

Proof. This follows directly from Theorem 8 of Wagenmaker et al. [2022a] by noting that r_h^k is non-increasing in k, $\mathcal{F}_{k-1}$-measurable, $r_h^k \in \mathscr{R}$, and using the covering number for $\mathscr{R}$ given in Lemma G.1.3. □

G.2 Deferred Proofs from Section 9.4

G.2.1 Concentration of Least Squares Estimates

Lemma G.2.1 (Lemma B.2 of Jin et al. [2020b]). *When running RFLIN-PLAN, for all h,*

$$\|\widehat{\boldsymbol{w}}_h\|_2 \le 2H\sqrt{dK_{\text{tot}}}.$$

Proof. Note that our construction of $\widehat{\boldsymbol{w}}_h$ is identical to the construction of $\boldsymbol{w}_h^K$ in Jin et al. [2020b], so we can apply Lemma B.2 of Jin et al. [2020b]. □

Lemma G.2.2. *Consider the function class*

$$\mathscr{F}(\boldsymbol{\Lambda}, \alpha) := \left\{ V(\cdot) = \min\left\{ \max_{a \in \mathcal{A}} \langle \boldsymbol{w}, \boldsymbol{\phi}(\cdot, a)\rangle + \widetilde{\beta}\|\boldsymbol{\phi}(\cdot, a)\|_{\boldsymbol{\Lambda}^{-1}}, H \right\}, \|\boldsymbol{w}\|_2 \le \alpha \right\}$$

for some fixed $\Lambda \succ 0$ and $\widetilde{\beta} > 0$, $\alpha > 0$. Then

$$\mathsf{N}(\mathscr{F}(\boldsymbol{\Lambda}, \alpha), \text{dist}_\infty, \epsilon) \le d \log(1 + 2\alpha/\epsilon).$$

Proof. Take some $V_1, V_2 \in \mathscr{F}(\boldsymbol{\Lambda}, \alpha)$. Since both $\min\{\cdot, H\}$ and $\max_a$ are contraction maps, and $\boldsymbol{\phi}(s, a) \in \mathcal{B}^d$ for all s, a,

$$\sup_s |V_1(s) - V_2(s)| \le \sup_{\boldsymbol{\phi} \in \mathcal{B}^d} |\langle \boldsymbol{w}_1 - \boldsymbol{w}_2, \boldsymbol{\phi}\rangle| \le \|\boldsymbol{w}_1 - \boldsymbol{w}_2\|_2.$$

Let $\mathcal{N}$ denote an ϵ-cover of the α-ball, $\mathcal{B}^d(\alpha)$. By Lemma F.1.10, $\log|\mathcal{N}| \le d\log(1 + 2\alpha/\epsilon)$. By the above, we then have that for any $V \in \mathscr{F}(\boldsymbol{\Lambda}, \alpha)$, there exists some $V' \in \mathcal{N}$ such that $\text{dist}_\infty(V, V') \le \epsilon$, which completes the proof. □

Lemma G.2.3 (Lemma D.4 of Jin et al. [2020b]). *Let $\{s_\tau\}_{\tau=1}^\infty$ be a stochastic process on state space $\mathcal{S}$ with corresponding filtration $\{\mathcal{F}_\tau\}_{\tau=0}^\infty$. Let $\{\boldsymbol{\phi}_\tau\}_{\tau=0}^\infty$ be an $\mathbb{R}^d$-valued stochastic process where $\boldsymbol{\phi}_\tau \in \mathcal{F}_{\tau-1}$ and $\|\boldsymbol{\phi}_\tau\|_2 \le 1$. Let $\boldsymbol{\Lambda}_k = I + \sum_{\tau=1}^k \boldsymbol{\phi}_\tau \boldsymbol{\phi}_\tau^\top$. Then for any $\delta > 0$, with probability at least $1 - \delta$, for all $k \ge 0$, and any $V \in \mathscr{F}$ so that $\sup_s |V(s)| \le H$, we have:*

$$\left\| \boldsymbol{\Lambda}_k^{-1/2} \sum_{\tau=1}^k \boldsymbol{\phi}_\tau (V(s_\tau) - \mathbb{E}[V(s_\tau)|\mathcal{F}_{\tau-1}]) \right\|^2 \le 4H^2 \left(\frac{d}{2}\log(1+k) + \log\frac{|\mathcal{N}_\epsilon|}{\delta} \right) + 8k^2\epsilon^2$$

where $\mathcal{N}_\epsilon$ is the ϵ-covering of $\mathscr{F}$ with respect to dist_∞.

Lemma G.2.4 (Full version of Lemma 9.2.1)**.** *Let $\mathcal{E}_Q$ denote the event that, for all $h \in [H]$ and $V \in \mathscr{F}_{h+1}$ simultaneously,*

$$\left\|\mathbf{\Lambda}_h^{-1/2}\sum_{i=1}^{\varsigma_\epsilon}\sum_{\tau=1}^{K_i}\boldsymbol{\phi}_{h,\tau}^i\left[V(s_{h+1,\tau}^i)-\mathbb{E}_h[V](s_{h,\tau}^i,a_{h,\tau}^i)\right]\right\|_2 \le cH\sqrt{d\log(1+dHK_{\mathrm{tot}})+\log H/\delta}$$

for a universal constant c and where $K_{\mathrm{tot}} = \sum_{i=1}^{\varsigma_\epsilon} K_i$ and $\mathscr{F}_{h+1} := \mathscr{F}(\mathbf{\Lambda}_{h+1}, 2H\sqrt{dK_{\mathrm{tot}}})$. Then $\mathbb{P}[\mathcal{E}_Q] \ge 1-\delta$.

Proof. This is a consequence of Lemma G.2.2 and Lemma G.2.3. Fix some ϵ, then by Lemma G.2.2 we have

$$\mathsf{N}(\mathscr{F}_{h+1}, \mathrm{dist}_\infty, \epsilon) \le d\log(1+2H\sqrt{dK_{\mathrm{tot}}}/\epsilon)$$

and by Lemma G.2.3, with probability $1-\delta$, for all $V \in \mathscr{F}_{h+1}$ simultaneously,

$$\begin{aligned}\left\|\mathbf{\Lambda}_h^{-1/2}\sum_{i=1}^{\varsigma_\epsilon}\sum_{\tau=1}^{K_i}\boldsymbol{\phi}_{h,\tau}^i[V(s_{h+1,\tau}^i)-\mathbb{E}_h[V](s_{h,\tau}^i,a_{h,\tau}^i)]\right\|_2^2 &\le 4H^2\left(\frac{d}{2}\log(1+K_{\mathrm{tot}})+\log\frac{|\mathcal{N}_\epsilon|}{\delta}\right)+8K_{\mathrm{tot}}^2\epsilon^2\\ &\le 4H^2\left(\frac{d}{2}\log(1+K_{\mathrm{tot}})+d\log(1+\frac{2H\sqrt{dK_{\mathrm{tot}}}}{\epsilon})+\log\frac{1}{\delta}\right)+8K_{\mathrm{tot}}^2\epsilon^2.\end{aligned}$$

Note also that, due to our data collection procedure collecting samples independently for each h, we have that $\mathbf{\Lambda}_h$ and $\{\{(\boldsymbol{\phi}_{h,\tau}^i, s_{h+1,\tau}^i)\}_{\tau=1}^{K_i}\}_{i=1}^{\varsigma_\epsilon}$ are uncorrelated with $\mathbf{\Lambda}_{h+1}$, so, unlike in Jin et al. [2020b], no union bound over possible $\mathbf{\Lambda}_{h+1}$ is needed. Choosing $\epsilon = \sqrt{\frac{H^2 d}{8K_{\mathrm{tot}}^2}}$, we can bound this as

$$\le cH^2\left(d\log(1+dHK_{\mathrm{tot}})+\log\frac{1}{\delta}\right)$$

for a universal constant c. The result follows by a union bound over all h. □

G.2.2 Optimism

As noted above, throughout this section we will let $V^\star, V^\pi, Q^\star, Q^\pi$ denote value functions defined with respect to a generic reward function r, and V, Q the value function estimates maintained by RFLin-Plan when called with reward function r.

Proof of Lemma 9.4.2. First, note that for any r satisfying Definition 8.1.1, by Lemma G.2.1, we will have that $V_h \in \mathscr{F}_h$, where $\mathscr{F}_h$ is defined as in Lemma G.2.4. Therefore, on $\mathcal{E}_Q$, we

have for any r,

$$\left\| \mathbf{\Lambda}_h^{-1/2} \sum_{i=1}^{\varsigma_\epsilon} \sum_{\tau=1}^{K_i} \boldsymbol{\phi}_{h,\tau}^i \left[V_{h+1}(s_{h+1,\tau}^i) - \mathbb{E}_h[V_{h+1}](s_{h,\tau}^i, a_{h,\tau}^i) \right] \right\|_2 \leq \widetilde{\beta}. \tag{G.2.1}$$

By definition:

$$\widehat{\boldsymbol{w}}_h = \mathbf{\Lambda}_h^{-1} \sum_{i=1}^{\varsigma_\epsilon} \sum_{\tau=1}^{K_i} \boldsymbol{\phi}_{h,\tau}^i (r(s_{h,\tau}^i, a_{h,\tau}^i) + V_{h+1}(s_{h+1,\tau}^i)).$$

Furthermore, we recall that $r_h(s_{h,\tau}^i, a_{h,\tau}^i) = \langle \boldsymbol{\theta}_h, \boldsymbol{\phi}_{h,\tau}^i \rangle$. Thus,

$$\begin{aligned}
\langle \boldsymbol{\phi}(s,a), \widehat{\boldsymbol{w}}_h \rangle &= \langle \boldsymbol{\phi}(s,a), \mathbf{\Lambda}_h^{-1} \sum_{i=1}^{\varsigma_\epsilon} \sum_{\tau=1}^{K_i} \boldsymbol{\phi}_{h,\tau}^i (r(s_{h,\tau}^i, a_{h,\tau}^i) + V_{h+1}(s_{h+1,\tau}^i)) \rangle \\
&= \underbrace{\langle \boldsymbol{\phi}(s,a), \mathbf{\Lambda}_h^{-1} \sum_{i=1}^{\varsigma_\epsilon} \sum_{\tau=1}^{K_i} \boldsymbol{\phi}_{h,\tau}^i (\boldsymbol{\phi}_{h,\tau}^i)^\top \boldsymbol{\theta}_h \rangle}_{(a)} \\
&\quad + \underbrace{\langle \boldsymbol{\phi}(s,a), \mathbf{\Lambda}_h^{-1} \sum_{i=1}^{\varsigma_\epsilon} \sum_{\tau=1}^{K_i} \boldsymbol{\phi}_{h,\tau}^i (V_{h+1}(s_{h+1,\tau}^i) - \mathbb{E}[V_{h+1}](s_{h,\tau}^i, a_{h,\tau}^i)) \rangle}_{(b)} \\
&\quad + \underbrace{\langle \boldsymbol{\phi}(s,a), \mathbf{\Lambda}_h^{-1} \sum_{i=1}^{\varsigma_\epsilon} \sum_{\tau=1}^{K_i} \boldsymbol{\phi}_{h,\tau}^i \mathbb{E}[V_{h+1}](s_{h,\tau}^i, a_{h,\tau}^i) \rangle}_{(c)}.
\end{aligned}$$

Now,

$$(a) = \langle \boldsymbol{\phi}(s,a), \boldsymbol{\theta}_h \rangle - \langle \boldsymbol{\phi}(s,a), \mathbf{\Lambda}_h^{-1} \boldsymbol{\theta}_h \rangle = r_h(s,a) - \langle \boldsymbol{\phi}(s,a), \mathbf{\Lambda}_h^{-1} \boldsymbol{\theta}_h \rangle$$

and, using the linear MDP assumption, Definition 8.1.1,

$$\begin{aligned}
(c) &= \langle \boldsymbol{\phi}(s,a), \mathbf{\Lambda}_h^{-1} \sum_{i=1}^{\varsigma_\epsilon} \sum_{\tau=1}^{K_i} \boldsymbol{\phi}_{h,\tau}^i (\boldsymbol{\phi}_{h,\tau}^i)^\top \int V_{h+1}(s') \mathrm{d}\boldsymbol{\mu}_h(s') \rangle \\
&= \langle \boldsymbol{\phi}(s,a), \int V_{h+1}(s') \mathrm{d}\boldsymbol{\mu}_h(s') \rangle - \langle \boldsymbol{\phi}(s,a), \mathbf{\Lambda}_h^{-1} \int V_{h+1}(s') \mathrm{d}\boldsymbol{\mu}_h(s') \rangle \\
&= \mathbb{E}_h[V_{h+1}](s,a) - \langle \boldsymbol{\phi}(s,a), \mathbf{\Lambda}_h^{-1} \int V_{h+1}(s') \mathrm{d}\boldsymbol{\mu}_h(s') \rangle.
\end{aligned}$$

Thus,

$$|\langle \boldsymbol{\phi}(s,a), \widehat{\boldsymbol{w}}_h\rangle - r_h(s,a) - \mathbb{E}_h[V_{h+1}](s,a)| \le (b) + \left|\langle \boldsymbol{\phi}(s,a), \boldsymbol{\Lambda}_h^{-1}\boldsymbol{\theta}_h\rangle\right| + \left|\langle \boldsymbol{\phi}(s,a), \boldsymbol{\Lambda}_h^{-1}\int V_{h+1}(s')\mathrm{d}\boldsymbol{\mu}_h(s')\rangle\right|.$$

By Cauchy-Schwartz and (G.2.1):

$$(b) \le \|\boldsymbol{\phi}(s,a)\|_{\boldsymbol{\Lambda}_h^{-1}} \left\|\boldsymbol{\Lambda}_h^{-1/2}\sum_{i=1}^{\varsigma_\epsilon}\sum_{\tau=1}^{K_i}\boldsymbol{\phi}_{h,\tau}^i(V_{h+1}(s_{h+1,\tau}^i) - \mathbb{E}_h[V_{h+1}](s_{h,\tau}^i, a_{h,\tau}^i))\right\|_2 \le \widetilde{\beta}\|\boldsymbol{\phi}(s,a)\|_{\boldsymbol{\Lambda}_h^{-1}}.$$

Furthermore, we can bound

$$\left|\langle \boldsymbol{\phi}(s,a), \boldsymbol{\Lambda}_h^{-1}\boldsymbol{\theta}_h\rangle\right| \le \sqrt{d}\|\boldsymbol{\phi}(s,a)\|_{\boldsymbol{\Lambda}_h^{-1}}$$

and

$$\begin{aligned}\left|\langle \boldsymbol{\phi}(s,a), \boldsymbol{\Lambda}_h^{-1}\int V_{h+1}(s')\mathrm{d}\boldsymbol{\mu}_h(s')\rangle\right| &\le \|\boldsymbol{\phi}(s,a)\|_{\boldsymbol{\Lambda}_h^{-1}}\|\boldsymbol{\Lambda}_h^{-1/2}\|_{\mathrm{op}}\|\int V_{h+1}(s')\mathrm{d}\boldsymbol{\mu}_h(s')\|_2 \\ &\le H\sqrt{d}\|\boldsymbol{\phi}(s,a)\|_{\boldsymbol{\Lambda}_h^{-1}}\end{aligned}$$

where the final inequality uses the linear MDP assumption, Definition 8.1.1, and that $\boldsymbol{\Lambda}_h \succeq I$. The result follows by combining these bounds. □

Proof of Lemma 9.4.1. We will prove this by induction, starting at step H. Since the value function at $H+1$ is 0 by definition, $Q_H^\star(s,a) = r_H(s,a)$. Thus, Lemma 9.4.2 gives

$$|\langle \boldsymbol{\phi}(s,a), \widehat{\boldsymbol{w}}_H\rangle - Q_H^\star(s,a)| \le \widetilde{\beta}\|\boldsymbol{\phi}(s,a)\|_{\boldsymbol{\Lambda}_H^{-1}}.$$

It follows that, since $\beta \ge \widetilde{\beta}$,

$$Q_H^\star(s,a) \le \min\{\langle \boldsymbol{\phi}(s,a), \widehat{\boldsymbol{w}}_H\rangle + \beta\|\boldsymbol{\phi}(s,a)\|_{\boldsymbol{\Lambda}_H^{-1}}, H\} = Q_H(s,a).$$

Now assume that $Q_{h+1}(s,a) \ge Q_{h+1}^\star(s,a)$ for all s,a. By the Bellman equation,

$$Q_h^\star(s,a) = r_h(s,a) + \mathbb{E}_h[V_{h+1}^\star](s,a).$$

Thus, Lemma 9.4.2 gives

$$|\langle \boldsymbol{\phi}(s,a), \widehat{\boldsymbol{w}}_h\rangle - Q_h^\star(s,a) - \mathbb{E}_h[V_{h+1} - V_{h+1}^\star](s,a)| \le \beta\|\boldsymbol{\phi}(s,a)\|_{\boldsymbol{\Lambda}_h^{-1}}.$$

By the inductive assumption, $\mathbb{E}_h[V_{h+1} - V^\star_{h+1}](s,a) \geq 0$, so we can rearrange this to get

$$Q^\star_h(s,a) \leq \min\{\langle \boldsymbol{\phi}(s,a), \widehat{\boldsymbol{w}}_h \rangle + \widetilde{\beta} \|\boldsymbol{\phi}(s,a)\|_{\boldsymbol{\Lambda}_h^{-1}}, H\} = Q_h(s,a).$$

□

Appendix H

Deferred Proofs from Chapter 10

H.1 Additional Notation

Divergences. For probability distributions $\mathbb{P}$ and $\mathbb{Q}$ over a measurable space $(\Omega, \mathscr{F})$ with a common dominating measure, we define the total variation distance as

$$D_{\mathsf{TV}}(\mathbb{P}, \mathbb{Q}) = \sup_{A \in \mathscr{F}} |\mathbb{P}(A) - \mathbb{Q}(A)| = \frac{1}{2} \int |d\mathbb{P} - d\mathbb{Q}|$$

and (squared) Hellinger distance as

$$D^2_{\mathsf{H}}(\mathbb{P}, \mathbb{Q}) = \int \left(\sqrt{d\mathbb{P}} - \sqrt{d\mathbb{Q}} \right)^2.$$

Interactive decision making. We formalize probability spaces in the same fashion as [Foster et al., 2021, 2022]. Decisions are associated with a measurable space $(\Pi, \mathscr{P})$, rewards are associated with the space $(\mathcal{R}, \mathscr{R})$, and observations are associated with the space $(\mathcal{O}, \mathscr{O})$. The history after round t is denoted by $\mathcal{H}^t = (\pi^1, r^1, o^1), \ldots, (\pi^t, r^t, o^t)$. We define

$$\Omega^t = \prod_{i=1}^{t} (\Pi \times \mathcal{R} \times \mathcal{O}), \quad \text{and} \quad \mathscr{F}^t = \bigotimes_{i=1}^{t} (\mathscr{P} \otimes \mathscr{R} \otimes \mathscr{O})$$

so that $\mathcal{H}^t$ is associated with the space $(\Omega^t, \mathscr{F}^t)$.

When the algorithm is clear from context, we let $\mathbb{P}^M$ denote the law it induces on $\mathcal{H}^T$ when $M : \Pi \to \triangle_{\mathcal{R} \times \mathcal{O}}$ is the underlying model, and let $\mathbb{E}^M[\cdot]$ denote the corresponding expectation. We will also overload notation somewhat and let $\mathbb{P}^{M,\pi}$ the density of $(r, o) \sim M(\pi)$.

Notation for complexity measures and allocations. We let $T(\pi)$ denote the number of times decision π is taken up to time T. For $\eta \in \mathbb{R}^{\Pi}_{+}$, we define

$$I^M(\eta\,; \mathcal{M}) = \inf_{M' \in \mathcal{M}^{\mathrm{alt}}(M)} \sum_{\pi \in \Pi} \eta(\pi) D_{\mathsf{KL}}(M(\pi) \,\|\, M'(\pi)), \tag{H.1.1}$$

so that we can write $\mathsf{glc}(\mathcal{M}, M^\star) = \inf_{\eta \in \mathbb{R}_+^\Pi} \{\sum_{\pi \in \Pi} \eta(\pi) \Delta^{M^\star}(\pi) \mid I^{M^\star}(\eta; \mathcal{M}) \geq 1\}$. We abbreviate $I^M(\eta) = I^M(\eta; \mathcal{M})$ whenever the class $\mathcal{M}$ is clear from context. We will occasionally overload notation and write $\Delta^M(\eta) = \sum_{\pi \in \Pi} \eta(\pi) \Delta^M(\pi)$ for $\eta \in \mathbb{R}_+^\Pi$. We also let $\mathcal{M}_{\mathrm{alt}}^\star := \mathcal{M}^{\mathrm{alt}}(M^\star)$ and

$$\mathcal{M}_\varepsilon^{\mathrm{gl}}(\lambda; \mathsf{n}_{\max}) := \{M \in \mathcal{M} \ : \ \lambda \in \Upsilon(M; \varepsilon, \mathsf{n}_{\max})\}. \tag{H.1.2}$$

Recall the definition

$$\Upsilon(M; \varepsilon) = \Big\{ \lambda \in \triangle_\Pi \mid \exists \mathsf{n} \in \mathbb{R}_+ \text{ s.t. } \mathbb{E}_{\pi \sim \lambda}[\Delta^M(\pi)] \leq \frac{(1+\varepsilon)\mathsf{g}^M}{\mathsf{n}}, \\ \inf_{M' \in \mathcal{M}^{\mathrm{alt}}(M)} \mathbb{E}_{\pi \sim \lambda}[D_{\mathsf{KL}}(M(\pi) \,\|\, M'(\pi))] \geq \frac{1-\varepsilon}{\mathsf{n}} \Big\}.$$

For a given $\lambda \in \Upsilon(M; \varepsilon)$, we refer to the $\mathsf{n} \in \mathbb{R}_+$ which realizes

$$\mathbb{E}_{\pi \sim \lambda}[\Delta^M(\pi)] \leq \frac{(1+\varepsilon)\mathsf{g}^M}{\mathsf{n}} \quad \text{and} \quad \inf_{M' \in \mathcal{M}^{\mathrm{alt}}(M)} \mathbb{E}_{\pi \sim \lambda}[D_{\mathsf{KL}}(M(\pi) \,\|\, M'(\pi))] \geq \frac{1-\varepsilon}{\mathsf{n}}$$

as the *normalization factor* of λ.

General divergences. While in the main text we have focused on results that hold for the KL divergence, throughout the appendix we will consider more general divergences $D(\cdot \,\|\, \cdot)$. In particular, rather than fixing the divergence in (10.2.2) to the KL divergence, we utilize divergence D. We also perform online estimation with respect to D rather than with respect to the KL divergence. While D could be an arbitrary non-negative function, we make the following assumptions on it.

First, we replace Assumption 10.4 with the following more general assumption.

Assumption H.1. *For all $M, M', M'' \in \mathcal{M}$, and $\pi \in \Pi$, there exists some L_{KL} such that*

$$|D_{\mathsf{KL}}(M(\pi) \,\|\, M''(\pi)) - D_{\mathsf{KL}}(M'(\pi) \,\|\, M''(\pi))| \leq L_{\mathsf{KL}} \sqrt{D\big(M(\pi) \,\|\, M'(\pi)\big)}.$$

Note that, by Jensen's inequality, Assumption H.1 immediately implies that, for $\xi \in \triangle(\mathcal{M})$,

$$\big|D_{\mathsf{KL}}(M(\pi) \,\|\, M''(\pi)) - \mathbb{E}_{\overline{M} \sim \xi}[D_{\mathsf{KL}}\big(\overline{M}(\pi) \,\|\, M''(\pi)\big)]\big| \leq L_{\mathsf{KL}} \sqrt{\mathbb{E}_{\overline{M} \sim \xi}[D\big(\overline{M}(\pi) \,\|\, M(\pi)\big)]}.$$

We in addition make the following assumption, which we note is met for the KL divergence.

Assumption H.2. *For all $M, M' \in \mathrm{co}(\mathcal{M})$, and π, we have*

$$D^2_{\mathsf{H}}(M(\pi), M'(\pi)) \le D\big(M(\pi) \,\|\, M'(\pi)\big).$$

Furthermore, $D\big(\cdot \,\|\, \cdot\big)$ is convex in its second argument.

A direct consequence of this assumption is that, when rewards are observed and bounded in $[0,1]$, we have

$$|f^M(\pi) - f^{M'}(\pi)| \le \sqrt{D\big(M(\pi) \,\|\, M'(\pi)\big)}.$$

Throughout the appendix, we assume that Assumption H.1 and Assumption H.2 hold for our divergence D.

We also generalize the definition of the Allocation-Estimation Coefficient to account for general divergences as

$$\mathsf{aec}^{\mathsf{D}}_{\varepsilon}(\mathcal{M}, \overline{M}) := \inf_{\lambda, \omega \in \triangle_\Pi} \sup_{M \in \mathcal{M} \setminus \mathcal{M}^{\mathrm{gl}}_{\varepsilon}(\lambda)} \frac{1}{\mathbb{E}_{\pi \sim \omega}[D\big(\overline{M}(\pi) \,\|\, M(\pi)\big)]} \tag{H.1.3}$$

and

$$\overline{\mathsf{aec}}^{\mathsf{D}}_{\varepsilon}(\mathcal{M}; \xi) := \inf_{\lambda, \omega \in \triangle_\Pi} \sup_{M \in \mathcal{M} \setminus \mathcal{M}^{\mathrm{gl}}_{\varepsilon}(\lambda)} \frac{1}{\mathbb{E}_{\overline{M} \sim \xi}[\mathbb{E}_{\pi \sim \omega}[D\big(\overline{M}(\pi) \,\|\, M(\pi)\big)]]}. \tag{H.1.4}$$

Other variants of the AEC are generalized similarly with a superscript D. Our guarantees will also depend on a notion of estimation error for general divergences, given by

$$\mathbf{Est}_{\mathsf{D}}(s) := \sum_{i=1}^{s} \mathbb{E}_{\widehat{M} \sim \xi^i}[\mathbb{E}_{\pi \sim p^i}[D\big(M^\star(\pi) \,\|\, \widehat{M}(\pi)\big)]]. \tag{H.1.5}$$

It will also be convenient to work with the following notion of estimation error:

$$\widehat{\mathbf{Est}}_{\mathsf{D}}(s) := \sum_{i=1}^{s} \mathbb{E}_{\widehat{M} \sim \xi^i}[\mathbb{E}_{\pi \sim p^i}[D\big(\widehat{M}(\pi) \,\|\, M^\star(\pi)\big)]]. \tag{H.1.6}$$

H.2 Technical Tools

H.2.1 Online Learning

In this section, we state online estimation guarantees for variants of the Tempered Aggregation algorithm of Chen et al. [2022]. Throughout the section we abbreviate $\mathbb{E}_{t-1}[\cdot] = \mathbb{E}[\cdot \mid \mathcal{H}^{t-1}, p^t, \xi^t]$. We recall that $\mathbb{P}^{M,\pi}(r, o)$ denotes the density over rewards and observations (r, o) under $M(\pi)$

Algorithm H.1 Tempered Aggregation

input: Finite class $\mathcal{M}$.
Initialize $\xi^1 \leftarrow \mathrm{Unif}(\mathcal{M})$.
for $t = 1, 2, 3, \ldots$ **do**
    Receive (π^t, r^t, o^t).
    Update estimator:

$$\xi^{t+1}(M) \propto \xi^t(M) \cdot \exp\left(\frac{1}{2}\log \mathbb{P}^{M,\pi^t}(r^t, o^t)\right) \quad \forall M \in \mathcal{M}.$$

Proposition H.1 (Tempered Aggregation, Finite Class Setting). *Assume* $|\mathcal{M}| \leq \infty$ *and* $M^\star \in \mathcal{M}$. *Then Algorithm H.1 produces estimates* $(\xi^t)_{t=1}^T$ *which satisfy, with probability at least* $1-\delta$,

$$\sum_{t=1}^{T} \mathbb{E}_{M\sim\xi^t}[\mathbb{E}_{\pi\sim p^t}[D^2_{\mathsf{H}}(M^\star(\pi), M(\pi))]] \leq 2\log\frac{|\mathcal{M}|}{\delta}.$$

Proof of Proposition H.1. We follow closely the proof of Theorem C.1 of Chen et al. [2022]. Define the random variable

$$A^t := -\log \mathbb{E}_{M\sim\xi^t}\left[\exp\left(\beta \log \frac{\mathbb{P}^{M,\pi^t}(r^t, o^t)}{\mathbb{P}^{M^\star,\pi^t}(r^t, o^t)}\right)\right].$$

We have

$$\begin{aligned}
\mathbb{E}_{t-1}[\exp(-A^t)] &= \mathbb{E}_{t-1}\left[\mathbb{E}_{M\sim\xi^t}\left[\exp\left(\frac{1}{2}\log \frac{\mathbb{P}^{M,\pi^t}(r^t, o^t)}{\mathbb{P}^{M^\star,\pi^t}(r^t, o^t)}\right)\right]\right] \\
&= \sum_{M\in\mathcal{M}} \xi^t(M)\mathbb{E}_{t-1}\left[\exp\left(\frac{1}{2}\log \frac{\mathbb{P}^{M,\pi^t}(r^t, o^t)}{\mathbb{P}^{M^\star,\pi^t}(r^t, o^t)}\right)\right] \\
&= \sum_{M\in\mathcal{M}} \xi^t(M)\mathbb{E}_{t-1}\left[\mathbb{E}_{o\sim M^\star(\pi^t)}\left[\sqrt{\frac{\mathbb{P}^{M,\pi^t}(r^t, o^t)}{\mathbb{P}^{M^\star,\pi^t}(r^t, o^t)}}\right]\right] \\
&= \sum_{M\in\mathcal{M}} \xi^t(M)\cdot\left(1 - \frac{1}{2}\mathbb{E}_{\pi\sim p^t}[D^2_{\mathsf{H}}(M^\star(\pi), M(\pi))]\right)
\end{aligned}$$

where the last equality holds by the definition of the Hellinger distance. This implies that

$$1 - \mathbb{E}_{t-1}[\exp(-A^t)] = \frac{1}{2}\mathbb{E}_{M\sim\xi^t}\left[\mathbb{E}_{\pi\sim p^t}[D^2_{\mathsf{H}}(M^\star(\pi), M(\pi))]\right].$$

By Lemma A.4 of Foster et al. [2021], we have that with probability at least $1-\delta$,

$$\begin{aligned}\sum_{t=1}^{T} A^t + \log\frac{1}{\delta} &\geq \sum_{t=1}^{T} -\log \mathbb{E}_{t-1}[\exp(-A^t)] \\ &\geq \sum_{t=1}^{T}\left(1 - \mathbb{E}_{t-1}[\exp(-A^t)]\right) \\ &= \frac{1}{2}\sum_{t=1}^{T} \mathbb{E}_{M\sim\xi^t}\left[\mathbb{E}_{\pi\sim p^t}[D^2_{\mathsf{H}}(M^\star(\pi), M(\pi))]\right].\end{aligned}$$

We turn to upper bounding $\sum_{t=1}^{T} A^t$. Following the proof of Theorem C.1 of Chen et al. [2022], we have

$$\sum_{t=1}^{T} A^t = -\log\left(\sum_{M\in\mathcal{M}} \xi^1(M)\exp\left(\sum_{t=1}^{T}\frac{1}{2}\log\frac{\mathbb{P}^{M,\pi^t}(r^t,o^t)}{\mathbb{P}^{M^\star,\pi^t}(r^t,o^t)}\right)\right).$$

Since $M^\star \in \mathcal{M}$, we can then bound

$$\sum_{t=1}^{T} A^t \leq -\log\left(\xi^1(M^\star)\exp\left(\sum_{t=1}^{T}\frac{1}{2}\log\frac{\mathbb{P}^{M^\star,\pi^t}(r^t,o^t)}{\mathbb{P}^{M^\star,\pi^t}(r^t,o^t)}\right)\right) = \log|\mathcal{M}|.$$

Combining these expressions gives the result. □

Proposition H.2 (Tempered Aggregation, Infinite Class Setting)**.** *Let $\mathcal{M}_{\mathrm{cov}}$ denote a (ρ,μ)-cover of $\mathcal{M}$ with covering number $\mathsf{N}_{\mathrm{cov}}(\mathcal{M},\rho,\mu)$. If we apply Algorithm H.1 to $\mathcal{M}_{\mathrm{cov}}$, we have that whenever $M^\star\in\mathcal{M}$, with probability at least $1-\delta-T\mu$,*

$$\sum_{t=1}^{T} \mathbb{E}_{M\sim\xi^t}[\mathbb{E}_{\pi\sim p^t}[D^2_{\mathsf{H}}(M^\star(\pi),M(\pi))]] \leq 2\log\frac{\mathsf{N}_{\mathrm{cov}}(\mathcal{M},\rho,\mu)}{\delta} + T\rho.$$

Proof of Proposition H.2. Defining A^t as in Proposition H.1, the first part of the proof is identical to that of Proposition H.1, and we conclude that, with probability at least $1-\delta$,

$$\sum_{t=1}^{T} A^t + \log\frac{1}{\delta} \geq \frac{1}{2}\sum_{t=1}^{T} \mathbb{E}_{M\sim\xi^t}[\mathbb{E}_{\pi\sim p^t}[D^2_{\mathsf{H}}(M^\star(\pi),M(\pi))]]$$

and

$$\sum_{t=1}^{T} A^t = -\log\left(\sum_{M\in\mathcal{M}_{\mathrm{cov}}} \xi^1(M)\exp\left(\sum_{t=1}^{T}\frac{1}{2}\log\frac{\mathbb{P}^{M,\pi^t}(r^t,o^t)}{\mathbb{P}^{M^\star,\pi^t}(r^t,o^t)}\right)\right).$$

Let $\mathcal{E}$ denote the event associated with $\mathcal{M}_{\text{cov}}$, as defined in Definition 10.2.1, which satisfies $\sup_{M' \in \mathcal{M}} \sup_{\pi} \mathbb{P}^{M'}(\mathcal{E} \mid \pi) \leq \mu$. Let $\widetilde{M} \in \mathcal{M}_{\text{cov}}$ denote the element in the cover which satisfies

$$\left| \log \frac{\mathbb{P}^{M^\star, \pi}(r, o)}{\mathbb{P}^{\widetilde{M}, \pi}(r, o)} \right| = \left| \log \mathbb{P}^{M^\star, \pi}(r, o) - \log \mathbb{P}^{\widetilde{M}, \pi}(r, o) \right| \leq \rho,$$

for all (r, o, π) with $\sup_{M' \in \mathcal{M}} \mathbb{P}^{M', \pi}(r, o \mid \mathcal{E}) > 0$. We then have

$$\sum_{t=1}^{T} \log \frac{\mathbb{P}^{M, \pi^t}(r^t, o^t)}{\mathbb{P}^{M^\star, \pi^t}(r^t, o^t)} = \sum_{t=1}^{T} \left(\log \frac{\mathbb{P}^{M, \pi^t}(r^t, o^t)}{\mathbb{P}^{\widetilde{M}, \pi^t}(r^t, o^t)} + \log \frac{\mathbb{P}^{\widetilde{M}, \pi^t}(r^t, o^t)}{\mathbb{P}^{M^\star, \pi^t}(r^t, o^t)} \right)$$

so we can bound

$$\sum_{t=1}^{T} A^t \leq \log |\mathcal{M}_{\text{cov}}| + \frac{1}{2} \sum_{t=1}^{T} \log \frac{\mathbb{P}^{M^\star, \pi^t}(r^t, o^t)}{\mathbb{P}^{\widetilde{M}, \pi^t}(r^t, o^t)}$$

which gives that with probability at least $1 - \delta$,

$$\sum_{t=1}^{T} \mathbb{E}_{M \sim \xi^t}[\mathbb{E}_{\pi \sim p^t}[D^2_{\mathsf{H}}(M^\star(\pi), M(\pi))]] \leq 2 \log |\mathcal{M}_{\text{cov}}| + 2 \log \frac{1}{\delta} + \sum_{t=1}^{T} \log \frac{\mathbb{P}^{M^\star, \pi^t}(r^t, o^t)}{\mathbb{P}^{\widetilde{M}, \pi^t}(r^t, o^t)}.$$

Let $\mathcal{E}_t$ denote the event that $\mathcal{E}$ occurs at step t. Denote the event $\mathcal{E}_1 := \cap_{t=1}^{T} \mathcal{E}_t$ and

$$\mathcal{E}_2 := \left\{ \sum_{t=1}^{T} \mathbb{E}_{M \sim \xi^t}[\mathbb{E}_{\pi \sim p^t}[D^2_{\mathsf{H}}(M^\star(\pi), M(\pi))]] \leq 2 \log |\mathcal{M}_{\text{cov}}| + 2 \log \frac{1}{\delta} + T\rho \right\}.$$

Then

$$\mathbb{P}^{M^\star}[\mathcal{E}_2] \leq \mathbb{P}^{M^\star}[\mathcal{E}_2 \cap \mathcal{E}_1] + \mathbb{P}^{M^\star}[\mathcal{E}_1^c].$$

By definition of $\mathcal{E}_t$ and a union bound we have $\mathbb{P}^{M^\star}[\mathcal{E}_1^c] \leq T\mu$. Furthermore, on the event $\mathcal{E}_1$ we can bound, for each $t \leq T$,

$$\log \frac{\mathbb{P}^{M^\star, \pi^t}(r^t, o^t)}{\mathbb{P}^{\widetilde{M}, \pi^t}(r^t, o^t)} \leq \rho.$$

Thus, it follows that on $\mathcal{E}_1$, $\sum_{t=1}^{T} \log \frac{\mathbb{P}^{M^\star, \pi^t}(r^t, o^t)}{\mathbb{P}^{\widetilde{M}, \pi^t}(r^t, o^t)} \leq T\rho$, which implies that $\mathbb{P}^{M^\star}[\mathcal{E}_2 \cap \mathcal{E}_1] \leq \delta$. The result follows.

□

H.2.2 Properties of Graves-Lai Program

In this section, we establish some basic properties of the Graves-Lai coefficient $\mathsf{g}^M \equiv \mathsf{glc}(\mathcal{M}, M)$. Through out the section, we omit dependence on the class $\mathcal{M}$ for various quantities of interest whenever it is clear from context. Throughout, we will use the fact that whenever $\Delta^M_{\min} > 0$, π_M is unique.

H.2.2.1 Basic Properties of Graves-Lai Program

Lemma H.2.1. *For any $M \in \mathcal{M}$ and $\mathsf{n} > 0$, we have*

$$\frac{\mathsf{g}^M}{\mathsf{n}} \le \inf_{\lambda \in \triangle_\Pi} \left\{ \Delta^M(\lambda) : I^M(\lambda) \ge \frac{1}{\mathsf{n}} \right\}.$$

Proof of Lemma H.2.1. Assume the contrary. Then there exists some $\widetilde{\lambda}$ such that

$$\Delta^M(\widetilde{\lambda}) < \mathsf{g}^M/\mathsf{n} \quad \text{and} \quad I^M(\widetilde{\lambda}) \ge 1/\mathsf{n}.$$

However, since both $\Delta^M(\lambda)$ and $I^M(\lambda)$ are linear in rescaling of λ, this implies that

$$\Delta^M(\mathsf{n}\widetilde{\lambda}) < \mathsf{g}^M \quad \text{and} \quad I^M(\mathsf{n}\widetilde{\lambda}) \ge 1.$$

By definition we have

$$\mathsf{g}^M = \inf_{\eta \in \mathbb{R}^\Pi_+} \Delta^M(\eta) \quad \text{s.t.} \quad I^M(\eta) \ge 1.$$

This is a contradiction, so the desired conclusion follows. □

Lemma H.2.2. *For any $M \in \mathcal{M}^+$ with $\Delta^M_{\min} > 0$, we can bound*

$$\mathsf{g}^M \le C^\xi_{\exp}(\tfrac{1}{4}(\Delta^M_{\min})^2)$$

for $\xi = \mathbb{I}_M$.

Proof of Lemma H.2.2. By definition we have

$$\begin{aligned}
\mathsf{g}^M &= \inf_{\eta \in \mathbb{R}^\Pi_+} \Delta^M(\eta) \quad \text{s.t.} \quad \inf_{M' \in \mathcal{M}^{\mathrm{alt}}(M)} \sum_\pi \eta(\pi) D_{\mathsf{KL}}(M(\pi) \,\|\, M'(\pi)) \ge 1 \\
&\le \inf_{\eta \in \mathbb{R}^\Pi_+} \|\eta\|_1 \quad \text{s.t.} \quad \inf_{M' \in \mathcal{M}^{\mathrm{alt}}(M)} \sum_\pi \eta(\pi) D_{\mathsf{KL}}(M(\pi) \,\|\, M'(\pi)) \ge 1.
\end{aligned}$$

Let $\xi \in \triangle_{\mathcal{M}}$ denote the distribution with $\xi(M) = 1$. Let and let $p_{\exp} := p^\xi_{\exp}(\varepsilon)$ denote the uniform exploration distribution defined with respect to ξ, and $C^\xi_{\exp}(\varepsilon)$ the corresponding uniform exploration coefficient, for some ε to be chosen.

Consider some $M' \in \mathcal{M}^{\text{alt}}(M)$. Since $\mathbb{E}_{\overline{M}\sim\xi}[\mathbb{E}_{\pi\sim p}[D_{\mathsf{KL}}\big(\overline{M}(\pi) \,\|\, M''(\pi)\big)]] = \mathbb{E}_{\pi\sim p}[D_{\mathsf{KL}}(M(\pi) \,\|\, M''(\pi))]$ for all M'' and $D_{\mathsf{KL}}(M(\pi) \,\|\, M(\pi)) = 0$ for all π, it follows from the definition of the uniform exploration coefficient that

$$\mathbb{E}_{p_{\text{exp}}}[D_{\mathsf{KL}}(M(\pi) \,\|\, M'(\pi))] \le 1/C^{\xi}_{\text{exp}}(\varepsilon) \implies \sup_{p\in\triangle_\Pi} \mathbb{E}_p[D_{\mathsf{KL}}(M(\pi) \,\|\, M'(\pi))] \le \varepsilon$$

or, alternatively,

$$\sup_{p\in\triangle_\Pi} \mathbb{E}_p[D_{\mathsf{KL}}(M(\pi) \,\|\, M'(\pi))] > \varepsilon \implies \mathbb{E}_{p_{\text{exp}}}[D_{\mathsf{KL}}(M(\pi) \,\|\, M'(\pi))] > 1/C^{\xi}_{\text{exp}}(\varepsilon).$$

If $M' \in \mathcal{M}^{\text{alt}}(M)$, then it follows that $\pi_M \notin \boldsymbol{\pi}_{M'}$. Take some $\pi_{M'} \in \boldsymbol{\pi}_{M'}$. By definition we have $f^M(\pi_M) \ge f^M(\pi_{M'}) + \Delta^M_{\text{min}}$ and $f^{M'}(\pi_{M'}) \ge f^{M'}(\pi_M)$. Thus,

$$\begin{aligned}
\Delta^M_{\text{min}} &\le f^M(\pi_M) - f^M(\pi_{M'}) + f^{M'}(\pi_{M'}) - f^{M'}(\pi_M) \\
&\le |f^M(\pi_M) - f^{M'}(\pi_M)| + |f^{M'}(\pi_{M'}) - f^M(\pi_{M'})| \\
&\le \sqrt{D\big(M(\pi_M) \,\|\, M'(\pi_M)\big)} + \sqrt{D\big(M(\pi_{M'}) \,\|\, M'(\pi_{M'})\big)}.
\end{aligned}$$

This implies that there exists some π such that $D\big(M(\pi) \,\|\, M'(\pi)\big) \ge (\Delta^M_{\text{min}}/2)^2$, so we can lower bound

$$\sup_{p\in\triangle_\Pi} \mathbb{E}_p[D_{\mathsf{KL}}(M(\pi) \,\|\, M'(\pi))] \ge \frac{1}{4}(\Delta^M_{\text{min}})^2.$$

Thus, setting $\varepsilon = \frac{1}{4}(\Delta^M_{\text{min}})^2$, we have that

$$\mathbb{E}_{p_{\text{exp}}}[D_{\mathsf{KL}}(M(\pi) \,\|\, M'(\pi))] > 1/C^{\xi}_{\text{exp}}(\tfrac{1}{4}(\Delta^M_{\text{min}})^2).$$

It follows that the allocation $\eta = C^{\xi}_{\text{exp}}(\frac{1}{4}(\Delta^M_{\text{min}})^2) \cdot p_{\text{exp}}$ realizes

$$\inf_{M'\in\mathcal{M}^{\text{alt}}(M)} \sum_\pi \eta(\pi) D_{\mathsf{KL}}(M(\pi) \,\|\, M'(\pi)) = \inf_{M'\in\mathcal{M}^{\text{alt}}(M)} C^{\xi}_{\text{exp}}(\tfrac{1}{4}(\Delta^M_{\text{min}})^2) \cdot \mathbb{E}_{p^M_{\text{exp}}}[D_{\mathsf{KL}}(M(\pi) \,\|\, M'(\pi))] \ge 1,$$

which proves the result. □

Lemma H.2.3. *Assume* $\mathsf{g}^M > 0, \Delta^M_{\text{min}} > 0,$ *and* $\mathsf{n}^M_{1/4} < \infty$. *Then it must be the case that*

$$\mathsf{g}^M \ge \Delta^M_{\text{min}} \cdot \frac{1}{\max_{M'\in\mathcal{M},\pi\in\Pi} D_{\mathsf{KL}}(M(\pi) \,\|\, M'(\pi))}.$$

Proof of Lemma H.2.3. Recall that

$$\mathsf{g}^M = \inf_{\eta\in\mathbb{R}^\Pi_+} \Delta^M(\eta) \quad \text{s.t.} \quad \inf_{M'\in\mathcal{M}^{\text{alt}}(M)} \sum_\pi \eta(\pi) D_{\mathsf{KL}}(M(\pi) \,\|\, M'(\pi)) \ge 1.$$

By the definition of $\mathsf{n}^M_{1/4}$, for any allocation $\eta \in \mathbb{R}^\Pi_+$ satisfying $\inf_{M' \in \mathcal{M}^{\mathrm{alt}}(M)} \sum_\pi \eta(\pi) D_{\mathsf{KL}}(M(\pi) \,\|\, M'(\pi)) \geq 3/4$, the allocation $\widetilde{\eta}$ defined as $\widetilde{\eta}(\pi) = \eta(\pi)$ for $\pi \neq \pi_M$, and $\widetilde{\eta}(\pi_M) = \mathsf{n}^M_{1/2}$ satisfies

$$\inf_{M' \in \mathcal{M}^{\mathrm{alt}}(M)} \sum_\pi \widetilde{\eta}(\pi) D_{\mathsf{KL}}(M(\pi) \,\|\, M'(\pi)) \geq 1/2,$$

and furthermore $\Delta^M(\eta) = \Delta^M(\widetilde{\eta})$. It follows that

$$\begin{aligned}
\mathsf{g}^M &\geq \inf_{\eta \in \mathbb{R}^\Pi_+} \Delta^M(\eta) \quad \text{s.t.} \quad \inf_{M' \in \mathcal{M}^{\mathrm{alt}}(M)} \sum_\pi \eta(\pi) D_{\mathsf{KL}}(M(\pi) \,\|\, M'(\pi)) \geq 3/4 \\
&\geq \inf_{\eta \in \mathbb{R}^\Pi_+} \Delta^M(\eta) \quad \text{s.t.} \quad \inf_{M' \in \mathcal{M}^{\mathrm{alt}}(M)} \sum_\pi \eta(\pi) D_{\mathsf{KL}}(M(\pi) \,\|\, M'(\pi)) \geq 1/2,\ \eta(\pi_M) \leq \mathsf{n}^M_{1/4}.
\end{aligned}$$

If for all $M' \in \mathcal{M}^{\mathrm{alt}}(M)$ we have $D_{\mathsf{KL}}(M(\pi_M) \,\|\, M'(\pi_M)) > 0$, this implies that we can distinguish M from M' by playing only π_M. Furthermore, by what we have just shown, this can be achieved by playing π_M at most $2\mathsf{n}^M_{1/4}$ times. It follows that, if $D_{\mathsf{KL}}(M(\pi_M) \,\|\, M'(\pi_M)) > 0$ for all $M' \in \mathcal{M}^{\mathrm{alt}}(M)$, then $\mathsf{g}^M = 0$. Thus, if $\mathsf{g}^M > 0$, there must exist some $M' \in \mathcal{M}^{\mathrm{alt}}(M)$ such that $D_{\mathsf{KL}}(M(\pi_M) \,\|\, M'(\pi_M)) = 0$.

We then have

$$\begin{aligned}
\mathsf{g}^M &\geq \inf_{\eta \in \mathbb{R}^\Pi_+} \Delta^M(\eta) \quad \text{s.t.} \quad \sum_{\pi \neq \pi_M} \eta(\pi) D_{\mathsf{KL}}(M(\pi) \,\|\, M'(\pi)) \geq 1 \\
&= \inf_{\eta \in \mathbb{R}^\Pi_+, \eta(\pi_M) = 0} \Delta^M(\eta) \quad \text{s.t.} \quad \sum_{\pi \neq \pi_M} \eta(\pi) D_{\mathsf{KL}}(M(\pi) \,\|\, M'(\pi)) \geq 1 \\
&\geq \inf_{\eta \in \mathbb{R}^\Pi_+, \eta(\pi_M) = 0} \Delta^M_{\min} \cdot \|\eta\|_1 \quad \text{s.t.} \quad \max_\pi D_{\mathsf{KL}}(M(\pi) \,\|\, M'(\pi)) \cdot \|\eta\|_1 \geq 1 \\
&= \Delta^M_{\min} \cdot \frac{1}{\max_\pi D_{\mathsf{KL}}(M(\pi) \,\|\, M'(\pi))}.
\end{aligned}$$

The result follows. □

H.2.2.2 Properties of the Information Content of Optimal Decisions

Lemma H.2.4. *Let $\varepsilon \in [0, 1/2)$ and $\bar{\mathsf{n}} > 0$ be given. We can bound, for any function $g(\omega, M)$ and $\mathcal{M}_0 \subseteq \mathcal{M}$ with $\inf_{M \in \mathcal{M}_0} \Delta^M_{\min} > 0$,*

$$\inf_{\omega, \lambda \in \triangle_\Pi} \sup_{M \in \mathcal{M}_0 \setminus \mathcal{M}^{\mathrm{gl}}_{2\varepsilon}(\lambda; \bar{\mathsf{n}})} g(\omega, M) \leq \inf_{\omega, \lambda \in \triangle_\Pi} \sup_{M \in \mathcal{M}_0 \setminus \mathcal{M}^{\mathrm{gl}}_{\varepsilon}(\lambda)} g(\omega, M)$$

as long as

$$\bar{\mathsf{n}} \geq \max_{M \in \mathcal{M}_0} \max \left\{ \mathsf{n}^M_\varepsilon, \frac{4\mathsf{g}^M}{\Delta^M_{\min}}, \frac{2\mathsf{g}^M}{\zeta \Delta^M_{\min}} \right\},$$

where

$$\zeta := \min_{M \in \mathcal{M}_0 : \mathsf{g}^M > 0} \min\left\{ \frac{\mathsf{g}^M}{\mathsf{g}^M + 2\mathsf{n}_\varepsilon^M}, \frac{\Delta_{\min}^M \varepsilon}{4} \right\}.$$

Proof of Lemma H.2.4. Note that each of these expressions only depend on λ through $\mathcal{M}_{2\varepsilon}^{\mathrm{gl}}(\lambda; \bar{\mathsf{n}})$ and $\mathcal{M}_{\varepsilon}^{\mathrm{gl}}(\lambda)$, respectively. To prove the result, it therefore suffices to show that, for every $\lambda \in \triangle_\Pi$, there exists $\lambda' \in \triangle_\Pi$ such that $\mathcal{M}_0 \cap \mathcal{M}_{\varepsilon}^{\mathrm{gl}}(\lambda) \subseteq \mathcal{M}_0 \cap \mathcal{M}_{2\varepsilon}^{\mathrm{gl}}(\lambda'; \bar{\mathsf{n}})$.

Fix $\lambda \in \triangle_\Pi$. Consider $M \in \mathcal{M}_0 \cap \mathcal{M}_{\varepsilon}^{\mathrm{gl}}(\lambda)$. By definition we have that there exists some $\mathsf{n} > 0$ such that

$$\Delta^M(\lambda) \le (1+\varepsilon)\mathsf{g}^M/\mathsf{n} \quad \text{and} \quad I^M(\lambda) \ge (1-\varepsilon)/\mathsf{n},$$

where here $I^M(\lambda) = I^M(\lambda; \mathcal{M})$. We consider two cases.

Case 1: $\lambda = \mathbb{I}_\pi$ for some $\pi \in \Pi$. First, suppose that $\pi \neq \pi_M$. Then we have

$$\Delta_{\min}^M \le \Delta^M(\lambda) \le (1+\varepsilon)\mathsf{g}^M/\mathsf{n} \implies \mathsf{n} \le (1+\varepsilon)\mathsf{g}^M/\Delta_{\min}^M.$$

It follows that as long as $\bar{\mathsf{n}} \ge (1+\varepsilon)\mathsf{g}^M/\Delta_{\min}^M$, then $M \in \mathcal{M}_{\varepsilon}^{\mathrm{gl}}(\lambda; \bar{\mathsf{n}})$.

Now, suppose that $\pi = \pi_M$. In this case, by the definition of n_ε^M, we immediately have that it suffices to take $\mathsf{n} = \mathsf{n}_\varepsilon^M$. Thus, for $\lambda = \mathbb{I}_\pi$, we have $\mathcal{M}_{2\varepsilon}^{\mathrm{gl}}(\lambda) = \mathcal{M}_{\varepsilon}^{\mathrm{gl}}(\lambda; \bar{\mathsf{n}})$ as long as

$$\bar{\mathsf{n}} \ge \max\{\mathsf{n}_\varepsilon^M, (1+\varepsilon)\mathsf{g}^M/\Delta_{\min}^M\}.$$

Case 2a: $\lambda \neq \mathbb{I}_\pi$ for any $\pi \in \Pi$. Fix some $\zeta \in (0, 1/2)$ to be chosen. Suppose that there exists some π' such that $\lambda(\pi') \ge 1 - \zeta$, and note that there can exist at most one such π'. Define λ' as

$$\lambda'(\pi') = 1 - \zeta, \quad \lambda'(\pi) = \frac{\zeta}{1 - \lambda(\pi')}\lambda(\pi), \quad \forall \pi \neq \pi'$$

and note that $\lambda' \in \triangle_\Pi$. Our goal is to show that $\mathcal{M}_{\varepsilon}^{\mathrm{gl}}(\lambda) \subseteq \mathcal{M}_{2\varepsilon}^{\mathrm{gl}}(\lambda'; \bar{\mathsf{n}})$.

Suppose that $\pi_M = \pi'$. Then

$$\Delta^M(\lambda') = \frac{\zeta}{1 - \lambda(\pi')} \cdot \Delta^M(\lambda) \le \frac{\zeta}{1 - \lambda(\pi')} \cdot \frac{(1+\varepsilon)\mathsf{g}^M}{\mathsf{n}}.$$

Denote $\mathsf{n}' := (\frac{\zeta}{1-\lambda(\pi')} \cdot \frac{1}{\mathsf{n}})^{-1}$. Since $I^M(\lambda) \ge (1-\varepsilon)/\mathsf{n}$, we have $I^M(\mathsf{n}\lambda) \ge 1 - \varepsilon$. Then, by the definition of n_ε^M, we have

$$\inf_{M' \in \mathcal{M}^{\mathrm{alt}}(M)} \sum_{\pi \neq \pi_M} \mathsf{n}\lambda(\pi) D_{\mathsf{KL}}(M(\pi) \,\|\, M'(\pi)) + \mathsf{n}_\varepsilon^M D_{\mathsf{KL}}(M(\pi_M) \,\|\, M'(\pi_M)) \ge 1 - 2\varepsilon.$$

However, note that

$$\begin{aligned}
I^M(\mathsf{n}'\lambda') &= \inf_{M'\in\mathcal{M}^{\mathrm{alt}}(M)} \sum_{\pi\neq\pi_M} \frac{\zeta\mathsf{n}'}{1-\lambda(\pi_M)}\lambda(\pi)D_{\mathsf{KL}}(M(\pi)\,\|\,M'(\pi)) + \mathsf{n}'(1-\zeta)D_{\mathsf{KL}}(M(\pi_M)\,\|\,M'(\pi_M)) \\
&= \inf_{M'\in\mathcal{M}^{\mathrm{alt}}(M)} \sum_{\pi\neq\pi_M} \mathsf{n}\lambda(\pi)D_{\mathsf{KL}}(M(\pi)\,\|\,M'(\pi)) + \frac{(1-\zeta)(1-\lambda(\pi'))\mathsf{n}}{\zeta}\cdot D_{\mathsf{KL}}(M(\pi_M)\,\|\,M'(\pi_M)) \\
&\overset{(a)}{\geq} \inf_{M'\in\mathcal{M}^{\mathrm{alt}}(M)} \sum_{\pi\neq\pi_M} \mathsf{n}\lambda(\pi)D_{\mathsf{KL}}(M(\pi)\,\|\,M'(\pi)) + \mathsf{n}_\varepsilon^M D_{\mathsf{KL}}(M(\pi_M)\,\|\,M'(\pi_M)) \\
&\overset{(b)}{\geq} 1-2\varepsilon
\end{aligned}$$

where (a) follows as long as

$$\frac{(1-\zeta)(1-\lambda(\pi'))\mathsf{n}}{\zeta} \geq \mathsf{n}_\varepsilon^M, \tag{H.2.1}$$

and (b) follows from what we have just shown. Rearranging, we have that (H.2.1) is equivalent to

$$\frac{(1-\lambda(\pi'))\mathsf{n}}{(1-\lambda(\pi'))\mathsf{n}+\mathsf{n}_\varepsilon^M} \geq \zeta.$$

Note that by Lemma H.2.1 and the definition of λ, we have

$$\frac{(1-\varepsilon)\mathsf{g}^M}{\mathsf{n}} \leq \inf_{\widetilde{\lambda}\in\triangle_\Pi}\left\{\Delta^M(\widetilde{\lambda}) : I^M(\widetilde{\lambda}) \geq \frac{1-\varepsilon}{\mathsf{n}}\right\} \leq \Delta^M(\lambda),$$

which implies that

$$(1-\varepsilon)\mathsf{g}^M \leq \Delta^M(\lambda)\cdot\mathsf{n} \leq (1-\lambda(\pi_M))\cdot\mathsf{n}.$$

As the function $\frac{x}{x+\mathsf{n}_\varepsilon^M}$ is increasing in x, a sufficient choice of ζ for this M is then

$$\min\{\frac{\mathsf{g}^M}{\mathsf{g}^M+2\mathsf{n}_\varepsilon^M}, 3/8\} \geq \zeta$$

Thus, for such a ζ, we have that $M\in\mathcal{M}_\varepsilon^{\mathrm{gl}}(\lambda';\mathsf{n}')$, which implies that $M\in\mathcal{M}_{2\varepsilon}^{\mathrm{gl}}(\lambda';\mathsf{n}')$. Note that $(1-\lambda(\pi'))\mathsf{n} \leq (1+\varepsilon)\mathsf{g}^M/\Delta_{\min}^M$ in this case, so we have that $M\in\mathcal{M}_{2\varepsilon}^{\mathrm{gl}}(\lambda';\bar{\mathsf{n}})$ as long as

$$\bar{\mathsf{n}} \geq \frac{2\mathsf{g}^M}{\zeta\Delta_{\min}^M}.$$

Consider now the case where $\pi_M \neq \pi'$. In this case, defining λ' as before, we can bound

$$\Delta^M(\lambda') \leq \zeta + (1-\zeta)\Delta^M(\pi') \leq \zeta + \lambda(\pi')\Delta^M(\pi') \leq \zeta + \Delta^M(\lambda) \leq \zeta + (1+\varepsilon)\mathsf{g}^M/\mathsf{n}.$$

Furthermore,

$$\begin{aligned}
I^M(\lambda') &= \inf_{M' \in \mathcal{M}^{\text{alt}}(M)} \sum_{\pi \neq \pi'} \frac{\zeta}{1-\lambda(\pi')} \lambda(\pi) D_{\mathsf{KL}}(M(\pi) \,\|\, M'(\pi)) + (1-\zeta) D_{\mathsf{KL}}(M(\pi') \,\|\, M'(\pi')) \\
&\geq \inf_{M' \in \mathcal{M}^{\text{alt}}(M)} \sum_{\pi \neq \pi'} (1-\zeta)\lambda(\pi) D_{\mathsf{KL}}(M(\pi) \,\|\, M'(\pi)) + (1-\zeta)\lambda(\pi') D_{\mathsf{KL}}(M(\pi') \,\|\, M'(\pi')) \\
&= (1-\zeta) I^M(\lambda) \\
&\geq (1-\zeta)(1-\varepsilon)/\mathsf{n}.
\end{aligned}$$

Since $\pi' \neq \pi_M$, we can lower bound $\Delta^M(\lambda) \geq (1-\zeta)\Delta^M_{\min}$, so

$$(1-\zeta)\Delta^M_{\min} \leq \Delta^M(\lambda) \leq (1+\varepsilon)\mathsf{g}^M/\mathsf{n} \implies \mathsf{n} \leq \frac{(1+\varepsilon)\mathsf{g}^M}{(1-\zeta)\Delta^M_{\min}} \leq \frac{4\mathsf{g}^M}{\Delta^M_{\min}}$$

We can therefore bound $\zeta \leq \varepsilon \mathsf{g}^M/\mathsf{n}$ as long as

$$\zeta \leq \frac{\Delta^M_{\min} \cdot \varepsilon}{4}.$$

Consider ζ that satisfies this inequality. Then $\Delta^M(\lambda') \leq (1+2\varepsilon)\mathsf{g}^M/\mathsf{n}$. We can also lower bound $(1-\zeta)(1-\varepsilon) \geq (1-2\varepsilon)$ as long as $\zeta \leq \frac{\varepsilon}{1-\varepsilon} \leq \varepsilon$. Thus, as long as

$$\zeta \leq \min\{1, \frac{\Delta^M_{\min}}{4}\} \cdot \varepsilon \quad \text{and} \quad \bar{\mathsf{n}} \geq \frac{4\mathsf{g}^M}{\Delta^M_{\min}},$$

we have that $M \in \mathcal{M}^{\text{gl}}_{2\varepsilon}(\lambda'; \bar{\mathsf{n}})$.

Case 2b: $\lambda \neq \mathbb{I}_\pi$ for any $\pi \in \Pi$. Finally, it remains to handle the case then there does not exist π' such that $\lambda(\pi') \geq 1-\zeta$. In this case, we can always lower bound

$$\zeta\Delta^M_{\min} \leq \Delta^M(\lambda) \leq (1+\varepsilon)\mathsf{g}^M/\mathsf{n} \implies \mathsf{n} \leq \frac{(1+\varepsilon)\mathsf{g}^M}{\zeta\Delta^M_{\min}},$$

so as long as $\bar{\mathsf{n}} \geq \frac{(1+\varepsilon)\mathsf{g}^M}{\zeta\Delta^M_{\min}}$, we have $M \in \mathcal{M}^{\text{gl}}_{2\varepsilon}(\lambda; \bar{\mathsf{n}})$.

Since we are in the regime where $\lambda \neq \mathbb{I}_\pi$, it must be the case that if $M \in \mathcal{M}_0 \cap \mathcal{M}^{\text{gl}}_\varepsilon(\lambda)$, then $\mathsf{g}^M > 0$. Thus, a sufficient choice of ζ is

$$\zeta = \min_{M \in \mathcal{M}_0 : \mathsf{g}^M > 0} \min\left\{ \frac{\mathsf{g}^M}{\mathsf{g}^M + 2\mathsf{n}^M_\varepsilon}, \frac{\Delta^M_{\min}\varepsilon}{4} \right\}.$$

Concluding the Proof. To show the result, we need that $\bar{\mathsf{n}}$ is large enough for each $M \in \mathcal{M}_0$. The argument above shows that it suffices to take

$$\bar{\mathsf{n}} \geq \max_{M \in \mathcal{M}_0} \max\left\{ \mathsf{n}_\varepsilon^M, \frac{4\mathsf{g}^M}{\Delta_{\min}^M}, \frac{2\mathsf{g}^M}{\zeta \Delta_{\min}^M} \right\}$$

for

$$\zeta := \min_{M \in \mathcal{M}_0 : \mathsf{g}^M > 0} \min\left\{ \frac{\mathsf{g}^M}{\mathsf{g}^M + 2\mathsf{n}_\varepsilon^M}, \frac{\Delta_{\min}^M \varepsilon}{4} \right\}.$$

This proves the result. □

Lemma H.2.5. *For every $M \in \mathcal{M}$ with $\Delta_{\min}^M > 0$, there exists some $\lambda \in \Upsilon(M; \varepsilon)$ with normalization factor n satisfying*

$$\mathsf{n} \leq \mathsf{g}^M / \Delta_{\min}^M + \mathsf{n}_\varepsilon^M,$$

i.e., we have $M \in \mathcal{M}_\varepsilon^{\mathrm{gl}}(\lambda; \mathsf{n})$.

