Estimation and Inference of Optimal Policies

Zhaoqi Li

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

Doctor of Philosophy

University of Washington

2024

Reading Committee:
Alex Luedtke, Chair
Lalit Jain, Chair
Kevin Jamieson

Program Authorized to Offer Degree:
Statistics

University of Washington

Abstract

Estimation and Inference of Optimal Policies

Zhaoqi Li

Co-Chairs of the Supervisory Committee:
Alex Luedtke
Department of Statistics

Lalit Jain
Department of Marketing and International Business

Many fields conduct experiments to learn policies that map individual characteristics to actions, with those achieving the best outcomes referred to as optimal policies. As getting human feedback from experiments is expensive, we are often interested in learning the optimal policy as quickly as possible. However, there are several challenges in developing practical approaches for policy learning. First, traditional methods usually only guarantee minimax optimality, while practitioners care more about performances for their particular problem instance. Therefore, a better notion of optimality than the worst case is needed. Second, existing optimal methods are generally hard to implement on a large scale, making deployment challenging for large companies. Third, real-world settings often involve multiple performance metrics of interest, such as mitigating side effects while ensuring good disease recovery in biomedical sciences or balancing short-term acquisition with long-term retention in digital marketing.

This dissertation tackles these challenges and provides several practical approaches for policy learning from various perspectives. To identify the optimal policy as fast as possible, we frame policy learning as pure exploration problems in bandits and develop algorithms that provably identify the optimal policy quickly for every problem instance, a concept we refer to

as *instance optimality*. In Chapter 2, we focus on the stochastic contextual bandit problem in the (ϵ, δ)-PAC setting: given a policy class, the goal is to return a policy whose expected reward is within ϵ of the optimal policy with probability greater than $1 - \delta$. We characterize the first *instance-dependent* PAC sample complexity of contextual bandits. We propose a new computationally efficient algorithm that achieves this sample complexity using only a polynomial number of calls to an argmax oracle.

Chapter 3 delves into the challenge of computational efficiency, focusing on developing algorithms that are easily implementable on a large scale. We focus on the linear bandit setting where we aim to return the arm with the largest reward given a set of arms and an unknown parameter vector. We introduce an algorithm that leverages the same oracles required by the widely-used Thompson sampling algorithm, namely sampling and argmax oracles, and achieves an asymptotically optimal exponential convergence rate. In addition, we demonstrate that our algorithm is easy to implement and performs empirically as well as existing optimal methods.

Chapter 4 explores the impact of the optimal policy on additional metrics when multiple objectives are of interest. We propose a novel margin condition that restricts how the subsidiary metric behaves for nearly optimal policies. Under this condition, we provide an efficient estimator for evaluating subsidiary metrics under a policy that is optimal for the primary one. Additionally, we introduce two alternative two-stage strategies that do not require a margin condition. Both methods first construct a set of candidate policies and then build a confidence interval over this set. We provide numerical simulations to assess the performance of these methods in various scenarios.

CONTENTS

List of Figures v

Glossary vi

Chapter 1: Introduction 1

1.1 Motivation and Challenges 1

1.2 Background 3

1.3 Contributions and Outline 4

Chapter 2: Instance-optimal PAC Algorithms for Contextual Bandits 7

2.1 Introduction 7

2.2 Problem statement and main results 11

2.3 Optimal Algorithms for Contextual Bandits 19

2.4 Conclusion 29

Chapter 3: Optimal Exploration is no harder than Thompson Sampling 30

3.1 Introduction 30

3.2 Motivating our approach 32

3.3 Best Arm Identification through Sampling 35

3.4 Related Work 43

3.5 Experiments 45
3.6 Conclusion 48

Chapter 4: Estimation of subsidiary performance metrics under optimal policies . 49
4.1 Introduction 49
4.2 Wald-type inference under a margin assumption 53
4.3 Inference of a general functional without margin assumption 57
4.4 Numerical experiment 66
4.5 Discussion 69

Chapter 5: Conclusion 71

Bibliography 73

Appendix A: Appendix to Chapter 2 86
A.1 Proof for Results in Section 2.2 86
A.2 Proof for sample complexity of Algorithm 1 and 2 93
A.3 Proof of the FW-GD subroutine 99
A.4 Proof of Theorem 2.3.4 108
A.5 Convergence analysis of FW-GD 119
A.6 Useful lemmas 142

Appendix B: Appendix to Chapter 3 146
B.1 Notations and General Description 146
B.2 Proof of Theorem 3.2.1 148

B.3 Proof of Theorem 3.3.4 151
B.4 Bounds and Events that Hold True Each Round 177
B.5 Technical Lemmas 178
B.6 Supplementary Plots 181

Appendix C: Appendix to Chapter 4 184
C.1 Proofs for Section 4.2 184
C.2 Proofs for Section 4.3 192
C.3 Multiplier bootstrap 203

LIST OF FIGURES

Figure Number Page

3.1 Best-arm identification rate for PEPS, LinGame [Degenne et al., 2020], LinGapE [Xu et al., 2018], Thompson sampling, and fixed weight strategy under three instances: Soare instance with $\omega = 0.1$, sphere instance with $d = 6$ and $|\mathcal{X}| = 20$, and Top-k instance with $d = 12$ and $k = 3$, with 500 repetitions for each instance. Confidence intervals with plus or minus two standard errors are shown. 47

4.1 Plot of primary and subsidiary performance metrics for an estimated policy given the threshold policy class $\Pi = \{\mathbf{1}(x \geq a) : a \in \mathbb{R}\}$. The estimator $\widehat{\pi}$ performs well in the sense that the Ω-regret $\Omega_{\pi^*}(P_0) - \Omega_{\widehat{\pi}}(P_0)$ is small, which is to be expected since π^* is defined to be an Ω-optimal rule. Nevertheless, in principle the Ψ-regret $\Psi_{\pi^*}(P_0) - \Psi_{\widehat{\pi}}(P_0)$ could still be large, since the Ψ-value function $\pi \mapsto \Psi_\pi(P_0)$ may be markedly different from the Ω-value function. Though a similar phenomenon can occur for unrestricted policy classes, which are our focus in this section, the infinite-dimensional nature of these classes precludes their visualization. 54

4.2 Example of first-stage elimination. Each black dot represents an estimate of $\Omega_\pi(P_0)$ and the horizontal bars denote the confidence bounds. Policies whose uniform upper confidence bound (UCB) is below the largest lower confidence bound (LCB) get eliminated. 59

4.3 An illustration of $s_{b,0}(X)$- and $q_{b,0}(X)$-value and Ω- and Ψ-value for a 1-dimensional threshold policy class $\Pi = \{\mathbf{1}_{[a,\infty)} : a \in [-1, 1]\}$ under different scenarios: the optimal policy for the primary outcome is nonunique, the optimal policy for the primary outcome is unique while the primary and subsidiary outcomes are correlated, and the optimal policy for the primary outcome is unique while the primary and subsidiary outcomes are not so correlated. The top figure represents Ω- and Ψ-value, while the bottom figure represents $s_{b,0}(X)$- and $q_{b,0}(X)$-value. 67

B.1 Average number of rejection samples taken until finding some $\theta \in \Theta^c_{\widehat{z}_t}$ 182

B.2 Average clock time per iteration for PEPS under three scenarios 182

GLOSSARY

:

ACKNOWLEDGMENTS

I would like to express my sincere gratitude to my academic advisors, Lalit Jain and Alex Luedtke. Lalit guided me to learn how to do research when I started my PhD research career late. Starting from defining simple problems, I gained confidence by working out proofs and applying simple changes to existing algorithms. You are also willing to spend all your available time with me, from checking every single proof to teaching me how to do a good academic presentation. I pushed myself to the limit and pushed the limit further throughout my PhD career, and I would not be even close to where I am right now without your help.

I am grateful to Alex for your consistent support over the years. You always answer my questions promptly and are always available to meet and help regardless of my ups and downs. Also, your attention to detail helped me significantly improve my writing skills and thinking about problems.

I would also like to thank Houssam Nassif for his support. During the darkest moments of my PhD, you offered me an opportunity to do an internship at Amazon, during which I discovered my current research interest. You also helped me find my current research advisors and again offered me a position during the most stressful time of my final year. I cannot finish my PhD without your support.

Then, I would like to thank Kevin Jamieson, who is not my official advisor but provides a lot of guidance for my research. You always present new ideas and give thoughtful feedback when I am stressed and unsure how to proceed. I have also learned a lot, both research-wise and career-wise, in your group meetings through countless valuable discussions over the years. I also thank Yen-Chi Chen and Byron Boots for serving on the committee for my general and final exams.

I am grateful for the department staff, Ellen Reynolds, Tracy Pham, Kristine Chan, and Vickie Graybeal. I'm especially grateful for Ellen as her always prompt response regardless of what question I have keeps my PhD journey on track.

I would also like to thank all my fellow students who shared a significant amount of experience with me during this journey, including but not limited to: Lang Liu, Hanyu Zhang, Zhen Miao, Andrew Wagenmaker, Si Cheng, Romain Camilleri, Anupreet Porwal, Gang Cheng, Jennifer Brennan, Zhihan Xiong, Yifang Chen, Kenny Zhang, Yidan Xu, and Yuhao Wan.

I would like to thank my friends, including but not limited to: Lang Li, Zun Yin, Selena Huang, Ruotong Wang, Qingyuan Jiang, Tiffany Fan, Shuangning Li, Yi Ren, and Qisheng Li, for supporting me. Although not contacted often, their presence always makes me feel that there is someone I can reach out to during downtimes.

Last but not least, I would like to thank Yilin Song for all the love, care, and joy you shared with me. I cannot go this far without you.

DEDICATION

to my family

Chapter 1

INTRODUCTION

1.1 Motivation and Challenges

A policy is a mapping from a set of individual-level features to actions. In many fields, we are interested in learning the best policy that gives us the highest subsequent outcome. For example, in digital marketing, given user features, companies look for the best strategy for targeting individual customers to maximize the expected revenue; in education, schools try to find the best resource allocation to ensure that all the students get the best possible education; in biomedical sciences, hospitals try to identify the best treatment for each patient to maximize the disease recovery rates. Systematic approaches for learning the optimal policy have large profits for companies and can significantly reduce the workload for school officers and clinicians. Therefore, developing practical approaches for policy learning is important.

To learn a good policy, companies first conduct experiments, in which they interact with the environment to collect data. Existing approaches, especially bandit and reinforcement learning approaches, can learn a good policy with a sufficient amount of data available. However, in practice, good data are usually expensive to collect. In particular, it is very costly to get human labels as humans are generally hard and expensive to recruit. Also, online interactions could be costly. In a large company like Amazon, deploying policies takes time, and human in the loop may be needed to get user feedback and analyze them. Therefore, it is important to consider if we can learn the optimal policy quickly in practical settings. Here "quickly" involves two aspects: one is with minimum interactions with the environment, and another is with fast computation.

We would like to design algorithms that make as few interactions with the environment as possible among any algorithm, in other words, achieving the optimal *sample complexity*.

One widely-used notion of optimality is called minimax optimality, which measures the best achievable performance in the worst-case scenario. However, practitioners usually care more about the actual performance of their particular problem instance, and their setting could be very different from the worst case. Algorithms that are optimal in the worst case can perform poorly in easier instances. To address this, a promising approach is to design algorithms that adjust to the difficulty of the specific problem, aiming for better overall performance. To develop performance guarantees for such algorithms, we use a stronger notion of optimality, called *instance-optimality*. This concept, originating from the work of [Lai and Robbins, 1985], ensures that the algorithm is optimal for *every* problem instance, not just optimal for the worst case.

When developing instance-optimal algorithms, a challenge is computational efficiency. Many existing instance-optimal bandit algorithms are hard to implement as they involve computing complicated optimization problems [Fiez et al., 2019, Degenne et al., 2020, Tao et al., 2018]. They are generally not computationally feasible in large-scale decision-making problems, which makes them hard to deploy for large companies. For example, Amazon Prime has over 100 million users, and millions of items are listed on their website. We would like algorithms that are easy to implement on these large-scale problems. Currently, only simple algorithms like Thompson sampling and some of its variants have been deployed there, and there is a need to develop efficient algorithms that can be easily implemented.

It is not the end goal after we have learned an optimal policy since there are usually multiple objectives of consideration in practical applications [Boominathan et al., 2020, Bica et al., 2021]. For example, in biomedical trials, while hospitals primarily care about disease recovery rates, they are also concerned about the potential side effects of a new drug. Strong medicine usually helps more in recovering the disease, but it might also have a potentially larger side effect. Also, in online advertising, companies primarily care about short-term acquisition as they seek to get more users to sign up [Chang et al., 2020]. However, they also care about long-term retention since they would love users to stay. These objectives might potentially lead to different strategies. For example, giving out free trials might be

beneficial for short-term subscriptions but users may quit after the trial; meanwhile, giving out discounts may be better for users to stay but fewer users may subscribe since it costs. Therefore, it is important to get an understanding of the optimal policy on other objectives, so practitioners have a clear picture about the overall impact of deploying this new policy.

This dissertation addresses the above challenges from different perspectives. First, for learning the optimal policy for a primary objective, I provide exploration strategies that are *instance-optimal* and *computationally fast* in large-scale experiments. Second, after we have learned the optimal policy for a performance metric, I provide estimators and methods to evaluate the effect of this policy on other subsidiary metrics. Section 1.2 introduces background together with some challenges in solving these problems, and Section 1.3 provides a detailed contribution of the dissertation with outlines for the remaining chapters.

1.2 Background

In traditional statistical inference, a common assumption is that the data are independent and identically distributed. However, this does not always hold in practical experiments. For example, companies would like their policies to evolve based on feedback from past experience; in clinical trials, the experiment might adjust dosages for new participants based on how current participants are responding. In these situations, the data are collected *adaptively* from past observations, and traditional methods based on the independence assumption, for example, the M-estimator and Z-estimator, will no longer work. Consequently, there is a growing need for experimental designs for making sequential decisions that will learn from experience and adapt to the environment.

A pivotal area for discussing sequential experimental design is the exploration of stochastic bandit problems, where the balance between exploration and exploitation plays a critical role in optimizing outcomes over time. Generally speaking, a bandit problem $\mathcal{M}$ is a sequential procedure between a *learner* and an *environment*. In each round, the learner chooses an action a from a given action set $\mathcal{A}$, and the environment reveals a (noisy) reward $r \in \mathbb{R}$. When choosing actions, the learner has access to the history of actions and rewards, so they

may choose actions adaptively based on the past. A significant part of the dissertation focuses on contextual bandits, where in each round, a context $c \in \mathcal{C}$ is shown to the learner, and a *policy* $\pi \in \Pi$ is a mapping from the context to actions. A learner adopts a policy to interact with an environment. An algorithm is *optimal* if it learns the best policy with the fewest possible rounds.

The exploration of stochastic bandits has traditionally been dominated by regret minimization algorithms, which aim to minimize the cumulative difference between the chosen actions and the best possible action [Abbasi-Yadkori et al., 2011, Russo et al., 2018]. However, for policy learning, these algorithms are only minimax optimal, which means that they are only optimal in the worst case. An ideal algorithm should perform better in easy cases. To capture the difference in such instances, we focus on a better notion of optimality, called *instance optimality*. We define a problem instance $\mathcal{M}$ to be the model class, which consists of the action set, policy set, reward function, and context distribution if applicable. An algorithm is *instance-optimal* if it performs the best among any algorithm for every problem instance. We aim to find instance optimal algorithms as they exploit the structure of each problem instance.

1.3 Contributions and Outline

This dissertation tackles the challenges described in the above sections. Below we provide an outline for each remaining chapter.

Instance-optimal and fast algorithm for pure exploration in linear bandits. Given a set of arms $\mathcal{Z} \subset \mathbb{R}^d$ and an unknown parameter vector $\theta_* \in \mathbb{R}^d$, the pure exploration linear bandit problem aims to return $\arg\max_{z \in \mathcal{Z}} z^\top \theta_*$, with high probability through noisy measurements of $x^\top \theta_*$ with $x \in \mathcal{X} \subset \mathbb{R}^d$. Existing (asymptotically) optimal methods require either a) potentially costly projections for each arm $z \in \mathcal{Z}$ or b) explicitly maintaining a subset of $\mathcal{Z}$ under consideration at each time. This complexity is at odds with the popular and simple Thompson Sampling algorithm for regret minimization, which just requires access

to a posterior sampling and argmax oracle, and does not need to enumerate $\mathcal{Z}$ at any point. Unfortunately, Thompson sampling is known to be sub-optimal for pure exploration. In this work, we pose a natural question: is there an algorithm that can explore optimally and only needs the same computational primitives as Thompson Sampling? We answer the question in the affirmative. We provide an algorithm that leverages only sampling and argmax oracles and achieves an exponential convergence rate, with the exponent equal to the exponent of the optimal fixed allocation asymptotically. In addition, we show that our algorithm can be easily implemented and performs as well empirically as existing asymptotically optimal methods.

Instance-optimal and computationally efficient algorithms for pure exploration in contextual bandits. In this work, we focus on the stochastic contextual bandit problem in the (ϵ, δ)-PAC setting: given a policy class Π the goal of the learner is to return a policy $\pi \in \Pi$ whose expected reward is within ϵ of the optimal policy with probability greater than $1 - \delta$. We characterize the first *instance-dependent* PAC sample complexity of contextual bandits through a quantity ρ_Π, and provide matching upper and lower bounds in terms of ρ_Π for the agnostic and linear contextual best-arm identification settings. We show that no algorithm can be simultaneously minimax-optimal for regret minimization and instance-dependent PAC for best-arm identification. Our main result is a new instance-optimal and computationally efficient algorithm that relies on a polynomial number of calls to an argmax oracle.

Estimation of subsidiary metrics under optimal policies. This paper presents two strategies for evaluating subsidiary metrics under a policy that is optimal for the primary one. The first relies on a novel margin condition that facilitates Wald-type inference. Under this and other regularity conditions, we show that the one-step corrected estimator is efficient. Despite the utility of this margin condition, it places strong restrictions on how the subsidiary metric behaves for nearly optimal policies, which may not hold in practice. We therefore introduce alternative, two-stage strategies that do not require a margin condition. The first stage constructs a set of candidate policies and the second builds a uniform confidence interval

over this set. We provide numerical simulations to evaluate the performance of these methods in different scenarios.

Other explorations. This dissertation does not include the work on dynamic pricing [Jain et al., 2023].

Chapter 2

INSTANCE-OPTIMAL PAC ALGORITHMS FOR CONTEXTUAL BANDITS

2.1 Introduction

We consider the stochastic contextual bandit problem in the PAC setting. Fix a distribution ν over a potentially countable[1] set of contexts $\mathcal{C}$. The action space is $\mathcal{A}$, and for computational tractability, we assume $|\mathcal{A}|$ is finite. We have a set of policies Π of interest where each policy $\pi \in \Pi$ is a map from contexts to an action space $\pi : \mathcal{C} \to \mathcal{A}$. The reward function is $r : \mathcal{C} \times \mathcal{A} \to \mathbb{R}$. At each time $t = 1, 2, \ldots$ a context $c_t \sim \nu$ arrives, the learner chooses an action $a_t \in \mathcal{A}$, and receives reward $r_t := r_t(c_t, a_t) \in \mathbb{R}$ with $\mathbb{E}[r_t|c_t, a_t] = r(c_t, a_t) \in \mathbb{R}$. The value of a policy $V(\pi)$ is the expected reward from playing action $\pi(c)$ in context c: $V(\pi) = \mathbb{E}_{c\sim\nu}[r(c, \pi(c))]$. Given a collection of policies Π, the objective is to identify the optimal policy $\pi_* := \arg\max_{\pi\in\Pi} V(\pi)$, with high probability. Formally, for any $\epsilon > 0$ and $\delta \in (0, 1)$, we seek to characterize the sample complexity of identifying a policy $\pi \in \Pi$ such that $V(\pi) \geq V(\pi_*) - \epsilon$, with probability at least $1 - \delta$. That is, we wish to minimize the total amount of interactions with the environment to learn an ϵ-optimal policy.

We study both the *agnostic* setting, where Π is an arbitrary set of policies with no assumed relationship with the reward function $r(c, a)$; and the *realizable* setting, where the policy class and the reward function follow a linear structure, known as the linear contextual bandit problem. In both cases, we are interested in *instance-dependent* sample complexity bounds. That is, the upper and lower bounds we seek do not simply depend on coarse quantities like $|\Pi|$, $|\mathcal{A}|$, and $1/\epsilon^2$, but more fine-grained relationships between the context distribution ν,

[1]Assuming the set of contexts is countable versus uncountable is for presentation purposes only, since it allow us the notational convenience of letting ν_c denote the probability of context c arriving.

geometry of policies Π, and the reward function $r : \mathcal{C} \times \mathcal{A} \to \mathbb{R}$. Our motivation is that instance-dependent bounds describe the difficulty of a particular problem instance, allowing optimal algorithms to adapt to the true difficulty of the problem, whether easy or hard. We seek algorithms that take advantage of "easy" instances instead of optimizing for the worst-case [Jun et al., 2021a].

2.1.1 Related work

Minimax regret bounds for general policy classes The vast majority of research in contextual bandits focuses on regret minimization. That is, for a time horizon T, the goal of the player is to minimize $\mathbb{E}\left[\sum_{t=1}^{T} r(c_t, \pi_*(c_t)) - r(c_t, a_t)\right]$. The landmark algorithm EXP4 for non-stochastic multi-armed bandits [Auer et al., 2002] achieves a regret bound of $\sqrt{|\mathcal{A}|T\log(|\Pi|)}$. Unfortunately, the running time of EXP4 is linear in $|\Pi|$ which is prohibitive for many problems of interest. The algorithms proposed in Dudik et al. [2011] and Agarwal et al. [2014] achieve the same regret bound with a computational complexity that is only polynomial in T and $\log(|\Pi|)$. Both approaches can be used to obtain an ϵ-optimal policy with probability at least $1-\delta$ using a sample complexity no more than $\frac{|\mathcal{A}|\log(|\Pi|/\delta)}{\epsilon^2}$. None of these works made any assumption on the connection between the reward function r and the policy class Π (i.e. the agnostic setting).

Instance-dependent regret bounds for general policy classes The epoch-greedy algorithm of Langford and Zhang [2007] achieved the first instance-dependent bounds on regret with a coarse guarantee depending only on the minimum policy gap $\Delta_{\mathsf{pol}} := V(\pi_*) - \max_{\pi \neq \pi_*} V(\pi)$. In the pursuit of more fine-grained regret bounds achievable by computationally efficient algorithms, many authors resort to the *realizability* assumption [Foster et al., 2018, Foster and Rakhlin, 2020, Simchi-Levi and Xu, 2021, Foster et al., 2021a]. The learner knows a hypothesis class $\mathcal{H}$ where each $f \in \mathcal{H}$ is a map $f : \mathcal{C} \times \mathcal{A} \to \mathbb{R}$, and there exists an $f^* \in \mathcal{H}$ such that $r(c,a) = f^*(c,a)$ for all $(c,a) \in \mathcal{C} \times \mathcal{A}$. Under this assumption, [Foster et al., 2021a] proves lower and upper bounds on the instance-dependent regret. Their bounds

are in term of the *uniform gap* $\Delta_{\mathsf{uniform}} := \min_{c \in \mathcal{C}} \min_{a \in \mathcal{A}} r(c, \pi_*(c)) - r(c, a)$. In general, for any policy class, they establish matching minimax lower and upper regret bounds of the form $\min\{\sqrt{|\mathcal{A}| T \log(|\mathcal{H}|)}, \frac{|\mathcal{A}| \log(|\mathcal{H}|)}{\Delta_{\mathsf{uniform}}} \mathfrak{C}_{\mathcal{H}}^{\mathsf{pol}}\}$, where $\mathfrak{C}_{\mathcal{H}}^{\mathsf{pol}}$ is the *policy disagreement coefficient*, a parameter depending on the geometry of $\mathcal{H}$ and the context distribution ν. That is, these bounds hold with respect to a worst-case family of instances parameterized by $\Delta_{\mathsf{uniform}}$ and $\mathfrak{C}_{\mathcal{H}}^{\mathsf{pol}}$. Using the standard online-to-batch conversion, this translates to a sample complexity (i.e. the time required to find an ϵ-good policy with constant probability) of roughly $\frac{|\mathcal{A}| \log(|\mathcal{H}|)}{\epsilon \ \Delta_{\mathsf{uniform}}} \mathfrak{C}_{\mathcal{H}}^{\mathsf{pol}}$. We show in Corollary 2.2.16 that this sample complexity is at least as large as our bounds. Further, unlike our bounds below, this sample complexity is unbounded as ϵ goes to 0. Recent work refines these kinds of regret bounds further, and provides minimax regret bounds in terms of the *decision-estimation coefficient* [Foster et al., 2021b].

Regret bounds for linear contextual bandits A special case of the realizable case assumes a linear structure for $\mathcal{H}$. Assume there exists a known feature map $\phi : \mathcal{C} \times \mathcal{A} \to \mathbb{R}^d$ and an unknown $\theta_* \in \mathbb{R}^d$ such that the true reward function is given as $r(c, a) = \langle \phi(c, a), \theta_* \rangle$. For this setting, popular optimism-based algorithms like LinUCB [Li et al., 2010] and Thompson sampling [Russo, 2016, Nabi et al., 2022] achieve a regret bound of $\min\{d\sqrt{T}, \frac{d^2}{\Delta_{\mathsf{uniform}}}\}$ [Abbasi-Yadkori et al., 2011]. Appealing to the online-to-batch conversion, this translates to a PAC guarantee of $\frac{d^2}{\epsilon \ \Delta_{\mathsf{uniform}}}$. More precise instance-dependent upper bounds on regret match instance-dependent lower bounds asymptotically as $T \to \infty$ [Hao et al., 2020, Tirinzoni et al., 2020]. These works are most similar to our setting and have qualitatively similar style algorithms. However, both approaches rely on asymptotics with large problem-dependent terms that may dominate the bounds in finite time. Our work is focused on upper bounds that nearly match lower bounds for all finite times.

PAC sample complexity for contextual bandits As we will describe, all contextual bandits with an arbitrary policy class can be reduced to PAC learning for linear bandits. Once we made this reduction, our sample complexity analysis draws inspiration from the nearly instance-optimal algorithm for linear best-arm identification [Fiez et al., 2019]. PAC

sample complexity of linear contextual bandits was also studied in [Zanette et al., 2021], who shows a minimax guarantee sample complexity that scales with $\frac{d^2}{\epsilon^2}\log(1/\delta)$. In their approach, [Agarwal et al., 2014] define their action sampling distribution as a convex combination over policies. Our sampling distribution, as well as the optimal sampling distribution, cannot be represented this way and is actually derived from the dual of the optimal experimental design objective.

2.1.2 Contributions

In this work, our contributions include:

1. In the agnostic setting, we introduce a quantity ρ_Π that characterizes the instance-dependent sample complexity of PAC learning for contextual bandits (see Equation 2.1). We show that ρ_Π appears in an information theoretic lower bound on the sample complexity of any PAC algorithm as $\epsilon \to 0$ in Theorem 2.2.2. To ground this, we describe it carefully in the setting of the trivial policy class (Section 2.2.2) and linear policy classes (Section 2.2.3). To do so, we reduce agnostic contextual bandits to the realizable linear case (also establishing matching upper and lower bounds in this setting).

2. We construct an instance on which any regret minimax-optimal algorithm necessarily has a sample complexity that scales quadratically with the optimal sample complexity (Theorem 2.2.6). This shows that no algorithm can be both regret minimax-optimal and instance-optimal PAC.

3. Finally, we propose Algorithm 4 whose sample complexity nearly matches the lower bound based on ρ_Π. By appealing to an argmax oracle, this algorithm has a runtime polynomial in ρ_Π, $1/\epsilon$, $\log(1/\delta)$, $|\mathcal{A}|$, and $\log(|\Pi|)$, assuming a unit cost of invoking the oracle.

2.2 Problem statement and main results

More formally, define $\mathcal{F}_t = \sigma(c_1, a_1, r_1, \ldots, c_t, a_t, r_t)$ as the natural σ-algebra filtration capturing all observed random variables up to time t. At each time t an *algorithm* defines a *sampling rule* $\mathcal{F}_t \mapsto \mathcal{A}$ which defines a_{t+1}, an $\mathcal{F}_t$-measurable stopping time $\tau \in \mathbb{N}$, and a *selection rule* $\mathcal{F}_t \mapsto \Pi$ that is only called once at the stopping time $t = \tau$.

Definition 2.2.1. Fix $\epsilon \geq 0$ and $\delta \in (0, 1)$. We say an algorithm is (ϵ, δ)-PAC for contextual bandits with policy class Π, if at the stopping time $\tau \in \mathbb{N}$ with $\mathbb{E}[\tau] < \infty$, the algorithm outputs $\widehat{\pi} \in \Pi$ satisfying $\mathbb{P}(V(\widehat{\pi}) \geq \max_{\pi \in \Pi} V(\pi) - \epsilon) \geq 1 - \delta$.

The *sample complexity* of an (ϵ, δ)-PAC algorithm for contextual bandits is the time at which the algorithm stops and outputs $\widehat{\pi}$. The following quantity governs the sample complexity :

$$\rho_{\Pi,\epsilon}(\Pi, v) := \min_{p_c \in \triangle_{\mathcal{A}},\ \forall c \in \mathcal{C}} \max_{\pi \in \Pi \setminus \pi_*} \frac{\mathbb{E}_{c \sim \nu}\left[\left(\frac{1}{p_{c,\pi(c)}} + \frac{1}{p_{c,\pi_*(c)}}\right) \mathbf{1}\{\pi_*(c) \neq \pi(c)\}\right]}{(\mathbb{E}_{c \sim \nu}[\, r(c, \pi_*(c)) - r(c, \pi(c))\,] \vee \epsilon)^2}. \tag{2.1}$$

Here, for any countable set $\mathcal{X}$ we have that $\triangle_{\mathcal{X}} = \{p \in \mathbb{R}^{|\mathcal{X}|} : \sum_{x \in \mathcal{X}} p_x = 1, p_x \geq 0\ \forall x \in \mathcal{X}\}$ so that p_c for every $c \in \mathcal{C}$ defines a probability distribution over actions $\mathcal{A}$. In addition we use the notation $a \vee b := \max\{a, b\}$. We begin with a necessary condition on the sample complexity for the particular case of exact policy identification ($\epsilon = 0$).

Theorem 2.2.2 (Lower bound)**.** *Fix $\epsilon = 0$ and $\delta \in (0, 1)$. Moreover, fix a contextual bandit instance $\mu = (\nu, r)$ and a collection of policies Π. Then any $(0, \delta)$-PAC algorithm for contextual bandits satisfies $\mathbb{E}_\mu[\tau] \geq \rho_{\Pi,0} \log(1/2.4\delta)$.*

The proof of the lower bound follows from standard information theoretic arguments [Kaufmann et al., 2016]. The lower bound implicitly applies to learners that know the distribution ν precisely. In practice, such knowledge would never be available however the learner may have a large dataset of offline data.

Assumption 1. Prior to starting the game, the learning algorithm is given a large dataset of contexts $\mathcal{D} = \{c_t\}_{t=1}^T$, where each $c_t \overset{i.i.d.}{\sim} \nu$ for all $t \in [T]$, and $T = O(\text{poly}(1/\epsilon, |\mathcal{A}|, \log(1/\delta), \log(|\Pi|)))$.

The above only assumes access to samples from the context distribution, not rewards or the value function. Importantly, since $\mathcal{C}$ could be uncountable, we do not assume $\mathcal{D}$ covers the support of ν. Assumption 1 is satisfied, for example, in an e-commerce setting where the context is the demographic information about visitors to the site for which massive troves of historical data may be available. Other works in PAC learning have made similar assumptions [Huang et al., 2015]. We would like our algorithm to be computationally efficient in the sense that it makes a polynomial number of calls to what we refer to as argmax oracle. Such an assumption is common in the contextual bandits literature [Agarwal et al., 2014, Krishnamurthy et al., 2017, Dudik et al., 2011].

Definition 2.2.3 (Argmax oracle (AMO))**.** The oracle $\mathsf{AMO}(\Pi, \{(c_t, s_t)\}_{t=1}^n)$ is an algorithm that given contexts and cost vectors $(c_1, s_1), \cdots, (c_n, s_n) \in \mathcal{C} \times \mathbb{R}^{|\mathcal{A}|}$, returns $\arg\max_{\pi \in \Pi} \sum_{t=1}^n s_t(\pi(c_t))$. The constrained argmax oracle C-AMO, given an upper bound l on the loss, returns $\arg\max_{\pi \in \Pi} \sum_{t=1}^n s_t(\pi(c_t))$ subject to $\sum_{t=1}^n s_t(\pi(c_t)) \leq l$.

In general we can implement AMO by calling to cost-sensitive classification [Dudik et al., 2011, Beygelzimer et al., 2005] and C-AMO through a Lagrangian relaxation and a cost-sensitive classification oracle [Agarwal et al., 2018, Cotter et al., 2019]. Our algorithm uses an argmax oracle as a subroutine at most a polynomial number of times in $\epsilon^{-1}, \log(1/\delta), |\mathcal{A}|$ and $\log(|\Pi|)$. In this sense, it is computationally efficient. The following sufficiency result holds for general $\epsilon \geq 0$.

Theorem 2.2.4 (Upper bound)**.** *Fix $\epsilon \geq 0$ and $\delta \in (0, 1)$. Under Assumption 1, there exists a computationally efficient (ϵ, δ)-PAC algorithm for contextual bandits that satisfies $\tau \leq \rho_{\Pi,\epsilon} \log(|\Pi| \log_2(1/\epsilon)/\delta) \log(1/\Delta_\epsilon)$, where $\Delta_\epsilon = \max\{\epsilon, \min_{\pi \in \Pi \setminus \pi_*} V(\pi_*) - V(\pi)\}$. Furthermore, this sample complexity never exceeds $\frac{|\mathcal{A}|(\log(|\Pi|)+\log(1/\delta))\log(1/\epsilon)}{\epsilon^2}$.*

The second part of the theorem follows from the first, since $\rho_{\Pi,\epsilon} \leq 2|\mathcal{A}|/\epsilon^2$ by taking $p_{c,a} = 1/|\mathcal{A}|$ for all $(c, a) \in \mathcal{C} \times \mathcal{A}$.

2.2.1 Inefficiency of low-regret algorithms

Computationally efficient algorithms are known to exist, such as ILOVETOCONBANDITS [Agarwal et al., 2014], which achieve a minimax-optimal cumulative regret of $\sqrt{T|\mathcal{A}|\log(|\Pi|/\delta)}$. Inspecting the proof in [Agarwal et al., 2014], one can extract a sample complexity of $\epsilon^{-2}|\mathcal{A}|\log(|\Pi|/\delta)$ from such results (which is also minimax optimal for PAC). The previous section showed that the sample complexity of our algorithm, Theorem 2.2.4, nearly matches the instance-dependent lower bound of Theorem 2.2.2. In other words, our algorithm achieves a nearly optimal instance-dependent PAC sample complexity. However, it is natural to wonder if perhaps with a tighter analysis, the minimax regret optimal algorithm in [Agarwal et al., 2014] also obtains the instance-optimal PAC sample complexity. In this section, we show that this is not the case. Indeed, we show that *any* algorithm that is minimax regret optimal must have a sample complexity that is at least quadratic in the optimal PAC sample complexity of some instance.

Definition 2.2.5 (Hard instance)**.** Fix $m \in \mathbb{N}$, $\Delta \in (0, 1]$ and let $\mathcal{C} = [m]$, $\mathcal{A} = \{0, 1\}$. For $i = 1, \ldots, m$, let $\pi_i(j) = \mathbf{1}\{i = j\}$ and define $r(i, j) = \Delta\mathbf{1}\{j = \pi_1(i)\}$. Then $V(\pi_1) = \Delta$ and $V(\pi_i) = \Delta(1 - 2/m)$ for all $i \in \mathcal{C} \setminus \{1\}$.

Note that for the hard instance, $m = |\Pi|$. If observations are corrupted by $\mathcal{N}(0, 1)$ additive noise, then a straightforward calculation shows that $\rho_{\Pi,0}(\Pi, v) = \frac{4/m}{(2\Delta/m)^2} = m\Delta^{-2}$ for the hard instance.

Theorem 2.2.6. *Fix $\delta \in (0, 1)$ and $\Delta \in (0, 1]$. We say an algorithm is an α-minimax regret algorithm if for some $\alpha > 0$ and all $T \in \mathbb{N}$:*

$$\max_{\mu'} \mathbb{E}_{\mu'}\Big[\sum_{t=1}^{T}(r_t(c_t, \pi_*(c_t)) - r_t(c_t, a_t))\Big] = \max_{\mu}\sum_{c,a}\mathbb{E}_{\mu'}[T_{c,a}(T)](r(c, \pi_*(c)) - r(c, a)) \leq \sqrt{\alpha|\mathcal{A}|T}$$

where the maximum is taken over all contextual bandit instances $\mu' = (\nu', r')$ and $T_{c,a}(T) = \sum_{t=1}^{T}\mathbf{1}\{c_t = c, a_t = a\}$. For any α-minimax regret algorithm, it is a $(0, \delta)$-PAC algorithm if at a stopping time τ it outputs the optimal policy π_* with probability at least $1 - \delta$. Any

α-minimax regret algorithm that is also $(0,\delta)$-PAC satisfies $\mathbb{E}_\mu[\tau] \geq m^2\Delta^{-2}\log^2(1/2.4\delta)/4\alpha$ for the instance $\mu = (\nu, r)$ defined in 2.2.5.

We point out that the minimax regret optimal rate takes $\alpha = \log(m) = \log(|\Pi|)$. Thus, taking $\Delta = 1$ and $\delta = 0.1$, the minimax regret optimal algorithm has a PAC sample complexity of $m^2/\log(m)$; whereas the PAC sample complexity of our algorithm, Theorem 2.2.4, is just $m\log(m)$. That is, algorithms with optimal minimax regret have a sample complexity that is at least nearly the optimal PAC sample complexity *squared*. This demonstrates that no algorithm can simultaneously be minimax regret optimal and obtain the optimal PAC sample complexity.

2.2.2 Trivial policy class

As a warm-up to discussing linear policy classes, let us consider the simplest policy class.

Definition 2.2.7 (Trivial policy class)**.** Assume $|\mathcal{C}| < \infty$ and let $\Pi = \{\pi(c) = a : (c,a) \in \mathcal{C} \times \mathcal{A}\}$ so that $|\Pi| = |\mathcal{A}|^{|\mathcal{C}|}$.

The trivial policy class has the flexibility to predict any action $a \in \mathcal{A}$ individually for each $c \in \mathcal{C}$. This allows us to show that $\rho_{\Pi,0}(\Pi, v) \leq \max_c \frac{2}{\nu_c}\sum_{a'}\Delta_{c,a'}^{-2}$ (see Appendix A.1.3). An immediate corollary of Theorem 2.2.4 is obtained by simply noting that $|\Pi| = |\mathcal{A}|^{|\mathcal{C}|}$.

Corollary 2.2.8 (Trivial class, upper)**.** *Fix $\epsilon > 0$ and $\delta \in (0,1)$. Let Π be the trivial policy class applied to some fixed $\mathcal{C}, \mathcal{A}$ spaces. Then under Assumption 1 there exists a computationally efficient (ϵ,δ)-PAC algorithm for contextual bandits that satisfies $\tau \leq \min\{A\epsilon^{-2}, \max_c \frac{1}{\nu_c}\sum_{a'}\Delta_{c,a'}^{-2}\}(|\mathcal{C}|\log(|\mathcal{A}|)+\log(1/\delta))\log(1/\Delta_\epsilon)$, where $\Delta_\epsilon = \max\{\epsilon, \min_{\pi\in\Pi\setminus\pi_*} V(\pi_*) - V(\pi)\}$. Furthermore, this sample complexity never exceeds $\frac{|\mathcal{A}|(|\mathcal{C}|\log(|\mathcal{A}|)+\log(1/\delta))}{\epsilon^2}\log(1/\epsilon)$.*

Ignoring log factors, the minimax sample complexity of the trivial class is just $\epsilon^{-2}|\mathcal{A}|(|\mathcal{C}| + \log(1/\delta))$. This is actually a somewhat surprising result, because it says $\lim_{\delta\to 0}\frac{\mathbb{E}[\tau]}{\log(1/\delta)} \to \epsilon^{-2}|\mathcal{A}|$ which is *independent* of $|\mathcal{C}|$. To see why this result is somewhat remarkable, if we played a best-arm identification algorithm for each of the $|\mathcal{C}|$ contexts, then this would lead to

a sample complexity of $\epsilon^{-2}|\mathcal{C}| \cdot |\mathcal{A}| \log(1/\delta)$. It is somewhat of a surprise that such a natural strategy is not optimal. For intuition for why we can avoid the multiplicative $|\mathcal{C}|$, note that to identify an ϵ-good policy among just two policies (π, π_*) using uniform exploration requires just $\epsilon^{-2}|\mathcal{A}| \log(1/\delta)$ samples. When we have more than two policies, a union bound achieves the claimed result.

The minimax sample complexity of Corollary 2.2.8 (i.e., the second statement) is nearly tight:

Theorem 2.2.9 (Trivial class, lower)**.** *Fix $\epsilon > 0$ and $\delta \in (0, 1/6)$. Let Π be the trivial policy class applied to some fixed $\mathcal{C}, \mathcal{A}$ spaces. Moreover, fix a contextual bandit instance $\mu = (\nu, r)$ and a collection of policies Π. Then any $(0, \delta)$-PAC algorithm for contextual bandits satisfies $\mathbb{E}_\mu[\tau] \geq \max_c \frac{1}{\nu_c} \sum_a \Delta_{c,a}^{-2} \log(1/2.4\delta)$. Furthermore, $\sup_\mu \mathbb{E}_\mu[\tau] \geq \epsilon^{-2}|\mathcal{A}|(|\mathcal{C}| + \log(1/\delta))$.*

2.2.3 Linear policy class

A particularly compelling model-class of policies is the set of linear policies.

Definition 2.2.10 (Linear policy class)**.** Fix a feature map $\phi : \mathcal{C} \times \mathcal{A} \to \mathbb{R}^d$ and assume it is known to the learner. Let $\Pi = \{\pi(c) = \arg\max_{a \in \mathcal{A}} \langle \phi(c, a), \theta \rangle, \forall \theta \in \mathbb{R}^d\}$.

We can consider two settings: the agnostic setting and the realizable setting. In the agnostic setting, there is no assumed relationship between the true reward function $r(c, a)$ and $\phi : \mathcal{C} \times \mathcal{A} \to \mathbb{R}^d$. In this case, Theorem 2.2.4 applies directly by taking a cover of Π.

Corollary 2.2.11 (Agnostic, upper bound)**.** *Fix $\epsilon \geq 0$ and $\delta \in (0, 1)$. Let Π be the linear policy class in $\mathbb{R}^d$. Under Assumption 1 there exists a computationally efficient (ϵ, δ)-PAC algorithm for contextual bandits that satisfies $\tau \leq \rho_{\Pi,\epsilon} \cdot (d \log(1/\epsilon) + \log(1/\delta)) \log(1/\Delta_\epsilon)$ where $\Delta_\epsilon = \max\{\epsilon, \min_{\pi \in \Pi \setminus \pi_*} V(\pi_*) - V(\pi)\}$. Furthermore, this sample complexity never exceeds $\frac{|\mathcal{A}|(d \log(1/\epsilon) + \log(1/\delta))}{\epsilon^2} \log(1/\epsilon)$.*

Comparing to the lower bound of Theorem 2.2.2, the instance dependent upper bound of Corollary 2.2.11 matches up to a factor of the dimension and negligible log factors. In contrast

to the "model-free" feel of the agnostic case, we can also consider a "model-based" type setting that we refer to as the realizable setting.

Definition 2.2.12 (Realizable)**.** We say the linear policy class is *realizable* if there exists a $\theta_* \in \mathbb{R}^d$ such that $r(c,a) = \langle \phi(c,a), \theta_* \rangle$ for all $c \in \mathcal{C}$ and $a \in \mathcal{A}$. Thus, for any $\pi \in \Pi$ we have $V(\pi) = \mathbb{E}_{c\sim\nu}[r(c,\pi(c))] = \mathbb{E}_{c\sim\nu}[\langle \phi(c,\pi(c)), \theta_* \rangle] = \langle \phi_\pi, \theta_* \rangle$ with $\phi_\pi := \mathbb{E}_{c\sim\nu}[\phi(c,\pi(c))]$. Finally, at the start of the game the learner knows this model.

The setting in Definition 2.2.12 is commonly referred to as the linear contextual bandit problem [Abbasi-Yadkori et al., 2011]. Clearly, we have that $\pi_*(c) = \arg\max_{a\in\mathcal{A}} \langle \phi(c,a), \theta_* \rangle$. We begin by defining a quantity fundamental to our sample complexity results:

$$\rho_{\mathsf{lin},\epsilon} := \min_{p_c \in \triangle_{\mathcal{A}}, \forall c \in \mathcal{C}} \max_{\pi \in \Pi \setminus \pi_*} \frac{\|\phi_\pi - \phi_{\pi_*}\|^2_{\mathbb{E}_{c\sim\nu}[\sum_{a\in\mathcal{A}} p_{c,a}\phi(c,a)\phi(c,a)^\top]^{-1}}}{\langle \phi_{\pi_*} - \phi_\pi, \theta_* \rangle^2 \vee \epsilon^2}.$$

Theorem 2.2.13 (Realizable, lower bound)**.** *Fix $\epsilon = 0$ and $\delta \in (0,1)$. Let Π be the linear policy class in $\mathbb{R}^d$ and assume it is realizable (see Definitions 2.2.10 and 2.2.12). Any $(0,\delta)$-PAC algorithm in this setting satisfies $\mathbb{E}[\tau] \geq \rho_{\mathsf{lin},0} \cdot \log(1/2.4\delta)$.*

We now state our nearly matching upper bound. However, in this case we note that the algorithm is not computationally efficient.

Theorem 2.2.14 (Realizable, upper bound)**.** *Fix $\epsilon \geq 0$ and $\delta \in (0,1)$. Let Π be the linear policy class in $\mathbb{R}^d$ and assume it is realizable (see Definitions 2.2.10 and 2.2.12). Under Assumption 1 there exists an (ϵ,δ)-PAC algorithm for this setting that with probability at least $1-\delta$ it satisfies*

$$\tau \leq \rho_{\mathsf{lin},\epsilon} \cdot (\min\{d\log(1/\epsilon), \log(|\Pi|)\} + \log(1/\delta))\log(1/\Delta_\epsilon)$$

where $\Delta_\epsilon = \max\{\epsilon, \min_{\pi\in\Pi\setminus\pi_*} \langle \phi_{\pi_*} - \phi_\pi, \theta_* \rangle\} = \max\{\epsilon, \min_{(c,a)\in\mathcal{C}\times\mathcal{A}:\pi_*(c)\neq a} \langle \phi(c,\pi_*(c)) - \phi(c,a), \theta_* \rangle\}$. Furthermore, this sample complexity never exceeds $\frac{d(d\log(1/\epsilon)+\log(1/\delta))\log(1/\epsilon)}{\epsilon^2}$.

Proof. To see the second part of the theorem statement, observe that

$$
\begin{aligned}
&\max_{\pi\in\Pi\setminus\pi_*} \|\phi_\pi - \phi_{\pi_*}\|^2_{\mathbb{E}_{c\sim\nu}[\sum_{a\in\mathcal{A}} p_{c,a}\phi(c,a)\phi(c,a)^\top]^{-1}} \\
&\quad= \max_{\pi\in\Pi\setminus\pi_*} \|\mathbb{E}_{c\sim\nu}[\phi(c,\pi(c)) - \phi(c,\pi_*(c))]\|^2_{\mathbb{E}_{c\sim\nu}[\sum_{a\in\mathcal{A}} p_{c,a}\phi(c,a)\phi(c,a)^\top]^{-1}} \\
&\quad\le \max_{\pi\in\Pi\setminus\pi_*} \mathbb{E}_{c\sim\nu}\left[\|\phi(c,\pi(c)) - \phi(c,\pi_*(c))\|^2_{\mathbb{E}_{c\sim\nu}[\sum_{a\in\mathcal{A}} p_{c,a}\phi(c,a)\phi(c,a)^\top]^{-1}}\right] \\
&\quad\le \max_{\pi\in\Pi} 4\,\mathbb{E}_{c\sim\nu}\left[\|\phi(c,\pi(c))\|^2_{\mathbb{E}_{c\sim\nu}[\sum_{a\in\mathcal{A}} p_{c,a}\phi(c,a)\phi(c,a)^\top]^{-1}}\right] \\
&\quad= \max_{q\in\triangle_\Pi} 4\,\mathbb{E}_{c\sim\nu}\left[\sum_{\pi\in\Pi} q_\pi \|\phi(c,\pi(c))\|^2_{\mathbb{E}_{c\sim\nu}[\sum_{a\in\mathcal{A}} p_{c,a}\phi(c,a)\phi(c,a)^\top]^{-1}}\right] \\
&\quad= \max_{q\in\triangle_\Pi} 4\,\mathrm{Tr}\left(\mathbb{E}_{c\sim\nu}\left[\sum_{\pi\in\Pi} q_\pi \phi(c,\pi(c))\phi(c,\pi(c))^\top\right]\mathbb{E}_{c\sim\nu}\left[\sum_{a\in\mathcal{A}} p_{c,a}\phi(c,a)\phi(c,a)^\top\right]^{-1}\right) \\
&\quad\le 4d
\end{aligned}
$$

where the last line takes $p_{c,a} = \sum_{\pi\in\Pi} \mathbf{1}\{\pi(c) = a\}q_\pi$, which is at least as good as the minimizing choice in the theorem. □

We remark that the algorithm that achieves this upper bound is very different than popular optimism-based algorithms for linear contextual bandits e.g., UCB or Thompson sampling [Abbasi-Yadkori et al., 2011]. Indeed, our algorithm computes an experimental design and is related to instance-dependent linear bandit algorithms developed for best-arm identification [Soare et al., 2014, Fiez et al., 2019, Degenne et al., 2020] and regret minimization [Hao et al., 2020, Tirinzoni et al., 2020]. To our knowledge, Theorem 2.2.14 provides the first instance-dependent sample complexity for the PAC setting of linear contextual bandits. The most relevant work to Theorem 2.2.14 is the work of [Zanette et al., 2021] which demonstrated a minimax sample complexity of $d^2/\epsilon^2 \log(1/\delta)$.

Remark 2.2.15 (Agnostic vs. Realizable)**.** Contrasting the above results, we note that the sample complexity of the agnostic case is always bounded by $|\mathcal{A}|d/\epsilon^2$. whereas it never exceeds d^2/ϵ^2 for the realizable case. This matches the intuition that when the number of

actions is much larger than the dimension, assuming realizability can significantly reduce the sample complexity.

2.2.4 Comparison to the Disagreement Coefficient

The work of [Foster et al., 2021a] provides regret bounds in terms of instance-dependent quantities inspired by the *disagreement coefficient*, a notion of complexity common in the active learning literature [Hanneke et al., 2014]. The following corollary relates our sample complexity to these notions of disagreement coefficients.

Define the *policy disagreement coefficient* as

$$\mathfrak{C}_{\Pi}^{\mathsf{pol}}(\epsilon_0) = \sup_{\epsilon \geq \epsilon_0} \frac{\mathbb{E}_{c\sim\nu}[\mathbf{1}\{\exists \pi \in \Pi_\epsilon : \pi(c) \neq \pi_*(c)\}]}{\epsilon}$$

where $\Pi_\epsilon := \{\pi \in \Pi : \mathbb{P}_\nu(\pi(c) \neq \pi_*(c)) \leq \epsilon\}$ and the *cost-sensitive disagreement coefficient* as

$$\mathfrak{C}_{\Pi}^{\mathsf{csc}}(\epsilon_0) = \sup_{\epsilon \geq \epsilon_0} \frac{\mathbb{E}_{c\sim\nu}[\mathbf{1}\{\exists \pi \in \Pi : \pi(c) \neq \pi_*(c), \mathbb{E}_{c\sim\nu}[\, r(c,\pi_*(c)) - r(c,\pi(c))\,] \leq \epsilon\}]}{\epsilon}.$$

The AdaCB algorithm of [Foster et al., 2021a] achieves a regret of roughly $R_T = O(\min_\delta\{\delta\Delta_{\mathsf{uniform}}T, \frac{|\mathcal{A}|\log(|\Pi|)\mathfrak{C}_{\Pi}^{\mathsf{pol}}(\delta)}{\Delta_{\mathsf{uniform}}}\})$ or $R_T = O\left(\min_\delta\left\{\delta T, |\mathcal{A}|\log(|\Pi|)\mathfrak{C}_{\Pi}^{\mathsf{csc}}(\delta)\right\}\right)$. Observe that at time T, given the outputs $\pi_1, \pi_2, \cdots, \pi_T$ from AdaCB algorithm, one could return a (randomized) policy $\tilde{\pi}$ which on observing a context, samples from the empirical distribution over the outputs. By Markov's inequality we have $\tilde{\pi}$, $V(\pi_*) - V(\tilde{\pi}) \leq O(\epsilon)$ with constant probability for $\epsilon = \frac{R_T}{T}$. Therefore, an upper bound on the regret translates to a PAC sample complexity of $\frac{|\mathcal{A}|\log(|\Pi|)}{\epsilon\Delta_{\mathsf{uniform}}}\mathfrak{C}_{\Pi}^{\mathsf{pol}}(\epsilon/\Delta_{\mathsf{uniform}})$ or $\frac{|\mathcal{A}|\log(|\Pi|)}{\epsilon}\mathfrak{C}_{\Pi}^{\mathsf{csc}}(\epsilon)$.

Finally, Corollary 2.2.16 shows that this sample complexity bound is at least as large as our upper bound, see Appendix A.1.5 for the proof.

Corollary 2.2.16. *Recall that* $\Delta_{\mathsf{uniform}} := \min_{c\in\mathcal{C}} \min_{a\in\mathcal{A}} r(c,\pi_*(c)) - r(c,a)$. *For any* $\epsilon_0 > 0$ *we have that*

1. $\rho_{\Pi,\epsilon_0} \leq \frac{2|\mathcal{A}|}{\epsilon_0\Delta_{\mathsf{uniform}}}\mathfrak{C}_{\Pi}^{\mathsf{pol}}(\epsilon_0/\Delta_{\mathsf{uniform}})$;
2. $\rho_{\Pi,\epsilon_0} \leq \frac{2|\mathcal{A}|}{\epsilon_0}\mathfrak{C}_{\Pi}^{\mathsf{csc}}(\epsilon_0)$.

Moreover, for all $\epsilon_0 \geq 0$ we have that $\rho_{\Pi,\epsilon_0} < \infty$ whenever $\Delta_{\mathsf{pol}} := V(\pi_*) - \max_{\pi \neq \pi_*} V(\pi) > 0$.

2.3 Optimal Algorithms for Contextual Bandits

2.3.1 Reduction to linear realizability and a simple elimination scheme

The astute reader may have noticed that if we ignore computation, Theorem 2.2.4 is actually an immediate corollary of Theorem 2.2.14 by taking $\phi(c,a) = \mathrm{vec}(\mathbf{e}_c \mathbf{e}_a^\top) \in \mathbb{R}^{|\mathcal{C}|\cdot|\mathcal{A}|}$ where $\mathbf{e}_i$ is a one-hot encoded vector so that $r(c,a) = \langle \phi(c,a), \theta_* \rangle$ with $\theta_* \in \mathbb{R}^{|\mathcal{C}|\cdot|\mathcal{A}|}$. This observation is key to our sample complexity results. Recall $\phi_\pi := \mathbb{E}_{c\sim\nu}[\phi(c,\pi(c))]$ from Definition 2.2.12, we have that $V(\pi) = \mathbb{E}[r(c,\pi(c))] = \mathbb{E}[\langle \phi(c,\pi(c)), \theta_* \rangle] = \langle \phi_\pi, \theta_* \rangle$. We stress that $\mathcal{C}$ can be uncountable, and thus we would never actually instantiate any of these vectors.

For notational convenience, define the feasible set of (context, action) probability distributions as $\Omega = \left\{ w \in \Delta_{\mathcal{C}\times\mathcal{A}} : \nu_c = \sum_{a\in\mathcal{A}} w_{a,c} \right\}$. Note that for each context, $p_c := \{w_{c,a}/\nu_c\}_{a\in\mathcal{A}} \in \Delta_{\mathcal{A}}$ defines a probability distribution over actions. Also define $A(w) := \sum_{c,a} w_{c,a} \phi(c,a)\phi(c,a)^\top$ for any $w \in \Omega$. Under this notation, recalling the right hand side from Theorems 2.2.13 and 2.2.14 we have

$$\min_{w\in\Omega} \max_{\pi\in\Pi\setminus\pi_*} \frac{\|\phi_\pi - \phi_{\pi_*}\|^2_{A(w)^{-1}}}{\langle \phi_{\pi_*} - \phi_\pi, \theta_* \rangle^2 \vee \epsilon^2} = \min_{p_c\in\triangle_{\mathcal{A}},\, \forall c\in\mathcal{C}} \max_{\pi\in\Pi\setminus\pi_*} \frac{\|\phi_\pi - \phi_{\pi_*}\|^2_{\mathbb{E}_{c\sim\nu}[\sum_{a\in\mathcal{A}} p_{c,a}\phi(c,a)\phi(c,a)^\top]^{-1}}}{\langle \phi_{\pi_*} - \phi_\pi, \theta_* \rangle^2 \vee \epsilon^2}$$

To show that the sample complexity of Theorem 2.2.4 is a corollary of Theorem 2.2.14, it suffices to show that equation (2.1) and the above display are equal. To see this, observe

$$\begin{aligned}
\|\phi_\pi - \phi_{\pi_*}\|^2_{A(w)^{-1}} &= \|\mathbb{E}_{c\sim\nu}[\mathrm{vec}(\mathbf{e}_c\mathbf{e}_{\pi(c)}^\top) - \mathrm{vec}(\mathbf{e}_c\mathbf{e}_{\pi_*(c)}^\top)]\|^2_{A(w)^{-1}} \\
&= \textstyle\sum_{c,a} \frac{\nu_c^2}{w_{c,a}} (\mathbf{1}\{\pi(c) = a\} + \mathbf{1}\{\pi_*(c) = a\} - 2\mathbf{1}\{\pi(c) = \pi'(c)\}) \\
&= \mathbb{E}_{c\sim\nu}\left[\left(\tfrac{1}{p_{c,\pi(c)}} + \tfrac{1}{p_{c,\pi_*(c)}} \right) \mathbf{1}\{\pi_*(c) \neq \pi(c)\} \right].
\end{aligned}$$

Due to this equivalence, the lower bound of Theorem 2.2.2 is also a corollary of Theorem 2.2.13. The lower bound of Theorem 2.2.13 follows almost immediately from the lower bound argument in [Fiez et al., 2019].

The conclusion of this section is that from a sample complexity analysis alone, all that is left is to prove Theorem 2.2.14. In the next section we propose an algorithm that achieves

this sample complexity but assumes precise knowledge of the context distribution ν (this is relaxed in following sections). While the algorithm is highly impractical for a number of reasons, its analysis provides a great deal of intuition and motivation for our final algorithm.

2.3.2 A simple, impractical, elimination-style algorithm

Algorithm 1 provides an initial elimination based method for the PAC-contextual bandit problem. The algorithm runs in stages. Before the start of each stage $\ell \in \mathbb{N}$, the algorithm defines a distribution $p_c^{(\ell)} \in \triangle_{\mathcal{A}}$ for each $c \in \mathcal{C}$. At each successive time $t \in [n_\ell]$, it plays the random action $a_t \sim p_{c_t}^{(\ell)}$ in response to context $c_t \sim \nu$, and receives random reward r_t with $\mathbb{E}[r_t|c_t, a_t] = \langle \phi(c_t, a_t), \theta_* \rangle$. Observe that

$$\mathbb{E}\left[\phi(c_t, a_t) r_t\right] = \mathbb{E}\left[\phi(c_t, a_t)\phi(c_t, a_t)^\top \theta_*\right] = \textstyle\sum_{c \in \mathcal{C}, a \in \mathcal{A}} w_{c,a}^{(\ell)} \phi(c, a)\phi(c, a)^\top \theta_* = A(w^{(\ell)})\theta_*$$

using the identity $w_{c,a}^{(\ell)} := \nu_c p_{c,a}^{(\ell)}$. Thus, if we set $O_t = A(w^{(\ell)})^{-1}\phi(c_t, a_t) r_t$ then $\mathbb{E}[O_t] = \theta_*$. A straightforward calculation also shows that $\mathrm{Cov}(O_t) = A(w^{(\ell)})^{-1}$ if r_t is perturbed with additive unit variance noise. Thus, an unbiased estimator of $\Delta(\pi, \pi_*) := V(\pi_*) - V(\pi) = \langle \phi_{\pi_*} - \phi_\pi, \theta_* \rangle$ is simply $\langle \phi_{\pi_*} - \phi_\pi, \frac{1}{n_\ell} \sum_t O_t \rangle$ which has variance $\frac{1}{n_\ell} \|\phi_{\pi_*} - \phi_\pi\|^2_{A(w^{(\ell)})^{-1}}$. Intuitively, $\langle \phi_{\pi_*} - \phi_\pi, \frac{1}{n_\ell} \sum_t O_t \rangle = \langle \phi_{\pi_*} - \phi_\pi, \theta_* \rangle \pm \sqrt{\frac{1}{n_\ell} \|\phi_{\pi_*} - \phi_\pi\|^2_{A(w^{(\ell)})^{-1}}}$ so we can safely conclude that a policy π is sub-optimal (i.e., $\pi \neq \pi_*$) if there exists any policy π' such that $\langle \phi_{\pi'} - \phi_\pi, \frac{1}{n_\ell} \sum_t O_t \rangle \gg \sqrt{\frac{1}{n_\ell} \|\phi_{\pi'} - \phi_\pi\|^2_{A(w^{(\ell)})^{-1}}}$. This is the intuition behind Contextual RAGE (Algorithm 1), which inherits its name from the best-arm identification algorithm of [Fiez et al., 2019] that inspired its strategy.

However, while $\langle \phi_{\pi_*} - \phi_\pi, \frac{1}{n_\ell} \sum_t O_t \rangle$ is unbiased and has controlled variance, it is potentially heavy-tailed because $w_{c,a}^{(\ell)}$ can be arbitrarily small. Instead of trying to control $w_{c,a}^{(\ell)}$ and appealing to Bernstein's inequality, we use the robust mean estimator of Catoni [Lugosi and Mendelson, 2019]. We can then show:

Lemma 2.3.1. *$\pi_* \in \Pi_\ell$ and $\max_{\pi \in \Pi_\ell} \langle \phi_{\pi_*} - \phi_\pi, \theta^* \rangle \leq 4\epsilon_\ell$ for all $\ell > 1$ w.p. at least $1 - \delta$.*

The lemma states that if Π_ℓ is the active set of policies still under consideration, the optimal policy π_* is never discarded from Π_ℓ, and moreover, the quality of all policies

remaining in Π_ℓ is getting better and better. The full proof of this lemma is in Appendix A.2. We are now ready to state the main sample complexity result.

Theorem 2.3.2. *Fix any policy class $\Pi = \{\pi : \mathcal{C} \to \mathcal{A}\}_\pi$, distribution over contexts ν, $\delta \in (0,1)$, $\epsilon \geq 0$, and feature map $\phi : \mathcal{C} \times \mathcal{A} \to \mathbb{R}^d$ such that $r(c,a) = \langle \phi(c,a), \theta_* \rangle$ (this is without loss generality, as one can always take $\phi(c,a) = vec(\mathbf{e}_c \mathbf{e}_a^\top)$). With probability at least $1-\delta$, if $\phi_\pi = \mathbb{E}_{c\sim\nu}[\phi(c,\pi(c))]$ and $\pi_* = \arg\max_\pi \langle \phi_\pi, \theta_* \rangle$ then* Contextual-RAGE *returns a policy $\widehat{\pi} \in \Pi$ such that $V(\widehat{\pi}) \geq V(\pi_*) - \epsilon$ after taking at most*

$$c \min_{w\in\Omega} \max_{\pi\in\Pi} \frac{\|\phi_\pi - \phi_{\pi_*}\|^2_{A(w)^{-1}}}{(\langle \phi_{\pi_*} - \phi_\pi, \theta^* \rangle \vee \epsilon)^2} \log(\log((\Delta \vee \epsilon)^{-1})|\Pi|/\delta) \log((\Delta \vee \epsilon)^{-1})$$

samples, where c is an absolute constant and $\Delta = \min_{\pi\in\Pi\setminus\pi_*} V(\pi_*) - V(\pi)$.

Proof. Define $S_\ell = \{\pi \in \Pi : \langle \phi_{\pi_*} - \phi_\pi, \theta^* \rangle \leq 4\epsilon_\ell\}$. The above lemma implies that with probability at least $1-\delta$ we have $\bigcap_{\ell=1}^\infty \{\Pi_\ell \subseteq S_\ell\}$. Observe that if for any $\mathcal{V} \subset \Pi$ we define $f(\mathcal{V}) = \min_{w\in\Omega} \rho(w, \mathcal{V})$ then

$$\rho(w^{(\ell)}, \Pi_\ell) = \min_{w\in\Omega} \max_{\pi,\pi'\in\Pi_\ell} \|\phi_\pi - \phi_{\pi'}\|^2_{A(w)^{-1}} \leq \min_{w\in\Omega} \max_{\pi,\pi'\in S_\ell} \|\phi_\pi - \phi_{\pi'}\|^2_{A(w)^{-1}} = \rho(S_\ell).$$

For $\ell \geq \lceil \log_2(4\Delta^{-1}) \rceil$ we have that $S_\ell = \{\pi_*\}$, thus the sample complexity to identify π_* is

$$\begin{aligned}
\sum_{\ell=1}^{\lceil \log_2(4\Delta^{-1}) \rceil} \tau_\ell &= \sum_{\ell=1}^{\lceil \log_2(4\Delta^{-1}) \rceil} \lceil 4\epsilon_\ell^{-2} \rho(w^{(\ell)}, \Pi_\ell) \log(2\ell^2|\Pi|/\delta) \rceil \\
&\leq \sum_{\ell=1}^{\lceil \log_2(4\Delta^{-1}) \rceil} 4\epsilon_\ell^{-2} \rho(S_\ell) \log(2\ell^2|\Pi|/\delta) + 1 \\
&\leq c \log(\log(\Delta^{-1})|\Pi|/\delta) \sum_{\ell=1}^{\lceil \log_2(4\Delta^{-1}) \rceil} \epsilon_\ell^{-2} \rho(S_\ell)
\end{aligned}$$

for some absolute constant $c > 0$. We now note that

$$\begin{aligned}
\min_{w\in\Omega}\max_{\pi\in\Pi}\frac{\|\phi_\pi-\phi_{\pi_*}\|^2_{A(w)^{-1}}}{(\langle\phi_{\pi_*}-\phi_\pi,\theta^*\rangle)^2} &= \min_{w\in\Omega}\max_{\ell\le\lceil\log_2(4\Delta^{-1})\rceil}\max_{\pi\in S_\ell}\frac{\|\phi_\pi-\phi_{\pi_*}\|^2_{A(w)^{-1}}}{(\langle\phi_{\pi_*}-\phi_\pi,\theta^*\rangle)^2} \\
&\ge \frac{1}{\lceil\log_2(4\Delta^{-1})\rceil}\min_{w\in\Omega}\sum_{\ell=1}^{\lceil\log_2(4\Delta^{-1})\rceil}\max_{\pi\in S_\ell}\frac{\|\phi_\pi-\phi_{\pi_*}\|^2_{A(w)^{-1}}}{(\langle\phi_{\pi_*}-\phi_\pi,\theta^*\rangle)^2} \\
&\ge \frac{1}{16\lceil\log_2(4\Delta^{-1})\rceil}\sum_{\ell=1}^{\lceil\log_2(4\Delta^{-1})\rceil}\epsilon_\ell^{-2}\min_{w\in\Omega}\max_{\pi\in S_\ell}\|\phi_\pi-\phi_{\pi_*}\|^2_{A(w)^{-1}} \\
&\ge \frac{1}{64\lceil\log_2(4\Delta^{-1})\rceil}\sum_{\ell=1}^{\lceil\log_2(4\Delta^{-1})\rceil}\epsilon_\ell^{-2}\min_{w\in\Omega}\max_{\pi,\pi'\in S_\ell}\|\phi_\pi-\phi_{\pi'}\|^2_{A(w)^{-1}} \\
&= \frac{1}{64\lceil\log_2(4\Delta^{-1})\rceil}\sum_{\ell=1}^{\lceil\log_2(4\Delta^{-1})\rceil}\epsilon_\ell^{-2}\rho(S_\ell)
\end{aligned}$$

where we have used the fact that $\max_{\pi,\pi'\in S_\ell}\|\phi_\pi-\phi_{\pi'}\|^2_{A(w)^{-1}} \le 4\max_{\pi\in S_\ell}\|\phi_\pi-\phi_{\pi_*}\|^2_{A(w)^{-1}}$ by the triangle inequality. □

2.3.3 Towards a more efficient algorithm

One major issue with Algorithm 1 is that it explicitly maintains a set of policies Π_ℓ from round to round. Since Π could be exponential in $|\mathcal{A}|$, this is a non-starter for any implementation. As a motivation for our approach, we consider a non-elimination algorithm, Algorithm 2, as an intermediate step. It does not maintain Π_ℓ and instead just solves the optimization problem (2.2) over Π. The design computed in (2.2) is chosen to ensure that for all $\pi \in \Pi$, $|\widehat{\Delta}_{\ell-1}(\pi,\widehat{\pi}_\ell) - \Delta(\pi,\pi_*)| \le 2\epsilon_{\ell-1} + \frac{1}{4}\Delta(\pi,\pi_*)$ with high probability (Lemma A.2.3). Equivalently, we estimate gaps up to a constant factor for policies with $\Delta(\pi,\pi_*) > \epsilon_\ell$, while our gap estimates are bounded by ϵ_ℓ for those policies satisfying $\Delta(\pi,\pi_*) \le \epsilon_\ell$. This ensures that our choice of $\widehat{\pi}_\ell$ is good enough, i.e. satisfies $V(\pi_*) - V(\widehat{\pi}_\ell) \le \epsilon_\ell$ with high probability. The full proof is in Appendix A.2.

Unfortunately, Algorithm 2 introduces additional problems. It is not clear whether solving (2.2) is computationally efficient. Also, we need to find an estimator $\widehat{\Delta}_l$ that is computationally efficient even if the policy space Π is infinite. In addition, it requires precise knowledge of ν to

even *define* the domain of distributions Ω optimized over, and store the solution $w \in \mathcal{C} \times \mathcal{A}$ explicitly. But in general, such precise knowledge will not be available and is only estimable using past data (Assumption 1).

2.3.4 An instance-optimal and computationally efficient algorithm

In this section we provide Algorithm 4, which witnesses the guarantees of Theorem 2.2.4 for the general agnostic contextual bandit problem. We now address the caveats of the previous approaches.

Access to Offline Data. By Assumption 1, we have access to a large amount of sampled offline contexts $\mathcal{D}$, where each $c_t \in \mathcal{D}$ is drawn IID from ν. Having access to $\mathcal{D}$ allows us to approximate $\mathbb{E}_{c\sim\nu}[\cdot]$ with expectations over the empirical distribution $\mathbb{E}_{c\sim\nu_{\mathcal{D}}}[\cdot]$, where $\nu_{\mathcal{D}}$ is the uniform distribution over historical data $\mathcal{D}$. The number of offline contexts we need only scales logarithmically over the size of the policy set Π, more specifically, $\mathrm{poly}(|\mathcal{A}|, \epsilon^{-1}, \log(|\Pi|), \log(1/\delta))$. We quantify the precise number of samples needed in Appendix A.3.2.

Computing the design efficiently. As described, the context space $\mathcal{C}$ may be infinite so maintaining a distribution $\omega \in \Omega \subset \Delta_{\mathcal{C}\times\mathcal{A}}$ is not possible. To overcome this issue, we consider the dual problem of equation (2.2). We can remove the square root by noticing that $2\sqrt{xy} = \min_{\gamma>0} \gamma x + \frac{y}{\gamma}$, and introducing an additional minimization over the variable $\gamma_\pi, \pi \in \Pi$. Then, the dual problem becomes

$$\max_{\lambda\in\Delta_\Pi} \min_{w\in\Omega} \min_{\gamma_\pi\geq 0} \sum_{\pi\in\Pi} \lambda_\pi \left(-\widehat{\Delta}_{l-1}(\pi, \widehat{\pi}_{l-1}) + \gamma_\pi \left\|\phi_\pi - \phi_{\widehat{\pi}_{l-1}}\right\|^2_{A(w)^{-1}} + \frac{\log(1/\delta_l)}{2\gamma_\pi n_l} \right). \tag{2.4}$$

Exchanging the order of the minimums on ω and γ, somewhat surprisingly we have the close-form expression (Lemma A.4.6)

$$\min_{\omega\in\Omega} \sum_{\pi\in\Pi} \lambda_\pi \gamma_\pi \|\phi_\pi - \phi_{\widehat{\pi}_{\ell-1}}\|^2_{A(w)^{-1}} = \mathbb{E}_{c\sim\nu} \left[\left(\sum_{a\in\mathcal{A}} \sqrt{(\lambda \odot \gamma)^\top t_a^{(c)}(\widehat{\pi}_{l-1})} \right)^2 \right],$$

Algorithm 1 Elimination Contextual RAGE

Input: Π, $\phi : \mathcal{C} \times \mathcal{A} \to \mathbb{R}^d$, $\delta \in (0,1)$

Initialize $\Pi_1 = \Pi$

for $\ell = 1, 2, \cdots, \lceil \log_2(1/\epsilon) \rceil$ **do**

$\epsilon_\ell := 2^{-\ell}, \delta_\ell := \delta/(2\ell^2|\Pi|)$

Let n_ℓ be the minimum value s.t.:

$$\min_{w \in \Omega} \max_{\pi, \pi' \in \Pi_\ell} \frac{\|\phi_\pi - \phi_{\pi'}\|^2_{A(w)^{-1}} \log(1/\delta_\ell)}{n_\ell} \leq \epsilon_\ell$$

with solution $w^{(\ell)}$.

For each $t \in [n_\ell]$, get $c_t \sim \nu$, pull $a_t \sim p^{(\ell)}_{c_t}$, observe reward r_t

Compute $O_t = A(w^{(\ell)})^{-1} \phi(c_t, a_t) r_t$.

For $\pi, \pi' \in \Pi_\ell$

$$\widehat{\Delta}_\ell(\pi, \pi') = \mathsf{Cat}(\{\langle \phi_\pi - \phi_{\pi'}, O_i \rangle\}_{i=1}^{n_\ell})$$

Update

$$\Pi_{\ell+1} = \Pi_\ell \backslash \{\pi' \in \Pi_l \mid \max_{\pi \in \Pi_\ell} : \widehat{\Delta}_\ell(\pi, \pi') > \epsilon_\ell\}$$

end for

Output: $\Pi_{\ell+1}$

Algorithm 2 Non-elimination Contextual RAGE

Input: Π, $\phi : \mathcal{C} \times \mathcal{A} \to \mathbb{R}^d$, $\delta \in (0,1)$

Initialize: $\widehat{\pi}_0 \in \Pi$ arbitrarily

for $\ell = 1, 2, \cdots, \lceil \log_2(1/\epsilon) \rceil$ **do**

$\epsilon_\ell := 2^{-\ell}$, $\delta_\ell := \delta/(2\ell^2|\Pi|)$

Let n_ℓ be the minimum value s.t.:

$$\min_{w \in \Omega} \max_{\pi \in \Pi} -\frac{1}{4} \widehat{\Delta}_{l-1}(\pi, \widehat{\pi}_{l-1}) + \sqrt{\frac{2\|\phi_\pi - \phi_{\widehat{\pi}_{l-1}}\|^2_{A(w)^{-1}} \log(1/\delta_l)}{n_\ell}} \leq \epsilon_\ell. \quad (2.2)$$

with solution $w^{(\ell)}$

For each $t \in [n_\ell]$, get $c_t \sim \nu$, pull $a_t \sim p^{(\ell)}_{c_t}$, observe reward r_t

Compute $O_t = A(w^{(\ell)})^{-1} \phi(c_t, a_t) r_t$.

For each $\pi \in \Pi$, let

$$\widehat{\Delta}_\ell(\pi, \widehat{\pi}_{\ell-1}) = \mathsf{Cat}(\{\langle \phi_\pi - \phi_{\widehat{\pi}_\ell}, O_i \rangle\}_{i=1}^{n_\ell}).$$

Set $\widehat{\pi}_\ell := \arg\min_{\pi \in \Pi} \widehat{\Delta}_\ell(\pi, \widehat{\pi}_{\ell-1})$ (2.3)

end for

Output: $\widehat{\pi}_l$

where for $\pi' \in \Pi$, $t_a^{(c)}(\pi') \in \{0,1\}^{|\Pi|}$ with $[t_a^{(c)}(\pi')]_\pi := \mathbf{1}\{\pi(c) = a, \pi'(c) \neq a\} + \mathbf{1}\{\pi(c) \neq a, \pi'(c) = a\}$ and $[\lambda \odot \gamma]_\pi := \lambda_\pi \gamma_\pi$. Interestingly, this value is achieved at a sampling distribution ω, which is a *non-linear* function of λ rather than a convex combination over policies (as in [Agarwal et al., 2014]). Because we have an expectation over contexts, this expectation can be replaced by an empirical estimate using historical data, thus avoiding any issues with an infinite context space. The final algorithm utilizing these observations found is in Algorithm 4.

The main challenge is finding a solution to the design problem (2.7). For starters, we can reduce it to a saddle point problem over (λ, γ) by considering only a dyadic sequence of $n \in \{2^k : k \in \mathbb{N}\}$. Our procedure uses an alternating ascent/descent method, with the caveat that λ lives in a simplex, and γ in a box. Both of the spaces are defined over a potentially infinite set of policies Π (which in the worst case may scale exponentially in $|\mathcal{C}|$), so we need to argue that we can find a sparse yet ϵ-good solution to (2.7).

To handle this, we use the Frank-Wolfe (FW) method on λ. Referring to the iterates of FW as λ^t, FW guarantees that the size of the support of λ^t in each iterate grows by at most 1. Thus, if initialized as a 1-sparse vector, we only need to maintain a sparse λ^t in each iteration. Each iterate of Frank-Wolfe involves computing

$$\arg\max_{\pi \in \Pi} [\nabla_\lambda h_\ell(\lambda, \gamma, n)]_\pi.$$

To do so, we show that we can appeal to a constrained argmax oracle (AMO) to run the Frank-Wolfe algorithm, a similar approach to [Agarwal et al., 2014]. At an iterate t, we use a gradient descent procedure for γ^t. We will show that in iterate t, the support of γ^t is contained in that of λ^t, and we can quantify the number of steps of gradient descent needed to find an ϵ-good solution. Though $h_l(\lambda, \gamma, n)$ might not be convex in γ, we nevertheless are able to argue that it has a unique minima and that gradient descent converges to this minima. We introduce our subroutine in Algorithm 3 and shows that it is computationally efficient with access to an argmax oracle (Definition 2.2.3) in Theorem 2.3.3.

Algorithm 3 FW-GD

Input: Π policy sets, number of actions $|\mathcal{A}|$, $\widehat{\pi}_{l-1} \in \Pi$, $\eta_l > 0$, $K \in \mathbb{N}$, threshold ϵ_l, $\gamma_{\min}$, $\gamma_{\max}$

Initialize $n_1 = 1$, $L = |\mathcal{A}|^2 \frac{((1+\eta_l)\gamma_{\max})^{5/2}}{\eta_l^{3/2}\gamma_{\min}^2}$

for $r = 1, 2, \cdots$ **do**

  Initialize $\lambda^0 = \mathbf{e}_0 \in \mathbb{R}^{\Pi}$, $\gamma^0 = \mathbf{1}_{|\Pi|} \cdot \sqrt{\frac{\log(1/\delta_l)}{|\mathcal{A}|^2 \eta_l n_r}} \in \mathbb{R}^{|\Pi|}$ // Never explicitly materialized

  for $t = 0, 1, 2, \cdots, K$ **do**

    Compute

$$\pi_t = \arg\max_{\pi \in \Pi} \left[\nabla_\lambda h_l(\lambda^t, \gamma^t, n_r)\right]_\pi \tag{2.5}$$

    Set the FW-gap

$$g_t = \left\langle \nabla_\lambda h_l(\lambda^t, \gamma^t, n_r), \mathbf{e}_{\pi_t} - \lambda^t \right\rangle = \left[\nabla_\lambda h_l(\lambda^t, \gamma^t, n_r)\right]_{\pi_t} - \sum_{\pi \in \mathrm{supp}(\lambda^t)} \left[\nabla_\lambda h_l(\lambda^t, \gamma^t, n_r)\right]_\pi$$

    Set $\beta_t = \min\left\{\frac{g_t}{L\|\lambda^t - \mathbf{e}_{\pi_t}\|_1^2}, 1\right\}$

    Set $\kappa_t = \frac{\epsilon_l}{(t+1)^2}$

    Set $\lambda^{t+1} = (1-\beta_t)\lambda^t + \beta_t \mathbf{e}_{\pi_t}$ // Only 1-sparse updates recorded

    Set $\gamma^{t+1} = \mathsf{GD}(\lambda^t, n_r, \kappa_t)$ // Only differences from γ_0 recorded

  end for

  if $h_l(\lambda^{K+1}, \gamma^{K+1}, n_r) \le \epsilon_l$ **then**

    break

  else

    $n_{r+1} = 2 \cdot n_r$

  end if

end for

Output: $\lambda^{K+1} \in \triangle_\Pi, \gamma^{K+1} \in \mathbb{R}_+^{|\Pi|}, n_r$

Theorem 2.3.3. *Let K_l be the number of iterations for FW-GD in the lth round and λ^*, γ^* be the exact solutions to the optimization problem $\max_{\lambda \in \Delta_\Pi} \min_{\gamma \in [\gamma_{\min}, \gamma_{\max}]^{|\Pi|}} h_l(\lambda, \gamma, n)$. Then, $K_l = \mathrm{poly}_1(|\mathcal{A}|, \epsilon_l^{-1}, \log(1/\delta))$ and the outputs $\lambda^{K_l+1}, \gamma^{K_l+1}$ satisfy $h_l(\lambda^*, \gamma^*, n) -$*

$h_l(\lambda^{K_l+1}, \gamma^{K_l+1}, n) \leq \epsilon_l$ with at most $O(K_l^2|\mathcal{D}|)$ calls to a constrained argmax oracle, where the size of the history $\mathcal{D}$ exceeding $\mathrm{poly}_2(\epsilon^{-1}, \log|\Pi|, \gamma_{\max}, \gamma_{\min}^{-1}, \eta^{-1}, |\mathcal{A}|, \log(1/\delta))$ *with probability at least $1-\delta$, where* poly_1, poly_2 *denote some polynomial.*

The full proof is in Appendix A.3. It is worth noting that we can bound the suboptimality error $h_l(\lambda^*, \gamma^*, n) - h_l(\lambda^{K_l+1}, \gamma^{K_l+1}, n)$ by the duality gap, as the primal objective is always at least as large as the optimum. Also, the Frank-Wolfe algorithm directly tackles the duality gap, so standard Frank-Wolfe analysis will show that the Frank-Wolfe output makes the duality gap small [Pedregosa et al., 2020].

Regularized Estimator. While Algorithms 1 and 2 use a robust mean estimator as in equation (2.3), this estimator is impractical with a very large number of policies Π. Instead, we use a regularized IPS estimator that can be computed using historical data and an argmax oracle.

Algorithm 4 puts it all together and Theorem 2.3.4 shows our main result. Note that for exposition purposes, we have omitted some additional regularization terms in the optimization problems that have no effect on the sample complexity, but ensure finite-time convergence. Appendix A.4 shows the full algorithm and the proof. In what follows, $\mathrm{poly}_1(|\mathcal{A}|, \epsilon^{-1}, \log(1/\delta)) \cdot \log(|\Pi|)$ and $\mathrm{poly}_2(|\mathcal{A}|, \epsilon^{-1}, \log(1/\delta), \log(|\Pi|))$ are polynomials in their arguments that specified in the appendix.