Proof of Lemma H.2.5. Consider some allocation $\eta \in \mathbb{R}_+^\Pi$ such that

$$\Delta^M(\eta) \leq \mathsf{g}^M \quad \text{and} \quad I^M(\eta) \geq 1. \tag{H.2.2}$$

Let η' denote the allocation satisfying $\eta'(\pi) = \eta(\pi)$ for $\pi \neq \pi_M$, and $\eta'(\pi_M) = 0$. Note that $\Delta^M(\eta) \geq \Delta_{\min}^M \|\eta'\|_1$, which implies that

$$\|\eta'\|_1 \leq \mathsf{g}^M / \Delta_{\min}^M.$$

Let η'' denote the allocation satisfying $\eta''(\pi) = \eta(\pi)$ for $\pi \neq \pi_M$ and $\eta''(\pi_M) = \mathsf{n}_\varepsilon^M$. Then by definition of n_ε^M, and since η satisfies (H.2.2), we have $I^M(\eta'') \geq 1 - \varepsilon$. Furthermore, it is straightforward to see that $\Delta^M(\eta'') \leq \mathsf{g}^M$. This implies that $\eta''/\|\eta''\|_1 \in \Upsilon(M; \varepsilon)$ with normalization factor $\|\eta''\|_1$. However, we can bound

$$\|\eta''\|_1 = \|\eta'\|_1 + \mathsf{n}_\varepsilon^M \leq \mathsf{g}^M / \Delta_{\min}^M + \mathsf{n}_\varepsilon^M.$$

This proves the result. □

H.2.2.3 Bounding Allocation-Estimation Coefficient via Uniform Exploration Coefficient

In this section we prove a generalized version of Proposition 10.2. In particular, rather than specializing to the KL divergence, we consider a general divergence D. We define the uniform exploration coefficient with respect to D as follows.

Definition H.2.1 (Uniform Exploration Coefficient, General Divergences)**.** For a randomized estimator $\xi \in \triangle_{\mathcal{M}}$ and divergence $D(\cdot \,\|\, \cdot)$, we define the uniform exploration coefficient

with respect to ξ at scale $\varepsilon > 0$ as the value of the following program:

$$C_{\mathrm{exp}}^{\mathsf{D},\xi}(\varepsilon) := \min_{C \in \mathbb{R}_+, p \in \triangle_\Pi} \left\{ C \;\middle|\; \forall M, M' \in \mathcal{M} \;:\; \begin{array}{r} \max_{M'' \in \{M, M'\}} \mathbb{E}_{\overline{M} \sim \xi}[\mathbb{E}_{\pi \sim p}[D\big(\overline{M}(\pi) \,\|\, M''(\pi)\big)]] \le 1/C \\ \implies \max_{p' \in \triangle_\Pi} \mathbb{E}_{\pi \sim p'}[D\big(M(\pi) \,\|\, M'(\pi)\big)] \le \varepsilon \end{array} \right\}.$$

We define $p_{\mathrm{exp}}^{\mathsf{D},\xi}(\varepsilon)$ as the minimizing distribution for this program, and let

$$C_{\mathrm{exp}}^{\mathsf{D}}(\mathcal{M}, \varepsilon) := \sup_{\xi \in \triangle_{\mathcal{M}}} C_{\mathrm{exp}}^{\mathsf{D},\xi}(\varepsilon)$$

denote the uniform exploration constant for class $\mathcal{M}$.

Lemma H.2.6 (Formal version of Proposition 10.2)**.** *Let $\varepsilon \in [0, 1/2)$ and $\mathcal{M}_0 \subseteq \mathcal{M}$ be given, and assume that $\inf_{M \in \mathcal{M}_0} \mathsf{g}^M > 0, \inf_{M \in \mathcal{M}_0} \Delta_{\min}^M > 0$, $\sup_{M \in \mathcal{M}_0} \mathsf{n}_{1/4}^M < \infty$, and Assumptions 10.5, H.1 and H.2 hold. Then for any $\xi \in \triangle_{\mathcal{M}_0}$, we can bound*

$$\min_{\lambda, \omega \in \triangle_\Pi} \sup_{M \in \mathcal{M}_0 \setminus \mathcal{M}_\varepsilon^{\mathrm{gl}}(\lambda)} \frac{1}{\mathbb{E}_{\overline{M} \sim \xi}[\mathbb{E}_\omega[D\big(\overline{M}(\pi) \,\|\, M(\pi)\big)]]} \le C_{\mathrm{exp}}^{\mathsf{D}}(\mathcal{M}_0, \delta)$$

for any $\delta > 0$ satisfying

$$\sqrt{\delta} \le \min_{M \in \mathcal{M}_0} \min \left\{ \min \left\{ \frac{1}{81 L_{\mathsf{KL}}}, \frac{\Delta_{\min}^M}{34 V_{\mathcal{M}}} \right\} \cdot \frac{\varepsilon}{2\mathsf{g}^M / \Delta_{\min}^M + \mathsf{n}_{\varepsilon/36}^M}, \frac{\Delta_{\min}^M}{3} \right\}.$$

Proof of Lemma H.2.6. Let $\delta > 0$ be some tolerance to be chosen. Let $\widetilde{M}$ denote some $M \in \mathcal{M}_0$ such that $\mathbb{E}_{\overline{M} \sim \xi}[\mathbb{E}_{p_{\mathrm{exp}}}[D\big(\overline{M}(\pi) \,\|\, M(\pi)\big)]] \le 1/C_{\mathrm{exp}}$, where we abbreviate $C_{\mathrm{exp}} := C_{\mathrm{exp}}^{\mathsf{D}}(\mathcal{M}_0, \delta)$ and $p_{\mathrm{exp}} := p_{\mathrm{exp}}^{\mathsf{D},\xi}(\delta)$ is a distribution that achieves the value of $C_{\mathrm{exp}}^{\mathsf{D}}(\mathcal{M}_0, \delta)$ for ξ; if such an $\widetilde{M}$ does not exist, we let $\widetilde{M} = \arg\min_{M \in \mathcal{M}_0} \mathbb{E}_{\overline{M} \sim \xi}[\mathbb{E}_{p_{\mathrm{exp}}}[D\big(\overline{M}(\pi) \,\|\, M(\pi)\big)]]$. Let $\varepsilon' > 0$ be some value to be chosen, and let $\widetilde{\lambda} \in \Upsilon(\widetilde{M}; \varepsilon')$ denote the allocation in $\Upsilon(\widetilde{M}; \varepsilon')$ with smallest normalizing factor n. Let $\widetilde{\mathsf{n}}$ denote the value of this normalizing factor, then:

$$\Delta^{\widetilde{M}}(\widetilde{\lambda}) \le (1 + \varepsilon')\mathsf{g}^{\widetilde{M}}/\widetilde{\mathsf{n}} \quad \text{and} \quad I^{\widetilde{M}}(\widetilde{\lambda}) \ge (1 - \varepsilon')/\widetilde{\mathsf{n}}.$$

We can bound:

$$\begin{aligned} \min_{\lambda \in \triangle_\Pi} \min_{\omega \in \triangle_\Pi} \sup_{M \in \mathcal{M}_\varepsilon^{\mathrm{gl}}(\lambda)^c \cap \mathcal{M}_0} \frac{1}{\mathbb{E}_{\overline{M} \sim \xi}[\mathbb{E}_\omega[D\big(\overline{M}(\pi) \,\|\, M(\pi)\big)]]} &\le \sup_{M \in \mathcal{M}_\varepsilon^{\mathrm{gl}}(\widetilde{\lambda})^c \cap \mathcal{M}_0} \frac{1}{\mathbb{E}_{\overline{M} \sim \xi}[\mathbb{E}_{p_{\mathrm{exp}}}[D\big(\overline{M}(\pi) \,\|\, M(\pi)\big)]]} \\ &\le \sup_{M \in \mathcal{M}_\varepsilon^{\mathrm{gl}}(\widetilde{\lambda})^c \cap \mathcal{M}_0} \frac{\mathbb{I}\{\mathbb{E}_{\overline{M} \sim \xi}[\mathbb{E}_{p_{\mathrm{exp}}}[D\big(\overline{M}(\pi) \,\|\, M(\pi)\big)]] \le 1/C_{\mathrm{exp}}\}}{\mathbb{E}_{\overline{M} \sim \xi}[\mathbb{E}_{p_{\mathrm{exp}}}[D\big(\overline{M}(\pi) \,\|\, M(\pi)\big)]]} + C_{\mathrm{exp}}, \end{aligned} \tag{H.2.3}$$

where here we take $\frac{0}{0} = 0$. If $\mathbb{E}_{\overline{M} \sim \xi}[\mathbb{E}_{p_{\mathrm{exp}}}[D\big(\overline{M}(\pi) \,\|\, \widetilde{M}(\pi)\big)]] > 1/C_{\mathrm{exp}}$, then by definition of

$\widetilde{M}$, for all $M \in \mathcal{M}_0$ we have

$$\mathbb{E}_{\overline{M}\sim\xi}[\mathbb{E}_{p_{\text{exp}}}[D\big(\overline{M}(\pi) \,\|\, M(\pi)\big)]] > 1/C_{\text{exp}},$$

so we can simply bound (H.2.3) $\leq C_{\text{exp}}$. Going forward, we assume this is not the case, so that $\mathbb{E}_{\overline{M}\sim\xi}[\mathbb{E}_{p_{\text{exp}}}[D\big(\overline{M}(\pi) \,\|\, \widetilde{M}(\pi)\big)]] \leq 1/C_{\text{exp}}$. Our goal is to show that, for small enough δ, $\widetilde{\lambda} \in \Upsilon(M;\varepsilon)$ for every other $M \in \mathcal{M}_0$ with $\mathbb{E}_{\overline{M}\sim\xi}[\mathbb{E}_{p_{\text{exp}}}[D\big(\overline{M}(\pi) \,\|\, \widetilde{M}(\pi)\big)]] \leq 1/C_{\text{exp}}$, so that $M \in \mathcal{M}_\varepsilon^{\text{gl}}(\widetilde{\lambda})$. This will imply that there does not exist $M \in \mathcal{M}_\varepsilon^{\text{gl}}(\widetilde{\lambda})^c \cap \mathcal{M}_0$ with $\mathbb{E}_{\overline{M}\sim\xi}[\mathbb{E}_{p_{\text{exp}}}[D\big(\overline{M}(\pi) \,\|\, \widetilde{M}(\pi)\big)]] \leq 1/C_{\text{exp}}$, which further implies that (H.2.3) $\leq C_{\text{exp}}$.

Fix any $M \in \mathcal{M}_0$. We note that by the definition of p_{exp}, if $\mathbb{E}_{\overline{M}\sim\xi}[\mathbb{E}_{p_{\text{exp}}}[D\big(\overline{M}(\pi) \,\|\, M(\pi)\big)]] \leq 1/C_{\text{exp}}$, then

$$\sup_{p\in\triangle_\Pi} \mathbb{E}_p[D\big(\widetilde{M}(\pi) \,\|\, M(\pi)\big)] \leq \delta,$$

This implies in particular that, for each π, $D\big(\widetilde{M}(\pi) \,\|\, M(\pi)\big) \leq \delta$.

Step 1: $\mathbb{E}_{\overline{M}\sim\xi}[\mathbb{E}_{p_{\text{exp}}}[D\big(\overline{M}(\pi) \,\|\, M(\pi)\big)]] \leq 1/C_{\text{exp}}$ implies $\pi_M = \pi_{\widetilde{M}}$. As noted, if

$$\mathbb{E}_{\overline{M}\sim\xi}[\mathbb{E}_{p_{\text{exp}}}[D\big(\overline{M}(\pi) \,\|\, M(\pi)\big)]] \leq 1/C_{\text{exp}},$$

we have $D\big(\widetilde{M}(\pi) \,\|\, M(\pi)\big) \leq \delta$ for all π. Assume that $\pi_M \neq \pi_{\widetilde{M}}$ (note that since $\inf_{M\in\mathcal{M}_0} \Delta_{\min}^M > 0$ by assumption, all $M \in \mathcal{M}_0$ have unique optimal). By definition we have $f^{\widetilde{M}}(\pi_{\widetilde{M}}) \geq f^{\widetilde{M}}(\pi_M) + \Delta_{\min}^{\widetilde{M}}$ and $f^M(\pi_M) \geq f^M(\pi_{\widetilde{M}})$. Thus,

$$\begin{aligned}
\Delta_{\min}^{\widetilde{M}} &\leq f^{\widetilde{M}}(\pi_{\widetilde{M}}) - f^{\widetilde{M}}(\pi_M) + f^M(\pi_M) - f^M(\pi_{\widetilde{M}}) \\
&\leq |f^{\widetilde{M}}(\pi_{\widetilde{M}}) - f^M(\pi_{\widetilde{M}})| + |f^M(\pi_M) - f^{\widetilde{M}}(\pi_M)| \\
&\leq \sqrt{D\big(\widetilde{M}(\pi_{\widetilde{M}}) \,\|\, M(\pi_{\widetilde{M}})\big)} + \sqrt{D\big(\widetilde{M}(\pi_M) \,\|\, M(\pi_M)\big)}.
\end{aligned}$$

This implies that there exists some π such that $D\big(\widetilde{M}(\pi) \,\|\, M(\pi)\big) \geq (\Delta_{\min}^{\widetilde{M}}/2)^2$. Assuming

$$\delta \leq \min_{M\in\mathcal{M}_0} (\Delta_{\min}^{\widetilde{M}}/3)^2, \tag{H.2.4}$$

this is a contradiction. Thus, it must be the case that $\pi_M = \pi_{\widetilde{M}}$, as long as (H.2.4) is satisfied.

Step 2: $\mathbb{E}_{\overline{M}\sim\xi}[\mathbb{E}_{p_{\text{exp}}}[D\big(\overline{M}(\pi) \,\|\, M(\pi)\big)]] \leq 1/C_{\text{exp}}$ implies $\widetilde{\lambda} \in \Upsilon(M;\epsilon)$. Under Assumption H.2, we can bound, $\forall \pi \in \Pi$,

$$|f^{\widetilde{M}}(\pi) - f^M(\pi)| \leq \sqrt{D\big(\widetilde{M}(\pi) \,\|\, M(\pi)\big)} \leq \sqrt{\delta}.$$

This implies that, for any $\lambda \in \triangle_\Pi$,

$$|\Delta^{\widetilde{M}}(\lambda) - \Delta^M(\lambda)| \le |f^{\widetilde{M}}(\pi_M) - f^M(\pi_M)| + \sum_\pi \lambda_\pi |f^{\widetilde{M}}(\pi) - f^M(\pi)| \le 4\sqrt{\delta}.$$

In addition, under Assumption H.1, we have

$$\begin{aligned} D_{\mathsf{KL}}(M(\pi) \,\|\, M'(\pi)) &\ge D_{\mathsf{KL}}\big(\widetilde{M}(\pi) \,\|\, M'(\pi)\big) - L_{\mathsf{KL}}\sqrt{D\big(\widetilde{M}(\pi) \,\|\, M(\pi)\big)} \\ &\ge D_{\mathsf{KL}}\big(\widetilde{M}(\pi) \,\|\, M'(\pi)\big) - 2L_{\mathsf{KL}}\sqrt{\delta}. \end{aligned}$$

This implies, for any $\lambda \in \triangle_\Pi$,

$$\begin{aligned} I^M(\lambda) &= \inf_{M' \in \mathcal{M}^{\mathrm{alt}}(M)} \sum_\pi \lambda(\pi) D_{\mathsf{KL}}(M(\pi) \,\|\, M'(\pi)) \\ &\ge \inf_{M' \in \mathcal{M}^{\mathrm{alt}}(M)} \sum_\pi \lambda(\pi) D_{\mathsf{KL}}\big(\widetilde{M}(\pi) \,\|\, M'(\pi)\big) - 2L_{\mathsf{KL}}\sqrt{\delta} \\ &= I^{\widetilde{M}}(\lambda) - 2L_{\mathsf{KL}}\sqrt{\delta} \end{aligned}$$

where the final equality uses that, given what we have already shown, $\pi_M = \pi_{\widetilde{M}}$, so that $\mathcal{M}^{\mathrm{alt}}(M) = \mathcal{M}^{\mathrm{alt}}(\widetilde{M})$. Repeating the calculation in the other direction, we get that $|I^M(\lambda) - I^{\widetilde{M}}(\lambda)| \le 2L_{\mathsf{KL}}\sqrt{\delta}$.

We next relate g^M to $\mathsf{g}^{\widetilde{M}}$. By definition we have

$$(1+\varepsilon')\mathsf{g}^M/\widetilde{\mathsf{n}} \ge \inf_{\lambda \in \triangle_\Pi} \Delta^M(\lambda) \quad \text{s.t.} \quad I^M(\lambda) \ge (1-\varepsilon')/\widetilde{\mathsf{n}}.$$

Applying our perturbation bounds we can lower bound this as

$$\begin{aligned} &\ge \inf_{\lambda \in \triangle_\Pi} \Delta^{\widetilde{M}}(\lambda) - 4\sqrt{\delta} \quad \text{s.t.} \quad I^{\widetilde{M}}(\lambda) \ge (1-\varepsilon')/\widetilde{\mathsf{n}} - 2L_{\mathsf{KL}}\sqrt{\delta} \\ &\ge \frac{\mathsf{g}^{\widetilde{M}}}{((1-\varepsilon')/\widetilde{\mathsf{n}} - 2L_{\mathsf{KL}}\sqrt{\delta})^{-1}} - 4\sqrt{\delta} \end{aligned}$$

where the last inequality follows from Lemma H.2.1. This implies that

$$\mathsf{g}^{\widetilde{M}} \le ((1-\varepsilon')/\widetilde{\mathsf{n}} - 2L_{\mathsf{KL}}\sqrt{\delta})^{-1}(1+\varepsilon') \cdot \frac{\mathsf{g}^M}{\widetilde{\mathsf{n}}} + 4((1-\varepsilon')/\widetilde{\mathsf{n}} - 2L_{\mathsf{KL}}\sqrt{\delta})^{-1}\sqrt{\delta}. \tag{H.2.5}$$

Assuming that

$$\sqrt{\delta} \le \frac{\varepsilon' - 2(\varepsilon')^2}{2(1+2\varepsilon')L_{\mathsf{KL}}\widetilde{\mathsf{n}}},$$

some algebra shows that

$$(\text{H.2.5}) \le (1+2\varepsilon')^2 \mathsf{g}^M + 4(1+2\varepsilon')\widetilde{\mathsf{n}}\sqrt{\delta}.$$

Now we can bound

$$\begin{aligned}\Delta^M(\widetilde{\lambda}) &\le \Delta^{\widetilde{M}}(\widetilde{\lambda}) + 4\sqrt{\delta} \\ &\le (1+\varepsilon')\mathsf{g}^{\widetilde{M}}/\widetilde{\mathsf{n}} + 4\sqrt{\delta} \\ &\le (1+2\varepsilon')^3 \mathsf{g}^M/\widetilde{\mathsf{n}} + 4(1+2\varepsilon')^2\sqrt{\delta} + 4\sqrt{\delta}\end{aligned}$$

and

$$I^M(\widetilde{\lambda}) \ge I^{\widetilde{M}}(\widetilde{\lambda}) - 2L_{\mathsf{KL}}\sqrt{\delta} \ge (1-\varepsilon')/\widetilde{\mathsf{n}} - 2L_{\mathsf{KL}}\sqrt{\delta}.$$

If ε' and δ are small enough so that

$$(1+2\varepsilon')^3 \le 1+\varepsilon/2 \quad \text{and} \quad 4(1+2\varepsilon')^2\sqrt{\delta} + 4\sqrt{\delta} \le \varepsilon \mathsf{g}^M/2\widetilde{\mathsf{n}}$$

and

$$1-\varepsilon' \ge 1-\varepsilon/2 \quad \text{and} \quad 2L_{\mathsf{KL}}\sqrt{\delta} \le \varepsilon/2\widetilde{\mathsf{n}},$$

then $\Delta^M(\widetilde{\lambda}) \le (1+\varepsilon)\mathsf{g}^M/\widetilde{\mathsf{n}}$ and $I^M(\widetilde{\lambda}) \ge (1-\varepsilon)/\widetilde{\mathsf{n}}$, which implies that $\widetilde{\lambda} \in \Upsilon(M;\varepsilon)$ with scaling factor $\widetilde{\mathsf{n}}$.

Step 3: Condition on δ. Altogether, we have assumed that δ satisfies (H.2.4) and, that for some $M \in \mathcal{M}_0$ with $\mathbb{E}_{\overline{M}\sim\xi}[\mathbb{E}_{p_{\text{exp}}}[D\big(\overline{M}(\pi) \,\|\, M(\pi)\big)]] \le 1/C_{\text{exp}}$, we have

$$\sqrt{\delta} \le \frac{\varepsilon' - 2(\varepsilon')^2}{2(1+2\varepsilon')L_{\mathsf{KL}}\widetilde{\mathsf{n}}}, \quad 4(1+2\varepsilon')^2\sqrt{\delta} + 4\sqrt{\delta} \le \varepsilon\mathsf{g}^M/2\widetilde{\mathsf{n}}, \quad 2L_{\mathsf{KL}}\sqrt{\delta} \le \varepsilon/2\widetilde{\mathsf{n}} \tag{H.2.6}$$

and

$$(1+2\varepsilon')^3 \le 1+\varepsilon/2, \quad 1-\varepsilon' \ge 1-\varepsilon/2.$$

Some algebra shows that, as long as $\varepsilon \le 1$, it suffices to take $\varepsilon' = \varepsilon/36$ to satisfy the latter two conditions. Furthermore, some calculation shows that a sufficient condition for (H.2.6) to be met is that

$$\sqrt{\delta} \le \min\left\{\frac{1}{81L_{\mathsf{KL}}}, \frac{\mathsf{g}^M}{17}\right\} \cdot \frac{\varepsilon}{\widetilde{\mathsf{n}}}.$$

By Lemma H.2.5 and our choice of $\widetilde{\mathsf{n}}$, we can bound

$$\widetilde{\mathsf{n}} \le 2\mathsf{g}^{\widetilde{M}}/\Delta_{\min}^{\widetilde{M}} + \mathsf{n}_{\varepsilon/36}^{\widetilde{M}}$$

so it suffices that we take

$$\sqrt{\delta} \le \min\left\{\frac{1}{81L_{\mathsf{KL}}}, \frac{\mathsf{g}^{M}}{17}\right\} \cdot \varepsilon\left(\frac{2\mathsf{g}^{\widetilde{M}}}{\Delta_{\min}^{\widetilde{M}}} + \mathsf{n}_{\varepsilon}^{\widetilde{M}}\right)^{-1}.$$

As $\widetilde{M}$ was chosen to an arbitrary model in $\mathcal{M}_0$ with $\mathbb{E}_{\overline{M}\sim\xi}[\mathbb{E}_{p_{\text{exp}}}[D\big(\overline{M}(\pi) \,\|\, \widetilde{M}(\pi)\big)]] \le 1/C_{\text{exp}}$, we take it to minimize $\mathsf{g}^{\widetilde{M}}$ over this constraint. It suffices then that

$$\sqrt{\delta} \le \min\left\{\frac{1}{81L_{\mathsf{KL}}}, \frac{\mathsf{g}^{\widetilde{M}}}{17}\right\} \cdot \varepsilon\left(\frac{2\mathsf{g}^{\widetilde{M}}}{\Delta_{\min}^{\widetilde{M}}} + \mathsf{n}_{\varepsilon/36}^{\widetilde{M}}\right)^{-1}.$$

Finally, by Lemma H.3.13 (under Assumption 10.5) and Lemma H.2.3, we can lower bound $\mathsf{g}^{\widetilde{M}} \ge \Delta_{\min}^{\widetilde{M}}/2V_{\mathcal{M}}$. Combining this condition with (H.2.4) gives the result.

□

H.3 Deferred Proofs from Section 10.2

Organization of Appendix H.3. In this section we prove the main results from Section 10.2. We consider a slightly generalized version of the setting in Section 10.2, where we allow for divergences other than just the KL divergence, as described below. This section is organized as follows.

- First, in Appendix H.3.1, we give the proof of our main result, Theorem 10.1. We break this proof into two principle components: bounding the regret of AE^2 in the exploit phase (Section H.3.1.1), and explore phase (Section H.3.1.2). The key results in this section are Lemma H.3.3, which formalizes the key algorithm intuition given in Section 10.2, showing that exploring via the AEC yields low regret, and Lemma H.3.4, which shows that, to enter the explore phase, the total "information gain" must be bounded as $O(\log T)$, which ultimately yields the optimal leading-order scaling. We combine these results with our estimation guarantees in Appendix H.3.1.3, where we give the proof of Theorem 10.1.

- In Appendix H.3.2, we extend AE^2 and Theorem 10.1 to the case where we assume no lower bound on the minimum gap over the model class, first presenting our main algorithm in this setting, Algorithm H.3, and then giving a proof of Theorem 10.2. The structure of this section is similar to Appendix H.3.1—the primary difference being a slightly different argument to handle the need to adapt to the minimum gap of the ground truth instance.

- Finally, in Appendix H.3.3 we present our estimation routine with covering, and prove that it achieves low estimation error, and in Appendix H.3.4 we provide the proofs of miscellaneous results used throughout Appendix H.3.

H.3.1 Regret Bound for Uniformly Regular Classes (Theorem 10.1)

In this section we prove Theorem H.1, which generalizes Theorem 10.1. To do so, we analyze Algorithm H.2, wich generalizes Algorithm 10.1 to allow for general divergences.

Throughout, we define

$$\underline{\mathsf{g}}^{\mathcal{M}} := \min_{M \in \mathcal{M}: \mathsf{g}^M > 0} \mathsf{g}^M.$$

Algorithm H.2 Allocation Estimation via Adaptive Exploration (AE^2, general divergences)

1: **input:** Optimality tolerance δ, model class $\mathcal{M}$, estimation oracle $\mathbf{Alg}_D$.
2: Initialize $s \leftarrow 1$, $\varepsilon \leftarrow \frac{\delta}{4+2\delta}$, $\mathsf{n}_{\max} \leftarrow \mathsf{n}_{\max}(\mathcal{M}, \varepsilon)$, $q \leftarrow \frac{4\mathsf{n}_{\max}+\delta\underline{\mathsf{g}}^{\mathcal{M}}}{4\mathsf{n}_{\max}+2\delta\underline{\mathsf{g}}^{\mathcal{M}}}$.
3: Compute $\xi^1 \leftarrow \mathbf{Alg}_D(\{\varnothing\})$ and $\widehat{M}^1 \leftarrow \mathbb{E}_{M\sim\xi^1}[M]$.
4: **for** $t = 1, 2, 3, \ldots$ **do**
5: **if** $\exists \pi_{\widehat{M}^s} \in \boldsymbol{\pi}_{\widehat{M}^s}$ s.t. $\forall M \in \mathcal{M}^{\mathrm{alt}}(\pi_{\widehat{M}^s})$, $\sum_{i=1}^{s-1} \mathbb{E}_{\widehat{M}\sim\xi^i}\left[\log \frac{\mathbb{P}^{\widehat{M},\pi^i}(r^i,o^i)}{\mathbb{P}^{M,\pi^i}(r^i,o^i)}\right] \geq \log(t\log t)$ **then** // Exploit
6: Play $\pi_{\widehat{M}^s}$.
7: **else** // Explore
8: Set $p^s \leftarrow q\lambda^s + (1-q)\omega^s$ for

$$\lambda^s, \omega^s \leftarrow \arg\min_{\lambda,\omega\in\triangle_\Pi} \sup_{M\in\mathcal{M}\setminus\mathcal{M}^{\mathrm{gl}}_\varepsilon(\lambda;\mathsf{n}_{\max})} \frac{1}{\mathbb{E}_{\widehat{M}\sim\xi^s}\left[\mathbb{E}_{\pi\sim\omega}\left[D\left(\widehat{M}(\pi)\,\|\,M(\pi)\right)\right]\right]}. \tag{H.3.1}$$

9: Draw $\pi^s \sim p^s$, observe r^s, o^s.
10: Compute estimate $\xi^{s+1} \leftarrow \mathbf{Alg}_D(\{(\pi^i, r^i, o^i)\}_{i=1}^s)$ and let $\widehat{M}^s = \mathbb{E}_{M\sim\xi^s}[M]$.
11: $s \leftarrow s+1$.

H.3.1.1 Bounding Regret of Exploit Phase

We refer to the *exploit phase* as the subset of rounds t in which Line 6 is reached, and refer to the *explore phase* as the subset of rounds in which Line 8 is reached.

Lemma H.3.1. *The total expected regret incurred by the exploit phase of Algorithm H.2 is bounded by* $2\log\log T + 3$.

Proof of Lemma H.3.1. Let $\mathcal{E}_t$ denote the event that we exploit at round t, and that $\pi_{\widehat{M}^s} \neq \pi_\star$. Then, since we can incur suboptimality of at most 1 at each round, the total expected regret incurred by the exploit phase is bounded by

$$\sum_{t=1}^{T} \mathbb{E}^{M^\star}[\mathbb{I}\{\mathcal{E}_t\}].$$

Let $\widetilde{\mathcal{E}}_t$ denote the event

$$\widetilde{\mathcal{E}}_t := \left\{ \forall s \geq 1 \ : \ \sum_{i=1}^{s} \mathbb{E}_{\widehat{M} \sim \xi^i} \left[\log \frac{\mathbb{P}^{\widehat{M}, \pi^i}(r^i, o^i)}{\mathbb{P}^{M^\star, \pi^i}(r^i, o^i)} \right] < \log(t \log t) \right\}.$$

By Lemma H.3.2, we have $\mathbb{P}^{M^\star}[\mathbb{I}\{\widetilde{\mathcal{E}}_t^c\}] \leq \frac{1}{t \log t}$, and we can bound

$$\mathbb{E}^{M^\star}[\mathbb{I}\{\mathcal{E}_t\}] \leq \mathbb{E}^{M^\star}[\mathbb{I}\{\mathcal{E}_t \cap \widetilde{\mathcal{E}}_t\}] + \mathbb{E}^{M^\star}[\mathbb{I}\{\widetilde{\mathcal{E}}_t^c\}] \leq \mathbb{E}^{M^\star}[\mathbb{I}\{\mathcal{E}_t \cap \widetilde{\mathcal{E}}_t\}] + \frac{1}{t \log t}.$$

Let s_t denote the exploration round at round t. If we exploit at round t, this implies that for all $M \in \mathcal{M}^{\mathrm{alt}}(\pi_{\widehat{M}^s})$, we have

$$\sum_{i=1}^{s_t - 1} \mathbb{E}_{\widehat{M} \sim \xi^i} \left[\log \frac{\mathbb{P}^{\widehat{M}, \pi^i}(r^i, o^i)}{\mathbb{P}^{M, \pi^i}(r^i, o^i)} \right] \geq \log(t \log t). \tag{H.3.2}$$

If $\pi_{\widehat{M}^s} \neq \pi_\star$, then $M^\star \in \mathcal{M}^{\mathrm{alt}}(\pi_{\widehat{M}^s})$, so (H.3.2) must hold for $M \leftarrow M^\star$. This contradicts $\widetilde{\mathcal{E}}_t$, however, so $\mathbb{E}^{M^\star}[\mathbb{I}\{\mathcal{E}_t \cap \widetilde{\mathcal{E}}_t\}] = 0$. Thus, $\mathbb{E}^{M^\star}[\mathbb{I}\{\mathcal{E}_t\}] \leq \frac{1}{t \log t}$, so

$$\sum_{t=1}^{T} \mathbb{E}^{M^\star}[\mathbb{I}\{\mathcal{E}_t\}] \leq 3 + \sum_{t=3}^{T} \frac{1}{t \log t} \leq 3 + 2 \int_e^T \frac{1}{t \log t} \mathrm{d}t = 3 + 2 \log \log T.$$

□

Lemma H.3.2. *For $\{(r^i, o^i, \xi^i)\}_{i=1}^s$ generated as in Algorithm H.2, we have that*

$$\mathbb{P}^{M^\star} \left[\exists s \geq 1 \ : \ \sum_{i=1}^{s} \mathbb{E}_{\widehat{M} \sim \xi^i} \left[\log \frac{\mathbb{P}^{\widehat{M}, \pi^i}(r^i, o^i)}{\mathbb{P}^{M^\star, \pi^i}(r^i, o^i)} \right] \geq \varepsilon \right] \leq e^{-\varepsilon}.$$

Proof of Lemma H.3.2. Denote

$$X_s := \exp\left(\sum_{i=1}^{s} \mathbb{E}_{\widehat{M} \sim \xi^i} \left[\log \frac{\mathbb{P}^{\widehat{M}, \pi^i}(r^i, o^i)}{\mathbb{P}^{M^\star, \pi^i}(r^i, o^i)} \right] \right).$$

We first show that X_s is a supermartingale. Letting $\mathcal{F}_{s-1}$ denote the filtration up to $s-1$, we have

$$\begin{aligned}
\mathbb{E}^{M^\star}[X_s \mid \mathcal{F}_{s-1}] &= \exp\left(\sum_{i=1}^{s-1} \mathbb{E}_{\widehat{M}\sim\xi^i}\left[\log \frac{\mathbb{P}^{\widehat{M},\pi^i}(r^i,o^i)}{\mathbb{P}^{M^\star,\pi^i}(r^i,o^i)}\right]\right) \cdot \mathbb{E}^{M^\star}\left[\exp\left(\mathbb{E}_{\widehat{M}\sim\xi^s}\left[\log \frac{\mathbb{P}^{\widehat{M},\pi^s}(r^s,o^s)}{\mathbb{P}^{M^\star,\pi^s}(r^s,o^s)}\right]\right) \mid \mathcal{F}_{s-1}\right] \\
&= X_{s-1} \cdot \mathbb{E}^{M^\star}\left[\exp\left(\mathbb{E}_{\widehat{M}\sim\xi^s}\left[\log \frac{\mathbb{P}^{\widehat{M},\pi^s}(r^s,o^s)}{\mathbb{P}^{M^\star,\pi^s}(r^s,o^s)}\right]\right) \mid \mathcal{F}_{s-1}\right] \\
&\overset{(a)}{\leq} X_{s-1} \cdot \mathbb{E}^{M^\star}\left[\mathbb{E}_{\widehat{M}\sim\xi^s}\left[\exp\left(\log \frac{\mathbb{P}^{\widehat{M},\pi^s}(r^s,o^s)}{\mathbb{P}^{M^\star,\pi^s}(r^s,o^s)}\right)\right] \mid \mathcal{F}_{s-1}\right] \\
&= X_{s-1} \cdot \mathbb{E}^{M^\star}\left[\frac{\mathbb{E}_{\widehat{M}\sim\xi^s}[\mathbb{P}^{\widehat{M},\pi^s}(r^s,o^s)]}{\mathbb{P}^{M^\star,\pi^s}(r^s,o^s)} \mid \mathcal{F}_{s-1}\right] \\
&= X_{s-1}
\end{aligned}$$

where (a) holds by Jensen's inequality, and the final equality holds since $\mathbb{E}_{\widehat{M}\sim\xi^s}[\mathbb{P}^{\widehat{M},\pi}(\cdot,\cdot)]$ is a valid distribution over $\mathcal{R}\times\mathcal{O}$. Thus, X_s is a supermartingale. Ville's Maximal Inequality then immediately gives that

$$\mathbb{P}^{M^\star}[\exists s \geq 1 \ : \ X_s \geq e^{\varepsilon}] \leq \frac{\mathbb{E}^{M^\star}[X_1]}{e^{\varepsilon}}.$$

To complete the proof, using the same calculation as above, we bound

$$\mathbb{E}^{M^\star}[X_1] = \mathbb{E}^{M^\star}\left[\exp\left(\mathbb{E}_{\widehat{M}\sim\xi^1}\left[\log \frac{\mathbb{P}^{\widehat{M},\pi^1}(r^1,o^1)}{\mathbb{P}^{M^\star,\pi^1}(r^1,o^1)}\right]\right)\right] \leq \mathbb{E}^{M^\star}\left[\mathbb{E}_{\widehat{M}\sim\xi^1}\left[\exp\left(\log \frac{\mathbb{P}^{\widehat{M},\pi^1}(r^1,o^1)}{\mathbb{P}^{M^\star,\pi^1}(r^1,o^1)}\right)\right]\right] \leq 1.$$

□

H.3.1.2 Bounding Regret of Explore Phase

Lemma H.3.3 (Main Explore Phase Regret Bound)**.** *Let s_T denote the total number of exploration rounds (which is a random variable), and assume that $\delta \in [0,1/2)$. Then running Algorithm H.2, if $\mathsf{g}^\star > 0$, we can bound*

$$\begin{aligned}
\mathbb{E}[s_T] \leq \frac{24\mathsf{n}_{\max}^2 + 8\mathsf{n}_{\max}\underline{\mathsf{g}}^{\mathcal{M}}}{(\delta\underline{\mathsf{g}}^{\mathcal{M}})^2} \cdot \overline{\mathsf{aec}}^{\mathsf{D}}_{\varepsilon/2}(\mathcal{M}) \cdot \mathbb{E}[\widehat{\mathbf{Est}}_{\mathsf{D}}(s_T)] + \frac{12\mathsf{n}_{\max}}{\delta\Delta_{\min}} \cdot \mathbb{E}[\mathbf{Est}_{\mathsf{KL}}(s_T)] \\
+ \frac{6\mathsf{n}_{\max}}{\delta} \cdot \mathbb{E}\left[\sum_{s=1}^{s_T} \inf_{M\in\mathcal{M}^{\mathrm{alt}}(M^\star)} \mathbb{E}_{p^s}[D_{\mathsf{KL}}(M^\star(\pi) \,\|\, M(\pi))] \cdot \mathbb{I}\{\pi_\star \in \boldsymbol{\pi}_{\widehat{M}^s}\}\right]
\end{aligned}$$

and the regret during exploration rounds is bounded as

$$\mathbb{E}\left[\sum_{s=1}^{s_T}\Delta^\star(\pi^s)\right] \le \frac{8\mathsf{n}_{\max}+2\underline{\mathsf{g}}^{\mathcal{M}}}{\delta\underline{\mathsf{g}}^{\mathcal{M}}}\cdot\overline{\mathsf{aec}}^{\mathsf{D}}_{\varepsilon/2}(\mathcal{M})\cdot\mathbb{E}[\widehat{\mathbf{Est}}_{\mathsf{D}}(s_T)] + \frac{2(1+\delta)\mathsf{g}^\star}{\Delta_{\min}}\cdot\mathbb{E}[\mathbf{Est}_{\mathsf{KL}}(s_T)]$$
$$+(1+\delta)\mathsf{g}^\star\cdot\mathbb{E}\left[\sum_{s=1}^{s_T}\inf_{M\in\mathcal{M}^{\mathrm{alt}}(M^\star)}\mathbb{E}_{p^s}[D_{\mathsf{KL}}(M^\star(\pi)\,\|\,M(\pi))]\cdot\mathbb{I}\{\pi_\star\in\boldsymbol{\pi}_{\widehat{M}^s}\}\right].$$

Proof of Lemma H.3.3. The expected regret during exploration can be written as $\mathbb{E}[\sum_{s=1}^{s_T}\Delta^\star(\pi^s)] = \mathbb{E}[\sum_{s=1}^{s_T}\mathbb{E}_{p^s}[\Delta^\star(\pi)]]$. By definition, for exploration rounds, we have $p^s \leftarrow q\lambda^s + (1-q)\omega^s$. For each $s \le s_T$, we consider three cases to bound the instantaneous expected regret, $\mathbb{E}_{p^s}[\Delta^\star(\pi)] = \Delta^\star(p^s)$.

Case 1: $M^\star \in \mathcal{M}\backslash\mathcal{M}^{\mathrm{gl}}_\varepsilon(\lambda^s;\mathsf{n}_{\max})$**.** Denote such rounds as $\mathcal{S}^1_{\mathrm{exp}}$. Write

$$\Delta^\star(p^s) = \left[\Delta^\star(p^s) - \gamma^s\mathbb{E}_{\widehat{M}\sim\xi^s}\left[\mathbb{E}_{p^s}[D\big(\widehat{M}(\pi)\,\|\,M^\star(\pi)\big)]\right]\right] + \gamma^s\mathbb{E}_{\widehat{M}\sim\xi^s}\left[\mathbb{E}_{p^s}[D\big(\widehat{M}(\pi)\,\|\,M^\star(\pi)\big)]\right]$$

for

$$\gamma^s := \frac{1+\delta}{1-q}\cdot\frac{1}{\mathbb{E}_{\widehat{M}\sim\xi^s}[\mathbb{E}_{\omega^s}[D\big(\widehat{M}(\pi)\,\|\,M^\star(\pi)\big)]]}. \tag{H.3.3}$$

In this case we have that

$$\begin{aligned}
&\gamma^s\mathbb{E}_{\widehat{M}\sim\xi^s}[\mathbb{E}_{p^s}[D\big(\widehat{M}(\pi)\,\|\,M^\star(\pi)\big)]]\\
&= \frac{1+\delta}{1-q}\cdot\frac{1}{\mathbb{E}_{\omega^s}[\mathbb{E}_{\widehat{M}\sim\xi^s}[D\big(\widehat{M}(\pi)\,\|\,M^\star(\pi)\big)]]}\mathbb{E}_{\widehat{M}\sim\xi^s}[\mathbb{E}_{p^s}[D\big(\widehat{M}(\pi)\,\|\,M^\star(\pi)\big)]]\\
&\ge \frac{1+\delta}{\mathbb{E}_{\widehat{M}\sim\xi^s}[\mathbb{E}_{p^s}[D\big(\widehat{M}(\pi)\,\|\,M^\star(\pi)\big)]]}\mathbb{E}_{\widehat{M}\sim\xi^s}[\mathbb{E}_{p^s}[D\big(\widehat{M}(\pi)\,\|\,M^\star(\pi)\big)]]\\
&= 1+\delta.
\end{aligned}$$

Thus, since the suboptimality gap is always bounded by 1, we can bound

$$\Delta^\star(p^s) - \gamma^s\mathbb{E}_{\widehat{M}\sim\xi^s}\left[\mathbb{E}_{p^s}[D\big(\widehat{M}(\pi)\,\|\,M^\star(\pi)\big)]\right] \le 1 - \gamma^s\mathbb{E}_{\widehat{M}\sim\xi^s}\left[\mathbb{E}_{p^s}[D\big(\widehat{M}(\pi)\,\|\,M^\star(\pi)\big)]\right] \le -\delta.$$

So,

$$\Delta^\star(p^s) \le -\delta + \gamma^s\mathbb{E}_{\widehat{M}\sim\xi^s}\left[\mathbb{E}_{p^s}[D\big(\widehat{M}(\pi)\,\|\,M^\star(\pi)\big)]\right].$$

Case 2: $M^\star \in \mathcal{M}_\varepsilon^{\mathrm{gl}}(\lambda^s; \mathsf{n}_{\max}), \pi_\star \in \boldsymbol{\pi}_{\widehat{M}^s}$**.** Denote such rounds as $\mathcal{S}^2_{\mathrm{exp}}$, and write

$$\Delta^\star(p^s) = [\Delta^\star(p^s) - (1+\delta)\mathsf{g}^\star \cdot \mathbb{E}_{p^s}[D_{\mathsf{KL}}(M^\star(\pi) \,\|\, M(\pi))]] + (1+\delta)\mathsf{g}^\star \cdot \mathbb{E}_{p^s}[D_{\mathsf{KL}}(M^\star(\pi) \,\|\, M(\pi))]$$

for any $M \in \mathcal{M}_{\mathrm{alt}}^\star$. In this case, since $M^\star \in \mathcal{M}_\varepsilon^{\mathrm{gl}}(\lambda^s; \mathsf{n}_{\max})$, we have that $\lambda^s \in \Upsilon(M^\star; \varepsilon)$. This then implies that

$$\Delta^\star(\lambda^s) \leq (1+\varepsilon)\mathsf{g}^\star/\mathsf{n}^\star \quad \text{and} \quad \inf_{M \in \mathcal{M}_{\mathrm{alt}}^\star} \mathbb{E}_{\lambda^s}[D_{\mathsf{KL}}(M^\star(\pi) \,\|\, M(\pi))] \geq (1-\varepsilon)/\mathsf{n}^\star$$

for some $\mathsf{n}^\star \leq \mathsf{n}_{\max}$. Since $M \in \mathcal{M}_{\mathrm{alt}}^\star$, it follows that $\mathbb{E}_{\lambda^s}[D_{\mathsf{KL}}(M^\star(\pi) \,\|\, M(\pi))] \geq (1-\varepsilon)/\mathsf{n}^\star$. Thus,

$$\begin{aligned}
\Delta^\star(p^s) - (1+\delta)\mathsf{g}^\star \mathbb{E}_{p^s}[D_{\mathsf{KL}}(M^\star(\pi) \,\|\, M(\pi))] &\leq q\left[\Delta^\star(\lambda^s) - (1+\delta)\mathsf{g}^\star \mathbb{E}_{\lambda^s}[D_{\mathsf{KL}}(M^\star(\pi) \,\|\, M(\pi))]\right] + 1 - q \\
&\leq q\left[(1+\varepsilon)\mathsf{g}^\star/\mathsf{n}^\star - (1+\delta)(1-\varepsilon)\mathsf{g}^\star/\mathsf{n}^\star\right] + 1 - q \\
&= q\left[2\varepsilon - \delta(1-\varepsilon)\right] \cdot \frac{\mathsf{g}^\star}{\mathsf{n}^\star} + 1 - q \\
&\overset{(a)}{=} -\frac{q\delta}{2}\frac{\mathsf{g}^\star}{\mathsf{n}^\star} + 1 - q \\
&\leq -\frac{q\delta}{2}\frac{\mathsf{g}^\star}{\mathsf{n}_{\max}} + 1 - q \\
&\overset{(b)}{\leq} -\frac{\delta}{4}\frac{\mathsf{g}^\star}{\mathsf{n}_{\max}},
\end{aligned}$$

where (a) follows from our choice of $\varepsilon = \frac{\delta/2}{2+\delta}$ and setting of $\mathsf{n}^\star$, and (b) follows from our setting of $q = \frac{4\mathsf{n}_{\max} + \delta \underline{\mathsf{g}}^{\mathcal{M}}}{4\mathsf{n}_{\max} + 2\delta \underline{\mathsf{g}}^{\mathcal{M}}}$, since $\mathsf{g}^\star > 0$, and some algebra. Thus,

$$\Delta^\star(p^s) \leq (1+\delta)\mathsf{g}^\star \mathbb{E}_{p^s}[D_{\mathsf{KL}}(M^\star(\pi) \,\|\, M(\pi))] - \frac{\delta}{4}\frac{\mathsf{g}^\star}{\mathsf{n}_{\max}}.$$

As this holds for every $M \in \mathcal{M}_{\mathrm{alt}}^\star$, we therefore have

$$\Delta^\star(p^s) \leq \inf_{M \in \mathcal{M}_{\mathrm{alt}}^\star} (1+\delta)\mathsf{g}^\star \mathbb{E}_{p^s}[D_{\mathsf{KL}}(M^\star(\pi) \,\|\, M(\pi))] - \frac{\delta}{4}\frac{\mathsf{g}^\star}{\mathsf{n}_{\max}}.$$

Case 3: $M^\star \in \mathcal{M}_\varepsilon^{\mathrm{gl}}(\lambda^s; \mathsf{n}_{\max}), \pi_\star \notin \boldsymbol{\pi}_{\widehat{M}^s}$**.** Denote such rounds as $\mathcal{S}^3_{\mathrm{exp}}$, and write

$$\begin{aligned}
\Delta^\star(p^s) = \left[\Delta^\star(p^s) - \frac{2(1+\delta)\mathsf{g}^\star}{\Delta_{\min}} \cdot \mathbb{E}_{\widehat{M} \sim \xi^s}[\mathbb{E}_{p^s}[D_{\mathsf{KL}}\big(M^\star(\pi) \,\|\, \widehat{M}(\pi)\big)]]\right] \\
+ \frac{2(1+\delta)\mathsf{g}^\star}{\Delta_{\min}} \cdot \mathbb{E}_{\widehat{M} \sim \xi^s}[\mathbb{E}_{p^s}[D_{\mathsf{KL}}\big(M^\star(\pi) \,\|\, \widehat{M}(\pi)\big)]].
\end{aligned}$$

Since $M^\star \in \mathcal{M}_\varepsilon^{\mathrm{gl}}(\lambda^s; \mathsf{n}_{\max})$, we have that $\lambda^s \in \Upsilon(M^\star; \varepsilon)$. This then implies that for any $M \in \mathcal{M}_{\mathrm{alt}}^\star$:

$$\Delta^\star(\lambda^s) \le (1+\varepsilon)\mathsf{g}^\star/\mathsf{n}^\star \quad \text{and} \quad \inf_{M \in \mathcal{M}_{\mathrm{alt}}^\star} \mathbb{E}_{\lambda^s}[D_{\mathsf{KL}}(M^\star(\pi) \,\|\, M(\pi))] \ge (1-\varepsilon)/\mathsf{n}^\star$$

for some $\mathsf{n}^\star \le \mathsf{n}_{\max}$. By Lemma H.3.9, since $\pi_\star \notin \boldsymbol{\pi}_{\widehat{M}^s}$, we can lower bound $\mathbb{E}_{\widehat{M}\sim\xi^s}[\mathbb{I}\{\widehat{M} \in \mathcal{M}_{\mathrm{alt}}^\star\}] \ge \frac{1}{2}\Delta_{\min}$. Thus, we have

$$\begin{aligned}
&\frac{2(1+\delta)\mathsf{g}^\star}{\Delta_{\min}} \mathbb{E}_{\widehat{M}\sim\xi^s}[\mathbb{E}_{\lambda^s}[D_{\mathsf{KL}}\big(M^\star(\pi) \,\|\, \widehat{M}(\pi)\big)]] \\
&\ge \frac{2(1+\delta)\mathsf{g}^\star}{\Delta_{\min}} \mathbb{E}_{\widehat{M}\sim\xi^s}[\mathbb{E}_{\lambda^s}[D_{\mathsf{KL}}\big(M^\star(\pi) \,\|\, \widehat{M}(\pi)\big) \cdot \mathbb{I}\{\widehat{M} \in \mathcal{M}_{\mathrm{alt}}^\star\}]] \\
&\ge \frac{(1+\delta)(1-\varepsilon)\mathsf{g}^\star}{\mathsf{n}^\star}.
\end{aligned}$$

This implies that

$$\begin{aligned}
&\Delta^\star(p^s) - \frac{2(1+\delta)\mathsf{g}^\star}{\Delta_{\min}} \cdot \mathbb{E}_{\widehat{M}\sim\xi^s}[\mathbb{E}_{p^s}[D_{\mathsf{KL}}\big(M^\star(\pi) \,\|\, \widehat{M}(\pi)\big)]] \\
&\le q[(1+\varepsilon)\mathsf{g}^\star/\mathsf{n}^\star - (1+\delta)(1-\varepsilon)\mathsf{g}^\star/\mathsf{n}^\star] + 1 - q \\
&\le -\frac{\delta}{4}\frac{\mathsf{g}^\star}{\mathsf{n}_{\max}},
\end{aligned}$$

where the final inequality follows by the same argument as in Case 2. Thus,

$$\Delta^\star(p^s) \le \frac{2(1+\delta)\mathsf{g}^\star}{\Delta_{\min}} \cdot \mathbb{E}_{\widehat{M}\sim\xi^s}[\mathbb{E}_{p^s}[D_{\mathsf{KL}}\big(M^\star(\pi) \,\|\, \widehat{M}(\pi)\big)]] - \frac{\delta}{4}\frac{\mathsf{g}^\star}{\mathsf{n}_{\max}}.$$

Completing the Proof. In total we have

$$\begin{aligned}
\mathbb{E}\left[\sum_{s=1}^{s_T} \Delta^\star(p^s)\right] \le \mathbb{E}\Bigg[&- \delta|\mathcal{S}_{\mathrm{exp}}^1| - \frac{\delta\mathsf{g}^\star}{4\mathsf{n}_{\max}}|\mathcal{S}_{\mathrm{exp}}^2 \cup \mathcal{S}_{\mathrm{exp}}^3| + \sum_{s\in\mathcal{S}_{\mathrm{exp}}^1} \gamma^s \mathbb{E}_{\widehat{M}\sim\xi^s}[\mathbb{E}_{p^s}[D\big(\widehat{M}(\pi) \,\|\, M^\star(\pi)\big)]] \\
&+ (1+\delta)\mathsf{g}^\star \sum_{s\in\mathcal{S}_{\mathrm{exp}}^2} \inf_{M\in\mathcal{M}_{\mathrm{alt}}^\star} \mathbb{E}_{p^s}[D_{\mathsf{KL}}(M^\star(\pi) \,\|\, M(\pi))] \\
&+ \frac{2(1+\delta)\mathsf{g}^\star}{\Delta_{\min}} \sum_{s\in\mathcal{S}_{\mathrm{exp}}^3} \mathbb{E}_{\widehat{M}\sim\xi^s}[\mathbb{E}_{p^s}[D_{\mathsf{KL}}\big(M^\star(\pi) \,\|\, \widehat{M}(\pi)\big)]]\Bigg] \\
\le \mathbb{E}\Bigg[&- \frac{\delta\mathsf{g}^\star}{4\mathsf{n}_{\max}} \cdot s_T + \frac{8\mathsf{n}_{\max} + 2\underline{\mathsf{g}}^{\mathcal{M}}}{\delta\underline{\mathsf{g}}^{\mathcal{M}}} \cdot \overline{\mathsf{aec}}_{\varepsilon/2}^{\mathsf{D}}(\mathcal{M}) \cdot \widehat{\mathbf{Est}}_{\mathsf{D}}(s_T) + \frac{2(1+\delta)\mathsf{g}^\star}{\Delta_{\min}} \cdot \mathbf{Est}_{\mathsf{KL}}(s_T)
\end{aligned}$$

$$+ (1+\delta)\mathsf{g}^\star \sum_{s=1}^{s_T} \inf_{M \in \mathcal{M}^\star_{\mathrm{alt}}} \mathbb{E}_{p^s}[D_{\mathsf{KL}}(M^\star(\pi) \,\|\, M(\pi))] \cdot \mathbb{I}\{\pi_\star \in \boldsymbol{\pi}_{\widehat{M}^s}\}\Bigg]$$

where the last inequality follows from Lemma H.3.10, which bounds

$$\gamma^s \le \frac{8\mathsf{n}_{\max} + 2\underline{\mathsf{g}}^{\mathcal{M}}}{\delta \underline{\mathsf{g}}^{\mathcal{M}}} \cdot \overline{\mathsf{aec}}^{\mathsf{D}}_{\varepsilon/2}(\mathcal{M}),$$

and also using that for any $s \in \mathcal{S}^2_{\mathrm{exp}}$ we have $\pi_\star \in \boldsymbol{\pi}_{\widehat{M}^s}$. Upper bounding $-\frac{\delta \mathsf{g}^\star}{4\mathsf{n}_{\max}} \cdot s_T \le 0$ proves the second claim in the lemma statement.

For the first claim, as regret is always nonnegative, it follows that

$$0 \le \mathbb{E}\Bigg[-\frac{\delta \mathsf{g}^\star}{4\mathsf{n}_{\max}} \cdot s_T + \frac{8\mathsf{n}_{\max} + 2\underline{\mathsf{g}}^{\mathcal{M}}}{\delta \underline{\mathsf{g}}^{\mathcal{M}}} \cdot \overline{\mathsf{aec}}^{\mathsf{D}}_{\varepsilon/2}(\mathcal{M}) \cdot \widehat{\mathbf{Est}}_{\mathsf{D}}(s_T) + \frac{2(1+\delta)\mathsf{g}^\star}{\Delta_{\min}} \cdot \mathbf{Est}_{\mathsf{KL}}(s_T)$$
$$+ (1+\delta)\mathsf{g}^\star \cdot \sum_{s=1}^{s_T} \inf_{M \in \mathcal{M}^\star_{\mathrm{alt}}} \mathbb{E}_{p^s}[D_{\mathsf{KL}}(M^\star(\pi) \,\|\, M(\pi))] \cdot \mathbb{I}\{\pi_\star \in \boldsymbol{\pi}_{\widehat{M}^s}\}\Bigg]$$

which implies

$$\mathbb{E}[s_T] \le \frac{4\mathsf{n}_{\max}}{\delta \mathsf{g}^\star} \cdot \mathbb{E}\Bigg[\frac{8\mathsf{n}_{\max} + 2\underline{\mathsf{g}}^{\mathcal{M}}}{\delta \underline{\mathsf{g}}^{\mathcal{M}}} \cdot \overline{\mathsf{aec}}^{\mathsf{D}}_{\varepsilon/2}(\mathcal{M}) \cdot \widehat{\mathbf{Est}}_{\mathsf{D}}(s_T) + \frac{2(1+\delta)\mathsf{g}^\star}{\Delta_{\min}} \cdot \mathbf{Est}_{\mathsf{KL}}(s_T)$$
$$+ (1+\delta)\mathsf{g}^\star \cdot \sum_{s=1}^{s_T} \inf_{M \in \mathcal{M}^\star_{\mathrm{alt}}} \mathbb{E}_{p^s}[D_{\mathsf{KL}}(M^\star(\pi) \,\|\, M(\pi))] \cdot \mathbb{I}\{\pi_\star \in \boldsymbol{\pi}_{\widehat{M}^s}\}\Bigg].$$

□

Lemma H.3.4. *When running both Algorithm H.2 and Algorithm H.3, for all $\alpha \in [0,1)$, we have*

$$\mathbb{E}\Bigg[\sum_{s=1}^{s_T} \inf_{M \in \mathcal{M}^{\mathrm{alt}}(M^\star)} s^\alpha \cdot \mathbb{E}_{p^s}[D_{\mathsf{KL}}(M^\star(\pi) \,\|\, M(\pi))] \cdot \mathbb{I}\{\pi_\star \in \boldsymbol{\pi}_{\widehat{M}^s}\}\Bigg]$$
$$\le \mathbb{E}[s_T^\alpha] \log T + \mathbb{E}\Bigg[V_{\mathcal{M}} s_T^{1/2+\alpha} \sqrt{1344 d_{\mathsf{cov}} \cdot \log(128 C_{\mathsf{cov}} s_T)}$$
$$+ (V_{\mathcal{M}} + L_{\mathsf{KL}}) \Bigg(4\frac{s_T^{1/2+\alpha/2}}{1-\alpha} + \sum_{s=1}^{s_T} s^{1/2+\alpha/2} \cdot \mathbb{E}_{\widehat{M} \sim \xi^s}[\mathbb{E}_{\pi \sim p^s}[D\big(\widehat{M}(\pi) \,\|\, M^\star(\pi)\big)]] \Bigg) \Bigg]$$
$$+ \mathbb{E}[s_T^\alpha] \log\log T + 7 V_{\mathcal{M}}.$$

Proof of Lemma H.3.4. Let $\widetilde{s}_T$ denote the final exploration round for which $\pi_\star \in \boldsymbol{\pi}_{\widehat{M}^s}$. Then,

upper bounding the KL divergence by $2V_{\mathcal{M}}$ via Lemma H.3.13, we have

$$\sum_{s=1}^{s_T} \inf_{M\in\mathcal{M}^\star_{\mathrm{alt}}} s^\alpha \mathbb{E}_{p^s}[D_{\mathsf{KL}}(M^\star(\pi)\,\|\,M(\pi))]\cdot\mathbb{I}\{\pi_\star\in\boldsymbol{\pi}_{\widehat{M}^s}\} \le \widetilde{s}_T^\alpha \sum_{s=1}^{\widetilde{s}_T-1} \inf_{M\in\mathcal{M}^\star_{\mathrm{alt}}} \mathbb{E}_{p^s}[D_{\mathsf{KL}}(M^\star(\pi)\,\|\,M(\pi))] + 2V_{\mathcal{M}} s_T^\alpha.$$

Under Assumption H.1 and via Jensen's inequality and AM-GM, we have, for any $\alpha_s > 0$ and $M\in\mathcal{M}$,

$$\begin{aligned}\mathbb{E}_{p^s}[D_{\mathsf{KL}}(M^\star(\pi)\,\|\,M(\pi))] &\le \mathbb{E}_{\widehat{M}\sim\xi^s}[\mathbb{E}_{p^s}[D_{\mathsf{KL}}\big(\widehat{M}(\pi)\,\|\,M(\pi)\big)]] + L_{\mathsf{KL}}\sqrt{\mathbb{E}_{\widehat{M}\sim\xi^s}[\mathbb{E}_{p^s}[D\big(\widehat{M}(\pi)\,\|\,M^\star(\pi)\big)]]}\\ &\le \mathbb{E}_{\widehat{M}\sim\xi^s}[\mathbb{E}_{p^s}[D_{\mathsf{KL}}\big(\widehat{M}(\pi)\,\|\,M(\pi)\big)]] + L_{\mathsf{KL}}\left(\frac{1}{\alpha_s} + \alpha_s\mathbb{E}_{\widehat{M}\sim\xi^s}[\mathbb{E}_{p^s}[D\big(\widehat{M}(\pi)\,\|\,M^\star(\pi)\big)]]\right),\end{aligned}$$

We now wish to bound

$$\mathbb{E}\left[\widetilde{s}_T^\alpha \sum_{s=1}^{\widetilde{s}_T-1} \inf_{M\in\mathcal{M}^\star_{\mathrm{alt}}} \mathbb{E}_{\widehat{M}\sim\xi^s}[\mathbb{E}_{p^s}[D_{\mathsf{KL}}\big(\widehat{M}(\pi)\,\|\,M(\pi)\big)]]\right].$$

Let $\mathcal{M}^j_{\mathrm{cov}}$ denote an (ρ_j,μ_j)-cover of $\mathcal{M}^\star_{\mathrm{alt}}$ and $\mathcal{E}^s_j$ the corresponding event at step $s\le 2^j$, for some ρ_j and μ_j to be chosen. Let $\mathcal{E}_j := \cap_{s\le 2^j}\mathcal{E}^s_j$. By definition $\mathbb{P}_{M^\star}[\mathcal{E}^s_j]\ge 1-\mu_j$, so $\mathbb{P}_{M^\star}[\mathcal{E}_j]\ge 1-2^j\mu_j$. Define an event

$$\begin{aligned}A_j := \Bigg\{ &\forall s\le 2^j, M\in\mathcal{M}^j_{\mathrm{cov}} \;:\; (s+1)^\alpha \sum_{i=1}^{s}\mathbb{E}_{\widehat{M}\sim\xi^i}[\mathbb{E}_{p^i}[D_{\mathsf{KL}}\big(\widehat{M}(\pi)\,\|\,M(\pi)\big)]]\\ &\le (s+1)^\alpha\sum_{i=1}^{s}\mathbb{E}_{\widehat{M}\sim\xi^i}\left[\log\frac{\mathbb{P}^{\widehat{M},\pi^i}(r^i,o^i)}{\mathbb{P}^{M,\pi^i}(r^i,o^i)}\right] + V_{\mathcal{M}}(s+1)^{\frac{1}{2}+\alpha}\sqrt{56\log\frac{2^j\mathsf{N}_{\mathrm{cov}}(\mathcal{M}^\star_{\mathrm{alt}},\rho_j,\mu_j)}{\delta_j}}\\ &+ V_{\mathcal{M}}\cdot\left(4\frac{(s+1)^{\frac{1+\alpha}{2}}}{1-\alpha} + \sum_{i=1}^{s} i^{\frac{1+\alpha}{2}}\cdot\mathbb{E}_{\widehat{M}\sim\xi^i}[\mathbb{E}_{\pi\sim p^i}[D\big(\widehat{M}(\pi)\,\|\,M^\star(\pi)\big)]]\right)\Bigg\}\end{aligned}$$

sfor some δ_j to be chosen. By invoking Lemma H.3.12 with $\beta_i = i^{1/2+\alpha/2}/(s+1)^\alpha$ and a union bound, we have $\mathbb{P}[A_j]\ge 1-\delta_j$ (while Lemma H.3.12 does not contain the $(s+1)^\alpha$ term, the bound in the expression for A_j simply gives the bound from Lemma H.3.12 multiplied through with $(s+1)^\alpha$, which is non-random, so this is admissible). We can decompose

$$\begin{aligned}&\mathbb{E}\left[\widetilde{s}_T^\alpha \sum_{s=1}^{\widetilde{s}_T-1} \inf_{M\in\mathcal{M}^\star_{\mathrm{alt}}} \mathbb{E}_{\widehat{M}\sim\xi^s}[\mathbb{E}_{p^s}[D_{\mathsf{KL}}\big(\widehat{M}(\pi)\,\|\,M(\pi)\big)]]\right]\\ &\quad\le \sum_{j=1}^{\lceil\log T\rceil}\mathbb{E}\left[\widetilde{s}_T^\alpha\cdot\inf_{M\in\mathcal{M}^\star_{\mathrm{alt}}}\sum_{s=1}^{\widetilde{s}_T-1}\mathbb{E}_{\widehat{M}\sim\xi^s}[\mathbb{E}_{p^s}[D_{\mathsf{KL}}\big(\widehat{M}(\pi)\,\|\,M(\pi)\big)]]\cdot\mathbb{I}\{\widetilde{s}_T\in[2^{j-1},2^j), A_j\cap\mathcal{E}_j\}\right]\end{aligned}$$

$$+ \sum_{j=1}^{\lceil \log T \rceil} \mathbb{E}\left[\widetilde{s}_T^{\alpha} \cdot \inf_{M \in \mathcal{M}^{\star}_{\mathrm{alt}}} \sum_{s=1}^{\widetilde{s}_T - 1} \mathbb{E}_{\widehat{M} \sim \xi^s}[\mathbb{E}_{p^s}[D_{\mathsf{KL}}\big(\widehat{M}(\pi) \,\|\, M(\pi)\big)]] \cdot \mathbb{I}\{\widetilde{s}_T \in [2^{j-1}, 2^j), A_j^c \cup \mathcal{E}_j^c\} \right].$$

We bound these terms separately. We can bound the second term as

$$\begin{aligned}
&\sum_{j=1}^{\lceil \log T \rceil} \mathbb{E}\left[\widetilde{s}_T^{\alpha} \cdot \inf_{M \in \mathcal{M}^{\star}_{\mathrm{alt}}} \sum_{s=1}^{\widetilde{s}_T - 1} \mathbb{E}_{\widehat{M} \sim \xi^s}[\mathbb{E}_{p^s}[D_{\mathsf{KL}}\big(\widehat{M}(\pi) \,\|\, M(\pi)\big)]] \cdot \mathbb{I}\{\widetilde{s}_T \in [2^{j-1}, 2^j), A_j^c \cup \mathcal{E}_j^c\} \right] \\
&\qquad \le \sum_{j=1}^{\lceil \log T \rceil} 2V_{\mathcal{M}} 2^{2j} (\mathbb{P}[A_j^c] + \mathbb{P}[\mathcal{E}_j^c]) \\
&\qquad \le \sum_{j=1}^{\lceil \log T \rceil} 2V_{\mathcal{M}} 2^{2j} (\delta_j + 2^j \mu_j)
\end{aligned}$$

where the first inequality follows by bounding Lemma H.3.13, and $\widetilde{s}_T \le 2^j$, and the second follows by our bound on the probability of A_j. Letting $\delta_j = \frac{1}{2^{3j}}, \mu_j = \frac{1}{2^{4j}}$, this term can be bounded by $4V_{\mathcal{M}}$. We turn now to the first term. Fix $j \in [\lceil \log T \rceil]$. By definition of the event A_j, and since $\widetilde{s}_T \le 2^j$, plugging in our choice of δ_j we can bound, for any $M \in \mathcal{M}^j_{\mathrm{cov}}$,

$$\begin{aligned}
&\mathbb{E}\left[\widetilde{s}_T^{\alpha} \cdot \inf_{M \in \mathcal{M}^{\star}_{\mathrm{alt}}} \sum_{s=1}^{\widetilde{s}_T - 1} \mathbb{E}_{\widehat{M} \sim \xi^s}[\mathbb{E}_{p^s}[D_{\mathsf{KL}}\big(\widehat{M}(\pi) \,\|\, M(\pi)\big)]] \cdot \mathbb{I}\{\widetilde{s}_T \in [2^{j-1}, 2^j), A_j \cap \mathcal{E}_j\} \right] \\
&\le \mathbb{E}\Bigg[\Bigg(\widetilde{s}_T^{\alpha} \cdot \inf_{M \in \mathcal{M}^{\star}_{\mathrm{alt}}} \sum_{s=1}^{\widetilde{s}_T - 1} \mathbb{E}_{\widehat{M} \sim \xi^s} \left[\log \frac{\mathbb{P}^{\widehat{M}, \pi^s}(r^s, o^s)}{\mathbb{P}^{M, \pi^s}(r^s, o^s)} \right] + V_{\mathcal{M}} \widetilde{s}_T^{1/2+\alpha} \sqrt{56 \log(2^{3j} \mathsf{N}_{\mathrm{cov}}(\mathcal{M}^{\star}_{\mathrm{alt}}, \rho_j, \mu_j))} \\
&\quad + V_{\mathcal{M}} \cdot \left(4 \frac{\widetilde{s}_T^{1/2+\alpha/2}}{1-\alpha} + \sum_{s=1}^{\widetilde{s}_T} s^{1/2+\alpha/2} \mathbb{E}_{\widehat{M} \sim \xi^s}[\mathbb{E}_{\pi \sim p^s}[D\big(\widehat{M}(\pi) \,\|\, M^{\star}(\pi)\big)]] \right) \Bigg) \cdot \mathbb{I}\{\widetilde{s}_T \in [2^{j-1}, 2^j), \mathcal{E}_j\} \Bigg] \\
&\le \mathbb{E}\Bigg[\Bigg(\widetilde{s}_T^{\alpha} \cdot \inf_{M \in \mathcal{M}^{\star}_{\mathrm{alt}}} \sum_{s=1}^{\widetilde{s}_T - 1} \mathbb{E}_{\widehat{M} \sim \xi^s} \left[\log \frac{\mathbb{P}^{\widehat{M}, \pi^s}(r^s, o^s)}{\mathbb{P}^{M, \pi^s}(r^s, o^s)} \right] + V_{\mathcal{M}} \widetilde{s}_T^{1/2+\alpha} \sqrt{168 \log(8 \widetilde{s}_T \mathsf{N}_{\mathrm{cov}}(\mathcal{M}^{\star}_{\mathrm{alt}}, \rho_j, \tfrac{\widetilde{s}_T^{-4}}{16}))} \\
&\quad + V_{\mathcal{M}} \cdot \left(4 \frac{\widetilde{s}_T^{1/2+\alpha/2}}{1-\alpha} + \sum_{s=1}^{\widetilde{s}_T} s^{1/2+\alpha/2} \mathbb{E}_{\widehat{M} \sim \xi^s}[\mathbb{E}_{\pi \sim p^s}[D\big(\widehat{M}(\pi) \,\|\, M^{\star}(\pi)\big)]] \right) \Bigg) \cdot \mathbb{I}\{\widetilde{s}_T \in [2^{j-1}, 2^j), \mathcal{E}_j\} \Bigg]
\end{aligned} \tag{H.3.4}$$

where the second inequality follows since, if $\widetilde{s}_T \in [2^{j-1}, 2^j)$, then we can bound $2^j \le 2\widetilde{s}_T$.