Theorem 2.3.4. *Fix any set of policies Π, context distribution ν and reward function $r(c,a) \in [0,1]$. With probability at least $1-\delta$, provided a history $\mathcal{D}$ whose size exceeds* $\mathrm{poly}_1(|\mathcal{A}|, \epsilon^{-1}, \log(1/\delta)) \cdot \log(|\Pi|)$*, Algorithm 4 returns a policy $\widehat{\pi}$ satisfying $V(\pi_*) - V(\widehat{\pi}_\ell) \leq \epsilon$ in a number of samples not exceeding $O(\rho_{\Pi,\epsilon} \log(|\Pi| \log_2(1/\Delta_\epsilon)/\delta) \log_2(1/\Delta_\epsilon))$ where $\Delta_\epsilon := \max\{\epsilon, \min_{\pi \in \Pi} V(\pi_*) - V(\pi)\}$.*

In addition, Algorithm 4 is computationally efficient and requires the amount of calls not exceeding $\mathrm{poly}_2(|\mathcal{A}|, \epsilon^{-1}, \log(1/\delta), \log(|\Pi|))$ *to a constrained argmax oracle.*

Algorithm 4 Contextual Oracle-efficient Dualized Algorithm (CODA)

Input: policies $\Pi = \{\pi : \mathcal{C} \to \mathcal{A}\}_\pi$, feature map $\phi : \mathcal{C} \times \mathcal{A} \to \mathbb{R}^d$, $\delta \in (0,1)$, historical data $\mathcal{D} = \{\nu_s\}_s$

1: initiate $\widehat{\pi}_0 \in \Pi$ arbitrarily, $\lambda_0 = \mathbf{e}_{\widehat{\pi}_0}$, $\widehat{\Delta}_0(\pi)$, γ_0, $\gamma_{\min}$, $\gamma_{\max}$ appropriately

2: **for** $l = 1, 2, \cdots$ **do**

3: $\epsilon_l = 2^{-l}$, $\delta_l = \delta/(l^2|\Pi|^2)$

4: Define

$$h_l(\lambda, \gamma, n) = \sum_{\pi\in\Pi} \lambda_\pi \left(-\widehat{\Delta}_{l-1}^{\gamma_{l-1}}(\pi, \widehat{\pi}_{l-1}) + \frac{\log(1/\delta_l)}{\gamma_\pi n} \right) + \mathbb{E}_{c\sim\nu_\mathcal{D}} \left[\left(\sum_{a\in\mathcal{A}} \sqrt{(\lambda \odot \gamma)^\top t_a^{(c)}(\widehat{\pi}_{l-1})} \right)^2 \right]. \tag{2.6}$$

5: Let $\lambda^l, \gamma^l, n_l = \mathsf{FW\text{-}GD}(\Pi, |\mathcal{A}|, \widehat{\pi}_{l-1}, \epsilon_l)$. These are the solutions to

$$n_\ell := \min\{n \in \mathbb{N} : \max_{\lambda\in\Delta_\Pi} \min_{\gamma\in[\gamma_{\min},\gamma_{\max}]^{|\Pi|}} h_l(\lambda, \gamma, n) \le \epsilon_\ell\} \tag{2.7}$$

6: For $i \in [n_\ell]$ get $c_i \sim \nu$, pull $a_i \sim p_{c_i}^{(\ell)}$ where $p_{c_s,a_s}^{(\ell)} \propto \sqrt{(\lambda_l \odot \gamma_l)^\top t_{a_s}^{(c_s)}(\widehat{\pi}_{l-1})}$, observe rewards r_s

7: For each $\pi \in \Pi$, define the IPS estimator

$$\widehat{\Delta}_l^{\gamma_l}(\pi, \widehat{\pi}_{l-1}) = \sum_{s=1}^{n_l} \frac{r_s}{p_{c_s,a_s}^{(\ell)} + [\gamma_l]_\pi} (\mathbf{1}\{\widehat{\pi}_{l-1}(c_s) = a_s\} - \mathbf{1}\{\pi(c_s) = a_s\})$$

8: set

$$\widehat{\pi}_l = \arg\min_{\pi\in\Pi} \widehat{\Delta}_l^{\gamma_l}(\pi, \widehat{\pi}_{l-1}) + \mathbb{E}_{c\sim\nu_\mathcal{D}} \left[\left(\frac{[\gamma_l]_\pi}{p_{c,\pi(c)}^{(\ell)}} + \frac{[\gamma_l]_\pi}{p_{c,\widehat{\pi}_{l-1}(c)}^{(\ell)}} \right) \mathbf{1}\{\widehat{\pi}_{l-1}(c) \neq \pi(c)\} \right] + \frac{\log(1/\delta_l)}{[\gamma_l]_\pi n_l} \tag{2.8}$$

9: **end for**

Output: $\widehat{\pi}_l$

2.4 Conclusion

This work provides the first instance-dependent lower bounds for the (ϵ, δ)-PAC contextual bandit problem. One limitation of this work is that our analysis of Algorithm 4 does not immediately extend to the realizable linear setting. That is, a computationally efficient algorithm that achieves the same bound is not known to exist. In all other settings discussed in this work, we proposed a computationally efficient algorithm. A second limitation is the assumption that we have access to a large pool of offline data. Because it seems necessary to plan with some information about the context distribution, it is not clear how one would completely remove such an assumption and achieve the same sample complexity bounds. As with any recommender system, there is the potential for unintended consequences from optimizing just a single metric. Moreover, other potential pitfalls can arise, such as negative feedback loops, if our assumptions fail to hold in real-world environments. Such consequences can be mitigated by tracking a diverse set of metrics.

Chapter 3

OPTIMAL EXPLORATION IS NO HARDER THAN THOMPSON SAMPLING

3.1 Introduction

The pure exploration bandit problem considers a sequential game between a learner with two sets of arms $\mathcal{X}, \mathcal{Z} \subset \mathbb{R}^d$ and nature. In each round, the learner chooses an arm $x \in \mathcal{X}$ and observes a noisy stochastic reward $y = x^\top \theta_* + \epsilon$ where $\theta_* \in \Theta$ is an unknown parameter vector and ϵ is assumed to be i.i.d Gaussian noise. The goal of the learner is to identify $z_* = \arg\max_{z \in \mathcal{Z}} z^\top \theta_*$ with high probability in a few measurements. The case of $\mathcal{X} = \mathcal{Z}$ is perhaps the most natural case to consider, and has enjoyed a fair amount of attention [Soare et al., 2014, Fiez et al., 2019, Degenne et al., 2020]. However, all proposed approaches share a common trait - complexity. Existing optimal algorithms rely on either explicitly enumerating a potentially large subset of $\mathcal{Z}$ or periodically solving a convex optimization program at every iteration. Consequently, it prompts us to question: is such complexity indeed indispensable for reaching asymptotic optimality?

Maintaining our focus on the specific instance where $\mathcal{X} = \mathcal{Z}$, we note that the pure exploration task can be addressed using any readily available regret minimization algorithm. That is, if an algorithm generates a series of plays $\{x_t\}_{t=1}^T$ such that $\max_{x \in \mathcal{X}} \sum_{t=1}^T \langle \theta_*, x - x_t \rangle \leq d\sqrt{T}$ then this immediately implies that $\widehat{x}_T$ drawn uniformly from the set $\{x_t\}_{t=1}^T$ is equal to $x_* = \arg\max_{x \in \mathcal{X}} \langle x, \theta_* \rangle$ with constant probability as soon as $T \geq d^2/\Delta_{\min}^2$, where $\Delta_{\min} = \min_{x \in \mathcal{X}, x \neq x_*} \theta_*^\top (x^* - x)$. One popular regret-minimization algorithm is Thompson Sampling (TS). Following its re-emergence from nearly seven decades of relative obscurity, it has rapidly ascended to become the most prevalently applied bandit algorithm in practical scenarios, as per the industrial experience of the authors. We postulate that its popularity is

due to (1) its simplicity to implement, (2) its flexibility to encode side-information in its prior, (3) its computational efficiency, and (4) strong empirical performance. The algorithm works by maintaining a distribution p_t over Θ given all observations up to the time t, and then plays $x_t = \arg\max_{x \in \mathcal{X}} \langle x, \theta_t \rangle$ where $\theta_t \sim p_t$. Once $y_t = \langle x_t, \theta_* \rangle + \epsilon_t$ is observed, the distribution is updated and the process repeats. As we can see, TS only relies on the ability to sample from a posterior distribution and compute a maximum inner product (an argmax oracle) - both operations which have been heavily studied and optimized. Unfortunately, TS is known to be sub-optimal for the pure exploration linear bandits problem due to its greedy exploration strategy. Indeed, there exist instances of $\mathcal{X}$ and θ_* for which the sample complexity of TS to identify the best arm scales *quadratically* in the optimal sample complexity achieved by other algorithms [Soare et al., 2014]. Even for regret minimization, it is know that TS is far from optimal from an instance-dependent perspective [Lattimore and Szepesvari, 2017]. But yet, due to its many favorable properties it is still the go-to algorithm in practice.

This paper aims to answer the following fundamental theoretical question: *Is there an algorithm that enjoys asymptotically optimal exploration that does not need to explicitly enumerate $\mathcal{Z}$ and only relies on posterior sampling and an argmax oracle?* We achieve this goal by not striving too far from the Thompson sampling algorithm itself and only assuming access to a sampling oracle and arg-max oracle. In fact, our proposed algorithm can be viewed as a generalization of Top-Two Thompson Sampling for the standard multi-armed bandit game [Russo, 2016] to the richer linear setting. At each iteration t, we maintain a sampling distribution centered at $\widehat{\theta}_t$ (a least squares estimator computed after t samples), and get a sample θ_t whose best arm is different than that of $\widehat{\theta}_t$ using a sampling oracle. Once such a θ_t is found, we update an online learner maintaining a distribution over $\mathcal{X}$ with rewards $\|\theta_t - \widehat{\theta}_t\|^2_{xx^\top}$. We prove that $\mathbb{P}(\widehat{z}_t \neq z_* | \{x_s\}_{s=1}^{t-1})$ decreases at an exponential rate with the exponent of the optimal fixed allocation. We also demonstrate that our method is not only theoretically sound by achieving an optimal sample complexity given oracle access, but is also computationally efficient empirically.

3.1.1 Problem Setting and Notation

We first define the linear bandit setting. Let $\mathcal{X}, \mathcal{Z} \in \mathbb{R}^d$ be two sets of arms and $\Theta \subset \mathbb{R}^d$ be the parameter space. At time t, we draw an action $x_t \in \mathcal{X}$, and receive the reward $y_t = x_t^\top \theta_* + \epsilon_t$ where $\theta_* \in \Theta$ and ϵ_t is i.i.d. Gaussian noise. The choice of arm x_t at time t is dependent on the filtration generated by $\{(x_s, y_s)\}_{s=1}^{t-1}$; furthermore, we denote the conditional probability given this filtration be $\mathbb{P}_\theta$.

Goal: We are interested in the best-arm identification task, i.e. we would like to find $z_* := \arg\max_{z \in \mathcal{Z}} z^\top \theta_*$ with high probability, while minimizing the number of measurements taken in $\mathcal{X}$.

We make the following assumption on the parameters that we will discuss further in Section 3.3.1.

Assumption 2. Θ is closed and bounded, with a non-empty interior.

Assumption 3. Assume that $\max_x \|x\|_2 \le L$.

Assumption 4. Assume that $\text{span}(\mathcal{Z}) \subset \text{span}(\mathcal{X})$ and the optimal arm $z_* \in \mathcal{Z}$ is unique.

Notation. For any matrix $A \in \mathbb{R}^{d \times d}$, we define the norm $\|x\|_A^2 := x^\top A x$. Given a set $\mathcal{S}$, we define the simplex $\triangle_{\mathcal{S}} := \{\lambda \in \mathbb{R}_{\ge 0}^{|\mathcal{S}|} : \sum_{i=1}^{|\mathcal{S}|} \lambda_i = 1\}$. Finally, given a (multivariate) normal distribution $\mathcal{N}(\theta, \Sigma^{-1})$ on $\mathbb{R}^d$ and some set Θ, we define the truncated normal distribution, denoted as $\text{TN}(\theta, \Sigma^{-1}; \Theta)$, to be the normal distribution restricted on Θ. For some $\lambda \in \triangle_{\mathcal{X}}$, we define $A(\lambda) := \sum_{x \in \mathcal{X}} \lambda_x x x^\top$. We define $\Delta_{\max} := \max_{x \in \mathcal{X}} \max_{\theta, \theta' \in \Theta} |x^\top(\theta - \theta')|$. We define the constants used in the algorithm as $C_{3,\ell} = \Delta_{\max} + L^2 \sqrt{d \log(T_\ell \ell^2)}$. The precise definition is in Appendix B.1.

3.2 Motivating our approach

Among all adaptive algorithms, it is known that for every $\theta_* \in \Theta$ there exists a $\lambda \in \triangle_{\mathcal{X}}$ such that sampling $x_1, x_2, \ldots, \overset{i.i.d.}{\sim} \lambda$ achieves the optimal sample complexity in the fixed

confidence setting [Soare et al., 2014, Fiez et al., 2019, Degenne et al., 2020]. Specifically, for any $\Theta \subset \mathbb{R}^d$ and $\mathcal{X}, \mathcal{Z} \subset \mathbb{R}^d$ define

$$\tau^* := \max_{\lambda \in \triangle_{\mathcal{X}}} \min_{\theta \in \Theta^c_{z_*}} \tfrac{1}{2} \|\theta - \theta_*\|^2_{A(\lambda)^{-1}} \tag{3.1}$$

where $\Theta^c_{z_*} = \{\theta \in \Theta : \exists z \in \mathcal{Z}, z^\top \theta \geq z_*^\top \theta\}$. Then it is known that to identify z_* with probability at least $1-\delta$, the expected sample complexity of any algorithm scales as $(\tau^*)^{-1} \log(2.4/\delta)$. Moreover, sampling according to the λ that achieves the maximum, when paired with an appropriate stopping time, achieves the optimal sample complexity asymptotically. As our setting is more naturally analyzed in the so-called fixed budget setting, we next state a result that can be viewed as a generalization of the result of Russo [2016] originally stated for the multi-armed bandit setting. Note that this is a lower bound similar to Glynn and Juneja [2004] and not a lower bound for the traditional fixed budget setting in multi-armed bandits [Karnin et al., 2013], since we only allow fixed λ not adapting to the observations.

Theorem 3.2.1. *Fix $\Theta = \mathbb{R}^d$ and any $\theta_* \in \Theta$. For some λ consider a procedure that draws $x_1, \ldots, x_T \sim \lambda$, then observes $y_t = \langle x_t, \theta_* \rangle + \epsilon_t$ for each t with $\epsilon_t \sim \mathcal{N}(0,1)$, and then computes $\widehat{z}_T = \arg\max_{z \in \mathcal{Z}} \langle z, \widehat{\theta}_T \rangle$ where $\widehat{\theta}_T = \arg\min_{\theta \in \Theta} \sum_{t=1}^T \|y_t - \langle \theta, x_t \rangle\|_2^2$. Then for any $\lambda \in \triangle_{\mathcal{X}}$ we have*

$$\limsup_{T \to \infty} -\frac{1}{T} \log\left(\mathbb{P}_{\theta_*, x_t \sim \lambda}(\widehat{z}_T \neq z_*)\right) \leq \tau^*.$$

The quantity τ^* is naturally interpreted from a hypothesis-testing lens. Given a fixed sampling distribution λ, note that $\mathbb{E}_{x \sim \lambda} KL(\mathcal{N}(\theta^\top x, 1) || \mathcal{N}(\theta_*^\top x, 1)) = \frac{1}{2}\|\theta - \theta_*\|^2_{A(\lambda)}$. Thus the min-max problem above aims to construct the distribution λ which maximizes the smallest KL divergence between θ and any alternative with a different best-arm. As noticed by many authors, this can be translated into a game-theoretic language. The max-player chooses a distribution over the set of possible measurements $\mathcal{X}$. At the same time, the min-player chooses an alternative θ whose best arm is not z_* in an attempt to fool the λ-player. This lower bound intuitively suggests a strategy for algorithm designers: devise a sampling method that ensures the resultant allocation aligns with the aforementioned objective.

In this pursuit (discussed extensively in Section 3.4) the game-theoretic perspective has been directly exploited by several works to give asymptotically optimal algorithms. The approaches of these works differ in detail but are similar in spirit and are motivated by the following oracle strategy that has access to θ_*. At each time, the max-player utilizes a no-regret online learner, such as exponential weights [Bubeck, 2011], to set λ_{t+1} based on an estimate of the best-response of the min-player, namely $\min_{\theta \in \Theta^c_{z_*}} \|\theta - \theta_*\|^2_{A(\lambda_t)}$. This guarantees that

$$\max_{\lambda \in \triangle_{\mathcal{X}}} \min_{\theta \in \Theta^c_{z_*}} \|\theta - \theta_*\|^2_{A(\lambda)} - \sum_{t=1}^{T} \min_{\theta \in \Theta^c_{z_*}} \|\theta - \theta_*\|^2_{A(\lambda_t)} \leq o(T)$$

which by a standard Jensen's inequality argument is sufficient to ensure that $\frac{1}{T}\sum_{t=1}^{T} \lambda_t$ is an approximate solution to the original saddle point problem. Then, the arm x_t pulled is sampled from λ_t at each time (or a deterministic tracking strategy is used).

The main computational challenge in this approach is that obtaining the best-response can be rather involved. The alternative set can be decomposed as a union of intersections of a convex set with a halfspace: $\Theta^c_{z_*} = \cup_{z \neq z_*} \Theta \cap \{\theta \in \mathbb{R}^d : z^\top \theta \geq z_*^\top \theta\}$. Thus computing the best-response involves computing $|\mathcal{Z}|$-many projections onto convex sets. For small values of $|\mathcal{Z}|$, this may be feasible. However, this computation may be onerous if $|\mathcal{Z}|$ is large or the projection step is very expensive, for example, in many combinatorial bandit settings such as shortest path problems in a graph [Chen et al., 2017]. As another example, in practical recommendation systems where $\mathcal{Z}$ represents items to be recommended, $|\mathcal{Z}|$ may be in the millions. Thus computing $|\mathcal{Z}|$ many projections under latency constraints may be impossible, even though Thompson Sampling can easily recommend good items [Biswas et al., 2019]. In addition, for both settings, there may be no easy closed-form expression for the projection.

Our method is based on the following equivalent formulation of τ^*. By linearizing the

min over alternatives with a distribution over $\Theta^c_{z_*}$, we can apply Sion's minimax theorem:

$$\begin{aligned}
&\max_{\lambda \in \triangle_{\mathcal{X}}} \inf_{\theta \in \Theta^c_{z_*}} \frac{1}{2} \|\theta - \theta_*\|^2_{A(\lambda)} \\
&= \max_{\lambda \in \triangle_{\mathcal{X}}} \min_{p \in \triangle(\Theta^c_{z_*})} \mathbb{E}_{\theta \sim p} \left[\frac{1}{2} \|\theta - \theta_*\|^2_{A(\lambda)} \right] \\
&= \min_{p \in \triangle(\Theta^c_{z_*})} \max_{\lambda \in \triangle_{\mathcal{X}}} \mathbb{E}_{\theta \sim p} \left[\frac{1}{2} \|\theta - \theta_*\|^2_{A(\lambda)} \right],
\end{aligned}$$

where $\triangle(\Theta^c_{z_*})$ denotes the set of distribution over the alternative set $\Theta^c_{z_*}$. This replaces the projections with an expectation over a distribution on $\Theta^c_{z_*}$. At first glance, the situation may seem worse - we have gone from finitely many projections to needing to maintain a distribution over a potentially infinite set!

However, imagine that Θ is finite and that we solve this saddle-point problem by maintaining a no-regret learner for the max-player as before, while similarly maintaining a no-regret learner for the min-player. Standard results in convex optimization guarantee that the average of the iterates of the two learners converge to a saddle point eventually [Liu and Orabona, 2022]. To be more precise, at each round t we draw an $x_t \sim \lambda_t$ and feed the (stochastic) loss $\sum_{\theta \in \Theta^c_{z_*}} p_{t,\theta} \|\theta - \theta_*\|^2_{x_t x_t^\top}$ to the learner for the min-player. Assuming the min-player learner is exponential weights, then the update is

$$p_{t+1,\theta} \propto p_{t,\theta} e^{-\eta \|\theta_* - \theta\|^2_{x_t x_t^\top}} \propto e^{-\eta \|\theta_* - \theta\|^2_{\sum_{s=1}^t x_s x_s^\top}}.$$

where η is an appropriate step-size. Hence, the resulting distribution p_{t+1} is reminiscent of the probability density function of a multivariate normal distribution $N(\theta_*, \eta^{-1}(\sum_{s=1}^t x_s x_s^\top)^{-1})$ restricted to $\Theta^c_{z_*}$. This observation motivates our algorithm - for the min-player we maintain an appropriate normal distribution and at each round, use samples from this distribution to generate a stochastic loss to feed the max-player. *This approach avoids explicitly maintaining $\mathcal{Z}$ or ever needing to compute a projection!* Of course, this discussion has relied on knowledge of θ_* and z_*. In the next section, we explain how our algorithm, PEPS, overcomes these restrictions.

3.3 Best Arm Identification through Sampling

Algorithm 5 Pure Exploration with Projection-Free Sampling (PEPS)

Input: Finite set of arms $\mathcal{X} \subset \mathbb{R}^d$, $\mathcal{Z} \subset \mathbb{R}^d$, time horizon T, $\eta_\lambda, \eta_p, \alpha$

Define $\lambda^G = \arg\min_{\lambda \in \triangle_\mathcal{X}} \max_{x \in \mathcal{X}} \|x\|^2_{A(\lambda)^{-1}}$, $\lambda_1 = \frac{1}{|\mathcal{X}|}\mathbf{1}$

Initialize $V_0 = I$, $S_0 = 0$, $p_1 = N(0, V_0)$, $\widehat{\theta}_1$ arbitrarily

for $t = 1, 2, \cdots, T$ **do**

  $\gamma_t = t^{-\alpha}$

  //Top Two Sampling

  Compute $\widehat{z}_t = \underset{z \in \mathcal{Z}}{\operatorname{argmax}}\, z^\top \widehat{\theta}_t$

  Sample $\theta_t = \mathsf{SAMPLE}(\mathrm{TN}(\widehat{\theta}_t, \eta_p^{-1} V_{t-1}^{-1}; \Theta^c_{\widehat{z}_t}))$

  //Take Sample and Observe Reward

  Sample $x_t \sim \tilde{\lambda}_t$ where $\tilde{\lambda}_t = (1 - \gamma_t)\lambda_t + \gamma_t \lambda^G$

  Observe $y_t = \langle \theta_*, x_t \rangle + \epsilon_t$ where $\epsilon_t \sim \mathcal{N}(0, 1)$

  //Update

  Update $V_t = V_{t-1} + x_t x_t^\top$, $S_t = S_{t-1} + x_t y_t$, and $\widehat{\theta}_{t+1} = V_t^{-1} S_t$

  Update $\lambda_{t+1} \propto \lambda_t e^{\eta_\lambda \tilde{g}_t}$ where $\tilde{g}_{t,x} = \left\| \theta_t - \widehat{\theta}_t \right\|^2_{xx^\top}, \forall x \in \mathcal{X}$

end for

Sample $\tilde{\theta} = \mathsf{SAMPLE}(\mathrm{TN}(\widehat{\theta}_{T+1}, V_T^{-1}; \Theta))$

Output: $\widehat{z}_\ell(\tilde{\theta}) = \arg\max_{z \in \mathcal{Z}} z^\top \tilde{\theta}$

Our main method PEPS is presented in Algorithm 5. Given a budget of T samples, we repeatedly sample θ_t utilizing a sampling oracle SAMPLE. We then sample an $x_t \sim \tilde{\lambda}_t$ where $\tilde{\lambda}_t$ is the distribution λ_t maintained by the λ-learner at time t mixed in with a diminishing amount γ_t of the G-optimal distribution λ^G. After playing x_t and observing a reward y_t, PEPS updates both the λ_t and the estimate $\widehat{\theta}_t$ with the covariance. In particular, given samples $\{x_s\}_{s=1}^t$, we let $\widehat{\theta}_{t+1} = V_t^{-1} S_t$ where $V_t = \sum_{s=1}^t x_s x_s^\top$ and $S_t = \sum_{s=1}^t x_s y_s$. Algorithm 5

Algorithm 6 Doubling trick

Input: Finite set of arms $\mathcal{X} \subset \mathbb{R}^d$, $\mathcal{Z} \subset \mathbb{R}^d$

1: **for** $\ell = 0, 1, \cdots, L$ **do**

2: Set $T_\ell = 2^\ell$, $\eta_\lambda = \sqrt{\frac{\log|\mathcal{X}|}{C_{3,\ell}^2 T_\ell}}$, $\eta_p = \sqrt{\frac{d\log(T_\ell C_{3,\ell})}{C_{3,\ell}^2 T_\ell}}$, $\alpha = 1/4$

3: $\widehat{z}_\ell = \mathrm{PEPS}(\mathcal{X}, \mathcal{Z}, T_\ell, \eta_\lambda, \eta_p, \alpha)$

4: **end for**

Output: $\widehat{z}_L$

depends on a finite time horizon T. To ensure that our algorithm is anytime and eventually converges to the optimal sampling scheme, we employ an outer loop Algorithm 6 utilizing a doubling scheme. Before we explain the theoretical guarantees, we first detail some of the aspects of the algorithm.

Updating the sampling distribution for θ_t. Our main innovation is introducing a distribution over $\Theta^c_{\widehat{z}_t}$ from which we can sample over. In particular, in each round, we sample θ_t from $\mathrm{TN}(\widehat{\theta}_t, \eta_p^{-1}V_{t-1}^{-1}; \Theta^c_{\widehat{z}_t})$, which is a *truncated normal distribution* with support $\Theta^c_{\widehat{z}_t}$ [Burkardt, 2014].

Following the discussion in the Section 3.2, it is tempting to see this update as a form of continuous exponential weights [Bubeck, 2011]. However, this is not quite true since the underlying action set $\Theta^c_{\widehat{z}_t}$ is changing each round. This creates several technical challenges in the proof. Note that similar to previous works, we could have maintained a learner for each $z \in \mathcal{Z}$ [Degenne et al., 2020]. However, our approach of maintaining a distribution prevents the need for this additional complexity of enumerating $\mathcal{Z}$.

From the perspective of exponential weights, η_p is a step size: the dependence on d in the numerator comes from the dimension of Θ; and $C_{3,\ell}^2$ is an upper bound on the stochastic loss $\|\theta_t - \widehat{\theta}_t\|^2_{x_t x_t^\top}$ that we guarantee with high probability due to forced exploration and boundedness of Θ.

We have the following regret guarantee on the online min learner. For notational conve-

nience, in this section, for some set $\mathcal{S}$ with nonempty interior, we let $p_t(\mathcal{S}) = \mathrm{TN}(\widehat{\theta}_t, \eta_p^{-1} V_{t-1}^{-1}; \mathcal{S})$ be the truncated normal distribution with support on $\mathcal{S}$.

Lemma 3.3.1 (informal)**.** *In round T_ℓ of epoch ℓ of Algorithm 6, we have with probability greater than $1 - 1/\ell^2$,*

$$\sum_{t=1}^{T_\ell} \mathbb{E}_{\theta \sim p_t(\Theta_{z_*}^c)} \left[\left\| \theta - \widehat{\theta}_t \right\|_{x_t x_t^\top}^2 \right] - \inf_{\theta \in \Theta_{z_*}^c} \|\theta - \theta_*\|_{V_{T_\ell}}^2$$
$$\leq O(d\sqrt{T_\ell} \log(LT_\ell)).$$

Sampling Oracle. Our algorithm involves a sampling oracle that takes samples from a truncated normal distribution.

Definition 3.3.2 (Sampling oracle (SAMPLE))**.** The oracle $\mathsf{SAMPLE}(p)$ is an algorithm that given some distribution p, returns a sample $\theta \sim p$.

There are various ways to implement this sampling oracle efficiently. The easiest way is to use rejection sampling. In particular, on line 7, for each round t, we repeatedly sample $\theta_t \sim N(\widehat{\theta}_t, \eta_p^{-1} V_{t-1}^{-1})$ until the best-arm of $\arg\max_{z \in \mathcal{Z}} z^\top \theta_t$ is not our current best guess $\widehat{z}_t = \arg\max_{z \in \mathcal{Z}} z^\top \widehat{\theta}_t$, and on line 17 we repeatedly sample $\tilde{\theta} \sim N(\widehat{\theta}_{T+1}, V_T^{-1})$ until $\tilde{\theta} \in \Theta$. Regarding the computation cost of rejection sampling, we suffer from some of the same challenges as Top-two sampling algorithms, which empirically work well in practice [Russo, 2016]. From a practical perspective, the rejection sampling step is only computationally costly if it requires many draws from the posterior to find a θ in the alternative $\Theta_{\widehat{z}_t}^c$. However, note that if we draw $O(1/\nu)$ vectors and none of them are in the alternative $\Theta_{\widehat{z}_t}^c$, by Markov's inequality, this arm they all agree on is the best arm with probability $1 - \nu$. Thus, as soon as it becomes computationally costly to sample an alternative, the problem is basically solved. We demonstrate empirically that the computational complexity is not at all onerous in Section 3.5 and Appendix B.6. Also, we note that our focus is on the query complexity given an effective way to sample, not the complexity of sampling from the distribution itself.

Since the sampling oracle only returns one sample at the end, our algorithm still achieves an asymptotically optimal *sample complexity* even if we draw $O(1/\nu)$ vectors inside the oracle.

Moreover, we remark that sampling from truncated normal distributions is a well-explored practice across statistics and machine learning, especially when sampling in a convex set. A variety of efficient methods such as Gibbs and hit-and-run procedures are available for this purpose [Devroye, 1986, Murphy, 2013, Li and Ghosh, 2015, Laddha and Vempala, 2023]. In particular, the hit-and-run algorithm ensures one gets a sample in the convex set with probability $1-\nu$ in $O(d^3 \log(1/\nu))$ samples in the worst case [Lovász, 1999]. Furthermore, novel approaches have improved the efficiency of traditional rejection techniques, especially when dealing with a convex support of the truncated normal distribution [Maatouk and Bay, 2016].

Update for λ_t. To update λ_t, which corresponds to the action of our max-player, we employ an exponential weighted learner (Hedge) over the set of actions $\mathcal{X}$. The reward vector $\tilde{g}_t \in \mathbb{R}^{|\mathcal{X}|}$ is stochastic with expectation $\mathbb{E}\tilde{g}_{t,x} = \mathbb{E}_{\theta \sim p_t(\Theta^c_{\widehat{z}_t})} \left\|\theta - \widehat{\theta}_t\right\|^2_{xx^\top}$ conditioning on the history of the algorithm $\{(x_s, y_s, \theta_s)\}_{s=1}^{t-1}$, and is bounded in high probability. We show that if we choose $\alpha = \frac{1}{4}$ and let $\tilde{\Delta}_{\max}$ be an upper bound on the loss function, we have the following regret guarantee:

Lemma 3.3.3 (informal)**.** *In round T_ℓ of epoch ℓ of Algorithm 6, we have with probability greater than $1 - 1/\ell^2$,*

$$\max_{\lambda \in \triangle_{\mathcal{X}}} \sum_{t=1}^{T_\ell} \mathbb{E}_{\theta \sim p_t(\Theta^c_{\widehat{z}_t})} \left\|\theta - \widehat{\theta}_t\right\|^2_{A(\lambda)} - \sum_{t=1}^{T_\ell} \mathbb{E}_{\theta \sim p_t(\Theta^c_{\widehat{z}_t})} \left\|\theta - \widehat{\theta}_t\right\|^2_{A(\lambda_t)} \leq O\left(\sqrt{(d + \tilde{\Delta}_{\max}) T_\ell \log \ell}\right)$$

Forced Exploration with G-optimal Design. To ensure adequate sampling in all directions, in each round we mix in some amount of the G-optimal distribution, denoted as $\lambda^G := \arg\min_{\lambda \in \Delta_{\mathcal{X}}} \max_{x \in \mathcal{X}} \|x\|^2_{A(\lambda)^{-1}}$. This ensures that $\max_{x \in \mathcal{X}} \|\widehat{\theta}_t - \theta\|_{xx^\top}$ is bounded

for any $\theta \in \Theta$ and $\widehat{z}_t$ is eventually z_* with probability 1. The rate at which the mixture of this distribution decays as $t^{-\alpha}$, for any $0 < \alpha < 1/2$, so it has no effect on asymptotic performance. We note that thanks to the implicit anti-concentration properties of sampling θ_t from a multivariate Gaussian, this step is probably unnecessary and just an artifact of the analysis [Agrawal and Goyal, 2017].

Argmax Oracle One advantage of our approach that is most reminiscient of Thompson Sampling is the calculation of $\widehat{z}_t$ at the start of each epoch. In practice, if we have an efficient arg max-oracle, this calculation can be computationally efficient and does not require maintaining $\mathcal{Z}$. By exploiting arg max oracles, we can tractably solve problems like shortest-path and matchings, even in settings where $|\mathcal{Z}|$ is super-exponential in d [Katz-Samuels et al., 2020].

Doubling Trick As presented, the regret guarantees for Lemmas 3.3.1 and 3.3.3 require fixed step sizes η_λ, η_p. To overcome this need for a fixed step size, we use a doubling trick and restart the algorithm every 2^ℓ samples [Shalev-Shwartz et al., 2012]. We believe the use of the doubling trick is purely a theoretical restriction and a more careful analysis could provide an anytime algorithm with no restarts.

3.3.1 Theoretical Guarantees

Recall that at the end of each epoch, $\widehat{z}_\ell(\theta) = \arg\max_{z \in \mathcal{Z}} z^\top \theta$ is the optimal answer for some $\theta \sim \pi_\ell$. Our main result is the following guarantee on Algorithm 6.

Theorem 3.3.4. *With probability* 1*,*

$$\lim_{\ell \to \infty} -\frac{1}{T_\ell} \log \mathbb{P}_{\theta \sim \pi_\ell}(\widehat{z}_\ell(\theta) \neq z_*) = \tau^*,$$

where $\pi_\ell := N(\widehat{\theta}_{T_\ell}, V_{T_\ell - 1}^{-1})$ *restricted to* Θ*.*

Thus our algorithm guarantees that asymptotically the probability that we do not identify the optimal arm decays at the rate of $e^{-T\tau^*}$, with τ^* being the optimal exponent as given in

Theorem 3.2.1. Such guarantees on the probability of a sampled arm are similar to those in the Bayesian best-arm literature, namely Russo [2016] and Jourdan et al. [2022]. In these works, a posterior distribution is maintained and they guarantee that the posterior probability that a non-optimal arm is sampled converges at an exponential rate, with the best possible exponent among all allocation rules. We provide a similar guarantee here for linear bandits. As a remark, this does not directly lead to a bound on the frequentist probability of error, which requires integration of the posterior probability over all randomness in the algorithm. We provide a small sketch of the proof now. A full proof is in Appendix B.3.

Proof sketch. We say that $a_n \doteq b_n$ if $\frac{1}{n}\log(a_n/b_n) \to 0$ as $n \to \infty$. We focus on a fixed round ℓ of Algorithm 6. Using the fact that the expectation of the empirical log-likelihood ratio (conditioned on the data collected) between θ_* and some $\theta \in \Theta$ is the KL divergence between them, we can show using a Laplace Approximation

$$\mathbb{P}_{\theta\sim\pi_\ell}(\widehat{z}_\ell \neq z_*) \doteq \exp\left(-T_\ell \inf_{\theta\in\Theta^c_{z_*}} \frac{1}{2}\left\|\theta-\theta_*\right\|^2_{A(\bar{e}_{T_\ell})}\right).$$

where $\bar{e}_{T_\ell} = \frac{1}{T_\ell}\sum_{t=1}^{T_\ell} e_{x_t}$. Letting $\bar{p}_{T_\ell} = \frac{1}{T_\ell}\sum_{t=1}^{T_\ell} p_t(\Theta^c_{\widehat{z}_t})$, we have

$$\begin{aligned}
&\max_{\lambda\in\triangle_\mathcal{X}} \mathbb{E}_{\theta\sim\bar{p}_{T_\ell}} \left\|\widehat{\theta}_t-\theta\right\|^2_{A(\lambda)} - \min_{p\in\triangle(\Theta^c_{z_*})} \mathbb{E}_{\theta\sim p}\left\|\widehat{\theta}_t-\theta\right\|^2_{A(\bar{e}_{T_\ell})} \\
&= \max_{\lambda\in\triangle_\mathcal{X}} \frac{1}{T_\ell}\sum_{t=1}^{T_\ell} \mathbb{E}_{\theta\sim p_t(\Theta^c_{\widehat{z}_t})}\left\|\widehat{\theta}_t-\theta\right\|^2_{A(\lambda)} \\
&\quad - \frac{1}{T_\ell}\sum_{t=1}^{T_\ell} \mathbb{E}_{\theta\sim p_t(\Theta^c_{\widehat{z}_t})}\left\|\theta-\widehat{\theta}_t\right\|^2_{A(\lambda_t)} && \text{(regret for max learner)} \\
&\quad + \frac{1}{T_\ell}\sum_{t=1}^{T_\ell} \mathbb{E}_{\theta\sim p_t(\Theta^c_{\widehat{z}_t})}\left\|\theta-\widehat{\theta}_t\right\|^2_{A(\lambda_t)} \\
&\quad - \frac{1}{T_\ell}\sum_{t=1}^{T_\ell} \mathbb{E}_{\theta\sim p_t(\Theta^c_{z_*})}\left\|\theta-\widehat{\theta}_t\right\|^2_{x_t x_t^\top} && \text{(error when } \widehat{z}_t \neq z_*) \\
&\quad + \frac{1}{T_\ell}\sum_{t=1}^{T_\ell} \mathbb{E}_{\theta\sim p_t(\Theta^c_{z_*})}\left\|\theta-\widehat{\theta}_t\right\|^2_{x_t x_t^\top} - \frac{1}{T_\ell}\inf_{\theta\in\Theta^c_{z_*}}\left\|\theta-\theta_*\right\|^2_{V_{T_\ell}}. && \text{(regret for the min learner)}
\end{aligned}$$

The regret guarantees in Lemmas 3.3.1 and 3.3.3 ensure the first and third sum are $o(1)$ and so go to 0 as $T_\ell \to \infty$. The fact that $p_t(\Theta^c_{\widehat{z}_t})$ is equal to $p_t(\Theta^c_{z_*})$ for large enough t

ensures that the middle term similarly goes to 0. Combining all terms and the fact that $\widehat{\theta}_t$ is close to θ_* guarantees that for any $\epsilon > 0$ there is a sufficiently large ℓ such that $\max_{\lambda \in \triangle_{\mathcal{X}}} \mathbb{E}_{\theta \sim \bar{p}_{T_\ell}} \|\theta_* - \theta\|^2_{A(\lambda)} - \min_{p \in \triangle(\Theta^c_{z_*})} \mathbb{E}_{\theta \sim p} \|\theta_* - \theta\|^2_{A(\bar{e}_{T_\ell})} \leq \epsilon$, which using minimax duality implies that $\inf_{\theta \in \Theta^c_{z_*}} \frac{1}{2} \|\theta - \theta_*\|^2_{A(\bar{e}_{T_\ell})} \geq \max_{\lambda \in \triangle_{\mathcal{X}}} \min_{p \in \triangle(\Theta^c_{z_*})} \mathbb{E}_{\theta \sim p} \left[\|\theta_* - \theta\|^2_{A(\lambda)} \right] - \epsilon$. Since the first term on the right-hand side is τ^*, we have shown that $\inf_{\theta \in \Theta^c_{z_*}} \frac{1}{2} \|\theta - \theta_*\|^2_{A(\bar{e}_{T_\ell})} \geq \tau^* - \epsilon$. Since by definition $\tau^* \geq \inf_{\theta \in \Theta^c_{z_*}} \frac{1}{2} \|\theta - \theta_*\|^2_{A(\bar{e}_{T_\ell})}$, choosing $\epsilon \to 0$ concludes the proof that $\mathbb{P}_{\theta \sim \pi_\ell}(\widehat{z}_\ell \neq z_*) \doteq \exp\left(-T_\ell \inf_{\theta \in \Theta^c_{z_*}} \frac{1}{2} \|\theta - \theta_*\|^2_{A(\bar{e}_{T_\ell})} \right) = \exp(-T_\ell \tau^*)$. □

Remark: Stopping times. Note that we are not providing a guarantee on the expected stopping time for any finite δ. Existing asymptotically optimal approaches which guarantee a finite stopping time in high probability, e.g. Degenne et al. [2020], utilize a generalized log-likelihood-ratio test of the form

$$\max_{z \in \mathcal{Z}} \min_{\theta \in \Theta^c_{\widehat{z}_t}} \|\theta - \widehat{\theta}_t\|_{V_t} \geq \beta(t, \delta)$$

where $\beta(t, \delta) = O(\sqrt{d \log((T + \|\theta_*\|_2)/\delta)})$ is an anytime confidence bound controlling the deviations of $\|\theta - \widehat{\theta}_t\|_{V_t}$ [Abbasi-Yadkori et al., 2011]. As a result, their algorithms saturate the lower bound for an expected stopping time, i.e. $\limsup_{\delta \to \infty} \mathbb{E}[\tau_\delta] / \log(1/\delta) \leq (\tau^*)^{-1}$. Unfortunately, this GLRT stopping rule itself requires a projection onto each element of $\mathcal{Z}$. We leave it as an open question whether an algorithm can be developed which is asymptotically optimal, requires no explicit projection, and has a finite expected stopping time in high probability.

Remark: Bounded assumptions on Θ. We assume Θ is closed and bounded. The boundedness assumption is needed since we would like to control that for each $\theta \in \Theta$, the rewards $x^\top \theta$ to be bounded for all arms $x \in \mathcal{X}$, which is used in our regret analysis for each learner. Learning algorithms such as AdaHedge [De Rooij et al., 2014] avoid the need for bounded rewards and we leave it as a future research direction to remove this condition.

3.4 Related Work

Pure Exploration Linear Bandits The pure exploration linear bandit problem was introduced in the seminal work of Soare et al. [2014]. In recent years, there has been renewed interest in this problem due to its ability to capture many best-arm-identification and pure exploration settings. Following the experimental design approach first considered by Soare et al. [2014], several different algorithmic frameworks were considered [Tao et al., 2018, Xu et al., 2018, Karnin et al., 2013].