Since $\widetilde{s}_T$ is an exploration round by definition, we know that for all $\pi_{\widehat{M}^{\widetilde{s}_T}} \in \boldsymbol{\pi}_{\widehat{M}^{\widetilde{s}_T}}$, there

exists some $M' \in \mathcal{M}^{\text{alt}}(\pi_{\widehat{M}^{\widetilde{s}_T}})$ such that

$$\sum_{s=1}^{\widetilde{s}_T - 1} \mathbb{E}_{\widehat{M} \sim \xi^s}\left[\log \frac{\mathbb{P}^{\widehat{M},\pi^s}(r^s, o^s)}{\mathbb{P}^{M',\pi^s}(r^s, o^s)}\right] \le \log(T \log T).$$

By assumption we have $\pi_\star \in \boldsymbol{\pi}_{\widehat{M}^{\widetilde{s}_T}}$, so it follows that there exists some (random) $M' \in \mathcal{M}^\star_{\text{alt}}$ such that the above inequality holds. Let $M'' \in \mathcal{M}^j_{\text{cov}}$ denote the model in $\mathcal{M}^j_{\text{cov}}$ such that

$$\left|\log \frac{\mathbb{P}^{M',\pi}(r, o)}{\mathbb{P}^{M'',\pi}(r, o)}\right| = \left|\log \mathbb{P}^{M',\pi}(r, o) - \log \mathbb{P}^{M'',\pi}(r, o)\right| \le \rho_j,$$

for all (r, o, π) for which $\sup_{\widetilde{M} \in \mathcal{M}} \mathbb{P}^{\widetilde{M},\pi}(r, o \mid \mathcal{E}_j) > 0$, which is guaranteed to exist by Definition 10.2.1. Note that by definition of $\mathcal{M}^j_{\text{cov}}$, we have $M'' \in \mathcal{M}^\star_{\text{alt}}$. We then have

$$\begin{aligned}
&\widetilde{s}_T^{\alpha} \cdot \inf_{M \in \mathcal{M}^\star_{\text{alt}}} \sum_{s=1}^{\widetilde{s}_T - 1} \mathbb{E}_{\widehat{M} \sim \xi^s}\left[\log \frac{\mathbb{P}^{\widehat{M},\pi^s}(r^s, o^s)}{\mathbb{P}^{M,\pi^s}(r^s, o^s)}\right] \cdot \mathbb{I}\{\widetilde{s}_T \in [2^{j-1}, 2^j), \mathcal{E}_j\} \\
&\quad \le \widetilde{s}_T^{\alpha} \sum_{s=1}^{\widetilde{s}_T - 1} \mathbb{E}_{\widehat{M} \sim \xi^s}\left[\log \frac{\mathbb{P}^{\widehat{M},\pi^s}(r^s, o^s)}{\mathbb{P}^{M'',\pi^s}(r^s, o^s)}\right] \cdot \mathbb{I}\{\widetilde{s}_T \in [2^{j-1}, 2^j), \mathcal{E}_j\} \\
&\quad = \widetilde{s}_T^{\alpha} \sum_{s=1}^{\widetilde{s}_T - 1} \left[\mathbb{E}_{\widehat{M} \sim \xi^s}\left[\log \frac{\mathbb{P}^{\widehat{M},\pi^s}(r^s, o^s)}{\mathbb{P}^{M',\pi^s}(r^s, o^s)}\right] + \log \frac{\mathbb{P}^{M',\pi^s}(r^s, o^s)}{\mathbb{P}^{M'',\pi^s}(r^s, o^s)}\right] \cdot \mathbb{I}\{\widetilde{s}_T \in [2^{j-1}, 2^j), \mathcal{E}_j\} \\
&\quad \le \widetilde{s}_T^{\alpha} \log(T \log T) + \rho_j \cdot \widetilde{s}_T^{1+\alpha},
\end{aligned}$$

where the inequality holds since on $\mathcal{E}_j$, we can ensure that $\log \frac{\mathbb{P}^{M',\pi^s}(r^s,o^s)}{\mathbb{P}^{M'',\pi^s}(r^s,o^s)} \le \rho_j$. Therefore, choosing $\rho_j = 2^{-3j}$, we have

$$\begin{aligned}
(\text{H.3.4}) &\le \mathbb{E}\Bigg[\Bigg(\widetilde{s}_T^{\alpha} \log(T \log T) + \rho_j \cdot \widetilde{s}_T^{1+\alpha} + V_{\mathcal{M}} \widetilde{s}_T^{1/2+\alpha} \sqrt{168 \widetilde{s}_T \log(8 \mathsf{N}_{\text{cov}}(\mathcal{M}^\star_{\text{alt}}, \rho_j, \tfrac{\widetilde{s}_T^{-4}}{16}))} \\
&\qquad + V_{\mathcal{M}} \cdot \left(4 \frac{\widetilde{s}_T^{1/2+\alpha/2}}{1-\alpha} + \sum_{s=1}^{\widetilde{s}_T} s^{1/2+\alpha/2} \mathbb{E}_{\widehat{M} \sim \xi^s}[\mathbb{E}_{\pi \sim p^s}[D(\widehat{M}(\pi) \,\|\, M^\star(\pi))]]\right)\Bigg) \cdot \mathbb{I}\{\widetilde{s}_T \in [2^{j-1}, 2^j)\}\Bigg] \\
&\le \mathbb{E}\Bigg[\Bigg(\widetilde{s}_T^{\alpha} \log(T \log T) + 2^{-j} + V_{\mathcal{M}} \widetilde{s}_T^{\alpha+1/2} \sqrt{168 \log(8 \widetilde{s}_T \mathsf{N}_{\text{cov}}(\mathcal{M}^\star_{\text{alt}}, \tfrac{\widetilde{s}_T^{-3}}{8}, \tfrac{\widetilde{s}_T^{-4}}{16}))} \\
&\qquad + V_{\mathcal{M}} \cdot \left(4 \frac{\widetilde{s}_T^{1/2+\alpha/2}}{1-\alpha} + \sum_{s=1}^{\widetilde{s}_T} s^{1/2+\alpha/2} \mathbb{E}_{\widehat{M} \sim \xi^s}[\mathbb{E}_{\pi \sim p^s}[D(\widehat{M}(\pi) \,\|\, M^\star(\pi))]]\right)\Bigg) \cdot \mathbb{I}\{\widetilde{s}_T \in [2^{j-1}, 2^j)\}\Bigg].
\end{aligned}$$

As this holds for each j, and the events $\{\widetilde{s}_T \in [2^{j-1}, 2^j)\}$ are disjoint, we can sum over j to

bound

$$\sum_{j=1}^{\lceil \log T \rceil} \mathbb{E}\left[\widetilde{s}_T^{\alpha} \cdot \inf_{M \in \mathcal{M}_{\text{alt}}^{\star}} \sum_{s=1}^{\widetilde{s}_T - 1} \mathbb{E}_{\widehat{M} \sim \xi^s}[\mathbb{E}_{p^s}[D_{\mathsf{KL}}(\widehat{M}(\pi) \,\|\, M(\pi))]] \cdot \mathbb{I}\{\widetilde{s}_T \in [2^{j-1}, 2^j), A_j \cap \mathcal{E}_j\}\right]$$
$$\leq \mathbb{E}\Bigg[\widetilde{s}_T^{\alpha} \log(T \log T) + 1 + V_{\mathcal{M}} \widetilde{s}_T^{\alpha+1/2} \sqrt{168 \log(8\widetilde{s}_T \mathsf{N}_{\mathrm{cov}}(\mathcal{M}_{\text{alt}}^{\star}, \tfrac{\widetilde{s}_T^{-3}}{8}, \tfrac{\widetilde{s}_T^{-4}}{16}))}$$
$$+ V_{\mathcal{M}} \cdot \left(4\frac{\widetilde{s}_T^{1/2+\alpha/2}}{1-\alpha} + \sum_{s=1}^{\widetilde{s}_T} s^{1/2+\alpha/2} \mathbb{E}_{\widehat{M} \sim \xi^s}[\mathbb{E}_{\pi \sim p^s}[D(\widehat{M}(\pi) \,\|\, M^{\star}(\pi))]]\right)\Bigg].$$

Finally, under Assumption 10.6 and Lemma H.3.15 we can bound

$$\begin{aligned}
\log(8\widetilde{s}_T \mathsf{N}_{\mathrm{cov}}(\mathcal{M}_{\text{alt}}^{\star}, \tfrac{\widetilde{s}_T^{-3}}{8}, \tfrac{\widetilde{s}_T^{-4}}{16})) &\leq \log(8\widetilde{s}_T) + \log \mathsf{N}_{\mathrm{cov}}(\mathcal{M}, \tfrac{\widetilde{s}_T^{-3}}{8}, \tfrac{\widetilde{s}_T^{-4}}{16}) \\
&\leq \log(8\widetilde{s}_T) + d_{\mathsf{cov}} \cdot \log\big(128 C_{\mathsf{cov}} \widetilde{s}_T^{7}\big) \\
&\leq 8 d_{\mathsf{cov}} \cdot \log(128 C_{\mathsf{cov}} \widetilde{s}_T).
\end{aligned}$$

The result follows. □

H.3.1.3 Completing the Proof

Theorem H.1 (Full version of Theorem 10.1)**.** *For any $\delta \leq 1/2$, if we set $D(\cdot \,\|\, \cdot) = D_{\mathsf{KL}}(\cdot \,\|\, \cdot)$ and instantiate $\mathbf{Alg}_D$ with Algorithm H.4, under Assumptions 10.1 to 10.7 and if $\mathsf{g}^{\star} > 0$, the expected regret of Algorithm H.2 is bounded by*

$$\mathbb{E}^{M^{\star}}[\mathbf{Reg}(T)] \leq (1+\delta)\mathsf{g}^{\star} \cdot \log T + C_{\mathsf{aec}} \cdot \overline{\mathsf{aec}}_{\varepsilon/2}(\mathcal{M}) \cdot \log^{3/2}(\log T) + \mathrm{lin}\left(C_{\mathsf{low}}, \sqrt{\log T}, \log^{3/2}(\log T)\right),$$

for

$$C_{\mathsf{low}} := \mathrm{lin}\left(\mathsf{g}^{\star}, \max_{M \in \mathcal{M}} \mathsf{g}^{M}, V_{\mathcal{M}}^{13/2}, L_{\mathsf{KL}}^2, d_{\mathsf{cov}}, \log C_{\mathsf{cov}}, \frac{1}{\delta^2}, \frac{1}{\Delta_{\min}^3}, \mathsf{n}_{\delta/6}^{\mathcal{M}}, \sqrt{\mathsf{aec}_{\varepsilon/2}(\mathcal{M})}\right)$$
$$C_{\mathsf{aec}} := c \cdot \frac{V_{\mathcal{M}}^2 d_{\mathsf{cov}} \log(C_{\mathsf{cov}}) \cdot \max_{M \in \mathcal{M}} \mathsf{g}^{M} \cdot (\delta^{-1} + V_{\mathcal{M}} \mathsf{n}_{\delta/6}^{\mathcal{M}})}{\delta \Delta_{\min}^3} \cdot \log(C_{\mathsf{low}})$$

and where $\mathrm{lin}(\cdot)$ denotes a function linear and poly-logarithmic in its arguments and $c > 0$ is a universal constant.

Proof of Theorem H.1. Letting $\mathcal{T}_{\text{exploit}}$ denote the exploitation rounds, we can bound the

total expected regret as

$$\sum_{t=1}^{T}\mathbb{E}[\Delta^\star(\pi^t)] = \mathbb{E}\left[\sum_{t\in\mathcal{T}_{\text{exploit}}}\Delta^\star(\pi^t)\right] + \mathbb{E}\left[\sum_{s=1}^{s_T}\Delta^\star(\pi^s)\right] \le 2\log\log T + 3 + \mathbb{E}\left[\sum_{s=1}^{s_T}\Delta^\star(\pi^s)\right],$$

where the inequality follows from Lemma H.3.1. It remains to bound the regret in the exploration rounds. By Lemma H.3.3, we can bound this as

$$\begin{aligned}\mathbb{E}\left[\sum_{s=1}^{s_T}\Delta^\star(\pi^s)\right] \le \frac{8\mathsf{n}_{\max} + 2\underline{\mathsf{g}}^{\mathcal{M}}}{\delta\underline{\mathsf{g}}^{\mathcal{M}}}\cdot\overline{\mathsf{aec}}^{\mathsf{D}}_{\varepsilon/2}(\mathcal{M})\cdot\mathbb{E}[\widehat{\mathbf{Est}}_{\mathsf{D}}(s_T)] + \frac{2(1+\delta)\mathsf{g}^\star}{\Delta_{\min}}\cdot\mathbb{E}[\mathbf{Est}_{\mathsf{KL}}(s_T)] \\ + (1+\delta)\mathsf{g}^\star\cdot\mathbb{E}\left[\sum_{s=1}^{s_T}\inf_{M\in\mathcal{M}^\star_{\text{alt}}}\mathbb{E}_{p^s}[D_{\mathsf{KL}}(M^\star(\pi)\,\|\,M(\pi))]\cdot\mathbb{I}\{\pi_\star\in\boldsymbol{\pi}_{\widehat{M}^s}\}\right].\end{aligned}$$

Applying Lemma H.3.4 with $\alpha = 0$, we can bound

$$\begin{aligned}\mathbb{E}\left[\sum_{s=1}^{s_T}\inf_{M\in\mathcal{M}^\star_{\text{alt}}}\mathbb{E}_{p^s}[D_{\mathsf{KL}}(M^\star(\pi)\,\|\,M(\pi))]\cdot\mathbb{I}\{\pi_\star\in\boldsymbol{\pi}_{\widehat{M}^s}\}\right] &\le \log T + \mathbb{E}\Bigg[V_{\mathcal{M}}\sqrt{1344 d_{\mathsf{cov}} s_T\cdot\log(128 C_{\mathsf{cov}} s_T)} \\ &\quad + (V_{\mathcal{M}} + L_{\mathsf{KL}})\left(4\sqrt{s_T} + \sum_{s=1}^{s_T}\sqrt{s}\cdot\mathbb{E}_{\widehat{M}\sim\xi^s}[\mathbb{E}_{\pi\sim p^s}[D\big(\widehat{M}(\pi)\,\|\,M^\star(\pi)\big)]]\right)\Bigg] + \log\log T + 7V_{\mathcal{M}} \\ &\le \log T + V_{\mathcal{M}}\sqrt{1344 d_{\mathsf{cov}}\mathbb{E}[s_T]\cdot\log(128 C_{\mathsf{cov}}\mathbb{E}[s_T])} \\ &\quad + (V_{\mathcal{M}} + L_{\mathsf{KL}})\left(4\sqrt{\mathbb{E}[s_T]} + \mathbb{E}\left[\sum_{s=1}^{s_T}\sqrt{s}\cdot\mathbb{E}_{\widehat{M}\sim\xi^s}[\mathbb{E}_{\pi\sim p^s}[D\big(\widehat{M}(\pi)\,\|\,M^\star(\pi)\big)]]\right]\right) + \log\log T + 7V_{\mathcal{M}}\end{aligned} \tag{H.3.5}$$

where the last inequality follows from Lemma H.3.16—applied with $\alpha = \beta = 1/2$, $a = 128C_{\mathsf{cov}}$—and the concavity of $\sqrt{x}$. Note that the condition of Lemma H.3.16 is met here since we assume $C_{\mathsf{cov}} \ge 1$. It follows from Lemma H.3.7 that

$$\mathbb{E}\left[\sum_{s=1}^{s_T}\sqrt{s}\cdot\mathbb{E}_{\widehat{M}\sim\xi^s}[\mathbb{E}_{\pi\sim p^s}[D\big(\widehat{M}(\pi)\,\|\,M^\star(\pi)\big)]]\right] \tag{H.3.6}$$

$$\begin{aligned}&\le \mathbb{E}\Big[(2+6V_{\mathcal{M}})(20 d_{\mathsf{cov}}\cdot\log(64 C_{\mathsf{cov}} s_T) + 1)\cdot 5\sqrt{2 s_T\log(2 s_T)}\Big] \\ &\quad + \mathbb{E}[32(1+V_{\mathcal{M}})\log(s_T)] + 8 \\ &\le (2+6V_{\mathcal{M}})(20 d_{\mathsf{cov}}\cdot\log(64 C_{\mathsf{cov}}\mathbb{E}[s_T]) + 1)\cdot 5\sqrt{2\mathbb{E}[s_T]\log(2\mathbb{E}[s_T])} \\ &\quad + 32(1+V_{\mathcal{M}})\log\mathbb{E}[s_T] + 8 \\ &= O\Big(V_{\mathcal{M}} d_{\mathsf{cov}}\log(C_{\mathsf{cov}}\mathbb{E}[s_T])\sqrt{\mathbb{E}[s_T]\log(\mathbb{E}[s_T])} + V_{\mathcal{M}}\sqrt{\mathbb{E}[s_T]}\Big),\end{aligned} \tag{H.3.7}$$

where the last inequality follows from Lemma H.3.16. Similarly, by Lemma H.3.8, we can bound both $\mathbb{E}[\widehat{\mathbf{Est}}_{\mathsf{D}}(s_T)]$ and $\mathbb{E}[\mathbf{Est}_{\mathsf{KL}}(s_T)]$ as

$$\begin{aligned}\mathbb{E}[\widehat{\mathbf{Est}}_{\mathsf{D}}(s_T)], \mathbb{E}[\mathbf{Est}_{\mathsf{KL}}(s_T)] &\le \mathbb{E}\Big[(2+5V_{\mathcal{M}})(20d_{\mathsf{cov}}\cdot\log(64C_{\mathsf{cov}}s_T)+1)\cdot\sqrt{\log(2s_T)}\Big] \\ &\quad + \mathbb{E}[32(1+V_{\mathcal{M}})\log(s_T)] + 8 \\ &\le (2+5V_{\mathcal{M}})(20d_{\mathsf{cov}}\cdot\log(64C_{\mathsf{cov}}\mathbb{E}[s_T])+1)\cdot\sqrt{\log(2\mathbb{E}[s_T])} \\ &\quad + 32(1+V_{\mathcal{M}})\log(\mathbb{E}[s_T]) + 8 \\ &= O\Big(V_{\mathcal{M}}d_{\mathsf{cov}}\log(C_{\mathsf{cov}}\mathbb{E}[s_T])\sqrt{\log(\mathbb{E}[s_T])} + V_{\mathcal{M}}\log(\mathbb{E}[s_T])\Big),\end{aligned} \tag{H.3.8}$$

where the second inequality follows from Lemma H.3.16 and Jensen's inequality, which lets us pass the expectation through. Together this gives a regret bound of

$$\begin{aligned}&\mathbb{E}\Big[\sum_{s=1}^{s_T}\Delta^\star(\pi^s)\Big] \\ &\le (1+\delta)\mathsf{g}^\star\cdot\log T + (1+\delta)\mathsf{g}^\star V_{\mathcal{M}}\sqrt{1344d_{\mathsf{cov}}\mathbb{E}[s_T]\cdot\log(128C_{\mathsf{cov}}\mathbb{E}[s_T])} + (1+\delta)\mathsf{g}^\star(\log\log T + 7V_{\mathcal{M}}) \\ &\quad + (1+\delta)\mathsf{g}^\star(V_{\mathcal{M}}+L_{\mathsf{KL}})\cdot O\Big(\sqrt{\mathbb{E}[s_T]} + V_{\mathcal{M}}d_{\mathsf{cov}}\cdot\log(C_{\mathsf{cov}}\mathbb{E}[s_T])\sqrt{\mathbb{E}[s_T]\log(\mathbb{E}[s_T])} + V_{\mathcal{M}}\log\mathbb{E}[s_T]\Big) \\ &\quad + \frac{8\mathsf{n}_{\max}+2\underline{\mathsf{g}}^{\mathcal{M}}}{\delta\underline{\mathsf{g}}^{\mathcal{M}}}\cdot\overline{\mathsf{aec}}^{\mathsf{D}}_{\varepsilon/2}(\mathcal{M})\cdot O\Big(V_{\mathcal{M}}d_{\mathsf{cov}}\log(C_{\mathsf{cov}}\mathbb{E}[s_T])\sqrt{\log(\mathbb{E}[s_T])} + V_{\mathcal{M}}\log(\mathbb{E}[s_T])\Big) \\ &\quad + \frac{2(1+\delta)\mathsf{g}^\star}{\Delta_{\min}}\cdot O\Big(V_{\mathcal{M}}d_{\mathsf{cov}}\log(C_{\mathsf{cov}}\mathbb{E}[s_T])\sqrt{\log(\mathbb{E}[s_T])} + V_{\mathcal{M}}\log(\mathbb{E}[s_T])\Big)\end{aligned}$$

By Lemma H.3.3 we can bound the total number of exploration rounds by

$$\begin{aligned}\mathbb{E}[s_T] &\le \frac{24\mathsf{n}_{\max}^2+8\mathsf{n}_{\max}\underline{\mathsf{g}}^{\mathcal{M}}}{(\delta\underline{\mathsf{g}}^{\mathcal{M}})^2}\cdot\overline{\mathsf{aec}}^{\mathsf{D}}_{\varepsilon/2}(\mathcal{M})\cdot\mathbb{E}[\widehat{\mathbf{Est}}_{\mathsf{D}}(s_T)] + \frac{12\mathsf{n}_{\max}}{\delta\Delta_{\min}}\cdot\mathbb{E}[\mathbf{Est}_{\mathsf{KL}}(s_T)] \\ &\quad + \frac{6\mathsf{n}_{\max}}{\delta}\cdot\mathbb{E}\Big[\sum_{s=1}^{s_T}\inf_{M\in\mathcal{M}^\star_{\mathrm{alt}}}\mathbb{E}_{p^s}[D_{\mathsf{KL}}(M^\star(\pi)\,\|\,M(\pi))]\cdot\mathbb{I}\{\pi_\star\in\boldsymbol{\pi}_{\widehat{M}^s}\}\Big] \\ &\le \frac{24\mathsf{n}_{\max}^2+8\mathsf{n}_{\max}\underline{\mathsf{g}}^{\mathcal{M}}}{(\delta\underline{\mathsf{g}}^{\mathcal{M}})^2}\cdot\overline{\mathsf{aec}}^{\mathsf{D}}_{\varepsilon/2}(\mathcal{M})\cdot O\Big(V_{\mathcal{M}}d_{\mathsf{cov}}\log(C_{\mathsf{cov}}\mathbb{E}[s_T])\sqrt{\log(\mathbb{E}[s_T])} + V_{\mathcal{M}}\log(\mathbb{E}[s_T])\Big) \\ &\quad + \frac{12\mathsf{n}_{\max}}{\delta\Delta_{\min}}\cdot O\Big(V_{\mathcal{M}}d_{\mathsf{cov}}\log(C_{\mathsf{cov}}\mathbb{E}[s_T])\sqrt{\log(\mathbb{E}[s_T])} + V_{\mathcal{M}}\log(\mathbb{E}[s_T])\Big) \\ &\quad + \frac{6\mathsf{n}_{\max}}{\delta}\cdot(\log T + V_{\mathcal{M}}\sqrt{1344d_{\mathsf{cov}}\mathbb{E}[s_T]\cdot\log(128C_{\mathsf{cov}}\mathbb{E}[s_T])} + \log\log T + 7V_{\mathcal{M}}) \\ &\quad + \frac{6\mathsf{n}_{\max}}{\delta}(V_{\mathcal{M}}+L_{\mathsf{KL}})\Big(4\sqrt{\mathbb{E}[s_T]} + O\Big(V_{\mathcal{M}}d_{\mathsf{cov}}\log(C_{\mathsf{cov}}\mathbb{E}[s_T])\sqrt{\mathbb{E}[s_T]\log(\mathbb{E}[s_T])} + V_{\mathcal{M}}\sqrt{\mathbb{E}[s_T]}\Big)\Big)\end{aligned}$$

where the second inequality follows from Equations (H.3.5), (H.3.7) and (H.3.8). Using Lemma H.3.17 and bounding $\underline{\mathsf{g}}^{\mathcal{M}} \le \mathsf{n}_{\max}$, we can solve this for $\mathbb{E}[s_T]$ to get

$$\mathbb{E}[s_T] \le \widetilde{O}\left(\frac{\mathsf{n}_{\max}}{\delta} \cdot \log T + \frac{V_{\mathcal{M}} d_{\mathsf{cov}} \log C_{\mathsf{cov}} \cdot \mathsf{n}_{\max}^2}{(\delta \underline{\mathsf{g}}^{\mathcal{M}})^2} \cdot \overline{\mathsf{aec}}^{\mathsf{D}}_{\varepsilon/2}(\mathcal{M}) + \frac{\mathsf{n}_{\max} V_{\mathcal{M}} d_{\mathsf{cov}} \log C_{\mathsf{cov}}}{\delta \Delta_{\min}} + \frac{\mathsf{n}_{\max}^2 (V_{\mathcal{M}} + L_{\mathsf{KL}})^2 (V_{\mathcal{M}}^2 d_{\mathsf{cov}}^2 \log^2 C_{\mathsf{cov}} + V_{\mathcal{M}}^2)}{\delta^2}\right).$$

Finally, by Lemma H.3.13 and Lemma H.2.3 we can lower bound $\underline{\mathsf{g}}^{\mathcal{M}} \ge \Delta_{\min}/2V_{\mathcal{M}}$. Plugging this into the above expression and using that

$$\mathsf{n}_{\max} = \mathsf{n}_{\max}(\mathcal{M}, \delta/6) := \frac{32}{\Delta_{\min}^2} \cdot \left(\frac{6}{\delta} + 2V_{\mathcal{M}} \mathsf{n}^{\mathcal{M}}_{\delta/6}\right) \cdot \max_{M \in \mathcal{M}} \mathsf{g}^M,$$

gives the result, after simplifying.

□

H.3.2 Regret Bound without Uniform Regularity (Theorem 10.2)

In this section, we prove Theorem 10.2 (as well as a more general result, Theorem H.2), which gives regret bounds for a variant of AE^2, Algorithm H.3, $\mathsf{AE}^2_\star$, which does not require uniform regularity, and adapts to the gap $\Delta^\star_{\min} := \Delta^{M^\star}_{\min}$ for the underlying model $M^\star$. Algorithm H.3 is a slightly more general version of Algorithm 10.2, incorporating general divergences D.

Throughout this section, we let $s^\star$ denote the first exploration round s such that $M^\star \in \mathcal{M}^\ell$ in Algorithm 10.2. Note that $s^\star$ is a deterministic quantity (though, the first round t in which $s = s^\star$ is not deterministic).

We remark briefly that, to bound (H.3.9) by a restriction of the AEC, it is critical that our estimator produced at each exploration round s, ξ^s, is only supported on $\mathcal{M}^\ell$. To accomplish this, we explicitly generate a cover of $\mathcal{M}^\ell$, $\mathcal{M}^\ell_{\mathsf{cov}}$, and run an estimation procedure on this cover. As Algorithm H.4, the estimation procedure used to prove Theorem 10.1, directly covers all of $\mathcal{M}$, to prove Theorem 10.2 we do not appeal directly to this algorithm, yet we note that the covering and estimation procedure employed by $\mathsf{AE}^2_\star$ are essentially identical to that in Algorithm H.4, modulo the choice of which set is being covered.

H.3.2.1 Bounding Regret of Explore Phase

Lemma H.3.5. *We have*

$$s^\star \le \left(\overline{\mathsf{aec}}^{\mathsf{D},\mathcal{M}}_{\varepsilon/2}(\mathcal{M}^\star)\right)^{\frac{1}{\alpha_{\mathcal{M}}}} + \left(\mathsf{n}^\star_\varepsilon + \frac{1}{\Delta^\star_{\min}} + \frac{4\mathsf{g}^\star}{\Delta^\star_{\min}} + \frac{2\mathsf{n}^\star_\varepsilon}{\mathsf{g}^\star} + \frac{4}{\Delta^\star_{\min}\varepsilon}\right)^{\frac{2}{\alpha_{\mathsf{n}}}}.$$

Algorithm H.3 Adaptive Exploration for Allocation Estimation for classes without uniform regularity ($\mathsf{AE}^2_\star$)

input: Optimality tolerance δ, divergence D, estimation oracle $\mathbf{Alg}_D$, growth parameters α_q, α_n, $\alpha_{\mathcal{M}}$.
$s \leftarrow 1$, $\ell \leftarrow 1$, $\varepsilon \leftarrow \frac{\delta}{4+2\delta}$.
$q^s \leftarrow 1 - s^{-\alpha_q}$, $\mathsf{n}^s \leftarrow s^{\alpha_{\mathsf{n}}}$.
Compute $\xi^1 \leftarrow \mathbf{Alg}_D(\{\varnothing\})$ and $\widehat{M}^1 \leftarrow \mathbb{E}_{M\sim\xi^1}[M]$.
for $t = 1, 2, 3, \ldots$ **do**
  if $s \geq 2^\ell$ **then** // Form active set and cover
    $\ell \leftarrow \ell + 1$.
    $\Delta^\ell \leftarrow \arg\min_{\Delta\geq 0} \Delta \quad \text{s.t.} \quad \overline{\mathsf{aec}}^{\mathrm{D},\mathcal{M}}_{\varepsilon/2}(\mathcal{M}_{\Delta,\frac{1}{\Delta}}) \leq s^{\alpha_{\mathcal{M}}}$.
    $\mathcal{M}^\ell \leftarrow \mathcal{M}_{\Delta^\ell,\frac{1}{\Delta^\ell}} \cap \left\{M \in \mathcal{M} \;:\; \mathsf{n}^M_\varepsilon + \frac{1}{\Delta^M_{\min}} + \frac{4\mathsf{g}^M}{\Delta^M_{\min}} + \frac{2\mathsf{n}^M_\varepsilon}{\mathsf{g}^M} + \frac{4}{\Delta^M_{\min}\varepsilon} \leq \sqrt{\mathsf{n}^s}\right\}$.
    $\mathcal{M}^\ell_{\mathrm{cov}} \leftarrow (\rho_\ell, \mu_\ell)$-cover of $\mathcal{M}^\ell$ for $\rho_\ell \leftarrow 2^{-\ell}, \mu_\ell \leftarrow 2^{-5\ell}$, $\mathfrak{D}^\ell \leftarrow \varnothing$.
  if $\exists \pi_{\widehat{M}^s} \in \boldsymbol{\pi}_{\widehat{M}^s}$ s.t. $\forall M \in \mathcal{M}^{\mathrm{alt}}(\pi_{\widehat{M}^s})$, $\sum_{i=1}^{s-1} \mathbb{E}_{\widehat{M}\sim\xi^i}\left[\log \frac{\mathbb{P}^{\widehat{M},\pi^i}(r^i,o^i)}{\mathbb{P}^{M,\pi^i}(r^i,o^i)}\right] \geq \log(t\log t)$ **then** // Exploit
    Play $\pi_{\widehat{M}^s}$.
  else // Explore
    Set $p^s \leftarrow q^s\lambda^s + (1-q^s)\omega^s$ for

$$\lambda^s, \omega^s \leftarrow \arg\min_{\lambda,\omega\in\triangle_\Pi} \sup_{M\in\mathcal{M}^\ell\setminus\mathcal{M}^{\mathrm{gl}}_\varepsilon(\lambda;\mathsf{n}^s)} \frac{1}{\mathbb{E}_{\widehat{M}\sim\xi^s}[\mathbb{E}_{\pi\sim\omega}[D\big(\widehat{M}(\pi)\,\|\,M(\pi)\big)]]}. \tag{H.3.9}$$

    Draw $\pi^s \sim p^s$, observe r^s, o^s, set $\mathfrak{D}^\ell \leftarrow \mathfrak{D}^\ell \cup \{(\pi^s, r^s, o^s)\}$.
    Compute estimate $\xi^{s+1} \leftarrow \mathbf{Alg}_D(\mathfrak{D}^\ell, \mathcal{M}^\ell_{\mathrm{cov}})$ and $\widehat{M}^{s+1} = \mathbb{E}_{M\sim\xi^{s+1}}[M]$.
    $s \leftarrow s + 1$.

Proof of Lemma H.3.5. We will have $M^\star \in \mathcal{M}^\ell$ as soon as $M^\star \in \mathcal{M}'$ and $M^\star \in \mathcal{M}''$ for

$$\mathcal{M}' \leftarrow \left\{M \in \mathcal{M} \;:\; \mathsf{n}^M_\varepsilon + \frac{1}{\Delta^M_{\min}} + \frac{4\mathsf{g}^M}{\Delta^M_{\min}} + \frac{2\mathsf{n}^M_\varepsilon}{\mathsf{g}^M} + \frac{4}{\Delta^M_{\min}\varepsilon} \leq \sqrt{\mathsf{n}^{2^{\ell-1}}}\right\}, \quad \mathcal{M}'' \leftarrow \mathcal{M}_{\Delta^\ell,\frac{1}{\Delta^\ell}}$$

where

$$\Delta^\ell = \arg\min_{\Delta\geq 0} \Delta \quad \text{s.t.} \quad \overline{\mathsf{aec}}^{\mathrm{D},\mathcal{M}}_{\varepsilon/2}(\mathcal{M}_{\Delta,\frac{1}{\Delta}}) \leq s^{\alpha_{\mathcal{M}}}.$$

Note that if this occurs for some ℓ, then the first exploration round s such that $M^\star \in \mathcal{M}^\ell$ is $s = 2^{\ell-1}$. From the definition $\mathsf{n}^{2^{\ell-1}} = 2^{\alpha_{\mathsf{n}}(\ell-1)}$, we have $M^\star \in \mathcal{M}'$ as soon as

$$\sqrt{2^{\alpha_{\mathsf{n}}(\ell-1)}} \geq \mathsf{n}^\star_\varepsilon + \frac{1}{\Delta^\star_{\min}} + \frac{4\mathsf{g}^\star}{\Delta^\star_{\min}} + \frac{2\mathsf{n}^\star_\varepsilon}{\mathsf{g}^\star} + \frac{4}{\Delta^\star_{\min}\varepsilon}$$

$$\iff 2^{\ell-1} \geq \left(\mathsf{n}^\star_\varepsilon + \frac{1}{\Delta^\star_{\min}} + \frac{4\mathsf{g}^\star}{\Delta^\star_{\min}} + \frac{2\mathsf{n}^\star_\varepsilon}{\mathsf{g}^\star} + \frac{4}{\Delta^\star_{\min}\varepsilon}\right)^{2/\alpha_{\mathsf{n}}}.$$

To have $M^\star \in \mathcal{M}''$, we need $\Delta^\ell \leq \Delta^\star_{\min}$ and $\mathsf{n}^\star_\varepsilon \leq \frac{1}{\Delta^\ell}$, which will be the case once

$$\overline{\mathsf{aec}}^{\mathsf{D},\mathcal{M}}_{\varepsilon/2}(\mathcal{M}^\star) \leq 2^{(\ell-1)\cdot\alpha_{\mathcal{M}}} \iff \left(\overline{\mathsf{aec}}^{\mathsf{D},\mathcal{M}}_{\varepsilon/2}(\mathcal{M}^\star)\right)^{\frac{1}{\alpha_{\mathcal{M}}}} \leq 2^{\ell-1}.$$

The result then follows by combining these bounds. □

Lemma H.3.6. *Let s_T denote the total number of exploration rounds. For $\delta \leq 1/2$, running Algorithm H.3, we can almost surely bound bound*

$$\begin{aligned} s_T \leq \frac{4}{\delta\mathsf{g}^\star}\Bigg(&\sum_{s=s^\star}^{s_T} 2s^{\alpha_q+\alpha_{\mathcal{M}}}\mathbb{E}_{\widehat{M}\sim\xi^s}[\mathbb{E}_{p^s}[D\big(\widehat{M}(\pi)\,\|\,M^\star(\pi)\big)]] \\ &+ \sum_{s=s^\star}^{s_T} s^{\alpha_{\mathsf{n}}}\cdot 2\mathsf{g}^\star \cdot \inf_{M\in\mathcal{M}^\star_{\mathrm{alt}}} \mathbb{E}_{p^s}[D_{\mathsf{KL}}(M^\star(\pi)\,\|\,M(\pi))]\cdot\mathbb{I}\{\pi_\star \in \boldsymbol{\pi}_{\widehat{M}^s}\} \\ &+ \sum_{s=s^\star}^{s_T} s^{3\alpha_{\mathsf{n}}/2}\cdot 4\mathsf{g}^\star\mathbb{E}_{\widehat{M}\sim\xi^s}[\mathbb{E}_{p^s}[D_{\mathsf{KL}}\big(M^\star(\pi)\,\|\,\widehat{M}(\pi)\big)]] + 4\mathsf{g}^\star(\tfrac{8}{\delta\mathsf{g}^\star})^{\frac{1+\alpha_{\mathsf{n}}}{\alpha_q-\alpha_{\mathsf{n}}}}\Bigg) + s^\star. \end{aligned}$$

In addition, the regret during exploration rounds is bounded as

$$\begin{aligned} \mathbb{E}\Bigg[\sum_{s=s^\star}^{s_T}\Delta^\star(\pi^s)\Bigg] \leq \mathbb{E}\Bigg[&\sum_{s=s^\star}^{s_T} 2s^{\alpha_q+\alpha_{\mathcal{M}}}\mathbb{E}_{\widehat{M}\sim\xi^s}[\mathbb{E}_{p^s}[D\big(\widehat{M}(\pi)\,\|\,M^\star(\pi)\big)] \\ &+ \sum_{s=s^\star}^{s_T}(1+\delta)\mathsf{g}^\star \cdot \inf_{M\in\mathcal{M}^\star_{\mathrm{alt}}} \mathbb{E}_{p^s}[D_{\mathsf{KL}}(M^\star(\pi)\,\|\,M(\pi))]\cdot\mathbb{I}\{\pi_\star \in \boldsymbol{\pi}_{\widehat{M}^s}\} \\ &+ \sum_{s=s^\star}^{s_T} s^{\alpha_{\mathsf{n}}/2}\cdot 4\mathsf{g}^\star\mathbb{E}_{\widehat{M}\sim\xi^s}[\mathbb{E}_{p^s}[D_{\mathsf{KL}}\big(M^\star(\pi)\,\|\,\widehat{M}(\pi)\big)]]\Bigg] + 4\mathsf{g}^\star(\tfrac{8}{\delta\mathsf{g}^\star})^{\frac{1}{\alpha_q-\alpha_{\mathsf{n}}}}. \end{aligned}$$

Proof of Lemma H.3.6. We first prove the bound on the regret during the exploration rounds, then use this result to prove the bound on s_T.

Recall that by definition, for exploration rounds, we have $p^s \leftarrow q^s\lambda^s + (1-q^s)\omega^s$. We consider three cases to bound the instantaneous expected regret, $\Delta^\star(p^s)$, for each $s \geq s^\star$.

Case 1: $M^\star \in \mathcal{M}^\ell \backslash \mathcal{M}^{\mathrm{gl}}_\varepsilon(\lambda^s; \mathsf{n}^s)$**.** Denote such rounds as $\mathcal{S}^1_{\mathrm{exp}}$. This case follows identically to Case 1 in Lemma H.3.3 with

$$\gamma^s = \frac{1+\mathsf{g}^\star\delta}{1+q^s}\cdot\frac{1}{\mathbb{E}_{\widehat{M}\sim\xi^s}[\mathbb{E}_{\omega^s}[D\big(\widehat{M}(\pi)\,\|\,M^\star(\pi)\big)]]}. \tag{H.3.10}$$

We therefore omit the proof and conclude that

$$\begin{aligned}\Delta^\star(p^s) &\le -\mathsf{g}^\star\delta + \gamma^s \mathbb{E}_{\widehat{M}\sim\xi^s}[\mathbb{E}_{p^s}[D\big(\widehat{M}(\pi)\,\|\,M^\star(\pi)\big)]] \\ &\le \gamma^s \mathbb{E}_{\widehat{M}\sim\xi^s}[\mathbb{E}_{p^s}[D\big(\widehat{M}(\pi)\,\|\,M^\star(\pi)\big)].\end{aligned}$$

As regret is always lower-bounded by 0, we have $\Delta^\star(p^s) \ge 0$, so for rounds $s \in \mathcal{S}^1_{\mathrm{exp}}$, we can also write

$$\mathsf{g}^\star\delta \le \gamma^s \mathbb{E}_{\widehat{M}\sim\xi^s}[\mathbb{E}_{p^s}[D\big(\widehat{M}(\pi)\,\|\,M^\star(\pi)\big)]. \tag{H.3.11}$$

Case 2: $M^\star \in \mathcal{M}^{\mathrm{gl}}_\varepsilon(\lambda^s;\mathsf{n}^s)$**,** $\pi_\star \in \boldsymbol{\pi}_{\widehat{M}^s}$**.** Denote such rounds as $\mathcal{S}^2_{\mathrm{exp}}$, and write

$$\Delta^\star(p^s) = [\Delta^\star(p^s) - (1+\delta)\mathsf{g}^\star \cdot \mathbb{E}_{p^s}[D_{\mathsf{KL}}(M^\star(\pi)\,\|\,M(\pi))]] + (1+\delta)\mathsf{g}^\star \cdot \mathbb{E}_{p^s}[D_{\mathsf{KL}}(M^\star(\pi)\,\|\,M(\pi))]$$

for an arbitrary model $M \in \mathcal{M}^\star_{\mathrm{alt}}$. In this case, since $M^\star \in \mathcal{M}^{\mathrm{gl}}_\varepsilon(\lambda^s;\mathsf{n}^s)$, we have that $\lambda^s \in \Upsilon(M^\star;\varepsilon)$. This implies that

$$\Delta^\star(\lambda^s) \le (1+\varepsilon)\mathsf{g}^\star/\mathsf{n}^\star_s \quad \text{and} \quad \inf_{M\in\mathcal{M}^\star_{\mathrm{alt}}} \mathbb{E}_{\lambda^s}[D_{\mathsf{KL}}(M^\star(\pi)\,\|\,M(\pi))] \ge (1-\varepsilon)/\mathsf{n}^\star_s$$

for some $\mathsf{n}^\star_s \le \mathsf{n}^s$.

Since $M \in \mathcal{M}^\star_{\mathrm{alt}}$, it follows that $\mathbb{E}_{\lambda^s}[D_{\mathsf{KL}}(M^\star(\pi)\,\|\,M(\pi))] \ge (1-\varepsilon^s)/\mathsf{n}^\star_s$. Thus,

$$\begin{aligned}\Delta^\star(p^s) - (1+\delta)\mathsf{g}^\star\mathbb{E}_{p^s}[D_{\mathsf{KL}}(M^\star(\pi)\,\|\,M(\pi))] &\le q^s\left[\Delta^\star(\lambda^s) - (1+\delta)\mathsf{g}^\star\mathbb{E}_{\lambda^s}[D_{\mathsf{KL}}(M^\star(\pi)\,\|\,M(\pi))]\right] + 1 - q^s \\ &\le q^s\left[(1+\varepsilon)\mathsf{g}^\star/\mathsf{n}^\star_s - (1+\delta)(1-\varepsilon)\mathsf{g}^\star/\mathsf{n}^\star_s\right] + 1 - q^s \\ &= q^s\left[2\varepsilon - \delta(1-\varepsilon)\right]\cdot\frac{\mathsf{g}^\star}{\mathsf{n}^\star_s} + 1 - q^s \\ &\overset{(a)}{=} -\frac{(1-s^{-\alpha_q})\delta}{2}\cdot\frac{\mathsf{g}^\star}{\mathsf{n}^\star_s} + s^{-\alpha_q} \\ &\overset{(b)}{\le} \begin{cases} s^{-\alpha_q} & s < (\frac{8\mathsf{n}^\star_s}{\delta\mathsf{g}^\star})^{1/\alpha_q} \\ -\frac{\delta\mathsf{g}^\star}{4\mathsf{n}^\star_s} & s \ge (\frac{8\mathsf{n}^\star_s}{\delta\mathsf{g}^\star})^{1/\alpha_q}\end{cases} \\ &\overset{(c)}{\le} \begin{cases} s^{-\alpha_q} & s < (\frac{8}{\delta\mathsf{g}^\star})^{\frac{1}{\alpha_q-\alpha_{\mathsf{n}}}} \\ -\frac{1}{4}\delta\mathsf{g}^\star\cdot s^{-\alpha_{\mathsf{n}}} & s \ge (\frac{8}{\delta\mathsf{g}^\star})^{\frac{1}{\alpha_q-\alpha_{\mathsf{n}}}}\end{cases} \\ &\overset{(d)}{\le} 2\mathsf{g}^\star\mathbb{I}\{s < (\tfrac{8}{\delta\mathsf{g}^\star})^{\frac{1}{\alpha_q-\alpha_{\mathsf{n}}}}\} - \frac{1}{4}\delta\mathsf{g}^\star\cdot s^{-\alpha_{\mathsf{n}}}\end{aligned}$$

where (a) follows from our choice of $\varepsilon = \frac{\delta/2}{2+\delta}$ and $q^s = 1 - s^{-\alpha_q}$, (b) follows from some algebra, (c) uses that $\mathsf{n}^\star_s \le \mathsf{n}^s$ and $\mathsf{n}^s = s^{\alpha_{\mathsf{n}}}$, and (d) follows since we will always have

$s^{-\alpha_q} \le 2\mathsf{g}^\star - \frac{1}{4}\delta\mathsf{g}^\star \cdot s^{-\alpha_{\mathsf{n}}}$. Thus:

$$\begin{aligned}\Delta^\star(p^s) &\le (1+\delta)\mathsf{g}^\star \mathbb{E}_{p^s}[D_{\mathsf{KL}}(M^\star(\pi) \,\|\, M(\pi))] + 2\mathsf{g}^\star \mathbb{I}\{s < (\tfrac{8}{\delta\mathsf{g}^\star})^{\frac{1}{\alpha_q - \alpha_{\mathsf{n}}}}\} - \frac{1}{4}\delta\mathsf{g}^\star \cdot s^{-\alpha_{\mathsf{n}}} \\ &\le (1+\delta)\mathsf{g}^\star \mathbb{E}_{p^s}[D_{\mathsf{KL}}(M^\star(\pi) \,\|\, M(\pi))] + 2\mathsf{g}^\star \mathbb{I}\{s < (\tfrac{8}{\delta\mathsf{g}^\star})^{\frac{1}{\alpha_q - \alpha_{\mathsf{n}}}}\}.\end{aligned}$$

As this holds for all $M \in \mathcal{M}^\star_{\mathrm{alt}}$, we have

$$\Delta^\star(p^s) \le (1+\delta)\mathsf{g}^\star \cdot \inf_{M \in \mathcal{M}^\star_{\mathrm{alt}}} \mathbb{E}_{p^s}[D_{\mathsf{KL}}(M^\star(\pi) \,\|\, M(\pi))] + 2\mathsf{g}^\star \mathbb{I}\{s < (\tfrac{8}{\delta\mathsf{g}^\star})^{\frac{1}{\alpha_q - \alpha_{\mathsf{n}}}}\}.$$

Since $\Delta^\star(p^s) \ge 0$, this also implies that for $s \in \mathcal{S}^2_{\mathrm{exp}}$:

$$\begin{aligned}\frac{1}{4}\delta\mathsf{g}^\star &\le s^{\alpha_{\mathsf{n}}} \cdot 2\mathsf{g}^\star \cdot \inf_{M \in \mathcal{M}^\star_{\mathrm{alt}}} \mathbb{E}_{p^s}[D_{\mathsf{KL}}(M^\star(\pi) \,\|\, M(\pi))] + 2\mathsf{g}^\star s^{\alpha_{\mathsf{n}}} \mathbb{I}\{s < (\tfrac{8}{\delta\mathsf{g}^\star})^{\frac{1}{\alpha_q - \alpha_{\mathsf{n}}}}\} \\ &\le s^{\alpha_{\mathsf{n}}} \cdot 2\mathsf{g}^\star \cdot \inf_{M \in \mathcal{M}^\star_{\mathrm{alt}}} \mathbb{E}_{p^s}[D_{\mathsf{KL}}(M^\star(\pi) \,\|\, M(\pi))] + 2\mathsf{g}^\star (\tfrac{8}{\delta\mathsf{g}^\star})^{\frac{\alpha_{\mathsf{n}}}{\alpha_q - \alpha_{\mathsf{n}}}} \cdot \mathbb{I}\{s < (\tfrac{8}{\delta\mathsf{g}^\star})^{\frac{1}{\alpha_q - \alpha_{\mathsf{n}}}}\}.\end{aligned} \tag{H.3.12}$$

Case 3: $M^\star \in \mathcal{M}^{\mathrm{gl}}_\varepsilon(\lambda^s; \mathsf{n}^s)$**,** $\pi_\star \notin \boldsymbol{\pi}_{\widehat{M}^s}$**.** Denote such rounds as $\mathcal{S}^3_{\mathrm{exp}}$, and write

$$\begin{aligned}\Delta^\star(p^s) = \Big[\Delta^\star(p^s) - 2(1+\delta)\mathsf{g}^\star\sqrt{\mathsf{n}^s} \cdot \mathbb{E}_{\widehat{M}\sim\xi^s}[\mathbb{E}_{p^s}[D_{\mathsf{KL}}\big(M^\star(\pi) \,\|\, \widehat{M}(\pi)\big)]]\Big] \\ + 2(1+\delta)\mathsf{g}^\star\sqrt{\mathsf{n}^s} \cdot \mathbb{E}_{\widehat{M}\sim\xi^s}[\mathbb{E}_{p^s}[D_{\mathsf{KL}}\big(M^\star(\pi) \,\|\, \widehat{M}(\pi)\big)]].\end{aligned}$$

Since $M^\star \in \mathcal{M}^{\mathrm{gl}}_\varepsilon(\lambda^s)$, we have that $\lambda^s \in \Upsilon(M^\star; \varepsilon)$. This implies that for any $M \in \mathcal{M}^\star_{\mathrm{alt}}$:

$$\Delta^\star(\lambda^s) \le (1+\varepsilon)\mathsf{g}^\star/\mathsf{n}^\star_s \quad \text{and} \quad \inf_{M \in \mathcal{M}^\star_{\mathrm{alt}}} \mathbb{E}_{\lambda^s}[D_{\mathsf{KL}}(M^\star(\pi) \,\|\, M(\pi))] \ge (1-\varepsilon)/\mathsf{n}^\star_s$$

for some $\mathsf{n}^\star_s \le \mathsf{n}^s$. Assume that we are at epoch ℓ. By construction we have that, for $M \in \mathcal{M}^\ell$, $\frac{1}{\Delta^M_{\min}} \le \sqrt{\mathsf{n}^{2^\ell}} \iff \Delta^M_{\min} \ge \frac{1}{\sqrt{\mathsf{n}^{2^\ell}}}$. Since n^s is increasing in s, this implies that $\Delta^M_{\min} \ge \frac{1}{\sqrt{\mathsf{n}^s}}$. As ξ^s is only supported on $\mathcal{M}^\ell$, since $\pi_\star \notin \boldsymbol{\pi}_{\widehat{M}^s}$, Lemma H.3.9 implies that $\mathbb{E}_{M\sim\xi^s}[\mathbb{I}\{M \in \mathcal{M}^\star_{\mathrm{alt}}\}] \ge \frac{1}{2\sqrt{\mathsf{n}^s}}$. Thus, we have

$$\begin{aligned}&2(1+\delta)\mathsf{g}^\star\sqrt{\mathsf{n}^s} \cdot \mathbb{E}_{\widehat{M}\sim\xi^s}[\mathbb{E}_{\lambda^s}[D_{\mathsf{KL}}\big(M^\star(\pi) \,\|\, \widehat{M}(\pi)\big)]] \\ &\ge 2(1+\delta)\mathsf{g}^\star\sqrt{\mathsf{n}^s} \cdot \mathbb{E}_{\widehat{M}\sim\xi^s}[\mathbb{E}_{\lambda^s}[D_{\mathsf{KL}}\big(M^\star(\pi) \,\|\, \widehat{M}(\pi)\big) \cdot \mathbb{I}\{M \in \mathcal{M}^\star_{\mathrm{alt}}\}]] \\ &\ge 2(1+\delta)\mathsf{g}^\star\sqrt{\mathsf{n}^s} \cdot \frac{1-\varepsilon}{2\sqrt{\mathsf{n}^s}\mathsf{n}^\star_s} \\ &\ge \frac{(1+\delta)(1-\varepsilon)\mathsf{g}^\star}{\mathsf{n}^\star_s}.\end{aligned}$$

This implies that

$$\begin{aligned}
&\Delta^\star(p^s) - 2(1+\delta)\mathsf{g}^\star\sqrt{\mathsf{n}^s}\cdot\mathbb{E}_{\widehat{M}\sim\xi^s}[\mathbb{E}_{p^s}[D_{\mathsf{KL}}\big(M^\star(\pi)\,\|\,\widehat{M}(\pi)\big)]]\\
&\le q^s[(1+\varepsilon)\mathsf{g}^\star/\mathsf{n}_s^\star - (1+\delta)(1-\varepsilon)\mathsf{g}^\star/\mathsf{n}_s^\star] + 1 - q^s\\
&\le 2\mathsf{g}^\star\mathbb{I}\{s < (\tfrac{8}{\delta\mathsf{g}^\star})^{\frac{1}{\alpha_q-\alpha_{\mathsf{n}}}}\} - \frac{1}{4}\delta\mathsf{g}^\star\cdot s^{-\alpha_{\mathsf{n}}},
\end{aligned}$$

where the final inequality follows by the same argument as in Case 2. Thus,

$$\begin{aligned}
\Delta^\star(p^s) &\le 2(1+\delta)\mathsf{g}^\star\sqrt{\mathsf{n}^s}\cdot\mathbb{E}_{\widehat{M}\sim\xi^s}[\mathbb{E}_{p^s}[D_{\mathsf{KL}}\big(M^\star(\pi)\,\|\,\widehat{M}(\pi)\big)]] + 2\mathsf{g}^\star\mathbb{I}\{s < (\tfrac{8}{\delta\mathsf{g}^\star})^{\frac{1}{\alpha_q-\alpha_{\mathsf{n}}}}\} - \frac{1}{4}\delta\mathsf{g}^\star\cdot s^{-\alpha_{\mathsf{n}}}\\
&= s^{\alpha_{\mathsf{n}}/2}\cdot 2(1+\delta)\mathsf{g}^\star\mathbb{E}_{\widehat{M}\sim\xi^s}[\mathbb{E}_{p^s}[D_{\mathsf{KL}}\big(M^\star(\pi)\,\|\,\widehat{M}(\pi)\big)]] + 2\mathsf{g}^\star\mathbb{I}\{s < (\tfrac{8}{\delta\mathsf{g}^\star})^{\frac{1}{\alpha_q-\alpha_{\mathsf{n}}}}\} - \frac{1}{4}\delta\mathsf{g}^\star\cdot s^{-\alpha_{\mathsf{n}}}\\
&\le s^{\alpha_{\mathsf{n}}/2}\cdot 2(1+\delta)\mathsf{g}^\star\mathbb{E}_{\widehat{M}\sim\xi^s}[\mathbb{E}_{p^s}[D_{\mathsf{KL}}\big(M^\star(\pi)\,\|\,\widehat{M}(\pi)\big)]] + 2\mathsf{g}^\star\mathbb{I}\{s < (\tfrac{8}{\delta\mathsf{g}^\star})^{\frac{1}{\alpha_q-\alpha_{\mathsf{n}}}}\}.
\end{aligned}$$

Just as in Case 2, using that $\Delta^\star(p^s)\ge 0$, this also implies that for $s\in\mathcal{S}^3_{\mathrm{exp}}$:

$$\frac{1}{4}\delta\mathsf{g}^\star \le s^{3\alpha_{\mathsf{n}}/2}\cdot 2(1+\delta)\mathsf{g}^\star\mathbb{E}_{\widehat{M}\sim\xi^s}[\mathbb{E}_{p^s}[D_{\mathsf{KL}}\big(M^\star(\pi)\,\|\,\widehat{M}(\pi)\big)]] + 2\mathsf{g}^\star(\tfrac{8}{\delta\mathsf{g}^\star})^{\frac{\alpha_{\mathsf{n}}}{\alpha_q-\alpha_{\mathsf{n}}}}\cdot\mathbb{I}\{s < (\tfrac{8}{\delta\mathsf{g}^\star})^{\frac{1}{\alpha_q-\alpha_{\mathsf{n}}}}\}. \tag{H.3.13}$$

Completing the Proof. In total we have

$$\begin{aligned}
\sum_{s=s^\star}^{s_T}\Delta^\star(p^s) \le \sum_{s\in\mathcal{S}^1_{\mathrm{exp}}}\gamma^s\mathbb{E}_{\widehat{M}\sim\xi^s}[\mathbb{E}_{p^s}[D\big(\widehat{M}(\pi)\,\|\,M^\star(\pi)\big)] + \sum_{s\in\mathcal{S}^2_{\mathrm{exp}}}(1+\delta)\mathsf{g}^\star\cdot\inf_{M\in\mathcal{M}^\star_{\mathrm{alt}}}\mathbb{E}_{p^s}[D_{\mathsf{KL}}(M^\star(\pi)\,\|\,M(\pi))]\\
+\sum_{s\in\mathcal{S}^3_{\mathrm{exp}}}s^{\alpha_{\mathsf{n}}/2}\cdot 2(1+\delta)\mathsf{g}^\star\mathbb{E}_{\widehat{M}\sim\xi^s}[\mathbb{E}_{p^s}[D_{\mathsf{KL}}\big(M^\star(\pi)\,\|\,\widehat{M}(\pi)\big)]] + 4\mathsf{g}^\star(\tfrac{8}{\delta\mathsf{g}^\star})^{\frac{1}{\alpha_q-\alpha_{\mathsf{n}}}}.
\end{aligned}$$

By Lemma H.3.11, for $s\in\mathcal{S}^1_{\mathrm{exp}}$, we can bound $\gamma^s\le(1+\mathsf{g}^\star\delta)\cdot s^{\alpha_q+\alpha_{\mathcal{M}}}$. This gives an upper bound on the above of

$$\begin{aligned}
&\le\sum_{s=s^\star}^{s_T}(1+\mathsf{g}^\star)s^{\alpha_q+\alpha_{\mathcal{M}}}\mathbb{E}_{\widehat{M}\sim\xi^s}[\mathbb{E}_{p^s}[D\big(\widehat{M}(\pi)\,\|\,M^\star(\pi)\big)]\\
&\quad+\sum_{s=s^\star}^{s_T}(1+\delta)\mathsf{g}^\star\cdot\inf_{M\in\mathcal{M}^\star_{\mathrm{alt}}}\mathbb{E}_{p^s}[D_{\mathsf{KL}}(M^\star(\pi)\,\|\,M(\pi))]\cdot\mathbb{I}\{\pi_\star\in\boldsymbol{\pi}_{\widehat{M}^s}\}\\
&\quad+\sum_{s=s^\star}^{s_T}s^{\alpha_{\mathsf{n}}/2}\cdot 4\mathsf{g}^\star\mathbb{E}_{\widehat{M}\sim\xi^s}[\mathbb{E}_{p^s}[D_{\mathsf{KL}}\big(M^\star(\pi)\,\|\,\widehat{M}(\pi)\big)]] + 4\mathsf{g}^\star(\tfrac{8}{\delta\mathsf{g}^\star})^{\frac{1}{\alpha_q-\alpha_{\mathsf{n}}}},
\end{aligned}$$

which proves the regret bound.

We now bound the number of exploration rounds. Since for every $s\in\mathcal{S}^1_{\mathrm{exp}}$ (H.3.11)

holds, for every $s \in \mathcal{S}^2_{\text{exp}}$ (H.3.12) holds, and for every $s \in \mathcal{S}^3_{\text{exp}}$ (H.3.13) holds, combining these inequalities gives

$$\begin{aligned}
&\frac{1}{4}\delta \mathsf{g}^\star |\mathcal{S}^1_{\text{exp}} \cup \mathcal{S}^2_{\text{exp}} \cup \mathcal{S}^3_{\text{exp}}| \\
&\quad \le \sum_{s \in \mathcal{S}^1_{\text{exp}}} \gamma^s \mathbb{E}_{\widehat{M} \sim \xi^s}[\mathbb{E}_{p^s}[D\big(\widehat{M}(\pi) \,\|\, M^\star(\pi)\big)] + \sum_{s \in \mathcal{S}^2_{\text{exp}}} s^{\alpha_{\mathsf{n}}} \cdot 2\mathsf{g}^\star \cdot \inf_{M \in \mathcal{M}^\star_{\text{alt}}} \mathbb{E}_{p^s}[D_{\mathsf{KL}}(M^\star(\pi) \,\|\, M(\pi))] \\
&\qquad + \sum_{s \in \mathcal{S}^3_{\text{exp}}} s^{3\alpha_{\mathsf{n}}/2} \cdot 4\mathsf{g}^\star \mathbb{E}_{\widehat{M} \sim \xi^s}[\mathbb{E}_{p^s}[D_{\mathsf{KL}}\big(M^\star(\pi) \,\|\, \widehat{M}(\pi)\big)]] + 4\mathsf{g}^\star (\tfrac{8}{\delta \mathsf{g}^\star})^{\frac{1+\alpha_{\mathsf{n}}}{\alpha_q - \alpha_{\mathsf{n}}}} \\
&\quad \le \sum_{s=s^\star}^{s_T} (1+\mathsf{g}^\star) s^{\alpha_q + \alpha_{\mathcal{M}}} \mathbb{E}_{\widehat{M} \sim \xi^s}[\mathbb{E}_{p^s}[D\big(\widehat{M}(\pi) \,\|\, M^\star(\pi)\big)] \\
&\qquad + \sum_{s=s^\star}^{s_T} s^{\alpha_{\mathsf{n}}} \cdot 2\mathsf{g}^\star \cdot \inf_{M \in \mathcal{M}^\star_{\text{alt}}} \mathbb{E}_{p^s}[D_{\mathsf{KL}}(M^\star(\pi) \,\|\, M(\pi))] \cdot \mathbb{I}\{\pi_\star \in \boldsymbol{\pi}_{\widehat{M}^s}\} \\
&\qquad + \sum_{s=s^\star}^{s_T} s^{3\alpha_{\mathsf{n}}/2} \cdot 4\mathsf{g}^\star \mathbb{E}_{\widehat{M} \sim \xi^s}[\mathbb{E}_{p^s}[D_{\mathsf{KL}}\big(M^\star(\pi) \,\|\, \widehat{M}(\pi)\big)]] + 4\mathsf{g}^\star (\tfrac{8}{\delta \mathsf{g}^\star})^{\frac{1+\alpha_{\mathsf{n}}}{\alpha_q - \alpha_{\mathsf{n}}}}.
\end{aligned}$$

Using that $|\mathcal{S}^1_{\text{exp}} \cup \mathcal{S}^2_{\text{exp}} \cup \mathcal{S}^3_{\text{exp}}| = s_T - s^\star$ and rearranging gives the bound on s_T.

□

H.3.2.2 Completing the Proof

Theorem H.2 (Full version of Theorem 10.2)**.** *Consider Algorithm H.3, and suppose we set $\delta \le 1/2$, $D(\cdot \,\|\, \cdot) = D_{\mathsf{KL}}(\cdot \,\|\, \cdot)$, $\alpha_{\mathcal{M}} = 1/2$, $\alpha_\zeta = 1/8$, and $\alpha_q = 1/4$, and instantiate $\mathbf{Alg}_D$ with Algorithm H.1. Then if Assumptions 10.1 to 10.6 hold and $\mathsf{g}^\star > 0$, the expected regret of is bounded by*

$$\mathbb{E}^{M^\star}[\mathbf{Reg}(T)] \le (1+\delta)\mathsf{g}^\star \log T + C_{\mathsf{aec}} \cdot \left(\overline{\mathsf{aec}}^{\mathsf{D},\mathcal{M}}_{\varepsilon/2}(\mathcal{M}^\star)\right)^3 \cdot \log^{3/2} \log T + C_{\mathsf{low}} \cdot \log^{6/7} T$$

for $\varepsilon \leftarrow \frac{\delta}{4+2\delta}$,

$$C_{\mathsf{aec}} := \widetilde{O}\left(\frac{(V_{\mathcal{M}} + L_{\mathsf{KL}}) V^3_{\mathcal{M}} d_{\mathsf{cov}} \log(C_{\mathsf{cov}})}{\delta \Delta^\star_{\min}}\right),$$

and

$$C_{\mathsf{low}} := \widetilde{O}\left(\mathrm{poly}\left(V_{\mathcal{M}}, L_{\mathsf{KL}}, \mathsf{n}^\star_\varepsilon, d_{\mathsf{cov}}, \log C_{\mathsf{cov}}, \mathsf{g}^\star, \frac{1}{\Delta^\star_{\min}}, \frac{1}{\delta}, \log\log T\right)\right).$$

Proof of Theorem H.2. The bound on the regret incurred in the exploit phase follows identically to Theorem H.1, since the exploit test is the same. We turn to bounding the regret in

the explore phase. First, since we can incur regret of at most 1 at every round, we bound

$$\mathbb{E}\left[\sum_{s=1}^{s_T}\Delta^{M^\star}(p^s)\right] \le \mathbb{E}\left[\sum_{s=s^\star}^{s_T}\Delta^{M^\star}(p^s)\right] + s^\star$$

By Lemma H.3.6, we can bound

$$\begin{aligned}
\mathbb{E}\left[\sum_{s=s^\star}^{s_T}\Delta^\star(\pi^s)\right] &\le \mathbb{E}\Bigg[\sum_{s=s^\star}^{s_T}(1+\mathsf{g}^\star)s^{\alpha_q+\alpha_{\mathcal{M}}}\mathbb{E}_{\widehat{M}\sim\xi^s}[\mathbb{E}_{p^s}[D\big(\widehat{M}(\pi)\,\|\,M^\star(\pi)\big)]] \\
&\quad + \sum_{s=s^\star}^{s_T}(1+\delta)\mathsf{g}^\star\cdot\inf_{M\in\mathcal{M}^\star_{\mathrm{alt}}}\mathbb{E}_{p^s}[D_{\mathsf{KL}}(M^\star(\pi)\,\|\,M(\pi))]\cdot\mathbb{I}\{\pi_\star\in\boldsymbol{\pi}_{\widehat{M}^s}\} \\
&\quad + \sum_{s=s^\star}^{s_T}s^{\alpha_{\mathsf{n}}/2}\cdot 4\mathsf{g}^\star\mathbb{E}_{\widehat{M}\sim\xi^s}[\mathbb{E}_{p^s}[D_{\mathsf{KL}}\big(M^\star(\pi)\,\|\,\widehat{M}(\pi)\big)]]\Bigg] + 4\mathsf{g}^\star\big(\tfrac{8}{\delta\mathsf{g}^\star}\big)^{\frac{1}{\alpha_q-\alpha_{\mathsf{n}}}} \\
&\le \mathbb{E}\Bigg[\sum_{s=s^\star}^{s_T}(1+\mathsf{g}^\star)s^{\alpha_q+\alpha_{\mathcal{M}}}\mathbb{E}_{\widehat{M}\sim\xi^s}[\mathbb{E}_{p^s}[D\big(\widehat{M}(\pi)\,\|\,M^\star(\pi)\big)]] \\
&\quad + \sum_{s=1}^{s_T}(1+\delta)\mathsf{g}^\star\cdot\inf_{M\in\mathcal{M}^\star_{\mathrm{alt}}}\mathbb{E}_{p^s}[D_{\mathsf{KL}}(M^\star(\pi)\,\|\,M(\pi))]\cdot\mathbb{I}\{\pi_\star\in\boldsymbol{\pi}_{\widehat{M}^s}\} \\
&\quad + \sum_{s=s^\star}^{s_T}s^{\alpha_{\mathsf{n}}/2}\cdot 4\mathsf{g}^\star\mathbb{E}_{\widehat{M}\sim\xi^s}[\mathbb{E}_{p^s}[D_{\mathsf{KL}}\big(M^\star(\pi)\,\|\,\widehat{M}(\pi)\big)]]\Bigg] + 4\mathsf{g}^\star\big(\tfrac{8}{\delta\mathsf{g}^\star}\big)^{\frac{1}{\alpha_q-\alpha_{\mathsf{n}}}}.
\end{aligned}$$

Applying Lemma H.3.4 with $\alpha = 0$, we have

$$\begin{aligned}
\mathbb{E}\left[\sum_{s=1}^{s_T}\inf_{M\in\mathcal{M}^\star_{\mathrm{alt}}}\mathbb{E}_{p^s}[D_{\mathsf{KL}}(M^\star(\pi)\,\|\,M(\pi))]\cdot\mathbb{I}\{\pi_\star\in\boldsymbol{\pi}_{\widehat{M}^s}\}\right] &\le \log T + \mathbb{E}\Bigg[V_{\mathcal{M}}s_T^{1/2}\cdot\sqrt{1344d_{\mathsf{cov}}\cdot\log(128C_{\mathsf{cov}}s_T)} \\
&\quad + (V_{\mathcal{M}}+L_{\mathsf{KL}})\left(4s_T^{1/2}+\sum_{s=1}^{s_T}s^{1/2}\cdot\mathbb{E}_{\widehat{M}\sim\xi^s}[\mathbb{E}_{\pi\sim p^s}[D\big(\widehat{M}(\pi)\,\|\,M^\star(\pi)\big)]]\right)\Bigg] + \log\log T + 7V_{\mathcal{M}}.
\end{aligned}$$

Note that for $s \ge s^\star$, our estimator is applied to a cover of a set containing $M^\star$. Furthermore, note that the estimation procedure used in Algorithm H.3 is, other than the different choice of set to cover, identical to that used in Algorithm H.4. It follows that the analysis of Algorithm H.4 can be applied to the estimation procedure of Algorithm H.3, with only the mild modification accounting for the difference in the size of the cover (as we are covering $\mathcal{M}^\ell$ instead of $\mathcal{M}$). However, as we can upper bound the size of the cover of $\mathcal{M}^\ell$ by the size of the cover of $\mathcal{M}$ via Lemma H.3.15, this change is inconsequential. Thus, by Lemma H.3.7,

we can bound

$$\begin{aligned}
&\mathbb{E}\left[\sum_{s=s^\star}^{s_T} 2s^{\alpha_q+\alpha_{\mathcal{M}}}\mathbb{E}_{\widehat{M}\sim\xi^s}[\mathbb{E}_{p^s}[D\big(\widehat{M}(\pi)\,\|\,M^\star(\pi)\big)]\right] \\
&\le O\Big(\mathbb{E}\Big[V_{\mathcal{M}}d_{\mathsf{cov}}\log(C_{\mathsf{cov}}s_T)s_T^{\alpha_q+\alpha_{\mathcal{M}}}\sqrt{\log(s_T)}\Big]\Big) \\
&\le O\Big(V_{\mathcal{M}}d_{\mathsf{cov}}\log(C_{\mathsf{cov}}\mathbb{E}[s_T])\mathbb{E}[s_T]^{\alpha_q+\alpha_{\mathcal{M}}}\sqrt{\log(\mathbb{E}[s_T])}\Big), \qquad \text{(H.3.14)}
\end{aligned}$$

where the second inequality uses Jensen's inequality and Lemma H.3.16 to pass the expectation through, which holds as long as $1/100 \le \alpha_q + \alpha_{\mathcal{M}} \le 3/4$. Similarly,

$$\begin{aligned}
&\mathbb{E}\left[\sum_{s=s^\star}^{s_T} s^{\alpha_{\mathsf{n}}/2}\cdot 4\mathsf{g}^\star\mathbb{E}_{\widehat{M}\sim\xi^s}[\mathbb{E}_{p^s}[D_{\mathsf{KL}}\big(M^\star(\pi)\,\|\,\widehat{M}(\pi)\big)]]\right] \\
&\le O\Big(\mathsf{g}^\star V_{\mathcal{M}}d_{\mathsf{cov}}\log(C_{\mathsf{cov}}\mathbb{E}[s_T])\mathbb{E}[s_T]^{\alpha_{\mathsf{n}}/2}\sqrt{\log(\mathbb{E}[s_T])}\Big), \qquad \text{(H.3.15)}
\end{aligned}$$

where we have again used Jensen's inequality and Lemma H.3.16 to pass the expectation through, which holds as long $1/100 \le \alpha_{\mathsf{n}}/2 \le 3/4$. For $s \le s^\star$, we do not have $M^\star \in \mathcal{M}^\ell$, and therefore the estimation guarantees are vacuous. In this regime, using that the KL divergence is always bounded by $2V_{\mathcal{M}}$ (Lemma H.3.13), we can simply upper bound the estimation error by $2V_{\mathcal{M}}$. Thus,

$$\begin{aligned}
&\mathbb{E}\left[\sum_{s=1}^{s_T} s^{1/2}\cdot\mathbb{E}_{\widehat{M}\sim\xi^s}[\mathbb{E}_{\pi\sim p^s}[D\big(\widehat{M}(\pi)\,\|\,M^\star(\pi)\big)]]\right] \\
&\le \mathbb{E}\left[\sum_{s=s^\star}^{s_T} s^{1/2}\cdot\mathbb{E}_{\widehat{M}\sim\xi^s}[\mathbb{E}_{\pi\sim p^s}[D\big(\widehat{M}(\pi)\,\|\,M^\star(\pi)\big)]]\right] + 2V_{\mathcal{M}}(s^\star)^{3/2} \\
&\le O\Big(V_{\mathcal{M}}d_{\mathsf{cov}}\log(C_{\mathsf{cov}}\mathbb{E}[s_T])\mathbb{E}[s_T]^{1/2}\sqrt{\log(\mathbb{E}[s_T])}\Big) + 2V_{\mathcal{M}}(s^\star)^{3/2},
\end{aligned}$$

which gives, applying Lemma H.3.16 again:

$$\begin{aligned}
\mathbb{E}\left[\sum_{s=1}^{s_T}\inf_{M\in\mathcal{M}^\star_{\mathrm{alt}}}\mathbb{E}_{p^s}[D_{\mathsf{KL}}(M^\star(\pi)\,\|\,M(\pi))]\cdot\mathbb{I}\{\pi_\star\in\boldsymbol{\pi}_{\widehat{M}^s}\}\right] &\le \log T + O\Big(V_{\mathcal{M}}\mathbb{E}[s_T]^{1/2}\sqrt{d_{\mathsf{cov}}\cdot\log(C_{\mathsf{cov}}\mathbb{E}[s_T])} \\
&+ (V_{\mathcal{M}}+L_{\mathsf{KL}})\Big(\mathbb{E}[s_T]^{1/2} + V_{\mathcal{M}}d_{\mathsf{cov}}\log(C_{\mathsf{cov}}\mathbb{E}[s_T])\mathbb{E}[s_T]^{1/2}\sqrt{\log(\mathbb{E}[s_T])}\Big)\Big) \\
&+ 2V_{\mathcal{M}}(s^\star)^{3/2} + \log\log T + 7V_{\mathcal{M}}. \qquad \text{(H.3.16)}
\end{aligned}$$

Combining Eqs. (H.3.14) to (H.3.16), we have

$$\begin{aligned}\mathbb{E}\Big[\sum_{s=s^\star}^{s_T}\Delta^\star(\pi^s)\Big] \le &(1+\delta)\mathsf{g}^\star\log T + O\Big(V_{\mathcal{M}}\sqrt{d_{\mathsf{cov}}\mathbb{E}[s_T]\log(C_{\mathsf{cov}}\mathbb{E}[s_T])}\\ &+ (V_{\mathcal{M}}+L_{\mathsf{KL}})V_{\mathcal{M}}d_{\mathsf{cov}}\log(C_{\mathsf{cov}}\mathbb{E}[s_T])\mathbb{E}[s_T]^{1/2}\sqrt{\log(\mathbb{E}[s_T])}\\ &+ V_{\mathcal{M}}d_{\mathsf{cov}}\log(C_{\mathsf{cov}}\mathbb{E}[s_T])\sqrt{\log\mathbb{E}[s_T]}((1+\mathsf{g}^\star)\mathbb{E}[s_T]^{\alpha_q+\alpha_{\mathcal{M}}}+\varepsilon\mathsf{g}^\star\mathbb{E}[s_T]^{\alpha_{\mathsf{n}}/2})\\ &+ \mathsf{g}^\star(\tfrac{1}{\delta\mathsf{g}^\star})^{\frac{1}{\alpha_q-\alpha_{\mathsf{n}}}}+\log\log T + V_{\mathcal{M}}(s^\star)^{3/2}\Big).\end{aligned}$$

To control this, it remains to bound s_T. By Lemma H.3.6 we have the following almost sure bound:

$$\begin{aligned}s_T \le \frac{4}{\delta\mathsf{g}^\star}\Bigg(&\underbrace{\sum_{s=s^\star}^{s_T}(1+\mathsf{g}^\star)s^{\alpha_q+\alpha_{\mathcal{M}}}\mathbb{E}_{\widehat{M}\sim\xi^s}[\mathbb{E}_{p^s}[D\big(\widehat{M}(\pi)\,\|\,M^\star(\pi)\big)]]}_{(a)}\\ &+\underbrace{\sum_{s=s^\star}^{s_T}s^{\alpha_{\mathsf{n}}}\cdot 2\mathsf{g}^\star\cdot\inf_{M\in\mathcal{M}^\star_{\mathsf{alt}}}\mathbb{E}_{p^s}[D_{\mathsf{KL}}(M^\star(\pi)\,\|\,M(\pi))]\cdot\mathbb{I}\{\pi_\star\in\boldsymbol{\pi}_{\widehat{M}^s}\}}_{(b)}\\ &+\underbrace{\sum_{s=1}^{s_T}s^{3\alpha_{\mathsf{n}}/2}\cdot 4\mathsf{g}^\star\mathbb{E}_{\widehat{M}\sim\xi^s}[\mathbb{E}_{p^s}[D_{\mathsf{KL}}\big(M^\star(\pi)\,\|\,\widehat{M}(\pi)\big)]]}_{(c)}+4\mathsf{g}^\star(\tfrac{4}{\delta\mathsf{g}^\star})^{\frac{1+\alpha_{\mathsf{n}}}{\alpha_q-\alpha_{\mathsf{n}}}}\Bigg)+s^\star.\end{aligned}$$

We bound the expectation of term (a) as in (H.3.14). To bound term (b), we apply Lemma H.3.4 with $\alpha=\alpha_{\mathsf{n}}$ to get

$$\begin{aligned}\mathbb{E}[(b)] \le \mathbb{E}[s_T^{\alpha_{\mathsf{n}}}]\log T + \mathbb{E}\Bigg[&V_{\mathcal{M}}s_T^{1/2+\alpha_{\mathsf{n}}}\cdot\sqrt{1344d_{\mathsf{cov}}\cdot\log(128C_{\mathsf{cov}}s_T)}\\ &+(V_{\mathcal{M}}+L_{\mathsf{KL}})\Big(4\frac{s_T^{1/2+\alpha_{\mathsf{n}}/2}}{1-\alpha_{\mathsf{n}}}+\sum_{s=1}^{s_T}s^{1/2+\alpha_{\mathsf{n}}/2}\cdot\mathbb{E}_{\widehat{M}\sim\xi^s}[\mathbb{E}_{\pi\sim p^s}[D\big(\widehat{M}(\pi)\,\|\,M^\star(\pi)\big)]]\Big)\Bigg]\\ &+\mathbb{E}[s_T^{\alpha_{\mathsf{n}}}]\log\log T+7V_{\mathcal{M}}.\end{aligned}$$

Again applying Lemma H.3.7 and Lemma H.3.16, we have

$$\mathbb{E}[(c)] \le O\Big(\mathsf{g}^\star V_{\mathcal{M}}d_{\mathsf{cov}}\log(C_{\mathsf{cov}}\mathbb{E}[s_T])\mathbb{E}[s_T]^{3\alpha_{\mathsf{n}}/2}\sqrt{\log(\mathbb{E}[s_T])}\Big)$$

and

$$\mathbb{E}\left[\sum_{s=1}^{s_T} s^{1/2+\alpha_{\mathsf{n}}/2} \cdot \mathbb{E}_{\widehat{M}\sim\xi^s}[\mathbb{E}_{\pi\sim p^s}[D\big(\widehat{M}(\pi) \,\|\, M^\star(\pi)\big)]]\right]$$
$$\le O\Big(V_{\mathcal{M}} d_{\mathsf{cov}} \log(C_{\mathsf{cov}}\mathbb{E}[s_T])\mathbb{E}[s_T]^{1/2+\alpha_{\mathsf{n}}/2}\sqrt{\log(\mathbb{E}[s_T])}\Big) + 2V_{\mathcal{M}}(s^\star)^{3/2+\alpha_{\mathsf{n}}/2},$$

as long as $1/100 \le \alpha_{\mathsf{n}} \le 1/4$. We therefore have (using Lemma H.3.16 to pass the expectation through):

$$\begin{aligned}\mathbb{E}[s_T] \le \frac{1}{\delta \mathsf{g}^\star} \cdot O\Bigg(&\mathbb{E}[s_T]^{\alpha_{\mathsf{n}}}\log T + V_{\mathcal{M}}\mathbb{E}[s_T]^{1/2+\alpha_{\mathsf{n}}}\sqrt{d_{\mathsf{cov}}\log(C_{\mathsf{cov}}\mathbb{E}[s_T])} + (V_{\mathcal{M}}+L_{\mathsf{KL}})\mathbb{E}[s_T]^{1/2+\alpha_{\mathsf{n}}/2}\\
&+ (V_{\mathcal{M}}+L_{\mathsf{KL}})V_{\mathcal{M}} d_{\mathsf{cov}}\log(C_{\mathsf{cov}}\mathbb{E}[s_T])\mathbb{E}[s_T]^{1/2+\alpha_{\mathsf{n}}/2}\sqrt{\log(\mathbb{E}[s_T])}\\
&+ V_{\mathcal{M}} d_{\mathsf{cov}}\log(C_{\mathsf{cov}}\mathbb{E}[s_T])(1+\mathsf{g}^\star)\mathbb{E}[s_T]^{\alpha_q+\alpha_{\mathcal{M}}}\sqrt{\log(\mathbb{E}[s_T])}\\
&+ \mathsf{g}^\star V_{\mathcal{M}} d_{\mathsf{cov}}\log(C_{\mathsf{cov}}\mathbb{E}[s_T])\mathbb{E}[s_T]^{3\alpha_{\mathsf{n}}/2}\sqrt{\log(\mathbb{E}[s_T])} + \mathsf{g}^\star(\tfrac{1}{\delta\mathsf{g}^\star})^{\frac{1+\alpha_{\mathsf{n}}}{\alpha_q-\alpha_{\mathsf{n}}}} + V_{\mathcal{M}}(s^\star)^{3/2+\alpha_{\mathsf{n}}/2}\Bigg).\end{aligned}$$

We now set $\alpha_{\mathcal{M}} = 1/2$, $\alpha_{\mathsf{n}} = 1/8$, and $\alpha_q = 1/4$, and note that all of the preceding parameter restrictions are satisfied for these choices. Furthermore, this parameter choice implies that—using Lemma H.3.17 to handle log factors—we have

$$\mathbb{E}[s_T] \le \widetilde{O}\left(\frac{1}{(\delta\mathsf{g}^\star)^{8/7}}\log^{8/7} T + \mathrm{poly}\left(V_{\mathcal{M}}, d_{\mathsf{cov}}, \log C_{\mathsf{cov}}, L_{\mathsf{KL}}, \mathsf{g}^\star, \frac{1}{\delta}, \frac{1}{\mathsf{g}^\star}\right) + \frac{V_{\mathcal{M}}(s^\star)^{3/2+\alpha_{\mathsf{n}}/2}}{\delta\mathsf{g}^\star}\right).$$

Plugging this into the regret bound given above, we have

$$\begin{aligned}\mathbb{E}\left[\sum_{s=s^\star}^{s_T}\Delta^\star(\pi^s)\right] \le (1+\delta)\mathsf{g}^\star\log T + \widetilde{O}\Bigg(&\mathrm{poly}\left(V_{\mathcal{M}}, d_{\mathsf{cov}}, \log C_{\mathsf{cov}}, L_{\mathsf{KL}}, \mathsf{g}^\star, \frac{1}{\delta}, \frac{1}{\mathsf{g}^\star}, \log\log T\right)\cdot\log^{6/7} T\\
&+ \frac{(1+1/\mathsf{g}^\star)(V_{\mathcal{M}}+L_{\mathsf{KL}})V_{\mathcal{M}}^2 d_{\mathsf{cov}}\log(C_{\mathsf{cov}}) + V_{\mathcal{M}}}{\delta}\cdot(s^\star)^{3/2}\cdot\log^{3/2}\log T\Bigg).\end{aligned}$$

Finally, by Lemma H.3.5, we can bound $s^\star$ as

$$s^\star \le \left(\overline{\mathsf{aec}}^{\mathsf{D},\mathcal{M}}_{\varepsilon/2}(\mathcal{M}^\star)\right)^{\frac{1}{\alpha_{\mathcal{M}}}} + \left(\mathsf{n}^\star_\varepsilon + \frac{1}{\Delta^\star_{\min}} + \frac{4\mathsf{g}^\star}{\Delta^\star_{\min}} + \frac{2\mathsf{n}^\star_\varepsilon}{\mathsf{g}^\star} + \frac{4}{\Delta^\star_{\min}\varepsilon}\right)^{\frac{2}{\alpha_{\mathsf{n}}}},$$

and, by Lemma H.2.3 and Lemma H.3.13, we can lower bound $\mathsf{g}^\star \ge \Delta^\star_{\min}/2V_{\mathcal{M}}$. Plugging this in gives the final bound. □

Algorithm H.4 Estimation with Adaptive Covering

input: Class $\mathcal{M}$.
$\ell \leftarrow 1$, $\mathfrak{D}^\ell \leftarrow \varnothing$.
$\mathcal{M}^\ell_{\mathrm{cov}} \leftarrow (\rho_\ell, \mu_\ell)$-cover of $\mathcal{M}$ for $\rho_\ell \leftarrow 2^{-\ell}, \mu_\ell \leftarrow 2^{-5\ell}$.
Initialize $\mathsf{TemperedAggregation}^\ell$ as an instance of Algorithm H.1 with $\mathcal{M}^\ell_{\mathrm{cov}}$.
for $s = 1, 2, 3, \ldots$ **do**
  Receive (π^s, r^s, o^s), $\mathfrak{D}^\ell \leftarrow \mathfrak{D}^\ell \cup \{(\pi^s, r^s, o^s)\}$.
  $\xi^s \leftarrow \mathsf{TemperedAggregation}^\ell(\mathfrak{D}^\ell)$, $\widehat{M}^s \leftarrow \mathbb{E}_{M\sim\xi^s}[M]$
  if $s \geq 2^\ell$ **then**.
    $\ell \leftarrow \ell + 1$, $\mathfrak{D}^\ell \leftarrow \varnothing$.
    $\mathcal{M}^\ell_{\mathrm{cov}} \leftarrow (\rho_\ell, \mu_\ell)$-cover of $\mathcal{M}$ for $\rho_\ell \leftarrow 2^{-\ell}, \mu_\ell \leftarrow 2^{-5\ell}$.
    Initialize $\mathsf{TemperedAggregation}^\ell$ as an instance of Algorithm H.1 with $\mathcal{M}^\ell_{\mathrm{cov}}$.

H.3.3 Estimation Guarantees

In this section, we analyze Algorithm H.4, which is a variant of the Tempered Aggregation algorithm (Algorithm H.1) designed for infinite classes. This algorithm is used within Algorithm H.2 in order to prove Theorem H.1.

Algorithm H.4 simply applies Algorithm H.1 to a sequence of covers for the class $\mathcal{M}$ on a doubling epoch schedule. In particular, at every epoch ℓ, Algorithm H.4 restarts the Tempered Aggregation algorithm (Algorithm H.1), clearing the Tempered Aggregation instance from the previous epoch from memory. We denote the ℓth instantiation of Tempered Aggregation as $\mathsf{TemperedAggregation}^\ell$.

Lemma H.3.7. *Let τ denote some stopping time with respect to the filtration $(\mathscr{F}^t)_{t=1}^T$ such that $\tau \leq T$ almost surely, and let $\alpha \in (0,1)$. When running Algorithm H.4 under Assumption 10.6, we have*

$$\mathbb{E}\left[\sum_{s=1}^{\tau} s^\alpha \cdot \mathbb{E}_{M\sim\xi^s}[\mathbb{E}_{\pi\sim p^s}[D^2_{\mathsf{H}}(M^\star(\pi), M(\pi))]]\right] \leq \mathbb{E}\left[(20 d_{\mathsf{cov}} \cdot \log(64 C_{\mathsf{cov}}\tau) + 1) \cdot \frac{2^{2\alpha}}{2^\alpha - 1}\tau^\alpha\right] + 4.$$

In addition, if Assumption 10.5 also holds, then

$$\begin{aligned}
&\mathbb{E}\left[\sum_{s=1}^{\tau} s^\alpha \cdot \mathbb{E}_{M\sim\xi^s}[\mathbb{E}_{\pi\sim p^s}[D_{\mathsf{KL}}(M^\star(\pi) \,\|\, M(\pi))]]\right] \\
&\leq \mathbb{E}\left[(2 + 6V_{\mathcal{M}})(20 d_{\mathsf{cov}} \cdot \log(64 C_{\mathsf{cov}}\tau) + 1) \cdot \frac{2^{2\alpha}}{2^\alpha - 1}\tau^\alpha \sqrt{\log(2\tau)}\right] \\
&\quad + \mathbb{E}[32(1 + V_{\mathcal{M}})\log(\tau)] + 8
\end{aligned} \tag{H.3.17}$$

and

$$\mathbb{E}\left[\sum_{s=1}^{\tau} s^{\alpha} \cdot \mathbb{E}_{M\sim\xi^s}[\mathbb{E}_{\pi\sim p^s}[D_{\mathsf{KL}}(M(\pi) \,\|\, M^\star(\pi))]]\right]$$
$$\leq \mathbb{E}\left[(2+6V_{\mathcal{M}})(20d_{\mathsf{cov}} \cdot \log(64C_{\mathsf{cov}}\tau)+1) \cdot \frac{2^{2\alpha}}{2^\alpha-1}\tau^\alpha\sqrt{\log(2\tau)}\right]$$
$$+ \mathbb{E}[32(1+V_{\mathcal{M}})\log(\tau)] + 8. \tag{H.3.18}$$

Proof of Lemma H.3.7. Let $\mathcal{S}^k$ denote the set of s values for which $\ell = k$ and note that $\mathcal{S}^k = \{2^{k-1}+1, 2^{k-1}+2, \ldots, 2^k\}$. We can bound

$$\mathbb{E}\left[\sum_{s=1}^{\tau} s^{\alpha}\mathbb{E}_{M\sim\xi^s}[\mathbb{E}_{\pi\sim p^s}[D^2_{\mathsf{H}}(M^\star(\pi), M(\pi))]]\right]$$
$$\leq \sum_{k=1}^{\lceil\log_2 T\rceil} \mathbb{E}\left[\sum_{s\in\mathcal{S}^k} s^{\alpha} \cdot \mathbb{E}_{M\sim\xi^s}[\mathbb{E}_{\pi\sim p^s}[D^2_{\mathsf{H}}(M^\star(\pi), M(\pi))]] \cdot \mathbb{I}\{s\leq\tau\}\right]$$
$$\leq \sum_{k=1}^{\lceil\log_2 T\rceil} 2^{\alpha k} \cdot \mathbb{E}\left[\sum_{s\in\mathcal{S}^k} \mathbb{E}_{M\sim\xi^s}[\mathbb{E}_{\pi\sim p^s}[D^2_{\mathsf{H}}(M^\star(\pi), M(\pi))]] \cdot \mathbb{I}\{s\leq\tau\}\right] \tag{H.3.19}$$

since we have that $\ell \leq \lceil\log_2 T\rceil$ by construction, and since $s \leq 2^k$ for $s \in \mathcal{S}^k$. Let A_k denote the event

$$A_k := \left\{\forall s\in\mathcal{S}^k \;:\; \sum_{i=2^{k-1}}^{s} \mathbb{E}_{M\sim\xi^i}[\mathbb{E}_{\pi\sim p^i}[D^2_{\mathsf{H}}(M^\star(\pi), M(\pi))]] \leq 2\log\frac{2^k\mathsf{N}_{\mathrm{cov}}(\mathcal{M},\rho_k,\mu_k)}{\delta_k} + 2^k\rho_k\right\}.$$

By Proposition H.2 and a union bound, $\mathbb{P}[A_k] \geq 1-\delta_k-2^{2k}\mu_k$. Since the Hellinger distance is always bounded by 2 and $|\mathcal{S}^k| \leq 2^k$, we can upper bound

$$\text{(H.3.19)} \leq \sum_{k=1}^{\lceil\log_2 T\rceil} 2^{\alpha k}\left(\sum_{s\in\mathcal{S}^k}\mathbb{E}\left[\mathbb{E}_{M\sim\xi^s}[\mathbb{E}_{\pi\sim p^s}[D^2_{\mathsf{H}}(M^\star(\pi), M(\pi))]] \cdot \mathbb{I}\{s\leq\tau, A_k\}\right] + 2\cdot 2^k\mathbb{E}[\mathbb{I}\{A_k^c\}]\right).$$

Choosing $\delta_k = 2^{-3k}$ and since $\mu_k = 2^{-5k}$, we have

$$\sum_{k=1}^{\lceil\log_2 T\rceil} 2^{\alpha k}\cdot 2^{k+1}\mathbb{E}[\mathbb{I}\{A_k^c\}] \leq \sum_{k=1}^{\lceil\log_2 T\rceil} 2^{2k+1}\cdot(\delta_k + 2^{2k}\mu_k) = \sum_{k=1}^{\lceil\log_2 T\rceil} 2^{2k+1}\cdot 2\cdot 2^{-3k} \leq 4.$$

Note that for $s \in \mathcal{S}^k$, if $s \leq \tau$, this implies that $2^{k-1} \leq \tau$. On the event A_k, for $\rho_k = 2^{-k}$,

when $\alpha > 0$ we can bound

$$\begin{aligned}
&\sum_{k=1}^{\lceil \log_2 T \rceil} 2^{\alpha k} \sum_{s \in \mathcal{S}^k} \mathbb{E}\big[\mathbb{E}_{M \sim \xi^s}[\mathbb{E}_{\pi \sim p^s}[D^2_{\mathsf{H}}(M^\star(\pi), M(\pi))]] \cdot \mathbb{I}\{s \le \tau, A_k\}\big] \\
&\quad \le \mathbb{E}\left[\sum_{k=1}^{\lceil \log_2 T \rceil} 2^{\alpha k}\left(2 \log \frac{2^k \mathsf{N}_{\mathrm{cov}}(\mathcal{M}, 2^{-k}, 2^{-5k})}{2^{-3k}} + 1\right) \cdot \mathbb{I}\{2^{k-1} \le \tau\}\right] \\
&\quad \le \mathbb{E}\left[\left(2 \log \frac{\mathsf{N}_{\mathrm{cov}}(\mathcal{M}, \tau^{-1}/2, \tau^{-5}/32)}{\tau^{-4}/16} + 1\right) \cdot \max_k \left(\frac{2^\alpha}{2^\alpha - 1} 2^{\alpha k} \cdot \mathbb{I}\{2^{k-1} \le \tau\}\right)\right] \\
&\quad \le \mathbb{E}\left[\left(2 \log \frac{\mathsf{N}_{\mathrm{cov}}(\mathcal{M}, \tau^{-1}/2, \tau^{-5}/32)}{\tau^{-4}/16} + 1\right) \cdot \frac{2^{2\alpha}}{2^\alpha - 1} \tau^\alpha\right]
\end{aligned} \tag{H.3.20}$$

where the final two inequalities follow since $2^k \le 2\tau$. Under Assumption 10.6 we have

$$\log \frac{\mathsf{N}_{\mathrm{cov}}(\mathcal{M}, \tau^{-1}/2, \tau^{-5}/32)}{\tau^{-4}/16} \le 10 d_{\mathsf{cov}} \cdot \log(64 C_{\mathsf{cov}} \tau),$$

which gives the first result.

Bound on KL Estimation Error. By Lemma H.3.14, for any $x > 0$ we have

$$D_{\mathsf{KL}}\big(M(\pi) \,\|\, \widetilde{M}(\pi)\big) \le (2 + 2V_{\mathcal{M}} + x) \cdot D^2_{\mathsf{H}}\big(M(\pi), \widetilde{M}(\pi)\big) + 32(1 + V^2_{\mathcal{M}}/x + V^3_{\mathcal{M}}/x^2) \cdot \exp(-x^2/8V^2_{\mathcal{M}}).$$

In particular choosing $x = V_{\mathcal{M}} \sqrt{8 \log s^2}$, we have

$$D_{\mathsf{KL}}\big(M(\pi) \,\|\, \widetilde{M}(\pi)\big) \le (2 + 2V_{\mathcal{M}} + V_{\mathcal{M}} \sqrt{8 \log s}) \cdot D^2_{\mathsf{H}}\big(M(\pi), \widetilde{M}(\pi)\big) + 32(1 + V_{\mathcal{M}})/s.$$

Repeating this for each step s, we can therefore bound

$$\begin{aligned}
&\mathbb{E}\left[\sum_{s=1}^{\tau} s^\alpha \mathbb{E}_{M \sim \xi^s}[\mathbb{E}_{\pi \sim p^s}[D_{\mathsf{KL}}(M^\star(\pi) \,\|\, M(\pi))]]\right] \\
&\le (2 + 6V_{\mathcal{M}}) \cdot \mathbb{E}\left[\sum_{s=1}^{\tau} s^\alpha \sqrt{\log s} \mathbb{E}_{M \sim \xi^s}[\mathbb{E}_{\pi \sim p^s}[D^2_{\mathsf{H}}(M^\star(\pi), M(\pi))]]\right] \\
&\qquad + 32(1 + V_{\mathcal{M}}) \cdot \mathbb{E}[\log \tau].
\end{aligned}$$

The result in (H.3.17) then follows from a calculation nearly identical to our above bound on Hellinger estimation error. Applying Lemma H.3.14 in a similar fashion with the arguments flipped gives (H.3.18). □

In the following, we extend Lemma H.3.7 to the case when $\alpha = 0$.

Lemma H.3.8. *Let τ denote some stopping time with respect to the filtration $(\mathscr{F}^t)_{t=1}^T$ such that $\tau \le T$ almost surely. When running Algorithm H.4, under Assumption 10.6, we have*

$$\mathbb{E}\left[\sum_{s=1}^{\tau} \mathbb{E}_{M\sim\xi^s}[\mathbb{E}_{\pi\sim p^s}[D^2_{\mathsf{H}}(M^\star(\pi), M(\pi))]]\right] \le \mathbb{E}[(20 d_{\mathsf{cov}} \cdot \log(64 C_{\mathsf{cov}}\tau) + 1)\cdot(2\log\tau + 1)] + 4.$$

In addition, if Assumption 10.5 also holds,

$$\begin{aligned}
&\mathbb{E}\left[\sum_{s=1}^{\tau} \mathbb{E}_{M\sim\xi^s}[\mathbb{E}_{\pi\sim p^s}[D_{\mathsf{KL}}(M^\star(\pi)\,\|\,M(\pi))]]\right] \\
&\le \mathbb{E}\Big[(2+5V_{\mathcal{M}})(20 d_{\mathsf{cov}} \cdot \log(64 C_{\mathsf{cov}}\tau) + 1)(2\log\tau+1)\cdot\sqrt{\log(2\tau)}\Big] \\
&\quad + \mathbb{E}[32(1+V_{\mathcal{M}})\log(\tau)] + 8
\end{aligned}$$

and

$$\begin{aligned}
&\mathbb{E}\left[\sum_{s=1}^{\tau} \mathbb{E}_{M\sim\xi^s}[\mathbb{E}_{\pi\sim p^s}[D_{\mathsf{KL}}(M(\pi)\,\|\,M^\star(\pi))]]\right] \\
&\le \mathbb{E}\Big[(2+5V_{\mathcal{M}})(20 d_{\mathsf{cov}} \cdot \log(64 C_{\mathsf{cov}}\tau) + 1)(2\log\tau+1)\cdot\sqrt{\log(2\tau)}\Big] \\
&\quad + \mathbb{E}[32(1+V_{\mathcal{M}})\log(\tau)] + 8.
\end{aligned}$$

Proof of Lemma H.3.8. This follows identically to Lemma H.3.7 but replacing (H.3.20) with

$$\begin{aligned}
&\sum_{k=1}^{\lceil\log_2 T\rceil} 2^{\alpha k} \sum_{s\in\mathcal{S}^k} \mathbb{E}\big[\mathbb{E}_{M\sim\xi^s}[\mathbb{E}_{\pi\sim p^s}[D^2_{\mathsf{H}}(M^\star(\pi), M(\pi))]]\cdot\mathbb{I}\{s\le\tau, A_k\}\big] \\
&\quad\le \mathbb{E}\left[\sum_{k=1}^{\lceil\log_2 T\rceil}\left(2\log\frac{2^k \mathsf{N}_{\mathrm{cov}}(\mathcal{M}, 2^{-k}, 2^{-5k})}{2^{-3k}} + 1\right)\cdot\mathbb{I}\{2^{k-1}\le\tau\}\right] \\
&\quad\le \mathbb{E}\left[\left(2\log\frac{\mathsf{N}_{\mathrm{cov}}(\mathcal{M}, \tau^{-1}/2, \tau^{-5}/32)}{\tau^{-4}/16} + 1\right)\cdot(2\log\tau+1)\right].
\end{aligned}$$

The bound on the KL estimation error also follows from the same reasoning as in Lemma H.3.7.

□

H.3.4 Supporting Lemmas

Lemma H.3.9. *Consider running either Algorithm H.2 or Algorithm H.3. Assume that $\pi_\star \notin \boldsymbol{\pi}_{\widehat{M}^s}$ and that $\min_{M\in\mathcal{M}\,:\,\xi^s(M)>0}\Delta^M_{\min} \ge \Delta$. Then $\mathbb{E}_{M\sim\xi^s}[\mathbb{I}\{M\in\mathcal{M}^\star_{\mathrm{alt}}\}] \ge \frac{1}{2}\Delta$.*

Proof of Lemma H.3.9. Recall that $\widehat{M}^s = \mathbb{E}_{M\sim\xi^s}[M]$, so $\pi \in \boldsymbol{\pi}_{\widehat{M}^s}$ implies that

$$\pi \in \arg\max_{\pi'\in\Pi} \mathbb{E}_{M\sim\xi^s}[f^M(\pi')].$$

If $\pi_\star \notin \boldsymbol{\pi}_{\widehat{M}^s}$, then there exists some $\widetilde{\pi}$ such that $\mathbb{E}_{M\sim\xi^s}[f^M(\widetilde{\pi})] > \mathbb{E}_{M\sim\xi^s}[f^M(\pi_\star)]$. Since $f^M(\pi) \in [0,1]$, this implies that

$$0 < \mathbb{E}_{M\sim\xi^s}[f^M(\widetilde{\pi}) - f^M(\pi_\star)] \le \mathbb{E}_{M\sim\xi^s}[\mathbb{I}\{M \in \mathcal{M}^\star_{\mathrm{alt}}\}] - \mathbb{E}_{M\sim\xi^s}[(f^M(\pi_\star) - f^M(\widetilde{\pi}))\cdot \mathbb{I}\{M \notin \mathcal{M}^\star_{\mathrm{alt}}\}].$$

For $M \notin \mathcal{M}^\star_{\mathrm{alt}}$, we have $f^M(\pi_\star) - f^M(\widetilde{\pi}) \ge \Delta^M_{\min} \ge \Delta$. Thus, the above implies

$$\begin{aligned} &0 < \mathbb{E}_{M\sim\xi^s}[\mathbb{I}\{M \in \mathcal{M}^\star_{\mathrm{alt}}\}] - \Delta\cdot\mathbb{E}_{M\sim\xi^s}[\mathbb{I}\{M \notin \mathcal{M}^\star_{\mathrm{alt}}\}] \\ \iff &\Delta\cdot(1 - \mathbb{E}_{M\sim\xi^s}[\mathbb{I}\{M \in \mathcal{M}^\star_{\mathrm{alt}}\}]) < \mathbb{E}_{M\sim\xi^s}[\mathbb{I}\{M \in \mathcal{M}^\star_{\mathrm{alt}}\}]. \end{aligned}$$

Rearranging gives $\mathbb{E}_{M\sim\xi^s}[\mathbb{I}\{M \in \mathcal{M}^\star_{\mathrm{alt}}\}] \ge \frac{\Delta}{1+\Delta} \ge \frac{1}{2}\Delta$. □

Lemma H.3.10. *When running Algorithm H.2, on rounds s for which $M^\star \in \mathcal{M}\backslash\mathcal{M}^{\mathrm{gl}}_\varepsilon(\lambda^s; \mathsf{n}_{\max})$, we have*

$$\gamma^s \le \frac{(1+\delta)(4\mathsf{n}_{\max} + 2\delta\underline{\mathsf{g}}^{\mathcal{M}})}{\delta\underline{\mathsf{g}}^{\mathcal{M}}} \cdot \overline{\mathsf{aec}}^{\mathsf{D}}_{\varepsilon/2}(\mathcal{M}),$$

for γ^s as defined in (H.3.3).

Proof of Lemma H.3.10. Recall that ω^s and λ^s are chosen to minimize

$$\sup_{M\in\mathcal{M}\backslash\mathcal{M}^{\mathrm{gl}}_\varepsilon(\lambda^s;\mathsf{n}_{\max})} \frac{1}{\mathbb{E}_{\widehat{M}\sim\xi^s}[\mathbb{E}_\omega[D\big(\widehat{M}(\pi)\,\|\,M(\pi)\big)]]}.$$

Since $M^\star \in \mathcal{M}\backslash\mathcal{M}^{\mathrm{gl}}_\varepsilon(\lambda^s; \mathsf{n}_{\max})$, we can therefore bound

$$\begin{aligned} \frac{1}{\mathbb{E}_{\widehat{M}\sim\xi^s}[\mathbb{E}_{\omega^s}[D\big(\widehat{M}(\pi)\,\|\,M^\star(\pi)\big)]]} &\le \inf_{\omega\in\triangle_\Pi} \sup_{M\in\mathcal{M}\backslash\mathcal{M}^{\mathrm{gl}}_\varepsilon(\lambda^s;\mathsf{n}_{\max})} \frac{1}{\mathbb{E}_{\widehat{M}\sim\xi^s}[\mathbb{E}_\omega[D\big(\widehat{M}(\pi)\,\|\,M(\pi)\big)]]} \\ &= \inf_{\lambda,\omega\in\triangle_\Pi} \sup_{M\in\mathcal{M}\backslash\mathcal{M}^{\mathrm{gl}}_\varepsilon(\lambda;\mathsf{n}_{\max})} \frac{1}{\mathbb{E}_{\widehat{M}\sim\xi^s}[\mathbb{E}_\omega[D\big(\widehat{M}(\pi)\,\|\,M(\pi)\big)]]}. \end{aligned}$$

Recall that we set

$$\mathsf{n}_{\max} = \left(\frac{1}{\Delta_{\min}\varepsilon} + \frac{2V_{\mathcal{M}}\mathsf{n}^{\mathcal{M}}_\varepsilon}{\Delta_{\min}}\right)\cdot\max_{M\in\mathcal{M}}\frac{32\mathsf{g}^M}{\Delta_{\min}}.$$

By Lemma H.3.13, under Assumption 10.5, we can bound $D_{\mathsf{KL}}(M(\pi)\,\|\,M'(\pi)) \le 2V_{\mathcal{M}}$ for all

$M, M' \in \mathcal{M}$ and $\pi \in \Pi$. It follows from Lemma H.2.3 that

$$\frac{2V_{\mathcal{M}}}{\Delta_{\min}} \geq \frac{1}{\min_{M \in \mathcal{M}: \mathsf{g}^M > 0} \mathsf{g}^M},$$

so

$$\mathsf{n}_{\max} \geq \left(\frac{1}{\Delta_{\min} \varepsilon} + \frac{\mathsf{n}_{\varepsilon}^{\mathcal{M}}}{\min_{M \in \mathcal{M}: \mathsf{g}^M > 0} \mathsf{g}^M} \right) \cdot \max_{M \in \mathcal{M}} \frac{32 \mathsf{g}^M}{\Delta_{\min}}.$$

Given this, straightforward calculation shows that $\mathsf{n}_{\max}$ meets the condition required by Lemma H.2.4, so Lemma H.2.4 implies

$$\begin{aligned}
\inf_{\lambda, \omega \in \triangle_\Pi} \sup_{M \in \mathcal{M} \setminus \mathcal{M}_\varepsilon^{\mathrm{gl}}(\lambda; \mathsf{n}_{\max})} \frac{1}{\mathbb{E}_{\widehat{M} \sim \xi^s}[\mathbb{E}_\omega[D\big(\widehat{M}(\pi) \,\|\, M(\pi)\big)]} &\leq \inf_{\lambda, \omega \in \triangle_\Pi} \sup_{M \in \mathcal{M} \setminus \mathcal{M}_{\varepsilon/2}^{\mathrm{gl}}(\lambda)} \frac{1}{\mathbb{E}_{\widehat{M} \sim \xi^s}[\mathbb{E}_\omega[D\big(\widehat{M}(\pi) \,\|\, M(\pi)\big)]} \\
&= \overline{\mathsf{aec}}_{\varepsilon/2}^{\mathsf{D}}(\mathcal{M}; \xi^s) \\
&\leq \overline{\mathsf{aec}}_{\varepsilon/2}^{\mathsf{D}}(\mathcal{M}).
\end{aligned}$$

By our choice of q we have $\frac{1}{1-q} = \frac{4\mathsf{n}_{\max} + 2\delta \underline{\mathsf{g}}^{\mathcal{M}}}{\delta \underline{\mathsf{g}}^{\mathcal{M}}}$. We can then bound

$$\gamma^s = \frac{1+\delta}{1-q} \cdot \frac{1}{\mathbb{E}_{\widehat{M} \sim \xi^s}[\mathbb{E}_{\omega^s}[D\big(\widehat{M}(\pi) \,\|\, M^\star(\pi)\big)]} \leq \frac{(1+\delta)(4\mathsf{n}_{\max} + 2\delta \underline{\mathsf{g}}^{\mathcal{M}})}{\delta \underline{\mathsf{g}}^{\mathcal{M}}} \cdot \overline{\mathsf{aec}}_{\varepsilon/2}^{\mathsf{D}}(\mathcal{M}).$$

□

Lemma H.3.11. *Consider running Algorithm H.3. Then on rounds s for which $M^\star \in \mathcal{M}^\ell \setminus \mathcal{M}_\varepsilon^{\mathrm{gl}}(\lambda^s; \mathsf{n}^s)$, we have*

$$\gamma^s \leq (1 + \mathsf{g}^\star \delta) \cdot s^{\alpha_q + \alpha_{\mathcal{M}}}$$

for γ^s as defined in (H.3.10), and $\alpha_q, \alpha_{\mathcal{M}}$ parameters of Algorithm H.3.

Proof of Lemma H.3.11. Assume we are at epoch ℓ. Recall that ω^s and λ^s minimize

$$\sup_{M \in \mathcal{M}^\ell \setminus \mathcal{M}_\varepsilon^{\mathrm{gl}}(\lambda^s; \mathsf{n}^s)} \frac{1}{\mathbb{E}_{\widehat{M} \sim \xi^s}[\mathbb{E}_\omega[D\big(\widehat{M}(\pi) \,\|\, M(\pi)\big)]}.$$

Since $M^\star \in \mathcal{M}^\ell \setminus \mathcal{M}_\varepsilon^{\mathrm{gl}}(\lambda^s; \mathsf{n}^s)$, we have can therefore bound

$$\begin{aligned}
\frac{1}{\mathbb{E}_{\widehat{M} \sim \xi^s}[\mathbb{E}_{\omega^s}[D\big(\widehat{M}(\pi) \,\|\, M^\star(\pi)\big)]]} &\leq \inf_{\omega \in \triangle_\Pi} \sup_{M \in \mathcal{M}^\ell \setminus \mathcal{M}_\varepsilon^{\mathrm{gl}}(\lambda^s; \mathsf{n}^s)} \frac{1}{\mathbb{E}_{\widehat{M} \sim \xi^s}[\mathbb{E}_\omega[D\big(\widehat{M}(\pi) \,\|\, M(\pi)\big)]]} \\
&= \inf_{\lambda, \omega \in \triangle_\Pi} \sup_{M \in \mathcal{M}^\ell \setminus \mathcal{M}_\varepsilon^{\mathrm{gl}}(\lambda; \mathsf{n}^s)} \frac{1}{\mathbb{E}_{\widehat{M} \sim \xi^s}[\mathbb{E}_\omega[D\big(\widehat{M}(\pi) \,\|\, M(\pi)\big)]}.
\end{aligned}$$

By construction, for every $M \in \mathcal{M}^\ell$, we have

$$\mathsf{n}_\varepsilon^M + \frac{1}{\Delta_{\min}^M} + \frac{4\mathsf{g}^M}{\Delta_{\min}^M} + \frac{2\mathsf{n}_\varepsilon^M}{\mathsf{g}^M} + \frac{4}{\Delta_{\min}^M \varepsilon} \le \sqrt{\mathsf{n}^s}.$$

This implies that

$$\zeta := \min_{M \in \mathcal{M}^\ell} \min\left\{ \frac{\mathsf{g}^M}{\mathsf{g}^M + 2\mathsf{n}_\varepsilon^M}, \frac{\Delta_{\min}^M \varepsilon}{4} \right\} \ge \frac{1}{\sqrt{\mathsf{n}^s}}$$

and

$$\max_{M \in \mathcal{M}^\ell} \max\left\{ \mathsf{n}_\varepsilon^M, \frac{4\mathsf{g}^M}{\Delta_{\min}^M} \right\} \le \sqrt{\mathsf{n}^s},$$

which together imply that

$$\max_{M \in \mathcal{M}^\ell} \max\left\{ \mathsf{n}_\varepsilon^M, \frac{4\mathsf{g}^M}{\Delta_{\min}^M}, \frac{2\mathsf{g}^M}{\zeta \Delta_{\min}^M} \right\} \le \mathsf{n}^s$$

By Lemma H.2.4 and since $\mathcal{M}^\ell$ is constructed such that $\inf_{M \in \mathcal{M}^\ell} \Delta_{\min}^M > 0$, we can therefore bound

$$\begin{aligned}\inf_{\lambda, \omega \in \triangle_\Pi} \sup_{M \in \mathcal{M}^\ell \setminus \mathcal{M}_\varepsilon^{\mathrm{gl}}(\lambda; \mathsf{n}^s)} \frac{1}{\mathbb{E}_{\widehat{M} \sim \xi^s}[\mathbb{E}_\omega[D\big(\widehat{M}(\pi) \,\|\, M(\pi)\big)]} &\le \inf_{\lambda, \omega \in \triangle_\Pi} \sup_{M \in \mathcal{M}^\ell \setminus \mathcal{M}_{\varepsilon/2}^{\mathrm{gl}}(\lambda)} \frac{1}{\mathbb{E}_{\widehat{M} \sim \xi^s}[\mathbb{E}_\omega[D\big(\widehat{M}(\pi) \,\|\, M(\pi)\big)]} \\ &\le \overline{\mathsf{aec}}_{\varepsilon/2}^{\mathsf{D}, \mathcal{M}}(\mathcal{M}_{\Delta^\ell, \frac{1}{\Delta^\ell}}; \xi^s)\end{aligned}$$

where the last inequality holds by the definition of $\mathcal{M}^\ell$ and Δ^ℓ. Note that by construction, ξ^s is only supported on $\mathcal{M}^\ell$, so we can bound $\overline{\mathsf{aec}}_{\varepsilon/2}^{\mathsf{D}, \mathcal{M}}(\mathcal{M}_{\Delta^\ell, \frac{1}{\Delta^\ell}}; \xi^s) \le \overline{\mathsf{aec}}_{\varepsilon/2}^{\mathsf{D}, \mathcal{M}}(\mathcal{M}_{\Delta^\ell, \frac{1}{\Delta^\ell}})$. By construction, we also have $\overline{\mathsf{aec}}_{\varepsilon/2}^{\mathsf{D}, \mathcal{M}}(\mathcal{M}_{\Delta^\ell, \frac{1}{\Delta^\ell}}) \le 2^{\ell \alpha_{\mathcal{M}}} \le s^{\alpha_{\mathcal{M}}}$. Lastly, by our choice for q^s we have $\frac{1}{1-q^s} = s^{\alpha_q}$. We can then bound

$$\gamma^s = \frac{1 + \mathsf{g}^\star \delta}{1 - q^s} \cdot \frac{1}{\mathbb{E}_{\widehat{M} \sim \xi^s}[\mathbb{E}_{\omega^s}[D\big(\widehat{M}(\pi) \,\|\, M^\star(\pi)\big)]} \le (1 + \delta) \cdot s^{\alpha_q + \alpha_{\mathcal{M}}}.$$

□

H.3.4.1 Likelihood Ratios

Lemma H.3.12. *Consider running either Algorithm H.2 or Algorithm H.3. Under Assumption 10.5, with probability at least $1 - \delta$, we can bound, for any $M \in \mathcal{M}$, s, and*

$\beta_i > 0$,

$$\sum_{i=1}^{s} \mathbb{E}_{\widehat{M}\sim\xi^i}[\mathbb{E}_{\pi\sim p^i}[D_{\mathsf{KL}}(\widehat{M}(\pi) \,\|\, M(\pi))]] \le \sum_{i=1}^{s} \mathbb{E}_{\widehat{M}\sim\xi^i}\left[\log \frac{\mathbb{P}^{\widehat{M},\pi^i}(r^i,o^i)}{\mathbb{P}^{M,\pi^i}(r^i,o^i)}\right] + V_{\mathcal{M}}\sqrt{56s\log 1/\delta} + V_{\mathcal{M}} \cdot \left(\sum_{i=1}^{s} 1/\beta_i + \sum_{i=1}^{s} \beta_i \mathbb{E}_{\widehat{M}\sim\xi^i}[\mathbb{E}_{\pi\sim p^i}[D(\widehat{M}(\pi) \,\|\, M^\star(\pi))]]\right).$$

Proof of Lemma H.3.12. By Theorem 2 of Shamir [2011], under Assumption 10.5 we have that with probability at least $1-\delta$,

$$\sum_{i=1}^{s} \mathbb{E}_{(r,o)\sim M^\star}\left[\mathbb{E}_{\widehat{M}\sim\xi^i}\left[\log \frac{\mathbb{P}^{\widehat{M},\pi}(r,o)}{\mathbb{P}^{M,\pi}(r,o)}\right] \mid \mathcal{H}^{i-1}\right] \le \sum_{i=1}^{s} \mathbb{E}_{\widehat{M}\sim\xi^i}\left[\log \frac{\mathbb{P}^{\widehat{M},\pi^i}(r^i,o^i)}{\mathbb{P}^{M,\pi^i}(r^i,o^i)}\right] + V_{\mathcal{M}}\sqrt{56s\log 1/\delta}. \tag{H.3.21}$$

Note that

$$\begin{aligned}\mathbb{E}\left[\mathbb{E}_{\widehat{M}\sim\xi^i}\left[\log \frac{\mathbb{P}^{\widehat{M},\pi^i}(r^i,o^i)}{\mathbb{P}^{M,\pi^i}(r^i,o^i)}\right] \mid \mathcal{H}^{i-1}\right] &= \mathbb{E}_{\pi\sim p^i}\left[\mathbb{E}_{(r,o)\sim M^\star(\pi)}\left[\mathbb{E}_{\widehat{M}\sim\xi^i}\left[\log \frac{\mathbb{P}^{\widehat{M},\pi}(r,o)}{\mathbb{P}^{M,\pi}(r,o)}\right]\right]\right] \\ &= \mathbb{E}_{\pi\sim p^i}\left[\mathbb{E}_{\widehat{M}\sim\xi^i}\left[\mathbb{E}_{(r,o)\sim M^\star(\pi)}\left[\log \frac{\mathbb{P}^{\widehat{M},\pi}(r,o)}{\mathbb{P}^{M,\pi}(r,o)}\right]\right]\right]\end{aligned}$$

By Lemma B.4 of Foster et al. [2022], we can bound, for any $\widehat{M} \in \mathcal{M}$,

$$\begin{aligned}&\left|\mathbb{E}_{(r,o)\sim M^\star(\pi)}\left[\log \frac{\mathbb{P}^{\widehat{M},\pi}(r,o)}{\mathbb{P}^{M,\pi}(r,o)}\right] - \mathbb{E}_{(r,o)\sim \widehat{M}(\pi)}\left[\log \frac{\mathbb{P}^{\widehat{M},\pi}(r,o)}{\mathbb{P}^{M,\pi}(r,o)}\right]\right| \\ &\quad\le \sqrt{\frac{1}{2}\left(\mathbb{E}_{(r,o)\sim M^\star(\pi)}\left[\log^2 \frac{\mathbb{P}^{\widehat{M},\pi}(r,o)}{\mathbb{P}^{M,\pi}(r,o)}\right] + \mathbb{E}_{(r,o)\sim \widehat{M}(\pi)}\left[\log^2 \frac{\mathbb{P}^{\widehat{M},\pi}(r,o)}{\mathbb{P}^{M,\pi}(r,o)}\right]\right) \cdot D^2_{\mathsf{H}}(M^\star(\pi), \widehat{M}(\pi))} \\ &\quad\le \sqrt{2V^2_{\mathcal{M}} \cdot D^2_{\mathsf{H}}(M^\star(\pi), \widehat{M}(\pi))}\end{aligned}$$

where the second inequality follows under the subgaussian assumption, Assumption 10.5. It follows that for any $\widehat{M} \in \mathcal{M}$,

$$\begin{aligned}&\mathbb{E}_{\pi\sim p^i}\left[\mathbb{E}_{(r,o)\sim M^\star(\pi)}\left[\log \frac{\mathbb{P}^{\widehat{M},\pi}(r,o)}{\mathbb{P}^{M,\pi}(r,o)}\right]\right] \\ &\ge \mathbb{E}_{\pi\sim p^i}\left[\mathbb{E}_{(r,o)\sim \widehat{M}(\pi)}\left[\log \frac{\mathbb{P}^{\widehat{M},\pi}(r,o)}{\mathbb{P}^{M,\pi}(r,o)}\right]\right] - \sqrt{2V^2_{\mathcal{M}} \cdot \mathbb{E}_{\pi\sim p^i}[D^2_{\mathsf{H}}(M^\star(\pi), \widehat{M}(\pi))]} \\ &\overset{(a)}{=} \mathbb{E}_{\pi\sim p^i}[D_{\mathsf{KL}}(\widehat{M}(\pi) \,\|\, M(\pi))] - \sqrt{2V^2_{\mathcal{M}} \cdot \mathbb{E}_{\pi\sim p^i}[D^2_{\mathsf{H}}(M^\star(\pi), \widehat{M}(\pi))]}\end{aligned}$$

$$\overset{(b)}{\geq} \mathbb{E}_{\pi\sim p^i}[D_{\mathsf{KL}}\big(\widehat{M}(\pi)\,\|\,M(\pi)\big)] - V_{\mathcal{M}}^2\cdot(\beta\mathbb{E}_{\pi\sim p^i}[D_{\mathsf{H}}^2\big(M^\star(\pi),\widehat{M}(\pi)\big)]+1/\beta)$$

where (a) follows by the definition of KL divergence, and (b) from AM-GM for any $\beta>0$. Since this bound holds uniformly for all $\widehat{M}\in\mathcal{M}$, this implies that

$$\mathbb{E}\left[\mathbb{E}_{\widehat{M}\sim\xi^i}\left[\log\frac{\mathbb{P}^{\widehat{M},\pi^i}(r^i,o^i)}{\mathbb{P}^{M,\pi^i}(r^i,o^i)}\right]\mid\mathcal{H}^{i-1}\right]\geq\mathbb{E}_{\widehat{M}\sim\xi^i}[\mathbb{E}_{\pi\sim p^i}[D_{\mathsf{KL}}\big(\widehat{M}(\pi)\,\|\,M(\pi)\big)]]$$
$$-V_{\mathcal{M}}^2\cdot(\beta\mathbb{E}_{\widehat{M}\sim\xi^i}[\mathbb{E}_{\pi\sim p^i}[D_{\mathsf{H}}^2\big(M^\star(\pi),\widehat{M}(\pi)\big)]]+1/\beta).$$