One of the first algorithms to achieve matching instance-optimal upper and lower bounds (within logarithmic factors) for the case of $\mathbb{R}^d$ was by Fiez et al. [2019] and depends on an elimination scheme. Shortly after, several works proposed asymptotically optimal algorithms. The first of these methods utilized the track and stop approach given in Jedra and Proutiere [2020], which fully solves the τ^* objective of Equation 3.1 using a plug-in estimator $\widehat{\theta}_t$ at each round. Due to the computational difficulty of this, several works proposed alternatives that iteratively updated the sampling distribution in each round. This includes the game theoretic viewpoint we utilize first proposed by Degenne et al. [2020, 2019], and a novel modification of Frank-Wolfe by Wang et al. [2021]. Other works have augmented these approaches by providing elimination schemes to reduce the set of alternative $\mathcal{Z}$ that need to be considered each round. Zaki et al. [2022] proposes a hybrid approach combining the elimination from Fiez et al. [2019] and Degenne et al. [2020] to remove the condition that Θ needs to be bounded. Tirinzoni and Degenne [2022] provide an elimination approach where they carefully exploit properties of $\mathcal{Z}$. Finally, we mention that the pure exploration problem has also been considered in the generalized linear bandit (logistic) settings in Kazerouni and Wein [2021] and Jun et al. [2021b]. Future work could explore extending sampling methods to these settings.

Oracle Based Approaches As discussed before, if $\mathcal{Z}$ is a large or combinatorial set, it may be impossible to maintain and appropriate oracles are needed. Katz-Samuels et al. [2020]

considers the linear combinatorial setting for matroid-like classes e.g. shortest-path, top-k, and bipartite matching. By exploiting ideas similar to Fiez et al. [2019], they provide an algorithm utilizing the argmax oracle to achieve near optimal sample complexity. A recent work by Li et al. [2022] reduces optimal policy learning in agnostic contextual bandits to pure exploration and provides a method analogous to Agarwal et al. [2014] which only relies on cost-sensitive classification.

Top Two Methods Our approach is perhaps most reminiscent of the Top-Two Thompson Sampling (TTTS) algorithm for best-arm identification in multi-armed bandits[1] of Russo [2016]. Similar to Thompson sampling Russo et al. [2018], TTTS maintains a posterior distribution over the means of the arms, and at each round samples a mean vector from the distribution and chooses the arm with the highest sampled mean. It then continues to sample mean vectors, until one is returned whose highest mean is different from the previous found one. Both arms are then pulled. As discussed in the introduction, our algorithm is similar in spirit - we sample until finding a parameter vector whose best-arm is different from our current estimate and then we utilize these vectors to update our learners. Top-two algorithms for multi-armed bandits perform well in practice and have been extensively studied in Bayesian and frequentist settings under various assumptions on noise [Qin et al., 2017, Shang et al., 2020, Jourdan et al., 2022, Qin and Russo, 2022, Lee et al., 2023]. However, they often depend on a parameter β, and only achieve a weaker notion of β-optimality. Our work is the first to propose and analyze an asymptotically optimal Top-two algorithm for the general linear bandit setting. We remark that the LinGapE algorithm [Xu et al., 2018] also uses a top-two approach and tends to perform well empricially, however it is unknown whether it is asymptotically optimal.

Online Learning and Thompson Sampling Finally we remark that the connection between Thompson Sampling and online learning has been previously explored in the early

[1] i.e. the arms are standard basis vectors $\mathcal{X} = \mathcal{Z} = \{e_1, \cdots, e_d\} \in \mathbb{R}^d$ and $\Theta = [0, 1]^d$

work of Li [2013]. This work focuses on the regret setting. Other works in the regret setting have explored connections between information-theoretic analysis of Thompson sampling and online stochastic mirror descent algorithms [Lattimore and Gyorgy, 2021, Zimmert and Lattimore, 2019]. We hope that our work provides a strong step in this direction for the structured pure exploration literature.

3.5 Experiments

In the following, we provide some preliminary experiments to demonstrate the performance of Algorithm 5. Note that the contribution of this paper is primarily theoretical - our goal is to demonstrate that asymptotically optimal algorithms for pure exploration can rely purely on sampling oracles. We hope that the preliminary experiments we provide encourage further exploration of this line of thinking and lead to algorithms that can be as easy to apply as Thompson sampling in practice.

With this in mind, we ran the following modification of some of the algorithms of the previous section. Firstly, we eschewed the doubling trick and instead just ran `PEPS` directly for a fixed horizon side T. Secondly, for the max-learner we made use of AdaHedge which is able to use an adaptive step size. Finally, we set $\eta_p = 1$. Though our algorithm only has theoretical guarantees over a bounded set Θ, we believe that this is primarily a limitation of our analysis and so we set $\Theta = \mathbb{R}^d$. We also remove the forced G-optimal exploration for the same reason. For the sampling oracle, we use rejection sampling method because of its simplicity. We demonstrate empirically that the computation cost is not onerous. We plot the number of rejection steps used each round along with clock time per iteration for our method in Appendix B.6. We also see that our method is running faster than the benchmark LinGame especially when the number of arms is large in Table B.2 in Appendix B.6. Further details on our experimental setup and additional evaluations are also in Appendix B.6.

The main algorithms we compare to are Thompson Sampling [Russo et al., 2018], LinGame [Degenne et al., 2020], and LinGapE Xu et al. [2018]. LinGame is based on the two-player game strategy with best-response detailed in Section 3.2. For a fair comparison, we run

	Soare's instance [Soare et al., 2014]			Sphere			TopK		
δ	0.1	0.05	0.01	0.1	0.05	0.01	0.2	0.1	0.05
PEPS	1027	1606	3284	294	476	794	7326	14188	22518
LinGame	828	1500	2688	186	282	638	8838	29963	>30000
LinGapE	708	1141	2281	316	433	690	7096	20570	>30000
Oracle	766	1232	2576	243	328	473	17363	>30000	>30000
TS	>5000	>5000	>5000	431	1046	2176	N/A	N/A	N/A

Table 3.1: The number of samples needed for $\mathbb{P}_{\theta\sim\pi_\ell}(\widehat{z}_\ell = z_*) > 1-\delta$ for various algorithms

LinGame and LinGapE without stopping. The goal of our experiments was to demonstrate that sampling and no-projection algorithms can be competitive against algorithms that explicitly project. From this perspective, we did not consider algorithms that eliminate. For a more extensive empirical comparison of existing algorithms, please see Tirinzoni and Degenne [2022]. We also include an oracle strategy that pulls arms from the allocation derived from the lower bound.

In summary, our algorithm achieves a similar performance compared to LinGame and LinGapE while beating LinTS in Soare and Sphere instances. For Top-k instance, our algorithm beats LinGame, LinTS, and LinGapE. Note that our algorithm is the first algorithm that relies purely on just sampling oracles and our theoretical analysis is only asymptotic, the experimental results are satisfactory since they show that our algorithm works decently well in practice. Now we detail the setting for each instance.

Soare's Instance [Soare et al., 2014]. The first instance we consider is the standard benchmark linear bandit instance described in Soare et al. [2014]. In this instance, the arm set $\mathcal{X} \subset \mathbb{R}^2$ with $|\mathcal{X}| = 3$. The first two arms are $x_1 = e_1, x_2 = e_2 \subset \mathbb{R}^2$, the canonical basis vectors, and an informative arm $x_3 = (\cos(\omega), \sin(\omega))$. The true parameter is $\theta_* = (1, 0) \in \mathbb{R}^d$.

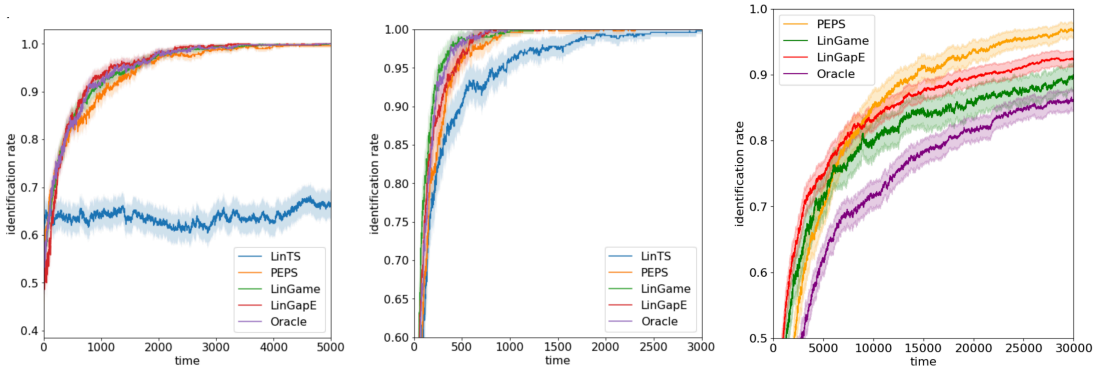

Figure 3.1: Best-arm identification rate for PEPS, LinGame [Degenne et al., 2020], LinGapE [Xu et al., 2018], Thompson sampling, and fixed weight strategy under three instances: Soare instance with $\omega = 0.1$, sphere instance with $d = 6$ and $|\mathcal{X}| = 20$, and Top-k instance with $d = 12$ and $k = 3$, with 500 repetitions for each instance. Confidence intervals with plus or minus two standard errors are shown.

In this problem, the optimal arm is always x_1. However, when the angle ω is small, it becomes challenging to distinguish the interfering arm x_{d+1} from x_1. An effective sampling strategy would pull arm x_2 instead of x_1 to reduce uncertainty between x_1 and x_{d+1} effectively. However, Thompson sampling will tend to pull x_1, which will take much longer to distinguish between the two competing arms. The experiments were carried out on a problem instance with $d = 2$ and $\omega = 0.1$. Our algorithm achieves a similar performance compared with LinGame and LinGapE while beats LinTS.

Sphere. Following Tao et al. [2018] and Degenne et al. [2020], we also consider a linear bandit instance where the arm set $\mathcal{X} \subset B^d := \{x \in \mathbb{R}^d : \|x\|_2 = 1\}$ is randomly drawn from a unit sphere of dimension d. For the true parameter, we select the two arms, x and x', that are closest to each other, and define $\theta_* = x + 0.01(x' - x)$, ensuring that x is the best arm. In our experiment, we run the three algorithms on a problem instance with $d = 6$ and $|\mathcal{X}| = 20$. As we can see, our algorithm still outperforms Thompson sampling and is competitive with

LinGame and LinGapE.

Top-k. The third instance we consider is the top-k combinatorial bandit problem where the goal is to identify the top-k means. In the linear setting, this can be expressed as $\mathcal{X} = \{e_1, \cdots, e_d\} \subset \mathbb{R}^d$ and $\mathcal{Z} = \{e_{i_1} + \cdots + e_{i_k} : i_1, \cdots, i_k \in \binom{[d]}{k}\} \subset \mathbb{R}^d$, i.e. $\mathcal{X}$ is the standard basis and $\mathcal{Z}$ is the set of indicator vectors of subsets of size k. Then, the best arm in this new arm set $\mathcal{Z}$ corresponds to the top-k arms in $\mathcal{X}$, which is the goal of top-k identification. Then we run BAI algorithms on this new arm set. We take $\theta = [1, .95, .90, \cdots, 1 - .05i, \cdots] \in \mathbb{R}^d$. As we can see, our algorithm outperforms LinGame and LinGapE in this instance.

We also present Table 3.1 describing the number of samples needed to reach a $1 - \delta$ idenfication rate for various δ values. Note that we do not run Thompson sampling for the Top-k instance (it is not defined when $\mathcal{X} \neq \mathcal{Z}$ so we put N/A there), and $> n$ in the table means that the algorithm fails to achieve $1 - \delta$ for the n iterations we run in the experiment. We can see that our algorithm, PEPS, achieves an $1 - \delta$ best-arm identification probability for all δ in all instances, with a rate similar to LinGame, outperforming LinTS in all three instances.

3.6 Conclusion

In this paper, we present the first sampling-based projection-free algorithm for pure exploration in linear bandits. Our algorithm only relies on a sampling oracle and an argmax oracle, so our algorithm is tractable in various settings. We show that our algorithm is asymptotically optimal in the sense that the probability that we do not identify the optimal arm decays exponentially with the optimal rate for a fixed allocation. We provide experiments demonstrating that our algorithm beats Thompson sampling and has competitive performance against benchmark algorithms such as LinGame [Degenne et al., 2020] in various problem instances. Our current approach has various limitations: for example, we need to assume that Θ is bounded. However, we hope that this work opens a line of investigation into better sampling-based algorithms for effective exploration.

Chapter 4

ESTIMATION OF SUBSIDIARY PERFORMANCE METRICS UNDER OPTIMAL POLICIES

4.1 Introduction

4.1.1 Literature Review

Many fields are interested in learning policies that map from individual-level characteristics to a choice of action. The policies that result in the best possible mean of a subsequent outcome are often referred to as optimal policies Athey and Wager [2021]. For example, in biomedical sciences the action may take the form of a treatment allocation and the outcome may be disease remission Ling et al. [2021], whereas in digital marketing the action and outcome may be a recommendation and click-through rate, respectively Hill et al. [2017]. There have been a variety of methods developed for estimating optimal policies. These methods include regression-based estimators such as Q-learning Qian and Murphy [2011], outcome-weighted learning Zhao et al. [2012], and doubly robust approaches [Dudík et al., 2011, Zhang et al., 2013], among others. Performance guarantees for these methods have been established by several authors Athey and Wager [2021], Qian and Murphy [2011], Zhao et al. [2012], Luedtke et al. [2020].

An estimated policy is unlikely to be implemented unless confidence intervals characterizing its performance are available Shi et al. [2020]. These performance metrics may take the form of the remission rate of all patients or the click-through rate of all customers. In both of these examples, the metric is the value of the optimal policy in the population, better known as the optimal value. Inference about the optimal value is well-studied when there is only one outcome of interest Luedtke et al. [2020], Liu et al. [2021]. Several works have shown that one-step estimators and targeted minimum loss based estimators are efficient

under conditions van der Laan and Luedtke [2015], Chambaz et al. [2017]. In particular, these works require a non-exceptional law condition that states that the conditional average action effect does not concentrate mass at zero Robins [2004]. Alternative strategies have been developed for constructing confidence intervals for the optimal value even when this condition fails Luedtke and Van Der Laan [2016], Chakraborty et al. [2013].

Though most existing methodological works on policy learning focus on optimizing for a single performance metric, in real-world settings there are often multiple other subsidiary performance metrics that are also of interest Boominathan et al. [2020], Bica et al. [2021]. These metrics may correspond to different summaries of the outcome, such as the median, rather than the mean, and time to disease remission Phillips et al. [2020]. Alternatively, they may summarize several different outcomes rather than just a single one. For example, when learning a treatment allocation, symptom reduction may be considered alongside prognosis Freemantle et al. [2003]. Most existing approaches for incorporating multiple outcomes involve combining them into a composite outcome and then using policy learning methods designed for single-outcome settings Butler et al. [2018]. In settings where the actions recommended by experts are recorded in the dataset, Murray et al. provided a means to construct a composite outcome in an automated fashion Murray et al. [2016]. However, when expert recommendations are not available, composite-outcome-based approaches require investigators to construct the composite outcome in some other way, which often ends up being somewhat arbitrary Luckett et al. [2021]. Some alternative approaches do not require the construction of a composite outcome. One such approach involves learning a policy that returns a set of recommended actions, rather than a single one Laber et al. [2014]. Each of the actions in this set should yield a desirable result for at least some of the outcomes. For settings where there is a primary outcome of interest, another approach involves using other secondary outcomes to define constraints that any selected policy must satisfy Linn et al. [2015], Wang et al. [2018].

In cases where a single outcome is of primary interest and others are only of secondary interest, a preferred approach may be to optimize only for this one outcome, while still making

inferences about the effect of the policy on the subsidiary outcomes. For example, just as the side effects of any new medical intervention must be assessed along with its effect on the primary outcome of interest [FDA, 2006], the side effects of a new treatment policy should be assessed as well [Linn et al., 2015]. As another example, if a company optimizes a policy for customer acquisition, it must also consider the impact the policy will have on customer retention Afeche et al. [2017]. In this work, we provide a systematic approach for assessing the impact of a policy that is optimized for some primary outcome on other, subsidiary outcomes.

4.1.2 Notation and objectives

Let $X \in \mathcal{X}$ be a feature, $A \in \{0, 1\}$ a binary action, and $Y \in \mathcal{Y}$ an outcome that is observed after the action. This outcome may be multivariate. Let $\mathcal{M}$ be a nonparametric model consisting of possible joint distributions P of (X, A, Y). Our sample consists of n independent and identically distributed draws $(X_i, A_i, Y_i)_{i=1}^n$ from $P_0 \in \mathcal{M}$. Let Π be a set of policies $\mathcal{X} \to \{0, 1\}$ that take as input a feature and take action 0 or 1. For a given policy $\pi \in \Pi$, let $\Omega_\pi(P)$ be a real-valued primary performance metric for the policy π under sampling from P, where we assume that larger values of this metric are considered preferable. Further let $\Psi_\pi(P)$ be a real-valued subsidiary performance metric for π under P. For example, when Y is a primary-subsidiary outcome pair $(Y^\star, Y^\dagger) \in \mathbb{R}^2$, these metrics could be the covariate-adjusted means of these two outcomes Luedtke et al. [2020], Luedtke and Van Der Laan [2016], that is, $\Omega_\pi(P) = \int \mathbb{E}_P[Y^\star | A = \pi(x), X = x] dP(x)$, and $\Psi_\pi(P) = \int \mathbb{E}_P[Y^\dagger | A = \pi(x), X = x] dP(x)$. Alternatively, the outcome Y may be univariate and the primary performance metric may be equal to the mean $\Omega_\pi(P) = \int \mathbb{E}_P[Y | A = \pi(x), X = x] dP(x)$ while the subsidiary metric may be equal to the covariate-adjusted probability that the outcome exceeds a specified value t, namely $\Psi_\pi(P) = \int P\{Y > t | A = \pi(x), X = x\} dP(x)$. We refer to $\Omega_\pi(P_0)$ and $\Psi_\pi(P_0)$ as the Ω-performance and Ψ-performance of the policy π.

For $P \in \mathcal{P}$, let Π_P^* denote the set of optimal policy with respect to the primary performance metric, that is, $\Pi_P^* := \{\pi \in \Pi : \Omega_\pi(P) = \sup_{\pi' \in \Pi} \Omega_{\pi'}(P)\}$. We denote a generic element of this set by π_P^*. We refer to elements of $\Pi^* := \Pi_{P_0}^*$ as Ω-optimal policies and denote a generic

element by π^*. We assume throughout that Π^* is nonempty, and note that in general this set may contain more than one policy. We are interested in making inferences about the subsidiary performance metric $\Psi_\pi(P_0)$ for Ω-optimal policies. Letting $\psi_0^\ell = \inf_{\pi\in\Pi^*} \Psi_\pi(P_0)$ and $\psi_0^u = \sup_{\pi\in\Pi^*} \Psi_\pi(P_0)$, our objective is to construct a confidence interval for the range of possible Ψ-performances under an Ω-optimal policy, that is, develop a confidence interval that is a superset of $[\psi_0^\ell, \psi_0^u]$ with a specified asymptotic probability.

When there is only one Ω-optimal policy, our objective is to determine the Ψ-performance of this policy, denoted by $\psi_0 := \psi_0^\ell = \psi_0^u$. When there are multiple Ω-optimal policies, ψ_0^ℓ may be less than ψ_0^u, and the upper and lower bounds of our interval inform on the most extreme Ψ-performances that can be attained from an Ω-optimal policy. For example, if larger values of Ψ are preferable, then the upper confidence bound on ψ_0^u informs about the best achievable Ψ-performance by an Ω-optimal policy. Such a policy can be shown to be one of several policies that fall along the Pareto front of the two-objective optimization problem that seeks to maximize Ω and Ψ. The Pareto front denotes the set of policies for which there is not a policy that performs better with respect to one of the two metrics and no worse with respect to the other. The difference between inferring about ψ_0^u and multi-objective optimization is that the policy with the best Ψ performance is primarily optimized with respect to one performance metric Ω, while multi-objective optimization optimizes several performance metrics simultaneously Gunantara [2018], Deb [2014], Bentley and Wakefield [1998].

We present our proposed confidence intervals in the next two sections. When doing so, we consider two separate cases. In Section 4.2, we begin with an easier and more specialized case, where the performance metrics Ω_π and Ψ_π are assumed to be the covariate-adjusted means of a primary outcome ($Y^\star$) and subsidiary outcome ($Y^\dagger$), there is assumed to be a unique Ω-optimal policy π^* over an unrestricted policy class Π, and a certain margin condition holds. In Section 4.3, we move on to a harder and more general case, where Ω_π and Ψ_π are arbitrary smooth parameters and there may be multiple Ω-optimal policies.

4.2 Wald-type inference under a margin assumption

In this section, we focus on the case where $\Omega_\pi(P) = \int \mathbb{E}_P[Y^\star|A = \pi(x), X = x]dP(x)$ and $\Psi_\pi(P) = \int \mathbb{E}_P[Y^\dagger|A = \pi(x), X = x]dP(x)$ for a primary-subsidiary outcome pair $(Y^\star, Y^\dagger)$ and, moreover, the policy class Π is unrestricted. We aim to build on existing works that evaluate the Ω-performance of an Ω-optimal policy [van der Laan and Luedtke, 2015, Luedtke and Van Der Laan, 2016]. These works have shown that a simple estimation strategy is efficient under a non-exceptional law condition that makes the Ω-optimal rule unique [Robins, 2004]. In this case, $\psi_0^\ell = \psi_0^u$ and we write $\psi_0 = \psi_0^\ell = \psi_0^u$. This strategy first obtains an estimate $\widehat{\pi}$ of the Ω-optimal rule, and then constructs a standard one-step estimator of $\Omega_{\widehat{\pi}}(P_0)$. Heuristically speaking, pursuing estimation of $\Omega_{\widehat{\pi}}(P_0)$, rather than $\Omega_{\pi^*}(P_0)$, introduces only negligible bias because $\widehat{\pi}$ should be a near-maximizer of $\Omega_\pi(P_0)$. Hence, similarly to the fact that $f(x) - f(x^*) = o(|x^* - x|)$ for a differentiable function $f : \mathbb{R} \to \mathbb{R}$ with maximizer x^*, the error induced by replacing π^* by $\widehat{\pi}$ in the functional $\pi \mapsto \Omega_\pi(P_0)$ should be second-order. In this section, we study the extent to which a standard one-step estimator of $\Psi_{\widehat{\pi}}(P_0)$ will yield an asymptotically normal and efficient estimator of $\Psi(P_0)$. This study is important since, if the standard one-step estimator satisfies these properties under only mild conditions, then there is little reason to develop alternative methods.

We now discuss a key condition that we will require to establish the efficiency of a standard one-step estimator for $\Psi(P_0)$, along with the validity of corresponding Wald-type confidence intervals. Define the function $q_b(P)(x) := \mathbb{E}_P[Y^\star|A = 1, X = x] - \mathbb{E}_P[Y^\star|A = 0, X = x]$ to be the conditional average treatment effect on the primary outcome and $s_b(P)(x) := \mathbb{E}_P[Y^\dagger|A = 1, X = x] - \mathbb{E}_P[Y^\dagger|A = 0, X = x]$ to be the conditional average treatment effect on the subsidiary outcome. We refer to these functions as the primary CATE and subsidiary CATE, respectively. We use the shorthand notation $q_{b,0} := q_b(P_0)$ and $s_{b,0} := s_b(P_0)$.

Condition 1 (Margin condition between $Y^\dagger$ and $Y^\star$)**.** For some $C_1 > 0$ and $\zeta > 2$,

$$P_0\left(|s_{b,0}(X)| \geq C_1 t\,|q_{b,0}(X)|\right) \leq t^{-\zeta}, \qquad \text{for all } t > 1. \tag{4.1}$$

When this condition holds, $|q_{b,0}(X)| \neq 0$ with P_0-probability one. Hence, this condition is a strengthening of the usual non-exceptional law condition [Robins, 2004] that is required when the Ψ and Ω performance metrics coincide. To ensure the validity of the standard one-step estimator, some form of strengthening appears to be needed to make up for the fact that π^* is defined as a maximizer in π of $\Omega_\pi(P_0)$, rather than $\Psi_\pi(P_0)$. Indeed, the estimation error of this estimator $\widehat{\psi}_{\widehat{\pi}}$ can be decomposed as

$$\widehat{\psi}_{\widehat{\pi}} - \Psi_{\pi^*}(P_0) = \left[\widehat{\psi}_{\widehat{\pi}} - \Psi_{\widehat{\pi}}(P_0)\right] + \left[\Psi_{\widehat{\pi}}(P_0) - \Psi_{\pi^*}(P_0)\right].$$

The fact that $\widehat{\psi}_{\widehat{\pi}}$ is a one-step estimator of $\Psi_{\widehat{\pi}}(P_0)$ should imply that the first term will be small. However, since π^* is not necessarily an optimizer for Ψ, it is possible that $\Omega_{\widehat{\pi}}(P_0)$ is close to $\Omega_{\pi^*}(P_0)$ while $\Psi_{\widehat{\pi}}(P_0)$ is far from $\Psi_{\pi^*}(P_0)$ — see Figure 4.1 for an illustration of this possibility. Therefore, we need a condition to characterize the flatness of the Ψ performance

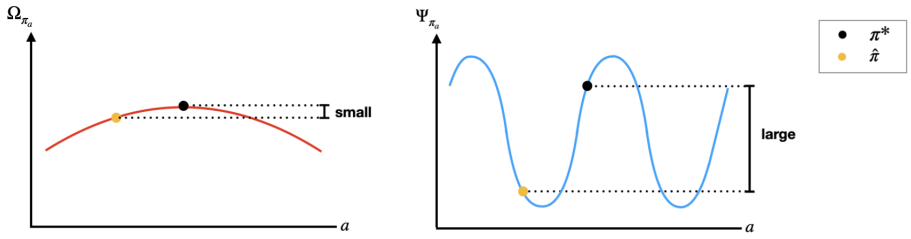

Figure 4.1: Plot of primary and subsidiary performance metrics for an estimated policy given the threshold policy class $\Pi = \{\mathbf{1}(x \geq a) : a \in \mathbb{R}\}$. The estimator $\widehat{\pi}$ performs well in the sense that the Ω-regret $\Omega_{\pi^*}(P_0) - \Omega_{\widehat{\pi}}(P_0)$ is small, which is to be expected since π^* is defined to be an Ω-optimal rule. Nevertheless, in principle the Ψ-regret $\Psi_{\pi^*}(P_0) - \Psi_{\widehat{\pi}}(P_0)$ could still be large, since the Ψ-value function $\pi \mapsto \Psi_\pi(P_0)$ may be markedly different from the Ω-value function. Though a similar phenomenon can occur for unrestricted policy classes, which are our focus in this section, the infinite-dimensional nature of these classes precludes their visualization.

surface relative to that of Ω. This flatness can be characterized by studying the absolute

CATE ratio $|q_{b,0}(X)| \,/\, |s_{b,0}(X)|$, where we use the convention that $b/0 = +\infty$ for $b > 0$ and we recall that $|q_{b,0}(X)| = 0$ with probability zero under (4.1). Condition 1 imposes that the absolute CATE ratio can only concentrate vanishingly little mass near zero when $X \sim P_0$. This certainly holds in the extreme case where, within each level x of the covariates, the magnitude of the expected effect of the action on the primary outcome, namely $|q_{b,0}(x)|$, is at least as large as the magnitude of its effect on the subsidiary outcome, namely $|s_{b,0}(x)|$. It also allows for scenarios where the magnitude $|s_{b,0}(x)|$ is much larger than $|q_{b,0}(x)|$ for certain features x with sufficiently small probability of occurrence. However, it can fail to hold when there are some feature levels where the action has no effect on the primary outcome and yet does have one on the subsidiary outcomes; this can occur, for example, if the primary outcome is cancer remission and the subsidiary outcome captures side effects induced by chemotherapy. Though Condition 1 may be strong, we were unable to show the validity of the standard one-step estimator without it. Therefore, in the remainder of this section we assume that this condition holds, and we refer the reader to the next section for a method that is valid even when it does not.

In the special case where $Y^\dagger = Y^\star$ a.s., the asymptotic normality and efficiency of the one-step estimator have previously been justified by establishing the pathwise differentiability of the Ω-performance of an Ω-optimal policy van der Laan and Luedtke [2015]. We follow a similar approach here when considering cases where $Y^\dagger$ and $Y^\star$ may differ. In particular, we establish the pathwise differentiability of $\Psi^* : P \mapsto \sup_{\pi \in \Pi^*_P} \Psi_\pi(P)$ in what follows. When doing this, we will need to impose Condition 1, along with an additional margin condition that is inspired by ones previously assumed in the policy learning [Qian and Murphy, 2011, Luedtke and Van Der Laan, 2016] and classification [Audibert and Tsybakov, 2007] literatures.

Condition 2 (Margin condition for $Y^\star$)**.** For some $\gamma > \frac{1}{\zeta}$,

$$P_0\left(0 < |q_{b,0}(X)| \leq t\right) \lesssim t^\gamma \qquad \forall t > 0. \tag{4.2}$$

This condition imposes that the unique Ω-optimal policy can be estimated well via a plug-in estimator [Qian and Murphy, 2011, Luedtke and Van Der Laan, 2016]. For some

generic $P \in \mathcal{P}$ and $\pi \in \Pi$, define $p_P(a|x) := P(A = a|X = x)$ and $D(\pi, P)(x, a, y^\dagger) = \frac{\mathbb{I}\{a=\pi(x)\}}{p_P(a|x)}[y^\dagger - s(a, x)] + s(\pi(x), x) - \Psi_\pi(P)$. We will use the shorthand $p_0 := p_{P_0}$ and $p_n := p_{\widehat{P}_n}$. The following result characterizes the pathwise differentiability of $\Psi^*(\cdot)$ at P_0.

Lemma 4.2.1. *Suppose that Ψ_π and Ω_π are covariate-adjusted means for each $\pi \in \Pi$, the policy class Π is unrestricted, and conditions 1 and 2 are satisfied. Then, Ψ^* is pathwise differentiable at P_0 relative to a nonparametric model with canonical gradient $D(\pi^*, P_0)$.*

We use the above result to argue that a one-step corrected estimator is efficient provided its influence function is equal to $D(\Pi^*, P_0)$. Consider some estimate $\widehat{P}_n$ of the true distribution P_0. The one-step corrected estimator takes the form $\psi_{OS,n} := \Psi_{\widehat{\pi}}(\widehat{P}_n) + P_n D(\widehat{\pi}, \widehat{P}_n)$. For simplicity, when studying this estimator, we focus on the case where $\widehat{\pi}$ is a plug-in estimator of the Ω-optimal policy, namely $\pi^*_{\widehat{P}_n}$. In principle, the policy estimator could be constructed using some other approach, such as outcome weighted learning [Zhao et al., 2012]. Let $q_{b,n}(x)$ and $s_{b,n}(x)$ be some estimates for the conditional average treatment effects $q_{b,0}(x)$ and $s_{b,0}(x)$ respectively. Also, let $s_n(a, x)$ be some estimate for $s_0(a, x)$. Define the $L_r(P)$ norm of a generic function $f : \mathcal{D} \to \mathbb{R}$ as $\|f\|_{r,P} := [\int_{\mathcal{D}} |f(t)|^r dP(t)]^{1/r}$. We first present some consistency conditions on these estimates.

Condition 3 (Consistent estimator of conditional average treatment effect on the primary outcome)**.** $\|q_{b,n} - q_{b,0}\|^{1+\gamma/2}_{\infty,P_0} = o_{P_0}(n^{-1/2})$.

Condition 4 (Consistent estimator of conditional average treatment effect on the subsidiary outcome)**.** $\max_{a\in\{0,1\}} \left\{ \left\| \frac{p_0(a|\cdot)}{p_n(a|\cdot)} - 1 \right\|_{2,P_0} \left\| s_{\widehat{P}_n}(a, \cdot) - s_{P_0}(a, \cdot) \right\|_{2,P_0} \right\} = o_{P_0}(n^{-1/2})$, where $s_P(a, x) := E_P[Y^\dagger \mid A = a, X = x]$.

Condition 4 is similar to Equation 15 in Luedtke and Van Der Laan [2016]. Discussion of this condition can be found in Luedtke and Van Der Laan [2016]. The following theorem states that the one-step estimator is efficient.

Theorem 4.2.2. *Under Conditions 1, 2, 3, 4, and also provided $D(\widehat{\pi}, \widehat{P}_n)$ falls in a P_0-Donsker class with probability tending to 1 and $\|D(\widehat{\pi}, \widehat{P}_n) - D(\pi^*, P_0)\|_{2,P_0} \xrightarrow{p} 0$, the one-step*

estimator $\psi_{OS,n}$ *for* $\widehat{\pi} = \pi^*_{\widehat{P}_n}$ *is an asymptotically linear estimator of* $\Psi^*(P_0)$ *with influence function* $D(\pi^*, P_0)$*, in the sense that*

$$\psi_{OS,n} - \Psi^*(P_0) = \frac{1}{n}\sum_{i=1}^{n} D(\pi^*, P_0)(X_i, A_i, Y_i^\dagger) + o_{P_0}(n^{-1/2}).$$

Moreover, $\psi_{OS,n}$ *is an asymptotically efficient estimator of* ψ_0*.*

The above can be used to construct Wald-type confidence intervals for ψ_0 of the form $\psi_{OS,n} \pm z_{1-\alpha/2}\sigma_n/\sqrt{n}$, where $z_{1-\alpha/2}$ is the $1-\alpha/2$ quantile of a standard normal random variable and $\sigma_n^2 := \frac{1}{n}\sum_{i=1}^{n} D(\widehat{\pi}, \widehat{P}_n)(X_i, A_i, Y_i^\dagger)^2$.

The Donsker condition stated in the above theorem can be removed if cross-fitting is used [Schick, 1986]. A 2-fold version of this approach first partitions the data in two halves. Then, it uses the first half of the data to learn $\widehat{\pi}$ and uses the remaining data to construct an estimator for $\Psi_{\widehat{\pi}}(P_0)$. The roles of the two halves are then swapped and the two estimators are subsequently averaged. Multi-fold versions of cross-fitting could also be used.

4.3 Inference of a general functional without margin assumption

4.3.1 Overview of the methods

The methods we present in this section are agnostic to whether Condition 1 holds and, more generally, whether there are multiple Ω-optimal policies. Because the parameter Ψ^* considered in the previous section may not even be well-defined when there are multiple such policies, we instead focus on inferring about the range $[\psi_0^l, \psi_0^u]$ of possible Ψ-performances of Ω-optimal policies. Unlike those in the previous section, the methods developed here critically rely on the policy class Π being restricted — in particular, being P_0-Donsker [Van Der Vaart and Wellner, 2013] — and this condition cannot be removed even if cross-fitting is employed (see Section 4.3.2 for a discussion). Also, in this section, we do not assume our performance criteria are covariate-adjusted means. Rather, they could take some other form, such as that of a covariate-adjusted median. In what follows we give an overview of our approach for inferring about $[\psi_0^l, \psi_0^u]$.

Our proposed method consists of two stages. The first spends $\beta < \alpha$ error probability to construct a confidence set $\widehat{\Pi}_\beta$ that contains the set of optimal policies Π^* with probability tending to at least $1 - \beta$. The second infers about the Ψ-performance of each remaining policy in this confidence set, returning a confidence interval for $[\psi_0^\ell, \psi_0^u]$ of the form

$$\left[\inf_{\pi \in \widehat{\Pi}_\beta} \left\{\widehat{\psi}_\pi - \frac{\widehat{\kappa}_\pi z_{\alpha,\beta}}{n^{1/2}}\right\}, \sup_{\pi \in \widehat{\Pi}_\beta} \left\{\widehat{\psi}_\pi + \frac{\widehat{\kappa}_\pi z_{\alpha,\beta}}{n^{1/2}}\right\}\right], \tag{4.3}$$

where $\widehat{\psi}_\pi$ is some estimate for $\Psi_\pi(P_0)$, $z_{\alpha,\beta}$ corresponds to $1 - (\alpha - \beta)/2$ quantile of normal distribution and $\widehat{\kappa}_\pi^2$ is an estimate of the asymptotic efficiency bound for estimating $\Psi_\pi(P_0)$. We provide a union bounding argument that shows that, under conditions, this confidence interval will cover $[\psi_0^\ell, \psi_0^u]$ with asymptotic probability $1 - \alpha$.

The first-stage confidence set $\widehat{\Pi}_\beta$ is constructed so that policies that perform poorly in terms of the primary performance metric are eliminated. In words, we maintain policies π whose uniform upper confidence bound for $\Psi_\pi(P_0)$ is greater than the largest non-uniform lower confidence bound across all policies in the set. Figure 4.2 shows an example of how the first-stage elimination is performed. More specifically, we define this set $\widehat{\Pi}_\beta$ after the first-stage filtration as

$$\widehat{\Pi}_\beta := \left\{\pi \in \Pi : L_n \leq \widehat{\omega}_\pi + \frac{\widehat{\sigma}_\pi t_\beta}{n^{1/2}}\right\}, \tag{4.4}$$

where $\widehat{\omega}_\pi$ is some estimate for $\Omega_\pi(P_0)$, $\widehat{\sigma}_\pi^2$ is an estimator of the asymptotic efficiency bound for estimating $\Omega_\pi(P_0)$, L_n is an asymptotically valid $1 - \beta/2$ lower bound for $\sup_{\pi \in \Pi} \Omega_\pi(P_0)$ (e.g., obtained via [Luedtke and Van Der Laan, 2016]), and t_β is selected in such a way that $\{\widehat{\omega}_\pi + \widehat{\sigma}_\pi t_\beta / n^{1/2} : \pi \in \Pi\}$ is an asymptotically valid $1 - \beta/2$ uniform upper confidence bound for $\{\Omega_\pi(P_0) : \pi \in \Pi\}$, in the sense that $\Omega_\pi(P_0) \leq \widehat{\omega}_\pi + \widehat{\sigma}_\pi t_\beta / n^{1/2}$ for all $\pi \in \Pi$ with probability tending to at least $1 - \beta/2$ as n goes to infinity.

It may at first be surprising that, in constructing the confidence interval for $[\psi_0^\ell, \psi_0^u]$, the only place a uniform confidence bound is used is in the upper bound of (4.4). Indeed, when we began studying this problem, the first approach that we considered was the same as that previously described, except with all confidence bounds replaced by uniform ones.

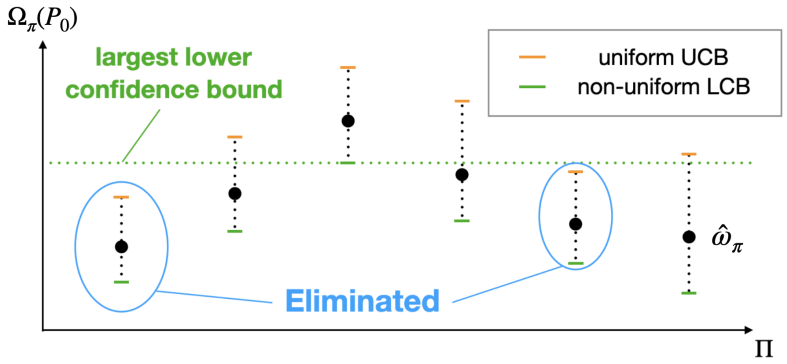

Figure 4.2: Example of first-stage elimination. Each black dot represents an estimate of $\Omega_\pi(P_0)$ and the horizontal bars denote the confidence bounds. Policies whose uniform upper confidence bound (UCB) is below the largest lower confidence bound (LCB) get eliminated.

In particular, L_n was defined as the maximum over $\pi \in \Pi$ of a uniform lower confidence bound for the Ω-value function and the minimal and maximal marginal confidence bounds in (4.3) were also replaced by minimal and maximal uniform confidence bounds. However, after analyzing this method, we discovered that less uniformity was needed than we initially expected. Indeed, the uniformity in defining L_n can be dropped since a simple union bounding argument shows that L_n only needs to satisfy that it falls below the optimal Ω-value with asymptotic probability at least $1 - \beta/2$; while selecting the maximum of a uniform confidence band for the value function does satisfy such a property, developing such a lower bound is now a well-studied problem, and so less conservative approaches have been developed [Luedtke and Van Der Laan, 2016]. The uniformity on the second stage can be dropped via an intersection-union method argument [Theorem 1 of Berger and Hsu, 1996], which we show can be applied since our interest concerns parameters defined as the maxima and minima over a set.

As mentioned earlier, justifying the above approach relies on a union-bounding argument across the β coverage error that could be made by the first-stage confidence interval in (4.4)

and the $1-\alpha-\beta$ coverage error that could be made by the second-stage confidence interval in (4.3). Relying on this union bound could result in unnecessarily wide confidence intervals, so we also present another two-stage method whose justification does not require a union bound. In the first stage of this approach we choose the quantiles $s_\alpha^\dagger$, $t_\alpha^\dagger$, and $u_\alpha^\dagger$ derived as extreme values of the joint distributions of estimators of $(\Omega_\pi(P_0))_{\pi\in\Pi}$ and $(\Psi_\pi(P_0))_{\pi\in\Pi}$ — see Section 4.3.3 for details. Then we construct $\widehat{\Pi}_\beta$ and the asymptotic interval the same ways as in (4.4) and (4.3), while replacing t_β and $z_{\alpha,\beta}$ with $t_\alpha^\dagger$ and $s_\alpha^\dagger$, respectively. Given that $s_\alpha^\dagger$, $t_\alpha^\dagger$, and $u_\alpha^\dagger$ are constructed based on a joint distribution, we refer to this approach as the joint approach. Because of the avoidance of the union bound, the joint approach is expected to provide tighter confidence intervals in scenarios when the primary and subsidiary outcomes are strongly correlated.

4.3.2 A union bounding approach

In this subsection, we provide additional details and theoretical results about the union bounding approach. We first need the following condition for an estimator of $\{\Omega_\pi(P_0):\pi\in\Pi\}$. In what follows, we let $\tilde{D}_\pi$ be the canonical gradient of Ψ_π relative to a locally nonparametric model, $\sigma_\pi(P_0) := [PD_\pi(P_0)^2]^{1/2}$, and $\kappa_\pi(P_0) := [P\tilde{D}_\pi(P_0)^2]^{1/2}$.

Condition 5 (Uniform asymptotic linearity of estimators of Ω-value and Ψ-value functions)**.** The estimators $\{\widehat{\omega}_\pi:\pi\in\Pi\}$ of $\{\Omega_\pi(P_0):\pi\in\Pi\}$ and $\{\widehat{\psi}_\pi:\pi\in\Pi\}$ of $\{\Psi_\pi(P_0):\pi\in\Pi\}$ satisfy

$$\sup_{\pi\in\Pi}[\widehat{\omega}_\pi-\Omega_\pi(P_0)-P_nD_\pi(P_0)]=o_p(n^{-1/2}),\qquad \sup_{\pi\in\Pi^*}\left[\widehat{\psi}_\pi-\Psi_\pi(P_0)-P_n\tilde{D}_\pi(P_0)\right]=o_p(n^{-1/2}). \tag{4.5}$$

These asymptotic linearity conditions can be established via consistency requirements similar to those in Condition 4 and a Donsker condition (see Section 2.1 of [Luedtke et al., 2020]). Note that the latter equality in (4.5) only requires uniformity over Π^*, rather than all of Π. Estimators satisfying (4.5) can be derived via one-step estimation [Pfanzagl, 1982],

targeted minimum loss-based estimation [Van Der Laan and Rubin, 2006], or double machine learning [Chernozhukov et al., 2018]. We now provide some conditions on the Ψ-value function, the policy class Π and necessary conditions for standard deviations and the primary outcome.

Condition 6 (Restricted policy class)**.** The policy class Π satisfies the following:

(1) Π has a bounded uniform entropy integral (Chapter 2.5.1 of [Van Der Vaart and Wellner, 2013]);

(2) Π is closed in $L^2(P_0)$, in the sense that, for all $\pi : \mathcal{X} \to \{0, 1\}$, a Π-valued sequence $(\pi_k)_{k=1}^{\infty}$ converges to π in $L^2(P_0)$ only if $\pi \in \Pi$;

(3) Π^* is non-empty.

Examples of such policy class Π in $L^2(P_0)$ include classes of binary decision trees with fixed depths while noting that Condition 6 applies to more complicated and general policy classes. We then provide some conditions for the standard deviations and the smoothness of the Ψ-value function.

Condition 7 (Non-vanishing standard deviations and consistent estimators thereof)**.** The standard deviations satisfy the following conditions: $\inf_{\pi\in\Pi} \sigma_\pi(P_0) > 0$, $\sup_{\pi\in\Pi} \sigma_\pi(P_0) < \infty$, $\inf_{\pi\in\Pi} \kappa_\pi(P_0) > 0$, and $\sup_{\pi\in\Pi} \kappa_\pi(P_0) < \infty$. In addition, $\widehat{\sigma}_\pi$ and $\widehat{\kappa}_\pi$ are uniformly consistent estimators of $\sigma_\pi(P_0)$ and $\kappa_\pi(P_0)$.

Condition 8 (Smoothness of performance metric in policy)**.** The map $\pi \mapsto \Psi_\pi(P_0)$ is continuous and, for all $\pi, \pi' \in \Pi$, $\|D_\pi - D_{\pi'}\|_{L^2(P_0)} \leq C_2 \|\pi - \pi'\|_{L^2(P_0)}$ for some constant C_2.

When Ω and Ψ are covariate-adjusted mean functionals as in Section 4.2 and the primary and subsidiary outcomes are bounded, Condition 8 is necessarily true. Let $\mathcal{F} := \{D_\pi(P_0)/\sigma_\pi(P_0) : \pi \in \Pi\}$ and $\tilde{\mathcal{F}} := \{\tilde{f}_\pi := \tilde{D}_\pi(P_0)/\kappa_\pi(P_0) : \pi \in \Pi\}$ denote the collections of canonical gradients that are standardized to have unit variance. Conditions 7 and 8 play a

crucial role in demonstrating that $\mathcal{F}$ and $\tilde{\mathcal{F}}$ are P_0-Donsker, which is required to validate the uniform confidence bands utilized in our union bounding approach.

We now show that the confidence set $\widehat{\Pi}_\beta$ contains the set of Ω-optimal policies Π^* with high probability asymptotically. Before presenting this result, we define the threshold t_β used to define this confidence set in (4.4). To this end, let $\{\mathbb{G}f : f \in \mathcal{F}\}$ be a mean-zero Gaussian process with a covariance function $(f_1, f_2) \mapsto Pf_1f_2$. Then, t_β is defined to be the $1-\beta/2$ quantile of $\sup_{f\in\mathcal{F}} \mathbb{G}f$ and Lemma C.2.1 in the appendix shows that $\left\{\widehat{\omega}_\pi \pm \frac{\widehat{\sigma}_\pi t_\beta}{n^{1/2}} : \pi \in \Pi\right\}$ is an asymptotically valid uniform β-level confidence band for $\{\omega_\pi : \pi \in \Pi\}$.

Lemma 4.3.1 (Asymptotic coverage of $\widehat{\Pi}_\beta$)**.** *If Conditions 5, 6, and 7 hold, then* $\limsup_n P\{\Pi^* \not\subseteq \widehat{\Pi}_\beta\} \leq \beta$.

The interval in (4.3) uses the remaining $\alpha - \beta$ error probability to construct a confidence interval for the random quantity $\widehat{\mathcal{I}}_\beta := [\inf_{\pi\in\widehat{\Pi}_\beta} \Psi_\pi(P_0), \sup_{\pi\in\widehat{\Pi}_\beta} \Psi_\pi(P_0)]$. On the event that $\Pi^* \subseteq \widehat{\Pi}_\beta$, it is true that $\widehat{\mathcal{I}}_\beta \supseteq [\psi_0^\ell, \psi_0^u]$, and so any interval that covers $\widehat{\mathcal{I}}_\beta$ also covers $[\psi_0^\ell, \psi_0^u]$. A union bound then gives our result. Our findings are summarized in Theorem 4.3.2.

Theorem 4.3.2 (Asymptotic coverage of CI_n)**.** *Under Conditions 5, 6, 7, and 8, for a fixed* $\alpha \in (0,1)$ *and any choice of* $\beta \in (0, \alpha)$*, the confidence interval* CI_n *as defined in* (4.3) *satisfies* $\liminf_{n\to\infty} \mathbb{P}(\{[\psi_0^\ell, \psi_0^u] \subseteq \mathrm{CI}_n\}) \geq 1 - \alpha$.

Also, as indicated in (4.3), the width of CI_n is determined by a quantile of a standard normal random variable — for example, when $\alpha = 0.06$ and $\beta = 0.01$, $z_{\alpha,\beta} \approx 1.96$. At first this may seem surprising, given that developing a uniform confidence band for $\{\Psi_\pi : \pi \in \Pi^*\}$ would require using a strictly larger scaling the standard error of $\widehat{\psi}_\pi$. However, our proof of Theorem 4.3.2 shows that using this larger scaling is not necessary for the sake of developing a confidence interval for $[\psi_0^\ell, \psi_0^u]$. The key to this argument involves showing that, under Condition 6, there exist π^ℓ and π^u in Π^* that attain the minimum and maximum Ψ-values, respectively. The existence of π^ℓ shows that the event where the lower bound of CI_n fails to

cover $\psi_0^\ell := \inf_{\pi \in \Pi^*} \Psi_\pi(P_0)$, intersected with $\Pi^* \subseteq \widehat{\Pi}_\beta$, satisfies

$$\begin{aligned}
&\left\{ \inf_{\pi \in \Pi^*} \Psi_\pi(P_0) < \inf_{\pi \in \widehat{\Pi}_\beta} \left[\widehat{\psi}_\pi - \widehat{\kappa}_\pi z_{\alpha,\beta}/n^{1/2} \right], \Pi^* \subseteq \widehat{\Pi}_\beta \right\} \\
&\subseteq \left\{ \inf_{\pi \in \Pi^*} \Psi_\pi(P_0) < \inf_{\pi \in \Pi^*} \left[\widehat{\psi}_\pi - \widehat{\kappa}_\pi z_{\alpha,\beta}/n^{1/2} \right] \right\} \\
&= \left\{ \psi_{\pi^\ell} < \inf_{\pi \in \Pi^*} \left[\widehat{\psi}_\pi - \widehat{\kappa}_\pi z_{\alpha,\beta}/n^{1/2} \right] \right\} \subseteq \left\{ \psi_{\pi^\ell} < \widehat{\psi}_{\pi^\ell} - \widehat{\kappa}_{\pi^\ell} z_{\alpha,\beta}/n^{1/2} \right\}.
\end{aligned}$$

The event on the right corresponds to the case where a marginal $1 - (\alpha - \beta)/2$-level lower Wald-type confidence interval fails to cover ψ_{π^ℓ}, and so occurs with asymptotic probability $(\alpha - \beta)/2$ under reasonable conditions. In our proof of Theorem 4.3.2, we establish the result using a union bounding argument that combines this with a similar guarantee for the upper bound of CI_n and the fact that $\Pi^* \not\subseteq \widehat{\Pi}_\beta$ happens with asymptotic probability at most β.

Under additional conditions, our confidence interval for $[\psi_0^\ell, \psi_0^u]$ not only ensures asymptotically valid coverage but also attains an optimal $n^{-1/2}$ convergence rate. In this part, we restrict the performance metrics to covariate-adjusted means and propose a boundedness condition on the primary and subsidiary CATE functions.

Condition 9 (Boundedness condition)**.** There exists some $C_3 < \infty$ such that for any $x \in \mathcal{X}$, we have $|s_{b,0}(x)| \leq C_3 |q_{b,0}(x)|$.

In most ways, Condition 9 is relatively stronger than Condition 1. Indeed, the limit of Condition 1 as $\zeta \to \infty$ corresponds to the condition that the subsidiary CATE is strictly less than a constant multiple of the primary CATE. Since Condition 1 allows for any $\zeta > 2$, it puts a much weaker constraint on how the subsidiary outcome behaves for nearly optimal policies. There is one sense, however, in which Condition 9 is weaker than Condition 1: it does not generally imply that the Ω-optimal policy is necessarily unique. This is true because it allows for equality between subsidiary and primary CATEs, and so both could be zero on some set of positive probability. Though the optimal policy need not be unique when Condition 9 holds, it must still be true that all Ω-optimal policies yield the same Ψ-value, and so in the following lemma we shall let $\psi_0 = \psi_0^\ell = \psi_0^u$.

Lemma 4.3.3 ($n^{-1/2}$ convergence rate of CI_n under conditions)**.** *Assume that the performance metrics are covariate-adjusted means as in Section 4.2, the unrestricted Ω-optimal policy over all possible maps from $\mathcal{X}$ to $\{0,1\}$ is in Π, and $L_n = \sup_{\pi\in\Pi}\left[\widehat{\omega}_\pi - \frac{\widehat{\sigma}_\pi t_\beta}{n^{1/2}}\right]$ in (4.4). Then, under Conditions 5, 6, 7, and 9, with probability at least $1-2\beta$ asymptotically, the width of the confidence interval for ψ_0 is $O_p(n^{-1/2})$.*

4.3.3 A joint approach

We now formally describe our joint approach. Consider the mean-zero Gaussian process $\{\mathbb{G}f : f \in \mathcal{F}\cup\tilde{\mathcal{F}}\}$ with covariance function $(f_1, f_2) \mapsto Pf_1f_2$. Our joint approach is the same as the two-stage procedure from Section 4.3.2, except that we require a particular choice of L_n and use cutoffs $(s^\dagger_\alpha, t^\dagger_\alpha, u^\dagger_\alpha)$ satisfying

$$\inf_{\pi\in\Pi}\mathbb{P}\left\{\inf_{f\in\mathcal{F}}\mathbb{G}f \geq -t^\dagger_\alpha, \sup_{f\in\mathcal{F}}\mathbb{G}f \leq s^\dagger_\alpha, \mathbb{G}\tilde{f}_\pi \geq -u^\dagger_\alpha, \mathbb{G}\tilde{f}_\pi \leq u^\dagger_\alpha\right\} \geq 1-\alpha. \tag{4.6}$$

More specifically, we define the set $\widehat{\Pi}^\dagger$ after the first-stage filtration as

$$\widehat{\Pi}^\dagger := \left\{\pi\in\Pi : \sup_{\pi\in\Pi}\left[\widehat{\omega}_\pi - \frac{\widehat{\sigma}_\pi s^\dagger_\alpha}{n^{1/2}}\right] \leq \widehat{\omega}_\pi + \frac{\widehat{\sigma}_\pi t^\dagger_\alpha}{n^{1/2}}\right\}. \tag{4.7}$$

Here we choose L_n to be the uppermost point of a uniform lower confidence band for $\{\Omega_\pi(P_0) : \pi\in\Pi\}$ with level $\beta^\dagger := \mathbb{P}\{\sup_{\pi\in\Pi}\mathbb{G}f_\pi > s^\dagger_{1-\alpha}\} < \alpha$. Note that in the union bounding approach, $\beta^\dagger = \beta$, while here $\beta^\dagger$ is implicitly defined through the joint cutoff (4.6). The resulting confidence interval is stated in Theorem 4.3.4.

Theorem 4.3.4. *Under Conditions 5, 6, 7, and 8, assuming the cutoffs $(s^\dagger_\alpha, t^\dagger_\alpha, u^\dagger_\alpha)$ satisfy (4.6), it holds that $\liminf_{n\to\infty}\mathbb{P}(\{[\psi^\ell_0, \psi^u_0] \subseteq \mathrm{CI}^\dagger_n\}) \geq 1-\alpha$, where*

$$\mathrm{CI}^\dagger_n := \left[\inf_{\pi\in\widehat{\Pi}^\dagger}\left\{\widehat{\psi}_\pi - \frac{\widehat{\kappa}_\pi u^\dagger_\alpha}{n^{1/2}}\right\}, \sup_{\pi\in\widehat{\Pi}^\dagger}\left\{\widehat{\psi}_\pi + \frac{\widehat{\kappa}_\pi u^\dagger_\alpha}{n^{1/2}}\right\}\right].$$

There are many possible choices of $(s^\dagger_\alpha, t^\dagger_\alpha, u^\dagger_\alpha)$ that satisfy (4.6). To select among these, we could choose the triple $(s^\dagger_\alpha, t^\dagger_\alpha, u^\dagger_\alpha)$ that provides the tightest confidence interval from this collection, resulting in what we refer to as an optimized joint method. This optimized

$(s_\alpha^\dagger, t_\alpha^\dagger, u_\alpha^\dagger)$ is justified since, for any choice of $(s_\alpha^\dagger, t_\alpha^\dagger, u_\alpha^\dagger)$ satisfying (4.6), the confidence interval $\mathrm{CI}_{\mathrm{n}}^\dagger$ has valid coverage. For any β, this optimized joint method yields a provably tighter confidence interval than the union bounding method that uses the same choice of L_n as in the left-hand side of (4.7). However, it is possible that the joint approach could potentially result in a wider confidence band in the first stage with the use of an alternative lower confidence bound for the Ω-optimal value, such as the one introduced in Luedtke and Van Der Laan [2016]. In practice, the optimized choice of $(s_\alpha^\dagger, t_\alpha^\dagger, u_\alpha^\dagger)$ is unknown, but it can be approximated via a multiplier bootstrap — see Appendix C.3 for details. Though our theorem focuses on a fixed and known triple $(s_\alpha^\dagger, t_\alpha^\dagger, u_\alpha^\dagger)$, adapting it to allow for the use of an estimated triple with an in-probability limit would be straightforward.

The cutoff in (4.6) considers the joint event regarding $\mathbb{G}\tilde{f}$ and $\mathbb{G}f$ for $f \in \mathcal{F}$ and $\tilde{f} \in \tilde{\mathcal{F}}$, thereby avoiding the use of the union bound required by the approach in Section 4.3.2. The tightness of this union bound relies on whether the event that Π^* is contained in the first stage policy set, namely $\{\Pi^* \subseteq \widehat{\Pi}_\beta\}$, and the event that $[\psi_0^\ell, \psi_0^u]$ is contained in the second stage confidence interval are disjoint. Of course, when these events are fully disjoint, the union bound will be tight. When they are independent, the (asymptotic) probability that both events occur is $\beta(\alpha - \beta)$, which will be small for choices of α and β commonly used in practice. Hence, the union bound will only be slightly loose in these cases. Finally, when the events fully overlap, the union bound will be as loose as possible. These scenarios can be better understood by relating them to primary and subsidiary outcomes. Generally, the dependence or independence between the events is likely to correlate with the extent to which primary and subsidiary outcomes depend on each other. The events tend to be independent when primary and subsidiary outcomes are independent, and dependent otherwise.

4.4 Numerical experiment

4.4.1 A 1D simulation

We conduct simulation studies to evaluate the length and coverage of $1-\alpha$ confidence intervals for bounds on a mean subsidiary outcome, $[\psi_0^\ell, \psi_0^u]$. Our first set of simulations focuses on a 1-dimensional threshold policy class, denoted as $\Pi = \{\mathbf{1}_{[a,\infty)} : a \in [-1, 1]\}$. We compare the confidence intervals from four approaches. The first is the union bounding approach described in Section 4.3.2, denoted as union. The second is the joint approach described in Section 4.3.3, denoted as joint. The third is the one-step estimator approach described in Section 4.2, denoted as one-step. To ensure that this approach applies, we design our scenarios so that the optimal policy for the unrestricted policy class lies in the threshold class Π. Consequently, in our simulation study, an estimate of the optimal policy in Π also estimates the optimal policy in the unrestricted class. The fourth is a one-step estimator with sample splitting, denoted as os-split. This approach is the same as one-step, except that we obtain an estimate $\widehat{\pi}_1$ of the Ω-optimal policy using only half of the data, and construct a Wald-type confidence interval for $\Psi_{\widehat{\pi}_1}(P_0)$ using the other half. Last, we present an oracle method, denoted as oracle, that knows the specific Ω-optimal policies that provide the upper and lower bounds, ψ_0^u and ψ_0^ℓ. The oracle method uses precisely those policies and construct a Wald-type confidence interval for $[\psi_0^\ell, \psi_0^u]$. Since we have no hope of getting optimal policies *a priori*, the oracle method cannot be used in practice.

We examine three distinct scenarios with an illustration of the Ω and Ψ values of each policy under various scenarios in the three panels in Figure 4.3. The left panel describes the situation where the set of Ω-optimal policies, Π^*, is not unique. In this scenario, $\Pi^* = \{\mathbf{1}_{[a,\infty)} : a \in [-0.5, 0]\}$, and the margin condition (Condition 1) is not satisfied for any ζ. The middle panel describes the situation where Π^* is unique the margin condition is satisfied for any $\zeta > 2$, as we can see that when π is around the optimal policy, $q_{b,0}(X)$ varies much faster than $s_{b,0}(X)$. The right panel describes the situation where Π^* is unique but the margin condition is not satisfied for any ζ, as we can see that as X varies, both $q_{b,0}(X)$ and

$s_{b,0}(X)$ vary linearly.

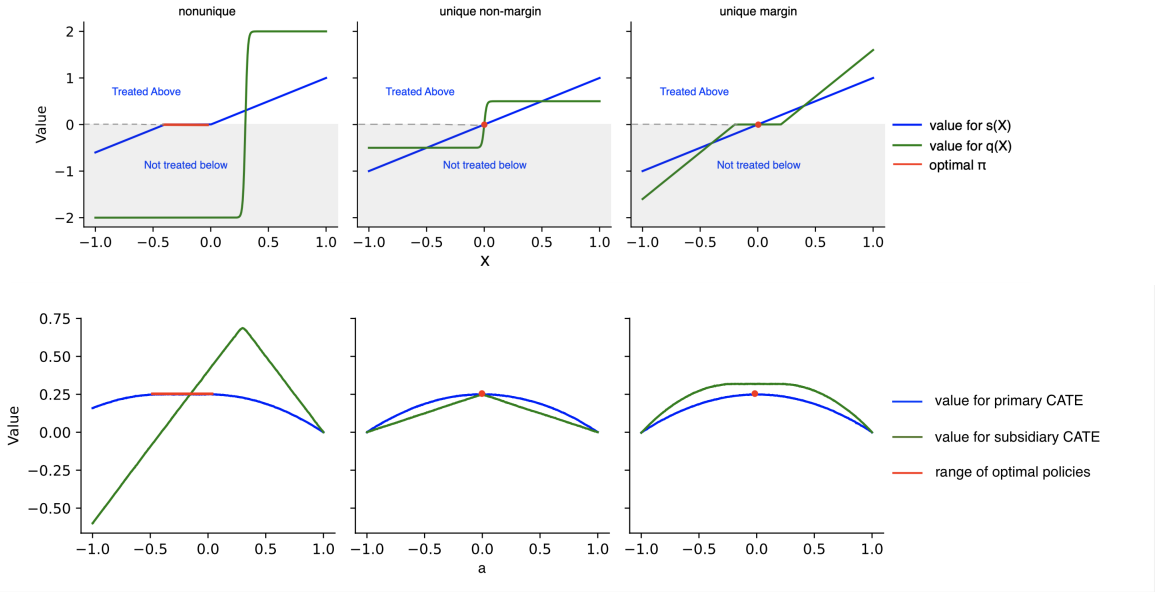

Figure 4.3: An illustration of $s_{b,0}(X)$- and $q_{b,0}(X)$-value and Ω- and Ψ-value for a 1-dimensional threshold policy class $\Pi = \{\mathbf{1}_{[a,\infty)} : a \in [-1, 1]\}$ under different scenarios: the optimal policy for the primary outcome is nonunique, the optimal policy for the primary outcome is unique while the primary and subsidiary outcomes are correlated, and the optimal policy for the primary outcome is unique while the primary and subsidiary outcomes are not so correlated. The top figure represents Ω- and Ψ-value, while the bottom figure represents $s_{b,0}(X)$- and $q_{b,0}(X)$-value.

For each scenario, we consider sample sizes n of 500 and 5000. To generate the set of policies, we construct a fine grid $(a_1, \cdots, a_N)$ for $N = 10^5$ over $[-1, 1]$ and denote the set of policy as $\Pi_N = \{\mathbf{1}_{[a_i,\infty)} : i \in [N]\}$. We use 1000 multiplier bootstrap replicates to estimate the supremum and infimum in generating the cutoffs. We let $\alpha = 0.05$ when constructing confidence intervals and use 1000 Monte Carlo replications to compute their coverage of the true interval $[\psi_0^\ell, \psi_0^u]$ as well as approximate their average widths. We estimate the conditional probability $p(a|x)$ via a kernel density estimator as implemented in the sklearn

package and the conditional probabilities $p(y|1,x)$ and $p(y|0,x)$ using gradient boosted trees as implemented in the xgboost package, both with the default settings. The Python code to reproduce the simulations is available at https://github.com/zhaoqil/EstimationSubsidiary.

Table 4.1 shows the coverages and the widths of confidence intervals of $[\psi_0^\ell, \psi_0^u]$ for different scenarios and different methods. We can see the the one-step estimator fails to provide a nominal coverage when the margin condition (Condition 1) is not satisfied. The other two methods produce similar coverages. We compare the confidence intervals with an oracle confidence interval, which is a lower bound on the width of any valid $1-\alpha$ confidence interval, and calculate the relative widths. We can see that the joint and union bounding methods generate confidence intervals about 2.3 times and 2.1 times as wide as the oracle confidence interval when the optimal policy is non-unique and unique, respectively. These results show that although our methods are conservative, they are relatively successful in maintaining a narrow confidence interval. In contrast, the one-step estimator produces a confidence interval that is about the same width as the oracle confidence interval, but it fails to provide valid coverage when the margin condition fails. Table 4.2 provides coverages and confidence interval widths with a larger sample size of 5000. In the non-unique setting, since there are multiple optimal policies for the primary outcome, $[\psi_0^\ell, \psi_0^u]$ will be an interval with some length. In our setting, we can see from the lower-left plot of Figure 4.3 that the length of $[\psi_0^\ell, \psi_0^u]$ is about 0.5, so any valid confidence interval for $[\psi_0^\ell, \psi_0^u]$ must have at least that length. Comparing the widths in Table 4.1 and 4.2, we can see that both the union bounding method and the joint method produce confidence intervals approaching that limit. In the setting where Ω-optimal policy is unique, the widths of the confidence intervals for all methods approach zero as n goes to infinity.

4.4.2 A 3D simulation

We also added a scenario where we have a 3D policy and the optimal policy is unique. The policy class is a restricted tree class, denoted as $\Pi = \{x \mapsto \mathbf{1}\{x \geq a_1, x \geq a_2, x \geq a_3\} : a_1, a_2, a_3 \in [-1, 1]\}\}$. We design our scenario so that the optimal policy for the unrestricted

	coverage				width				
	union	joint	one-step	os-split	union	joint	one-step	os-split	oracle
non-unique	1.000	1.000	0.000	0.000	1.549	1.538	0.240	0.317	0.668
unique non-margin	0.980	0.980	0.812	0.751	0.148	0.143	0.068	0.089	0.068
unique margin	0.978	0.981	0.949	0.953	0.149	0.144	0.074	0.108	0.074

Table 4.1: Coverage and width of $[\psi_0^\ell, \psi_0^u]$ for different scenarios with sample size $n = 500$

	coverage				width				
	union	joint	one-step	os-split	union	joint	one-step	os-split	oracle
non-unique	1.000	1.000	0.000	0.000	1.091	1.061	0.061	0.096	0.561
unique non-margin	0.981	0.986	0.810	0.734	0.036	0.035	0.017	0.027	0.016
unique margin	0.983	0.989	0.946	0.949	0.040	0.036	0.023	0.037	0.023

Table 4.2: Coverage and width of $[\psi_0^\ell, \psi_0^u]$ for different scenarios with sample size $n = 5000$

policy class lies in the tree class. We compare the outcome interval from three approaches: union, joint, and one-step. The method os-split provides a wider interval while having a worse coverage than one-step in 1D simulation results, so we drop it from the simulation. For each scenario, we consider a sample size n of 500. We again use 1000 multiplier bootstrap replicates to estimate the supremum and infimum. In this scenario, instead of generating a fine grid and computing the maximum over the grid, we use the nlopt package to numerically approximate the maximum. We let $\alpha = 0.05$ and use 500 Monte Carlo replications to compute the coverage and approximate the average confidence interval widths. Table 4.3 shows the results. The joint methods achieves slightly shorter widths in this setting (5-6%), and the results are otherwise similar to those from Section 4.4.1.

4.5 Discussion

The problem studied in existing works aiming to infer about the optimal value of an optimal rule can be viewed as a special case of our setup, where the subsidiary and primary outcomes

	coverage			width			
	union	joint	one-step	union	joint	one-step	oracle
3D margin	0.970	0.948	0.940	0.199	0.186	0.124	0.124
3D non-margin	1.000	0.988	0.594	0.185	0.175	0.092	0.092

Table 4.3: Coverage and width for 3D policy class with sample size $n = 500$

coincide. In these cases, our two-stage approaches provide ways to make inference without the margin condition considered in such works Qian and Murphy [2011], Luedtke and Van Der Laan [2016]. Instead, we need uniform asymptotic linearity for the value functions and an appropriately restricted policy class. The margin condition could fail if the subsidiary metric varies too much across the set of policies that are nearly optimal for the primary metric Luedtke et al. [2020]. However, if the policy class is Donsker and the estimator is established via debiased machine learning, the uniform asymptotic linearity condition will be plausible even when a margin condition does not hold.

In our numerical experiments, our union bounding and joint approaches produced valid confidence intervals, even if they were somewhat conservative. Under margin conditions, these intervals attain a parametric $n^{-1/2}$ rate, matching those based on an efficient one-step estimator, although with a less favorable leading constant. However, when the margin conditions fail, intervals based on the one-step estimator fail to achieve valid coverage.

Chapter 5

CONCLUSION

In this dissertation, we addressed some challenges to estimation and inference for optimal policies. There are several promising directions for future work.

In Chapter 2, we present an instance-optimal and computationally efficient algorithm for pure exploration in contextual bandits. Although this algorithm is computationally efficient, it requires an enormous amount of calls to the oracle to run. In particular, each iteration of the Frank-Wolfe algorithm would require a call to the cost-sensitive classification oracle. A follow-up work Krishnamurthy et al. [2024] proposes a computationally efficient algorithm for contextual bandits in the simple regret minimization setting. Their algorithm only requires exponentially fewer calls to the oracle to run, while their objective is slightly different from PAC learning, so a future direction is to develop similar fast algorithms for PAC learning. Also, the computational efficiency only applies to agnostic settings instead of linear realizable settings. It remains an interesting direction to think about computationally efficient algorithms for that setting.

In Chapter 3, we present a simple yet asymptotically optimal algorithm for pure exploration in linear bandits. Given its simplicity, it would be interesting to develop similar sampling-based algorithms for pure exploration in contextual bandits or reinforcement learning. Most existing works on reinforcement learning focus on regret minimization, and existing pure exploration algorithms in linear Markov Decision Processes (MDP) and contextual bandits are either computationally inefficient or rely on oracles that are computationally intensive in many settings Wagenmaker and Jamieson [2022], Li et al. [2022]. However, developing optimal pure exploration algorithms in reinforcement learning is more challenging as the complexity of finding the optimal allocation, i.e. the sample complexity lower bound, remains

understudied. Existing algorithms are generally empirical and lack theoretical justifications Schulman et al. [2017], Rafailov et al. [2024]. A recent work Wagenmaker and Foster [2023] makes a step towards this direction in a general interactive decision-making framework.

In Chapter 4, we develop estimators and methods for the inference of subsidiary metrics under an optimal policy for the primary metric. The estimator that achieves efficiency relies on a strong margin condition that assumes that the subsidiary metric is relatively flat compared to the primary metric, while the two-stage and the union bounding approach produces relatively wide confidence intervals empirically. In future research, it would be interesting to develop an adaptive procedure that is leading-constant-optimal under margin conditions and, even without them, can produce intervals that provide valid coverage. As for other future work, it is worth exploring methods for inferring subsidiary metrics using observations from adaptive experiments, which are non-independent but have a martingale structure. Observations from longitudinal settings could also be considered. Additionally, one could examine simultaneous inference for multiple subsidiary metrics rather than one.

BIBLIOGRAPHY

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, 2020.

Liyuan Xu, Junya Honda, and Masashi Sugiyama. A fully adaptive algorithm for pure exploration in linear bandits. In *International Conference on Artificial Intelligence and Statistics*, pages 843–851. PMLR, 2018.

Tze Leung Lai and Herbert Robbins. Asymptotically efficient adaptive allocation rules. *Advances in applied mathematics*, 6(1):4–22, 1985.

Tanner Fiez, Lalit Jain, Kevin Jamieson, and Lillian Ratliff. Sequential experimental design for transductive linear bandits. In *Advances in Neural Information Processing Systems*, 2019.

Chao Tao, Saúl Blanco, and Yuan Zhou. Best arm identification in linear bandits with linear dimension dependency. In *International Conference on Machine Learning*, pages 4877–4886. PMLR, 2018.

Soorajnath Boominathan, Michael Oberst, Helen Zhou, Sanjat Kanjilal, and David Sontag. Treatment policy learning in multiobjective settings with fully observed outcomes. In *Proc. 26th ACM SIGKDD Int. Conf. Knowl. Discov. & Data Min.*, pages 1937–1947, 2020.

Ioana Bica, Ahmed M Alaa, Craig Lambert, and Mihaela Van Der Schaar. From real-world patient data to individualized treatment effects using machine learning: current and future methods to address underlying challenges. *Clin. Pharmacol. Ther.*, 109(1):87–100, 2021.

Shuhua Chang, Zhaowei Zhang, Xinyu Wang, and Yan Dong. Optimal acquisition and retention strategies in a duopoly model of competition. *European Journal of Operational Research*, 282(2):677–695, 2020.

Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. *Advances in neural information processing systems*, 24, 2011.

Daniel J Russo, Benjamin Van Roy, Abbas Kazerouni, Ian Osband, Zheng Wen, et al. A tutorial on thompson sampling. *Foundations and Trends® in Machine Learning*, 11(1): 1–96, 2018.

Lalit Jain, Zhaoqi Li, Erfan Loghmani, Blake Mason, and Hema Yoganarasimhan. Effective adaptive exploration of prices and promotions in choice-based demand models. *Available at SSRN 4438537*, 2023.

Kwang-Sung Jun, Lalit Jain, Blake Mason, and Houssam Nassif. Improved confidence bounds for the linear logistic model and applications to bandits. In *International Conference on Machine Learning (ICML)*, pages 5148–5157, 2021a.

Peter Auer, Nicolo Cesa-Bianchi, Yoav Freund, and Robert E Schapire. The nonstochastic multiarmed bandit problem. *SIAM journal on computing*, 32(1):48–77, 2002.

Miroslav Dudik, Daniel Hsu, Satyen Kale, Nikos Karampatziakis, John Langford, Lev Reyzin, and Tong Zhang. Efficient optimal learning for contextual bandits. In *Proceedings of the Twenty-Seventh Conference on Uncertainty in Artificial Intelligence*, pages 169–178, 2011.

Alekh Agarwal, Daniel Hsu, Satyen Kale, John Langford, Lihong Li, and Robert Schapire. Taming the monster: A fast and simple algorithm for contextual bandits. In *International Conference on Machine Learning*, pages 1638–1646. PMLR, 2014.

John Langford and Tong Zhang. The epoch-greedy algorithm for multi-armed bandits with side information. *Advances in neural information processing systems*, 20, 2007.

Dylan Foster, Alekh Agarwal, Miroslav Dudik, Haipeng Luo, and Robert Schapire. Practical contextual bandits with regression oracles. In *International Conference on Machine Learning*, pages 1539–1548. PMLR, 2018.

Dylan Foster and Alexander Rakhlin. Beyond ucb: Optimal and efficient contextual bandits with regression oracles. In *International Conference on Machine Learning*, pages 3199–3210. PMLR, 2020.

David Simchi-Levi and Yunzong Xu. Bypassing the monster: A faster and simpler optimal algorithm for contextual bandits under realizability. *Mathematics of Operations Research*, 2021.

Dylan Foster, Alexander Rakhlin, David Simchi-Levi, and Yunzong Xu. Instance-dependent complexity of contextual bandits and reinforcement learning: A disagreement-based perspective. In *Conference on Learning Theory*, pages 2059–2059. PMLR, 2021a.

Dylan J Foster, Sham M Kakade, Jian Qian, and Alexander Rakhlin. The statistical complexity of interactive decision making. *arXiv preprint arXiv:2112.13487*, 2021b.

Lihong Li, Wei Chu, John Langford, and Robert E Schapire. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th international conference on World wide web*, pages 661–670, 2010.

Daniel Russo. Simple bayesian algorithms for best arm identification. In *Conference on Learning Theory*, pages 1417–1418, 2016.

Sareh Nabi, Houssam Nassif, Joseph Hong, Hamed Mamani, and Guido Imbens. Bayesian meta-prior learning using Empirical Bayes. *Management Science*, 68(3):1737–1755, 2022.

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.

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 Zanette, Kefan Dong, Jonathan Lee, and Emma Brunskill. Design of experiments for stochastic contextual linear bandits. *Advances in Neural Information Processing Systems*, 34, 2021.

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.

Tzu-Kuo Huang, Alekh Agarwal, Daniel J Hsu, John Langford, and Robert E Schapire. Efficient and parsimonious agnostic active learning. *Advances in Neural Information Processing Systems*, 28, 2015.

Akshay Krishnamurthy, Alekh Agarwal, Tzu-Kuo Huang, Hal Daumé III, and John Langford. Active learning for cost-sensitive classification. In *International Conference on Machine Learning*, pages 1915–1924. PMLR, 2017.

Alina Beygelzimer, Varsha Dani, Tom Hayes, John Langford, and Bianca Zadrozny. Error limiting reductions between classification tasks. In *Proceedings of the 22nd international conference on Machine learning*, pages 49–56, 2005.

Alekh Agarwal, Alina Beygelzimer, Miroslav Dudík, John Langford, and Hanna Wallach. A reductions approach to fair classification. In *International Conference on Machine Learning*, pages 60–69. PMLR, 2018.

Andrew Cotter, Maya Gupta, Heinrich Jiang, Nathan Srebro, Karthik Sridharan, Serena Wang, Blake Woodworth, and Seungil You. Training well-generalizing classifiers for fairness metrics and other data-dependent constraints. In *International Conference on Machine Learning*, pages 1397–1405. PMLR, 2019.

Marta Soare, Alessandro Lazaric, and Rémi Munos. Best-arm identification in linear bandits. *Advances in Neural Information Processing Systems*, 27, 2014.

Steve Hanneke et al. Theory of disagreement-based active learning. *Foundations and Trends® in Machine Learning*, 7(2-3):131–309, 2014.

Gábor Lugosi and Shahar Mendelson. Mean estimation and regression under heavy-tailed distributions: A survey. *Foundations of Computational Mathematics*, 19(5):1145–1190, 2019.

Fabian Pedregosa, Geoffrey Negiar, Armin Askari, and Martin Jaggi. Linearly convergent frank-wolfe with backtracking line-search. In *Proceedings of the 23rdInternational Conference on Artificial Intelligence and Statistics*, 2020. URL `https://arxiv.org/pdf/1806.05123.pdf`.

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.

Peter Glynn and Sandeep Juneja. A large deviations perspective on ordinal optimization. In *Proceedings of the 2004 Winter Simulation Conference, 2004.*, volume 1. IEEE, 2004.

Zohar Karnin, Tomer Koren, and Oren Somekh. Almost optimal exploration in multi-armed bandits. In *International Conference on Machine Learning*, pages 1238–1246. PMLR, 2013.

Sébastien Bubeck. Introduction to online optimization. *Lecture notes*, 2:1–86, 2011.

Lijie Chen, Anupam Gupta, Jian Li, Mingda Qiao, and Ruosong Wang. Nearly optimal sampling algorithms for combinatorial pure exploration. In *Conference on Learning Theory*, pages 482–534. PMLR, 2017.

Ari Biswas, Thai T Pham, Michael Vogelsong, Benjamin Snyder, and Houssam Nassif. Seeker:

Real-time interactive search. In *International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 2867–2875, 2019.

Mingrui Liu and Francesco Orabona. On the initialization for convex-concave min-max problems. In *International Conference on Algorithmic Learning Theory*, pages 743–767. PMLR, 2022.

John Burkardt. The truncated normal distribution. *Department of Scientific Computing Website, Florida State University*, 1:35, 2014.

Luc Devroye. *Non-Uniform Random Variate Generation*. Springer-Verlag, New York, NY, USA, 1986.

Kevin P. Murphy. *Machine learning : a probabilistic perspective*. MIT Press, Cambridge, Mass. [u.a.], 2013.

Yifang Li and Sujit K. Ghosh. Efficient sampling methods for truncated multivariate normal and student-t distributions subject to linear inequality constraints. *Journal of Statistical Theory and Practice*, 9(4):712–732, 2015.

Aditi Laddha and Santosh S Vempala. Convergence of gibbs sampling: coordinate hit-and-run mixes fast. *Discrete & Computational Geometry*, pages 1–20, 2023.

László Lovász. Hit-and-run mixes fast. *Mathematical programming*, 86:443–461, 1999.

Hassan Maatouk and Xavier Bay. A new rejection sampling method for truncated multivariate gaussian random variables restricted to convex sets. In *Monte Carlo and Quasi-Monte Carlo Methods: MCQMC, Leuven, Belgium, April 2014*, pages 521–530. Springer, 2016.

Shipra Agrawal and Navin Goyal. Near-optimal regret bounds for thompson sampling. *Journal of the ACM (JACM)*, 64(5):1–24, 2017.

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.

Shai Shalev-Shwartz et al. Online learning and online convex optimization. *Foundations and Trends® in Machine Learning*, 4(2):107–194, 2012.

Marc Jourdan, Rémy Degenne, Dorian Baudry, Rianne de Heide, and Emilie Kaufmann. Top two algorithms revisited. *arXiv preprint arXiv:2206.05979*, 2022.

Steven De Rooij, Tim Van Erven, Peter D Grünwald, and Wouter M Koolen. Follow the leader if you can, hedge if you must. *The Journal of Machine Learning Research*, 15(1): 1281–1316, 2014.

Yassir Jedra and Alexandre Proutiere. Optimal best-arm identification in linear bandits. *Advances in Neural Information Processing Systems*, 33:10007–10017, 2020.

Rémy Degenne, Wouter M Koolen, and Pierre Ménard. Non-asymptotic pure exploration by solving games. *Advances in Neural Information Processing Systems*, 32, 2019.

Po-An Wang, Ruo-Chun Tzeng, and Alexandre Proutiere. Fast pure exploration via frank-wolfe. *Advances in Neural Information Processing Systems*, 34:5810–5821, 2021.

Mohammadi Zaki, Avinash Mohan, and Aditya Gopalan. Improved pure exploration in linear bandits with no-regret learning. In *IJCAI International Joint Conference on Artificial Intelligence*, pages 3709–3715. International Joint Conferences on Artificial Intelligence, 2022.

Andrea Tirinzoni and Rémy Degenne. On elimination strategies for bandit fixed-confidence identification. *arXiv preprint arXiv:2205.10936*, 2022.

Abbas Kazerouni and Lawrence M Wein. Best arm identification in generalized linear bandits. *Operations Research Letters*, 49(3):365–371, 2021.

Kwang-Sung Jun, Lalit Jain, Blake Mason, and Houssam Nassif. Improved confidence bounds for the linear logistic model and applications to bandits. In *International Conference on Machine Learning*, pages 5148–5157. PMLR, 2021b.

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.

Chao Qin, Diego Klabjan, and Daniel Russo. Improving the expected improvement algorithm. *Advances in Neural Information Processing Systems*, 30, 2017.

Xuedong Shang, Rianne Heide, Pierre Menard, Emilie Kaufmann, and Michal Valko. Fixed-confidence guarantees for bayesian best-arm identification. In *International Conference on Artificial Intelligence and Statistics*, pages 1823–1832. PMLR, 2020.

Chao Qin and Daniel Russo. Adaptivity and confounding in multi-armed bandit experiments. *arXiv preprint arXiv:2202.09036*, 2022.

Jongyeong Lee, Junya Honda, and Masashi Sugiyama. Thompson exploration with best challenger rule in best arm identification. *arXiv preprint arXiv:2310.00539*, 2023.

Lihong Li. Generalized thompson sampling for contextual bandits. *arXiv preprint arXiv:1310.7163*, 2013.