Combining this with (H.3.21), and using Assumption H.2 proves the result. □

Lemma H.3.13. *Under Assumption 10.5, we have, for any $M,M'\in\mathcal{M}$,*

$$D_{\mathsf{KL}}(M'(\pi)\,\|\,M(\pi))\leq 2V_{\mathcal{M}},\quad\forall\pi\in\Pi.$$

Proof of Lemma H.3.13. We have

$$D_{\mathsf{KL}}(M'(\pi)\,\|\,M(\pi))=\mathbb{E}_{(r,o)\sim M'(\pi)}\left[\log\frac{\mathbb{P}^{M',\pi}(r,o)}{\mathbb{P}^{M,\pi}(r,o)}\right]\leq\mathbb{E}_{(r,o)\sim M'(\pi)}\left[\left|\log\frac{\mathbb{P}^{M',\pi}(r,o)}{\mathbb{P}^{M,\pi}(r,o)}\right|\right]\leq 2V_{\mathcal{M}}$$

where the last inequality holds under Assumption 10.5. □

Lemma H.3.14. *Under Assumption 10.5, for any $x>0$ and $M,\widetilde{M}\in\mathcal{M}$, we have*

$$D_{\mathsf{KL}}\big(M(\pi)\,\|\,\widetilde{M}(\pi)\big)\leq(2+2V_{\mathcal{M}}+x)\cdot D_{\mathsf{H}}^2\Big(M(\pi),\widetilde{M}(\pi)\Big)+32(1+V_{\mathcal{M}}^2/x+V_{\mathcal{M}}^3/x^2)\cdot\exp(-x^2/8V_{\mathcal{M}}^2).$$

Proof of Lemma H.3.14. Fix π. Define

$$\mathcal{E}:=\left\{\left|\log\frac{\mathbb{P}^{M,\pi}(r,o)}{\mathbb{P}^{\widetilde{M},\pi}(r,o)}-D_{\mathsf{KL}}\big(M(\pi)\,\|\,\widetilde{M}(\pi)\big)\right|\leq x\right\}$$

and for $j\in\mathbb{N}$,

$$\mathcal{E}_j:=\left\{e^{j-1}\cdot x<\left|\log\frac{\mathbb{P}^{M,\pi}(r,o)}{\mathbb{P}^{\widetilde{M},\pi}(r,o)}-D_{\mathsf{KL}}\big(M(\pi)\,\|\,\widetilde{M}(\pi)\big)\right|\leq e^j\cdot x\right\}.$$

Note that $\mathcal{E},(\mathcal{E}_j)_{j=1}^\infty$ form a partition of the probability space. Furthermore, since $D_{\mathsf{KL}}\big(M(\pi)\,\|\,\widetilde{M}(\pi)\big)=\mathbb{E}_{o\sim M(\pi)}[\log\frac{\mathbb{P}^{M,\pi}(r,o)}{\mathbb{P}^{\widetilde{M},\pi}(r,o)}]$, under Assumption 10.5 we have that $\mathbb{P}^{M,\pi}(\mathcal{E}_j)\leq 2\exp(-x^2e^{2(j-1)}/V_{\mathcal{M}}^2)$ and $\mathbb{P}^{M,\pi}(\mathcal{E}^c)\leq 2\exp(-x^2/V_{\mathcal{M}}^2)$. Now,

$$D_{\mathsf{KL}}\big(M(\pi)\,\|\,\widetilde{M}(\pi)\big)=\int\log\frac{\mathbb{P}^{M,\pi}(r,o)}{\mathbb{P}^{\widetilde{M},\pi}(r,o)}\mathrm{d}\mathbb{P}^{M,\pi}(r,o)$$

$$= \int_{\mathcal{E}} \log \frac{\mathbb{P}^{M,\pi}(r,o)}{\mathbb{P}^{\widetilde{M},\pi}(r,o)} \mathrm{d}\mathbb{P}^{M,\pi}(r,o) + \sum_{j=1}^{\infty} \int_{\mathcal{E}_j} \log \frac{\mathbb{P}^{M,\pi}(r,o)}{\mathbb{P}^{\widetilde{M},\pi}(r,o)} \mathrm{d}\mathbb{P}^{M,\pi}(r,o).$$

Using that $D_{\mathsf{KL}}\big(M(\pi) \,\|\, \widetilde{M}(\pi)\big) \le 2V_{\mathcal{M}}$ by Lemma H.3.13, we can bound

$$\begin{aligned}
\sum_{j=1}^{\infty} \int_{\mathcal{E}_j} \log \frac{\mathbb{P}^{M,\pi}(r,o)}{\mathbb{P}^{\widetilde{M},\pi}(r,o)} \mathrm{d}\mathbb{P}^{M,\pi}(r,o) &\le \sum_{j=1}^{\infty} (e^j x + 2V_{\mathcal{M}}) \cdot \int_{\mathcal{E}_j} \mathrm{d}\mathbb{P}^{M,\pi}(r,o) \\
&= \sum_{j=1}^{\infty} (e^j x + 2V_{\mathcal{M}}) \cdot \mathbb{P}^{M,\pi}(\mathcal{E}_j) \\
&\le \sum_{j=1}^{\infty} (e^j x + 2V_{\mathcal{M}}) \cdot 2\exp(-x^2 e^{2(j-1)}/V_{\mathcal{M}}^2) \\
&\le \int_0^{\infty} (e^j x + 2V_{\mathcal{M}}) \cdot 2\exp(-x^2 e^{2(j-1)}/V_{\mathcal{M}}^2) \mathrm{d}j \\
&\le 4(x + V_{\mathcal{M}}) \int_0^{\infty} e^j \exp(-e^j \cdot x^2 e^{-2}/V_{\mathcal{M}}^2) \mathrm{d}j \\
&= 4(x + V_{\mathcal{M}}) V_{\mathcal{M}}^2 e^2 \exp(-x^2 e^{-2}/V_{\mathcal{M}}^2)/x^2 \\
&\le 32(V_{\mathcal{M}}^2/x + V_{\mathcal{M}}^3/x^2) \cdot \exp(-x^2/8V_{\mathcal{M}}^2).
\end{aligned}$$

We turn now to the first term. Note that we can write

$$\begin{aligned}
&\int_{\mathcal{E}} \log \frac{\mathbb{P}^{M,\pi}(r,o)}{\mathbb{P}^{\widetilde{M},\pi}(r,o)} \mathrm{d}\mathbb{P}^{M,\pi}(r,o) \\
&= \int_{\mathcal{E}, \mathbb{P}^{\widetilde{M},\pi}(r,o) > \mathbb{P}^{M,\pi}(r,o)} \left(\log \frac{\mathbb{P}^{M,\pi}(r,o)}{\mathbb{P}^{\widetilde{M},\pi}(r,o)} + \frac{\mathbb{P}^{\widetilde{M},\pi}(r,o)}{\mathbb{P}^{M,\pi}(r,o)} - 1 \right) \mathrm{d}\mathbb{P}^{M,\pi}(r,o) - \mathbb{P}^{\widetilde{M},\pi}(\mathcal{E}^c) \\
&\quad + \int_{\mathcal{E}, \mathbb{P}^{\widetilde{M},\pi}(r,o) \le \mathbb{P}^{M,\pi}(r,o)} \left(\frac{\mathbb{P}^{M}(o \mid \pi)}{\mathbb{P}^{\widetilde{M},\pi}(r,o)} \log \frac{\mathbb{P}^{M,\pi}(r,o)}{\mathbb{P}^{\widetilde{M},\pi}(r,o)} + 1 - \frac{\mathbb{P}^{M,\pi}(r,o)}{\mathbb{P}^{\widetilde{M},\pi}(r,o)} \right) \mathrm{d}\mathbb{P}^{\widetilde{M},\pi}(r,o) + \mathbb{P}^{M,\pi}(\mathcal{E}^c)
\end{aligned}$$

Following the proof of Lemma 4 of Yang and Barron [1998] and using that $\log \frac{\mathbb{P}^{M,\pi}(r,o)}{\mathbb{P}^{\widetilde{M},\pi}(r,o)} \le D_{\mathsf{KL}}\big(M(\pi) \,\|\, \widetilde{M}(\pi)\big) + x \le 2V_{\mathcal{M}} + x$ on $\mathcal{E}$, we can bound this as

$$\begin{aligned}
&\le (2 + 2V_{\mathcal{M}} + x) \int_{\mathcal{E}} (\sqrt{\mathrm{d}\mathbb{P}^{M,\pi}(r,o)} - \sqrt{\mathrm{d}\mathbb{P}^{\widetilde{M},\pi}(r,o)})^2 + \mathbb{P}^{M,\pi}(\mathcal{E}^c) \\
&\le (2 + 2V_{\mathcal{M}} + x) \int (\sqrt{\mathrm{d}\mathbb{P}^{M,\pi}(r,o)} - \sqrt{\mathrm{d}\mathbb{P}^{\widetilde{M},\pi}(r,o)})^2 + \mathbb{P}^{M,\pi}(\mathcal{E}^c) \\
&\le (2 + 2V_{\mathcal{M}} + x) \cdot D_{\mathsf{H}}^2\big(M(\pi), \widetilde{M}(\pi)\big) + 2\exp(-x^2/V_{\mathcal{M}}^2).
\end{aligned}$$

□

H.3.4.2 Covering Numbers

Lemma H.3.15. *For any subset $\mathcal{M}' \subseteq \mathcal{M}$, there exists some (ρ, μ)-cover $\mathcal{M}'_{\mathrm{cov}} \subseteq \mathcal{M}'$ for $\mathcal{M}'$ such that $|\mathcal{M}'_{\mathrm{cov}}| \le \mathsf{N}_{\mathrm{cov}}(\mathcal{M}, \rho/2, \mu)$.*

Proof of Lemma H.3.15. Let $\mathcal{M}_{\mathrm{cov}}$ denote a $(\rho/2, \mu)$-cover of $\mathcal{M}$ with event $\mathcal{E}$. Throughout the proof we use the shorthand $(r, o, \pi) \in \mathcal{E}$ to denote that there exists $M \in \mathcal{M}$ such that $\mathbb{P}^{M,\pi}(r, o \mid \mathcal{E}) > 0$. By definition, it follows that for any $M \in \mathcal{M}$, there exists $M' \in \mathcal{M}_{\mathrm{cov}}$ such that

$$\sup_{r,o,\pi \,:\, (r,o,\pi)\in\mathcal{E}} \left| \log \frac{\mathbb{P}^{M,\pi}(r,o)}{\mathbb{P}^{M',\pi}(r,o)} \right| \le \rho/2. \tag{H.3.22}$$

Let $\mathcal{M}'_{\mathrm{cov}} = \varnothing$ and consider running the following procedure for every $M' \in \mathcal{M}_{\mathrm{cov}}$:

1. Choose a single $M \in \mathcal{M}'$ such that $\sup_{r,o,\pi:(r,o,\pi)\in\mathcal{E}} \left| \log \frac{\mathbb{P}^{M,\pi}(r,o)}{\mathbb{P}^{M',\pi}(r,o)} \right| \le \rho/2$ (if such an M exists).

2. If there exists an $M \in \mathcal{M}'$ in step 1, set $\mathcal{M}'_{\mathrm{cov}} \leftarrow \mathcal{M}'_{\mathrm{cov}} \cup \{M\}$. Otherwise $\mathcal{M}'_{\mathrm{cov}}$ remains unchanged.

By construction $\mathcal{M}'_{\mathrm{cov}} \subseteq \mathcal{M}'$, and $|\mathcal{M}'_{\mathrm{cov}}| \le |\mathcal{M}_{\mathrm{cov}}|$. We claim that $\mathcal{M}'_{\mathrm{cov}}$ is a (ρ, μ)-cover of $\mathcal{M}'$. To see why, take some $M \in \mathcal{M}'$. Let $M' \in \mathcal{M}_{\mathrm{cov}}$ denote the point realizing (H.3.22) for M. Let M'' denote the point chosen in the above procedure for M'. Note that there must exist some M'' chosen for this M' since (H.3.22) holds for M, so in particular $M \in \mathcal{M}'$ satisfies the condition of step 1 in the above procedure. Then,

$$\begin{aligned}
\sup_{r,o,\pi \,:\, (r,o,\pi)\in\mathcal{E}} \left| \log \frac{\mathbb{P}^{M,\pi}(r,o)}{\mathbb{P}^{M'',\pi}(o|\pi)} \right| &= \sup_{r,o,\pi \,:\, (r,o,\pi)\in\mathcal{E}} \left| \log \frac{\mathbb{P}^{M,\pi}(r,o)}{\mathbb{P}^{M',\pi}(r,o)} + \log \frac{\mathbb{P}^{M',\pi}(r,o)}{\mathbb{P}^{M'',\pi}(o|\pi)} \right| \\
&\le \sup_{r,o,\pi \,:\, (r,o,\pi)\in\mathcal{E}} \left| \log \frac{\mathbb{P}^{M,\pi}(r,o)}{\mathbb{P}^{M',\pi}(r,o)} \right| + \sup_{r,o,\pi \,:\, (r,o,\pi)\in\mathcal{E}} \left| \log \frac{\mathbb{P}^{M'',\pi}(o|\pi)}{\mathbb{P}^{M',\pi}(r,o)} \right| \\
&\le \rho/2 + \rho/2 = \rho
\end{aligned}$$

where the last inequality follows by our choice of M' and the definition of M''. Thus, it follows that $\mathcal{M}'_{\mathrm{cov}}$ is a (ρ, μ)-cover of $\mathcal{M}'$. □

H.3.4.3 Further Lemmas

Lemma H.3.16. *For $a > 0$, $\alpha \le [0, 3/4]$, and $\beta > 0$, the function $x^\alpha \log^\beta(ax)$ is concave in x for $x \ge \frac{1}{a}\exp\left(\frac{4\beta}{\alpha}\right)$ when $\beta \le 1$, and for $x \ge \max\left\{\frac{1}{a}\exp\left(\sqrt{\frac{8(\beta-1)\beta}{\alpha}}\right), \frac{1}{a}\exp\left(\frac{4\beta}{\alpha}\right)\right\}$ when $\beta > 1$.*

Proof of Lemma H.3.16. By some calculation, we have

$$\frac{\mathrm{d}^2}{\mathrm{d}^2 x}\left(x^\alpha \log^\beta(ax)\right) = (-1+\beta)\beta x^{-2+\alpha}\log^{\beta-2}(ax) + (-\beta+2\alpha\beta)x^{-2+\alpha}\log^{\beta-1}(ax) + (-1+\alpha)\alpha x^{-2+\alpha}\log^\beta(ax).$$

If we restrict to $x \geq 1/a$, then to show that the function is concave it then suffices to show that

$$(-1+\beta)\beta\log^{-2}(ax) + (-\beta+2\alpha\beta)\log^{-1}(ax) + (-1+\alpha)\alpha \leq 0$$

which, since $\alpha \leq 3/4$, is implied by

$$(-1+\beta)\beta\log^{-2}(ax) + \frac{1}{2}\beta\log^{-1}(ax) \leq \frac{1}{4}\alpha$$

which is further implied by

$$(-1+\beta)\beta\log^{-2}(ax) \leq \frac{1}{8}\alpha \quad \text{and} \quad \frac{1}{2}\beta\log^{-1}(ax) \leq \frac{1}{8}\alpha.$$

The former condition is met for $x > 1/a$ for all $\beta \in (0,1]$. For $\beta > 1$, it is met as long as

$$\frac{8(\beta-1)\beta}{\alpha} \leq \log^2(ax) \iff x \geq \frac{1}{a}\exp\left(\sqrt{\frac{8(\beta-1)\beta}{\alpha}}\right).$$

The latter condition is met for

$$x \geq \frac{1}{a}\exp\left(\frac{4\beta}{\alpha}\right).$$

□

Lemma H.3.17. *For all $B, C, n \geq 1$, if $x \leq C\log^n(Bx)$, then $x \leq C(2n)^n\log^n(2nBC)$.*

Proof of Lemma H.3.17. This is a direct consequence of Lemma G.1.2. □

H.4 Deferred Proofs from Section 10.3

In this section, we provide proofs for the examples given in Section 10.2. We begin in Appendix H.4.1 by introducing a condition which implies $\mathsf{n}_\varepsilon^{M^\star}$ is bounded, and is easy to verify for many classes of interest. Next, in Appendix H.4.2, we consider a variety of structured bandit settings, and in Appendix H.4.3 extend this to contextual bandits with finitely many arms. In Appendix H.4.4, we provide proofs for the informative arm example of Section 10.1.3. Finally, in Appendix H.4.5, we consider tabular MDPs.

H.4.1 Preliminaries: Regular Models

To bound the quantity $\mathsf{n}^\star_\varepsilon = \mathsf{n}^{M^\star}_\varepsilon$ for the examples we consider, it will be helpful to introduce the following notion of a *regular model*.

Definition H.4.1 (Regular Model)**.** We say instance $M \in \mathcal{M}$ is a *regular model* if there exists some constant $L^M_{\mathcal{M}} > 0$ such that, for any $M' \in \mathcal{M}^{\mathrm{alt}}(M)$ with $D_{\mathsf{KL}}(M(\pi_M) \,\|\, M'(\pi_M)) > 0$, there exists $M'' \in \mathcal{M}^{\mathrm{alt}}(M)$ such that $D_{\mathsf{KL}}(M(\pi_M) \,\|\, M''(\pi_M)) = 0$ and, for all $\pi \in \Pi$,

$$|D_{\mathsf{KL}}(M(\pi) \,\|\, M'(\pi)) - D_{\mathsf{KL}}(M(\pi) \,\|\, M''(\pi))| \le \sqrt{L^M_{\mathcal{M}} D_{\mathsf{KL}}(M(\pi_M) \,\|\, M'(\pi_M))} + L^M_{\mathcal{M}} D_{\mathsf{KL}}(M(\pi_M) \,\|\, M'(\pi_M)). \tag{H.4.1}$$

Our definition of a regular model is a direct generalization of existing notions of class regularity found in the literature [Degenne et al., 2020b]. As we will see, for a variety of standard bandit classes (including multi-armed bandits, linear bandits, and Lipschitz bandits), as well as tabular MDPs, one can show that$M^\star$ is a regular model with $L^\star_{\mathcal{M}} = L^{M^\star}_{\mathcal{M}}$ bounded by a polynomial function of problem parameters. Intuitively, $M^\star$ will be a regular model if, for any instance $M' \in \mathcal{M}^{\mathrm{alt}}(M^\star)$ for which it is sufficient to pull $\pi_\star$ in order to distinguish $M^\star$ and M' (thereby ruling out M' while incurring no regret), then there exists some other instance $M'' \in \mathcal{M}^{\mathrm{alt}}(M^\star)$ which is "close" to M' in a certain sense, and which cannot be distinguished from $M^\star$ by simply pulling $\pi_\star$. As the following result shows, the quantity $\mathsf{n}^\star_\varepsilon$ can be bounded whenever $M^\star$ is a regular model.

Proposition H.3. *If M is a regular model with $\Delta^M_{\min} > 0$, we can bound*

$$\mathsf{n}^M_\varepsilon \le \frac{2\mathsf{g}^M}{\Delta^M_{\min}} \cdot \left(1 + L^M_{\mathcal{M}} + \frac{2\mathsf{g}^M}{\varepsilon \Delta^M_{\min}} \cdot L^M_{\mathcal{M}}\right).$$

Given Proposition H.3, for many of the examples in this section, rather than bounding $\mathsf{n}^\star_\varepsilon$ directly, we first show that $M^\star$ is a regular model with $L^\star_{\mathcal{M}}$ well-bounded, and then use Proposition H.3 to obtain a bound on $\mathsf{n}^\star_\varepsilon$.

Proof of Proposition H.3. To prove this result, it suffices to show that, for every normalized allocation $\lambda \in \Upsilon(M, \varepsilon)$ with normalization factor n, there exists some allocation $\eta \in \mathbb{R}^\Pi_+$ such that 1) $\eta(\pi) = \mathsf{n}\lambda(\pi)$ for $\pi \neq \pi_M$, and $\eta(\pi_M) \le \bar{\mathsf{n}}$ for some well-bounded $\bar{\mathsf{n}}$, and 2) $\Delta^M(\eta) \le (1 + 2\varepsilon)\mathsf{g}^M$ and $I^M(\eta) \ge 1 - 2\varepsilon$.

Fix some $\lambda \in \Upsilon(M, \varepsilon)$ with normalization factor $\mathsf{n} > 0$. Note that if M is a regular model, then $I^M(\mathbb{I}_{\pi_M}) = 0$. Since $\lambda \in \Upsilon(M, \varepsilon)$, this implies that $\lambda(\pi_M) < 1$. Let λ' denote the allocation $\lambda'(\pi_M) = 1 - \zeta$ and $\lambda'(\pi) = \frac{\zeta}{1 - \lambda(\pi_M)}\lambda(\pi)$ for $\pi \neq \pi_M$, for some ζ to be chosen.

We have

$$\Delta^M(\lambda') = \frac{\zeta}{1 - \lambda(\pi_M)} \Delta^M(\lambda) \le \frac{\zeta}{1 - \lambda(\pi_M)}(1 + \varepsilon)\mathsf{g}^M/\mathsf{n}. \tag{H.4.2}$$

Take $M' \in \mathcal{M}^{\text{alt}}(M)$ such that $D_{\mathsf{KL}}(M(\pi_M) \,\|\, M'(\pi_M)) > 0$ and let $M'' \in \mathcal{M}^{\text{alt}}(M)$ denote the instance guaranteed to exist under Definition H.4.1. We then have that, for any π,

$$\begin{aligned} D_{\mathsf{KL}}(M(\pi) \,\|\, M'(\pi)) &\geq D_{\mathsf{KL}}(M(\pi) \,\|\, M''(\pi)) - \sqrt{L^M_{\mathcal{M}} D_{\mathsf{KL}}(M(\pi_M) \,\|\, M'(\pi_M))} - L^M_{\mathcal{M}} D_{\mathsf{KL}}(M(\pi_M) \,\|\, M'(\pi_M)) \\ &\geq D_{\mathsf{KL}}(M(\pi) \,\|\, M''(\pi)) - (1+\alpha) L^M_{\mathcal{M}} D_{\mathsf{KL}}(M(\pi_M) \,\|\, M'(\pi_M)) - \frac{1}{\alpha} \end{aligned}$$

where the last inequality follows for any $\alpha > 0$ by AM-GM. Then

$$\begin{aligned} &\sum_{\pi} \lambda'(\pi) D_{\mathsf{KL}}(M(\pi) \,\|\, M'(\pi)) \\ &= \sum_{\pi \neq \pi_M} \frac{\zeta}{1-\lambda(\pi_M)} \lambda(\pi) D_{\mathsf{KL}}(M(\pi) \,\|\, M'(\pi)) + (1-\zeta) D_{\mathsf{KL}}(M(\pi_M) \,\|\, M'(\pi_M)) \\ &\geq \sum_{\pi \neq \pi_M} \frac{\zeta \lambda(\pi)}{1-\lambda(\pi_M)} \Big(D_{\mathsf{KL}}(M(\pi) \,\|\, M''(\pi)) - (1+\alpha) L^M_{\mathcal{M}} D_{\mathsf{KL}}(M(\pi_M) \,\|\, M'(\pi_M)) - \frac{1}{\alpha} \Big) \\ &\quad + (1-\zeta) D_{\mathsf{KL}}(M(\pi_M) \,\|\, M'(\pi_M)) \\ &= \sum_{\pi \neq \pi_M} \frac{\zeta}{1-\lambda(\pi_M)} \lambda(\pi) D_{\mathsf{KL}}(M(\pi) \,\|\, M''(\pi)) + \Big((1-\zeta) - (1+\alpha) L^M_{\mathcal{M}} \zeta \Big) D_{\mathsf{KL}}(M(\pi_M) \,\|\, M'(\pi_M)) - \frac{\zeta}{\alpha}. \end{aligned}$$

We have $M'' \in \mathcal{M}^{\text{alt}}(M)$, so by definition

$$\sum_{\pi} \lambda(\pi) D_{\mathsf{KL}}(M(\pi) \,\|\, M''(\pi)) \geq (1-\varepsilon)/\mathsf{n}.$$

However, $D_{\mathsf{KL}}(M(\pi_M) \,\|\, M''(\pi_M)) = 0$ by assumption, so it follows that

$$\sum_{\pi \neq \pi_M} \frac{\zeta}{1-\lambda(\pi_M)} \lambda(\pi) D_{\mathsf{KL}}(M(\pi) \,\|\, M''(\pi)) \geq \frac{\zeta}{1-\lambda(\pi_M)} \cdot \frac{1-\varepsilon}{\mathsf{n}}.$$

By assumption, $\Delta^M(\lambda) \leq (1+\varepsilon)\mathsf{g}^M/\mathsf{n}$, and we can also lower bound $\Delta^M(\lambda) \geq (1-\lambda(\pi_M))\Delta^M_{\min}$. Rearranging these implies that

$$(1-\lambda(\pi_M))\mathsf{n} \leq (1+\varepsilon)\mathsf{g}^M/\Delta^M_{\min}.$$

Set $\alpha = \frac{(1+\varepsilon)\mathsf{g}^M}{\varepsilon \Delta^M_{\min}}$, then we have

$$\begin{aligned} \frac{\zeta}{1-\lambda(\pi_M)} \cdot \frac{1-\varepsilon}{\mathsf{n}} - \frac{\zeta}{\alpha} &= \frac{\zeta}{1-\lambda(\pi_M)} \cdot \frac{1-\varepsilon}{\mathsf{n}} - \frac{\zeta\varepsilon}{(1+\varepsilon)\mathsf{g}^M/\Delta^M_{\min}} \\ &\geq \frac{\zeta}{1-\lambda(\pi_M)} \cdot \frac{1-\varepsilon}{\mathsf{n}} - \frac{\zeta\varepsilon}{(1-\lambda(\pi_M))\mathsf{n}} \end{aligned}$$

$$= \frac{\zeta}{1-\lambda(\pi_M)} \cdot \frac{1-2\varepsilon}{\mathsf{n}}.$$

Furthermore, with this choice of α, if

$$\zeta \leq \frac{1}{1+(1+\alpha)L^M_{\mathcal{M}}} = \frac{1}{1+(1+(1+\varepsilon)\mathsf{g}^M/\varepsilon\Delta^M_{\min})L^M_{\mathcal{M}}},$$

we have

$$\Big((1-\zeta)-(1+\alpha)L^M_{\mathcal{M}}\zeta\Big)D_{\mathsf{KL}}(M(\pi_M)\,\|\,M'(\pi_M)) \geq 0.$$

We therefore have that, for any $M' \in \mathcal{M}^{\text{alt}}(M)$ with $D_{\mathsf{KL}}(M(\pi_M)\,\|\,M'(\pi_M)) > 0$, that

$$\sum_\pi \lambda'(\pi) D_{\mathsf{KL}}(M(\pi)\,\|\,M'(\pi)) \geq \frac{\zeta}{1-\lambda(\pi_M)} \cdot \frac{1-2\varepsilon}{\mathsf{n}}.$$

Now consider $M' \in \mathcal{M}^{\text{alt}}(M)$ with $D_{\mathsf{KL}}(M(\pi_M)\,\|\,M'(\pi_M)) = 0$. In this case we have

$$\sum_\pi \lambda'(\pi) D_{\mathsf{KL}}(M(\pi)\,\|\,M'(\pi)) = \frac{\zeta}{1-\lambda(\pi_M)} \cdot \sum_\pi \lambda(\pi) D_{\mathsf{KL}}(M(\pi)\,\|\,M'(\pi)) \geq \frac{\zeta}{1-\lambda(\pi_M)} \cdot \frac{1-\varepsilon}{\mathsf{n}}$$

where the inequality follows since $\lambda \in \Upsilon(M,\varepsilon)$ with normalization factor n by assumption. Together these bounds then imply that:

$$I^M(\lambda') \geq \frac{\zeta}{1-\lambda(\pi_M)} \cdot \frac{1-2\varepsilon}{\mathsf{n}}.$$

Combining this with our bound on $\Delta^M(\lambda')$ in (H.4.2) implies that $\lambda' \in \Upsilon(M;2\varepsilon)$ with parameter $\mathsf{n}' = \frac{1-\lambda(\pi_M)}{\zeta} \cdot \mathsf{n}$.

To conclude, define the allocation $\eta := \mathsf{n}'\lambda'$. Then for $\pi \neq \pi_M$:

$$\eta(\pi) = \frac{1-\lambda(\pi_M)}{\zeta} \cdot \mathsf{n}\lambda'(\pi) = \mathsf{n}\lambda(\pi)$$

and

$$\eta(\pi_M) = \frac{1-\lambda(\pi_M)}{\zeta} \cdot \mathsf{n}\lambda'(\pi_M) \leq \frac{1-\lambda(\pi_M)}{\zeta} \cdot \mathsf{n}$$

It follows then that $\Delta^M(\eta) = \mathsf{n}\Delta^M(\lambda) \leq (1+\varepsilon)\mathsf{g}^M$, and $I^M(\eta) \geq \mathsf{n}'I^M(\lambda') \geq 1-2\varepsilon$. Therefore, η satisfies the desired condition. Since $\lambda \in \Upsilon(M,\varepsilon)$, we have

$$\Delta^M(\lambda) \leq (1+\varepsilon)\mathsf{g}^M/\mathsf{n} \implies (1-\lambda(\pi_M))\mathsf{n} \leq (1+\varepsilon)\mathsf{g}^M/\Delta^M_{\min},$$

and thus

$$\eta(\pi_M) \le \frac{(1+\varepsilon)\mathsf{g}^M}{\Delta^M_{\min}\cdot\zeta} \le \frac{2\mathsf{g}^M(1+(1+2\mathsf{g}^M/\varepsilon\Delta^M_{\min})L^\star_{\mathcal{M}})}{\Delta^M_{\min}}.$$

which proves the result. □

H.4.2 Structured Bandits with Gaussian Noise

In this section, we consider the problem of structured bandits with Gaussian noise, in which $\mathcal{O} = \{\varnothing\}$, and the mean reward functions belong to a given function class $\mathcal{F}$. Concretely, we consider the model class

$$\mathcal{M} = \big\{M(\pi) = \mathcal{N}(f(\pi), \sigma^2) \;:\; f \in \mathcal{F}\big\}.$$

We set

$$D\big(M(\pi)\,\|\,\overline{M}(\pi)\big) \leftarrow D_{\mathsf{KL}}\big(M(\pi)\,\|\,\overline{M}(\pi)\big) = \frac{1}{2\sigma^2}(f^M(\pi) - f^{\overline{M}}(\pi))^2 \tag{H.4.3}$$

for D the divergence used by $\mathsf{AE}^2_\star$. In general in the following examples we take $\sigma = 1$ for simplicity.

We begin by verifying that the basic regularity conditions required by our results are satisfied for generic classes $\mathcal{F}$, then provide bounds on the AEC for specific classes of interest.

Lemma H.4.1. *For bandits with Gaussian noise:*

1. *For $D \leftarrow D_{\mathsf{KL}}$, Assumptions 10.5, H.1 and H.2 hold with parameters*

$$L_{\mathsf{KL}} = V_{\mathcal{M}} = \frac{2\sqrt{2}}{\sigma}.$$

2. *We can bound $\mathsf{N}_{\mathrm{cov}}(\mathcal{M}, \rho, \mu)$ by the covering number of $\mathcal{M}$ in the distance $d(M, M') := \sup_{\pi\in\Pi}|f^M(\pi) - f^{M'}(\pi)|$ at tolerance $\frac{\sigma^2\cdot\rho}{2+\sqrt{2\sigma^2\log(2/\mu)}}$. Furthermore, it suffices to take $\mathcal{E} := \{|r| \le 1 + \sqrt{2\sigma^2\log(2/\mu)}\}$.*

Proof of Lemma H.4.1. In this setting, we have that for any $M, M' \in \mathcal{M}$ and any $\pi \in \Pi$,

$$D_{\mathsf{KL}}(M(\pi)\,\|\,M'(\pi)) = \frac{1}{2\sigma^2}(f^M(\pi) - f^{M'}(\pi))^2.$$

For $\overline{M} \in \mathcal{M}$, we therefore have

$$\big|D_{\mathsf{KL}}(M(\pi)\,\|\,M'(\pi)) - D_{\mathsf{KL}}\big(\overline{M}(\pi)\,\|\,M'(\pi)\big)\big| = \frac{1}{2\sigma^2}\big|(f^M(\pi) - f^{M'}(\pi))^2 - (f^{\overline{M}}(\pi) - f^{M'}(\pi))^2\big|$$

$$\leq \frac{2}{\sigma^2}|f^M(\pi) - f^{\overline{M}}(\pi)|$$
$$= \frac{2\sqrt{2}}{\sigma}\sqrt{D\big(M(\pi) \,\|\, \overline{M}(\pi)\big)},$$

where the inequality follows from the Mean Value Theorem and the assumption that $f^{\overline{M}}(\pi) \in [0,1]$ for all $\pi \in \Pi$. This verifies that Assumption H.1 holds with $L_{\mathsf{KL}} = \frac{2\sqrt{2}}{\sigma}$.

To show that Assumption 10.5 is met, we note that for all $M, \overline{M}, \overline{M}' \in \mathcal{M}$,

$$\begin{aligned}
\log \frac{\mathbb{P}^{\overline{M},\pi}(r,o)}{\mathbb{P}^{M,\pi}(r,o)} &= \frac{1}{2\sigma^2}(r - f^M(\pi))^2 - \frac{1}{2\sigma^2}(r - f^{\overline{M}}(\pi))^2 \\
&= \frac{1}{2\sigma^2}\left[f^M(\pi)^2 + f^{\overline{M}}(\pi)^2 - 2r(f^M(\pi) - f^{\overline{M}}(\pi))\right] \\
&= \frac{1}{2\sigma^2}\left[f^M(\pi)^2 + f^{\overline{M}}(\pi)^2 - 2f^{\overline{M}'}(\pi)(f^M(\pi) - f^{\overline{M}}(\pi)) + 2(f^{\overline{M}'}(\pi) - r)(f^M(\pi) - f^{\overline{M}}(\pi))\right] \\
&= \mathbb{E}_{(r,o)\sim\overline{M}'(\pi)}\left[\log \frac{\mathbb{P}^{\overline{M},\pi}(r,o)}{\mathbb{P}^{M,\pi}(r,o)}\right] + \frac{1}{\sigma^2}(f^{\overline{M}'}(\pi) - r)(f^M(\pi) - f^{\overline{M}}(\pi)).
\end{aligned}$$

It follows that

$$\log \frac{\mathbb{P}^{\overline{M},\pi}(r,o)}{\mathbb{P}^{M,\pi}(r,o)} - \mathbb{E}_{(r,o)\sim\overline{M}'(\pi)}\left[\log \frac{\mathbb{P}^{\overline{M},\pi}(r,o)}{\mathbb{P}^{M,\pi}(r,o)}\right]$$

is sub-gaussian with parameter $\mathbb{E}_{r\sim\overline{M}'(\pi)}[(\frac{1}{\sigma^2}(f^{\overline{M}'}(\pi) - r)(f^M(\pi) - f^{\overline{M}}(\pi)))^2] \leq \frac{4}{\sigma^2}$, which verifies Assumption 10.5 with $V_{\mathcal{M}}^2 = \frac{8}{\sigma^2}$.

Finally, we bound the covering number. Let $\mathcal{E} := \{|r| \leq 1 + \sqrt{2\sigma^2 \log(2/\mu)}\}$. Elementary manipulations show that $\mathbb{P}^{\overline{M},\pi}(\mathcal{E}^c) \leq \mu$ for any $\overline{M} \in \mathcal{M}$ and π. Using the same calculation as above, we have

$$\log \frac{\mathbb{P}^{M,\pi}(r,o)}{\mathbb{P}^{M',\pi}(r,o)} = \frac{1}{2\sigma^2}(r - f^M(\pi))^2 - \frac{1}{2\sigma^2}(r - f^{M'}(\pi))^2 \leq \frac{1 + |r|}{\sigma^2} \cdot |f^M(\pi) - f^{M'}(\pi)|$$

where the inequality follows from the Mean Value Theorem. We therefore have that for any $M, M' \in \mathcal{M}$,

$$\sup_{r,o,\pi \,:\, |r| \leq 1+\sqrt{2\sigma^2\log(2/\mu)}} \left|\log \frac{\mathbb{P}^{M,\pi}(r,o)}{\mathbb{P}^{M',\pi}(r,o)}\right| \leq \frac{2 + \sqrt{2\sigma^2 \log(2/\mu)}}{\sigma^2} \cdot \sup_{\pi} |f^M(\pi) - f^{M'}(\pi)|.$$

It follows that if we can form a $\frac{\sigma^2 \cdot \rho}{2+\sqrt{2\sigma^2\log(2/\mu)}}$-cover of $\mathcal{M}$ in the distance $d(M, M') = \sup_{\pi\in\Pi} |f^M(\pi) - f^{M'}(\pi)|$, this will serve as an (ρ, μ) cover of $\mathcal{M}$. $\square$

H.4.2.1 Discrete Structured Bandits

As a first example of bandits with Gaussian noise, we present an additional class that satisfies the uniformly regular assumption.

Example H.4.1 (Discrete Structured Bandits)**.** Fix $\Delta_{\min} > 0$, and consider a discrete reward space $\mathcal{R} \subseteq [0,1]$ satisfying $\min_{r,r' \in \mathcal{R}} |r - r'| \geq \Delta_{\min}$. Consider any function class $\mathcal{F} \subseteq (\Pi \to \mathcal{R})$ defined such that each $f \in \mathcal{F}$ has a unique optimal decision. Let our model class be defined as

$$\mathcal{M} = \{M(\pi) = \mathcal{N}(f(\pi), 1) \mid f \in \mathcal{F}\}.$$

I t is straightforward to show that Assumption 10.4 and Assumption 10.5 are met with $L_{\mathsf{KL}}, V_{\mathcal{M}} \leq 4$, and that Assumption 10.6 is satisfied with d_{cov} scaling with the log-covering number of $\mathcal{F}$ in the distance $d(f, f') = \sup_{\pi \in \Pi} |f(\pi) - f'(\pi)|$, and $C_{\mathsf{cov}} = O(1)$. Furthermore, Assumption 10.7 is satisfied by construction of $\mathcal{R}$ and $\mathcal{F}$, and we can bound $\mathsf{n}_{\varepsilon}^{\mathcal{M}} \leq \frac{2}{\Delta_{\min}^2}$.[1] We thus have the following corollary to Theorem 10.1.

Corollary H.3. *In the discrete structured bandits setting considered above, if* $\mathsf{g}^\star > 0$*, the regret of* AE^2 *is bounded as*

$$\mathbb{E}^{M^\star}[\mathbf{Reg}(T)] \leq (1+\varepsilon)\mathsf{g}^\star \cdot \log(T) + \overline{\mathsf{aec}}_{\varepsilon/12}(\mathcal{M}) \cdot \mathrm{poly}\left(\max_{M \in \mathcal{M}} \mathsf{g}^M, d_{\mathsf{cov}}, \tfrac{1}{\varepsilon}, \tfrac{1}{\Delta_{\min}}, \log\log T\right) \cdot \log^{1/2}(T)$$

As we show in Example 10.3.2, the AEC in this setting can be bounded in terms of the eluder dimension of $\mathcal{F}$.

Proof for Example H.4.1. We provide calculations for the discrete structured bandits setting of Example H.4.1. First, note that Assumptions 10.4 to 10.6 all hold by Lemma H.4.1. To bound $\mathsf{n}^{\mathcal{M}}$, consider $M \in \mathcal{M}$ and $M' \in \mathcal{M}^{\mathrm{alt}}(M)$. Note that either $D_{\mathsf{KL}}(M(\pi_M) \,\|\, M'(\pi_M)) = 0$, in which case there is no advantage to playing π_M, or, due to the discretization of the means, $D_{\mathsf{KL}}(M(\pi_M) \,\|\, M'(\pi_M)) \geq \frac{1}{2}\Delta_{\min}^2$. Thus for any allocation $\eta \in \mathbb{R}_+^{\Pi}$, as long as $\eta(\pi_M) \geq \frac{2}{\Delta_{\min}^2}$, we have $\eta(\pi_M) D_{\mathsf{KL}}(M(\pi_M) \,\|\, M'(\pi_M)) \geq 1$. It follows that there is no advantage to choosing $\eta(\pi_M)$ larger than $\frac{2}{\Delta_{\min}^2}$, so we can bound $\mathsf{n}^{\mathcal{M}} \leq \frac{2}{\Delta_{\min}^2}$. □

H.4.2.2 Multi-Armed Bandits (Example 10.3.1)

In this section we prove the result in Example 10.3.1. First, note that for any $M \in \mathcal{M}$, we have $\mathsf{g}^M \leq A/\Delta_{\min}^M$. Assumptions 10.5, H.1 and H.2 are met due to Lemma H.4.1, and with constants $L_{\mathsf{KL}}, V_{\mathcal{M}} \leq 4$, $d_{\mathsf{cov}} = O(A)$, and $C_{\mathsf{cov}} = O(1)$. By Lemma H.4.2—stated and proven

[1] It is not difficult to see that, given the construction of $\mathcal{R}$, once the optimal arm has been played $\frac{2}{\Delta_{\min}^2}$ times, no additional information can be extracted from playing it.

below—$M^\star$ is a regular model with $L^\star_{\mathcal{M}} = \sqrt{2}$ as long as $f^{M^\star}(\pi_\star) < 1$. It remains to bound the AEC.

Proof of Proposition 10.3. It is immediate to see that $C_{\mathrm{exp}}(\mathcal{M}, \varepsilon) \le O(\frac{A}{\varepsilon})$ by choosing the exploration distribution to be uniform over A. By Lemma H.2.6, we then

$$\overline{\mathsf{aec}}^{\mathcal{M}}_{\varepsilon}(\mathcal{M}^\star) \le c_1 \cdot \frac{A^3}{\varepsilon^2 \Delta_\star^6}$$

for a universal constant c_1. By Proposition H.3, we have

$$\mathsf{n}^\star_{\varepsilon/36} \le c_2 \cdot \frac{\mathsf{g}^\star}{\Delta^\star_{\min}} \cdot \left(1 + \frac{\mathsf{g}^\star}{\varepsilon \Delta^\star_{\min}}\right) \le c_2 \cdot \frac{A^2}{\varepsilon (\Delta^\star_{\min})^4},$$

for a universal constant c_2, so we can lower bound $\Delta_\star \ge c_3 \varepsilon (\Delta^\star_{\min})^4 / A^2$, giving

$$\overline{\mathsf{aec}}^{\mathcal{M}}_{\varepsilon}(\mathcal{M}^\star) \le c_4 \cdot \frac{A^{15}}{\varepsilon^8 (\Delta^\star_{\min})^{24}}.$$

□

Lemma H.4.2. *In the multi-armed bandit setting of Example 10.3.1, any model $M^\star \in \mathcal{M}$ is a regular model with $L^\star_{\mathcal{M}} = \sqrt{3}$ as long as $f^{M^\star}(\pi_\star) < 1$.*

Proof of Lemma H.4.2. Let $M' \in \mathcal{M}^{\mathrm{alt}}(M^\star)$ and assume that $D_{\mathsf{KL}}(M^\star(\pi_\star) \,\|\, M'(\pi_\star)) > 0$.

Case 1: $f^{M'}(\pi_{M'}) + f^{M^\star}(\pi_\star) - f^{M'}(\pi_\star) \le 1$**.** Let M'' denote the Gaussian bandit instance given by $f^{M''}(\pi) = f^{M'}(\pi) + f^{M^\star}(\pi_\star) - f^{M'}(\pi_\star)$; by our assumption, $M'' \in \mathcal{M}$, and $M'' \in \mathcal{M}^{\mathrm{alt}}(M^\star)$. Furthermore, $f^{M''}(\pi_\star) = f^{M^\star}(\pi_\star)$ which implies $D_{\mathsf{KL}}(M^\star(\pi_\star) \,\|\, M''(\pi_\star)) = 0$ as desired. Finally, for any π, we have

$$f^{M'}(\pi) - f^{M''}(\pi) = f^{M^\star}(\pi_\star) - f^{M'}(\pi_\star).$$

This implies that for all π, since all models in $\mathcal{M}$ have unit Gaussian rewards, using the expression for the KL divergence given in (H.4.3):

$$|D_{\mathsf{KL}}(M^\star(\pi) \,\|\, M'(\pi)) - D_{\mathsf{KL}}(M^\star(\pi) \,\|\, M''(\pi))| \le |f^{M^\star}(\pi_\star) - f^{M'}(\pi_\star)| = \sqrt{2 D_{\mathsf{KL}}(M^\star(\pi_\star) \,\|\, M'(\pi_\star))}$$

which implies the condition of Definition H.4.1 is met with $L^\star_{\mathcal{M}} = \sqrt{2}$.

Case 2: $f^{M'}(\pi_{M'}) + f^{M^\star}(\pi_\star) - f^{M'}(\pi_\star) > 1$**.** For this case the model M'' constructed in Case 1 will not be in $\mathcal{M}$. Assume first that $f^{M^\star}(\pi_\star) \ge f^{M'}(\pi_{M'})$ and in this case define M''

to be the instance

$$f^{M''}(\pi) = \begin{cases} \min\{f^{M^\star}(\pi_\star), f^{M'}(\pi) + f^{M^\star}(\pi_\star) - f^{M'}(\pi_\star)\} & \pi \neq \pi_{M'} \\ f^{M^\star}(\pi_\star) + \delta & \pi = \pi_{M'} \end{cases}$$

or some $\delta > 0$ such that $f^{M^\star}(\pi_\star) + \delta < 1$ (note that such a δ exists since we have assumed $f^{M^\star}(\pi_\star) < 1$). Note that we now have $M'' \in \mathcal{M}$, and $f^{M''}(\pi) < f^{M''}(\pi_{M'})$ for all $\pi \neq \pi_{M'}$, so $M'' \in \mathcal{M}^{\text{alt}}(M^\star)$. Furthermore, we have $f^{M''}(\pi_\star) = f^{M^\star}(\pi_\star)$, so $D_{\mathsf{KL}}(M^\star(\pi_\star) \,\|\, M''(\pi_\star)) = 0$. For $\pi \neq \pi_{M'}$, if $\min\{f^{M^\star}(\pi_\star), f^{M'}(\pi) + f^{M^\star}(\pi_\star) - f^{M'}(\pi_\star)\} = f^{M^\star}(\pi_\star)$, this implies that

$$f^{M^\star}(\pi_\star) \leq f^{M'}(\pi) + f^{M^\star}(\pi_\star) - f^{M'}(\pi_\star) \implies f^{M^\star}(\pi_\star) - f^{M'}(\pi) \leq f^{M^\star}(\pi_\star) - f^{M'}(\pi_\star).$$

Since we have assumed $f^{M^\star}(\pi_\star) \geq f^{M'}(\pi_{M'})$, this implies that

$$|f^{M''}(\pi) - f^{M'}(\pi)| = |f^{M^\star}(\pi_\star) - f^{M'}(\pi)| \leq |f^{M^\star}(\pi_\star) - f^{M'}(\pi_\star)|$$

So by the expression given for the KL divergence in (H.4.3), we have

$$|D_{\mathsf{KL}}(M^\star(\pi) \,\|\, M'(\pi)) - D_{\mathsf{KL}}(M^\star(\pi) \,\|\, M''(\pi))| \leq \sqrt{3 D_{\mathsf{KL}}(M^\star(\pi_\star) \,\|\, M'(\pi_\star))}. \tag{H.4.4}$$

For $\pi \neq \pi_{M'}$ with $\min\{f^{M^\star}(\pi_\star), f^{M'}(\pi) + f^{M^\star}(\pi_\star) - f^{M'}(\pi_\star)\} = f^{M'}(\pi) + f^{M^\star}(\pi_\star) - f^{M'}(\pi_\star)$, the bound on $|D_{\mathsf{KL}}(M^\star(\pi) \,\|\, M'(\pi)) - D_{\mathsf{KL}}(M^\star(\pi) \,\|\, M''(\pi))|$ follows identically to Case 1. For $\pi = \pi_{M'}$, since we have assumed that $f^{M^\star}(\pi_\star) \geq f^{M'}(\pi_{M'})$ we have

$$|f^{M''}(\pi_{M'}) - f^{M'}(\pi_{M'})| = f^{M^\star}(\pi_\star) - f^{M'}(\pi_{M'}) + \delta \leq f^{M^\star}(\pi_\star) - f^{M'}(\pi_\star) + \delta.$$

For small enough δ, this implies that (H.4.4) is satisfied for $\pi = \pi_{M'}$ as well.

Consider now the case where $f^{M^\star}(\pi_\star) < f^{M'}(\pi_{M'})$. In this case define M'' by

$$f^{M''}(\pi) = \begin{cases} \min\{f^{M'}(\pi_{M'}) - \delta, f^{M'}(\pi) + f^{M^\star}(\pi_\star) - f^{M'}(\pi_\star)\} & \pi \neq \pi_{M'} \\ f^{M'}(\pi_{M'}) & \pi = \pi_{M'} \end{cases}$$

for $\delta > 0$ small enough that $f^{M'}(\pi_{M'}) - \delta > f^{M^\star}(\pi_\star)$. Note that M'' and $M'' \in \mathcal{M}^{\text{alt}}(M^\star)$ by construction, and that $f^{M''}(\pi_\star) = f^{M^\star}(\pi_\star)$ by our choice of δ, so $D_{\mathsf{KL}}(M^\star(\pi_\star) \,\|\, M''(\pi_\star)) = 0$. For $\pi \neq \pi_{M'}$, if $\min\{f^{M'}(\pi_{M'}) - \delta, f^{M'}(\pi) + f^{M^\star}(\pi_\star) - f^{M'}(\pi_\star)\} = f^{M'}(\pi_{M'}) - \delta$, then we have

$$\begin{aligned} |f^{M''}(\pi) - f^{M'}(\pi)| &= |f^{M'}(\pi_{M'}) - \delta - f^{M'}(\pi)| \\ &\leq f^{M'}(\pi_{M'}) - f^{M'}(\pi) + \delta \\ &\leq |f^{M^\star}(\pi_\star) - f^{M'}(\pi_\star)| + \delta \end{aligned}$$

where for the final inequality we have used that $\min\{f^{M'}(\pi_{M'}) - \delta, f^{M'}(\pi) + f^{M^\star}(\pi_\star) - f^{M'}(\pi_\star) = f^{M'}(\pi_{M'}) - \delta$. It follows that (H.4.4) is satisfied for this π for sufficiently small δ. If we instead

have $\min\{f^{M'}(\pi_{M'}) - \delta, f^{M'}(\pi) + f^{M^\star}(\pi_\star) - f^{M'}(\pi_\star)\} = f^{M'}(\pi) + f^{M^\star}(\pi_\star) - f^{M'}(\pi_\star)$, then the bound on $|D_{\mathsf{KL}}(M^\star(\pi) \,\|\, M'(\pi)) - D_{\mathsf{KL}}(M^\star(\pi) \,\|\, M''(\pi))|$ follows identically to Case 1.

For $\pi = \pi_{M'}$, we have $|f^{M''}(\pi_{M'}) - f^{M'}(\pi_{M'})| = 0$. This proves the result.

□

H.4.2.3 Structured Bandits with Bounded Eluder Dimension (Example 10.3.2)

In this section, we give generic bounds on the uniform exploration coefficient and Allocation-Estimation Coefficient for structured bandit classes with bounded eluder dimension (cf. Definition 10.3.2). These result are used by subsequent examples, including linear bandits.

Lemma H.4.3. *Let $\mathcal{M}\{M(\pi) = \mathcal{N}(f(\pi), 1) \mid f \in \mathcal{F}\}$. Then for all $\varepsilon > 0$, we have*

$$C_{\mathrm{exp}}(\mathcal{M}, \varepsilon) \leq \frac{16 d_{\mathrm{E}}(\mathcal{F}, \sqrt{\varepsilon}/2)}{\varepsilon}.$$

Proof of Lemma H.4.3. Let $\xi \in \triangle(\mathcal{M})$. Recall the expression for KL divergence between Gaussians of unit variance:

$$\mathbb{E}_{\overline{M}\sim\xi}[\mathbb{E}_p[D_{\mathsf{KL}}\big(\overline{M}(\pi) \,\|\, M(\pi)\big)]] = \frac{1}{2}\mathbb{E}_{\overline{M}\sim\xi}[\mathbb{E}_p[(f^M(\pi) - f^{\overline{M}}(\pi))^2]].$$

Abbreviate $d_{\mathrm{E}} := d_{\mathrm{E}}(\mathcal{F}, \sqrt{\varepsilon}/2)$ and let $\{\pi_1, \ldots, \pi_{d_{\mathrm{E}}}\}$ denote a maximal sequence of ε-independent points. By the definition of the eluder dimension, for any $\pi \in \Pi$ and any $M, \overline{M} \in \mathcal{M}$, we have:

$$\sqrt{\sum_{i=1}^{d_{\mathrm{E}}}(f^M(\pi_i) - f^{\overline{M}}(\pi_i))^2} \leq \sqrt{\varepsilon/2} \implies |f^M(\pi) - f^{\overline{M}}(\pi)| \leq \sqrt{\varepsilon/2}.$$

Now, set p to be the uniform distribution over $\{\pi_1, \ldots, \pi_{d_{\mathrm{E}}}\}$. Assume that $M, M' \in \mathcal{M}$ are such that

$$\max_{M''\in\{M,M'\}} \mathbb{E}_{\overline{M}\sim\xi}\,\mathbb{E}_p[D_{\mathsf{KL}}\big(\overline{M}(\pi) \,\|\, M''(\pi)\big)] = \frac{1}{2}\max_{M''\in\{M,M'\}} \mathbb{E}_{\overline{M}\sim\xi}\,\mathbb{E}_p[(f^{M''}(\pi) - f^{\overline{M}}(\pi))^2] \leq \varepsilon/(16 d_{\mathrm{E}}),$$

Markov's inequality implies that for each $M'' \in \{M, M'\}$, with probability at least $3/4$ over the draw of $\overline{M} \sim \xi$,

$$\mathbb{E}_p[(f^{M''}(\pi) - f^{\overline{M}}(\pi))^2] \leq \varepsilon/(2 d_{\mathrm{E}}).$$

Taking a union bound, we conclude that with probability at least $1/2$ over the draw of $\overline{M} \sim \xi$,

$$\max_{M''\in\{M,M'\}} \mathbb{E}_p[(f^{M''}(\pi) - f^{\overline{M}}(\pi))^2] \leq \varepsilon/(2 d_{\mathrm{E}}). \tag{H.4.5}$$

Going forward, let $\overline{M} \in \mathcal{M}$ be any model such that (H.4.5) holds; we have just proven that

such a model exists. It follows from the maximality of $\pi_1, \dots, \pi_{d_{\mathrm{E}}}$ and the definition of p that for all $\widetilde{\pi} \in \Pi$, and $M'' \in \mathcal{M}$,

$$\mathbb{E}_p[(f^{M''}(\pi) - f^{\overline{M}}(\pi))^2] \le \varepsilon/(2d_{\mathrm{E}}) \implies (f^{M''}(\widetilde{\pi}) - f^{\overline{M}}(\widetilde{\pi}))^2 \le \varepsilon/2.$$

In particular, since this holds for both $M \in \{M', M''\}$, and since (H.4.5) holds, we have that for all π,

$$\begin{aligned} D_{\mathsf{KL}}(M'(\pi) \,\|\, M''(\pi)) &= \frac{1}{2}(f^{M'}(\pi) - f^{M''}(\pi))^2 \\ &\le (f^{M'}(\pi) - f^{\overline{M}}(\pi))^2 + (f^{\overline{M}}(\pi) - f^{M''}(\pi))^2 \\ &\le \varepsilon. \end{aligned}$$

As this is the condition required by Definition 10.3.1, it suffices to take $C^{\xi}_{\mathrm{exp}}(\varepsilon) = 16 d_{\mathrm{E}}(\mathcal{F}, \sqrt{\varepsilon}/2)/\varepsilon$. Since this bound holds uniformly for all choices of ξ, the result follows. □

Proof of Proposition 10.4. The bound

$$C_{\mathrm{exp}}(\mathcal{M}^\star, \delta) \le \frac{16 d_{\mathrm{E}}(\mathcal{F}, \sqrt{\delta}/2)}{\delta}.$$

follows from Lemma H.4.3, since $d_{\mathrm{E}}(\mathcal{F}', \delta) \le d_{\mathrm{E}}(\mathcal{F}, \delta)$ for all $\mathcal{F}' \subseteq \mathcal{F}$.

By Lemma H.2.6 we can bound $\overline{\mathsf{aec}}^{\mathcal{M}}_{\varepsilon}(\mathcal{M}^\star)$:

$$\overline{\mathsf{aec}}^{\mathcal{M}}_{\varepsilon}(\mathcal{M}^\star) \le C_{\mathrm{exp}}(\mathcal{M}^\star, \delta) \quad \text{for} \quad \delta = \min_{M \in \mathcal{M}^\star} \min\left\{ \min\left\{ \frac{1}{81 L_{\mathsf{KL}}}, \frac{\Delta^M_{\min}}{34 V_{\mathcal{M}}} \right\} \cdot \frac{\varepsilon}{2\mathsf{g}^M/\Delta^M_{\min} + \mathsf{n}^M_{\varepsilon/36}}, \frac{\Delta^M_{\min}}{3} \right\}^2.$$

Lemma Lemma H.2.2 implies that for all $M \in \mathcal{M}^\star$,

$$\mathsf{g}^M \le C_{\mathrm{exp}}(\mathcal{M}^\star, \tfrac{1}{4}(\Delta^M_{\min})^2) \le \frac{64 d_{\mathrm{E}}(\mathcal{F}, \frac{1}{2}\Delta^M_{\min})}{(\Delta^M_{\min})^2} \le \frac{64 d_{\mathrm{E}}(\mathcal{F}, \frac{1}{2}\Delta_\star)}{\Delta_\star^2}$$

where we have used that the eluder dimension increases as its scale ε decreases; by Lemma H.4.1, it suffices to take $L_{\mathsf{KL}} = V_{\mathcal{M}} = 2$. A sufficient value for δ is therefore

$$\delta = c \cdot \frac{\varepsilon^2 \Delta_\star^8}{d_{\mathrm{E}}(\mathcal{F}, \frac{1}{2}\Delta_\star)^2}$$

The result follows.

□

H.4.2.4 Linear Bandits (Example 10.3.3)

Proof of Proposition 10.5. The result follows directly from Proposition 10.4, since it is known that linear bandits have eluder dimension which scales as $d_{\mathrm{E}}(\mathcal{F}, \varepsilon) = O(d \cdot \log 1/\varepsilon)$ [Russo and Van Roy, 2013]. □

In what follows, we prove Proposition 10.6, providing sufficient conditions under which it is possible to bound the regularity constant $L^\star_{\mathcal{M}}$ (and hence $\mathsf{n}^\star_\varepsilon$) for linear bandits.

We begin with an geometric assumption on Θ, $\mathcal{X}$, and $\theta^\star$, which we will show ensures that $M^\star(\pi) = \mathcal{N}(\langle x_\pi, \theta^\star \rangle, 1)$ is a regular model. To state our condition, we denote, for any vectors x and y, we define x_y and $x_{\bar{y}}$ to be unique vectors satisfying $x = x_y + x_{\bar{y}}$ for $x_y \parallel y$ and $x_{\bar{y}} \perp y$.

Assumption H.3 (Regular Linear Bandits)**.** *The sets $\Theta, \mathcal{X}$ and model parameter $\theta^\star$ satisfy:*

1. *Θ is a convex polytope.*
2. *For all $\theta \in \Theta$, we have that there exists some $\delta_\theta > 0$ such that $\{\theta' \in \mathbb{R}^d \ : \ \|\theta' - \theta\|_2 \leq \delta_\theta\} \subseteq \Theta$.*
3. *Letting $x^\star \in \mathcal{X}$ denote the optimal action for $\theta^\star$, we have*

$$\left\{\theta \in \mathbb{R}^d \ : \ \|\theta - \theta^\star\|_2 \leq \max_{x \in \mathcal{X}, x \neq x^\star} \Delta^\star(x)/\|x_{\bar{x}^\star}\|_2\right\} \subseteq \Theta.$$

The first two points above are quite mild. The primary restriction of Assumption H.3 is Point 3, which requires that $\theta^\star$ is located sufficiently far within the interior of Θ. Using Assumption H.3 we can state the full version of Proposition 10.6.

Proposition H.4 (Full Version of Proposition 10.6)**.** *If $\Theta, \mathcal{X}$, and $\theta^\star$ satisfy Assumption H.3, then $\mathsf{n}^\star_\varepsilon$ is bounded by a polynomial function of $d, 1/\Delta^\star_{\min}, 1/\varepsilon, \mathsf{g}^\star$, and a geometry-dependent term scaling with the structure of $\mathcal{X}$ and Θ.*

Remark H.4.1 (Comparison to Existing Work)**.** We remark that Assumption H.3 is similar to the conditions required by existing works which achieve instance-optimality in linear bandits with polynomial lower-order terms [Tirinzoni et al., 2020, Kirschner et al., 2021]. Though neither of these works explicitly states such a condition, closer inspection of their analysis reveals it is indeed required. In particular, the proof of Lemma 1 of Tirinzoni et al. [2020] relies on a result from Degenne et al. [2020b] which shows that a condition analogous to Definition H.4.1 is met for linear bandits. However, the proof given in Degenne et al. [2020b] appears to only hold when Θ is unbounded, or a condition such as Assumption H.3 holds. As Tirinzoni et al. [2020] assumes that Θ is bounded, their results therefore only appear to hold if a condition similar to Assumption H.3 also holds. Similarly, in the proof of Lemma

10 of Kirschner et al. [2021], it is assumed that for every arm $x \neq x_{\pi_\star}$, there exists some instance in the alternate set with optimal arm x. To satisfy this condition, it appears that an assumption similar to Assumption H.3 is required.

Thus, while not stated explicitly in the existing literature, it therefore seems that all existing results which obtain reasonable lower-order terms require an assumption similar to Assumption H.3. Removing this assumption (or showing it is necessary) is an interesting direction for future work.

Proof of Proposition H.4. Under Assumption H.3, this follows directly from Lemma H.4.4 and Proposition H.3. □

Lemma H.4.4. *Under Assumption H.3, the linear bandit model $M^\star$ is regular for some $L^\star_{\mathcal{M}} < \infty$ whose value depends on the geometry of Θ and $\mathcal{X}$.*

Proof of Lemma H.4.4. Fix some $\theta^\star \in \Theta$ and let $x^\star$ denote its optimal arm. Let $\Theta^{\text{alt}}(\theta^\star) \subseteq \Theta$ denote parameters with optimal arm $x \neq x^\star$. Assume there exists some $\theta \in \Theta^{\text{alt}}(\theta^\star)$ such that $\langle \theta - \theta^\star, x^\star \rangle \neq 0$ (if this is not the case, $M^\star$ immediately satisfies Definition H.4.1 and we are done). Let $\Theta_x = \{\theta \in \Theta \ : \ \langle x, \theta \rangle \geq \langle x^\star, \theta \rangle\}$, $\Theta^\star = \{\theta \in \mathbb{R}^d \ : \ \langle \theta - \theta^\star, x^\star \rangle = 0\}$, and $\Theta^\star_x = \Theta_x \cap \Theta^\star$. We first show that, under Assumption H.3, $\Theta^\star_x \neq \varnothing$ for all $x \in \mathcal{X}$, $x \neq x^\star$. We then use this to show that $M^\star(\pi) = \mathcal{N}(\langle x_\pi, \theta^\star \rangle, 1)$ is a regular model.