Tor Lattimore and Andras Gyorgy. Mirror descent and the information ratio. In *Conference on Learning Theory*, pages 2965–2992. PMLR, 2021.

Julian Zimmert and Tor Lattimore. Connections between mirror descent, thompson sampling and the information ratio. *Advances in Neural Information Processing Systems*, 32, 2019.

Susan Athey and Stefan Wager. Policy learning with observational data. *Econometrica*, 89 (1):133–161, 2021.

Yaobin Ling, Pulakesh Upadhyaya, Luyao Chen, Xiaoqian Jiang, and Yejin Kim. Heterogeneous treatment effect estimation using machine learning for healthcare application: tutorial and benchmark. *arXiv preprint arXiv:2109.12769*, 2021.

Daniel N Hill, Houssam Nassif, Yi Liu, Anand Iyer, and SVN Vishwanathan. An efficient bandit algorithm for realtime multivariate optimization. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1813–1821, 2017.

Min Qian and Susan A Murphy. Performance guarantees for individualized treatment rules. *Ann. Stat.*, 39(2):1180, 2011.

Yingqi Zhao, Donglin Zeng, A John Rush, and Michael R Kosorok. Estimating individualized treatment rules using outcome weighted learning. *J. Am. Stat. Assoc.*, 107(499):1106–1118, 2012.

Miroslav Dudík, John Langford, and Lihong Li. Doubly robust policy evaluation and learning. *arXiv preprint arXiv:1103.4601*, 2011.

Baqun Zhang, Anastasios A Tsiatis, Eric B Laber, and Marie Davidian. Robust estimation of optimal dynamic treatment regimes for sequential treatment decisions. *Biometrika*, 100 (3):681–694, 2013.

Alex Luedtke, Antoine Chambaz, et al. Performance guarantees for policy learning. In *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, pages 2162–2188. Institut Henri Poincaré, 2020.

Chengchun Shi, Sheng Zhang, Wenbin Lu, and Rui Song. Statistical inference of the value function for reinforcement learning in infinite horizon settings. *arXiv preprint arXiv:2001.04515*, 2020.

Lin Liu, Zach Shahn, James M Robins, and Andrea Rotnitzky. Efficient estimation of optimal regimes under a no direct effect assumption. *J. Am. Stat. Assoc.*, 116(533):224–239, 2021.

Mark J van der Laan and Alexander R Luedtke. Targeted learning of the mean outcome under an optimal dynamic treatment rule. *Journal of causal inference*, 3(1):61–95, 2015.

Antoine Chambaz, Wenjing Zheng, and Mark J van der Laan. Targeted sequential design for targeted learning inference of the optimal treatment rule and its mean reward. *Ann. Stat.*, 45(6):2537, 2017.

James M Robins. Optimal structural nested models for optimal sequential decisions. In *Proceedings of the second seattle Symposium in Biostatistics*, pages 189–326. Springer, 2004.

Alexander R Luedtke and Mark J Van Der Laan. Statistical inference for the mean outcome under a possibly non-unique optimal treatment strategy. *Ann. Stat.*, 44(2):713 – 742, 2016.

Bibhas Chakraborty, Eric B Laber, and Yingqi Zhao. Inference for optimal dynamic treatment regimes using an adaptive m-out-of-n bootstrap scheme. *Biometrics*, 69(3):714–723, 2013.

Rachel Phillips, Odile Sauzet, and Victoria Cornelius. Statistical methods for the analysis of adverse event data in randomised controlled trials: a scoping review and taxonomy. *BMC medical research methodology*, 20(1):1–13, 2020.

Nick Freemantle, Melanie Calvert, John Wood, Joanne Eastaugh, and Carl Griffin. Composite outcomes in randomized trials: greater precision but with greater uncertainty? *Jama*, 289 (19):2554–2559, 2003.

Emily L Butler, Eric B Laber, Sonia M Davis, and Michael R Kosorok. Incorporating patient preferences into estimation of optimal individualized treatment rules. *Biometrics*, 74(1): 18–26, 2018.

Thomas A Murray, Peter F Thall, and Ying Yuan. Utility-based designs for randomized comparative trials with categorical outcomes. *Statistics in medicine*, 35(24):4285–4305, 2016.

Daniel J Luckett, Eric B Laber, Siyeon Kim, and Michael R Kosorok. Estimation and optimization of composite outcomes. *Journal of Machine Learning Research*, 22(167):1–40, 2021.

Eric B Laber, Daniel J Lizotte, and Bradley Ferguson. Set-valued dynamic treatment regimes for competing outcomes. *Biometrics*, 70(1):53–61, 2014.

Kristin A Linn, Eric B Laber, and Leonard A Stefanski. Chapter 15: Estimation of dynamic treatment regimes for complex outcomes: balancing benefits and risks. In *Adaptive treatment strategies in practice: Planning trials and analyzing data for personalized medicine*, pages 249–262. SIAM, 2015.

Yuanjia Wang, Haoda Fu, and Donglin Zeng. Learning optimal personalized treatment rules in consideration of benefit and risk: with an application to treating type 2 diabetes patients with insulin therapies. *Journal of the American Statistical Association*, 113(521):1–13, 2018.

FDA. Guidance for industry: Adverse reactions section of labeling for human prescription drug and biological products – content and format https://www.fda.gov/media/72139/download. 2006.

Philipp Afeche, Mojtaba Araghi, and Opher Baron. Customer acquisition, retention, and service access quality: Optimal advertising, capacity level, and capacity allocation. *Manuf. Serv. Oper. Manag.*, 19(4):674–691, 2017.

Nyoman Gunantara. A review of multi-objective optimization: Methods and its applications. *Cogent Engineering*, 5(1):1502242, 2018.

Kalyanmoy Deb. Multi-objective optimization. In *Search methodologies*, pages 403–449. Springer, 2014.

Peter J Bentley and Jonathan P Wakefield. Finding acceptable solutions in the pareto-optimal range using multiobjective genetic algorithms. In *Soft computing in engineering design and manufacturing*, pages 231–240. Springer, 1998.

Jean-Yves Audibert and Alexandre B Tsybakov. Fast learning rates for plug-in classifiers. *Ann. Stat.*, 35(2):608–633, 2007.

Anton Schick. On asymptotically efficient estimation in semiparametric models. *Ann. Stat.*, pages 1139–1151, 1986.

Aad W Van Der Vaart and Jon A Wellner. *Weak convergence and empirical processes: with applications to statistics*. Springer, 2013.

Roger L Berger and Jason C Hsu. Bioequivalence trials, intersection-union tests and equivalence confidence sets. *Statistical Science*, 11(4):283–319, 1996.

Johann Pfanzagl. Lecture notes in statistics. *Contributions to a general asymptotic statistical theory*, 13, 1982.

Mark J Van Der Laan and Daniel Rubin. Targeted maximum likelihood learning. *The international journal of biostatistics*, 2(1), 2006.

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters, 2018.

Sanath Kumar Krishnamurthy, Ruohan Zhan, Susan Athey, and Emma Brunskill. Proportional response: Contextual bandits for simple and cumulative regret minimization. *Advances in Neural Information Processing Systems*, 36, 2024.

Andrew Wagenmaker and Kevin Jamieson. Instance-dependent near-optimal policy identification in linear mdps via online experiment design. *arXiv preprint arXiv:2207.02575*, 2022.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36, 2024.

Andrew J 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.

Shie Mannor and John N Tsitsiklis. Lower bounds on the sample complexity of exploration in the multi-armed bandit problem. In *Learning Theory and Kernel Machines*, pages 418–432. Springer, 2003.

Aad W Van der Vaart. *Asymptotic statistics*, volume 3. Cambridge university press, 2000.

Chi Jin, Praneeth Netrapalli, Rong Ge, Sham M Kakade, and Michael I Jordan. On nonconvex optimization for machine learning: Gradients, stochasticity, and saddle points. *Journal of the ACM (JACM)*, 68(2):1–29, 2021.

John Milnor and David W Weaver. *Topology from the differentiable viewpoint*, volume 21. Princeton university press, 1997.

Francesco Orabona. A modern introduction to online learning. *arXiv preprint arXiv:1912.13213*, 2019.

Tor Lattimore and Csaba Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020.

Victor Chernozhukov, Denis Chetverikov, and Kengo Kato. Gaussian approximations and multiplier bootstrap for maxima of sums of high-dimensional random vectors. *Ann. Stat.*, 2013.

Appendix A

APPENDIX TO CHAPTER 2

Appendix

In the appendix we present algorithms and proofs not included in the main text. Broadly speaking,

- Section A presents proofs for lower bounds;
- Section B presents proofs for the proposed computationally inefficient algorithms 1 and 2;
- Section C presents results to justify the computational efficiency of Algorithm 4;
- Section D presents arguments for Algorithm 4 hitting the sample complexity lower bound;
- Section E-F provides technical proofs to argue about convergence of our subroutines.

The table below summarises the notations we used in the proof.

A.1 Proof for Results in Section 2.2

A.1.1 Proof of Theorem 2.2.2

We quickly point out that the proof of Theorem 2.2.2 is identical to the proof of the linear policy class case proof of Theorem 2.2.13. Please see that argument below.

$t_a^{(c)}(\pi')$	$\{\mathbf{1}\{\pi(c)=a,\pi'(c)\neq a\}+\mathbf{1}\{\pi(c)\neq a,\pi'(c)=a\}\}_{\pi\in\Pi}\in\mathbb{R}^{\Pi}$
S_ℓ	$\{\pi\in\Pi:\langle\phi_{\pi_*}-\phi_\pi,\theta^*\rangle=V(\pi_*)-V(\pi)=\Delta(\pi,\pi_*)\leq\epsilon_\ell\}$
$w(\lambda,\gamma)$	$[w(\lambda,\gamma)]_{a,c}=\nu_c\cdot p_{c,a}=\nu_c\cdot\frac{\sqrt{(\lambda\odot\gamma)^\top(t_a^{(c)}+\eta)}}{\sum_{a'\in\mathcal{A}}\sqrt{(\lambda\odot\gamma)^\top(t_{a'}^{(c)}+\eta)}}$
$\widehat{\Delta}_l^\gamma(\pi,\pi')$	$\sum_{s=1}^{n_l}\frac{r_s}{p^{(\ell)}_{c_s,a_s}+\gamma_\pi}(\mathbf{1}\{\pi'(c_s)=a_s\}-\mathbf{1}\{\pi(c_s)=a_s\})$
$h_l(\lambda,\gamma,n)$	$\sum_{\pi\in\Pi}\lambda_\pi\cdot\left(-\widehat{\Delta}_{l-1}^{\gamma^{l-1}}(\pi,\widehat{\pi}_{l-1})+\frac{\log(1/\delta_l)}{\gamma_\pi n}\right)$ $+\gamma_\pi\mathbb{E}_{c\sim\nu_{\mathcal{D}}}\left[\left(\sum_{a\in\mathcal{A}}\sqrt{(\lambda\odot\gamma)^\top(t_a^{(c)}+\eta_l)}\right)^2\right]$
$\mathcal{P}_l(w,\gamma)$	$\max_{\pi\in\Pi}\left(-\widehat{\Delta}_{l-1}^{\gamma^{l-1}}(\pi,\widehat{\pi}_{l-1})+\gamma\left\|\phi_\pi-\phi_{\widehat{\pi}_{l-1}}\right\|^2_{A(w)^{-1}}+\frac{\log(1/\delta)}{\gamma n_l}\right)$

Table A.1: Glossary

A.1.2 Proof of Theorem 2.2.6

Proof of Theorem 2.2.6 . To relate the random stopping time to the regret bound, note that

$$\sum_{c,a}\mathbb{E}_\mu[T_{c,a}(\tau)](r(c,\pi_*(c))-r(c,a))\leq\mathbb{E}_\mu\left[\sqrt{\alpha\left|\mathcal{A}\right|\tau}\right]\leq\sqrt{\alpha|\mathcal{A}|\mathbb{E}_\mu[\tau]}$$

where the last inequality follows by Jensen's inequality. Since $\pi_1:=\pi_*$ for our particular instance, if $\bar{c}=\arg\min_{c\in[m]}\mathbb{E}_\mu[T_{c,\pi_c(c)}(\tau)]$ then

$$\begin{aligned}\sum_{c,a}\mathbb{E}_\mu[T_{c,a}(\tau)](r(c,\pi_1(c))-r(c,a))&=\sum_{c,a}\mathbb{E}_\mu[T_{c,a}(\tau)]\Delta\mathbf{1}\{a\neq\pi_1(c)\}\\&\geq\sum_c\max_a\mathbb{E}_\mu[T_{c,a}(\tau)]\Delta\mathbf{1}\{a\neq\pi_1(c)\}\\&\geq m\min_c\max_a\mathbb{E}_\mu[T_{c,a}(\tau)]\Delta\mathbf{1}\{a\neq\pi_1(c)\}\\&=m\mathbb{E}_\mu[T_{\bar{c},\pi_{\bar{c}}(\bar{c})}(\tau)]\Delta.\end{aligned}$$

Combining the two equations above, and rearranging, we observe that

$$\mathbb{E}_\mu[T_{\bar{c},\pi_{\bar{c}}(\bar{c})}(\tau)]\leq\frac{1}{m\Delta}\sqrt{\alpha|\mathcal{A}|\mathbb{E}_\mu[\tau]}.$$

Define an instance $\mu'=(\nu,r')$ such that $r'(c,a)=r(c,a)$ for all $(c,a)\in[m]\times\{0,1\}\setminus(\bar{c},1)$, and set $r'(\bar{c},1)=r'(\bar{c},\pi_{\bar{c}}(\bar{c}))=2\Delta$ under μ' (instead of $r(\bar{c},\pi_{\bar{c}}(\bar{c}))=0$ under μ). Note that

under μ', we now have that $\pi_{\bar{c}}$ is the unique optimal policy. If the algorithm is $(0,\delta)$-PAC then by [Kaufmann et al., 2016, Lemma 1] we have that

$$\begin{aligned}\log(1/2.4\delta) &\leq \sum_{c,a} KL(\mathcal{N}(r(c,a),1)|\mathcal{N}(r'(c,a),1)) \cdot \mathbb{E}_\mu[T_{c,a}(\tau)] \\ &= KL(\mathcal{N}(0,1)|\mathcal{N}(2\Delta,1)) \cdot \mathbb{E}_\mu[T_{\bar{c},\pi_{\bar{c}}(\bar{c})}(\tau)] = 2\Delta^2 \cdot \mathbb{E}_\mu[T_{\bar{c},\pi_{\bar{c}}(\bar{c})}(\tau)] \\ &\leq 2\Delta^2 \cdot \frac{1}{m\Delta}\sqrt{\alpha|\mathcal{A}|\mathbb{E}_\mu[\tau]} = \sqrt{\frac{4\alpha\mathbb{E}_\mu[\tau]}{m^2\Delta^{-2}}}.\end{aligned}$$

The result follows by rearranging. □

A.1.3 Trivial Class: Proof of Theorem 2.2.9

Firstly note that

$$\begin{aligned}\rho_{\Pi,0}(\Pi,v) &= \min_{p_c\in\triangle_\mathcal{A},\ \forall c\in\mathcal{C}} \max_{\pi\in\Pi\setminus\pi_*} \frac{\mathbb{E}_{c\sim\nu}\left[\left(\frac{1}{p_{c,\pi(c)}}+\frac{1}{p_{c,\pi_*(c)}}\right)\mathbf{1}\{\pi_*(c)\neq\pi(c)\}\right]}{(\mathbb{E}_{c\sim\nu}[\,r(c,\pi_*(c))-r(c,\pi(c))\,])^2} \\ &= \min_{p_c\in\triangle_\mathcal{A},\ \forall c\in\mathcal{C}} \max_{\pi\in\Pi\setminus\pi_*} \frac{\sum_{c\in\mathcal{C}}\nu_c\left(\frac{1}{p_{c,\pi(c)}}+\frac{1}{p_{c,\pi_*(c)}}\right)\mathbf{1}\{\pi_*(c)\neq\pi(c)\}}{(\sum_{c\in\mathcal{C}}\nu_c\Delta_{c,\pi(c)}\mathbf{1}\{\pi_*(c)\neq\pi(c)\})^2} \\ &= \min_{p_c\in\triangle_\mathcal{A},\ \forall c\in\mathcal{C}} \max_{\substack{\alpha\in\{0,1\}^{|\mathcal{C}|\times|\mathcal{A}|}\setminus\mathbf{0}:\\ \sum_a\alpha_{c,a}\in\{0,1\}}} \frac{\sum_{c,a}\alpha_{c,a}\nu_c\left(\frac{1}{p_{c,\pi(c)}}+\frac{1}{p_{c,\pi_*(c)}}\right)\mathbf{1}\{\pi_*(c)\neq a\}}{(\sum_{c,a}\alpha_{c,a}\nu_c\Delta_{c,\pi(c)}\mathbf{1}\{\pi_*(c)\neq\pi(c)\})^2} \\ &= \min_{p_c\in\triangle_\mathcal{A},\ \forall c\in\mathcal{C}} \max_{c,a:\pi_*(c)\neq a} \frac{\nu_c\left(\frac{1}{p_{c,a}}+\frac{1}{p_{c,\pi_*(c)}}\right)}{(\nu_c\Delta_{c,a})^2} \\ &\leq \max_c \frac{2}{\nu_c}\sum_{a'}\Delta_{c,a'}^{-2}\end{aligned}$$

where the last equality follows from repeated application of the inequality $\frac{a_1+a_2}{(b_1+b_2)^2} \leq \frac{a_1}{b_1^2} \vee \frac{a_2}{b_2^2}$.

Proof of Theorem 2.2.9. The proof of the instance-dependent lower bound for $\epsilon = 0$ follows directly from Theorem 2.2.2. The second minimax statement is, to our best knowledge, novel.

First, note that $\sup_\mu \mathbb{E}_\mu[\tau] \geq \epsilon^{-2}|\mathcal{A}|\log(1/\delta)$ by a reduction to multi-armed bandits by just setting $\nu_1 = 1$ and $\nu_c = 0$ for all $c \neq 1$ [Mannor and Tsitsiklis, 2003, Kaufmann et al., 2016]. If U denotes the set of instances that achieves this supremum, and V is another set of

instances, we note that $\sup_\mu \mathbb{E}_\mu[\tau] = \sup_P \mathbb{E}_{\mu\sim P}\mathbb{E}_\mu[\tau] \geq \frac{1}{2}\sup_{\mu\in U}\mathbb{E}_\mu[\tau] + \frac{1}{2}\sup_{\mu\in V}\mathbb{E}_\mu[\tau]$ for some other set of instances V. Thus, it remains to show that $\sup_\mu \mathbb{E}_\mu[\tau] \geq \epsilon^{-2}|\mathcal{A}|\cdot|\mathcal{C}|$.

Consider the following construction of $|\Pi| = |\mathcal{A}|^{|\mathcal{C}|}$ instances. For each context $c \in \mathcal{C}$ let $\nu_c = 1/|\mathcal{C}|$, and for each $\pi \in \Pi$ let $r_\pi(c,a) = \alpha\epsilon\mathbf{1}\{\pi(c) = a\}$ for some $\alpha > 0$ to be determined later. Clearly, policy π is the unique optimal policy under the reward function $r_\pi(s,a)$. Assume that observations are perturbed by Gaussian $\mathcal{N}(0,1)$ noise.

Fix $p \in (1/2, 1)$ to be determined later. Let $S := \{c \in \mathcal{C} : \mathbb{P}_{\mu_\pi}(\pi(c) = \widehat{\pi}(c)) > p\}$ and suppose $|S| \leq |\mathcal{C}|/8$. Then

$$\begin{aligned}
\mathbb{P}_{\mu_\pi}(V(\pi) - V(\widehat{\pi}) \leq \epsilon) &= \mathbb{P}_{\mu_\pi}(\frac{1}{|\mathcal{C}|}\sum_{c\in\mathcal{C}}\alpha\epsilon\mathbf{1}\{\widehat{\pi}(c) \neq \pi(c)\} \leq \epsilon) \\
&= \mathbb{P}_{\mu_\pi}(\sum_{c\in\mathcal{C}}\mathbf{1}\{\widehat{\pi}(c) \neq \pi(c)\} \leq |\mathcal{C}|/\alpha) \\
&= \mathbb{P}_{\mu_\pi}(\sum_{c\in\mathcal{C}}\mathbf{1}\{\widehat{\pi}(c) = \pi(c)\} \geq |\mathcal{C}|(1 - 1/\alpha)) \\
&\leq \mathbb{P}_{\mu_\pi}(\sum_{c\in\mathcal{C}\setminus S}\mathbf{1}\{\widehat{\pi}(c) = \pi(c)\} \geq |\mathcal{C}|(1 - 1/\alpha - 1/8)) \\
&\leq \frac{\sum_{c\in\mathcal{C}\setminus S}\mathbb{P}_{\mu_\pi}(\widehat{\pi}(c) = \pi(c))}{|\mathcal{C}|(1 - 1/\alpha - 1/8)} \leq \frac{p}{1 - 1/\alpha - 1/8} \leq 5/6
\end{aligned}$$

with $p = 5/8$ and $\alpha = 8$. This implies that for $\delta \in (0, 1/8)$, any (ϵ,δ)-PAC algorithm must satisfy $\min_\pi |\{c \in \mathcal{C} : \mathbb{P}_{\mu_\pi}(\pi(c) = \widehat{\pi}(c)) > p\}| \geq |\mathcal{C}|/8$.

Assume the algorithm is permutation invariant (note that any reasonable algorithm satisfies this, including UCB, Thompson Sampling, elimination, etc.). Let $\mu_\pi^{(i)} = (\nu, r_0)$ where $r_\pi^{(i)}(c,i) = r_\pi^{(i)}(c,\pi(c)) = \alpha\epsilon$, and $r_\pi^{(i)}(c,j) = 0$ for $j \notin \{i, \pi(c)\}$. Note that $\mathbb{P}_{\mu_\pi}(\pi(c) = \widehat{\pi}(c)) \geq p = 5/6$ and also by the symmetric algorithm assumption that $\mathbb{P}_{\mu_\pi^{(i)}}(\pi(c) = \widehat{\pi}(c)) \leq 1/2$ because there are two identical best-arms. Note that $\sum_{j\in\mathcal{A}}\mathbb{E}_{\mu_\pi}[T_{c,j}]KL(\mu_\pi(j), \mu_\pi^{(i)}(j)) = \mathbb{E}_{\mu_\pi}[T_{c,i}]\alpha^2\epsilon^2/2$ for $i \neq \pi(c)$. Putting these two pieces together and applying Lemma 1

of [Kaufmann et al., 2016], we have:

$$\begin{aligned}\mathbb{E}_{\mu_\pi}[T_{c,i}]\alpha^2\epsilon^2/2 &= \sum_{j\in\mathcal{A}} \mathbb{E}_{\mu_\pi}[T_{c,j}] KL(\mu_\pi(j), \mu_\pi^{(i)}(j)) \\ &\geq d(\mathbb{P}_{\mu_\pi}(\pi(c) = \widehat{\pi}(c)), \mathbb{P}_{\mu_\pi^{(i)}}(\pi(c) = \widehat{\pi}(c))) \\ &\geq d(5/6, 1/2) = \frac{1}{6}\log(5^5/3^6) \geq 1/10.\end{aligned}$$

Thus, $\mathbb{E}_{\mu_\pi}[\sum_{i\neq\pi_*(c)} T_{c,i}] \geq \frac{1}{5}\alpha^{-2}\epsilon^{-2}(|\mathcal{A}|-1)$ and this must occur on at least $|\mathcal{C}|/8$ contexts. Pick one context c of these arbitrarily. Then

$$\frac{1}{5}\alpha^{-2}\epsilon^{-2}(|\mathcal{A}|-1) \leq \mathbb{E}_{\mu_\pi}[\sum_{i\neq\pi_*(c)} T_{c,i}] = \mathbb{E}_{\mu_\pi}[\sum_{t=1}^{\tau} \mathbf{1}\{c_t = c\}] = \mathbb{E}_{\mu_\pi}[\tau]\nu_c = \mathbb{E}_{\mu_\pi}[\tau]/|\mathcal{C}|.$$

Consequently, $\mathbb{E}[\tau] \geq \frac{1}{5}\alpha^{-2}\epsilon^{-2}(|\mathcal{A}|-1)|\mathcal{C}|$.

□

A.1.4 Proofs of Linear Policy Class

We begin by defining a quantity fundamental to our sample complexity results:

$$\rho_{\mathsf{lin},\epsilon} := \min_{p_c\in\triangle_{\mathcal{A}}, \forall c\in\mathcal{C}} \max_{\pi\in\Pi\setminus\pi_*} \frac{\|\phi_\pi - \phi_{\pi_*}\|^2_{\mathbb{E}_{c\sim\nu}[\sum_{a\in\mathcal{A}} p_{c,a}\phi(c,a)\phi(c,a)^\top]^{-1}}}{\langle\phi_{\pi_*} - \phi_\pi, \theta_*\rangle^2 \vee \epsilon^2}. \tag{A.1}$$

We quickly point out that the proof of Theorem 2.2.2 is identical to the proof of the linear policy class case proof of Theorem 2.2.13.

Proof of Theorem 2.2.13. For any $\theta \in \mathbb{R}^d$ let $\mathbb{P}_\theta(\cdot)$ and $\mathbb{E}_\theta[\cdot]$ denote the probability and expectation laws under θ and ν such that $c_t \sim \nu$ and playing action $a_t \in \mathcal{A}$ results in reward $r_t \sim \mathcal{N}(\langle\phi(c_t, a_t), \theta\rangle, 1)$. If an algorithm is $(0,\delta)$-PAC then $\sup_{\theta\in\mathbb{R}^d} \mathbb{P}_\theta(V(\widehat{\pi}(c)) < V(\pi_*(c))) \leq \delta$. Now, of course, under θ we have that

$$\begin{aligned}V(\widehat{\pi}(c)) < V(\pi_*(c)) &\iff \mathbb{E}_{c\sim\nu}[\langle\theta, \phi(c,\widehat{\pi}(c)) - \phi(c,\pi_*(c))\rangle] < 0 \\ &\iff \langle\theta, \phi_{\widehat{\pi}} - \phi_{\pi_*}\rangle < 0 \\ &\iff \exists c : \nu_c\langle\theta, \phi(c,\widehat{\pi}(c)) - \phi(c,\pi_*(c))\rangle < 0.\end{aligned}$$

Fix $\theta_* \in \mathbb{R}^d$ and recall that under θ we have that $\pi_*(c) = \arg\max_{a \in \mathcal{A}} \langle \phi(c,a), \theta \rangle$. Fix any $\theta \in \mathbb{R}^d$ and $\max_{c,a} \nu_c \langle \theta, \phi(c,a) - \phi(c, \pi_*(c)) \rangle > 0$. Then by [Kaufmann et al., 2016, Lemma 1] we have that

$$
\begin{aligned}
& d(\mathbb{P}_{\theta_*}(V(\widehat{\pi}) = V(\pi_*)), \mathbb{P}_{\theta}(V(\widehat{\pi}) = V(\pi_*))) \\
& \leq \sum_{c',a'} \mathbb{E}_{\theta_*}[T_{c',a'}(\tau)] KL(\mathcal{N}(\langle \theta_*, \phi(c',a') \rangle, 1) | \mathcal{N}(\langle \theta, \phi(c',a') \rangle, 1)) \\
& = \sum_{c',a'} \mathbb{E}_{\theta_*}[T_{c',a'}(\tau)] \|\theta_* - \theta\|^2_{\phi(c',a')\phi(c',a')^\top} / 2 \\
& = \mathbb{E}_{\theta_*}[\tau] \sum_{c',a'} \frac{\mathbb{E}_{\theta_*}[T_{c',a'}(\tau)]}{\mathbb{E}_{\theta_*}[\tau]} \|\theta_* - \theta\|^2_{\phi(c',a')\phi(c',a')^\top} / 2 \\
& \leq \max_{p_c \in \triangle_{\mathcal{A}}, \forall c \in \mathcal{C}} \mathbb{E}_{\theta_*}[\tau] \sum_{c',a'} \nu_{c'} p_{c',a'} \|\theta_* - \theta\|^2_{\phi(c',a')\phi(c',a')^\top} / 2 \\
& = \max_{p_c \in \triangle_{\mathcal{A}}, \forall c \in \mathcal{C}} \mathbb{E}_{\theta_*}[\tau] \|\theta_* - \theta\|^2_{\mathbb{E}_{c \sim \nu}[\sum_a p_{c,a} \phi(c,a)\phi(c,a)^\top]} / 2
\end{aligned}
$$

where the last inequality follows from Wald's identity:

$$
\sum_{a' \in \mathcal{A}} \mathbb{E}_{\theta_*}[T_{c',a'}(\tau)] = \sum_{a' \in \mathcal{A}} \mathbb{E}_{\theta_*} \left[\sum_{t=1}^{\tau} \mathbf{1}\{a_t = a', c_t = c'\} \right] = \mathbb{E}_{\theta_*} \left[\sum_{t=1}^{\tau} \mathbf{1}\{c_t = c'\} \right] = \mathbb{E}_{\theta_*}[\tau] \nu_{c'}.
$$

Noting that $d(\mathbb{P}_{\theta_*}(V(\widehat{\pi}) = V(\pi_*)), \mathbb{P}_{\theta}(V(\widehat{\pi}) \geq d(1-\delta, \delta) \geq \log(1/2.4\delta)$ and we can minimize over θ, given the conditions, we have that

$$
\begin{aligned}
\log(1/2.4\delta) \leq & \max_{p_c \in \triangle_{\mathcal{A}}, \forall c \in \mathcal{C}} \min_{\theta : \exists c : \nu_c \langle \theta, \phi(c,a) - \phi(c, \pi_*(c)) \rangle > 0} \mathbb{E}_{\theta_*}[\tau] \|\theta_* - \theta\|^2_{\mathbb{E}_{c \sim \nu}[\sum_a p_{c,a} \phi(c,a)\phi(c,a)^\top]} / 2 \\
= & \mathbb{E}_{\theta_*}[\tau] \max_{p_c \in \triangle_{\mathcal{A}}, \forall c \in \mathcal{C}} \min_{\substack{c,a \in \mathcal{C} \times \mathcal{A} \\ \pi_*(c) \neq a}} \frac{\langle \phi(c, \pi_*(c)) - \phi(c,a), \theta_* \rangle^2}{2 \|\phi(c,a) - \phi(c, \pi_*(c))\|_{\mathbb{E}_{c \sim \nu}[\sum_a p_{c,a} \phi(c,a)\phi(c,a)^\top]^{-1}}}.
\end{aligned}
$$

After rearranging we conclude that

$$
\mathbb{E}_{\theta_*}[\tau] \geq \min_{p_c \in \triangle_{\mathcal{A}}, \forall c \in \mathcal{C}} \max_{\substack{c,a \in \mathcal{C} \times \mathcal{A} \\ \pi_*(c) \neq a}} \frac{2 \|\phi(c,a) - \phi(c, \pi_*(c))\|_{\mathbb{E}_{c \sim \nu}[\sum_a p_{c,a} \phi(c,a)\phi(c,a)^\top]^{-1}}}{\langle \phi(c, \pi_*(c)) - \phi(c,a), \theta_* \rangle^2} \log(1/2.4\delta).
$$

To see that equation (A.1) is a lower bound, follow the exact same sequence of steps but taking any $\theta \in \mathbb{R}^d$ and $\max_{\pi \in \Pi} \mathbb{E}_{c \sim \nu}[\langle \theta, \phi(c, \pi(c)) - \phi(c, \pi_*(c)) \rangle] > 0$. $\square$

A.1.5 Proof for Corollary 2.2.16

Proof. Observe that

$$\begin{aligned}
\rho_{\Pi,\epsilon_0} &:= \min_{p_c\in\triangle_{\mathcal{A}},\ \forall c\in\mathcal{C}} \max_{\pi\in\Pi\setminus\pi_*} \frac{\mathbb{E}_{c\sim\nu}\left[\left(\frac{1}{p_{c,\pi(c)}}+\frac{1}{p_{c,\pi_*(c)}}\right)\mathbf{1}\{\pi_*(c)\neq\pi(c)\}\right]}{(\mathbb{E}_{c\sim\nu}[\,r(c,\pi_*(c))-r(c,\pi(c))\,]\vee\epsilon_0)^2} \\
&= \min_{p_c\in\triangle_{\mathcal{A}},\ \forall c\in\mathcal{C}} \max_{\epsilon\geq\epsilon_0} \max_{\pi\in\Pi\setminus\pi_*:\Delta(\pi)\leq\epsilon} \frac{\mathbb{E}_{c\sim\nu}\left[\left(\frac{1}{p_{c,\pi(c)}}+\frac{1}{p_{c,\pi_*(c)}}\right)\mathbf{1}\{\pi_*(c)\neq\pi(c)\}\right]}{\epsilon^2} \\
&= \min_{p_c\in\triangle_{\mathcal{A}},\ \forall c\in\mathcal{C}} \max_{\epsilon\geq\epsilon_0} \max_{\pi\in\Pi\setminus\pi_*:\Delta(\pi)\leq\epsilon} \frac{\mathbb{E}_{c\sim\nu}\left[\left(\frac{1}{p_{c,\pi(c)}}+\frac{1}{p_{c,\pi_*(c)}}\right)\mathbf{1}\{\pi_*(c)\neq\pi(c),\Delta(\pi)\leq\epsilon\}\right]}{\epsilon^2} \\
&\leq \min_{p_c\in\triangle_{\mathcal{A}},\ \forall c\in\mathcal{C}} \max_{\epsilon\geq\epsilon_0} \max_{\pi\in\Pi\setminus\pi_*:\Delta(\pi)\leq\epsilon} \frac{\mathbb{E}_{c\sim\nu}\left[\left(\frac{1}{p_{c,\pi(c)}}+\frac{1}{p_{c,\pi_*(c)}}\right)\mathbf{1}\{\exists\pi\in\Pi:\pi_*(c)\neq\pi(c),\Delta(\pi)\leq\epsilon\}\right]}{\epsilon^2} \\
&\overset{(i)}{\leq} \max_{\epsilon\geq\epsilon_0} \max_{\pi\in\Pi\setminus\pi_*:\Delta(\pi)\leq\epsilon} \frac{\mathbb{E}_{c\sim\nu}\left[(|\mathcal{A}|+|\mathcal{A}|)\,\mathbf{1}\{\exists\pi\in\Pi:\pi_*(c)\neq\pi(c),\Delta(\pi)\leq\epsilon\}\right]}{\epsilon^2} \\
&= \max_{\epsilon\geq\epsilon_0} \frac{2|\mathcal{A}|\mathbb{E}_{c\sim\nu}\left[\mathbf{1}\{\exists\pi\in\Pi:\pi_*(c)\neq\pi(c),\Delta(\pi)\leq\epsilon\}\right]}{\epsilon^2} \leq \frac{2|\mathcal{A}|}{\epsilon_0}\mathfrak{C}^{\mathsf{csc}}_{\Pi}(\epsilon_0),
\end{aligned}$$

where (i) follows from taking $p_c \in \triangle_{\mathcal{A}}$ to be the uniform distribution over all actions for each $c \in \mathcal{C}$. To relate this to the policy disagreement coefficient, note that

$$\begin{aligned}
\Delta(\pi) = \mathbb{E}_{c\sim\nu}[r(c,\pi_*(c))-r(c,\pi(c))] &\geq \mathbb{E}_{c\sim\nu}[\mathbf{1}\{\pi(c)\neq\pi_*(c)\}(\min_{c\in\mathcal{C}}\min_{a\in\mathcal{A}} r(c,\pi_*(c))-r(c,a))] \\
&= \mathbb{P}_\nu(\pi(c)\neq\pi_*(c))\Delta_{\mathsf{uniform}}.
\end{aligned}$$

Therefore,

$$\begin{aligned}
&\max_{\epsilon\geq\epsilon_0} \frac{2|\mathcal{A}|\mathbb{E}_{c\sim\nu}\left[\mathbf{1}\{\exists\pi\in\Pi:\pi_*(c)\neq\pi(c),\Delta(\pi)\leq\epsilon\}\right]}{\epsilon^2} \\
&\leq \max_{\epsilon\geq\epsilon_0} \frac{2|\mathcal{A}|\mathbb{E}_{c\sim\nu}\left[\mathbf{1}\{\exists\pi\in\Pi:\pi_*(c)\neq\pi(c),\mathbb{P}_\nu(\pi(c)\neq\pi_*(c))\leq\frac{\epsilon}{\Delta_{\mathsf{uniform}}}\}\right]}{\epsilon^2} \\
&\leq \frac{2|\mathcal{A}|}{\epsilon_0\Delta_{\mathsf{uniform}}}\mathfrak{C}^{\mathsf{pol}}_{\Pi}(\epsilon_0/\Delta_{\mathsf{uniform}}).
\end{aligned}$$

□

A.2 Proof for sample complexity of Algorithm 1 and 2

Proof of Lemma 2.3.1. For any $\mathcal{V} \subseteq \Pi$ and $\pi \in \mathcal{V}$ define the event

$$\mathcal{E}_{\pi,\ell}(\mathcal{V}) = \{|\widehat{o}_{\pi_*,\pi,\ell}(\mathcal{V}) - \langle \phi_{\pi_*} - \phi_\pi, \theta_* \rangle| \le \epsilon_\ell\}$$

where it is implicit that $\widehat{o}_{\pi_*,\pi,\ell} := \widehat{o}_{\pi_*,\pi,\ell}(\mathcal{V})$ is the resulting estimate after round ℓ if Π_ℓ had been equal to $\mathcal{V}$. Define $w_\ell(\mathcal{V})$ and $\tau_\ell(\mathcal{V})$ analogously. By the properties of the Catoni estimator, we have for any $\mathcal{V} \subset \Pi$ with probability at least $1 - \frac{\delta}{2\ell^2|\Pi|}$ that

$$\begin{aligned}
|\widehat{o}_{\pi_*,\pi,\ell}(\mathcal{V}) - \langle \phi_{\pi_*} - \phi_\pi, \theta_* \rangle| &\le \|\phi_{\pi_*} - \phi_\pi\|_{A(w_\ell(\mathcal{V}))^{-1}} \sqrt{\frac{2\log(2\ell^2|\Pi|/\delta)}{\tau_\ell(\mathcal{V}) - \log(2\ell^2|\Pi|/\delta)}} \\
&\le \sqrt{\frac{\|\phi_{\pi_*} - \phi_\pi\|^2_{A(w_\ell(\mathcal{V}))^{-1}}}{2\epsilon_\ell^{-2}\rho(w_\ell(\mathcal{V}), \mathcal{V})\log(2\ell^2|\Pi|/\delta)}} \sqrt{2\log(2\ell^2|\Pi|/\delta)} = \epsilon_\ell.
\end{aligned}$$

Consequently,

$$\begin{aligned}
\mathbb{P}\left(\bigcup_{\ell=1}^{\infty}\bigcup_{\pi\in\Pi_\ell}\{\mathcal{E}^c_{\pi,\ell}(\Pi_\ell)\}\right) &\le \sum_{\ell=1}^{\infty}\mathbb{P}\left(\bigcup_{\pi\in\Pi_\ell}\{\mathcal{E}^c_{\pi,\ell}(\Pi_\ell)\}\right) \\
&= \sum_{\ell=1}^{\infty}\sum_{\mathcal{V}\subseteq\Pi}\mathbb{P}\left(\bigcup_{\pi\in\mathcal{V}}\{\mathcal{E}^c_{\pi,\ell}(\mathcal{V})\}, \Pi_\ell = \mathcal{V}\right) \\
&= \sum_{\ell=1}^{\infty}\sum_{\mathcal{V}\subseteq\Pi}\mathbb{P}\left(\bigcup_{\pi\in\mathcal{V}}\{\mathcal{E}^c_{\pi,\ell}(\mathcal{V})\}\right)\mathbb{P}(\Pi_\ell = \mathcal{V}) \\
&\le \sum_{\ell=1}^{\infty}\sum_{\mathcal{V}\subseteq\Pi}\tfrac{\delta|\mathcal{V}|}{2\ell^2|\Pi|}\mathbb{P}(\Pi_\ell = \mathcal{V}) \le \delta.
\end{aligned}$$

Thus, assume $\bigcap_{\ell=1}^{\infty}\bigcap_{\pi\in\Pi_\ell}\{\mathcal{E}_{\pi,\ell}(\Pi_\ell)\}$ holds. For any $\pi \in \Pi_\ell$ we have

$$\begin{aligned}
\widehat{o}_{\pi,\pi_*,\ell} &= \widehat{o}_{\pi,\pi_*,\ell} - \langle \phi_\pi - \phi_{\pi_*}, \theta_* \rangle + \phi_{\pi_*}, \theta_* \rangle \\
&\le \epsilon_\ell + \langle \phi_\pi - \phi_{\pi_*}, \theta_* \rangle \le \epsilon_\ell
\end{aligned}$$

which implies that π_* would survive to round $\ell+1$. And for any $\pi' \in \Pi_\ell$ such that $\langle \phi_{\pi_*} - \phi_{\pi'}, \theta^* \rangle > 2\epsilon_\ell$ we have

$$\begin{aligned}\max_{\pi\in\Pi_\ell} \widehat{o}_{\pi,\pi',\ell} &\geq \widehat{o}_{\pi_*,\pi',\ell} \\ &= \langle \phi_{\pi'} - \phi_{\pi_*}, \theta_* \rangle - \widehat{o}_{\pi',\pi_*,\ell} + \langle \phi_{\pi_*} - \phi_{\pi'}, \theta_* \rangle \\ &> -\epsilon_\ell + 2\epsilon_\ell = \epsilon_\ell\end{aligned}$$

which implies this π' would be kicked out. Note that this implies that $\max_{\pi\in\Pi_{\ell+1}} \langle \phi_{\pi_*} - \phi_\pi, \theta^* \rangle \leq 2\epsilon_\ell = 4\epsilon_{\ell+1}$. □

In the remaining of this section we provide a proof for the sample complexity of Algorithm 2.

Theorem A.2.1. *Under $\mathcal{E}$, for all $\ell \in \mathbb{N}$, the following holds:*

1. $\widehat{\pi}_\ell \in S_\ell := \{\pi \in \Pi : V(\pi_*) - V(\pi) \leq \epsilon_\ell\}$;
2. $n_\ell \lesssim \min_{w\in\Omega} \max_{\pi\in\Pi} \frac{\|\phi_{\pi_*} - \phi_\pi\|^2_{A(w)^{-1}} \log(1/\delta_l)}{\epsilon_l^2 + \Delta(\pi)^2}$.

Without loss of generality, we assume that $\forall t$, the reward $r_t \in [0,1]$. Note that by the result about Catoni estimator in [Lugosi and Mendelson, 2019], we have for all $\ell \in \mathbb{N}$ and $\pi, \pi' \in \Pi$, that

$$|\mathsf{Cat}(\{\langle \phi_\pi - \phi_{\pi'}, O_t \rangle\}_{t=1}^{n_\ell}) - \langle \phi_\pi - \phi_{\pi'}, \theta_* \rangle| \leq \|\phi_\pi - \phi_{\pi'}\|_{A(w^{(\ell)})^{-1}} \sqrt{\frac{2\log(2\ell^2|\Pi|/\delta)}{n_\ell - \log(2\ell^2|\Pi|/\delta)}}.$$

Therefore, in the ℓth round, we have for any $\pi, \pi' \in \Pi$,

$$\begin{aligned}\left|\widehat{\Delta}_l(\pi,\pi') - \Delta(\pi,\pi')\right| &= |\mathsf{Cat}(\{\langle \phi_\pi - \phi_{\pi'}, O_i \rangle\}_{i=1}^{n_\ell}) - \langle \phi_\pi - \phi_{\pi'}, \theta_* \rangle| \\ &\leq \sqrt{\frac{2\|\phi_\pi - \phi_{\pi'}\|^2_{A(w^{(\ell)})^{-1}} \log(2\ell^2|\Pi|/\delta)}{n_\ell}}. \qquad \text{(A.2)}\end{aligned}$$

Then, let $\delta_l = \frac{\delta}{2l^2|\Pi|}$ we define the event

$$\mathcal{E}_l = \bigcap_{\pi,\pi'\in\Pi} \left\{ \left|\widehat{\Delta}_l(\pi,\pi') - \Delta(\pi,\pi')\right| \leq \sqrt{\frac{2\|\phi_\pi - \phi_{\pi'}\|^2_{A(w^{(\ell)})^{-1}} \log(1/\delta_l)}{n_\ell}} \right\},$$

and $\mathcal{E} = \bigcap_{l=0}^{\infty} \mathcal{E}_l$. First, by equation A.2, we have that $\mathcal{E}$ happens with probability at least $1-\delta$. In order to show the sample complexity lower bound, we use proof by induction. Note that in a step of Lemma A.2.4, we can show that $n_l \lesssim \min_{w\in\Omega} \max_{\pi\in\Pi} \frac{\left\|\phi_{\widehat{\pi}_{l-1}} - \phi_\pi\right\|^2_{A(w)^{-1}} \log(1/\delta_l)}{\epsilon_l^2 + \Delta(\pi)^2}$, so we induct on this result. Assume in round $l-1$, $\widehat{\pi}_{l-1} \in S_{l-1} = \{\pi \in \Pi : \Delta(\pi, \pi_*) \le \epsilon_{l-1}\}$ and $n_{l-1} \lesssim \min_{w\in\Omega} \max_{\pi\in\Pi} \frac{\left\|\phi_{\widehat{\pi}_{l-2}} - \phi_\pi\right\|^2_{A(w)^{-1}} \log((l-1)^2|\Pi|^2/\delta)}{\epsilon_{l-1}^2 + \Delta(\pi)^2}$. Then, the following lemma gives us an upper bound on the UCB.

Lemma A.2.2. *We have for any $\pi \in \Pi$,*

$$\sqrt{\frac{\|\phi_{\widehat{\pi}_l} - \phi_\pi\|^2_{A(w^{(\ell)})^{-1}} \log(1/\delta_l)}{n_l}} \le \frac{1}{28}\left(4\epsilon_l + \widehat{\Delta}_{l-1}(\pi, \widehat{\pi}_{l-1})\right).$$

Proof. By definition of n_l and $w^{(\ell)}$ and $\pi^{(\ell)}$ being the saddle point, we have

$$\begin{aligned}
&-\frac{1}{4}\widehat{\Delta}_{l-1}(\pi^{(\ell)}, \widehat{\pi}_{l-1}) + 28\sqrt{\frac{2\|\phi_{\pi^{(\ell)}} - \phi_{\widehat{\pi}_{l-1}}\|^2_{A(w^{(\ell)})^{-1}} \log(1/\delta_l)}{n_\ell}} \\
&= \max_{\pi\in\Pi} -\frac{1}{4}\widehat{\Delta}_{l-1}(\pi, \widehat{\pi}_{l-1}) + \sqrt{\frac{1568\|\phi_\pi - \phi_{\widehat{\pi}_{l-1}}\|^2_{A(w^{(\ell)})^{-1}} \log(1/\delta_l)}{n_\ell}} \le \epsilon_l.
\end{aligned}$$

Solving for n_l gives us

$$n_l \ge \max_{\pi\in\Pi} \frac{1568\left\|\phi_\pi - \phi_{\widehat{\pi}_{l-1}}\right\|^2_{A(w^{(\ell)})^{-1}} \log(1/\delta_l)}{(4\epsilon_l + \widehat{\Delta}_{l-1}(\pi, \widehat{\pi}_{l-1}))^2}.$$

We have for any $\pi \in \Pi$,

$$\begin{aligned}
2n_l &\ge 3136 \max_{\pi\in\Pi} \frac{\left\|\phi_{\widehat{\pi}_{l-1}} - \phi_\pi\right\|^2_{A(w^{(\ell)})^{-1}} \log(1/\delta_l)}{(4\epsilon_l + \widehat{\Delta}_{l-1}(\pi, \widehat{\pi}_{l-1}))^2} \\
&\ge 1568 \frac{\left\|\phi_{\widehat{\pi}_{l-1}} - \phi_\pi\right\|^2_{A(w^{(\ell)})^{-1}} \log(1/\delta_l)}{(4\epsilon_l + \widehat{\Delta}_{l-1}(\pi, \widehat{\pi}_{l-1}))^2} \\
&\quad + 1568 \frac{\left\|\phi_{\widehat{\pi}_{l-1}} - \phi_{\widehat{\pi}_l}\right\|^2_{A(w^{(\ell)})^{-1}} \log(1/\delta_l)}{(4\epsilon_l + \widehat{\Delta}_{l-1}(\widehat{\pi}_l, \widehat{\pi}_{l-1}))^2} \\
&\overset{(i)}{\ge} 1568 \frac{\left(\left\|\phi_{\widehat{\pi}_{l-1}} - \phi_\pi\right\|^2_{A(w^{(\ell)})^{-1}} + \left\|\phi_{\widehat{\pi}_{l-1}} - \phi_{\widehat{\pi}_l}\right\|^2_{A(w^{(\ell)})^{-1}}\right) \log(1/\delta_l)}{\max\{(4\epsilon_l + \widehat{\Delta}_{l-1}(\widehat{\pi}_l, \widehat{\pi}_{l-1}))^2, (4\epsilon_l + \widehat{\Delta}_{l-1}(\pi, \widehat{\pi}_{l-1}))^2\}} \\
&\overset{(ii)}{\ge} 1568 \frac{\|\phi_{\widehat{\pi}_l} - \phi_\pi\|^2_{A(w^{(\ell)})^{-1}} \log(1/\delta_l)}{\max\{(4\epsilon_l + \widehat{\Delta}_{l-1}(\widehat{\pi}_l, \widehat{\pi}_{l-1}))^2, (4\epsilon_l + \widehat{\Delta}_{l-1}(\pi, \widehat{\pi}_{l-1}))^2\}}.
\end{aligned}$$

where (i) holds by lower bounding the ratio with a larger denominator, and (ii) holds by triangular inequality. Therefore, using the fact that $\widehat{\Delta}(\pi, \widehat{\pi}_{l-1}) \geq 0$ for any $\pi \in \Pi$ since $\widehat{\pi}_{l-1} = \arg\max_{\pi\in\Pi} \widehat{V}_{l-1}(\pi)$, we have $\sqrt{\max\{(4\epsilon_l + \widehat{\Delta}_{l-1}(\widehat{\pi}_l, \widehat{\pi}_{l-1}))^2, (4\epsilon_l + \widehat{\Delta}_{l-1}(\pi, \widehat{\pi}_{l-1}))^2\}} = \max\{4\epsilon_l + \widehat{\Delta}_{l-1}(\widehat{\pi}_l, \widehat{\pi}_{l-1}), 4\epsilon_l + \widehat{\Delta}_{l-1}(\pi, \widehat{\pi}_{l-1})\}$, so we have

$$\sqrt{\frac{\|\phi_{\widehat{\pi}_l} - \phi_\pi\|^2_{A(w^{(\ell)})^{-1}} \log(1/\delta_l)}{n_l}} \leq \frac{1}{28}\left(4\epsilon_l + \max\{\widehat{\Delta}_{l-1}(\pi, \widehat{\pi}_{l-1}), \widehat{\Delta}_{l-1}(\widehat{\pi}_l, \widehat{\pi}_{l-1})\}\right).$$

□

With the above results, the following lemma controls the difference between the empirical gap and the true gap.

Lemma A.2.3. *With inductive hypotheses, we have for any $\pi \in \Pi$,*

$$|\widehat{\Delta}_{l-1}(\pi, \widehat{\pi}_{l-1}) - \Delta(\pi, \pi_*)| \leq 2\epsilon_{l-1} + \frac{1}{4}\Delta(\pi, \pi_*).$$

Proof. We prove this by induction. First, in round $l = 0$, this holds by choosing a sufficiently large n_0. Then, in round $l - 1$,

$$\begin{aligned}
&|\widehat{\Delta}_{l-1}(\pi, \widehat{\pi}_{l-1}) - \Delta(\pi, \pi_*)| \\
&= |\widehat{\Delta}_{l-1}(\pi, \widehat{\pi}_{l-1}) - \Delta(\pi, \widehat{\pi}_{l-1}) - \Delta(\widehat{\pi}_{l-1}, \pi_*)| \\
&\leq \sqrt{\frac{2\left\|\phi_\pi - \phi_{\widehat{\pi}_{l-1}}\right\|^2_{A(w^{(\ell-1)})^{-1}} \log(1/\delta_{l-1})}{n_{l-1}}} + \epsilon_{l-1} \\
&\overset{(i)}{\leq} \frac{\sqrt{2}}{28}\left(4\epsilon_{l-1} + \max\{\widehat{\Delta}_{l-2}(\pi, \widehat{\pi}_{l-2}), \widehat{\Delta}_{l-2}(\widehat{\pi}_{l-1}, \widehat{\pi}_{l-2})\}\right) + \epsilon_{l-1} \\
&\overset{(ii)}{\leq} \frac{\sqrt{2}}{28}\left(4\epsilon_{l-1} + 2\epsilon_{l-2} + \frac{5}{4}\Delta(\pi, \widehat{\pi}_{l-2}) + 2\epsilon_{l-2} + \frac{5}{4}\Delta(\widehat{\pi}_{l-1}, \widehat{\pi}_{l-2})\right) + \epsilon_{l-1} \\
&\leq \frac{\sqrt{2}}{28}\left(4\epsilon_{l-1} + 4\epsilon_{l-2} + \frac{5}{4}\Delta(\pi, \pi_*) + \frac{5}{4}\Delta(\widehat{\pi}_{l-1}, \pi_*)\right) + \epsilon_{l-1} \\
&\leq \frac{\sqrt{2}}{28}\left(4\epsilon_{l-1} + 4\epsilon_{l-2} + \frac{5}{4}\Delta(\pi, \pi_*) + \frac{5}{4}\epsilon_{l-1}\right) + \epsilon_{l-1} \\
&\leq 2\epsilon_{l-1} + \frac{1}{4}\Delta(\pi, \pi_*),
\end{aligned}$$

where (i) follows from the preceding lemma and (ii) follows from the inductive hypothesis that

$$|\widehat{\Delta}_{l-2}(\pi, \widehat{\pi}_{l-2}) - \Delta(\pi, \pi_*)| \leq 2\epsilon_{l-2} + \frac{1}{4}\Delta(\pi, \pi_*).$$

□

We make use of these two lemmas to state a lower bound on n_l.

Lemma A.2.4. *Under $\mathcal{E}$, the choice for n_l in the algorithm satisfies*

$$n_l \lesssim \min_{w \in \Omega} \max_{\pi \in \Pi} \frac{\|\phi_{\pi_*} - \phi_\pi\|^2_{A(w)^{-1}} \log(1/\delta_l)}{\epsilon_l^2 + \Delta(\pi)^2}.$$

Proof. By inductive hypothesis on n_{l-1} and under $\mathcal{E}_l$, we have for any $\pi \in \Pi$,

$$\begin{aligned}
\Delta(\pi, \pi_*) &= \Delta(\pi, \widehat{\pi}_{l-1}) + \Delta(\widehat{\pi}_{l-1}, \pi_*) \\
&\overset{(i)}{\leq} \widehat{\Delta}_{l-1}(\pi, \widehat{\pi}_{l-1}) + \sqrt{\frac{2\left\|\phi_{\widehat{\pi}_{l-1}} - \phi_\pi\right\|^2_{A(w^{(\ell-1)})^{-1}} \log((l-1)^2|\Pi|^2/\delta)}{n_{l-1}}} + \epsilon_{l-1} \\
&\overset{(ii)}{\leq} \widehat{\Delta}_{l-1}(\pi, \widehat{\pi}_{l-1}) + \frac{\sqrt{2}}{28}\left(4\epsilon_{l-1} + \widehat{\Delta}_{l-2}(\pi, \widehat{\pi}_{l-2})\right) + \epsilon_{l-1} \\
&\leq \widehat{\Delta}_{l-1}(\pi, \widehat{\pi}_{l-1}) + \frac{\sqrt{2}}{28}\left(4\epsilon_{l-1} + \frac{5}{4}\Delta(\pi, \pi_*) + 2\epsilon_{l-2}\right) + \epsilon_{l-1} \\
&\leq \widehat{\Delta}_{l-1}(\pi, \widehat{\pi}_{l-1}) + \frac{1}{4}\Delta(\pi, \pi_*) + 2\epsilon_{l-1}.
\end{aligned}$$

where (i) follows from $\mathcal{E}_{l-1}$ and (ii) follows from Lemma A.2.2. Therefore,

$$\begin{aligned}
&\min_{w\in\Omega}\max_{\pi\in\Pi} -\frac{1}{4}\widehat{\Delta}_{l-1}(\pi,\widehat{\pi}_{l-1}) + 28\sqrt{\frac{2\left\|\phi_\pi - \phi_{\widehat{\pi}_{l-1}}\right\|^2_{A(w)^{-1}}\log(1/\delta_l)}{n_l}} \\
&\leq \min_{w\in\Omega}\max_{\pi\in\Pi} -\frac{3}{16}\Delta(\pi,\pi_*) + \frac{1}{2}\epsilon_l + 28\sqrt{\frac{2\left\|\phi_\pi - \phi_{\widehat{\pi}_{l-1}}\right\|^2_{A(w)^{-1}}\log(1/\delta_l)}{n_l}} \\
&\leq \min_{w\in\Omega}\max_{\pi\in\Pi}\Big(-\frac{3}{16}\Delta(\pi,\pi_*) + 28\sqrt{\frac{2\left\|\phi_{\pi_*} - \phi_\pi\right\|^2_{A(w)^{-1}}\log(1/\delta_l)}{n_l}} \\
&\qquad + 28\sqrt{\frac{2\left\|\phi_{\pi_*} - \phi_{\widehat{\pi}_{l-1}}\right\|^2_{A(w)^{-1}}\log(1/\delta_l)}{n_l}}\Big) + \frac{1}{2}\epsilon_l \\
&\leq \min_{w\in\Omega}\max_{\pi\in\Pi}\Bigg(-\frac{3}{16}\Delta(\pi,\pi_*) + 28\sqrt{\frac{2\left\|\phi_{\pi_*} - \phi_\pi\right\|^2_{A(w)^{-1}}\log(1/\delta_l)}{n_l}} \\
&\qquad + 28\sqrt{\max_{\pi'\in S_{l-1}}\frac{2\left\|\phi_{\pi_*} - \phi_{\pi'}\right\|^2_{A(w)^{-1}}\log(1/\delta_l)}{n_l}}\Bigg) + \frac{1}{2}\epsilon_l
\end{aligned}$$

which is less than ϵ_l whenever

$$n_l \gtrsim \min_{w\in\Omega}\max_{\pi\in\Pi}\frac{\left\|\phi_{\pi_*} - \phi_\pi\right\|^2_{A(w)^{-1}}\log(1/\delta_l)}{\epsilon_l^2 + \Delta(\pi,\pi_*)^2}.$$

□

Then we finish our first goal. The next goal is to show that $\widehat{\pi}_l \in S_l$.

Lemma A.2.5. *Under $\mathcal{E}_l$, we have $\Delta(\widehat{\pi}_l, \pi_*) \leq \epsilon_l$.*

Proof. On $\mathcal{E}_l$, we have

$$
\begin{aligned}
&\Delta(\widehat{\pi}_l, \widehat{\pi}_{l-1}) \\
&\leq \widehat{\Delta}_l(\widehat{\pi}_l, \widehat{\pi}_{l-1}) + \sqrt{\frac{2\left\|\phi_{\widehat{\pi}_l} - \phi_{\widehat{\pi}_{l-1}}\right\|^2_{A(w^{(\ell)})^{-1}} \log(1/\delta_l)}{n_l}} && \text{(by event } \mathcal{E}_l\text{)} \\
&\leq \widehat{\Delta}_l(\pi_*, \widehat{\pi}_{l-1}) + \sqrt{\frac{2\left\|\phi_{\widehat{\pi}_l} - \phi_{\widehat{\pi}_{l-1}}\right\|^2_{A(w^{(\ell)})^{-1}} \log(1/\delta_l)}{n_l}} && \text{(by minimality of } \widehat{\pi}_l\text{)} \\
&\leq \Delta(\pi_*, \widehat{\pi}_{l-1}) + \sqrt{\frac{2\left\|\phi_{\widehat{\pi}_{l-1}} - \phi_{\pi_*}\right\|^2_{A(w^{(\ell)})^{-1}} \log(1/\delta_l)}{n_l}} + \sqrt{\frac{2\left\|\phi_{\widehat{\pi}_l} - \phi_{\widehat{\pi}_{l-1}}\right\|^2_{A(w^{(\ell)})^{-1}} \log(1/\delta_l)}{n_l}} \\
& && \text{(by event } \mathcal{E}_l\text{)} \\
&\leq \Delta(\pi_*, \widehat{\pi}_{l-1}) + \frac{\sqrt{2}}{28}\left(4\epsilon_l + \widehat{\Delta}_{l-1}(\pi_*, \widehat{\pi}_{l-1}) + 4\epsilon_l + \widehat{\Delta}_{l-1}(\widehat{\pi}_l, \widehat{\pi}_{l-1})\right) && \text{(by Lemma A.2.2)} \\
&\leq \Delta(\pi_*, \widehat{\pi}_{l-1}) + \frac{\sqrt{2}}{28}\left(4\epsilon_l + 2\epsilon_{l-1} + \frac{5}{4}\Delta(\pi_*, \widehat{\pi}_{l-1}) + 4\epsilon_l + 2\epsilon_{l-1} + \frac{5}{4}\Delta(\widehat{\pi}_l, \widehat{\pi}_{l-1})\right) \\
& && \text{(by Lemma A.2.3)} \\
&\leq \Delta(\pi_*, \widehat{\pi}_{l-1}) + \frac{3}{56}\left(8\epsilon_{l-1} + \frac{5}{4}\Delta(\widehat{\pi}_l, \pi_*)\right).
\end{aligned}
$$

Therefore, $\frac{209}{224}\Delta(\widehat{\pi}_l, \pi_*) \leq \frac{6}{7}\epsilon_l$ and $\Delta(\widehat{\pi}_l, \pi_*) \leq \epsilon_l$, so $\widehat{\pi}_l \in S_l$. □

A.3 Proof of the FW-GD *subroutine*

In this section, we aim to prove Theorem 2.3.3. Specifically, Section A.3.1 quantifies the number of oracle calls, and Section A.3.2 quantifies the number of offline data needed in order to approximate the expectation over the context distribution. In particular, the size of the history follows directly from Lemma A.3.5 and A.3.6. We will see that $\eta, \gamma_{\max}, \gamma_{\min}$ all scale at most polynomially on $|\mathcal{A}|$ and ϵ^{-1}. We leave the convergence analysis of the algorithm in Section A.5. In particular, we will see in Theorem A.5.1 that $K_l = \text{poly}(|\mathcal{A}|, \epsilon_l^{-1})$, which shows that the total number of oracle calls is at most $\text{poly}(|\mathcal{A}|, \epsilon^{-1}, \log(1/\delta), \log(|\Pi|))$. Combining all results above gives Theorem 2.3.3.

A.3.1 Proof of computational efficiency

In this section, we address the technical issues on computational efficiency of our algorithm. Fix an iteration t and let K_l be the number of iterations for FW-GD in the lth round.

Lemma A.3.1. *Equation* (2.5) *can be computed with* $(t + T_{l-1})|\mathcal{D}|$ *call to a cost-sensitive classification oracle.*

Proof. We consider the tth iteration of the lth round for some n_r. In this iteration, we compute

$$[\nabla_\lambda h_l(\lambda^t,\gamma^t, n_r)]_\pi = \sum_{i=1}^{n_l} \frac{r_i}{p^{(\ell)}_{c_i,a_i} + [\gamma^{l-1}]_\pi} \left(\mathbf{1}\{\pi(c_i) = a_i\} - \mathbf{1}\{\widehat{\pi}_{l-1}(c_i) = a_i\}\right) + \frac{\log(1/\delta_l)}{[\gamma^t]_\pi n}$$
$$+ \mathbb{E}_{c\sim\nu_\mathcal{D}}\left[\left(\sum_{a\in\mathcal{A}} \sqrt{(\lambda^t \odot \gamma^t)^\top (t^{(c)}_a + \eta_l)}\right)\left(\sum_{a'\in\mathcal{A}} \frac{[\gamma^t]_\pi (t^{(c)}_{a'} + \eta_l)_\pi}{\sqrt{(\lambda^t \odot \gamma^t)^\top (t^{(c)}_{a'} + \eta_l)}}\right)\right].$$

Define $\gamma_0 := \sqrt{\frac{\log(1/\delta_l)}{|\mathcal{A}|^2 \eta_l n_r}}$. Initially, each coordinate of γ^t is γ_0. In round t of the algorithm, at most t coordinates of γ will change, and these coordinates will be in $\mathrm{supp}(\lambda^t)$. Also, for any $j \notin \mathrm{supp}(\lambda^{l-1})$, $\gamma^{l-1}_j = \gamma_0$. Therefore, let $t^{(c)}_a(\cdot, \widehat{\pi}_{l-1}) \in \mathbb{R}^{|\Pi|}$, in round l,

$$\begin{aligned}
&\operatorname*{argmax}_{\pi\in\Pi\setminus(\mathrm{supp}(\lambda^t)\cup\mathrm{supp}(\lambda^{l-1}))} \left[\nabla_\lambda h_l(\lambda^t, \gamma^t, n_r)\right]_\pi \\
&= \operatorname*{argmax}_{\pi\in\Pi\setminus(\mathrm{supp}(\lambda^t)\cup\mathrm{supp}(\lambda^{l-1}))} \sum_{i=1}^{n_l} \frac{r_i}{p^{(\ell)}_{c_i,a_i} + \gamma_0} \mathbf{1}\{\pi(c_i) = a_i\} + \frac{\log(1/\delta_l)}{\gamma_0 n_r} \\
&\quad + \mathbb{E}_{c\sim\nu_\mathcal{D}}\left[\left(\sum_{a\in\mathcal{A}} \sqrt{(\lambda^t \odot \gamma^t)^\top (t^{(c)}_a(\widehat{\pi}_{l-1}) + \eta_l)}\right)\left(\sum_{a'\in\mathcal{A}} \frac{\gamma_0 (t^{(c)}_{a'}(\widehat{\pi}_{l-1}) + \eta_l)_\pi}{\sqrt{(\lambda^t \odot \gamma^t)^\top (t^{(c)}_{a'}(\widehat{\pi}_{l-1}) + \eta_l)}}\right)\right] \\
&= \operatorname*{argmax}_{\pi\in\Pi\setminus(\mathrm{supp}(\lambda^t)\cup\mathrm{supp}(\lambda^{l-1}))} \sum_{i=1}^{n_l} \frac{r_i}{p^{(\ell)}_{c_i,a_i} + \gamma_0} \mathbf{1}\{\pi(c_i) = a_i\} \\
&\quad + \mathbb{E}_{c\sim\nu_\mathcal{D}}\left[\sum_{a'\in\mathcal{A}} \frac{\sum_{a\in\mathcal{A}} \sqrt{(\lambda^t \odot \gamma^t)^\top (t^{(c)}_a(\widehat{\pi}_{l-1}) + \eta_l)}}{\sqrt{(\lambda^t \odot \gamma^t)^\top (t^{(c)}_{a'}(\widehat{\pi}_{l-1}) + \eta_l)}} \gamma_0 t^{(c)}_{a'}(\widehat{\pi}_{l-1})_\pi\right] \\
&= \operatorname*{argmax}_{\pi\in\Pi\setminus(\mathrm{supp}(\lambda^t)\cup\mathrm{supp}(\lambda^{l-1}))} \sum_{i=1}^{n_l+|\mathcal{D}|} L_i(\pi(c_i))
\end{aligned}$$

which is a cost-sensitive classification problem with cost vector

$$L_i(a) = \begin{cases} \frac{r_i}{p^{(\ell)}_{c_i,a_i}+\gamma_0}\mathbf{1}\{a = a_i\} & \text{for } i = 1, \cdots, n_l \\ \left(\frac{\gamma_0}{s_{a,c_i}} + \frac{\gamma_0}{s_{\widehat{\pi}_{l-1}(c_i),c_i}}\right)\mathbf{1}\{a \neq \widehat{\pi}_{l-1}(c_i)\} & \text{for } i = n_l+1, \cdots, n_l + |\mathcal{D}| \end{cases}$$

where $s_{a,c} = \frac{\sqrt{(\lambda^t \odot \gamma^t)^\top (t_a^{(c)}(\widehat{\pi}_{l-1})+\eta_l)}}{\sum_{a' \in \mathcal{A}} \sqrt{(\lambda^t \odot \gamma^t)^\top (t_{a'}^{(c)}(\widehat{\pi}_{l-1})+\eta_l)}}$. Note that $s_{a,c}$ is computable since λ^t has at most t non-zero elements in step t. Then, let $\pi^\sharp := \text{supp}(\lambda^t) \cup \text{supp}(\lambda^{l-1})$, we have

$$\begin{aligned} &\arg\max_{\pi \in \Pi} \left[\nabla_\lambda h_l(\lambda^t, \gamma^t, n_r)\right]_\pi \\ &= \arg\max \left\{ \operatorname*{argmax}_{\pi \in \Pi^\sharp} \left[\nabla_\lambda h_l(\lambda^t, \gamma^t, n_r)\right]_\pi, \operatorname*{argmax}_{\pi \in \Pi \setminus \Pi^\sharp} \left[\nabla_\lambda h_l(\lambda^t, \gamma^t, n_r)\right]_\pi \right\}. \end{aligned}$$

The first piece could be found directly since $\text{supp}(\lambda^t) \cup \text{supp}(\lambda^{l-1}) \leq t + T_{l-1}$. The second piece could be computed with $(t + T_{l-1})|\mathcal{D}|$ calls to a constrained cost-sensitive classification oracle, stated in Lemma A.3.2 below. □

Lemma A.3.2. *For any set $B_t \subset \Pi$, we can compute $\operatorname*{argmax}_{\pi \in \Pi \setminus B_t} [\nabla_\lambda h_l(\lambda^t, \gamma^t, n_r)]_\pi$ using $|B_t| \cdot |\mathcal{D}|$ calls to a constrained cost-sensitive classification oracle defined in Definition 2.2.3.*

Proof. Algorithm 7 below shows that we could compute this argmax via the C-AMO oracle. First, by construction of the algorithm, we have that $\pi_e \notin B_t$, so $\pi_e \in \Pi \setminus B_t$. It remains to show that π_e achieves the maximum. We prove this via contradiction. Assume that there is some other $\pi' \neq \pi_e$ that satisfies $\pi' \notin B_t$ and $\nabla_\lambda[h_l(\lambda, \gamma, n)]_{\pi'} > \nabla_\lambda[h_l(\lambda, \gamma, n)]_{\pi_e}$. By construction of our algorithm, we know that $\nabla_\lambda[h_l(\lambda, \gamma, n)]_{\pi_k}$ is non-increasing in k. We find the largest $0 \leq j \leq i-1$ such that

$$\nabla_\lambda[h_l(\lambda, \gamma, n)]_{\pi_{j+1}} \leq \nabla_\lambda[h_l(\lambda, \gamma, n)]_{\pi'} \leq \nabla_\lambda[h_l(\lambda, \gamma, n)]_{\pi_j}.$$

First, since j is the largest, we have $\nabla_\lambda[h_l(\lambda, \gamma, n)]_{\pi_{j+1}} < \nabla_\lambda[h_l(\lambda, \gamma, n)]_{\pi'}$, i.e. the first inequality is strict. By assumption that $\pi' \notin B_t$ and $\pi' \neq \pi_e$, we have $\pi' \neq \pi_k$, $\forall 0 \leq k \leq i$. So $\exists c_0 \in \mathcal{D}$ such that $\pi'(c_0) \neq \pi_j(c_0)$. Then we get a contradiction since in iteration j, at line 6 we should return π'_{c_0} instead of π_{j+1}. Therefore, there does not exist such π' and π_e achieves the maximum. □

Algorithm 7 Constrained cost-sensitive classification

Input: policy set Π, set of policies to avoid B_t, objective function h_l, context history $\mathcal{D}$, tolerance ϵ

$\pi_0 = \operatorname{argmax}_{\pi \in \Pi} [\nabla_\lambda h_l(\lambda, \gamma, n)]_\pi$, $i = 0$
while $\pi_i \in B_t$ **do**
  for $c \in \mathcal{D}$ **do**
    compute $\pi'_c = \operatorname{argmax}_{\substack{\pi \in \Pi \\ \pi(c) \neq \pi_i(c)}} [\nabla_\lambda h_l(\lambda, \gamma, n)]_\pi$ s.t. $[\nabla_\lambda h_l(\lambda, \gamma, n)]_\pi \leq [\nabla_\lambda h_l(\lambda, \gamma, n)]_{\pi_i}$
  end for
  $\pi_{i+1} = \operatorname{argmax}_{c \in \mathcal{D}} [\nabla_\lambda h_l(\lambda, \gamma, n)]_{\pi'_c}$
  $i = i + 1$
end while
$\pi_e = \pi_i$

Output: π_e

Lemma A.3.3. *We can compute equation* (2.8) *with* $K_l|\mathcal{D}|$ *calls to a constrained argmax oracle.*

Proof. We follow the proof technique in Lemma A.3.1 and break the argmin into two pieces with $\pi \in \text{supp}(\lambda^l)$ and $\pi \in \Pi \setminus \text{supp}(\lambda^l)$. We only show how to compute the second piece as the first piece could be compute directly. We know that $\widehat{\Delta}_l^{\gamma^l}(\pi, \widehat{\pi}_{l-1}) = \sum_{i=1}^{n_l} \frac{r_i}{p^{(\ell)}_{c_i, a_i} + [\gamma^l]_\pi} (\mathbf{1}\{\widehat{\pi}_{l-1}(c_i) = a_i\} - \mathbf{1}\{\pi(c_i) = a_i\})$. Then, similar to proof of Lemma A.3.1,

let $\gamma_\pi = \gamma_0$ for all $\pi \in \Pi \setminus \text{supp}(\lambda^l)$, we have

$$\begin{aligned}
&\operatorname*{argmin}_{\pi\in\Pi\setminus\text{supp}(\lambda^l)} \widehat{\Delta}_l^{\gamma^l}(\pi, \widehat{\pi}_{l-1}) + \mathbb{E}_{c\sim\nu_{\mathcal{D}}}\left[\left(\frac{[\gamma^l]_\pi}{p^{(\ell)}_{c,a}} + \frac{[\gamma^l]_\pi}{s_{a',c}}\right)\mathbf{1}\{\widehat{\pi}_{l-1}(c) \neq \pi(c)\}\right] + \frac{\log(1/\delta_l)}{[\gamma^l]_\pi n_l} \\
&= \operatorname*{argmin}_{\pi\in\Pi\setminus\text{supp}(\lambda^l)} \sum_{i=1}^{n_l} \frac{r_i}{p^{(\ell)}_{c_i,a_i} + [\gamma^l]_\pi}\left(\mathbf{1}\{\widehat{\pi}_{l-1}(c_i) = a_i\} - \mathbf{1}\{\pi(c_i) = a_i\}\right) \\
&\quad + \mathbb{E}_{c\sim\nu_{\mathcal{D}}}\left[\left(\frac{[\gamma^l]_\pi}{p^{(\ell)}_{c,a}} + \frac{[\gamma^l]_\pi}{p^{(\ell)}_{c,a'}}\right)\mathbf{1}\{\widehat{\pi}_{l-1}(c) \neq \pi(c)\}\right] \\
&= \operatorname*{argmin}_{\pi\in\Pi\setminus\text{supp}(\lambda^l)} \sum_{i=1}^{n_l} -\frac{r_i}{p^{(\ell)}_{c_i,a_i} + \gamma_0}\mathbf{1}\{\pi(c_i) = a_i\} \\
&\quad + \mathbb{E}_{c\sim\nu_{\mathcal{D}}}\left[\left(\frac{\gamma_0}{p^{(\ell)}_{c,a}} + \frac{\gamma_0}{p^{(\ell)}_{c,a'}}\right)\mathbf{1}\{\widehat{\pi}_{l-1}(c) \neq \pi(c)\}\right] \\
&= \operatorname*{argmin}_{\pi\in\Pi\setminus\text{supp}(\lambda^l)} \sum_{i=1}^{n_l} \frac{r_i}{p^{(\ell)}_{c_i,a_i} + \gamma_0}\mathbf{1}\{\pi(c_i) = a_i\} \\
&\quad - \mathbb{E}_{c\sim\nu_{\mathcal{D}}}\left[\left(\frac{\gamma_0}{p^{(\ell)}_{c,a}} + \frac{\gamma_0}{p^{(\ell)}_{c,a'}}\right)\mathbf{1}\{\widehat{\pi}_{l-1}(c) \neq \pi(c)\}\right]
\end{aligned}$$

which is a cost-sensitive classification problem with cost vector

$$L_i(a) = \begin{cases} \frac{r_i}{p^{(\ell)}_{c_i,a_i}+\gamma_0}\mathbf{1}\{a = a_i\} & \text{for } i = 1, \cdots, n_l \\ -\left(\frac{\gamma_0}{p^{(\ell)}_{c_i,a}} + \frac{\gamma_0}{p^{(\ell)}_{c_i,\widehat{\pi}_{l-1}(c_i)}}\right)\mathbf{1}\{a \neq \widehat{\pi}_{l-1}(c_i)\} & \text{for } i = n_l+1, \cdots, n_l + |\mathcal{D}|. \end{cases}$$

□

A.3.2 Quantify the offline data

We first prove a general result for an empirical process bound of the difference of the expectation and the truth in Lemma A.3.4.

Lemma A.3.4. *Let $m = |\mathcal{D}|$ and define some set $\mathcal{K} \subset \gamma_{\max}\triangle_\Pi$. Consider some function $u : \mathcal{C} \times \mathcal{K} \to \mathbb{R}$ with $c, \kappa \mapsto u(c, \kappa)$ and define $\mathcal{F} \triangleq \{c \mapsto u(c, \kappa) : \kappa \in \mathcal{K}\}$. If*

1. *u satisfies that for any $c \in \mathcal{C}$ and $\kappa \in \mathcal{K}$, $u(c, \kappa) \in [0, b]$ where $b < \infty$ is a uniform upper bound;*

2. there exists $L < \infty$ such that $\|u(\cdot, \kappa_1) - u(\cdot, \kappa_2)\|_{\mathcal{F}} \leq L \|\kappa_1 - \kappa_2\|_1$.

Then, with probability at least $1 - \delta$,

$$\sup_{\kappa \in \mathcal{K}} |\mathbb{E}_{c \sim \nu_{\mathcal{D}}} [u(c, \kappa)] - \mathbb{E}_{c \sim \nu} [u(c, \kappa)]| \leq \sqrt{\frac{b^2}{2m} \log\left(\frac{2}{\delta}\right)} + \frac{16}{\sqrt{m}} L \gamma_{\max} \sqrt{2k \log(3e|\Pi|/k)}.$$

Proof. By the bounded condition on u we have $\{\mathbb{E}_{c \sim \nu_{\mathcal{D}}} [u(c, \kappa)] : \kappa \in \mathcal{K}\}$ satisfies the bounded difference property with parameter b. Then we use McDiarmid's inequality to get with probability at least $1 - \delta$,

$$\begin{aligned}
&\sup_{\kappa \in \mathcal{K}} |\mathbb{E}_{c \sim \nu_{\mathcal{D}}} [u(c, \kappa)] - \mathbb{E}_{c \sim \nu} [u(c, \kappa)]| \\
&\leq \sqrt{\frac{b^2}{2m} \log\left(\frac{2}{\delta}\right)} + \mathbb{E}\left[\sup_{\kappa \in \mathcal{K}} |\mathbb{E}_{c \sim \nu_{\mathcal{D}}} [u(c, \kappa)] - \mathbb{E}_{c \sim \nu} [u(c, \kappa)]|\right].
\end{aligned}$$

Also, note that by definition of $\mathcal{F}$ and classical results on entropy integral [Van der Vaart, 2000],

$$\mathbb{E}\left[\sup_{\kappa \in \mathcal{K}} |\mathbb{E}_{c \sim \nu_{\mathcal{D}}} [u(c, \kappa)] - \mathbb{E}_{c \sim \nu} [u(c, \kappa)]|\right] \leq \frac{8}{\sqrt{n}} \sup_{Q} \int_0^\infty \sqrt{\log N(\mathcal{F}, L_2(Q), \epsilon)} d\epsilon,$$

where $N(\mathcal{F}, L_2(Q), \epsilon)$ is the covering number. By condition 2 and property of covering numbers,

$$\sup_{Q} N(\mathcal{F}, L_2(Q), \epsilon) \leq N(\mathcal{F}, \|\cdot\|_{\mathcal{F}}, \epsilon) \leq N(\mathcal{K}, \|\cdot\|_1, \epsilon/L).$$

Denote B_1^k as the l_1 ball with dimension k. We know that for $\epsilon \leq 1$, $N(B_1^k, \|\cdot\|_1, \epsilon) \leq \left(\frac{3}{\epsilon}\right)^k$. Since $\mathcal{K} \subset \gamma_{\max} \triangle_{\Pi}^{(k)} \subset \gamma_{\max} B_1^k$, and there are $\binom{\Pi}{k}$ ways to choose such a support $\gamma_{\max} B_1^k$, by union bound over k-dimensional subspaces we have

$$\begin{aligned}
N(\mathcal{K}, \|\cdot\|_1, \epsilon/L) &\leq \binom{\Pi}{k} N(\gamma_{\max} B_1^k, \|\cdot\|_1, \epsilon/L) \\
&\leq \binom{\Pi}{k} N(B_1^k, \|\cdot\|_1, \epsilon/(L\gamma_{\max})) \\
&\leq \left(\frac{e|\Pi|}{k}\right)^k \left(\frac{3L\gamma_{\max}}{\epsilon}\right)^k \leq \left(\frac{3L\gamma_{\max} e|\Pi|}{\epsilon k}\right)^k.
\end{aligned}$$

Therefore,

$$\begin{aligned}
\sup_Q \int_0^\infty \sqrt{\log N(\mathcal{F}, L_2(Q), \epsilon)} d\epsilon &\leq \int_0^\infty \sqrt{\log N(\mathcal{K}, \|\cdot\|_1, \epsilon/L)} d\epsilon \\
&\leq \int_0^{L\gamma_{\max}} \sqrt{k \log\left(\frac{3L\gamma_{\max} e|\Pi|}{\epsilon k}\right)} d\epsilon \\
&= L\gamma_{\max} \int_0^1 \sqrt{k \log\left(\frac{3e|\Pi|}{\epsilon k}\right)} d\epsilon \\
&\leq L\gamma_{\max} \sqrt{\int_0^1 k \log\left(\frac{3e|\Pi|}{\epsilon k}\right) d\epsilon} \\
&\leq L\gamma_{\max} \sqrt{2k \log(3e|\Pi|/k)}.
\end{aligned}$$

Combining all results yields

$$\begin{aligned}
\mathbb{E}\left[\sup_{\lambda \in \triangle_\Pi^{(k)}} \left|\mathbb{E}_{c\sim\nu_\mathcal{D}}\left[u(c,\kappa)\right] - \mathbb{E}_{c\sim\nu}\left[u(c,\kappa)\right]\right|\right] &\leq \frac{16}{\sqrt{m}} \sup_Q \int_0^\infty \sqrt{\log N(\mathcal{F}, L_2(Q), \epsilon)} d\epsilon \\
&\leq \frac{16}{\sqrt{m}} L\gamma_{\max} \sqrt{2k \log(3e|\Pi|/k)}.
\end{aligned}$$

Therefore, our result follows. □

Then, we take two special kind of $u(c, \kappa)$, and get the bounds for our estimate of the expectation over ν with the offline history $\mathcal{D}$.