Part 1: $\Theta^\star_x \neq \varnothing$. Fix some $x \in \mathcal{X}$ with $x \neq x^\star$. Consider $\theta = \theta^\star + a x_{\bar{x}^\star}$ for some $a \in \mathbb{R}$ to be chosen. By construction we have $\langle \theta, x^\star \rangle = \langle \theta^\star, x^\star \rangle$, which implies that $\theta \in \Theta^\star$ for all $a \in \mathbb{R}$. We wish to choose a large enough that $\langle \theta, x \rangle \geq \langle \theta, x^\star \rangle$. Note that

$$\langle \theta, x \rangle = \langle \theta^\star, x \rangle + a \langle x_{\bar{x}^\star}, x_{\bar{x}^\star} + x_{x^\star} \rangle = \langle \theta^\star, x \rangle + a \|x_{\bar{x}^\star}\|_2^2$$

and $\langle \theta, x^\star \rangle = \langle \theta^\star, x^\star \rangle$. Thus, to satisfy $\langle \theta, x \rangle \geq \langle \theta, x^\star \rangle$, we need

$$a \|x_{\bar{x}^\star}\|_2^2 \geq \langle \theta^\star, x^\star - x \rangle \iff a \geq \Delta^\star(x) / \|x_{\bar{x}^\star}\|_2^2.$$

Let $a = \Delta^\star(x)/\|x_{\bar{x}^\star}\|_2^2$, then it follows that $\langle \theta, x \rangle \geq \langle \theta, x^\star \rangle$. Furthermore, we can bound

$$\|\theta - \theta^\star\|_2 \leq a \|x_{\bar{x}^\star}\|_2 = \Delta^\star(x) / \|x_{\bar{x}^\star}\|_2.$$

Under Assumption H.3, it follows that $\theta \in \Theta$.

Part 2: $M^\star$ is a Regular Model. Let $\overline{\Theta}_x = \{\theta - \theta^\star \ : \ \theta \in \Theta_x\}$. Note that, since Θ is a convex polytope, and Θ_x simply adds a linear inequality constraint, Θ_x is also convex. Let $\overline{\Theta}^\star_x = \{\phi \in \overline{\Theta}_x \ : \ \langle \phi, x^\star \rangle = 0\}$. From Part 1, we have $\overline{\Theta}^\star_x \neq \varnothing$. Lemma 23 of Kirschner et al. [2021] then gives that there exists a geometry-dependent constant $C(\Theta, \mathcal{X})$ such that, for all $\phi \in \overline{\Theta}_x$:

$$\min_{\phi' \in \overline{\Theta}^\star_x} \|\phi - \phi'\|_2 \leq C(\Theta, \mathcal{X}) \cdot |\langle \phi, x^\star \rangle|.$$

This implies that for all $\theta \in \Theta_x$, we have:

$$\min_{\theta' \in \Theta^\star_x} \|\theta - \theta'\|_2 \le C(\Theta, \mathcal{X}) \cdot |\langle \theta - \theta^\star, x^\star \rangle|.$$

Now consider some $\theta \in \Theta^{\text{alt}}(\theta^\star)$, and assume that $\langle \theta - \theta^\star, x^\star \rangle \neq 0$ (by assumption such a θ exists). Assume that θ has optimal arm x, which implies that $\theta \in \Theta_x$. By what we have just shown, we know that there exists some $\theta' \in \Theta$ with $\langle \theta', x \rangle > \langle \theta', x^\star \rangle$ so that $\theta' \in \Theta^{\text{alt}}(\theta^\star)$, $\langle \theta' - \theta^\star, x^\star \rangle = 0$, and

$$\|\theta - \theta'\|_2 \le C(\Theta, \mathcal{X}) \cdot |\langle \theta - \theta^\star, x^\star \rangle|.$$

Note that, for any $x' \in \mathcal{X}$, we have

$$\begin{aligned} |D_{\mathsf{KL}}(\theta^\star(x') \,\|\, \theta(x')) - D_{\mathsf{KL}}(\theta^\star(x') \,\|\, \theta'(x'))| &= \frac{1}{2} |\langle \theta^\star - \theta, x' \rangle^2 - \langle \theta^\star - \theta', x' \rangle^2| \\ &\le 2 \max_{\theta'' \in \Theta} |\langle \theta'', x' \rangle| \cdot |\langle \theta - \theta', x' \rangle| \\ &\le 2 \max_{\theta'' \in \Theta} \|\theta''\|_2 \|x'\|_2^2 \cdot \|\theta - \theta'\|_2. \end{aligned}$$

Furthermore, note that

$$\sqrt{D_{\mathsf{KL}}(\theta^\star(x^\star) \,\|\, \theta(x^\star))} = \frac{1}{\sqrt{2}} |\langle \theta^\star - \theta, x^\star \rangle|,$$

so

$$\begin{aligned} &|D_{\mathsf{KL}}(\theta^\star(x') \,\|\, \theta(x')) - D_{\mathsf{KL}}(\theta^\star(x') \,\|\, \theta'(x'))| \\ &\quad \le \left(2\sqrt{2} C(\Theta, \mathcal{X}) \max_{\theta'' \in \Theta, x'' \in \mathcal{X}} \|\theta''\|_2 \|x''\|_2^2 \right) \cdot \sqrt{D_{\mathsf{KL}}(\theta^\star(x^\star) \,\|\, \theta(x^\star))}. \end{aligned}$$

As $\theta \in \Theta^\star(\theta^\star)$ was arbitrary, we have therefore shown that $M^\star$ is a regular model with

$$L^\star_{\mathcal{M}} = \left(2\sqrt{2} \cdot C(\Theta, \mathcal{X}) \cdot \max_{\theta'' \in \Theta, x'' \in \mathcal{X}} \|\theta''\|_2 \|x''\|_2^2 \right)^2.$$

□

H.4.2.5 Generalized Linear Models (Example 10.3.4)

Proof Sketch for Example 10.3.4. The bound on the AEC follows as in Proposition 10.5, using that the eluder dimension for generalized linear models is bounded as $O(d \cdot (\frac{g_{\max}}{g_{\min}})^2 \cdot \log \frac{g_{\max}}{\varepsilon})$ [Russo and Van Roy, 2013]. For the other regularity assumptions, note that by the Mean

Value Theorem, we have

$$|D_{\mathsf{KL}}(M(\pi) \,\|\, M'(\pi)) - D_{\mathsf{KL}}(M(\pi) \,\|\, M''(\pi))| = \frac{1}{2}|(g(\langle\theta, x\rangle) - g(\langle\theta', x\rangle))^2 - (g(\langle\theta, x\rangle) - g(\langle\theta'', x\rangle))^2| \\ \leq 2g_{\max}|\langle\theta' - \theta'', x\rangle|$$

and

$$\sqrt{D_{\mathsf{KL}}(M(\pi) \,\|\, M'(\pi))} = \frac{1}{\sqrt{2}}|g(\langle\theta, x\rangle) - g(\langle\theta', x\rangle)| \geq \frac{g_{\min}}{\sqrt{2}}|\langle\theta - \theta', x\rangle|.$$

In light of these inequalities, bounds on all relevant regularity parameters for generalized linear bandits follow from similar reasoning to the proofs for linear bandits. In particular, the conclusion of Lemma H.4.4 holds for generalized linear bandits under Assumption H.3, with $L^\star_{\mathcal{M}}$ as in Lemma H.4.4, but scaled by $(\frac{g_{\max}}{g_{\min}})^2)$.

□

H.4.3 Contextual Bandits with Finitely Many Actions (Example 10.3.5)

In this setting we take $D\big(M(\pi) \,\|\, \overline{M}(\pi)\big) \leftarrow D_{\mathsf{KL}}\big(M(\pi) \,\|\, \overline{M}(\pi)\big)$ for D the divergence employed by $\mathsf{AE}^2_\star$. Note that we have

$$D_{\mathsf{KL}}\big(M(\pi) \,\|\, \overline{M}(\pi)\big) = \frac{1}{2}\mathbb{E}_{x\sim p_{\mathcal{X}}}[\mathbb{E}_{a\sim\pi(x)}[(f^M(x,a) - f^{\overline{M}}(x,a))^2]].$$

Lemma H.4.5. *In the contextual bandits setting of Example 10.3.5:*

1. *Assumptions 10.5, H.1 and H.2 hold with parameters*

$$L_{\mathsf{KL}} = V_{\mathcal{M}} = 2\sqrt{2}$$

   and $D\big(\,\cdot \,\|\, \cdot\,\big) = D_{\mathsf{KL}}(\cdot \,\|\, \cdot)$.

2. *We can bound* $\mathsf{N}_{\mathrm{cov}}(\mathcal{M}, \rho, \mu)$ *by the covering number of* $\mathcal{M}$ *in the distance* $d(M, M') = \sup_{x\in\mathcal{X}, a\in\mathcal{A}} |f^M(x,a) - f^{M'}(x,a)|$ *at tolerance* $\frac{\sigma^2\cdot\rho}{2+\sqrt{2\log(2/\mu)}}$. *Furthermore, it suffices to take* $\mathcal{E} := \{|r| \leq 1 + \sqrt{2\log(2/\mu)}\}$.

Proof of Lemma H.4.5. Using the expression for the KL divergence given above, for any $M, M', \overline{M} \in \mathcal{M}$ and $\pi \in \Pi$, we have

$$\big|D_{\mathsf{KL}}(M(\pi) \,\|\, M'(\pi)) - D_{\mathsf{KL}}\big(\overline{M}(\pi) \,\|\, M'(\pi)\big)\big| \\ = \frac{1}{2}\left|\mathbb{E}_{x\sim p_{\mathcal{X}}}[\mathbb{E}_{a\sim\pi(x)}[(f^M(x,a) - f^{M'}(x,a))^2]] - \mathbb{E}_{x\sim p_{\mathcal{X}}}[\mathbb{E}_{a\sim\pi(x)}[(f^{\overline{M}}(x,a) - f^{M'}(x,a))^2]]\right|$$

$$
\begin{aligned}
&\leq \frac{1}{2}\mathbb{E}_{x\sim p_{\mathcal{X}}}\left[\mathbb{E}_{a\sim\pi(x)}\left[\left|(f^{M}(x,a)-f^{M'}(x,a))^2-(f^{\overline{M}}(x,a)-f^{M'}(x,a))^2\right|\right]\right]\\
&\leq 2\mathbb{E}_{x\sim p_{\mathcal{X}}}\left[\mathbb{E}_{a\sim\pi(x)}\left[\left|f^{M}(x,a)-f^{\overline{M}}(x,a)\right|\right]\right]\\
&\leq 2\sqrt{\mathbb{E}_{x\sim p_{\mathcal{X}}}\left[\mathbb{E}_{a\sim\pi(x)}\left[\left(f^{M}(x,a)-f^{\overline{M}}(x,a)\right)^2\right]\right]}\\
&= 2\sqrt{2}\sqrt{D_{\mathsf{KL}}\big(M(\pi)\,\|\,\overline{M}(\pi)\big)}.
\end{aligned}
$$

This verifies Assumption H.1 holds with $L_{\mathsf{KL}} = 2\sqrt{2}$. Assumption H.2 is immediate.

To show that Assumption 10.5 is met, we note that

$$
\log\frac{\mathbb{P}^{\overline{M},\pi}(r,o)}{\mathbb{P}^{M,\pi}(r,o)} = \log\frac{\mathbb{P}^{\overline{M},\pi}(r\mid o)\mathbb{P}^{\overline{M},\pi}(o)}{\mathbb{P}^{M,\pi}(r\mid o)\mathbb{P}^{M,\pi}(o)} = \log\frac{\mathbb{P}^{\overline{M},\pi}(r\mid o)}{\mathbb{P}^{M,\pi}(r\mid o)},
$$

where the second equality holds because the context distribution is identical for all models. As the reward likelihoods conditioned on the context are Gaussian, a calculation similar to Lemma H.4.1 shows that Assumption 10.5 with $V_{\mathcal{M}} = 2\sqrt{2}$. The covering number bound also follows from the same reasoning as Lemma H.4.1. □

Lemma H.4.6. *For the contextual bandit setting described above, we can bound*

$$
C_{\exp}(\mathcal{M},\varepsilon) \leq \frac{4A}{\varepsilon}.
$$

Proof of Lemma H.4.6. Fix some $\xi \in \triangle_{\mathcal{M}}$ and let $\pi_{\exp}$ be uniform over $\mathcal{A}$ for each context. Then, for any $p' \in \triangle_{\Pi}$, we have

$$
\begin{aligned}
&\mathbb{E}_{\pi\sim p'}[D_{\mathsf{KL}}(M(\pi)\,\|\,M'(\pi))]\\
&= \frac{1}{2}\mathbb{E}_{\pi\sim p'}[\mathbb{E}_{x\sim p_{\mathcal{X}}}[\mathbb{E}_{a\sim\pi(x)}[(f^{M}(x,a)-f^{M'}(x,a))^2]]]\\
&\leq \mathbb{E}_{\overline{M}\sim\xi}[\mathbb{E}_{\pi\sim p'}[\mathbb{E}_{x\sim p_{\mathcal{X}}}[\mathbb{E}_{a\sim\pi(x)}[(f^{M}(x,a)-f^{\overline{M}}(x,a))^2+(f^{\overline{M}}(x,a)-f^{M'}(x,a))^2]]]]\\
&\leq \sum_{a\in\mathcal{A}}\mathbb{E}_{\overline{M}\sim\xi}[\mathbb{E}_{x\sim p_{\mathcal{X}}}[(f^{M}(x,a)-f^{\overline{M}}(x,a))^2+(f^{\overline{M}}(x,a)-f^{M'}(x,a))^2]]\\
&= A\mathbb{E}_{\overline{M}\sim\xi}[\mathbb{E}_{x\sim p_{\mathcal{X}}}[\mathbb{E}_{a\sim p_{\exp}(x)}[(f^{M}(x,a)-f^{\overline{M}}(x,a))^2+(f^{\overline{M}}(x,a)-f^{M'}(x,a))^2]]]\\
&\leq 2A\mathbb{E}_{\overline{M}\sim\xi}[D_{\mathsf{KL}}\big(\overline{M}(\pi_{\exp})\,\|\,M(\pi_{\exp})\big)+D_{\mathsf{KL}}\big(\overline{M}(\pi_{\exp})\,\|\,M'(\pi_{\exp})\big)].
\end{aligned}
$$

It follows that if

$$
\mathbb{E}_{\overline{M}\sim\xi}[D_{\mathsf{KL}}\big(\overline{M}(\pi_{\exp})\,\|\,M(\pi_{\exp})\big)] \leq \frac{\varepsilon}{4A} \quad\text{and}\quad \mathbb{E}_{\overline{M}\sim\xi}[D_{\mathsf{KL}}\big(\overline{M}(\pi_{\exp})\,\|\,M'(\pi_{\exp})\big)] \leq \frac{\varepsilon}{4A},
$$

then we can bound $\mathbb{E}_{\pi\sim p'}[D_{\mathsf{KL}}(M(\pi)\,\|\,M'(\pi))] \leq \varepsilon$. Thus, choosing $p_{\exp} \in \triangle(\Pi)$ to place probability mass 1 on $\pi_{\exp}$, a sufficient bound on $C_{\exp}(\mathcal{M},\varepsilon)$ is $4A/\varepsilon$. □

Proof of Proposition 10.8. The bound on $C_{\exp}(\mathcal{M}^\star, \varepsilon)$ follows from Lemma H.4.6. Hence, by Lemma H.2.6 we can bound $\overline{\mathsf{aec}}_\varepsilon^{\mathcal{M}}(\mathcal{M}^\star)$ as:

$$\overline{\mathsf{aec}}_\varepsilon^{\mathcal{M}}(\mathcal{M}^\star) \le \frac{4A}{\delta} \quad \text{for} \quad \delta = \min_{M \in \mathcal{M}^\star} \min \left\{ \min \left\{ \frac{1}{81 L_{\mathsf{KL}}}, \frac{\Delta_{\min}^M}{34 V_{\mathcal{M}}} \right\} \cdot \frac{\varepsilon}{2\mathsf{g}^M / \Delta_{\min}^M + \mathsf{n}_{\varepsilon/36}^M}, \frac{\Delta_{\min}^M}{3} \right\}^2.$$

By Lemma H.2.2, we have that for all $M \in \mathcal{M}^\star$,

$$\mathsf{g}^M \le C_{\exp}(\mathcal{M}^\star, \frac{1}{4}(\Delta_{\min}^M)^2) \le \frac{16A}{\Delta_\star^2}.$$

By Lemma H.4.5, we can take $L_{\mathsf{KL}} = V_{\mathcal{M}} = 2\sqrt{2}$. A sufficient choice for δ is therefore

$$\delta = c \cdot \frac{\varepsilon^2 \Delta_\star^8}{A^2}.$$

The result follows. □

H.4.4 Informative Arms (Example 10.2.1)

In this section, we provide calculations for the bandits with informative arms setting in Example 10.2.1. We first show that Assumptions 10.4 to 10.6 are satisfied.

- Lemma H.4.1 If $\pi \in [A]$, then the response is simply Gaussian, so by Lemma H.4.1, the condition of Assumption 10.4 is met with $L_{\mathsf{KL}} = 2$. If $\pi = \pi_i^\circ$, then by the Mean Value Theorem we have

$$\begin{aligned}
&\left| D_{\mathsf{KL}}(M(\pi) \,\|\, M'(\pi)) - D_{\mathsf{KL}}\big(\overline{M}(\pi) \,\|\, M'(\pi)\big) \right| \\
&\quad = \left| \sum_{a \in [A]} \mathbb{P}^{M,\pi}(a) \log \frac{\mathbb{P}^{M,\pi}(a)}{\mathbb{P}^{M',\pi}(a)} - \sum_{a \in [A]} \mathbb{P}^{\overline{M},\pi}(a) \log \frac{\mathbb{P}^{\overline{M},\pi}(a)}{\mathbb{P}^{M',\pi}(a)} \right| \\
&\quad \le \left(1 + \max_{a \in [A]} \max \left\{ \left| \log \frac{\mathbb{P}^{M,\pi}(a)}{\mathbb{P}^{M',\pi}(a)} \right|, \left| \log \frac{\mathbb{P}^{\overline{M},\pi}(a)}{\mathbb{P}^{M',\pi}(a)} \right| \right\} \right) \cdot \sum_{a \in [A]} |\mathbb{P}^{M,\pi}(a) - \mathbb{P}^{\overline{M},\pi}(a)| \\
&\quad = \left(1 + \max_{a \in [A]} \max \left\{ \left| \log \frac{\mathbb{P}^{M,\pi}(a)}{\mathbb{P}^{M',\pi}(a)} \right|, \left| \log \frac{\mathbb{P}^{\overline{M},\pi}(a)}{\mathbb{P}^{M',\pi}(a)} \right| \right\} \right) \cdot D_{\mathsf{TV}}\big(\mathbb{P}^{M,\pi}, \mathbb{P}^{\overline{M},\pi}\big) \\
&\quad \le \left(1 + \max_{a \in [A]} \max \left\{ \left| \log \frac{\mathbb{P}^{M,\pi}(a)}{\mathbb{P}^{M',\pi}(a)} \right|, \left| \log \frac{\mathbb{P}^{\overline{M},\pi}(a)}{\mathbb{P}^{M',\pi}(a)} \right| \right\} \right) \cdot \sqrt{\frac{1}{2} D_{\mathsf{KL}}(\mathbb{P}^{M,\pi} \,\|\, \mathbb{P}^{\overline{M},\pi})}.
\end{aligned}$$

  Using the bound on the log-likelihood ratio given above, this verifies that Assumption 10.4 holds with $L_{\mathsf{KL}} = \max\{2, 1 + \log \frac{A}{1-\beta}\}$.

- If $\pi \in [A]$, then the since the response is Gaussian, by Lemma H.4.1, the condition

of Assumption 10.5 is met with $V_{\mathcal{M}} = 2$. If $\pi = \pi_i^\circ$, then for $M \in \mathcal{M}$, either the observation is distributed as $1/A$, so $\mathbb{P}^{M,\pi}(r, o) = 1/A$ for all $o \in [A]$, or i is the informative arm for instance M, in which case $\mathbb{P}^{M,\pi}(r, o) = (1-\beta)/A$ for $o \neq \pi_M$, and $\mathbb{P}^{M,\pi}(r, o) = \beta + (1-\beta)/A$ for $o = \pi_M$ (note that we can disregard $o = \perp$ since it occurs with probability 0 if an informative arm is pulled). The log-likehood ratio is then at most

$$\log \frac{\beta + (1-\beta)/A}{(1-\beta)/A} \leq \log \frac{A}{1-\beta}.$$

Thus, Assumption 10.5 is satisfied with $V_{\mathcal{M}} = \max\{2, \log \frac{A}{1-\beta}\}$.

- Using Lemma H.4.1, it is easy to see that Assumption 10.6 is met with $d_{\mathsf{cov}} = O(A)$ and $C_{\mathsf{cov}} = O(1)$.

To bound the parameter $\mathsf{n}^{\mathcal{M}}$, consider $M \in \mathcal{M}$ and $M' \in \mathcal{M}^{\mathrm{alt}}(M)$. Note that either $D_{\mathsf{KL}}(M(\pi_M) \,\|\, M'(\pi_M)) = 0$, in which case there is no advantage to playing π_M, or, due to the discretization of the means, $D_{\mathsf{KL}}(M(\pi_M) \,\|\, M'(\pi_M)) \geq \frac{1}{2}\Delta_{\min}^2$. Thus, for any allocation $\eta \in \mathbb{R}_+^{\Pi}$, as long as $\eta(\pi_M) \geq \frac{2}{\Delta_{\min}^2}$, we have $\eta(\pi_M) D_{\mathsf{KL}}(M(\pi_M) \,\|\, M'(\pi_M)) \geq 1$. It follows that there is no advantage to choosing $\eta(\pi_M)$ larger than $\frac{2}{\Delta_{\min}^2}$, so we can bound $\mathsf{n}^{\mathcal{M}} \leq \frac{2}{\Delta_{\min}^2}$.

H.4.4.1 Bounding the Allocation-Estimation Coefficient

We begin with some basic observations. First, since we restrict $\mathcal{M}$ to only contain instances with a single optimal decision, if $f^M(\pi_M) = \lfloor \frac{1}{\Delta_{\min}} \rfloor \Delta_{\min}$ for some $M \in \mathcal{M}$, this implies that $f^M(\pi) < \lfloor \frac{1}{\Delta_{\min}} \rfloor \Delta_{\min}$ for all $\pi \neq \pi_M$. Fix some $M \in \mathcal{M}$ satisfying $f^M(\pi_M) = \lfloor \frac{1}{\Delta_{\min}} \rfloor \Delta_{\min}$. It follows that, for every $M' \in \mathcal{M}^{\mathrm{alt}}(M)$, it must be the case that $f^{M'}(\pi_M) < \lfloor \frac{1}{\Delta_{\min}} \rfloor \Delta_{\min}$. Therefore, since M and M' have different reward means at π_M, and since this holds for all $M' \in \mathcal{M}^{\mathrm{alt}}(M)$, M can be distinguished from every $M' \in \mathcal{M}^{\mathrm{alt}}(M)$ by playing π_M. In this case, then, $\mathsf{g}^M = 0$, so any ε-optimal Graves-Lai allocation must put all its mass on π_M, implying $\Upsilon(M, \varepsilon) = \{\mathbb{I}_{\pi_M}\}$. Denote such instances M with $f^M(\pi_M) = \lfloor \frac{1}{\Delta_{\min}} \rfloor \Delta_{\min}$ as $\overline{\mathcal{M}}$. Note that for M with $f^M(\pi_M) < \lfloor \frac{1}{\Delta_{\min}} \rfloor \Delta_{\min}$, we have $\mathbb{I}_{\pi_M^\circ} \in \Upsilon(M, \varepsilon)$.

We proceed to bound the value of the Fix $\overline{M} \in \mathrm{co}(\mathcal{M})$. For a first case, assume that that $f^{\overline{M}}(\pi_{\overline{M}}) \leq \lfloor \frac{1}{\Delta_{\min}} \rfloor \Delta_{\min} - \frac{1}{2}\Delta_{\min}$ and let $k = \arg\max_{i \in [N]} \mathbb{P}^{\overline{M}, \pi_i^\circ}(o = \pi_{\overline{M}})$ denote the index of the most informative arm for $\overline{M}$. Let $\lambda = \mathbb{I}_{\pi_k^\circ}$, and note that $\mathcal{M}_\varepsilon^{\mathrm{gl}}(\lambda)$ contains only instances in $\mathcal{M}$ that have informative arm k. Let $\mathcal{M}' = \{M \in \mathcal{M} \,:\, \pi_M^\circ \neq \pi_k^\circ\} \cup \overline{\mathcal{M}}$. Then $\mathcal{M} \backslash \mathcal{M}_\varepsilon^{\mathrm{gl}}(\lambda) \subseteq \mathcal{M}'$. Let $\omega = \frac{1}{2}\mathrm{Unif}(\{\pi_i^\circ\}_{i \in [N]}) + \frac{1}{2}\mathrm{Unif}([A])$. Then,

$$\mathsf{aec}_\varepsilon(\mathcal{M}, \overline{M}) \leq \sup_{M \in \mathcal{M}'} \frac{1}{\mathbb{E}_{\pi \sim \omega}[D_{\mathsf{KL}}\big(\overline{M}(\pi) \,\|\, M(\pi)\big)]}$$

$$\leq \sup_{M\in\mathcal{M},\pi^\circ_M\neq\pi^\circ_k} \frac{2N}{\sum_{i=1}^N D_{\mathsf{KL}}\big(\overline{M}(\pi^\circ_i)\,\|\,M(\pi^\circ_i)\big)} + \sup_{M\in\overline{\mathcal{M}}} \frac{2A}{\sum_{\pi\in[A]} D_{\mathsf{KL}}\big(\overline{M}(\pi)\,\|\,M(\pi)\big)}$$

If $\pi^\circ_{M} \neq \pi^\circ_k$, this implies that $o \sim M(\pi^\circ_i)$ is uniform on $[A]$ for $\pi^\circ_i \neq \pi^\circ_{M}$, and $o \sim M(\pi^\circ_i)$ is distributed as $\beta\mathbb{I}_{\pi_M} + (1-\beta)\mathrm{Unif}([A])$ for $\pi^\circ_i = \pi^\circ_{M}$. Note that since $k = \arg\max_{i\in[N]} \mathbb{P}^{\overline{M},\pi^\circ_i}(o = \pi_{\overline{M}})$ and $\overline{M} \in \mathrm{co}(\mathcal{M})$, we can have at most $\mathbb{P}^{\overline{M},\pi^\circ_i}(o) \leq 1/A + \beta/2$ for all o if $i \neq k$, since if this were not the case, then i must be k. It follows from Pinsker's inequality that for M with $\pi^\circ_{M} \neq \pi^\circ_k$:

$$\begin{aligned}
D_{\mathsf{KL}}\big(\overline{M}(\pi^\circ_{M})\,\|\,M(\pi^\circ_{M})\big) &\geq 2D_{\mathsf{TV}}\big(M(\pi^\circ_{M}), \overline{M}(\pi^\circ_{M})\big)^2 \\
&\geq 2|\mathbb{P}^{M,\pi^\circ_M}(o = \pi_{M}) - \mathbb{P}^{\overline{M},\pi^\circ_M}(o = \pi_{M})|^2 \\
&= 2|\beta - 1/A - \beta/2|^2 \\
&\geq 2|\beta/4|^2
\end{aligned}$$

where the last inequality uses our assumption that $\beta \geq 4/A$. For $M \in \overline{\mathcal{M}}$, since $f^{\overline{M}}(\pi_{\overline{M}}) \leq \lfloor \frac{1}{\Delta_{\min}} \rfloor \Delta_{\min} - \frac{1}{2}\Delta_{\min}$, we have that

$$D_{\mathsf{KL}}\big(\overline{M}(\pi)\,\|\,M(\pi)\big) \geq \frac{1}{8}\Delta^2_{\min}.$$

Thus, we can bound

$$\mathsf{aec}_\varepsilon(\mathcal{M}, \overline{M}) \leq \frac{64N}{\beta^2} + \frac{16A}{\Delta^2_{\min}}.$$

Now, consider the second case where $\overline{M}$ has $f^{\overline{M}}(\pi_{\overline{M}}) > \lfloor \frac{1}{\Delta_{\min}} \rfloor \Delta_{\min} - \frac{1}{2}\Delta_{\min}$. Note that in this case we must have $|\boldsymbol{\pi}_{\overline{M}}| = 1$. Set $\lambda = \mathbb{I}_{\pi_{\overline{M}}}$. Then we have that $\mathcal{M}^{\mathrm{gl}}_\varepsilon(\lambda)$ contains every instance except the single instance with $f^{M}(\pi_{\overline{M}}) = \lfloor \frac{1}{\Delta_{\min}} \rfloor \Delta_{\min}$. Let $\omega = \mathbb{I}_{\pi_{\overline{M}}}$. Note that for any instance with $f^{M}(\pi_{\overline{M}}) < \lfloor \frac{1}{\Delta_{\min}} \rfloor \Delta_{\min}$, i.e. every instance in $\mathcal{M}\backslash\mathcal{M}^{\mathrm{gl}}_\varepsilon(\lambda)$, we have

$$D_{\mathsf{KL}}\big(\overline{M}(\pi_{\overline{M}})\,\|\,M(\pi_{\overline{M}})\big) \geq \frac{1}{8}\Delta^2_{\min}.$$

It follows that with such an $\overline{M}$, we can bound

$$\mathsf{aec}_\varepsilon(\mathcal{M}, \overline{M}) \leq \frac{8}{\Delta^2_{\min}}.$$

H.4.5 Tabular Reinforcement Learning (Section 10.3.2)

In this section, we prove all of the claims in Section 10.3.2 concerning tabular reinforcement learning.

Throughout this section we let $M_{sh}(a)$ denote the joint distribution of the next state and reward if we play action a in state s at step h on MDP $M \in \mathcal{M}$. We also define

$$w_h^{M,\pi}(s,a) = \mathbb{P}^{M,\pi}(s_h = s, a_h = a)$$

as the state-action visitation probabilities on MDP M under policy π (and define $w_h^{M,\pi}(s)$ analogously). We let $r_h^M(s,a) = \mathbb{E}_{r\sim R_h^M(s,a)}[r]$ denote the mean reward on MDP M at (s,a,h), and let $r_h(s_h,a_h)$ the denote the realized (random reward) at step h. We let $r := (r_1(s_1,a_1),\dots,r_H(s_H,a_H))$ denote the vector of all random rewards in a given episode. $\tau = (s_1,\dots,s_H)$ denotes a trajectory of states, and $\tau_h = s_h$ the hth state in the trajectory. We denote the Q-value function on M for policy π by

$$Q_h^{M,\pi}(s,a) = \mathbb{E}^{M,\pi}\left[\sum_{h'=h}^{H} r_{h'}(s_{h'},a_{h'}) \mid s_h = s, a_h = a\right]$$

and the value function by $V_h^{M,\pi}(s) = \mathbb{E}_{a\sim\pi_h(s)}[Q_h^{M,\pi}(s,a)]$. We denote the value of a policy by $V_1^{M,\pi} := V_1^{M,\pi}(s_1)$. For any function $V : \mathcal{S} \to \mathbb{R}$ we denote

$$\mathbb{P}_h^M[V](s,a) = \mathbb{E}_{s'\sim P_h^M(\cdot|s,a)}[V(s')].$$

For all results in this section concerning general divergences, we take $D\big(\cdot \,\|\, \cdot\big) \leftarrow D_{\mathsf{KL}}(\cdot\,\|\,\cdot)$.

Proof of Proposition 10.9. To bound the AEC, we first move from KL divergence to Hellinger distance. Since we always have $D_{\mathsf{KL}}\big(\overline{M}(\pi)\,\|\,M(\pi)\big) \geq D^2_{\mathsf{H}}\big(\overline{M}(\pi), M(\pi)\big)$, we upper bound

$$\overline{\mathsf{aec}}_\varepsilon^{\mathcal{M}}(\mathcal{M}^\star) \leq \sup_{\xi\in\triangle_{\mathcal{M}}} \inf_{\lambda,\omega\in\triangle_\Pi} \sup_{M\in\mathcal{M}^\star\setminus\mathcal{M}_\varepsilon^{\mathrm{gl}}(\lambda)} \frac{1}{\mathbb{E}_{\overline{M}\sim\xi}[\mathbb{E}_\omega[D^2_{\mathsf{H}}\big(\overline{M}(\pi), M(\pi)\big)]]}.$$

We then apply Lemma H.2.6 to bound this by $C^{\mathsf{D}}_{\mathrm{exp}}(\mathcal{M}^\star,\delta)$, with $D\big(\cdot\,\|\,\cdot\big) \leftarrow D^2_{\mathsf{H}}(\cdot,\cdot)$. The bound on $C^{\mathsf{D}}_{\mathrm{exp}}(\mathcal{M}^\star,\varepsilon)$ then follows directly from Lemma H.4.10, and gives

$$\overline{\mathsf{aec}}_\varepsilon^{\mathcal{M}}(\mathcal{M}^\star) \leq \frac{SAH^2\cdot\log^2 H}{\delta^2} \text{ for } \delta = \min_{M\in\mathcal{M}^\star} \min\left\{ \min\left\{\frac{1}{81L_{\mathsf{KL}}}, \frac{\Delta^M_{\min}}{34V_{\mathcal{M}}}\right\}\cdot\frac{\varepsilon}{2\mathsf{g}^M/\Delta^M_{\min} + \mathsf{n}^M_{\varepsilon/36}}, \frac{\Delta^M_{\min}}{3}\right\}^2.$$

By Lemma H.4.8, we have that Assumptions 10.5 and H.1 hold with

$$L_{\mathsf{KL}} = V_{\mathcal{M}} = 4H + \max_{\overline{M},\overline{M}'\in\mathcal{M}} \max_{\pi\in\Pi}\max_{\tau\in\mathcal{T}} \left|\log\frac{\mathbb{P}^{\overline{M}',\pi}(\tau)}{\mathbb{P}^{\overline{M},\pi}(\tau)}\right|.$$

As we assume every transition has probability at least $P_{\min}$, we can lower bound $\mathbb{P}^{\overline{M},\pi}(\tau) \geq$

$P_{\min}^H$, so it suffices to take

$$L_{\mathsf{KL}} = V_{\mathcal{M}} = H(4 + \log 1/P_{\min}).$$

By Lemma H.2.2, for $M \in \mathcal{M}^\star$ we can bound

$$\mathsf{g}^M \le C_{\exp}(\mathcal{M}^\star, \tfrac{1}{4}(\Delta_{\min}^M)^2) \le c \cdot \frac{SAH^2 \cdot \log^2 H}{\Delta_\star^4}.$$

A sufficient choice of δ is therefore

$$\delta = c \cdot \frac{\varepsilon^2 \Delta_\star^{12}}{S^2A^2H^6 \cdot \log^4 H \cdot \log^2 1/P_{\min}}.$$

□

H.4.5.1 Tabular MDPs are Regular Classes

Lemma H.4.7. *If $M^\star$ is such that $r_h^{M^\star}(s,a) \in [0, 1/H^2)$ for all (s,a,h), then in the setting of $\mathcal{M} \leftarrow \mathcal{M}_{\mathrm{tab}}(P_{\min})$, $M^\star$ is a regular model with constant*

$$L_{\mathcal{M}}^\star = \frac{96}{\Delta_{\min}^\star}$$

Proof of Lemma H.4.7. Take some MDP $M' \in \mathcal{M}^{\mathrm{alt}}(M^\star)$ such that $D_{\mathsf{KL}}(M^\star(\pi_\star) \,\|\, M'(\pi_\star)) > 0$. Let M'' be such that

$$D_{\mathsf{KL}}(M_{sh}^\star(\pi_\star(s,h)) \,\|\, M_{sh}''(\pi_\star(s,h))) = 0, \quad \forall s,h$$

and

$$D_{\mathsf{KL}}(M_{sh}'(a) \,\|\, M_{sh}''(a)) = 0, \quad \forall s,h,a \neq \pi_\star(s,h),$$

so that M'' is the MDP which is identical to $M^\star$ on optimal actions, and identical to M' on suboptimal actions (recall that optimal actions for $M^\star$ are unique). By construction, we have that $D_{\mathsf{KL}}(M^\star(\pi_\star) \,\|\, M''(\pi_\star)) = 0$. Furthermore, it is not difficult to see that $M'' \in \mathcal{M}$. In particular, to verify that $P_h^{M''}(s' \mid s,a) \ge P_{\min}$ for each (s,a,h,s'), we note that since $M^\star, M' \in \mathcal{M}$, for every (s,a,h,s'), we have $P_h^{M^\star}(s' \mid s,a) \ge P_{\min}$ and $P_h^{M'}(s' \mid s,a) \ge P_{\min}$. By construction, we have that $P_h^{M''}(\cdot \mid s,a)$ is identical to either $P_h^{M^\star}(\cdot \mid s,a)$ or $P_h^{M'}(\cdot \mid s,a)$ for each (s,a,h), so it follows that $P_h^{M''}(s' \mid s,a) \ge P_{\min}$. The remaining conditions for inclusion in $\mathcal{M}$ are immediate. We consider two cases.

Case 1: $M'' \in \mathcal{M}^{\text{alt}}(M^\star)$**.** For $\pi \in \Pi$, by Lemma H.4.15 we have

$$D_{\mathsf{KL}}(M^\star(\pi) \,\|\, M'(\pi)) = \sum_{s,a,h} w_h^{M^\star,\pi}(s,a) D_{\mathsf{KL}}(M^\star_{sh}(a) \,\|\, M'_{sh}(a)),$$

and

$$\begin{aligned} D_{\mathsf{KL}}(M^\star(\pi) \,\|\, M''(\pi)) &= \sum_{s,a,h} w_h^{M^\star,\pi}(s,a) D_{\mathsf{KL}}(M^\star_{sh}(a) \,\|\, M''_{sh}(a)) \\ &= \sum_{s,a,h} w_h^{M^\star,\pi}(s,a) D_{\mathsf{KL}}(M^\star_{sh}(a) \,\|\, M'_{sh}(a)) \cdot \mathbb{I}\{a \neq \pi_\star(s,h)\} \end{aligned}$$

so that

$$\begin{aligned} &|D_{\mathsf{KL}}(M^\star(\pi) \,\|\, M'(\pi)) - D_{\mathsf{KL}}(M^\star(\pi) \,\|\, M''(\pi))| \\ &= \sum_{s,h} w_h^{M^\star,\pi}(s,\pi_\star(s,h)) D_{\mathsf{KL}}(M^\star_{sh}(\pi_\star(s,h)) \,\|\, M'_{sh}(\pi_\star(s,h))) \\ &\le \sup_{s,h} \frac{w_h^{M^\star,\pi}(s,\pi_\star(s,h))}{w_h^{M^\star,\pi_\star}(s,\pi_\star(s,h))} \cdot D_{\mathsf{KL}}(M^\star(\pi_\star) \,\|\, M'(\pi_\star)) \\ &\le \sup_{s,h} \frac{1}{w_h^{M^\star,\pi_\star}(s)} \cdot D_{\mathsf{KL}}(M^\star(\pi_\star) \,\|\, M'(\pi_\star)) \\ &\le \frac{1}{\Delta^\star_{\min}} \cdot D_{\mathsf{KL}}(M^\star(\pi_\star) \,\|\, M'(\pi_\star)) \end{aligned}$$

where the last inequality follows from Lemma H.4.13. Thus, in this case, $M^\star$ is a regular model with

$$L^\star_{\mathcal{M}} = \frac{1}{\Delta^\star_{\min}}.$$

Case 2: $M'' \notin \mathcal{M}^{\text{alt}}(M^\star)$**.** Let $(\widetilde{s}, \widetilde{a}, \widetilde{h})$ be such that $Q_{\widetilde{h}}^{M',\pi_\star}(\widetilde{s},\widetilde{a}) > Q_{\widetilde{h}}^{M',\pi_\star}(\widetilde{s},\pi_\star(\widetilde{s},\widetilde{h}))$, and note that such a tuple is guaranteed to exist by Lemma H.4.11 since $M' \in \mathcal{M}^{\text{alt}}(M^\star)$. Let $\widetilde{M}''$ denote an MDP that is identical to M'' everywhere except for at $(\widetilde{s},\widetilde{h},\widetilde{a})$, where we set $r_{\widetilde{h}}^{\widetilde{M}''}(\widetilde{s},\widetilde{a})$ so that

$$Q_{\widetilde{h}}^{\widetilde{M}'',\pi_\star}(\widetilde{s},\widetilde{a}) = Q_{\widetilde{h}}^{\widetilde{M}'',\pi_\star}(\widetilde{s},\pi_\star(\widetilde{s},\widetilde{h})) + \delta \tag{H.4.6}$$

for some $\delta > 0$ to be chosen. This will ensure $\widetilde{a}$ is the optimal action in $(\widetilde{s},\widetilde{h})$, so $\pi_{\widetilde{M}''} \neq \pi_\star$. By construction we have that $M^\star$ and $\widetilde{M}''$ behave identically on $\pi_\star$, which implies that $Q_{\widetilde{h}}^{\widetilde{M}'',\pi_\star}(\widetilde{s},\pi_\star(\widetilde{s},\widetilde{h})) = Q_{\widetilde{h}}^{M^\star,\pi_\star}(\widetilde{s},\pi_\star(\widetilde{s},\widetilde{h}))$. Furthermore, by assumption we have $r_h^{M^\star}(s,a) < 1/H^2$ for all (s,a,h), which implies $Q_{\widetilde{h}}^{M^\star,\pi_\star}(\widetilde{s},\pi_\star(\widetilde{s},\widetilde{h})) < 1/H$. As $Q_{\widetilde{h}}^{\widetilde{M}'',\pi_\star}(\widetilde{s},\widetilde{a}) = r_{\widetilde{h}}^{\widetilde{M}''}(\widetilde{s},\widetilde{a}) +$

$\mathbb{P}_{\widetilde{h}}^{\widetilde{M}''}[V_{\widetilde{h}+1}^{\widetilde{M}'',\pi_\star}](\widetilde{s},\widetilde{a}) \geq r_{\widetilde{h}}^{\widetilde{M}''}(\widetilde{s},\widetilde{a})$, it follows that for small enough δ, we can ensure (H.4.6) is met with $r_{\widetilde{h}}^{\widetilde{M}''}(\widetilde{s},\widetilde{a}) < 1/H$, so that $\widetilde{M}'' \in \mathcal{M}$.

If we can show that, for all π, $|D_{\mathsf{KL}}(M^\star(\pi) \,\|\, M'(\pi)) - D_{\mathsf{KL}}\big(M^\star(\pi) \,\|\, \widetilde{M}''(\pi)\big)|$ is bounded by some function of $D_{\mathsf{KL}}(M^\star(\pi_\star) \,\|\, M'(\pi_\star))$, we are then done. We proceed to show this. First, note that, similar to Case 1:

$$\begin{aligned}
&|D_{\mathsf{KL}}(M^\star(\pi) \,\|\, M'(\pi)) - D_{\mathsf{KL}}\big(M^\star(\pi) \,\|\, \widetilde{M}''(\pi)\big)| \\
&\qquad = \Big| \sum_{s,h} w_h^{M^\star,\pi}(s,\pi_\star(s,h)) D_{\mathsf{KL}}(M^\star_{sh}(\pi_\star(s,h)) \,\|\, M'_{sh}(\pi_\star(s,h)))\Big| \\
&\qquad\quad + w_{\widetilde{h}}^{M^\star,\pi}(\widetilde{s},\widetilde{a})\Big| D_{\mathsf{KL}}\big(M^\star_{\widetilde{s}\widetilde{h}}(\widetilde{a}) \,\|\, M'_{\widetilde{s}\widetilde{h}}(\widetilde{a})\big) - D_{\mathsf{KL}}\big(M^\star_{\widetilde{s}\widetilde{h}}(\widetilde{a}) \,\|\, \widetilde{M}''_{\widetilde{s}\widetilde{h}}(\widetilde{a})\big)\Big| \\
&\qquad \leq \sup_{s,h} \frac{1}{w_h^{M^\star,\pi_\star}(s)} \cdot D_{\mathsf{KL}}(M^\star(\pi_\star) \,\|\, M'(\pi_\star)) \\
&\qquad\quad + \frac{1}{2} w_{\widetilde{h}}^{M^\star,\pi}(\widetilde{s},\widetilde{a})\Big| (r_{\widetilde{h}}^{M^\star}(\widetilde{s},\widetilde{a}) - r_{\widetilde{h}}^{M'}(\widetilde{s},\widetilde{a}))^2 - (r_{\widetilde{h}}^{M^\star}(\widetilde{s},\widetilde{a}) - r_{\widetilde{h}}^{\widetilde{M}''}(\widetilde{s},\widetilde{a}))^2\Big|
\end{aligned} \tag{H.4.7}$$

where the inequality follows by what we showed in Case 1, and since M' and $\widetilde{M}''$ have identical transitions at $(\widetilde{s},\widetilde{a},\widetilde{h})$, so the contribution to the KL divergence from the transitions cancels, leaving only the KL divergence between unit-variance Gaussians. By the Mean Value Theorem and since rewards are in $[0,1/H]$, we have

$$\frac{1}{2}\Big|(r_{\widetilde{h}}^{M^\star}(\widetilde{s},\widetilde{a}) - r_{\widetilde{h}}^{M'}(\widetilde{s},\widetilde{a}))^2 - (r_{\widetilde{h}}^{M^\star}(\widetilde{s},\widetilde{a}) - r_{\widetilde{h}}^{\widetilde{M}''}(\widetilde{s},\widetilde{a}))^2\Big| \leq \frac{1}{H}|r_{\widetilde{h}}^{M'}(\widetilde{s},\widetilde{a}) - r_{\widetilde{h}}^{\widetilde{M}''}(\widetilde{s},\widetilde{a})|.$$

Thus, it suffices to bound $|r_{\widetilde{h}}^{M'}(\widetilde{s},\widetilde{a}) - r_{\widetilde{h}}^{\widetilde{M}''}(\widetilde{s},\widetilde{a})|$.

By assumption $Q_{\widetilde{h}}^{M',\pi_\star}(\widetilde{s},\widetilde{a}) > Q_{\widetilde{h}}^{M',\pi_\star}(\widetilde{s},\pi_\star(\widetilde{s},\widetilde{h}))$. We can then ensure

$$Q_{\widetilde{h}}^{\widetilde{M}'',\pi_\star}(\widetilde{s},\widetilde{a}) - Q_{\widetilde{h}}^{\widetilde{M}'',\pi_\star}(\widetilde{s},\pi_\star(\widetilde{s},\widetilde{h})) \leq Q_{\widetilde{h}}^{M',\pi_\star}(\widetilde{s},\widetilde{a}) - Q_{\widetilde{h}}^{M',\pi_\star}(\widetilde{s},\pi_\star(\widetilde{s},\widetilde{h}))$$

for δ sufficiently small. This is equivalent to, abbreviating $\widetilde{a}_{M^\star} := \pi_\star(\widetilde{s},\widetilde{h})$:

$$\begin{aligned}
&r_{\widetilde{h}}^{\widetilde{M}''}(\widetilde{s},\widetilde{a}) + \mathbb{P}_{\widetilde{h}}^{\widetilde{M}''}[V_{\widetilde{h}+1}^{\widetilde{M}'',\pi_\star}](\widetilde{s},\widetilde{a}) - r_{\widetilde{h}}^{\widetilde{M}''}(\widetilde{s},\widetilde{a}_{M^\star}) - \mathbb{P}_{\widetilde{h}}^{\widetilde{M}''}[V_{\widetilde{h}+1}^{\widetilde{M}'',\pi_\star}](\widetilde{s},\widetilde{a}_{M^\star}) \\
&\qquad \leq r_{\widetilde{h}}^{M'}(\widetilde{s},\widetilde{a}) + \mathbb{P}_{\widetilde{h}}^{M'}[V_{\widetilde{h}+1}^{M',\pi_\star}](\widetilde{s},\widetilde{a}) - r_{\widetilde{h}}^{M'}(\widetilde{s},\widetilde{a}_{M^\star}) - \mathbb{P}_{\widetilde{h}}^{M'}[V_{\widetilde{h}+1}^{M',\pi_\star}](\widetilde{s},\widetilde{a}_{M^\star}).
\end{aligned}$$

By construction we have that $V_h^{\widetilde{M}'',\pi_\star}(s) = V_h^{M^\star,\pi_\star}(s)$ for all (s,h), $r_{\widetilde{h}}^{\widetilde{M}''}(\widetilde{s},\widetilde{a}_{M^\star}) = r_{\widetilde{h}}^{M^\star}(\widetilde{s},\widetilde{a}_{M^\star})$, and $\mathbb{P}_{\widetilde{h}}^{\widetilde{M}''}[V_{\widetilde{h}+1}^{\widetilde{M}'',\pi_\star}](\widetilde{s},\widetilde{a}_{M^\star}) = \mathbb{P}_{\widetilde{h}}^{M^\star}[V_{\widetilde{h}+1}^{M^\star,\pi_\star}](\widetilde{s},\widetilde{a}_{M^\star})$, since $\widetilde{M}''$ behaves identically to M on actions taken by $\pi_\star$. Furthermore, we have $\mathbb{P}_{\widetilde{h}}^{\widetilde{M}''}[V_{\widetilde{h}+1}^{\widetilde{M}'',\pi_\star}](\widetilde{s},\widetilde{a}) = \mathbb{P}_{\widetilde{h}}^{M'}[V_{\widetilde{h}+1}^{M^\star,\pi_\star}](\widetilde{s},\widetilde{a})$ since $\widetilde{M}''$ behaves identically to M' on actions not taken by $\pi_\star$, other than the reward at $(\widetilde{s},\widetilde{a},\widetilde{h})$. Using these

simplifications and rearranging, we get

$$\begin{aligned}
&|r_{\widetilde{h}}^{\widetilde{M}''}(\widetilde{s},\widetilde{a}) - r_{\widetilde{h}}^{M'}(\widetilde{s},\widetilde{a})| \\
&\le |r_{\widetilde{h}}^{M'}(\widetilde{s},\widetilde{a}_{M^\star}) - r_{\widetilde{h}}^{M^\star}(\widetilde{s},\widetilde{a}_{M^\star})| + |\mathbb{P}_{\widetilde{h}}^{M^\star}[V_{\widetilde{h}+1}^{M^\star,\pi_\star}](\widetilde{s},\widetilde{a}_{M^\star}) - \mathbb{P}_{\widetilde{h}}^{M'}[V_{\widetilde{h}+1}^{M',\pi_\star}](\widetilde{s},\widetilde{a}_{M^\star})| \\
&\quad + |\mathbb{P}_{\widetilde{h}}^{M'}[V_{\widetilde{h}+1}^{M^\star,\pi_\star}](\widetilde{s},\widetilde{a}) - \mathbb{P}_{\widetilde{h}}^{M'}[V_{\widetilde{h}+1}^{M',\pi_\star}](\widetilde{s},\widetilde{a})| \\
&\le |r_{\widetilde{h}}^{M'}(\widetilde{s},\widetilde{a}_{M^\star}) - r_{\widetilde{h}}^{M^\star}(\widetilde{s},\widetilde{a}_{M^\star})| + |\mathbb{P}_{\widetilde{h}}^{M^\star}[V_{\widetilde{h}+1}^{M^\star,\pi_\star}](\widetilde{s},\widetilde{a}_{M^\star}) - \mathbb{P}_{\widetilde{h}}^{M'}[V_{\widetilde{h}+1}^{M^\star,\pi_\star}](\widetilde{s},\widetilde{a}_{M^\star})| \\
&\quad + |\mathbb{P}_{\widetilde{h}}^{M'}[V_{\widetilde{h}+1}^{M^\star,\pi_\star}](\widetilde{s},\widetilde{a}_{M^\star}) - \mathbb{P}_{\widetilde{h}}^{M'}[V_{\widetilde{h}+1}^{M',\pi_\star}](\widetilde{s},\widetilde{a}_{M^\star})| + |\mathbb{P}_{\widetilde{h}}^{M'}[V_{\widetilde{h}+1}^{M^\star,\pi_\star}](\widetilde{s},\widetilde{a}) - \mathbb{P}_{\widetilde{h}}^{M'}[V_{\widetilde{h}+1}^{M',\pi_\star}](\widetilde{s},\widetilde{a})|.
\end{aligned}$$

Since rewards are unit-variance Gaussian, we have

$$|r_{\widetilde{h}}^{M'}(\widetilde{s},\widetilde{a}_{M^\star}) - r_{\widetilde{h}}^{M^\star}(\widetilde{s},\widetilde{a}_{M^\star})| \le \sqrt{2D_{\mathsf{KL}}\big(M^\star_{\widetilde{h},\widetilde{s}}(\widetilde{a}_{M^\star}) \,\|\, M'_{\widetilde{h},\widetilde{s}}(\widetilde{a}_{M^\star})\big)} \le \sqrt{\frac{2}{w_{\widetilde{h}}^{M^\star,\pi_\star}(\widetilde{s})} D_{\mathsf{KL}}(M^\star(\pi_\star) \,\|\, M'(\pi_\star))}.$$

Since $V_{\widetilde{h}+1}^{M^\star,\pi_\star} \in [0,1]$, we have

$$\begin{aligned}
|\mathbb{P}_{\widetilde{h}}^{M^\star}[V_{\widetilde{h}+1}^{M^\star,\pi_\star}](\widetilde{s},\widetilde{a}_{M^\star}) - \mathbb{P}_{\widetilde{h}}^{M'}[V_{\widetilde{h}+1}^{M^\star,\pi_\star}](\widetilde{s},\widetilde{a}_{M^\star})| &\le \sum_{s'} |P_{\widetilde{h}}^{M}(s' \mid \widetilde{s},\widetilde{a}_{M^\star}) - P_{\widetilde{h}}^{M'}(s' \mid \widetilde{s},\widetilde{a}_{M^\star})| \\
&\le 2D_{\mathsf{TV}}\big(P_{\widetilde{h}}^{M}(\cdot \mid \widetilde{s},\widetilde{a}_{M^\star}), P_{\widetilde{h}}^{M'}(\cdot \mid \widetilde{s},\widetilde{a}_{M^\star})\big) \\
&\le \sqrt{2D_{\mathsf{KL}}\big(P_{\widetilde{h}}^{M}(\cdot \mid \widetilde{s},\widetilde{a}_{M^\star}) \,\|\, P_{\widetilde{h}}^{M'}(\cdot \mid \widetilde{s},\widetilde{a}_{M^\star})\big)} \\
&\le \sqrt{\frac{2}{w_{\widetilde{h}}^{M^\star,\pi_\star}(\widetilde{s})} D_{\mathsf{KL}}\big(P_{\widetilde{h}}^{M}(\cdot \mid \widetilde{s},\widetilde{a}_{M^\star}) \,\|\, P_{\widetilde{h}}^{M'}(\cdot \mid \widetilde{s},\widetilde{a}_{M^\star})\big)} \\
&\le \sqrt{\frac{2}{w_{\widetilde{h}}^{M^\star,\pi_\star}(\widetilde{s})} D_{\mathsf{KL}}(M^\star(\pi_\star) \,\|\, M'(\pi_\star))}.
\end{aligned}$$

By Lemma H.4.12 we have

$$\begin{aligned}
|\mathbb{P}_{\widetilde{h}}^{M'}[V_{\widetilde{h}+1}^{M^\star,\pi_\star}](\widetilde{s},\widetilde{a}_{M^\star}) - \mathbb{P}_{\widetilde{h}}^{M'}[V_{\widetilde{h}+1}^{M',\pi_\star}](\widetilde{s},\widetilde{a}_{M^\star})| &\le \sum_{s'} P_{\widetilde{h}}^{M'}(s' \mid \widetilde{s},\widetilde{a}_{M^\star}) |V_{\widetilde{h}+1}^{M^\star,\pi_\star}(s') - V_{\widetilde{h}+1}^{M',\pi_\star}(s')| \\
&\le \sum_{s'} P_{\widetilde{h}}^{M'}(s' \mid \widetilde{s},\widetilde{a}_{M^\star}) \cdot \sqrt{\frac{8H}{w_{\widetilde{h}+1}^{M^\star,\pi_\star}(s')} \cdot D_{\mathsf{KL}}(M^\star(\pi_\star) \,\|\, M'(\pi_\star))} \\
&\le \sup_s \sqrt{\frac{8H}{w_{\widetilde{h}+1}^{M^\star,\pi_\star}(s)} \cdot D_{\mathsf{KL}}(M^\star(\pi_\star) \,\|\, M'(\pi_\star))}
\end{aligned}$$

and similarly

$$|\mathbb{P}_{\widetilde{h}}^{M'}[V_{\widetilde{h}+1}^{M^\star,\pi_\star}](\widetilde{s},\widetilde{a}) - \mathbb{P}_{\widetilde{h}}^{M'}[V_{\widetilde{h}+1}^{M',\pi_\star}](\widetilde{s},\widetilde{a})| \le \sup_s \sqrt{\frac{8H}{w_{\widetilde{h}+1}^{M^\star,\pi_\star}(s)} \cdot D_{\mathsf{KL}}(M^\star(\pi_\star) \,\|\, M'(\pi_\star))}.$$

Altogether then:

$$\begin{aligned}|r_{\widetilde{h}}^{\widetilde{M}''}(\widetilde{s},\widetilde{a}) - r_{\widetilde{h}}^{M'}(\widetilde{s},\widetilde{a})| &\le \left(\sqrt{\frac{8}{w_{\widetilde{h}}^{M^\star,\pi_\star}(\widetilde{s})}} + \sup_s \sqrt{\frac{32H}{w_{\widetilde{h}+1}^{M^\star,\pi_\star}(s)}}\right) \cdot \sqrt{D_{\mathsf{KL}}(M^\star(\pi_\star) \,\|\, M'(\pi_\star))} \\ &\le \sup_{s,h} \sqrt{\frac{96H}{w_h^{M^\star,\pi_\star}(s)}} \cdot \sqrt{D_{\mathsf{KL}}(M^\star(\pi_\star) \,\|\, M'(\pi_\star))}.\end{aligned}$$

Combining this with (H.4.7), we have

$$\begin{aligned}&|D_{\mathsf{KL}}(M^\star(\pi) \,\|\, M'(\pi)) - D_{\mathsf{KL}}\big(M^\star(\pi) \,\|\, \widetilde{M}''(\pi)\big)| \\ &\le \sup_{s,h} \frac{1}{w_h^{M^\star,\pi_\star}(s)} \cdot D_{\mathsf{KL}}(M^\star(\pi_\star) \,\|\, M'(\pi_\star)) + \sup_{s,h} \sqrt{\frac{96}{H w_h^{M^\star,\pi_\star}(s)}} \cdot \sqrt{D_{\mathsf{KL}}(M^\star(\pi_\star) \,\|\, M'(\pi_\star))} \\ &\le \frac{1}{\Delta_{\min}^\star} \cdot D_{\mathsf{KL}}(M^\star(\pi_\star) \,\|\, M'(\pi_\star)) + \sqrt{\frac{96}{H\Delta_{\min}^\star}} \cdot \sqrt{D_{\mathsf{KL}}(M^\star(\pi_\star) \,\|\, M'(\pi_\star))}\end{aligned}$$

where the second inequality uses Lemma H.4.13. Thus, in this case $M^\star$ is a regular model with

$$L_{\mathcal{M}}^\star = \frac{96}{\Delta_{\min}^\star}.$$

□

H.4.5.2 Tabular MDPs Satisfy Basic Assumptions

Lemma H.4.8. *Tabular MDPs with unit-variance Gaussian rewards satisfy Assumptions 10.5, H.1 and H.2 with*

$$L_{\mathsf{KL}} = V_{\mathcal{M}} = 8H + \max_{\overline{M},\overline{M}' \in \mathcal{M}} \max_{\pi \in \Pi} \max_{\tau \in \mathcal{T}} \left|\log \frac{\mathbb{P}^{\overline{M}',\pi}(\tau)}{\mathbb{P}^{\overline{M},\pi}(\tau)}\right|,$$

and $D(\cdot \,\|\, \cdot) = D_{\mathsf{KL}}(\cdot \,\|\, \cdot)$, where $\mathcal{T} := \mathcal{S}^H$ and $\mathbb{P}^{\overline{M},\pi}(\tau)$ denotes the probability of observing state sequence $\tau \in \mathcal{T}$ on $\overline{M}$ when playing policy π.

Proof of Lemma H.4.8. We verify each assumption separately.

Verifying Assumption H.1. Fix some $M, M', \overline{M} \in \mathcal{M}$. Our goal is to bound

$$\left|D_{\mathsf{KL}}(M(\pi) \,\|\, M'(\pi)) - D_{\mathsf{KL}}\big(\overline{M}(\pi) \,\|\, M'(\pi)\big)\right|.$$

Let $\widetilde{M}$ denote the MDP with transitions identical to M but rewards identical to $\overline{M}$. Then

$$\begin{aligned}\left|D_{\mathsf{KL}}(M(\pi) \,\|\, M'(\pi)) - D_{\mathsf{KL}}\big(\overline{M}(\pi) \,\|\, M'(\pi)\big)\right| &\le \left|D_{\mathsf{KL}}(M(\pi) \,\|\, M'(\pi)) - D_{\mathsf{KL}}\big(\widetilde{M}(\pi) \,\|\, M'(\pi)\big)\right| \\ &\quad + \left|D_{\mathsf{KL}}\big(\widetilde{M}(\pi) \,\|\, M'(\pi)\big) - D_{\mathsf{KL}}\big(\overline{M}(\pi) \,\|\, M'(\pi)\big)\right|\end{aligned}$$

We bound these terms separately. First, by Lemma H.4.15 we have

$$\begin{aligned}&D_{\mathsf{KL}}(M(\pi) \,\|\, M'(\pi)) = \sum_{s,a,h} w_h^{M,\pi}(s,a) D_{\mathsf{KL}}(M_{sh}(\pi(s,h)) \,\|\, M'_{sh}(\pi(s,h))) \\ &= \sum_{s,a,h} w_h^{M,\pi}(s,a)\left[\frac{1}{2}(r_h^M(s,a) - r_h^{M'}(s,a))^2 + D_{\mathsf{KL}}\big(P_h^M(\cdot \mid s,\pi(s,h)) \,\|\, P_h^{M'}(\cdot \mid s,\pi(s,h))\big)\right]\end{aligned}$$

and, given our definition of $\widetilde{M}$,

$$\begin{aligned}&D_{\mathsf{KL}}\big(\widetilde{M}(\pi) \,\|\, M'(\pi)\big) \\ &= \sum_{s,a,h} w_h^{M,\pi}(s,a)\left[\frac{1}{2}(r_h^{\overline{M}}(s,a) - r_h^{M'}(s,a))^2 + D_{\mathsf{KL}}\big(P_h^M(\cdot \mid s,\pi(s,h)) \,\|\, P_h^{M'}(\cdot \mid s,\pi(s,h))\big)\right].\end{aligned}$$

Thus,

$$\begin{aligned}&\left|D_{\mathsf{KL}}(M(\pi) \,\|\, M'(\pi)) - D_{\mathsf{KL}}\big(\widetilde{M}(\pi) \,\|\, M'(\pi)\big)\right| \\ &\quad= \left|\frac{1}{2}\sum_{s,a,h} w_h^{M,\pi}(s,a)\big[(r_h^M(s,a) - r_h^{M'}(s,a))^2 - (r_h^{\overline{M}}(s,a) - r_h^{M'}(s,a))^2\big]\right| \\ &\quad\overset{(a)}{\le} \sum_{s,a,h} w_h^{M,\pi}(s,a)|r_h^M(s,a) - r_h^{\overline{M}}(s,a)| \\ &\quad\le \sum_{s,a,h} w_h^{M,\pi}(s,a)\sqrt{2D_{\mathsf{KL}}\big(M_{sh}(a) \,\|\, \overline{M}_{sh}(a)\big)} \\ &\quad\le \sqrt{2H \cdot \sum_{s,a,h} w_h^{M,\pi}(s,a) D_{\mathsf{KL}}\big(M_{sh}(a) \,\|\, \overline{M}_{sh}(a)\big)} \\ &\quad= \sqrt{2H \cdot D_{\mathsf{KL}}\big(M(\pi) \,\|\, \overline{M}(\pi)\big)},\end{aligned}$$

where (a) holds by the Mean Value Theorem and the assumption that reward means are in $[0,1]$, and the final equality holds by Lemma H.4.15.

We turn now to bounding the second term. Let $\mathcal{T} = \mathcal{S}^H$ denote the space of all possible state trajectories. Let $\mathbb{P}^{M,\pi}(\tau = \cdot)$ denote the probability of observing $\tau \in \mathcal{T}$ when playing policy π on M. We then have

$$\begin{aligned}
D_{\mathsf{KL}}\big(\widetilde{M}(\pi) \,\|\, M'(\pi)\big) &= \int \log \frac{\mathbb{P}^{\widetilde{M},\pi}(r,\tau)}{\mathbb{P}^{M',\pi}(r,\tau)} \mathrm{d}\mathbb{P}^{\widetilde{M},\pi}(r,\tau) \\
&= \int\int \log \frac{\mathbb{P}^{\overline{M},\pi}(r \mid \tau)\mathbb{P}^{M,\pi}(\tau)}{\mathbb{P}^{M',\pi}(r \mid \tau)\mathbb{P}^{M',\pi}(\tau)} \mathrm{d}\mathbb{P}^{\overline{M},\pi}(r \mid \tau)\mathrm{d}\mathbb{P}^{M,\pi}(\tau) \\
&= \int \log \frac{\mathbb{P}^{M,\pi}(\tau)}{\mathbb{P}^{M',\pi}(\tau)} \mathrm{d}\mathbb{P}^{M,\pi}(\tau) + \int \left(\int \log \frac{\mathbb{P}^{\overline{M},\pi}(r \mid \tau)}{\mathbb{P}^{M',\pi}(r \mid \tau)} \mathrm{d}\mathbb{P}^{\overline{M},\pi}(r \mid \tau) \right) \mathrm{d}\mathbb{P}^{M,\pi}(\tau) \\
&= \sum_{\tau \in \mathcal{T}} \mathbb{P}^{M,\pi}(\tau) \log \frac{\mathbb{P}^{M,\pi}(\tau)}{\mathbb{P}^{M',\pi}(\tau)} + \sum_{\tau \in \mathcal{T}} \mathbb{P}^{M,\pi}(\tau) D_{\mathsf{KL}}\big(\mathbb{P}^{\overline{M},\pi}(r \mid \tau) \,\|\, \mathbb{P}^{M',\pi}(r \mid \tau)\big).
\end{aligned}$$

It follows that

$$\begin{aligned}
&\Big| D_{\mathsf{KL}}\big(\widetilde{M}(\pi) \,\|\, M'(\pi)\big) - D_{\mathsf{KL}}\big(\overline{M}(\pi) \,\|\, M'(\pi)\big) \Big| \\
&\leq \left| \sum_{\tau \in \mathcal{T}} \mathbb{P}^{M,\pi}(\tau) \log \frac{\mathbb{P}^{M,\pi}(\tau)}{\mathbb{P}^{M',\pi}(\tau)} - \sum_{\tau \in \mathcal{T}} \mathbb{P}^{\overline{M},\pi}(\tau) \log \frac{\mathbb{P}^{\overline{M},\pi}(\tau)}{\mathbb{P}^{M',\pi}(\tau)} \right| \\
&\quad + \sum_{\tau \in \mathcal{T}} |\mathbb{P}^{M,\pi}(\tau) - \mathbb{P}^{\overline{M},\pi}(\tau)| D_{\mathsf{KL}}\big(\mathbb{P}^{M,\pi}(r \mid \tau) \,\|\, \mathbb{P}^{M',\pi}(r \mid \tau)\big).
\end{aligned}$$

Note that, since rewards at each state are independent,

$$D_{\mathsf{KL}}\big(\mathbb{P}^{M,\pi}(r \mid \tau) \,\|\, \mathbb{P}^{M',\pi}(r \mid \tau)\big) = \sum_{h=1}^{H} D_{\mathsf{KL}}\big(\mathbb{P}^{M,\pi}(r_h \mid \tau) \,\|\, \mathbb{P}^{M',\pi}(r_h \mid \tau)\big) \leq H,$$

since rewards have means are in $[0,1]$ and are unit Gaussian. This implies

$$\begin{aligned}
\sum_{\tau \in \mathcal{T}} |\mathbb{P}^{M,\pi}(\tau) - \mathbb{P}^{\overline{M},\pi}(\tau)| D_{\mathsf{KL}}\big(\mathbb{P}^{M,\pi}(r \mid \tau) \,\|\, \mathbb{P}^{M',\pi}(r \mid \tau)\big) &\leq H \sum_{\tau \in \mathcal{T}} |\mathbb{P}^{M,\pi}(\tau) - \mathbb{P}^{\overline{M},\pi}(\tau)| \\
&= H D_{\mathsf{TV}}\big(M(\pi), \overline{M}(\pi)\big) \\
&\leq H \sqrt{\frac{1}{2} D_{\mathsf{KL}}\big(M(\pi) \,\|\, \overline{M}(\pi)\big)}.
\end{aligned}$$

Note that $\frac{\mathrm{d}}{\mathrm{d}x} x \log \frac{x}{y} = 1 + \log \frac{x}{y}$, so by the Mean Value Theorem we have

$$\begin{aligned}
&\left| \mathbb{P}^{M,\pi}(\tau) \log \frac{\mathbb{P}^{M,\pi}(\tau)}{\mathbb{P}^{M',\pi}(\tau)} - \mathbb{P}^{\overline{M},\pi}(\tau) \log \frac{\mathbb{P}^{\overline{M},\pi}(\tau)}{\mathbb{P}^{M',\pi}(\tau)} \right| \\
&\quad \le \left(1 + \max \left\{ \left| \log \frac{\mathbb{P}^{M,\pi}(\tau)}{\mathbb{P}^{M',\pi}(\tau)} \right|, \left| \log \frac{\mathbb{P}^{\overline{M},\pi}(\tau)}{\mathbb{P}^{M',\pi}(\tau)} \right| \right\} \right) \cdot |\mathbb{P}^{M,\pi}(\tau) - \mathbb{P}^{\overline{M},\pi}(\tau)| \\
&\quad \le \left(1 + \max_{\overline{M}' \in \mathcal{M}} \max_{\tau' \in \mathcal{T}} \left| \log \frac{\mathbb{P}^{\overline{M}',\pi}(\tau')}{\mathbb{P}^{M',\pi}(\tau')} \right| \right) \cdot |\mathbb{P}^{M,\pi}(\tau) - \mathbb{P}^{\overline{M},\pi}(\tau)|.
\end{aligned}$$

It follows that

$$\begin{aligned}
&\left| \sum_{\tau \in \mathcal{T}} \mathbb{P}^{M,\pi}(\tau) \log \frac{\mathbb{P}^{M,\pi}(\tau)}{\mathbb{P}^{M',\pi}(\tau)} - \sum_{\tau \in \mathcal{T}} \mathbb{P}^{\overline{M},\pi}(\tau) \log \frac{\mathbb{P}^{\overline{M},\pi}(\tau)}{\mathbb{P}^{M',\pi}(\tau)} \right| \\
&\quad \le \left(1 + \max_{\overline{M}' \in \mathcal{M}} \max_{\tau' \in \mathcal{T}} \left| \log \frac{\mathbb{P}^{\overline{M}',\pi}(\tau')}{\mathbb{P}^{M',\pi}(\tau')} \right| \right) \cdot \sum_{\tau \in \mathcal{T}} |\mathbb{P}^{M,\pi}(\tau) - \mathbb{P}^{\overline{M},\pi}(\tau)| \\
&\quad = \left(1 + \max_{\overline{M}' \in \mathcal{M}} \max_{\tau' \in \mathcal{T}} \left| \log \frac{\mathbb{P}^{\overline{M}',\pi}(\tau')}{\mathbb{P}^{M',\pi}(\tau')} \right| \right) \cdot D_{\mathsf{TV}}\big(M(\pi), \overline{M}(\pi)\big) \\
&\quad \le \left(1 + \max_{\overline{M}' \in \mathcal{M}} \max_{\tau' \in \mathcal{T}} \left| \log \frac{\mathbb{P}^{\overline{M}',\pi}(\tau')}{\mathbb{P}^{M',\pi}(\tau')} \right| \right) \cdot \sqrt{\frac{1}{2} D_{\mathsf{KL}}\big(M(\pi) \,\|\, \overline{M}(\pi)\big)}.
\end{aligned}$$

This verifies Assumption H.1 with

$$L_{\mathsf{KL}} = 1 + \sqrt{2H} + H + \max_{M' \in \mathcal{M}, \overline{M}' \in \mathcal{M}} \max_{\pi \in \Pi} \max_{\tau \in \mathcal{T}} \left| \log \frac{\mathbb{P}^{\overline{M}',\pi}(\tau)}{\mathbb{P}^{M',\pi}(\tau)} \right|.$$

Verifying Assumption H.2. That $D^2_{\mathsf{H}}\big(\overline{M}(\pi), \overline{M}'(\pi)\big) \le D\big(\overline{M}(\pi) \,\|\, \overline{M}'(\pi)\big)$ is immediate, since KL always upper bounds Hellinger squared.

Verifying Assumption 10.5. We have

$$\log \frac{\mathbb{P}^{\overline{M},\pi}(r,o)}{\mathbb{P}^{M,\pi}(r,o)} = \log \frac{\mathbb{P}^{\overline{M},\pi}(\tau)}{\mathbb{P}^{M,\pi}(\tau)} + \log \frac{\mathbb{P}^{\overline{M},\pi}(r \mid \tau)}{\mathbb{P}^{M,\pi}(r \mid \tau)} = \log \frac{\mathbb{P}^{\overline{M},\pi}(\tau)}{\mathbb{P}^{M,\pi}(\tau)} + \sum_{h=1}^{H} \log \frac{\mathbb{P}^{\overline{M},\pi}(r_h \mid \tau)}{\mathbb{P}^{M,\pi}(r_h \mid \tau)}.$$

Using the same calculation as in Lemma H.4.1, we have that $\log \frac{\mathbb{P}^{\overline{M},\pi}(r_h|\tau)}{\mathbb{P}^{M,\pi}(r_h|\tau)}$ is 8-subgaussian, since rewards are unit-variance Gaussian. As $\log \frac{\mathbb{P}^{\overline{M},\pi}(r_h|\tau)}{\mathbb{P}^{M,\pi}(r_h|\tau)}$ and $\log \frac{\mathbb{P}^{\overline{M},\pi}(r_{h'}|\tau)}{\mathbb{P}^{M,\pi}(r_{h'}|\tau)}$ are independent for $h \neq h'$, it follows that $\log \frac{\mathbb{P}^{\overline{M},\pi}(r|\tau)}{\mathbb{P}^{M,\pi}(r|\tau)}$ is $8H$-subgaussian.

Furthermore, bounding

$$\log \frac{\mathbb{P}^{\overline{M},\pi}(\tau)}{\mathbb{P}^{M,\pi}(\tau)} \le \sup_{\overline{M},M\in\mathcal{M}} \sup_{\pi\in\Pi} \sup_{\tau\in\mathcal{T}} \left|\log \frac{\mathbb{P}^{\overline{M},\pi}(\tau)}{\mathbb{P}^{M,\pi}(\tau)}\right| =: V_{\mathcal{T}},$$

we have that $\log \frac{\mathbb{P}^{\overline{M},\pi}(\tau)}{\mathbb{P}^{M,\pi}(\tau)}$ is $V_{\mathcal{T}}^2$-subgaussian. Since the sum of subgaussian random variables is subgaussian, it follows that $\log \frac{\mathbb{P}^{\overline{M},\pi}(r,o)}{\mathbb{P}^{M,\pi}(r,o)} = \log \frac{\mathbb{P}^{\overline{M},\pi}(\tau)}{\mathbb{P}^{M,\pi}(\tau)} + \log \frac{\mathbb{P}^{\overline{M},\pi}(r|\tau)}{\mathbb{P}^{M,\pi}(r|\tau)}$ is $(V_{\mathcal{T}}^2 + 8H)$-subgaussian, which verifies Assumption 10.5.

□

Lemma H.4.9. *Let $P_{\min} := \inf_{M\in\mathcal{M}} \inf_{h,s',s,a} P_h^M(s' \mid s,a)$ and assume $\mathcal{M}$ is such that $P_{\min} > 0$. We can construct a (ρ,μ)-cover of $\mathcal{M}$ with respect to $\mathcal{E} := \{|r_h| \le 1 + \sqrt{2\log(2H/\mu)}, \forall h \in [H]\}$, with*

$$\mathsf{N}_{\mathrm{cov}}(\mathcal{M},\rho,\mu) \le \frac{1}{\min\{\frac{\rho P_{\min}}{4H}, 2P_{\min}\}^{S^2AH}} \cdot \frac{(2H(2+\sqrt{\log(H/\mu)}))^{SAH}}{\rho^{SAH}}.$$

Proof of Lemma H.4.9. Throughout this proof, we use $M = \{(P_h^M)_{h=1}^H, (r_h^M)_{h=1}^H\} \in \mathcal{M}$ to denote the MDP in $\mathcal{M}$ with $R_h^M(s,a) = \mathcal{N}(r_h^M(s,a),1)$; for brevity, we $\mathcal{S}$, $\mathcal{A}$, and s_1 to be fixed and the dependence on them.

Observe that for any models $M, M' \in \mathcal{M}$, we have

$$\begin{aligned}\left|\log \frac{\mathbb{P}^{M,\pi}(r,o)}{\mathbb{P}^{M',\pi}(r,o)}\right| &\le \left|\log \frac{\mathbb{P}^{M,\pi}(\tau)}{\mathbb{P}^{M',\pi}(\tau)}\right| + \left|\log \frac{\mathbb{P}^{M,\pi}(r \mid \tau)}{\mathbb{P}^{M',\pi}(r \mid \tau)}\right| \\ &= \sum_{h=1}^H \left|\log \frac{P_h^M(\tau_{h+1} \mid \tau_h, \pi(\tau_h,h))}{P_h^{M'}(\tau_{h+1} \mid \tau_h, \pi(\tau_h,h))}\right| + \sum_{h=1}^H \left|\log \frac{\mathbb{P}^{M,\pi}(r_h \mid \tau)}{\mathbb{P}^{M',\pi}(r_h \mid \tau)}\right|.\end{aligned}$$

Let $\mathcal{I}_\varepsilon = \{\varepsilon, 2\varepsilon, \dots, \lfloor 1/\varepsilon \rfloor \varepsilon\}$, so that $|\mathcal{I}_\varepsilon| \le 1/\varepsilon$. Let $\mathcal{P}_\varepsilon$ denote an ε cover of $\triangle_{\mathcal{S}}$ in the ℓ_∞-norm, so that for any $P \in \triangle_{\mathcal{S}}$, there exists some $P' \in \mathcal{P}_\varepsilon$ such that $\sup_{s\in\mathcal{S}} |P_s - P'_s| \le \varepsilon$. It suffices to choose $\mathcal{P}_\varepsilon = \mathcal{I}_\varepsilon^S \cap \triangle_{\mathcal{S}}$, so we can bound $|\mathcal{P}_\varepsilon| \le 1/\varepsilon^S$. Let

$$\mathcal{M}_{\mathrm{cov}} = \left\{M = \{(P_h^M)_{h=1}^H, (r_h^M)_{h=1}^H\} \;:\; P_h^M(\cdot \mid s,a) \in \mathcal{P}_{\varepsilon_1}, r_h^M(s,a) \in \mathcal{I}_{\varepsilon_2}, \quad \forall s,a,h\right\}$$

for parameters $\varepsilon_1, \varepsilon_2 > 0$ to be chosen. Then

$$\mathcal{M}_{\mathrm{cov}} = (|\mathcal{P}_{\varepsilon_1}||\mathcal{I}_{\varepsilon_2}|)^{SAH} \le \frac{1}{\varepsilon_1^{S^2AH}} \cdot \frac{1}{\varepsilon_2^{SAH}}.$$

We will show that $\mathcal{M}_{\mathrm{cov}}$ is a (ρ,μ)-cover of $\mathcal{M}$ for appropriately chosen $\mathcal{E}$ and $\varepsilon_1, \varepsilon_2 > 0$.

Let $\mathcal{E} := \{|r_h| \le 1 + \sqrt{2\log(2H/\mu)}, \forall h \in [H]\}$. As we assume rewards are unit-variance Gaussian and have means in $[0,1]$, it is straightforward to see that $\mathbb{P}[\mathcal{E}^c] \le \mu$. Fix M and let

$M' \in \mathcal{M}_{\mathrm{cov}}$ denote the instance such that

$$|r_h^M(s,a) - r_h^{M'}(s,a)| \le \varepsilon_2 \quad \text{and} \quad \sup_{s' \in \mathcal{S}} |P_h^M(s' \mid s,a) - P_h^{M'}(s' \mid s,a)| \le \varepsilon_1, \quad \forall s,a,h.$$

Note that such an instance is guaranteed to exist by definition of $\mathcal{M}_{\mathrm{cov}}$.

By a similar argument as in Lemma H.4.1, we can bound, on $\mathcal{E}$,

$$\begin{aligned}
\sum_{h=1}^H \left| \log \frac{\mathbb{P}^{M,\pi}(r_h \mid \tau)}{\mathbb{P}^{M',\pi}(r_h \mid \tau)} \right| &\le \sum_{h=1}^H (1 + |r_h|) \cdot \sup_{s,a} |r_h^M(s,a) - r_h^{M'}(s,a)| \\
&\le H(2 + \sqrt{2\log(2H/\mu)}) \cdot \sup_{s,a,h} |r_h^M(s,a) - r_h^{M'}(s,a)| \\
&\le H(2 + \sqrt{2\log(2H/\mu)}) \cdot \varepsilon_2.
\end{aligned}$$

We also have

$$\begin{aligned}
\sum_{h=1}^H \left| \log \frac{P_h^M(\tau_{h+1} \mid \tau_h, \pi(\tau_h,h))}{P_h^{M'}(\tau_{h+1} \mid \tau_h, \pi(\tau_h,h))} \right| &\le H \cdot \sup_{h,s',s,a} \left| \log \frac{P_h^M(s' \mid s,a)}{P_h^{M'}(s' \mid s,a)} \right| \\
&\le H \cdot \sup_{|x| \le \varepsilon_1} \sup_{h,s',s,a} \left| \log \frac{P_h^M(s' \mid s,a)}{P_h^M(s' \mid s,a) - x} \right| \\
&\le H \cdot \sup_{h,s',s,a} \frac{\varepsilon_1}{P_h^M(s' \mid s,a) - \varepsilon_1}
\end{aligned}$$

where the last inequality holds as long as $\inf_{h,s',s,a} P_h^M(s' \mid s,a) - \varepsilon_1 > 0$. Denoting $P_{\min} := \inf_{M \in \mathcal{M}} \inf_{h,s',s,a} P_h^M(s' \mid s,a)$, for $\mathcal{M}_{\mathrm{cov}}$ to be a (ρ,μ)-cover, it therefore suffices that

$$H(2 + \sqrt{2\log(2H/\mu)}) \cdot \varepsilon_2 \le \rho/2, \quad \frac{2H\varepsilon_1}{P_{\min}} \le \rho/2, \quad \text{and} \quad P_{\min} \ge \varepsilon_1/2$$

so it suffices to take

$$\varepsilon_1 = \min\{\frac{\rho P_{\min}}{4H}, 2P_{\min}\} \quad \text{and} \quad \varepsilon_2 = \frac{\rho}{2H(2 + \sqrt{\log(H/\mu)})}.$$

The result now follows from our bound on $|\mathcal{M}_{\mathrm{cov}}|$. □

H.4.5.3 Tabular MDPs have Bounded Uniform Exploration Coefficient

Lemma H.4.10. *For the tabular MDP class $\mathcal{M}$ in (10.3.3), we can bound, for all $\varepsilon > 0$,*

$$C_{\mathrm{exp}}^{\mathsf{D}}(\mathcal{M}, \varepsilon) \le \frac{320000SAH^2 \cdot \log^2 H}{\varepsilon^2}.$$

for $D(\cdot \,\|\, \cdot) \leftarrow D^2_{\mathsf{H}}(\cdot,\cdot)$.

Proof of Lemma H.4.10. Let $\xi \in \Delta(\mathcal{M})$ be given. Define

$$p_{\mathrm{exp}} = \arg\min_{p\in\triangle_\Pi} \max_{q\in\triangle_\Pi} \sum_{s,a,h} \mathbb{E}_{\pi\sim q}\left[\mathbb{E}_{\overline{M}\sim\xi}\left[\frac{(w_h^{\overline{M},\pi}(s,a))^2}{\mathbb{E}_{\pi'\sim p}[w_h^{\overline{M},\pi'}(s,a)]}\right]\right].$$

We first show that, for any $M \in \mathcal{M}$ and any π,

$$\mathbb{E}_{\overline{M}\sim\xi}[D^2_{\mathsf{H}}(\overline{M}(\pi), M(\pi))] \le \sqrt{SAH^2 \cdot \mathbb{E}_{\overline{M}\sim\xi}[\mathbb{E}_{\pi\sim p_{\mathrm{exp}}}[D^2_{\mathsf{H}}(\overline{M}(\pi), M(\pi))]]}.$$

Consider any policy π. We can bound

$$\begin{aligned}
&\mathbb{E}_{\overline{M}\sim\xi}[D^2_{\mathsf{H}}(\overline{M}(\pi), M(\pi))] \\
&\overset{(a)}{\le} 100\log(H)\cdot \sum_{s,a,h} \mathbb{E}_{\overline{M}\sim\xi}\left[w_h^{\overline{M},\pi}(s,a) D^2_{\mathsf{H}}(\overline{M}_{sh}(a), M_{sh}(a))\right] \\
&\overset{(b)}{\le} 100\log(H)\cdot \sqrt{\sum_{s,a,h} \mathbb{E}_{\overline{M}\sim\xi}\left[\frac{w_h^{\overline{M},\pi}(s,a)^2}{\mathbb{E}_{\pi'\sim p_{\mathrm{exp}}}[w_h^{\overline{M},\pi'}(s,a)]}\right]} \\
&\qquad \cdot \sqrt{\sum_{s,a,h} \mathbb{E}_{\pi'\sim p_{\mathrm{exp}}}\left[\mathbb{E}_{\overline{M}\sim\xi}\left[w_h^{\overline{M},\pi'}(s,a) D^4_{\mathsf{H}}(\overline{M}_{sh}(a), M_{sh}(a))\right]\right]} \\
&\overset{(c)}{\le} 200\log(H)\cdot \sqrt{\sum_{s,a,h} \mathbb{E}_{\overline{M}\sim\xi}\left[\frac{w_h^{\overline{M},\pi}(s,a)^2}{\mathbb{E}_{\pi'\sim p_{\mathrm{exp}}}[w_h^{\overline{M},\pi'}(s,a)]}\right]} \\
&\qquad \cdot \sqrt{\sum_{s,a,h} \mathbb{E}_{\pi'\sim p_{\mathrm{exp}}}\left[\mathbb{E}_{\overline{M}\sim\xi}\left[w_h^{\overline{M},\pi'}(s,a) D^2_{\mathsf{H}}(\overline{M}_{sh}(a), M_{sh}(a))\right]\right]}
\end{aligned}$$

where (a) follows from Lemma A.13 of Foster et al. [2021], (b) follows from Cauchy-Schwarz and Jensen's inequality, and (c) follows because the Hellinger distance is always bounded by 2. Now note that, by definition of p_{exp}, we have

$$\sum_{s,a,h} \mathbb{E}_{\overline{M}\sim\xi}\left[\frac{w_h^{\overline{M},\pi}(s,a)^2}{\mathbb{E}_{\pi'\sim p_{\mathrm{exp}}}[w_h^{\overline{M},\pi'}(s,a)]}\right] \le \min_{p\in\triangle_\Pi}\max_{q\in\triangle_\Pi} \sum_{s,a,h} \mathbb{E}_{\pi\sim q}\left[\mathbb{E}_{\overline{M}\sim\xi}\left[\frac{(w_h^{\overline{M},\pi}(s,a))^2}{\mathbb{E}_{\pi'\sim p}[w_h^{\overline{M},\pi'}(s,a)]}\right]\right]$$

and by the minimax theorem, we can bound

$$\min_{p\in\triangle_\Pi}\max_{q\in\triangle_\Pi} \sum_{s,a,h} \mathbb{E}_{\pi\sim q}\left[\mathbb{E}_{\overline{M}\sim\xi}\left[\frac{(w_h^{\overline{M},\pi}(s,a))^2}{\mathbb{E}_{\pi'\sim p}[w_h^{\overline{M},\pi'}(s,a)]}\right]\right] = \max_{q\in\triangle_\Pi}\min_{p\in\triangle_\Pi} \sum_{s,a,h} \mathbb{E}_{\pi\sim q}\left[\mathbb{E}_{\overline{M}\sim\xi}\left[\frac{(w_h^{\overline{M},\pi}(s,a))^2}{\mathbb{E}_{\pi'\sim p}[w_h^{\overline{M},\pi'}(s,a)]}\right]\right]$$

$$
\begin{aligned}
&\leq \max_{q \in \triangle_\Pi} \sum_{s,a,h} \mathbb{E}_{\pi \sim q}\left[\mathbb{E}_{\overline{M} \sim \xi}\left[\frac{(w_h^{\overline{M},\pi}(s,a))^2}{\mathbb{E}_{\pi' \sim q}[w_h^{\overline{M},\pi'}(s,a)]}\right]\right] \\
&\leq \max_{q \in \triangle_\Pi} \sum_{s,a,h} \mathbb{E}_{\pi \sim q}\left[\mathbb{E}_{\overline{M} \sim \xi}\left[\frac{w_h^{\overline{M},\pi}(s,a)}{\mathbb{E}_{\pi' \sim q}[w_h^{\overline{M},\pi'}(s,a)]}\right]\right] \\
&= SAH.
\end{aligned}
$$

By Lemma A.9 of Foster et al. [2021], since $\mu^{\overline{M}}(s,a,h) := \frac{1}{H} w_h^{\overline{M},\pi'}(s,a)$ forms a valid distribution on $\mathcal{S} \times \mathcal{A} \times [H]$, we can upper bound

$$
\sum_{s,a,h} w_h^{\overline{M},\pi'}(s,a) D^2_{\mathsf{H}}\big(\overline{M}_{sh}(a), M_{sh}(a)\big) \leq H D^2_{\mathsf{H}}\big(\overline{M}(\pi'), M(\pi')\big).
$$

Altogether then, we have shown that for all $\pi \in \Pi$,

$$
\mathbb{E}_{\overline{M} \sim \xi}[D^2_{\mathsf{H}}\big(\overline{M}(\pi), M(\pi)\big)] \leq 200 \log(H) \cdot \sqrt{SAH^2 \cdot \mathbb{E}_{\overline{M} \sim \xi}[\mathbb{E}_{\pi \sim p_{\mathrm{exp}}}[D^2_{\mathsf{H}}\big(\overline{M}(\pi), M(\pi)\big)]]}
$$

as desired. Since the Hellinger distance is a metric and satisfies the triangle inequality, this in particular implies that, for any M, M',

$$
\begin{aligned}
D^2_{\mathsf{H}}(M(\pi), M'(\pi)) &\leq 2\mathbb{E}_{\overline{M} \sim \xi}[D^2_{\mathsf{H}}\big(\overline{M}(\pi), M(\pi)\big)] + 2\mathbb{E}_{\overline{M} \sim \xi}[D^2_{\mathsf{H}}\big(\overline{M}(\pi), M'(\pi)\big)] \\
&\leq 400 \log H \cdot \sqrt{SAH^2 \cdot \mathbb{E}_{\overline{M} \sim \xi}[\mathbb{E}_{\pi \sim p_{\mathrm{exp}}}[D^2_{\mathsf{H}}\big(\overline{M}(\pi), M(\pi)\big)]]} \\
&\quad + 400 \log H \sqrt{SAH^2 \cdot \mathbb{E}_{\overline{M} \sim \xi}[\mathbb{E}_{\pi \sim p_{\mathrm{exp}}}[D^2_{\mathsf{H}}\big(\overline{M}(\pi), M'(\pi)\big)]]}.
\end{aligned}
$$

Thus, if

$$
\mathbb{E}_{\overline{M} \sim \xi}[\mathbb{E}_{\pi \sim p_{\mathrm{exp}}}[D^2_{\mathsf{H}}\big(\overline{M}(\pi), M''(\pi)\big)]] \leq \frac{\varepsilon^2}{320000 SAH^2 \cdot \log^2 H}
$$

for both $M'' \in \{M, M'\}$, then $D^2_{\mathsf{H}}(M(\pi), M'(\pi)) \leq \varepsilon$. It follows that a sufficient choice for $C^{\mathsf{D}}_{\mathrm{exp}}$ is $320000 SAH^2 \log^2 H / \varepsilon^2$. □

H.4.5.4 Supporting Lemmas

Lemma H.4.11. *If M has a unique optimal policy π_M and $M' \in \mathcal{M}^{\mathrm{alt}}(M)$, then there exists some $(\widetilde{s}, \widetilde{a}, \widetilde{h})$ such that*

$$
Q^{M',\pi_M}_{\widetilde{h}}(\widetilde{s}, \widetilde{a}) > V^{M',\pi_M}_{\widetilde{h}}(\widetilde{s}).
$$

Proof of Lemma H.4.11. Assume that this is not the case, i.e. that for all (s, a, h),

$$Q_h^{M',\pi_M}(s,a) \le V_h^{M',\pi_M}(s) = Q_h^{M',\pi_M}(s, \pi_M(s,h)).$$

Our goal is to show that in this case $\pi_{M'} = \pi_M$, which contradicts the fact that $M' \in \mathcal{M}^{\text{alt}}(M)$. We proceed by induction.

Base Case. Let $h = H$ and assume that for all (s, a),

$$Q_H^{M',\pi_M}(s,a) \le Q_H^{M',\pi_M}(s, \pi_M(s,h)).$$

This contradicts the assumption that π_M is unique.

Inductive Case. Assume that $\pi_{M'}(s,h') = \pi_M(s,h')$ for all s and $h' > h$. This then implies that $V_{h+1}^{M',\pi_M}(s) = V_{h+1}^{M',\pi_{M'}}(s)$ for all s. It follows that for all a

$$Q_h^{M',\pi_M}(s,a) = r_h^{M'}(s,a) + \mathbb{P}_h^{M'}[V_{h+1}^{M',\pi_M}](s,a) = r_h^{M'}(s,a) + \mathbb{P}_h^{M'}[V_{h+1}^{M',\pi_{M'}}](s,a) = Q_h^{M',\pi_{M'}}(s,a)$$

so in particular $Q_h^{M',\pi_M}(s,\pi_M(s,h)) = Q_h^{M',\pi_{M'}}(s,\pi_M(s,h))$. Since we have assumed that for all (s,a)

$$Q_h^{M',\pi_M}(s,a) \le Q_h^{M',\pi_M}(s,\pi_M(s,h))$$

we have

$$Q_h^{M',\pi_{M'}}(s,\pi_{M'}(s,h)) \le Q_h^{M',\pi_{M'}}(s,\pi_M(s,h)).$$

However, since each $M \in \mathcal{M}$ has a unique optimal action at each state, this is a contradiction unless $\pi_{M'}(s,h) = \pi_M(s,h)$, which proves the inductive hypothesis. The result follows. □

Lemma H.4.12. *For MDPs M, M' with unit variance Gaussian rewards, we have*

$$V_h^{M',\pi}(s) - V_h^{M,\pi}(s) \le \sqrt{\frac{8H}{w_h^{M,\pi}(s)} \cdot D_{\mathsf{KL}}(M(\pi) \,\|\, M'(\pi))}.$$

Proof of Lemma H.4.12. In the Gaussian reward setting, we have

$$r_{h'}^{M'}(s',a') - r_{h'}^{M}(s',a') \le \sqrt{(r_{h'}^{M'}(s',a') - r_{h'}^{M}(s',a'))^2} \le \sqrt{2D_{\mathsf{KL}}\big(M_{h',s'}(a') \,\|\, M'_{h',s'}(a')\big)}.$$

Furthermore, since $V_{h'+1}^{M',\pi}(s') \in [0,1]$, we have

$$\mathbb{P}_{h'}^{M'}[V_{h'+1}^{M',\pi}](s',a') - \mathbb{P}_{h'}^{M}[V_{h'+1}^{M',\pi}](s',a') \le \sum_{s''} |P_{h'}^{M'}(s'' \mid s',a') - P_{h'}^{M}(s'' \mid s',a')|$$

$$
\begin{aligned}
&= 2D_{\mathsf{TV}}\big(P^{M'}_{h'}(\cdot \mid s', a'), P^{M}_{h'}(\cdot \mid s', a')\big) \\
&\le \sqrt{2D_{\mathsf{KL}}\big(P^{M'}_{h'}(\cdot \mid s', a') \,\|\, P^{M}_{h'}(\cdot \mid s', a')\big)} \\
&\le \sqrt{2D_{\mathsf{KL}}\big(M_{h',s'}(a') \,\|\, M'_{h',s'}(a')\big)}.
\end{aligned}
$$

By Lemma H.4.14, Jensen's inequality, and Lemma H.4.15, it follows that

$$
\begin{aligned}
V^{M',\pi}_h(s) - V^{M,\pi}_h(s) &\le \sum_{h'=h}^{H} \sum_{s',a'} w^{M,\pi}_{h'}(s', a' \mid s_h = s) \cdot 2\sqrt{2D_{\mathsf{KL}}\big(M_{h',s'}(a') \,\|\, M'_{h',s'}(a')\big)} \\
&\le 2\sqrt{2H \sum_{h'=h}^{H} \sum_{s',a'} w^{M,\pi}_{h'}(s', a' \mid s_h = s) D_{\mathsf{KL}}\big(M_{h',s'}(a') \,\|\, M'_{h',s'}(a')\big)} \\
&\le 2\sqrt{\frac{2H}{w^{M,\pi}_h(s)} \cdot \sum_{h'=h}^{H} \sum_{s',a'} w^{M,\pi}_{h'}(s', a') D_{\mathsf{KL}}\big(M_{h',s'}(a') \,\|\, M'_{h',s'}(a')\big)} \\
&\le 2\sqrt{\frac{2H}{w^{M,\pi}_h(s)} D_{\mathsf{KL}}(M(\pi) \,\|\, M'(\pi))}
\end{aligned}
$$

where we have used that, for $h < h'$,

$$
w^{M,\pi}_{h'}(s', a') = \sum_{s''} w^{M,\pi}_{h'}(s', a' \mid s_h = s'') w^{M}_{\pi}(s'', h) \ge w^{M,\pi}_{h'}(s', a' \mid s_h = s) w^{M,\pi}_{h}(s).
$$

□

Lemma H.4.13. *For any $M \in \mathcal{M}$ for which π_M is unique, we have*

$$
\Delta^{M}_{\min} \le \min_{s,h} w^{M,\pi_M}_h(s).
$$

Proof of Lemma H.4.13. Let $\widetilde{\pi}$ denote the policy that differs from policy π_M only at the state $\widetilde{s}$ and layer $\widetilde{h}$ given by $(\widetilde{s}, \widetilde{h}) = \arg\min_{s,h} w^{M,\pi_M}_h(s)$. Note that this implies, in particular, that $w^{M,\pi_M}_{\widetilde{h}}(\widetilde{s}) = w^{M,\widetilde{\pi}}_{\widetilde{h}}(\widetilde{s})$ since $\widetilde{\pi}$ and π_M take identical actions up to step $\widetilde{h}$. By the Performance-Difference Lemma [Kakade, 2003], we have

$$
\begin{aligned}
V^{M,\pi_M}_1 - V^{M,\widetilde{\pi}}_1 &= \sum_{h=1}^{H} \sum_{s,a} w^{M,\widetilde{\pi}}_h(s,a)(V^{M,\pi_M}_h(s) - Q^{M,\pi_M}_h(s,a)) \\
&\overset{(a)}{=} w^{M,\widetilde{\pi}}_{\widetilde{h}}(\widetilde{s}, \widetilde{\pi}(\widetilde{s}, \widetilde{h}))(V^{M,\pi_M}_{\widetilde{h}}(\widetilde{s}) - Q^{M,\pi_M}_{\widetilde{h}}(\widetilde{s}, \widetilde{\pi}(\widetilde{s}, \widetilde{h}))) \\
&= w^{M,\pi_M}_{\widetilde{h}}(\widetilde{s})(V^{M,\pi_M}_{\widetilde{h}}(\widetilde{s}) - Q^{M,\pi_M}_{\widetilde{h}}(\widetilde{s}, \widetilde{\pi}(\widetilde{s}, \widetilde{h})))
\end{aligned}
$$

$$\leq w_{\widetilde{h}}^{M,\pi_M}(\widetilde{s})$$

where (a) holds since $V_h^{M,\pi_M}(s) = Q_h^{M,\pi_M}(s,a)$ for all (s,a,h) with $w^{M,\widetilde{\pi}}(s,a,h) > 0$ other than at $(\widetilde{s},\widetilde{h})$. By assumption, the optimal policy is unique, so $V_1^{M,\pi_M} - V_1^{M,\widetilde{\pi}} > 0$, and thus

$$\Delta_{\min}^M = \min_{\pi\in\Pi\,:\,V_1^{M,\pi_M}-V_1^{M,\pi}>0} V_1^{M,\pi_M} - V_1^{M,\pi} \leq V_1^{M,\pi_M} - V_1^{M,\widetilde{\pi}} \leq w_{\widetilde{h}}^{M,\pi_M}(\widetilde{s}) = \min_{s,h} w_h^{M,\pi_M}(s).$$

□

Lemma H.4.14 (Lemma E.15 of Dann et al. [2017]). *For MDPs M, M' and policy π, we have*

$$V_h^{M',\pi}(s) - V_h^{M,\pi}(s) = \sum_{h'=h}^{H}\sum_{s',a'} w_{h'}^{M,\pi}(s',a' \mid s_h = s)\cdot\Big[r_{h'}^{M'}(s',a') - r_{h'}^{M}(s',a') + \mathbb{P}_{h'}^{M'}[V_{h'+1}^{M',\pi}](s',a') - \mathbb{P}_{h'}^{M}[V_{h'+1}^{M',\pi}](s',a')\Big].$$

Lemma H.4.15. *For MDPs M, M' and policy π, we have*

$$D_{\mathsf{KL}}(M(\pi) \,\|\, M'(\pi)) = \sum_{h=1}^{H}\sum_{s,a} w_h^{M,\pi}(s,a) D_{\mathsf{KL}}(M_{hs}(a) \,\|\, M'_{hs}(a)).$$

Proof of Lemma H.4.15. This is a standard calculation; see e.g. [Simchowitz and Jamieson, 2019, Tirinzoni et al., 2021]. □

H.5 Deferred Proofs from Section 10.4

H.5.1 Technical Lemmas

Throughout this section, when the class $\mathcal{M}$ is clear from context, we define

$$\Upsilon(M;\varepsilon,\bar{\mathsf{n}}) := \{\lambda \in \triangle_\Pi \;:\; \exists \mathsf{n} \leq \bar{\mathsf{n}} \text{ s.t. } \Delta^M(\lambda) \leq (1+\varepsilon)\mathsf{g}^M/\mathsf{n}, I^M(\lambda;\mathcal{M}) \geq (1-\varepsilon)/\mathsf{n}\}. \tag{H.5.1}$$

Lemma H.5.1 (Derandomization). *Let $\bar{\mathsf{n}} > 0$ be given. For any $p \in \triangle(\triangle(\Pi))$, defining $\overline{\lambda}_p = \mathbb{E}_{\lambda\sim p}[\lambda] \in \Delta(\Pi)$, we have that for all $M \in \mathcal{M}$,*

$$\mathbb{I}\big\{\overline{\lambda}_p \notin \Upsilon(M;\varepsilon,\bar{\mathsf{n}})\big\} \leq \delta^{-1}\cdot\mathbb{P}_{\lambda\sim p}[\lambda \notin \Upsilon(M;\varepsilon/2,\bar{\mathsf{n}})],$$

where $\delta := \frac{\varepsilon}{2}\cdot\min\Big\{1, \frac{\mathsf{g}^M}{\bar{\mathsf{n}}}\Big\}$.

Proof of Lemma H.5.1. Let $M \in \mathcal{M}$ be fixed and abbreviate $I^M(\lambda) = I^M(\lambda;\mathcal{M})$. Fix

$p \in \triangle(\triangle(\Pi))$. For any $\lambda \in \Upsilon(M; \varepsilon/2, \bar{\mathsf{n}})$, let $\mathsf{n}_\lambda > 0$ denote the least $\mathsf{n} > 0$ such that

$$\Delta^M(\lambda) \leq (1 + \varepsilon/2)\mathsf{g}^M/\mathsf{n}, \quad \text{and} \quad I^M(\lambda; \mathcal{M}) \geq (1 - \varepsilon/2)/\mathsf{n}.$$

Define

$$\mathsf{n} = \left(\mathbb{E}_{\lambda \sim p}\left[\frac{1}{\mathsf{n}_\lambda} \mid \lambda \in \Upsilon(M; \varepsilon/2, \bar{\mathsf{n}})\right]\right)^{-1},$$

and note that by Jensen's inequality,

$$\mathsf{n} \leq \mathbb{E}_{\lambda \sim p}[\mathsf{n}_\lambda \mid \lambda \in \Upsilon(M; \varepsilon/2, \bar{\mathsf{n}})] \leq \bar{\mathsf{n}}.$$

We first observe that since $\Delta^M \in [0, 1]$,

$$\begin{aligned}
\Delta^M(\bar{\lambda}_p) &\leq \mathbb{E}_{\lambda \sim p}[\Delta^M(\lambda) \mid \lambda \in \Upsilon(M; \varepsilon/2, \bar{\mathsf{n}})] + \mathbb{P}_{\lambda \sim p}[\lambda \notin \Upsilon(M; \varepsilon/2, \bar{\mathsf{n}})] \\
&\leq (1 + \varepsilon/2)\mathsf{g}^M \cdot \mathbb{E}_{\lambda \sim p}\left[\frac{1}{\mathsf{n}_\lambda} \mid \lambda \in \Upsilon(M; \varepsilon/2, \bar{\mathsf{n}})\right] + \mathbb{P}_{\lambda \sim p}[\lambda \notin \Upsilon(M; \varepsilon/2, \bar{\mathsf{n}})] \\
&= (1 + \varepsilon/2)\frac{\mathsf{g}^M}{\mathsf{n}} + \mathbb{P}_{\lambda \sim p}[\lambda \notin \Upsilon(M; \varepsilon/2, \bar{\mathsf{n}})].
\end{aligned}$$

Next, note that the map $\lambda \mapsto I^M(\lambda)$ is concave and non-negative (it is an infimum over non-negative linear functions), so we have

$$\begin{aligned}
I^M(\bar{\lambda}_p) &\geq \mathbb{E}_{\lambda \sim p}[I^M(\lambda)] \\
&\geq \mathbb{E}_{\lambda \sim p}[I^M(\lambda) \mid \lambda \in \Upsilon(M; \varepsilon/2, \bar{\mathsf{n}})] \cdot \mathbb{P}_{\lambda \sim p}[\lambda \in \Upsilon(M; \varepsilon/2, \bar{\mathsf{n}})] \\
&\geq (1 - \varepsilon/2)\,\mathbb{E}_{\lambda \sim p}\left[\frac{1}{\mathsf{n}_\lambda} \mid \lambda \in \Upsilon(M; \varepsilon/2, \bar{\mathsf{n}})\right] \cdot \mathbb{P}_{\lambda \sim p}[\lambda \in \Upsilon(M; \varepsilon/2, \bar{\mathsf{n}})] \\
&= (1 - \varepsilon/2)\frac{1}{\mathsf{n}} \cdot \mathbb{P}_{\lambda \sim p}[\lambda \in \Upsilon(M; \varepsilon/2, \bar{\mathsf{n}})] \\
&= (1 - \varepsilon/2)\frac{1}{\mathsf{n}} \cdot (1 - \mathbb{P}_{\lambda \sim p}[\lambda \notin \Upsilon(M; \varepsilon/2, \bar{\mathsf{n}})]).
\end{aligned}$$

It follows that as long as

$$\mathbb{P}_{\lambda \sim p}[\lambda \notin \Upsilon(M; \varepsilon/2, \bar{\mathsf{n}})] \leq \delta := \frac{\varepsilon}{2} \cdot \min\left\{1, \frac{\mathsf{g}^M}{\bar{\mathsf{n}}}\right\} \leq \frac{\varepsilon}{2} \cdot \min\left\{1, \frac{\mathsf{g}^M}{\mathsf{n}}\right\},$$

we have

$$\Delta^M(\bar{\lambda}_p) \leq (1 + \varepsilon)\mathsf{g}^M/\mathsf{n}, \quad \text{and} \quad I^M(\bar{\lambda}_p; \mathcal{M}) \geq (1 - \varepsilon)/\mathsf{n},$$

so that $\bar{\lambda}_p \in \Upsilon(M; \varepsilon, \mathsf{n}) \subseteq \Upsilon(M; \varepsilon, \bar{\mathsf{n}})$. We conclude that

$$\mathbb{I}\{\bar{\lambda}_p \notin \Upsilon(M; \varepsilon, \bar{\mathsf{n}})\} \leq \mathbb{I}\{\mathbb{P}_{\lambda \sim p}[\lambda \notin \Upsilon(M; \varepsilon/2, \bar{\mathsf{n}})] > \delta\} \leq \delta^{-1} \cdot \mathbb{P}_{\lambda \sim p}[\lambda \notin \Upsilon(M; \varepsilon/2, \bar{\mathsf{n}})].$$

□

Lemma H.5.2. *Let $M \in \mathcal{M}$ be given, and suppose Assumption 10.7 holds. Fix $T \in \mathbb{N}$ and consider an algorithm $\mathbb{A}$ such that for all $M' \in \mathcal{M}^{\text{alt}}(M) \cup \{M\}$,*

$$\mathbb{E}^{M',\mathbb{A}}[\mathbf{Reg}(T)] \leq R^{M'} \cdot \log(T).$$

for some bound $R^M \geq 2$. Define $\eta^M \in \mathbb{R}_+^{\Pi}$ via

$$\eta^M(\pi) = \mathbb{E}^{M,\mathbb{A}}\left[\frac{T(\pi)}{\log(T)}\right].$$

Then if

$$\log(T) \geq \frac{6}{\varepsilon}\log\left(\sup_{M' \in \mathcal{M}^{\text{alt}}(M)\cup\{M\}} \frac{R^{M'}}{\Delta_{\min}^{M'}} \cdot \log(T)\right),$$

we must have

$$I^M(\eta^M; \mathcal{M}) \geq (1-\varepsilon).$$

Proof of Lemma H.5.2. Throughout this proof we will use that π_M is uniquely defined for all $M \in \mathcal{M}$ by Assumption 10.7. Note that

$$\mathbb{E}^{M,\mathbb{A}}[\mathbf{Reg}(T)] \leq R^M \log T \implies \sum_{\pi \neq \pi_M} \mathbb{E}^{M,\mathbb{A}}[T(\pi)] \leq \frac{R^M \log T}{\Delta_{\min}^M}.$$

Fix some $M' \in \mathcal{M}^{\text{alt}}(M)$. Then $\pi_M \neq \pi_{M'}$ (recall that under Assumption 10.7, each $M \in \mathcal{M}$ has a unique optimal), so

$$\mathbb{E}^{M,\mathbb{A}}[T(\pi_{M'})] \leq \frac{R^M \log T}{\Delta_{\min}^M}, \quad \mathbb{E}^{M',\mathbb{A}}[T(\pi_{M'})] \geq T - \frac{R^{M'} \log T}{\Delta_{\min}^{M'}}.$$

By Lemma H.1 of Simchowitz and Jamieson [2019], we have that for any $\mathcal{H}^T$-measurable variable $Z \in [0,1]$, that

$$\sum_{\pi} \mathbb{E}^{M,\mathbb{A}}[T(\pi)] D_{\mathsf{KL}}(M(\pi), M'(\pi)) \geq d(\mathbb{E}^{M,\mathbb{A}}[Z], \mathbb{E}^{M',\mathbb{A}}[Z])$$

for $d(x,y) = x\log\frac{x}{y} + (1-x)\log\frac{1-x}{1-y}$. Choosing $Z = T(\pi_{M'})/T$, and using that

$$d(x,y) \geq (1-x)\log\frac{1}{1-y} - \log 2,$$

we have

$$\sum_{\pi} \mathbb{E}^{M,\mathbb{A}}[T(\pi)] D_{\mathsf{KL}}(M(\pi), M'(\pi)) \geq \left(1 - \frac{R^M \log T}{\Delta^M_{\min} T}\right) \log \frac{T}{T - (T - \frac{R^{M'} \log T}{\Delta^{M'}_{\min}})} - \log 2$$
$$= \left(1 - \frac{R^M \log T}{\Delta^M_{\min} T}\right) \left(\log T - \log \frac{R^{M'} \log T}{\Delta^{M'}_{\min}}\right) - \log 2.$$

Now, if

$$\log(T) \geq \frac{2}{\varepsilon} \log\left(\frac{\sup_{M' \in \mathcal{M}^{\mathrm{alt}}(M) \cup \{M\}} R^{M'} \log T}{\Delta^{M'}_{\min}} \vee 2\right),$$

we have $\log(2) \leq \varepsilon \log(T)$,

$$\log T - \log\left(\frac{R^{M'} \log T}{\Delta^{M'}_{\min}}\right) \geq (1 - \varepsilon) \log(T),$$

and $1 - \frac{R^M \log T}{\Delta^M_{\min} T} \geq (1 - \varepsilon)$, so we can lower bound

$$\sum_{\pi} \mathbb{E}^{M,\mathbb{A}}[T(\pi)] D_{\mathsf{KL}}(M(\pi), M'(\pi)) \geq \left((1 - \varepsilon)^2 - \varepsilon\right) \log(T) \geq (1 - 3\varepsilon) \log(T).$$

As this is true for every $M' \in \mathcal{M}^{\mathrm{alt}}(M)$, the result follows. □

Lemma 10.4.1. *Let $\varepsilon \in (0, 2)$, and suppose that Assumption 10.7 holds. Fix $T \in \mathbb{N}$ and consider an algorithm $\mathbb{A}$ such that for all $M \in \mathcal{M}$,*

$$\mathbb{E}^{M,\mathbb{A}}[\mathbf{Reg}(T)] \leq (1 + \varepsilon) \mathsf{g}^M \cdot \log(T).$$

For each $M \in \mathcal{M}$, define $\eta^M \in \mathbb{R}^{\Pi}_+$ via $\eta^M(\pi) = \mathbb{E}^{M,\mathbb{A}}\left[\frac{T(\pi)}{\log(T)}\right]$, where $T(\pi)$ denotes the number of pulls of decision π, and define $\lambda^M = \eta^M / \|\eta^M\|_1$. Then if

$$\log(T) \geq \frac{6}{\varepsilon} \log\left(\sup_{M \in \mathcal{M}} \frac{2\mathsf{g}^M}{\Delta^M_{\min}} \cdot \log(T)\right),$$

we have that for all $M \in \mathcal{M}$,

$$\lambda^M \in \Upsilon(M; \varepsilon). \tag{10.4.8}$$

Proof of Lemma 10.4.1. Immediate consequence of Lemma H.5.2. □

H.5.2 Proof of Theorem 10.9

Theorem H.4 (Full Statement of Theorem 10.9)**.** *Let $\varepsilon > 0$ and $\mathcal{M}_0 \subseteq \mathcal{M}$ be given. Let $\{\bar{\mathsf{n}}^M\}_{M \in \mathcal{M}_0}$ be a collection of non-negative scalars indexed by $\mathcal{M}_0$, and set $\delta := \frac{\varepsilon}{2} \cdot$*

$\min\Big\{1, \inf_{M\in\mathcal{M}_0} \frac{\mathsf{g}^M}{\bar{\mathsf{n}}^M}\Big\}$. *Unless*

$$T > \frac{\delta}{8}\cdot \sup_{\overline{M}\in\mathcal{M}^+} \mathsf{aec}^{\mathcal{M}}_{2\varepsilon}(\mathcal{M}_0, \overline{M}),$$

any algorithm must have, for some $M \in \mathcal{M}_0$*:*

$$\mathbb{P}^{M,\mathbb{A}}\Big[\widehat{\lambda} \notin \Upsilon(M;\varepsilon,\bar{\mathsf{n}}^M)\Big] \geq \frac{\delta}{6}.$$

Proof of Theorem H.4. Fix $\varepsilon > 0$, and let an algorithm $\mathbb{A}$ be given. For any $\overline{M} \in \mathcal{M}^+$, define $q^{\overline{M}} = \mathbb{P}^{\overline{M},\mathbb{A}}(\widehat{\lambda} = \cdot) \in \triangle(\triangle(\Pi))$, and let $\omega^{\overline{M}} := \mathbb{E}^{\overline{M},\mathbb{A}}\Big[\frac{1}{T}\sum_{t=1}^T p^t\Big] \in \triangle(\Pi)$.

Fix $\alpha > 0$ and $\overline{M} \in \mathcal{M}^+$ be fixed. Define

$$M = \arg\max_{M\in\mathcal{M}_0}\Big\{\mathbb{P}_{\lambda\sim q^{\overline{M}}}[\lambda \notin \Upsilon(M;\varepsilon,\bar{\mathsf{n}}^M)] \mid \mathbb{E}_{\pi\sim\omega^{\overline{M}}}\big[D_{\mathsf{KL}}\big(\overline{M}(\pi)\,\|\,M(\pi)\big)\big] \leq \alpha^2\Big\};$$

we assume that such an $M \in \mathcal{M}_0$ does exist, as otherwise the claim we will prove is trivial. It is immediate from this definition that we have

$$\begin{aligned}
&\mathbb{P}^{\overline{M},\mathbb{A}}\Big[\widehat{\lambda} \notin \Upsilon(M;\varepsilon,\bar{\mathsf{n}}^M)\Big]\\
&= \mathbb{P}_{\lambda\sim q^{\overline{M}}}[\lambda \notin \Upsilon(M;\varepsilon,\bar{\mathsf{n}}^M)]\\
&= \sup_{M\in\mathcal{M}_0}\Big\{\mathbb{P}_{\lambda\sim q^{\overline{M}}}[\lambda \notin \Upsilon(M;\varepsilon,\bar{\mathsf{n}}^M)] \mid \mathbb{E}_{\pi\sim\omega^{\overline{M}}}\big[D_{\mathsf{KL}}\big(\overline{M}(\pi)\,\|\,M(\pi)\big)\big] \leq \alpha^2\Big\}\\
&\geq \inf_{q\in\triangle(\triangle(\Pi)),\omega\in\triangle(\Pi)} \sup_{M\in\mathcal{M}_0}\big\{\mathbb{P}_{\lambda\sim q}[\lambda \notin \Upsilon(M;\varepsilon,\bar{\mathsf{n}}^M)] \mid \mathbb{E}_{\pi\sim\omega}\big[D_{\mathsf{KL}}\big(\overline{M}(\pi)\,\|\,M(\pi)\big)\big] \leq \alpha^2\big\}\\
&=: \mathsf{opt},
\end{aligned}$$

with the convention that this value is zero if the set $\big\{M \in \mathcal{M}_0 \mid \mathbb{E}_{\pi\sim\omega}\big[D_{\mathsf{KL}}\big(\overline{M}(\pi)\,\|\,M(\pi)\big)\big] \leq \alpha^2\big\}$ is empty. In addition, we have

$$\mathbb{E}_{\pi\sim\omega^{\overline{M}}}\big[D_{\mathsf{KL}}\big(\overline{M}(\pi)\,\|\,M(\pi)\big)\big] \leq \alpha^2. \tag{H.5.2}$$

Now, define $\delta := \frac{\varepsilon}{2}\cdot\min\Big\{1, \inf_{M\in\mathcal{M}_0}\frac{\mathsf{g}^M}{\bar{\mathsf{n}}^M}\Big\}$, and let $\bar{\lambda}_q := \mathbb{E}_{\lambda\sim q}[\lambda]$. By Lemma H.5.1, we have

$$\begin{aligned}
\mathsf{opt} &= \inf_{q\in\triangle(\triangle(\Pi)),\omega\in\triangle(\Pi)} \sup_{M\in\mathcal{M}_0}\big\{\mathbb{P}_{\lambda\sim q}[\lambda \notin \Upsilon(M;\varepsilon,\bar{\mathsf{n}}^M)] \mid \mathbb{E}_{\pi\sim\omega}\big[D_{\mathsf{KL}}\big(\overline{M}(\pi)\,\|\,M(\pi)\big)\big] \leq \alpha^2\big\}\\
&\geq \delta\cdot \inf_{q\in\triangle(\triangle(\Pi)),\omega\in\triangle(\Pi)} \sup_{M\in\mathcal{M}_0}\big\{\mathbb{I}\big\{\bar{\lambda}_q \notin \Upsilon(M;2\varepsilon,\bar{\mathsf{n}}^M)\big\} \mid \mathbb{E}_{\pi\sim\omega}\big[D_{\mathsf{KL}}\big(\overline{M}(\pi)\,\|\,M(\pi)\big)\big] \leq \alpha^2\big\}\\
&\geq \delta\cdot \inf_{\lambda\in\triangle(\Pi),\omega\in\triangle(\Pi)} \sup_{M\in\mathcal{M}_0}\big\{\mathbb{I}\{\lambda \notin \Upsilon(M;2\varepsilon,\bar{\mathsf{n}}^M)\} \mid \mathbb{E}_{\pi\sim\omega}\big[D_{\mathsf{KL}}\big(\overline{M}(\pi)\,\|\,M(\pi)\big)\big] \leq \alpha^2\big\}
\end{aligned}$$

$$\geq \delta \cdot \inf_{\lambda \in \triangle(\Pi), \omega \in \triangle(\Pi)} \sup_{M \in \mathcal{M}_0} \left\{ \mathbb{I}\{\lambda \notin \Upsilon(M; 2\varepsilon)\} \mid \mathbb{E}_{\pi \sim \omega}\left[D_{\mathsf{KL}}\left(\overline{M}(\pi) \,\|\, M(\pi) \right) \right] \leq \alpha^2 \right\}$$
$$= \delta \cdot \mathbb{I}\left\{ \alpha^2 \geq \left(\mathsf{aec}^{\mathcal{M}}_{2\varepsilon}(\mathcal{M}_0, \overline{M}) \right)^{-1} \right\}. \tag{H.5.3}$$

We conclude that

$$\mathbb{P}^{\overline{M}, \mathbb{A}}\left[\widehat{\lambda} \notin \Upsilon(M; \varepsilon, \bar{\mathsf{n}}^M) \right] \geq \delta \cdot \mathbb{I}\left\{ \alpha^2 \geq \left(\mathsf{aec}^{\mathcal{M}}_{2\varepsilon}(\mathcal{M}_0, \overline{M}) \right)^{-1} \right\}. \tag{H.5.4}$$

To proceed, using Lemma A.11 of [Foster et al., 2021], we have

$$\begin{aligned}
\mathbb{P}^{M, \mathbb{A}}\left[\widehat{\lambda} \notin \Upsilon(M; \varepsilon, \bar{\mathsf{n}}^M) \right] &\geq \frac{1}{3} \mathbb{P}^{\overline{M}, \mathbb{A}}\left[\widehat{\lambda} \notin \Upsilon(M; \varepsilon, \bar{\mathsf{n}}^M) \right] - \frac{4}{3} D_{\mathsf{KL}}(\mathbb{P}^{\overline{M}, \mathbb{A}} \,\|\, \mathbb{P}^{M, \mathbb{A}}) \\
&\geq \frac{\delta}{3} \mathbb{I}\left\{ \alpha^2 \geq \left(\mathsf{aec}^{\mathcal{M}}_{2\varepsilon}(\mathcal{M}_0, \overline{M}) \right)^{-1} \right\} - \frac{4}{3} D_{\mathsf{KL}}(\mathbb{P}^{\overline{M}, \mathbb{A}} \,\|\, \mathbb{P}^{M, \mathbb{A}}).
\end{aligned}$$

Using (H.5.2) gives

$$D_{\mathsf{KL}}(\mathbb{P}^{\overline{M}, \mathbb{A}} \,\|\, \mathbb{P}^{M, \mathbb{A}}) = \mathbb{E}^{M, \mathbb{A}}\left[\sum_{t=1}^{T} \mathbb{E}_{\pi \sim p^t} D_{\mathsf{KL}}\left(\overline{M}(\pi) \,\|\, M(\pi) \right) \right] = T \cdot \mathbb{E}_{\pi \sim \omega^{\overline{M}}}\left[D_{\mathsf{KL}}\left(\overline{M}(\pi) \,\|\, M(\pi) \right) \right] \leq \alpha^2 T,$$

so we have

$$\mathbb{P}^{M, \mathbb{A}}\left[\widehat{\lambda} \notin \Upsilon(M; \varepsilon, \bar{\mathsf{n}}^M) \right] \geq \frac{\delta}{3} \mathbb{I}\left\{ \alpha^2 \geq \left(\mathsf{aec}^{\mathcal{M}}_{2\varepsilon}(\mathcal{M}_0, \overline{M}) \right)^{-1} \right\} - \frac{4}{3} \alpha^2 T.$$

We set $\alpha^2 = \frac{\delta}{8T}$, so that

$$\begin{aligned}
\mathbb{P}^{M, \mathbb{A}}\left[\widehat{\lambda} \notin \Upsilon(M; \varepsilon, \bar{\mathsf{n}}^M) \right] &\geq \frac{\delta}{6} \cdot \mathbb{I}\left\{ \alpha^2 \geq \left(\mathsf{aec}^{\mathcal{M}}_{2\varepsilon}(\mathcal{M}_0, \overline{M}) \right)^{-1} \right\} \\
&= \frac{\delta}{6} \cdot \mathbb{I}\left\{ T \leq \frac{\delta}{8} \cdot \mathsf{aec}^{\mathcal{M}}_{2\varepsilon}(\mathcal{M}_0, \overline{M}) \right\}.
\end{aligned}$$

By taking the supremum over all possible choices for $\overline{M} \in \mathcal{M}^+$, we conclude that unless

$$T > \frac{\delta}{8} \cdot \sup_{\overline{M} \in \mathcal{M}^+} \mathsf{aec}^{\mathcal{M}}_{2\varepsilon}(\mathcal{M}_0, \overline{M}),$$

the algorithm must have $\mathbb{P}^{M, \mathbb{A}}\left[\widehat{\lambda} \notin \Upsilon(M; \varepsilon, \bar{\mathsf{n}}^M) \right] \geq \frac{\delta}{6}$. □

H.5.3 Proof of Theorem 10.10

Recall that for $M \in \mathcal{M}^+$ we define

$$\boldsymbol{\pi}_M = \arg\max_{\pi \in \Pi} f^M(\pi)$$

as the set $\boldsymbol{\pi}_M \subseteq \Pi$ of all optimal decisions π with $f^M(\pi) = \max_{\pi' \in \Pi} f^M(\pi')$. Unless otherwise stated, the results in this subsection do not make use of Assumption 10.7. For $\overline{M} \in \mathcal{M}^+$ and $\mathcal{M}_0 \subseteq \mathcal{M}$, we define

$$\mathcal{M}_0^{\mathrm{opt}}(\overline{M}) = \big\{ M \in \mathcal{M}_0 \mid \boldsymbol{\pi}_M \subseteq \boldsymbol{\pi}_{\overline{M}},\ \ D_{\mathsf{KL}}\big(\overline{M}(\pi) \,\|\, M(\pi)\big) = 0 \ \forall \pi \in \boldsymbol{\pi}_{\overline{M}} \big\}.$$

For a subset $\Pi' \subseteq \Pi$, let

$$N_{\neg \Pi'} = \big| \big\{ t \in [T] \mid \pi^t \notin \Pi' \big\} \big|.$$

Note that for all $M \in \mathcal{M}_0^{\mathrm{opt}}(\overline{M})$, since $\boldsymbol{\pi}_M \subseteq \boldsymbol{\pi}_{\overline{M}}$, we have

$$N_{\neg \boldsymbol{\pi}_{\overline{M}}} \leq N_{\neg \boldsymbol{\pi}_M}.$$

Theorem 10.10 (Main lower bound—strong variant)**.** *Let $\varepsilon > 0$, $\mathsf{n}_{\max} > 0$, and $\mathcal{M}_0 \subseteq \mathcal{M}$ be given, and define $\delta = \frac{\varepsilon}{2} \cdot \min\{1, \inf_{M \in \mathcal{M}_0} \mathsf{g}^M / \mathsf{n}_{\max}\}$. Unless*

$$\sup_{M \in \mathcal{M}_0} \frac{\mathsf{g}^M}{\Delta_{\min}^M} \cdot \log(T) \geq \Omega(\delta^2) \cdot \sup_{\overline{M} \in \mathcal{M}^+} \mathsf{aec}_{2\varepsilon}^{\mathcal{M}}(\mathcal{M}_0^{\mathrm{opt}}(\overline{M}), \overline{M}),$$

there is no algorithm that simultaneously ensures that

1. $\mathbb{E}^{M, \mathbb{A}}[\mathbf{Reg}(T)] \leq 2 \cdot \mathsf{g}^M \log(T), \ \ \forall M \in \mathcal{M}_0.$
2. $\mathbb{P}^{M, \mathbb{A}}\Big[\widehat{\lambda} \notin \Upsilon(M; \varepsilon, \mathsf{n}_{\max})\Big] \leq \frac{\delta}{12}, \ \ \forall M \in \mathcal{M}_0.$

Proof of Theorem 10.10. For each $M \in \mathcal{M}$, if $\mathbb{E}^{M, \mathbb{A}}[\mathbf{Reg}(T)] \leq 2\mathsf{g}^M \log(T)$, then $\mathbb{E}^{M, \mathbb{A}}[N_{\neg \boldsymbol{\pi}_M}] \leq 2\frac{\mathsf{g}^M}{\Delta_{\min}^M} \log(T)$. The result now follows by appealing to Theorem H.5 with $\bar{\mathsf{n}}^M = \mathsf{n}_{\max}$ and $R = 2 \sup_{M \in \mathcal{M}_0} \frac{\mathsf{g}^M}{\Delta_{\min}^M} \log(T)$. □

Theorem H.5. *Let $T \in \mathbb{N}$, $\varepsilon > 0$, $R \geq 1$, and $\mathcal{M}_0 \subseteq \mathcal{M}$ be given. Let $\{\bar{\mathsf{n}}^M\}_{M \in \mathcal{M}_0}$ be a collection of non-negative scalars indexed by $\mathcal{M}_0$. Define $\delta = \frac{\varepsilon}{2} \cdot \min\Big\{1, \inf_{M \in \mathcal{M}_0} \frac{\mathsf{g}^M}{\bar{\mathsf{n}}^M}\Big\}$. Unless*

$$R \geq \frac{\delta^2}{192} \cdot \sup_{\overline{M} \in \mathcal{M}^+} \mathsf{aec}_{2\varepsilon}^{\mathcal{M}}(\mathcal{M}_0^{\mathrm{opt}}(\overline{M}), \overline{M}),$$

there is no algorithm that simultaneously ensures that

1. $\mathbb{E}^{M,\mathbb{A}}[N_{\neg \boldsymbol{\pi}_M}] \leq R, \ \ \forall M \in \mathcal{M}_0.$

2. $\mathbb{P}^{M,\mathbb{A}}\Big[\widehat{\lambda} \notin \Upsilon(M;\varepsilon,\bar{\mathsf{n}}^M)\Big] \leq \frac{\delta}{12}, \ \ \forall M \in \mathcal{M}_0.$

Proof of Theorem H.5. Let $\varepsilon > 0$ be fixed. To prove the result, it suffices to lower bound the constrained minimax value

$$f^M := \sup_{\overline{M} \in \mathcal{M}^+} \inf_{\mathbb{A}} \left\{ \sup_{M \in \mathcal{M}_0^{\mathrm{opt}}(\overline{M})} \mathbb{P}^{M,\mathbb{A}}\Big[\widehat{\lambda} \notin \Upsilon(M;\varepsilon,\bar{\mathsf{n}}^M)\Big] \mid \mathbb{E}^{M',\mathbb{A}}[N_{\neg \boldsymbol{\pi}_{M'}}] \leq R \ \ \forall M' \in \mathcal{M}_0^{\mathrm{opt}}(\overline{M}) \right\}. \tag{H.5.5}$$

We begin by appealing to the following technical lemma.

Lemma H.5.3. *Let $\overline{M} \in \mathcal{M}^+$ and $T \in \mathbb{N}$ be given. Consider any algorithm $\mathbb{A}$ with the property that for all $M \in \mathcal{M}_0^{\mathrm{opt}}(\overline{M})$,*

$$\mathbb{E}^{M,\mathbb{A}}[N_{\neg \boldsymbol{\pi}_M}] \leq R$$

for some $R \geq 1$. For any $\beta \in (0,1)$, there exists a modified algorithm $\mathbb{A}'$ with the following properties:

- $\mathbb{P}^{M,\mathbb{A}'}\Big[N_{\neg \boldsymbol{\pi}_{\overline{M}}} > \lceil \frac{R}{\beta} \rceil\Big] = 0$ *for all models $M \in \mathcal{M}^+$.*
- *For all $M \in \mathcal{M}_0^{\mathrm{opt}}(\overline{M})$,*

$$\mathbb{P}^{M,\mathbb{A}}\Big[\widehat{\lambda} \notin \Upsilon(M;\varepsilon,\bar{\mathsf{n}}^M)\Big] \geq \mathbb{P}^{M,\mathbb{A}'}\Big[\widehat{\lambda} \notin \Upsilon(M;\varepsilon,\bar{\mathsf{n}}^M)\Big] - \beta.$$

By Lemma H.5.3, for any $\beta \in (0,1)$, we have

$$f^M \geq \sup_{\overline{M} \in \mathcal{M}^+} \inf_{\mathbb{A}} \left\{ \sup_{M \in \mathcal{M}_0^{\mathrm{opt}}(\overline{M})} \mathbb{P}^{M,\mathbb{A}}\Big[\widehat{\lambda} \notin \Upsilon(M;\varepsilon,\bar{\mathsf{n}}^M)\Big] \mid \mathbb{P}^{M',\mathbb{A}}\left[N_{\neg \boldsymbol{\pi}_{\overline{M}}} > \left\lceil \frac{R}{\beta} \right\rceil\right] = 0 \ \ \forall M' \in \mathcal{M}^+ \right\} - \beta.$$

Now, consider an arbitrary choice for $\overline{M}$ above. We lower bound the minimax value using another technical lemma.

Lemma H.5.4. *Let $T \in \mathbb{N}$ and $\varepsilon > 0$ be given. Let $\{\bar{\mathsf{n}}^M\}_{M \in \mathcal{M}}$ be a collection of non-negative scalars indexed by $\mathcal{M}$. Consider any algorithm $\mathbb{A}$ with the property that*

$$\mathbb{P}^{\overline{M},\mathbb{A}}\big[N_{\neg \boldsymbol{\pi}_{\overline{M}}} > R\big] = 0$$

for some $R \geq 1$. *For any* $\overline{M} \in \mathcal{M}^+$, *if we set* $\delta := \frac{\varepsilon}{2} \cdot \min\left\{1, \inf_{M \in \mathcal{M}_0^{\mathrm{opt}}(\overline{M})} \frac{\mathsf{g}^M}{\bar{\mathsf{n}}^M}\right\}$, *then unless*

$$R > \frac{\delta}{8} \cdot \mathsf{aec}^{\mathcal{M}}_{2\varepsilon}(\mathcal{M}_0^{\mathrm{opt}}(\overline{M}), \overline{M}),$$

the algorithm must have

$$\sup_{M \in \mathcal{M}_0^{\mathrm{opt}}(\overline{M})} \mathbb{P}^{M, \mathbb{A}}\left[\widehat{\lambda} \notin \Upsilon(M; \varepsilon, \bar{\mathsf{n}}^M)\right] \geq \frac{\delta}{6}.$$

Bounding $\left\lceil \frac{R}{\beta} \right\rceil \leq \frac{2R}{\beta}$, it follows from Lemma H.5.4 that unless

$$\frac{2R}{\beta} > \frac{\delta}{8} \cdot \mathsf{aec}^{\mathcal{M}}_{2\varepsilon}(\mathcal{M}_0^{\mathrm{opt}}(\overline{M}), \overline{M}),$$

where $\delta := \frac{\varepsilon}{2} \cdot \min\left\{1, \inf_{M \in \mathcal{M}_0^{\mathrm{opt}}(\overline{M})} \frac{\mathsf{g}^M}{\bar{\mathsf{n}}^M}\right\}$, we have

$$f^M \geq \inf_{\mathbb{A}} \left\{ \sup_{M \in \mathcal{M}_0^{\mathrm{opt}}(\overline{M})} \mathbb{P}^{M, \mathbb{A}}\left[\widehat{\lambda} \notin \Upsilon(M; \varepsilon, \bar{\mathsf{n}}^M)\right] \mid \mathbb{P}^{M', \mathbb{A}}\left[N_{\neg \boldsymbol{\pi}_{\overline{M}}} > \left\lceil \frac{R}{\beta} \right\rceil\right] = 0 \ \ \forall M' \in \mathcal{M}^+ \right\} - \beta \geq \frac{\delta}{6} - \beta.$$

To conclude, we set $\beta = \frac{\delta}{12}$ and maximize over $\overline{M} \in \mathcal{M}^+$. □

Proof of Lemma H.5.3. Fix $\beta \in (0, 1)$ and let $C := \lceil \frac{R}{\beta} \rceil$. Fix $\mathbb{A} = (p, q)$ and consider the algorithm $\mathbb{A}' = (p', q')$ defined implicitly as follows. For $t = 1, \ldots, T$:

- Sample $\pi^t \sim p^t(\cdot \mid \mathcal{H}^{t-1})$.
- If $|\{i \leq t \mid \pi^i \notin \boldsymbol{\pi}_{\overline{M}}\}| = C$, break and play an arbitrary decision $\pi \in \boldsymbol{\pi}_{\overline{M}}$ until round T.

Return $\widehat{\lambda} \sim q(\cdot \mid \mathcal{H}^T)$.

It is immediate from this construction that $\mathbb{A}' = (p', q')$ has $N_{\neg \boldsymbol{\pi}_{\overline{M}}} \leq C$ almost surely under all possible models $M \in \mathcal{M}^+$. We now focus on bounding the performance. Let T_0 be the greatest value of t for which $|\{i \leq t \mid \pi^i \notin \boldsymbol{\pi}_{\overline{M}}\}| \leq C$. First, observe that for all $M \in \mathcal{M}_0^{\mathrm{opt}}(\overline{M})$, since the algorithms behave identically in law whenever $T_0 = T$,

$$\begin{aligned}
\mathbb{P}^{M, \mathbb{A}'}\left[\widehat{\lambda} \in \Upsilon(M; \varepsilon, \bar{\mathsf{n}}^M)\right] &\geq \mathbb{P}^{M, \mathbb{A}'}\left[\widehat{\lambda} \in \Upsilon(M; \varepsilon, \bar{\mathsf{n}}^M) \wedge T_0 = T\right] \\
&= \mathbb{P}^{M, \mathbb{A}}\left[\widehat{\lambda} \in \Upsilon(M; \varepsilon, \bar{\mathsf{n}}^M) \wedge T_0 = T\right] \\
&= \mathbb{P}^{M, \mathbb{A}}\left[\widehat{\lambda} \in \Upsilon(M; \varepsilon, \bar{\mathsf{n}}^M) \wedge N_{\neg \boldsymbol{\pi}_{\overline{M}}} \leq C\right].
\end{aligned}$$

By the union bound, we have

$$\mathbb{P}^{M,\mathbb{A}}\Big[\widehat{\lambda} \in \Upsilon(M;\varepsilon,\bar{\mathsf{n}}^M) \wedge N_{\neg\boldsymbol{\pi}_{\overline{M}}} \le C\Big] \ge \mathbb{P}^{M,\mathbb{A}}\Big[\widehat{\lambda} \in \Upsilon(M;\varepsilon,\bar{\mathsf{n}}^M)\Big] - \mathbb{P}^{M,\mathbb{A}}\big[N_{\neg\boldsymbol{\pi}_{\overline{M}}} > C\big].$$

Finally, we observe that by Markov's inequality we have

$$\mathbb{P}^{M,\mathbb{A}}\big[N_{\neg\boldsymbol{\pi}_{\overline{M}}} > C\big] \le \frac{\mathbb{E}^{M,\mathbb{A}}\big[N_{\neg\boldsymbol{\pi}_{\overline{M}}}\big]}{C} \le \frac{\mathbb{E}^{M,\mathbb{A}}[N_{\neg\boldsymbol{\pi}_M}]}{C} \le \beta,$$

where we have used that $N_{\neg\boldsymbol{\pi}_{\overline{M}}} \le N_{\neg\boldsymbol{\pi}_M}$, since $\boldsymbol{\pi}_M \subseteq \boldsymbol{\pi}_{\overline{M}}$. Rearranging, we obtain

$$\mathbb{P}^{M,\mathbb{A}}\Big[\widehat{\lambda} \notin \Upsilon(M;\varepsilon,\bar{\mathsf{n}}^M)\Big] \ge \mathbb{P}^{M,\mathbb{A}'}\Big[\widehat{\lambda} \notin \Upsilon(M;\varepsilon,\bar{\mathsf{n}}^M)\Big] - \beta.$$

□

Proof of Lemma H.5.4. Fix $\varepsilon > 0$, and let an algorithm $\mathbb{A}$ be given. For any $\overline{M} \in \mathcal{M}^+$, define $q^{\overline{M}} = \mathbb{P}^{\overline{M},\mathbb{A}}(\widehat{\lambda} = \cdot) \in \triangle(\triangle(\Pi))$, and let $\omega^{\overline{M}} := \mathbb{E}^{\overline{M},\mathbb{A}}\Big[\frac{1}{N_{\neg\boldsymbol{\pi}_{\overline{M}}}} \sum_{t:\pi^t \notin \boldsymbol{\pi}_{\overline{M}}} p^t\Big] \in \triangle(\Pi)$, with the convention that the value inside the expectation is zero whenever $N_{\neg\boldsymbol{\pi}_{\overline{M}}} = 0$.[2]

Fix $\alpha > 0$ and $\overline{M} \in \mathcal{M}^+$ be fixed. Define

$$M = \operatorname*{arg\,max}_{M \in \mathcal{M}_0^{\mathrm{opt}}(\overline{M})} \Big\{\mathbb{P}_{\lambda\sim q^{\overline{M}}}[\lambda \notin \Upsilon(M;\varepsilon,\bar{\mathsf{n}}^M)] \mid \mathbb{E}_{\pi\sim\omega^{\overline{M}}}\big[D_{\mathsf{KL}}\big(\overline{M}(\pi) \,\|\, M(\pi)\big)\big] \le \alpha^2\Big\};$$

we assume that such an $M \in \mathcal{M}_0^{\mathrm{opt}}(\overline{M})$ does exist, as otherwise the claim we will prove is trivial. It is immediate from this definition that we have

$$\begin{aligned}
&\mathbb{P}^{\overline{M},\mathbb{A}}\Big[\widehat{\lambda} \notin \Upsilon(M;\varepsilon,\bar{\mathsf{n}}^M)\Big] \\
&= \mathbb{P}_{\lambda\sim q^{\overline{M}}}[\lambda \notin \Upsilon(M;\varepsilon,\bar{\mathsf{n}}^M)] \\
&= \sup_{M \in \mathcal{M}_0^{\mathrm{opt}}(\overline{M})} \Big\{\mathbb{P}_{\lambda\sim q^{\overline{M}}}[\lambda \notin \Upsilon(M;\varepsilon,\bar{\mathsf{n}}^M)] \mid \mathbb{E}_{\pi\sim\omega^{\overline{M}}}\big[D_{\mathsf{KL}}\big(\overline{M}(\pi) \,\|\, M(\pi)\big)\big] \le \alpha^2\Big\} \\
&\ge \inf_{q\in\triangle(\triangle(\Pi)),\omega\in\triangle(\Pi)} \sup_{M \in \mathcal{M}_0^{\mathrm{opt}}(\overline{M})} \big\{\mathbb{P}_{\lambda\sim q}[\lambda \notin \Upsilon(M;\varepsilon,\bar{\mathsf{n}}^M)] \mid \mathbb{E}_{\pi\sim\omega}\big[D_{\mathsf{KL}}\big(\overline{M}(\pi) \,\|\, M(\pi)\big)\big] \le \alpha^2\big\} =: \mathsf{opt},
\end{aligned}$$

with the convention that this value is zero if the set $\big\{M \in \mathcal{M}_0^{\mathrm{opt}}(\overline{M}) \mid \mathbb{E}_{\pi\sim\omega}\big[D_{\mathsf{KL}}\big(\overline{M}(\pi) \,\|\, M(\pi)\big)\big] \le \alpha^2\big\}$ is empty. In addition, we have

$$\mathbb{E}_{\pi\sim\omega^{\overline{M}}}\big[D_{\mathsf{KL}}\big(\overline{M}(\pi) \,\|\, M(\pi)\big)\big] \le \alpha^2. \tag{H.5.6}$$

[2] If $N_{\neg\boldsymbol{\pi}_{\overline{M}}} = 0$ almost surely under $\overline{M}$, we can take $R = 0$, in which case the statement of the lemma is vacuous.

Now, define $\delta := \frac{\varepsilon}{2} \cdot \min\Big\{1, \inf_{M \in \mathcal{M}_0^{\mathrm{opt}}(\overline{M})} \frac{\mathsf{g}^M}{\bar{\mathsf{n}}^M}\Big\}$, and let $\overline{\lambda}_q := \mathbb{E}_{\lambda \sim q}[\lambda]$. By Lemma H.5.1, we have

$$\begin{aligned}
\mathsf{opt} &= \inf_{q \in \triangle(\triangle(\Pi)), \omega \in \triangle(\Pi)} \sup_{M \in \mathcal{M}_0^{\mathrm{opt}}(\overline{M})} \Big\{ \mathbb{P}_{\lambda \sim q}[\lambda \notin \Upsilon(M; \varepsilon, \bar{\mathsf{n}}^M)] \;\Big|\; \mathbb{E}_{\pi \sim \omega}\big[D_{\mathsf{KL}}\big(\overline{M}(\pi) \,\|\, M(\pi)\big)\big] \le \alpha^2 \Big\} \\
&\ge \delta \cdot \inf_{q \in \triangle(\triangle(\Pi)), \omega \in \triangle(\Pi)} \sup_{M \in \mathcal{M}_0^{\mathrm{opt}}(\overline{M})} \Big\{ \mathbb{I}\big\{\overline{\lambda}_q \notin \Upsilon(M; 2\varepsilon, \bar{\mathsf{n}}^M)\big\} \;\Big|\; \mathbb{E}_{\pi \sim \omega}\big[D_{\mathsf{KL}}\big(\overline{M}(\pi) \,\|\, M(\pi)\big)\big] \le \alpha^2 \Big\} \\
&\ge \delta \cdot \inf_{\lambda \in \triangle(\Pi), \omega \in \triangle(\Pi)} \sup_{M \in \mathcal{M}_0^{\mathrm{opt}}(\overline{M})} \Big\{ \mathbb{I}\{\lambda \notin \Upsilon(M; 2\varepsilon, \bar{\mathsf{n}}^M)\} \;\Big|\; \mathbb{E}_{\pi \sim \omega}\big[D_{\mathsf{KL}}\big(\overline{M}(\pi) \,\|\, M(\pi)\big)\big] \le \alpha^2 \Big\} \\
&\ge \delta \cdot \inf_{\lambda \in \triangle(\Pi), \omega \in \triangle(\Pi)} \sup_{M \in \mathcal{M}_0^{\mathrm{opt}}(\overline{M})} \Big\{ \mathbb{I}\{\lambda \notin \Upsilon(M; 2\varepsilon)\} \;\Big|\; \mathbb{E}_{\pi \sim \omega}\big[D_{\mathsf{KL}}\big(\overline{M}(\pi) \,\|\, M(\pi)\big)\big] \le \alpha^2 \Big\} \\
&= \delta \cdot \mathbb{I}\Big\{\alpha^2 \ge \big(\mathsf{aec}^{\mathcal{M}}_{2\varepsilon}(\mathcal{M}_0^{\mathrm{opt}}(\overline{M}), \overline{M})\big)^{-1}\Big\}. \qquad \text{(H.5.7)}
\end{aligned}$$

Hence, we have

$$\mathbb{P}^{\overline{M}, \mathbb{A}}\Big[\widehat{\lambda} \notin \Upsilon(M; \varepsilon, \bar{\mathsf{n}}^M)\Big] \ge \delta \cdot \mathbb{I}\Big\{\alpha^2 \ge \big(\mathsf{aec}^{\mathcal{M}}_{2\varepsilon}(\mathcal{M}_0^{\mathrm{opt}}(\overline{M}), \overline{M})\big)^{-1}\Big\}. \qquad \text{(H.5.8)}$$

To proceed, using Lemma A.11 of [Foster et al., 2021], we have

$$\begin{aligned}
\mathbb{P}^{M, \mathbb{A}}\Big[\widehat{\lambda} \notin \Upsilon(M; \varepsilon, \bar{\mathsf{n}}^M)\Big] &\ge \frac{1}{3}\mathbb{P}^{\overline{M}, \mathbb{A}}\Big[\widehat{\lambda} \notin \Upsilon(M; \varepsilon, \bar{\mathsf{n}}^M)\Big] - \frac{4}{3} D_{\mathsf{KL}}(\mathbb{P}^{\overline{M}, \mathbb{A}} \,\|\, \mathbb{P}^{M, \mathbb{A}}) \\
&\ge \frac{\delta}{3}\mathbb{I}\Big\{\alpha^2 \ge \big(\mathsf{aec}^{\mathcal{M}}_{2\varepsilon}(\mathcal{M}_0^{\mathrm{opt}}(\overline{M}), \overline{M})\big)^{-1}\Big\} - \frac{4}{3} D_{\mathsf{KL}}(\mathbb{P}^{\overline{M}, \mathbb{A}} \,\|\, \mathbb{P}^{M, \mathbb{A}}).
\end{aligned}$$

Now, recall that from the definition, we have that for all $M \in \mathcal{M}_0^{\mathrm{opt}}(\overline{M})$,

$$\begin{aligned}
D_{\mathsf{KL}}(\mathbb{P}^{\overline{M}, \mathbb{A}} \,\|\, \mathbb{P}^{M, \mathbb{A}}) &= \mathbb{E}^{M, \mathbb{A}}\left[\sum_{t : \pi^t \notin \boldsymbol{\pi}_{\overline{M}}} \mathbb{E}_{\pi \sim p^t} D_{\mathsf{KL}}\big(\overline{M}(\pi) \,\|\, M(\pi)\big)\right] \\
&= \mathbb{E}^{M, \mathbb{A}}\left[\frac{N_{\neg \boldsymbol{\pi}_{\overline{M}}}}{N_{\neg \boldsymbol{\pi}_{\overline{M}}}} \sum_{t : \pi^t \notin \boldsymbol{\pi}_{\overline{M}}} \mathbb{E}_{\pi \sim p^t} D_{\mathsf{KL}}\big(\overline{M}(\pi) \,\|\, M(\pi)\big)\right] \\
&\le R \cdot \mathbb{E}^{M, \mathbb{A}}\left[\frac{1}{N_{\neg \boldsymbol{\pi}_{\overline{M}}}} \sum_{t : \pi^t \notin \boldsymbol{\pi}_{\overline{M}}} \mathbb{E}_{\pi \sim p^t} D_{\mathsf{KL}}\big(\overline{M}(\pi) \,\|\, M(\pi)\big)\right] \\
&= R \cdot \mathbb{E}_{\pi \sim \omega^{\overline{M}}}\big[D_{\mathsf{KL}}\big(\overline{M}(\pi) \,\|\, M(\pi)\big)\big] \le \alpha^2 R,
\end{aligned}$$

where the first inequality uses that $\mathbb{P}^{\overline{M}, \mathbb{A}}\big[N_{\neg \boldsymbol{\pi}_{\overline{M}}} > R\big] = 0$, and the second inequality uses (H.5.6).

With this, we have

$$\mathbb{P}^{M,\mathbb{A}}\Big[\widehat{\lambda} \notin \Upsilon(M;\varepsilon,\bar{\mathsf{n}}^M)\Big] \geq \frac{\delta}{3}\mathbb{I}\Big\{\alpha^2 \geq \big(\mathsf{aec}^{\mathcal{M}}_{2\varepsilon}(\mathcal{M}_0^{\mathrm{opt}}(\overline{M}),\overline{M})\big)^{-1}\Big\} - \frac{4}{3}\alpha^2 R.$$

We set $\alpha^2 = \frac{\delta}{8R}$, so that

$$\begin{aligned}\mathbb{P}^{M,\mathbb{A}}\Big[\widehat{\lambda} \notin \Upsilon(M;\varepsilon,\bar{\mathsf{n}}^M)\Big] &\geq \frac{\delta}{6}\cdot\mathbb{I}\Big\{\alpha^2 \geq \big(\mathsf{aec}^{\mathcal{M}}_{2\varepsilon}(\mathcal{M}_0^{\mathrm{opt}}(\overline{M}),\overline{M})\big)^{-1}\Big\}\\ &= \frac{\delta}{6}\cdot\mathbb{I}\Big\{R \leq \frac{\delta}{8}\cdot\mathsf{aec}^{\mathcal{M}}_{2\varepsilon}(\mathcal{M}_0^{\mathrm{opt}}(\overline{M}),\overline{M})\Big\}.\end{aligned}$$

We conclude that unless

$$R > \frac{\delta}{8}\cdot\sup_{\overline{M}\in\mathcal{M}^+}\mathsf{aec}^{\mathcal{M}}_{2\varepsilon}(\mathcal{M}_0^{\mathrm{opt}}(\overline{M}),\overline{M}),$$

the algorithm must have $\mathbb{P}^{M,\mathbb{A}}\Big[\widehat{\lambda} \notin \Upsilon(M;\varepsilon,\bar{\mathsf{n}}^M)\Big] \geq \frac{\delta}{6}$.

□

H.5.4 Proofs for Lower Bound Examples

Proof of Example 10.4.2. Let $\Delta \in (0,1)$ and $A \geq 2$ be given and set

$$\mathcal{M} = \Big\{M(\pi) = \mathcal{N}(f^M(\pi), 1/2) \mid f^M \in [0,1]^A\Big\}$$

and

$$\mathcal{M}_0 = \{M \in \mathcal{M} \ : \ \Delta^M_{\min} \geq \Delta/2\}$$

Define $\overline{M} \in \mathcal{M}$ via $f^{\overline{M}}(\pi) = \Delta\mathbb{I}\{\pi = A\}$. Fix $\varepsilon \in (0,1/2)$ and define a subclass

$$\mathcal{M}' = \{\overline{M}\} \cup \{M_i\}_{i\in[A-1]}$$

via

$$f^{M_i}(\pi) = \Delta\mathbb{I}\{\pi = A\} + \varepsilon\cdot\Delta\mathbb{I}\{\pi = i\}.$$

Since $\varepsilon \leq 1/2$, we have $\mathcal{M}' \subseteq \mathcal{M}^{\mathrm{opt}}(\overline{M})$ and $\mathcal{M}' \subseteq \mathcal{M}_0$. In addition, we have

$$D_{\mathsf{KL}}\big(\overline{M}(\pi) \,\|\, M_i(\pi)\big) = (f^{\overline{M}}(\pi) - f^{M_i}(\pi))^2.$$

Let $\mathcal{M}'' \subseteq \mathcal{M}$ denote the set of instances such that, for $M' \in \mathcal{M}''$, $D_{\mathsf{KL}}(M_i(A) \,\|\, M'(A)) = 0$, and $M' \in \mathcal{M}^{\text{alt}}(M_i)$, for all $i \in [A-1]$. Then,

$$\begin{aligned} I^{M_i}(\lambda; \mathcal{M}) &= \inf_{M' \in \mathcal{M}^{\text{alt}}(M_i)} \mathbb{E}_{\pi \sim \lambda}[D_{\mathsf{KL}}(M_i(\pi) \,\|\, M'(\pi))] \\ &\leq \inf_{M' \in \mathcal{M}''} \mathbb{E}_{\pi \sim \lambda}[D_{\mathsf{KL}}(M_i(\pi) \,\|\, M'(\pi))] \\ &= \min_{j \in [A-1]} \big\{ \lambda_i \cdot (1-\varepsilon)^2 \Delta^2 \mathbb{I}\{j = i\} + \lambda_j \cdot \Delta^2 \mathbb{I}\{j \neq i\} \big\}. \end{aligned} \tag{H.5.9}$$

We also have that

$$\mathsf{g}^{M_i} = \mathsf{g} := \frac{(A-2)}{\Delta} + \frac{1}{(1-\varepsilon)\Delta},$$

where we have used again that $\varepsilon \leq 1/2$.

Fix any pair $\lambda, \omega \in \triangle_\Pi$ and consider the value

$$\sup_{M \in \mathcal{M}' \setminus \mathcal{M}_\varepsilon^{\text{gl}}(\lambda)} \left\{ \frac{1}{\mathbb{E}_{\pi \sim \omega}\big[D_{\mathsf{KL}}\big(\overline{M}(\pi) \,\|\, M(\pi)\big)\big]} \right\}.$$

Pick $\delta \leq \varepsilon/32$, and let $\mathcal{J} \subseteq [A-1]$ be the set of models i for which $\lambda \in \Upsilon(M_i; \delta)$. For each such model, by definition, there exists $\mathsf{n}_i > 0$ such that

$$\Delta^{M_i}(\lambda) \leq (1+\delta)\frac{\mathsf{g}}{\mathsf{n}_i}, \quad \text{and} \quad I^{M_i}(\lambda; \mathcal{M}) \geq (1-\delta)\frac{1}{\mathsf{n}_i}.$$

Define $\bar{\mathsf{n}} = \max_{i \in \mathcal{J}} \mathsf{n}_i$ and $\underline{\mathsf{n}} = \min_{i \in \mathcal{J}} \mathsf{n}_i$. Using (H.5.9), it is immediate that for all $j \in [A-1]$, $\lambda_j \geq (1-\delta)\frac{\Delta^{-2}}{\underline{\mathsf{n}}}$, and that for all $i \in \mathcal{J}$,

$$\lambda_i \geq (1-\delta)\frac{(1-\varepsilon)^{-2}\Delta^{-2}}{\mathsf{n}_i}.$$

In particular, this implies that for all $i \in \mathcal{J}$,

$$\begin{aligned} (1-\delta)\frac{(A-2)\Delta^{-1}}{\underline{\mathsf{n}}} + (1-\delta)\frac{(1-\varepsilon)^{-1}\Delta^{-1}}{\mathsf{n}_i} &\leq \Delta^{M_i}(\lambda) \\ &\leq (1+\delta)\frac{\mathsf{g}}{\mathsf{n}_i} \\ &= (1+\delta)\frac{(A-2)\Delta^{-1}}{\mathsf{n}_i} + (1+\delta)\frac{(1-\varepsilon)^{-1}\Delta^{-1}}{\mathsf{n}_i}, \end{aligned}$$

or by rearranging,

$$(1-\delta)\frac{(A-2)\Delta^{-1}}{\underline{\mathsf{n}}} \leq (1+\delta)\frac{(A-2)\Delta^{-1}}{\mathsf{n}_i} + 2\delta\frac{(1-\varepsilon)^{-1}\Delta^{-1}}{\mathsf{n}_i},$$

$$\leq (1+\delta)\frac{(A-2)\Delta^{-1}}{\mathsf{n}_i} + 4\delta\frac{\Delta^{-1}}{\mathsf{n}_i},$$
$$\leq (1+2\delta)\frac{(A-2)\Delta^{-1}}{\mathsf{n}_i}$$

as long as $A \geq 6$. Since this holds uniformly for all $i \in \mathcal{J}$, rearranging once more gives

$$\bar{\mathsf{n}} \leq \frac{(1+2\delta)}{1-\delta}\underline{\mathsf{n}} \leq (1+2\delta)^2\underline{\mathsf{n}} \leq (1+8\delta)\underline{\mathsf{n}},$$

where we have used that $\delta \leq 1/2$.

Now, observe that for all $i \in \mathcal{J}$, we have

$$\begin{aligned}
\Delta^{M_i}(\lambda) &\geq (1-\delta)(A-|\mathcal{J}|-1)\frac{1}{\Delta\underline{\mathsf{n}}} + (1-\delta)|\mathcal{J}|\frac{1}{(1-\varepsilon)^2\Delta\bar{\mathsf{n}}} + (1-\delta)\frac{1}{(1-\varepsilon)\Delta\mathsf{n}_i},\\
&\geq (1-\delta)(A-|\mathcal{J}|-1)\frac{1}{\Delta\underline{\mathsf{n}}} + \frac{(1-\delta)}{1+8\delta}|\mathcal{J}|\frac{1}{(1-\varepsilon)^2\Delta\underline{\mathsf{n}}} + (1-\delta)\frac{1}{(1-\varepsilon)\Delta\mathsf{n}_i},\\
&\geq \frac{(1-\delta)}{\Delta\underline{\mathsf{n}}}\left((A-|\mathcal{J}|-1) + |\mathcal{J}|\frac{1-8\delta}{(1-\varepsilon)^2}\right) + (1-\delta)\frac{1}{(1-\varepsilon)\Delta\mathsf{n}_i},\\
&\geq \frac{(1-\delta)}{\Delta\underline{\mathsf{n}}}\left((A-|\mathcal{J}|-1) + |\mathcal{J}|\frac{1}{(1-\varepsilon)}\right) + (1-\delta)\frac{1}{(1-\varepsilon)\Delta\mathsf{n}_i},
\end{aligned}$$

where we have used that $\frac{1}{1+x} \geq 1-x$, and that $\delta \leq \varepsilon/8$. Suppose that $|\mathcal{J}| \geq \frac{A}{2}$. Then we have

$$\begin{aligned}
(A-|\mathcal{J}|-1) + |\mathcal{J}|\frac{1}{(1-\varepsilon)} \geq \frac{A}{2}\left(1+\frac{1}{1-\varepsilon}\right) - 1 &\geq (1+\varepsilon/2)A - 1\\
&\geq (1+\varepsilon/2)(A-1),
\end{aligned}$$

so that

$$\Delta^{M_i}(\lambda) \geq \frac{(1-\delta)(1+\varepsilon/2)}{\Delta\underline{\mathsf{n}}}(A-1) + (1-\delta)\frac{1}{(1-\varepsilon)\Delta\mathsf{n}_i}.$$

Noting that $\underline{\mathsf{n}} \leq \mathsf{n}_i$ and $\delta \leq \varepsilon/8$, we further have

$$\Delta^{M_i}(\lambda) \geq (1+\varepsilon/4)\frac{A-1}{\Delta}\frac{1}{\mathsf{n}_i} + (1-\delta)\frac{1}{(1-\varepsilon)\Delta\mathsf{n}_i}.$$

Observe that the right-hand side above is greater than $(1+\delta)\frac{\mathsf{g}}{\mathsf{n}_i}$ if and only if

$$(1+\varepsilon/4)(A-1) > (1+\delta)(A-1) + \frac{2\delta}{1-\varepsilon},$$

which is satisfied if $\delta \leq \varepsilon/32$. In this case, we have

$$\Delta^{M_i}(\lambda) > (1+\delta)\frac{\mathsf{g}}{\mathsf{n}_i},$$

which contradicts the assumption that $i \in \mathcal{J}$. It follows that we must have $|\mathcal{J}| < A/2$.

Now, to conclude, select $i = \arg\min_{i \in [A-1]\setminus\mathcal{J}} \omega_i$, and consider the value

$$\sup_{M \in \mathcal{M}' \setminus \mathcal{M}_{\varepsilon}^{\mathrm{gl}}(\lambda)} \left\{ \frac{1}{\mathbb{E}_{\pi\sim\omega}\left[D_{\mathsf{KL}}\left(\overline{M}(\pi) \,\|\, M(\pi)\right)\right]} \right\} \geq \frac{1}{\mathbb{E}_{\pi\sim\omega}\left[D_{\mathsf{KL}}\left(\overline{M}(\pi) \,\|\, M_i(\pi)\right)\right]} = \frac{1}{\omega_i \cdot \varepsilon^2\Delta^2},$$

where the first inequality follows because $\lambda \notin \Upsilon(M_i;\delta)$ by definition, and the equality follows from the construction of $\overline{M}$ and M_i. Since $\sum_{i\in[A-1]\setminus\mathcal{J}} \omega_i \leq 1$ and $|[A-1]\setminus\mathcal{J}| \geq \frac{A}{2}$, we must have $\omega_i \leq \frac{2}{A}$, so that

$$\frac{1}{\omega_i \cdot \varepsilon^2\Delta^2} \geq \frac{A}{2\varepsilon^2\Delta^2}$$

as desired. To complete the proof, note that this holds uniformly for all choices for λ and ω, and that

$$\begin{aligned}
\mathsf{aec}_{\varepsilon}^{\mathcal{M}}(\mathcal{M}_0, \overline{M}) &= \inf_{\lambda,\omega\in\triangle_{\Pi}} \sup_{M\in\mathcal{M}_0\setminus\mathcal{M}_{\varepsilon}^{\mathrm{gl}}(\lambda)} \left\{ \frac{1}{\mathbb{E}_{\pi\sim\omega}\left[D_{\mathsf{KL}}\left(\overline{M}(\pi) \,\|\, M(\pi)\right)\right]} \right\} \\
&\geq \inf_{\lambda,\omega\in\triangle_{\Pi}} \sup_{M\in\mathcal{M}'\setminus\mathcal{M}_{\varepsilon}^{\mathrm{gl}}(\lambda)} \left\{ \frac{1}{\mathbb{E}_{\pi\sim\omega}\left[D_{\mathsf{KL}}\left(\overline{M}(\pi) \,\|\, M(\pi)\right)\right]} \right\}.
\end{aligned}$$

To obtain the parameter setting in the theorem statement, we rescale $\Delta \leftarrow 2\Delta$ and $\varepsilon \leftarrow 32\varepsilon$. □

Proof of Example 10.4.3. We reduce the lower bound to that of multi-armed bandits via a standard tree construction [Osband and Van Roy, 2016, Domingues et al., 2021]; as the argument is standard, we only sketch the approach. Assume without loss of generality that H is a multiple of 2. Set $H = \log_2(S/2)$. Consider a sub-class $\mathcal{M}' \subseteq \mathcal{M}$ defined as follows. All models $M \in \mathcal{M}'$ have identical, deterministic dynamics given by a binary tree. Each layer h has 2^{h-1} states, so that layer H has $S/4$ states, and the total number of states is $S-1$. The agent begins from a root state s_1 deterministically. For $h \leq H-1$, there are two available actions, left and right. Choosing left leads to the left successor for the current layer, and right leads to the right successor for the current layer. There are no rewards for layer $h \leq H-1$. For layer H, there are A available actions, and rewards are arbitrary, subject to the constraint that the mean lies in $[0,1]$ and the noise follows $\mathcal{N}(0,1/2)$.

It is clear that the class $\mathcal{M}'$ is equivalent to the class of multi-armed bandit instances

with $SA/4$ actions. As a consequence, the lower bound follows from Example 10.4.2.

□

Proof of Example 10.4.1. Let $\Delta \in (0, 1/6)$, $\beta \in (0, 1)$, and $A, N \geq 2$ be given. Consider the reference model $\overline{M} \in \mathcal{M}^+$ defined as follows:

- For each bandit arm $k \in [A]$, we have $f^{\overline{M}}(k) = \frac{1}{2} + \Delta \mathbb{I}\{k = A\}$ and $r \sim \mathcal{N}(f^{\overline{M}}(k), 1)$. There are no observations, i.e. $o = \perp$ almost surely.

- For each revealing arm π_k°, we receive zero reward almost surely (so $f^{\overline{M}}(\pi_k^\circ) = 0$) and $o \sim \mathrm{Unif}([A])$.

We define a subclass

$$\mathcal{M}' = \{M_j\}_{j \in [N]} \subset \mathcal{M}_0^{\mathrm{opt}}(\overline{M})$$

as follows

- For each bandit arm $k \in [A]$, we have $f^{M_j}(k) = \frac{1}{2} + \Delta \mathbb{I}\{k = A\}$ and $r \sim \mathcal{N}(f^{M_j}(k), 1)$. There are no observations, i.e. $o = \perp$ almost surely.

- For each revealing arm π_k°, we receive zero reward almost surely (so $f^{M_j}(\pi_k^\circ) = 0$). We have

$$o \sim \begin{cases} \mathrm{Unif}([A]), & k \neq j, \\ \beta \mathbb{I}_i + (1 - \beta)\mathrm{Unif}([A]), & k = j. \end{cases}$$

Note that $\mathcal{M}' \subseteq \mathcal{M}_0$. For all $j \in [N]$, a direct calculation gives

$$D_{\mathsf{KL}}\big(\overline{M}(\pi_j^\circ) \,\|\, M_j(\pi_j^\circ)\big) = \frac{A-1}{A} \log\left(\frac{1}{1-\beta}\right) + \frac{1}{A} \log\left(\frac{1}{1+\beta(A-1)}\right) =: \alpha$$

and

$$D_{\mathsf{KL}}\big(\overline{M}(\pi) \,\|\, M_j(\pi)\big) = \alpha \cdot \mathbb{I}\{\pi = \pi_j^\circ\}. \tag{H.5.10}$$

In addition, it is straightforward to see that $\alpha \leq 2\beta$ whenever $\beta \leq 1/2$. Next we calculate that for any $j \in [N]$ and $M \in \mathcal{M}$ with $\pi_M^\circ = \pi_j^\circ$, and $\pi_M \neq A$,

$$D_{\mathsf{KL}}\big(M_j(\pi_j^\circ) \,\|\, M(\pi_j^\circ)\big) = \beta \log\left(1 + \frac{\beta A}{1-\beta}\right) =: \gamma,$$

which has $\gamma \le O(\beta)$ whenever $\beta \le 1/A$ and $\gamma \ge \beta \log(1+\beta A)$. Lastly, we have that for all $i \in [A]$, all $M, M' \in \mathcal{M}$ have

$$D_{\mathsf{KL}}(M(i) \,\|\, M'(i)) = \frac{1}{2}(f^{M}(i) - f^{M'}(i))^2.$$

Let $\mathcal{M}''$ denote the set of instances such that for $M' \in \mathcal{M}''$, $\pi_{M'} \neq A$, and $f^{M'}(A) = \frac{1}{2} + \Delta$, so that $D_{\mathsf{KL}}(M_j(A) \,\|\, M'(A)) = 0$ and $\mathcal{M}'' \subseteq \mathcal{M}^{\mathrm{alt}}(M_j)$ for all $j \in [N]$. Using the above calculations and the definition of $I^{M_j}(\lambda; \mathcal{M})$, we can then compute, for all $j \in [N]$,

$$I^{M_j}(\lambda; \mathcal{M}) \le \inf_{M' \in \mathcal{M}''} \mathbb{E}_\lambda[D_{\mathsf{KL}}(M_j(\pi) \,\|\, M'(\pi))] = \frac{\Delta^2}{2} \cdot \min_{k \in [A-1]} \lambda_k + \gamma \cdot \lambda_{\pi_j^\circ} \tag{H.5.11}$$

and

$$\mathsf{g}^{M_j} = \mathsf{g} := \min\left\{2\frac{A-1}{\Delta}, \left(\frac{1}{2} + \Delta\right)\frac{1}{\gamma}\right\} = \left(\frac{1}{2} + \Delta\right)\frac{1}{\gamma}$$

whenever $\gamma \ge \Delta/2(A-1)$.

Fix any pair $\lambda, \omega \in \triangle_\Pi$ and consider the value

$$\sup_{M \in \mathcal{M}' \setminus \mathcal{M}^{\mathrm{gl}}_{1/2}(\lambda)} \left\{ \frac{1}{\mathbb{E}_{\pi \sim \omega}\left[D_{\mathsf{KL}}\left(\overline{M}(\pi) \,\|\, M(\pi)\right)\right]} \right\}.$$

Let $\mathcal{J} \subseteq [N]$ be the set of models j for which $\lambda \in \Upsilon(M_j; 1/2)$. For each such model, by definition, there exists $\mathsf{n}_j > 0$ such that

$$\Delta^{M_j}(\lambda) \le (1 + 1/2)\frac{\mathsf{g}}{\mathsf{n}_j} \le \frac{1}{\gamma \mathsf{n}_j}, \quad \text{and} \quad I^{M_j}(\lambda; \mathcal{M}) \ge \frac{1}{2\mathsf{n}_j}.$$

Define $\bar{\mathsf{n}} = \max_{j \in \mathcal{J}} \mathsf{n}_j$ and $\underline{\mathsf{n}} = \min_{j \in \mathcal{J}} \mathsf{n}_j$. Let us begin with some basic observations.

- Since $\Delta^{M_j} = \Delta^{\overline{M}}$ for all j, we have $\Delta^{M_j}(\lambda) \le \frac{1}{\gamma \mathsf{n}_{j'}}$ for all $j, j' \in \mathcal{J}$, and hence

$$\Delta^{M_j}(\lambda) \le \frac{1}{\gamma \bar{\mathsf{n}}}. \tag{H.5.12}$$

- Any $j \in \mathcal{J}$ must have

$$\lambda_{\pi_j^\circ} \gamma \ge \frac{1}{4\mathsf{n}_j} \ge \frac{1}{4\bar{\mathsf{n}}}. \tag{H.5.13}$$

  To see this, observe that if it were not the case, we would need $\min_{i \in [A-1]} \lambda_i \frac{\Delta^2}{2} \ge \frac{1}{4\mathsf{n}_j}$ to satisfy the constraint that $I^{M_j}(\lambda; \mathcal{M}) \ge \frac{1}{2\mathsf{n}_j}$ (by (H.5.11)). But if this were to occur, we

would have

$$\frac{A-1}{2\Delta \mathsf{n}_j} \le \Delta^{M_j}(\lambda) \le \frac{1}{\gamma \mathsf{n}_j},$$

which would contradict the assumption that $\gamma \ge 2\Delta/(A-1)$.

Combing the inequalities (H.5.12) and (H.5.13), it follows that any $j \in \mathcal{J}$ must have

$$\frac{|\mathcal{J}|}{8\gamma \mathsf{n}_j} \le \frac{1}{2}\sum_{k\in\mathcal{J}} \lambda_{\pi_k^\circ} \le \Delta^{M_j}(\lambda) \le \frac{1}{\gamma \mathsf{n}_j},$$

which implies that $|\mathcal{J}| \le 8$. Hence, as long as $N \ge 16$, we have $|[N] \setminus \mathcal{J}| \ge N/2$.

To conclude, select $k = \arg\min_{j\in[N]\setminus\mathcal{J}} \omega_{\pi_j^\circ}$, and consider the value

$$\sup_{M\in\mathcal{M}'\setminus\mathcal{M}^{\mathrm{gl}}_{1/2}(\lambda)} \left\{ \frac{1}{\mathbb{E}_{\pi\sim\omega}\left[D_{\mathsf{KL}}\left(\overline{M}(\pi) \,\|\, M(\pi)\right)\right]} \right\} \ge \frac{1}{\mathbb{E}_{\pi\sim\omega}\left[D_{\mathsf{KL}}\left(\overline{M}(\pi) \,\|\, M_k(\pi)\right)\right]} = \frac{1}{\omega_{\pi_k^\circ}\cdot\alpha},$$

where the first inequality follows because $\lambda \notin \Upsilon(M_k; 1/2)$ by definition, and the equality follows from (H.5.10). Since $\sum_{j\in[N]\setminus\mathcal{J}} \omega_{\pi_j^\circ} \le 1$ and $|[N]\setminus\mathcal{J}| \ge \frac{N}{2}$, we must have $\omega_{\pi_k^\circ} \le \frac{2}{N}$, so that

$$\frac{1}{\omega_{\pi_k^\circ}\cdot\alpha} \ge \frac{N}{2\alpha}.$$

as desired. Since this holds uniformly for all choices for λ and ω, the proof is completed. □

H.5.5 Lower Bound on Regret for Algorithms with Well-Behaved Tails

In this section, we present an additional result, Theorem H.6, which shows that for algorithms for which the tail behavior is "well-behaved" in a certain sense, the Allocation-Estimation Coefficient directly leads to lower bounds on the least possible value of T for which any algorithm can achieve (approximate) instance-optimality.

Theorem H.6. *Let the time horizon $T \in \mathbb{N}$, $\varepsilon \in (0, 1/2)$, and $\mathcal{M}_0 \subseteq \mathcal{M}$ be given. Suppose that there exists an algorithm $\mathbb{A}$ with the property that for all $M \in \mathcal{M}_0$,*

1. $\mathbb{E}^{M,\mathbb{A}}[\mathbf{Reg}(T)] \le (1+\varepsilon)\mathsf{g}^M \log(T)$.
2. *For all $\pi \in \Pi$, if $\mathbb{E}^{M,\mathbb{A}}[T(\pi)] \neq 0$, then $\mathbb{E}^{M,\mathbb{A}}[T(\pi)] \ge 1$.*
3. $\sqrt{\mathbb{E}^{M,\mathbb{A}}[(\mathbf{Reg}(T))^2]} \le 2\mathsf{g}^M \log(T)$.

In addition, assume that 1) $\mathsf{g}^M \geq 1$ *for all* $M \in \mathcal{M}_0$*, 2) Assumption 10.7 holds, 3) Assumption 10.5 holds with parameter* $V_{\mathcal{M}} \geq 1$*, and 4) that*

$$\log(T) \geq \frac{12}{\varepsilon} \log\left(\sup_{M \in \mathcal{M}} \frac{2\mathsf{g}^M}{\Delta^M_{\min}} \cdot \log(T) \right).$$

Then if we define $\delta = \varepsilon \cdot \min\{1, \inf_{M \in \mathcal{M}_0} \frac{\mathsf{g}^M}{3\mathsf{g}^M/\Delta^M_{\min} + \mathsf{n}^{\mathcal{M}}_{\varepsilon}}\}$*, it must be the case that*

$$\log^3(T) \geq \frac{\delta^2}{C} \cdot \sup_{\overline{M} \in \mathcal{M}^+} \mathsf{aec}^{\mathcal{M}}_{4\varepsilon}(\mathcal{M}^{\mathrm{opt}}_0(\overline{M}), \overline{M}),$$

for $C \leq O\left((\sup_{M \in \mathcal{M}} \frac{\mathsf{g}^M}{\Delta^M_{\min}})^4 \cdot \frac{V^2_{\mathcal{M}} \log(\delta^{-1})}{\varepsilon^2} \right)$.

We prove Theorem H.6 by combining Theorem 10.10 with an another technical result, Proposition H.5 (stated and proven in the sequel), which shows that for algorithms that satisfy the assumptions of Theorem H.6, it is possible to use the empirical decision frequencies to compute an allocation that is approximately optimal with high probability.

Proof of Theorem H.6. Define $\bar{\mathsf{n}}^M = 3\frac{\mathsf{g}^M}{\Delta^M_{\min}} + \mathsf{n}^{\mathcal{M}}_{\varepsilon}$, and set

$$\delta := \varepsilon \cdot \min\left\{ 1, \inf_{M \in \mathcal{M}_0} \frac{\mathsf{g}^M}{\bar{\mathsf{n}}^M} \right\}.$$

Assume that

$$\log(T) \geq \frac{12}{\varepsilon} \log\left(\sup_{M \in \mathcal{M}} \frac{2\mathsf{g}^M}{\Delta^M_{\min}} \cdot \log(T) \right).$$

Let $\mathbb{A}$ be the algorithm in the statement of the proposition, and let $\mathbb{A}'$ be the modified algorithm created through Proposition H.5 with parameter δ. By assumption, we have that $\sqrt{\mathbb{E}^{M,\mathbb{A}}[N_{\neg \boldsymbol{\pi}_M}]} \leq R := 2 \sup_{M \in \mathcal{M}} \frac{\mathsf{g}^M}{\Delta^M_{\min}} \log(T)$. We define $n = c \cdot \frac{\log(24\delta^{-1})}{\varepsilon^2} \cdot \frac{R^3 V^2_{\mathcal{M}}}{\log(T)}$ for a sufficiently large numerical constant c and $T' = T \cdot n$. Proposition H.5 implies that for time T', the algorithm $\mathbb{A}'$ satisfies

$$\sqrt{\mathbb{E}^{M,\mathbb{A}'}[N_{\neg \boldsymbol{\pi}_M}]} \leq R' := R \cdot n$$

and

$$\mathbb{P}^{M,\mathbb{A}'}\left[\widehat{\lambda} \in \Upsilon(M; 2\varepsilon, \bar{\mathsf{n}}^M) \right] \geq 1 - \frac{\delta}{24}.$$

On the other hand, since the precondition of Theorem H.5 is now satisfied with parameter R', we have that unless

$$R' \geq \frac{\delta^2}{192} \cdot \sup_{\overline{M} \in \mathcal{M}^+} \mathsf{aec}^{\mathcal{M}}_{4\varepsilon}(\mathcal{M}^{\mathrm{opt}}_0(\overline{M}), \overline{M}), \tag{H.5.14}$$

the algorithm must have

$$\mathbb{P}^{M,\mathbb{A}'}\left[\widehat{\lambda} \in \Upsilon(M; 2\varepsilon, \bar{\mathsf{n}}^M)\right] \leq 1 - \frac{\delta}{12}.$$

As $\frac{\delta}{12} > \frac{\delta}{24}$, this is a contradiction unless (H.5.14) holds.

□

Proposition H.5. *Let the time horizon $T \in \mathbb{N}$ and $\mathcal{M}_0 \subseteq \mathcal{M}$ be given. Let $\mathbb{A}$ be an algorithm with the property that for all $M \in \mathcal{M}_0$,*

1. $\mathbb{E}^{M,\mathbb{A}}[\mathbf{Reg}(T)] \leq (1+\varepsilon)\mathsf{g}^M \log(T)$ *for some* $\varepsilon \in (0,1)$.
2. *For all* $\pi \in \Pi$, *if* $\mathbb{E}^{M,\mathbb{A}}[T(\pi)] \neq 0$, *then* $\mathbb{E}^{M,\mathbb{A}}[T(\pi)] \geq 1$.
3. $\sqrt{\mathbb{E}^{M,\mathbb{A}}\left[N^2_{\neg \boldsymbol{\pi}_M}\right]} \leq R$ *for some* $R \geq 2$.

In addition, assume that 1) $\mathsf{g}^M \geq 1$ *for all* $M \in \mathcal{M}_0$, *2) Assumption 10.7 holds, 3) Assumption 10.5 holds with parameter* $V_{\mathcal{M}} \geq 1$, *and 4) that*

$$\log(T) \geq \frac{12}{\varepsilon} \log\left(\sup_{M \in \mathcal{M}} \frac{2\mathsf{g}^M}{\Delta^M_{\min}} \cdot \log(T)\right).$$

Then for any $\delta \in (0, e^{-1})$, *if we define* $n = c \cdot \frac{\log(\delta^{-1})}{\varepsilon^2} \cdot \frac{R^3 V^2_{\mathcal{M}}}{\log(T)}$ *for a sufficiently large numerical constant* c, *there exists an algorithm* $\mathbb{A}'$ *that, using* $T' := T \cdot n$ *rounds, returns a normalized allocation* $\widehat{\lambda} \in \triangle_\Pi$ *such that*

$$\mathbb{P}^{M,\mathbb{A}'}\left[\widehat{\lambda} \in \Upsilon(M; 2\varepsilon, \bar{\mathsf{n}}^M)\right] \geq 1 - \delta,$$

for $\bar{\mathsf{n}}^M \leq 3\frac{\mathsf{g}^M}{\Delta^M_{\min}} + \mathsf{n}^{\mathcal{M}}_\varepsilon$ *and that* $\sqrt{\mathbb{E}^{M,\mathbb{A}'}\left[N^2_{\neg \boldsymbol{\pi}_M}\right]} \leq R \cdot n$ *and*

$$\mathbb{E}^{M,\mathbb{A}'}[\mathbf{Reg}(T')] \leq (1+\varepsilon)\mathsf{g}^M \log(T) \cdot n$$

for all $M \in \mathcal{M}_0$.

Proof of Proposition H.5. We first state a technical lemma regarding robust mean estimation.

Lemma H.5.5. *Let $X \in \mathbb{R}^d$ be a random variable with $\mu := \mathbb{E}[X]$. Assume that $\|\mu\|_0 \leq s$, where s is a known upper bound. For any $\delta \leq e^{-1}$, there exists an estimator $\widehat{\mu}_n$ that, given n independent samples from X, ensures that with probability at least $1-\delta$,*

$$\|\widehat{\mu}_n - \mu\|_1 \leq 24\sqrt{\frac{2s \cdot \mathbb{E}\|X-\mu\|_2^2 \cdot \log(\delta^{-1})}{n}}.$$

In addition, $\|\widehat{\mu}_n\|_0 \le s$ *with probability* 1.

Throughout, we will use that since Assumption 10.7 holds, π_M is unique for all $M \in \mathcal{M}$. Fix $M \in \mathcal{M}_0$. Let $\widehat{\eta} \in \mathbb{R}_+^{\Pi}$ denote the vector of empirical decision frequencies when $\mathbb{A}$ is run with horizon T, i.e. $\widehat{\eta}(\pi) = T(\pi)$. Let

$$\eta^M = \mathbb{E}^{M,\mathbb{A}}[\widehat{\eta}].$$

For parameters $n \in \mathbb{N}$ and $\delta \le e^{-1}$ we define $\mathbb{A}'$ as follows:

- Run $\mathbb{A}$ a total of n times independently (so that $T' = T \cdot n$), and let $\widehat{\eta}^1, \dots, \widehat{\eta}^n$ be the empirical decision frequencies.
- Apply the algorithm from Lemma H.5.5 to $\widehat{\eta}^1, \dots, \widehat{\eta}^n$ with parameters δ and $s = 2R$, and let $\check{\eta} \in \mathbb{R}_+^{\Pi}$ be the resulting vector (note that we can take $\check{\eta}$ to have non-negative entries without loss of generality).
- Set $\widehat{\pi} = \arg\max_{\pi \in \Pi} \check{\eta}$, and set $\widetilde{\eta}(\pi) = \check{\eta}(\pi)\mathbb{I}\{\pi \neq \widehat{\pi}\}$ and $\widetilde{\eta}(\widehat{\pi}) = \mathsf{n}_\varepsilon^{\mathcal{M}} \cdot \log T$ (note that $\mathsf{n}_\varepsilon^{\mathcal{M}}$ is a class-dependent quantity, and so is known to the learner).
- Set $\widehat{\lambda} = \widetilde{\eta}/\|\widetilde{\eta}\|_1$.

It is immediate that this construction satisfies $\sqrt{\mathbb{E}^{M,\mathbb{A}'}\left[N_{\neg\pi_M}^2\right]} \le R \cdot n$ and

$$\mathbb{E}^{M,\mathbb{A}'}[\mathbf{Reg}(T')] \le (1+\varepsilon)\mathbf{g}^M \log(T) \cdot n,$$

so it remains to show that $\widehat{\lambda}$ is a near-optimal allocation with high probability when n is chosen appropriately.

We start by applying Lemma H.5.5. To do so, we carry out some prerequisite calculations. First, by assumption, we have $\eta^M(\pi) \ge 1$ if $\eta^M \neq 0$. Using this, along with Assumption 10.7, we have

$$\|\eta^M\|_0 \le 1 + \sum_{\pi \neq \pi_M} \eta^M(\pi) \le 1 + \mathbb{E}^{M,\mathbb{A}}[N_{\neg\pi_M}] \le 1 + R \le 2R.$$

Second,

$$\begin{aligned}\mathbb{E}^{M,\mathbb{A}}\|\widehat{\eta} - \eta^M\|_2^2 &\le \mathbb{E}^{M,\mathbb{A}}\left[\left(\widehat{\eta}(\pi_M) - \eta^M(\pi_M)\right)^2\right] + \sum_{\pi \neq \pi_M} \mathbb{E}^{M,\mathbb{A}}\left[\widehat{\eta}(\pi)^2\right] \\ &\le \mathbb{E}^{M,\mathbb{A}}\left[\left(\widehat{\eta}(\pi_M) - \eta^M(\pi_M)\right)^2\right] + \mathbb{E}^{M,\mathbb{A}}\left[N_{\neg\pi_M}^2\right].\end{aligned}$$

Furthermore,

$$\begin{aligned}\mathbb{E}^{M,\mathbb{A}}\Big[(\widehat{\eta}(\pi_{M}) - \eta^{M}(\pi_{M}))^2\Big] &= \mathbb{E}^{M,\mathbb{A}}\left[\left(\sum_{\pi\neq\pi_M}\widehat{\eta}(\pi) - \sum_{\pi\neq\pi_M}\eta^{M}(\pi)\right)^2\right] \\ &\leq \mathbb{E}^{M,\mathbb{A}}\left[\left(\sum_{\pi\neq\pi_M}\widehat{\eta}(\pi)\right)^2\right] \\ &= \mathbb{E}^{M,\mathbb{A}}\big[N^2_{\neg\pi_M}\big].\end{aligned}$$

so that

$$\mathbb{E}^{M,\mathbb{A}}\|\widehat{\eta} - \eta^{M}\|_2^2 \leq 2\,\mathbb{E}^{M,\mathbb{A}}\big[N^2_{\neg\pi_M}\big] \leq 2R^2.$$

As a result, Lemma H.5.5 implies that with probability $1-\delta$,

$$\|\check{\eta} - \eta^{M}\|_1 \leq \sqrt{C_1\frac{\log(\delta^{-1})}{n}} =: \varepsilon_{\mathrm{stat}} \tag{H.5.15}$$

where $C_1 = O(R^3)$.

Next, we appeal to Lemma H.5.2, which implies that as long as

$$\log(T) \geq \frac{12}{\varepsilon}\log\left(\sup_{M\in\mathcal{M}}\frac{2\mathsf{g}^{M}}{\Delta^{M}_{\min}}\cdot\log(T)\right), \tag{H.5.16}$$

we have

$$\Delta^{M}(\eta^{M}) \leq (1+\varepsilon/2)\mathsf{g}^{M}\log(T), \quad \text{and} \quad I^{M}(\eta^{M};\mathcal{M}) \geq (1-\varepsilon/2)\log(T).$$

Applying (H.5.15) and using that $\Delta^{M} \in [0,1]$ and $D_{\mathsf{KL}}(M(\pi)\,\|\,M'(\pi)) \leq 2V_{\mathcal{M}}$ by Lemma H.3.13, we have

$$\Delta^{M}(\check{\eta}) \leq (1+\varepsilon/2)\mathsf{g}^{M}\log(T) + \varepsilon_{\mathrm{stat}}, \quad \text{and} \quad I^{M}(\check{\eta};\mathcal{M}) \geq (1-\varepsilon/2)\log(T) - \varepsilon_{\mathrm{stat}}\cdot 2V_{\mathcal{M}}.$$

Thus, as long as

$$\varepsilon_{\mathrm{stat}} \leq \frac{1}{2}\varepsilon\cdot\min\big\{\mathsf{g}^{M}\log(T), (2V_{\mathcal{M}})^{-1}\log(T)\big\},$$

we have

$$\Delta^{M}(\check{\eta}) \leq (1+\varepsilon)\mathsf{g}^{M}\log(T), \quad \text{and} \quad I^{M}(\check{\eta};\mathcal{M}) \geq (1-\varepsilon)\log(T). \tag{H.5.17}$$

Next, we claim that $\widehat{\pi} = \pi_{M}$. To see this, note that since $\mathbb{E}^{M,\mathbb{A}}[\mathbf{Reg}(T)] \leq 2\mathsf{g}^{M}\log(T)$, we

have

$$\eta^M(\pi_M) \geq T - 2\frac{\mathsf{g}^M}{\Delta^M_{\min}}\frac{\log(T)}{T} \geq \frac{3}{4}T$$

as long as $T > 8\frac{\mathsf{g}^M}{\Delta^M_{\min}}\log(T)$, which is implied by the condition (H.5.16). Hence, as long as $\varepsilon_{\text{stat}} \leq \frac{T}{4}$, we have

$$\check{\eta}(\pi_M) > \frac{T}{2},$$

which implies that $\widehat{\pi} = \pi_M$. By definition of $\mathsf{n}^{\mathcal{M}}_{\varepsilon}$, and since $\check{\eta}$ satisfies (H.5.17) above, we then have that setting $\check{\eta}(\widehat{\pi}) = \mathsf{n}^{\mathcal{M}}_{\varepsilon}\log(T)$ does not affect the regret, and only decreases the information gain by a factor of $\varepsilon\log(T)$. It follows that

$$\Delta^M(\widetilde{\eta}) \leq (1+\varepsilon)\mathsf{g}^M\log(T), \quad \text{and} \quad I^M(\widetilde{\eta};\mathcal{M}) \geq (1-2\varepsilon)\log(T).$$

We conclude that $\widehat{\lambda} \in \Upsilon(M; 2\varepsilon, \bar{\mathsf{n}})$ for

$$\bar{\mathsf{n}} = \|\widetilde{\eta}\|_1/\log(T).$$

To wrap up, we compute that

$$\begin{aligned}\|\widetilde{\eta}\|_1 &\leq \sum_{\pi\neq\pi_M}\eta^M(\pi) + \varepsilon_{\text{stat}} + \mathsf{n}^{\mathcal{M}}_{\varepsilon}\log(T) \leq 2\frac{\mathsf{g}^M}{\Delta^M_{\min}}\log(T) + \varepsilon_{\text{stat}} + \mathsf{n}^{\mathcal{M}}_{\varepsilon}\log(T)\\ &\leq 3\frac{\mathsf{g}^M}{\Delta^M_{\min}}\log(T) + \mathsf{n}^{\mathcal{M}}_{\varepsilon}\log(T)\end{aligned}$$

whenever $\varepsilon_{\text{stat}} \leq \mathsf{g}^M\log(T)$.

□

Proof of Lemma H.5.5. From Proposition 1 of [Lugosi and Mendelson, 2019], we have that for any $\delta \leq e^{-1}$, there exists an estimator $\widetilde{\mu}_n$ that, given n independent samples of X, ensures that with probability at least $1-\delta$,

$$\|\widetilde{\mu}_n - \mu\|_2 \leq 12\sqrt{\frac{\mathbb{E}\|X-\mu\|_2^2 \cdot \log(\delta^{-1})}{n}}.$$

Given $\widetilde{\mu}_n$, we define $\widehat{\mu}_n = \arg\min_{u\in\mathbb{R}^d:\|u\|_0\leq s}\|u - \widetilde{\mu}_n\|_2$. Since $\|\mu\|_0 \leq s$, we have $\|\widetilde{\mu}_n - \widehat{\mu}_n\|_2 = \min_{u:\|u\|_0\leq s}\|\widetilde{\mu}_n - u\|_2 \leq \|\widetilde{\mu}_n - \mu\|_2$. It follows that

$$\|\widehat{\mu}_n - \mu\|_2 \leq \|\widetilde{\mu}_n - \widehat{\mu}_n\|_2 + \|\widetilde{\mu}_n - \mu\| \leq 2\|\widetilde{\mu}_n - \mu\|_2.$$

Finally, we note that since $\widehat{\mu}_n$ and μ are both s-sparse, $\|\widehat{\mu}_n - \mu\|_1 \leq \sqrt{2s}\|\widehat{\mu}_n - \mu\|_2$. □