Lemma A.3.5. *Let $m = |\mathcal{D}|$. Then, with probability at least $1 - \delta$, we have*

$$\begin{aligned}
&\sup_{(\lambda,\gamma)\in\gamma_{\max}\triangle_\Pi^{(k)}} \left| \mathbb{E}_{c\sim\nu_\mathcal{D}}\left[\left(\sum_{a\in\mathcal{A}} \sqrt{(\lambda\odot\gamma)^\top (t_a^{(c)} + \eta_l)}\right)^2\right] - \mathbb{E}_{c\sim\nu}\left[\left(\sum_{a\in\mathcal{A}} \sqrt{(\lambda\odot\gamma)^\top (t_a^{(c)} + \eta_l)}\right)^2\right] \right| \\
&\leq \sqrt{\frac{|\mathcal{A}|^4 \gamma_{\max}^2 (1+\eta_l)^2}{2m} \log\left(\frac{2}{\delta}\right)} + \frac{16}{\sqrt{m}} |\mathcal{A}|^2 \gamma_{\max} \sqrt{\frac{2k(1+\eta_l)\gamma_{\max}}{\eta_l \gamma_{\min}} \log\left(\frac{3e|\Pi|}{k}\right)}.
\end{aligned}$$

Proof. Define $\kappa \in \mathcal{K}$ such that $\kappa_\pi = \lambda_\pi \gamma_\pi$. Then, $\mathcal{K} \subset \gamma_{\max}\triangle_\Pi$ since $\sum_{\pi\in\Pi} \kappa_\pi = \sum_{\pi\in\Pi} \lambda_\pi\gamma_\pi \leq \gamma_{\max}$. Then, let $u(c,\kappa) = \left(\sum_{a\in\mathcal{A}} \sqrt{\kappa^\top (t_a^{(c)} + \eta_l)}\right)^2$. We aim to use the result of Lemma A.3.4 to get our bound. First, since for any $\kappa \in \mathcal{K}$ and any $c \in \mathcal{D}$, $u(c,\kappa) \in [|\mathcal{A}|^2 \gamma_{\min}\eta_l, |\mathcal{A}|^2(1 +$

$\eta_l)\gamma_{\max}]$, so condition 1 is satisfied. Also, note that $u(c,\kappa)$ is Lipschitz in κ, i.e.

$$
\begin{aligned}
&\|u(\cdot,\kappa_1)-u(\cdot,\kappa_2)\|_{\mathcal{F}} \\
&= \sup_{c\in\mathcal{C}} |u(c,\kappa_1)-u(c,\kappa_2)| \\
&= \sup_{c\in\mathcal{C}} \left| \left(\sum_{a\in\mathcal{A}} \sqrt{\kappa_1^\top (t_a^{(c)}+\eta_l)}\right)^2 - \left(\sum_{a\in\mathcal{A}} \sqrt{\kappa_2^\top (t_a^{(c)}+\eta_l)}\right)^2 \right| \\
&\le \sup_{c\in\mathcal{C}} \left| \left(\sum_{a\in\mathcal{A}} \sqrt{\kappa_1^\top (t_a^{(c)}+\eta_l)} + \sqrt{\kappa_2^\top (t_a^{(c)}+\eta_l)}\right) \left(\sum_{a\in\mathcal{A}} \sqrt{\kappa_1^\top (t_a^{(c)}+\eta_l)} - \sqrt{\kappa_2^\top (t_a^{(c)}+\eta_l)}\right) \right| \\
&= \sup_{c\in\mathcal{C}} \left(\sum_{a\in\mathcal{A}} \sqrt{\kappa_1^\top (t_a^{(c)}+\eta_l)} + \sqrt{\kappa_2^\top (t_a^{(c)}+\eta_l)}\right) \left(\sum_{a\in\mathcal{A}} \frac{\left|(\kappa_1-\kappa_2)^\top t_a^{(c)}\right|}{\sqrt{\kappa_1^\top (t_a^{(c)}+\eta_l)} + \sqrt{\kappa_2^\top (t_a^{(c)}+\eta_l)}}\right) \\
&\le \sup_{c\in\mathcal{C}} \left(\sum_{a\in\mathcal{A}} \sqrt{\kappa_1^\top (t_a^{(c)}+\eta_l)} + \sqrt{\kappa_2^\top (t_a^{(c)}+\eta_l)}\right) \left(\sum_{a\in\mathcal{A}} \frac{\|\kappa_1-\kappa_2\|_1}{\sqrt{\kappa_1^\top (t_a^{(c)}+\eta_l)} + \sqrt{\kappa_2^\top (t_a^{(c)}+\eta_l)}}\right) \\
&\le |\mathcal{A}|^2 \sqrt{\frac{(1+\eta_l)\gamma_{\max}}{\eta_l\gamma_{\min}}} \|\kappa_1-\kappa_2\|_1 .
\end{aligned}
$$

Therefore, condition 2 is satisfied with $L = |\mathcal{A}|^2\sqrt{\frac{(1+\eta_l)\gamma_{\max}}{\eta_l\gamma_{\min}}}$. Plugging in the result in Lemma A.3.4, we get

$$
\begin{aligned}
&\sup_{\lambda\in\triangle_\Pi^{(k)}} |\mathbb{E}_{c\sim\nu_{\mathcal{D}}}[u(c,\kappa)] - \mathbb{E}_{c\sim\nu}[u(c,\kappa)]| \\
&\le \sqrt{\frac{|\mathcal{A}|^4\gamma_{\max}^2(1+\eta_l)^2}{2m}\log\left(\frac{2}{\delta}\right)} + \frac{16}{\sqrt{m}}|\mathcal{A}|^2\gamma_{\max}\sqrt{\frac{2k(1+\eta_l)\gamma_{\max}}{\eta_l\gamma_{\min}}\log\left(\frac{3e|\Pi|}{k}\right)}.
\end{aligned}
$$

□

Lemma A.3.6. *For any $\pi \in \Pi$, with probability at least $1-\delta$,*

$$\sup_{(\lambda,\gamma)\in\gamma_{\max}\triangle_\Pi} \left| \mathbb{E}_{c\sim\nu_\mathcal{D}} \left[\sum_{a\in\mathcal{A}} \frac{\sum_{a'\in\mathcal{A}} \sqrt{(\lambda\odot\gamma)^\top (t_{a'}^{(c)}+\eta_l)}}{\sqrt{(\lambda\odot\gamma)^\top (t_a^{(c)}+\eta_l)}} (\gamma_\pi[t_a^{(c)}]_\pi) \right] \right.$$
$$\left. -\mathbb{E}_{c\sim\nu} \left[\sum_{a\in\mathcal{A}} \frac{\sum_{a'\in\mathcal{A}} \sqrt{(\lambda\odot\gamma)^\top (t_{a'}^{(c)}+\eta_l)}}{\sqrt{(\lambda\odot\gamma)^\top (t_a^{(c)}+\eta_l)}} (\gamma_\pi[t_a^{(c)}]_\pi) \right] \right|$$
$$\leq \gamma_{\max} \left(\sqrt{\frac{|\mathcal{A}|^4(1+\eta)\gamma_{\max}}{2\eta\gamma_{\min} m} \log\left(\frac{2}{\delta}\right)} + \frac{8|\mathcal{A}|^2\gamma_{\max}}{\sqrt{m}(\eta_l\gamma_{\min})^{3/2}} \sqrt{2k\log(3e|\Pi|/k)} \right).$$

Proof. First, note that

$$\frac{\sum_{a'\in\mathcal{A}} \sqrt{(\lambda\odot\gamma)^\top (t_{a'}^{(c)}+\eta_l)}}{\sqrt{(\lambda\odot\gamma)^\top (t_a^{(c)}+\eta_l)}} (\gamma_\pi[t_a^{(c)}]_\pi) \leq \gamma_{\max} \frac{\sum_{a'\in\mathcal{A}} \sqrt{(\lambda\odot\gamma)^\top (t_{a'}^{(c)}+\eta_l)}}{\sqrt{(\lambda\odot\gamma)^\top (t_a^{(c)}+\eta_l)}} [t_a^{(c)}]_\pi.$$

Then, we define $u(c,\kappa) = \sum_{a\in\mathcal{A}} \frac{\sum_{a'\in\mathcal{A}} \sqrt{\kappa^\top (t_{a'}^{(c)}+\eta_l)}}{\sqrt{\kappa^\top (t_a^{(c)}+\eta_l)}} [t_a^{(c)}]_\pi$. First, note that for any $c\in\mathcal{C}$ and $\kappa\in\mathcal{K}$, $u(c,\kappa) \in \left[0, |\mathcal{A}|^2 \frac{\sqrt{(1+\eta)\gamma_{\max}}}{\sqrt{\eta\gamma_{\min}}}\right]$, so condition 1 in Lemma A.3.4 is satisfied. Also,

$$\|u(c,\kappa_1) - u(c,\kappa_2)\|_\mathcal{F} = \sup_{c\in\mathcal{C}} |u(c,\kappa_1) - u(c,\kappa_2)| \tag{A.3}$$
$$= \sup_{c\in\mathcal{C}} \left| \sum_{a\in\mathcal{A}} \frac{\sum_{a'\in\mathcal{A}} \sqrt{\kappa_1^\top (t_{a'}^{(c)}+\eta_l)}}{\sqrt{\kappa_1^\top (t_a^{(c)}+\eta_l)}} [t_a^{(c)}]_\pi - \sum_{a\in\mathcal{A}} \frac{\sum_{a'\in\mathcal{A}} \sqrt{\kappa_2^\top (t_{a'}^{(c)}+\eta_l)}}{\sqrt{\kappa_2^\top (t_a^{(c)}+\eta_l)}} [t_a^{(c)}]_\pi \right|$$
$$= \sup_{c\in\mathcal{C}} \left| \sum_{a\in\mathcal{A}} \left[\frac{\sum_{a'\in\mathcal{A}} \sqrt{\kappa_1^\top (t_{a'}^{(c)}+\eta_l)}\sqrt{\kappa_2^\top (t_a^{(c)}+\eta_l)} - \sqrt{\kappa_2^\top (t_{a'}^{(c)}+\eta_l)}\sqrt{\kappa_1^\top (t_a^{(c)}+\eta_l)}}{\sqrt{\kappa_1^\top (t_a^{(c)}+\eta_l)}\sqrt{\kappa_2^\top (t_a^{(c)}+\eta_l)}} [t_a^{(c)}]_\pi \right] \right|$$
$$\leq \sup_{c\in\mathcal{C}} \sum_{a\in\mathcal{A}} \left[\frac{\sum_{a'\in\mathcal{A}} \left| \sqrt{\kappa_1^\top (t_{a'}^{(c)}+\eta_l)}\sqrt{\kappa_2^\top (t_a^{(c)}+\eta_l)} - \sqrt{\kappa_2^\top (t_{a'}^{(c)}+\eta_l)}\sqrt{\kappa_1^\top (t_a^{(c)}+\eta_l)} \right|}{\sqrt{\kappa_1^\top (t_a^{(c)}+\eta_l)}\sqrt{\kappa_2^\top (t_a^{(c)}+\eta_l)}} \right]. \tag{A.4}$$

Note that by triangular inequality

$$\begin{aligned}
&\left|\sqrt{\kappa_2^\top(t_a^{(c)}+\eta_l)}\sqrt{\kappa_1^\top(t_{a'}^{(c)}+\eta_l)}-\sqrt{\kappa_1^\top(t_a^{(c)}+\eta_l)}\sqrt{\kappa_2^\top(t_{a'}^{(c)}+\eta_l)}\right| \\
&\leq \left|\sqrt{\kappa_2^\top(t_a^{(c)}+\eta_l)}-\sqrt{\kappa_1^\top(t_a^{(c)}+\eta_l)}\right|\sqrt{\kappa_1^\top(t_{a'}^{(c)}+\eta_l)} \\
&\quad+\sqrt{\kappa_1^\top(t_a^{(c)}+\eta_l)}\left|\sqrt{\kappa_1^\top(t_{a'}^{(c)}+\eta_l)}-\sqrt{\kappa_2^\top(t_{a'}^{(c)}+\eta_l)}\right|.
\end{aligned}$$

Also note that

$$\begin{aligned}
\left|\sqrt{\kappa_2^\top(t_a^{(c)}+\eta_l)}-\sqrt{\kappa_1^\top(t_a^{(c)}+\eta_l)}\right| &= \frac{\left|\sum_{\pi\in\Pi}([\kappa_1]_\pi-[\kappa_2]_\pi)(t_a^{(c)}+\eta_l)_\pi\right|}{\sqrt{\kappa_2^\top(t_a^{(c)}+\eta_l)}+\sqrt{\kappa_1^\top(t_a^{(c)}+\eta_l)}} \\
&\leq \frac{1}{2\sqrt{\eta_l\gamma_{\min}}}\left\|\kappa_2-\kappa_1\right\|_1.
\end{aligned}$$

Therefore, (A.4) is bounded by $|\mathcal{A}|^2\frac{1}{\eta_l\gamma_{\min}}\frac{1}{2\sqrt{\eta_l\gamma_{\min}}}\left\|\kappa_2-\kappa_1\right\|_1$, so condition 2 is satisfied with $L=\frac{|\mathcal{A}|^2}{2(\eta_l\gamma_{\min})^{3/2}}$. Then, by Lemma A.3.4, with probability at least $1-\delta$,

$$\begin{aligned}
&\sup_{(\lambda,\gamma)\in\gamma_{\max}\triangle_\Pi}\left|\mathbb{E}_{c\sim\nu_\mathcal{D}}\left[\sum_a\frac{\sum_{a'\in\mathcal{A}}\sqrt{(\lambda\odot\gamma)^\top(t_{a'}^{(c)}+\eta_l)}}{\sqrt{(\lambda\odot\gamma)^\top(t_a^{(c)}+\eta_l)}}(\gamma_\pi[t_a^{(c)}]_\pi)\right]\right. \\
&\left.-\mathbb{E}_{c\sim\nu}\left[\sum_a\frac{\sum_{a'\in\mathcal{A}}\sqrt{(\lambda\odot\gamma)^\top(t_{a'}^{(c)}+\eta_l)}}{\sqrt{(\lambda\odot\gamma)^\top(t_a^{(c)}+\eta_l)}}(\gamma_\pi[t_a^{(c)}]_\pi)\right]\right| \\
&\leq\gamma_{\max}\left(\sqrt{\frac{|\mathcal{A}|^4(1+\eta)\gamma_{\max}}{2\eta\gamma_{\min}m}\log\left(\frac{2}{\delta}\right)}+\frac{8|\mathcal{A}|^2\gamma_{\max}}{\sqrt{m}(\eta_l\gamma_{\min})^{3/2}}\sqrt{2k\log(3e|\Pi|/k)}\right).
\end{aligned}$$

□

A.4 Proof of Theorem 2.3.4

We first write down Algorithm 4 in full detail in Algorithm 8. We aim to show that Algorithm 8 achieves the sample complexity lower bound. The two big goals here is to show that $\widehat{\pi}_l\in S_l$ for all l, which shows that we get the optimal policy, and n_l achieves the sample complexity lower bound.

Algorithm 8 Full CODA Algorithm

Input: policies $\Pi = \{\pi : \mathcal{C} \to \mathcal{A}\}_\pi$, feature map $\phi : \mathcal{C} \times \mathcal{A} \to \mathbb{R}^d$, $\delta \in (0,1)$, historical data $\mathcal{D} = \{\nu_s\}_s$

initiate $\widehat{\pi}_0 \in \Pi$ arbitrarily, $\lambda_0 = \mathbf{e}_{\widehat{\pi}_0}$, $\widehat{\Delta}_0(\pi)$, γ_0 appropriately

for $l = 1, 2, \cdots$ **do**

$\epsilon_l = 2^{-l}$, $\eta_l = C_1 \epsilon_l^2 |\mathcal{A}|^{-4}$, $\delta_l = \delta/(l^2|\Pi|^2)$, K_l appropriately

$t_a^{(c)}(\pi') = \{\mathbf{1}\{\pi(c) = a, \pi'(c) \neq a\} + \mathbf{1}\{\pi(c) \neq a, \pi'(c) = a\}\}_{\pi \in \Pi} \in \mathbb{R}^\Pi$

Define $\gamma_{\min} := \frac{1}{3}\sqrt{\frac{\eta_l \log(1/\delta_l)}{n}}$, $\gamma_{\max} := \sqrt{\frac{\log(1/\delta_l)}{|\mathcal{A}|^2 \eta_l n}}$

Define

$$h_l(\lambda, \gamma, n) = \sum_{\pi \in \Pi} \lambda_\pi \left(-\widehat{\Delta}_{l-1}^{\gamma^{l-1}}(\pi, \widehat{\pi}_{l-1}) + \frac{\log(1/\delta_l)}{\gamma_\pi n} \right) + \mathbb{E}_{c \sim \nu_\mathcal{D}} \left[\left(\sum_{a \in \mathcal{A}} \sqrt{(\lambda \odot \gamma)^\top (t_a^{(c)}(\widehat{\pi}_{l-1}) + \eta_l)} \right)^2 \right]. \tag{A.5}$$

Let $\lambda^l, \gamma^l, n_l = \mathsf{FW\text{-}GD}(\Pi, |\mathcal{A}|, \widehat{\pi}_{l-1}, \eta_l, K_l, \epsilon_l, \gamma_{\min}, \gamma_{\max})$. These are the solutions to

$$n_\ell := \min\{n \in \mathbb{N} : \max_{\lambda \in \triangle_\Pi} \min_{\gamma \in [\gamma_{\min}, \gamma_{\max}]^{|\Pi|}} h_l(\lambda, \gamma, n) \leq \epsilon_\ell\} \tag{A.6}$$

Receive contexts $c_1, c_2, \cdots, c_{n_l} \sim \nu$.

For each c_s, $s = 1, 2, \cdots, n_l$, pull arms $a_s \sim p_{c_s}^{(\ell)}$ where $p_{c_s, a_s}^{(\ell)} \propto \sqrt{(\lambda^l \odot \gamma^l)^\top (t_{a_s}^{(c_s)}(\widehat{\pi}_{l-1}) + \eta_l)}$, and observe rewards r_s where $t_{a_s}^{(c_s)}(\widehat{\pi}_{l-1}) \in \mathbb{R}^{|\Pi|}$

For each $\pi \in \Pi$, define the IPS estimator

$$\widehat{\Delta}_l^{\gamma^l}(\pi, \widehat{\pi}_{l-1}) = \sum_{s=1}^{n_l} \frac{r_s}{p_{c_s, a_s}^{(\ell)} + [\gamma^l]_\pi} (\mathbf{1}\{\widehat{\pi}_{l-1}(c_s) = a_s\} - \mathbf{1}\{\pi(c_s) = a_s\})$$

set

$$\widehat{\pi}_l = \arg\min_{\pi \in \Pi} \widehat{\Delta}_l^{\gamma^l}(\pi, \widehat{\pi}_{l-1}) + \mathbb{E}_{c \sim \nu_\mathcal{D}} \left[\left(\frac{[\gamma^l]_\pi}{p_{c, \pi(c)}^{(\ell)}} + \frac{[\gamma^l]_\pi}{p_{c, \widehat{\pi}_{l-1}(c)}^{(\ell)}} \right) \mathbf{1}\{\widehat{\pi}_{l-1}(c) \neq \pi(c)\} \right] + \frac{\log(1/\delta_l)}{[\gamma^l]_\pi n_l}. \tag{A.7}$$

end for

Output: $\widehat{\pi}_l$

Theorem A.4.1. *With probability at least $1 - \delta$, Algorithm 8 returns a policy $\widehat{\pi}$ satisfying*

$V(\pi_*) - V(\widehat{\pi}_\ell) \le \epsilon$ *in a number of samples not exceeding* $O(\rho_{*,\epsilon} \log(|\Pi| \log_2(1/\Delta_\epsilon)/\delta) \log_2(1/\Delta_\epsilon)$ *where* $\Delta_\epsilon := \max\{\epsilon, \min_{\pi \in \Pi} V(\pi_*) - V(\pi)\}$.

Proof. We first define our key events. Recall

$$\widehat{\Delta}_l^{\gamma^l}(\pi, \widehat{\pi}_{l-1}) = \sum_{s=1}^{n_l} \frac{r_s}{p^{(\ell)}_{c_s,a_s} + [\gamma^l]_\pi} (\mathbf{1}\{\widehat{\pi}_{l-1}(c_s) = a_s\} - \mathbf{1}\{\pi(c_s) = a_s\})$$

and $\Delta(\pi, \pi') = V(\pi') - V(\pi)$. Define $w(\lambda, \gamma) \in \mathbb{R}^{|\mathcal{A}| \times |\mathcal{C}|}$ with

$$[w(\lambda, \gamma)]_{a,c} := \nu_c \cdot p_{c,a} = \nu_c \cdot \frac{\sqrt{(\lambda \odot \gamma)^\top (t_a^{(c)}(\widehat{\pi}_{l-1}) + \eta_l)}}{\sum_{a' \in \mathcal{A}} \sqrt{(\lambda \odot \gamma)^\top (t_{a'}^{(c)}(\widehat{\pi}_{l-1}) + \eta_l)}}.$$

Then define the events

$$\mathcal{E}_l := \bigcap_{\pi, \pi' \in \Pi} \left\{ \left| \widehat{\Delta}_l^{\gamma^l}(\pi, \pi') - \Delta(\pi, \pi') \right| \le 2[\gamma^l]_\pi \left\| \phi_\pi - \phi_{\pi'} \right\|^2_{A(w(\lambda^l, \gamma^l))^{-1}} + \frac{2 \log(1/\delta_l)}{[\gamma^l]_\pi n_l} \right\},$$

and the good event $\mathcal{E} = \bigcap_{l=1}^\infty \mathcal{E}_l$. Lemma A.4.3 shows that $\mathcal{E}$ happens with probability at least $1 - \delta$, and Lemma A.4.7 shows that under this event $\mathcal{E}$,

$$n_l \lesssim \min_{w \in \Omega} \max_{\pi \in \Pi} \frac{\|\phi_{\pi_*} - \phi_\pi\|^2_{A(w)^{-1}} \log(1/\delta_l)}{\epsilon_l^2 + \Delta(\pi, \pi_*)^2}.$$

Therefore, the total number of samples is no more than

$$\begin{aligned}
&\sum_{l=1}^{\log_2(1/\Delta_\epsilon)} \min_{w \in \Omega} \max_{\pi \in \Pi} \frac{\|\phi_{\pi_*} - \phi_\pi\|^2_{A(w)^{-1}} \log(l^2 |\Pi|^2/\delta)}{\epsilon_l^2 + \Delta(\pi, \pi_*)^2} \\
&\overset{(i)}{\le} \sum_{l=1}^{\log_2(1/\Delta_\epsilon)} \min_{w \in \Omega} \max_{\pi \in \Pi \setminus \pi_*} \frac{2 \|\phi_{\pi_*} - \phi_\pi\|^2_{A(w)^{-1}} \log(l^2 |\Pi|^2/\delta)}{\epsilon_l^2 + \Delta(\pi, \pi_*)^2} \\
&\overset{(ii)}{\le} \sum_{l=1}^{\log_2(1/\Delta_\epsilon)} \min_{p^{(c)} \in \triangle_{\mathcal{A}}, \forall c \in \mathcal{C}} \max_{\pi \in \Pi \setminus \pi_*} \frac{\mathbb{E}_{c \sim \nu} \left[\left(\frac{1}{p^{(c)}_{\pi_*(c)}} + \frac{1}{p^{(c)}_{\pi(c)}} \right) \mathbf{1}\{\pi_*(c) \ne \pi(c)\} \right] \log(l^2 |\Pi|^2/\delta)}{\Delta(\pi, \pi_*)^2 + \epsilon_l^2} \\
&\lesssim \rho_{\star,\epsilon}(\Pi, v) \log(\log_2(1/\Delta_\epsilon) |\Pi|/\delta) \log_2(1/\Delta_\epsilon).
\end{aligned}$$

where (i) follows from the fact that π_* gives zero for the RHS, and (ii) follows from Lemma A.6.1. □

In what follows, we will fill in the road map to the proof of Lemma A.4.3 and A.4.7. First, Lemma A.4.2 controls the estimation error of the gap and shows that $\mathbb{P}(\mathcal{E}_\ell) > 1 - \delta_\ell$, which leads to the high-probability of the good event $\mathcal{E}$ (Lemma A.4.3). Lemma A.4.4 applies the duality machinery in Section A.5 and controls the variance term. Lemma A.4.5 applies the result of Lemma A.4.4 and shows an upper bound for the difference between estimate gap and the true gap, which is a very similar result of Lemma A.2.3. Lemma A.4.6 is an important lemma showing the analytical solution of w given some λ and γ. With all of these results above, we get Lemma A.4.7 which gives the upper bound on the sample complexity.

Lemma A.4.2. *For any $l > 0$, $\pi, \pi' \in \Pi$, with probability at least $1 - \delta_l$,*

$$\left|\widehat{\Delta}_l^{\gamma^l}(\pi, \pi') - \Delta(\pi, \pi')\right| \leq 2[\gamma^l]_\pi \left\|\phi_\pi - \phi_{\pi'}\right\|^2_{A(w(\lambda^l, \gamma^l))^{-1}} + \frac{2\log(1/\delta_l)}{[\gamma^l]_\pi n_l}.$$

Proof. Define

$$\widehat{V}_l^{\gamma^l}(\pi) := \sum_{s=1}^{n_l} \frac{r_s}{p^{(\ell)}_{c_s,a_s} + [\gamma^l]_\pi} \mathbf{1}\{\pi(c_s) = a_s\},$$

so that

$$\widehat{\Delta}_l^{\gamma^l}(\pi, \pi') = \widehat{V}_l^{\gamma^l}(\pi') - \widehat{V}_l^{\gamma^l}(\pi).$$

First, note that below.

$$\begin{aligned} V(\pi) &= \mathbb{E}_{c\sim\nu}\left[r(c, \pi(c))\right] \\ &= \mathbb{E}_{c\sim\nu}\left[\mathbb{E}_{a\sim p_c^{(\ell)}}\left[r(c,a)\frac{\mathbf{1}\{\pi(c) = a\}}{p^{(\ell)}_{c,a}}\middle| c\right]\right] = \mathbb{E}\left[\frac{1}{t}\sum_{s=1}^{t}\frac{r_s}{p^{(\ell)}_{c_s,a_s}}\mathbf{1}\{\pi(c_s) = a_s\}\right]. \end{aligned}$$

Therefore,

$$
\begin{aligned}
&\left|\mathbb{E}\left[\widehat{V}_l^{\gamma^l}(\pi)-\widehat{V}_l^{\gamma^l}(\pi')\right]-[V(\pi)-V(\pi')]\right| \\
&\leq\left|\mathbb{E}\left[\frac{1}{n_l}\sum_{s=1}^{n_l}\left(\frac{1}{p^{(\ell)}_{c_s,a_s}+[\gamma^l]_\pi}-\frac{1}{p^{(\ell)}_{c_s,a_s}}\right)\left(\mathbf{1}\{\pi(c_s)=a_s\}-\mathbf{1}\{\pi'(c_s)=a_s\}\right)\right]\right| \\
&=\left|\mathbb{E}\left[\frac{1}{n_l}\sum_{s=1}^{n_l}\frac{-[\gamma^l]_\pi}{p^{(\ell)}_{c_s,a_s}\left(p^{(\ell)}_{c_s,a_s}+[\gamma^l]_\pi\right)}\left(\mathbf{1}\{\pi(c_s)=a_s\}-\mathbf{1}\{\pi'(c_s)=a_s\}\right)\right]\right| \\
&\leq\mathbb{E}\left[\frac{1}{n_l}\sum_{s=1}^{n_l}\frac{[\gamma^l]_\pi\left(\mathbf{1}\{\pi'(c_s)=a_s,\pi(c_s)\neq a_s\}+\mathbf{1}\{\pi'(c_s)\neq a_s,\pi(c_s)=a_s\}\right)}{p^{(\ell)}_{c_s,a_s}\left(p^{(\ell)}_{c_s,a_s}+[\gamma^l]_\pi\right)}\right] \\
&=[\gamma^l]_\pi\mathbb{E}\left[\frac{1}{p^{(\ell)}_{c,a}\left(p^{(\ell)}_{c,a}+[\gamma^l]_\pi\right)\nu_c^2}[\phi_\pi-\phi_{\pi'}]^2_{a,c}\right] \\
&=[\gamma^l]_\pi\sum_{c\in\mathcal{C}}\nu_c\sum_{a\in\mathcal{A}}p^{(\ell)}_{c,a}\frac{1}{p^{(\ell)}_{c,a}\nu_c^2\left(p^{(\ell)}_{c,a}+[\gamma^l]_\pi\right)}[\phi_\pi-\phi_{\pi'}]^2_{a,c} \\
&\leq[\gamma^l]_\pi\left\|\phi_\pi-\phi_{\pi'}\right\|^2_{A(w(\lambda^l,\gamma^l))^{-1}}
\end{aligned}
$$

where the last inequality follows since $\nu_c p^{(\ell)}_{c,a}=[w(\lambda^l,\gamma^l)]_{a,c}$. Meanwhile, note that

$$
\frac{r_s}{p^{(\ell)}_{c_s,a_s}+[\gamma^l]_\pi}\left(\mathbf{1}\{\pi(c_s)=a_s\}-\mathbf{1}\{\pi'(c_s)=a_s\}\right)\leq\frac{1}{[\gamma^l]_\pi},
$$

and

$$
\begin{aligned}
&\mathbb{E}\left[\left(\frac{r_s}{p^{(\ell)}_{c_s,a_s}+[\gamma^l]_\pi}\left(\mathbf{1}\{\pi(c_s)=a_s\}-\mathbf{1}\{\pi'(c_s)=a_s\}\right)\right)^2\right] \\
&\leq\mathbb{E}\left[\frac{1}{(p^{(\ell)}_{c_s,a_s}+[\gamma^l]_\pi)^2}\left(\mathbf{1}\left\{\pi\left(c_s\right)=a_s\right\}-\mathbf{1}\left\{\pi'\left(c_s\right)=a_s\right\}\right)^2\right] \\
&=\mathbb{E}\left[\frac{1}{(p^{(\ell)}_{c_s,a_s}+[\gamma^l]_\pi)^2\nu_c^2}[\phi_\pi-\phi_{\pi'}]^2_{a,c}\right] \\
&\leq\left\|\phi_\pi-\phi_{\pi'}\right\|^2_{A(w(\lambda^l,\gamma^l))^{-1}}
\end{aligned}
$$

by a similar argument as before. Therefore, by Bernstein's inequality, we have with probability

at least $1-\delta$,

$$\left|\widehat{V}_l^{\gamma^l}(\pi)-\widehat{V}_l^{\gamma^l}(\pi')-\mathbb{E}\left[\widehat{V}_l^{\gamma^l}(\pi)-\widehat{V}_l^{\gamma^l}(\pi')\right]\right| \le \sqrt{\|\phi_\pi-\phi_{\pi'}\|^2_{A(w(\lambda^l,\gamma^l))^{-1}}\frac{2\log(1/\delta)}{n_l}}+\frac{\log(1/\delta)}{[\gamma^l]_\pi n_l}.$$

Combining this with the deviation on expectation gives us

$$\begin{aligned}
&\left|\widehat{\Delta}_l^{\gamma^l}(\pi,\pi')-\Delta(\pi,\pi')\right| \\
&\le [\gamma^l]_\pi \|\phi_\pi-\phi_{\pi'}\|^2_{A(w(\lambda^l,\gamma^l))^{-1}}+\sqrt{\|\phi_\pi-\phi_{\pi'}\|^2_{A(w(\lambda^l,\gamma^l))^{-1}}\frac{2\log(1/\delta)}{n_l}}+\frac{2\log(1/\delta)}{[\gamma^l]_\pi n_l} \\
&\le 2[\gamma^l]_\pi \|\phi_\pi-\phi_{\pi'}\|^2_{A(w(\lambda^l,\gamma^l))^{-1}}+\frac{4\log(1/\delta)}{[\gamma^l]_\pi n_l}.
\end{aligned}$$

□

Lemma A.4.3. $\mathbb{P}(\mathcal{E}) \ge 1-\delta$.

Proof. By Lemma A.4.2 and a union bound over all policies, we have

$$\mathbb{P}\left(\mathcal{E}_l \mid \mathcal{E}_{l-1},\cdots,\mathcal{E}_1\right) \ge 1-\frac{\delta}{l^2}.$$

Since $\mathcal{E}=\bigcap_{l=0}^\infty \mathcal{E}_l$,

$$\begin{aligned}
\mathbb{P}(\mathcal{E}^c) &= \mathbb{P}((\cap_{l=0}^\infty \mathcal{E}_l)^c) = \mathbb{P}\left(\cup_{l=0}^\infty \mathcal{E}_l^c\right) = \mathbb{P}\left(\cup_{l=0}^\infty \left(\mathcal{E}_l^c \backslash \left(\cup_{j<l}\mathcal{E}_j^c\right)\right)\right) \\
&\le \sum_{l=0}^\infty \mathbb{P}\left(\mathcal{E}_l^c \backslash \left(\cup_{j<l}\mathcal{E}_j^c\right)\right) \le \sum_{l=0}^\infty \mathbb{P}\left(\mathcal{E}_l^c \mid (\cap_{j<l}\mathcal{E}_j)\right) \le \sum_{l=0}^\infty \frac{\delta}{l^2} \le \delta.
\end{aligned}$$

Therefore, $\mathbb{P}(\mathcal{E}) \ge 1-\delta$. □

Lemma A.4.4. *Under $\mathcal{E}$, we have for any $\pi \in \Pi$,*

$$[\gamma^l]_\pi \left\|\phi_\pi-\phi_{\widehat{\pi}_{l-1}}\right\|^2_{A(w(\lambda^l,\gamma^l))^{-1}}+\frac{\log(1/\delta_l)}{[\gamma^l]_\pi n_l} \le \frac{1}{6}\epsilon_l+\frac{1}{64}\widehat{\Delta}_{l-1}^{\gamma^{l-1}}(\pi,\widehat{\pi}_{l-1}).$$

Proof. We know that the choice of n_l ensures

$$h_l(\lambda^l,\gamma^l,n_l) \le \epsilon_l.$$

Also, by Theorem A.5.1 we have

$$\frac{1}{3}\epsilon_l \geq \max_{\pi\in\Pi}\left(-\frac{1}{8}\widehat{\Delta}_{l-1}^{\gamma^{l-1}}(\pi,\widehat{\pi}_{l-1}) + 8[\gamma^l]_\pi \left\|\phi_\pi - \phi_{\widehat{\pi}_{l-1}}\right\|^2_{A(w(\lambda^l,\gamma^l))^{-1}} + \frac{8\log(1/\delta_l)}{[\gamma^l]_\pi n_l}\right) - h_l(\lambda^l,\gamma^l,n_l).$$

Combining the above two displays gives us

$$\begin{aligned}\epsilon_l &\geq h_l(\lambda^l,\gamma^l,n_l)\\ &\geq \max_{\pi\in\Pi}\left(-\frac{1}{8}\widehat{\Delta}_{l-1}^{\gamma^{l-1}}(\pi,\widehat{\pi}_{l-1}) + 8[\gamma^l]_\pi \left\|\phi_{\widehat{\pi}_{l-1}} - \phi_\pi\right\|^2_{A(w(\lambda^l,\gamma^l))^{-1}} + \frac{8\log(1/\delta_l)}{[\gamma^l]_\pi n_l}\right) - \frac{1}{3}\epsilon_l.\end{aligned}$$

Therefore, for any $\pi \in \Pi$,

$$[\gamma^l]_\pi \left\|\phi_\pi - \phi_{\widehat{\pi}_{l-1}}\right\|^2_{A(w(\lambda^l,\gamma^l))^{-1}} + \frac{\log(1/\delta_l)}{[\gamma^l]_\pi n_l} \leq \frac{1}{6}\epsilon_l + \frac{1}{64}\widehat{\Delta}_{l-1}^{\gamma^{l-1}}(\pi,\widehat{\pi}_{l-1}).$$

□

Lemma A.4.5. *Under $\mathcal{E}$, for all $l \in \mathbb{N}$, the following holds:*

1. $|\widehat{\Delta}_{l-1}^{\gamma^{l-1}}(\pi,\widehat{\pi}_{l-1}) - \Delta(\pi,\pi_*)| \leq 2\epsilon_{l-1} + \frac{1}{4}\Delta(\pi,\pi_*)$.

2. $\widehat{\pi}_l \in S_l := \{\pi \in \Pi : \Delta(\pi,\pi_*) \leq \epsilon_l\}$.

Proof. We prove this by induction. First, in round $l = 0$, this holds since our rewards are bounded by 1. Then, assume that in round $l-1$, we have $\widehat{\pi}_{l-1} \in S_{l-1}$ and

$$|\widehat{\Delta}_{l-2}^{\gamma_{l-2}}(\pi,\widehat{\pi}_{l-2}) - \Delta(\pi,\pi_*)| \leq 2\epsilon_{l-2} + \frac{1}{4}\Delta(\pi,\pi_*).$$

Then, on round l,

$$
\begin{aligned}
&|\widehat{\Delta}_{l-1}^{\gamma^{l-1}}(\pi, \widehat{\pi}_{l-1}) - \Delta(\pi, \pi_*)| \\
&= |\widehat{\Delta}_{l-1}^{\gamma^{l-1}}(\pi, \widehat{\pi}_{l-1}) - \Delta(\pi, \widehat{\pi}_{l-1}) - \Delta(\widehat{\pi}_{l-1}, \pi_*)| \\
&\leq 2[\gamma^{l-1}]_\pi \left\|\phi_\pi - \phi_{\widehat{\pi}_{l-1}}\right\|^2_{A(w(\lambda^{l-1},\gamma^{l-1}))^{-1}} + \frac{2\log(1/\delta_{l-1})}{[\gamma^{l-1}]_\pi n_{l-1}} + \epsilon_{l-1} \\
&\qquad \text{(from event } \mathcal{E} \text{ and inductive hypothesis)} \\
&\leq \frac{2}{3}\epsilon_l + \frac{1}{64}\widehat{\Delta}_{l-2}^{\gamma_{l-2}}(\pi, \widehat{\pi}_{l-2}) + \frac{1}{64}\widehat{\Delta}_{l-2}^{\gamma_{l-2}}(\widehat{\pi}_{l-1}, \widehat{\pi}_{l-2}) + \epsilon_{l-1} \qquad \text{(from Lemma A.4.4)} \\
&\leq \frac{5}{3}\epsilon_{l-1} + \frac{1}{64}\left(2\epsilon_{l-2} + \frac{5}{4}\Delta(\pi, \pi_*) + 2\epsilon_{l-2} + \frac{5}{4}\Delta(\widehat{\pi}_{l-1}, \pi_*)\right) \qquad \text{(from inductive hypothesis)} \\
&\leq \frac{5}{3}\epsilon_{l-1} + \frac{1}{64}\left(2\epsilon_{l-2} + \frac{5}{4}\Delta(\pi, \pi_*) + 2\epsilon_{l-2} + \frac{5}{4}\epsilon_{l-1}\right) \\
&\leq 2\epsilon_{l-1} + \frac{1}{4}\Delta(\pi, \pi_*).
\end{aligned}
$$

Also,

$$
\begin{aligned}
\Delta(\widehat{\pi}_l, \widehat{\pi}_{l-1}) &\leq \widehat{\Delta}_l^{\gamma^l}(\widehat{\pi}_l, \widehat{\pi}_{l-1}) + [\gamma^l]_{\widehat{\pi}_l} \left\|x_{\widehat{\pi}_l} - \phi_{\widehat{\pi}_{l-1}}\right\|^2_{A(w(\lambda^l,\gamma^l))^{-1}} + \frac{\log(1/\delta_l)}{[\gamma^l]_{\widehat{\pi}_l} n_l} \qquad \text{(from } \mathcal{E}) \\
&\leq \widehat{\Delta}_l^{\gamma^l}(\pi_*, \widehat{\pi}_{l-1}) + [\gamma^l]_{\pi_*} \left\|\phi_{\pi_*} - \phi_{\widehat{\pi}_{l-1}}\right\|^2_{A(w(\lambda^l,\gamma^l))^{-1}} + \frac{\log(1/\delta_l)}{[\gamma^l]_{\pi_*} n_l} \\
&\qquad \text{(eqn (2.8), the minimum)} \\
&\leq \Delta(\pi_*, \widehat{\pi}_{l-1}) + 2[\gamma^l]_{\pi_*} \left\|\phi_{\pi_*} - \phi_{\widehat{\pi}_{l-1}}\right\|^2_{A(w(\lambda^l,\gamma^l))^{-1}} + \frac{2\log(1/\delta_l)}{[\gamma^l]_{\pi_*} n_l} \qquad \text{(from } \mathcal{E}) \\
&\leq \Delta(\pi_*, \widehat{\pi}_{l-1}) + \frac{1}{3}\epsilon_l + \frac{1}{32}\widehat{\Delta}_{l-1}^{\gamma^{l-1}}(\pi_*, \widehat{\pi}_{l-1}) \qquad \text{(from Lemma A.4.4)} \\
&\leq \Delta(\pi_*, \widehat{\pi}_{l-1}) + \frac{1}{3}\epsilon_l + \frac{1}{32}\left(2\epsilon_{l-1} + \frac{5}{4}\Delta(\pi_*, \pi_*)\right). \qquad \text{(from the above)}
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
\Delta(\widehat{\pi}_l, \pi_*) &= \Delta(\widehat{\pi}_l, \widehat{\pi}_{l-1}) - \Delta(\pi_*, \widehat{\pi}_{l-1}) \\
&\leq \frac{1}{3}\epsilon_l + \frac{1}{16}2\epsilon_l \\
&\leq \epsilon_\ell
\end{aligned}
$$

Therefore, $\Delta(\widehat{\pi}_l, \pi_*) \leq \epsilon_l$, so $\widehat{\pi}_l \in S_l$.

□

Lemma A.4.6. *For any $\lambda \in \triangle_\Pi$, $\gamma \in \mathbb{R}^{|\Pi|}$, and $\pi' \in \Pi$, we have*

$$\min_{w\in\Omega} \sum_{\pi\in\Pi} \lambda_\pi \gamma_\pi \|\phi_\pi - \phi_{\pi'}\|^2_{A(w)^{-1}} = \mathbb{E}_{c\sim\nu}\left[\left(\sum_{a\in\mathcal{A}} \sqrt{(\lambda\odot\gamma)^\top t_a^{(c)}(\pi')}\right)^2\right].$$

where $w_{a,c} = \nu_c p_a^{(c)}$ and $p_a^{(c)} \propto \sqrt{\sum_{\pi\in\Pi} \lambda_\pi\gamma_\pi(\mathbf{1}\{\pi'(c)=a, \pi(c)\neq a\} + \mathbf{1}\{\pi'(c)\neq a, \pi(c)=a\})}$ and $\odot$ denotes element-wise multiplication.

Proof. For any $\lambda \in \triangle_\Pi$,

$$\begin{aligned}
&\min_{w\in\Omega} \sum_{\pi\in\Pi} \lambda_\pi \gamma_\pi \|\phi_\pi - \phi_{\pi'}\|^2_{A(w)^{-1}} \\
&= \min_{w\in\Omega} \sum_{\pi\in\Pi}\sum_{a,c} \frac{\lambda_\pi\gamma_\pi}{w_{a,c}} (\phi_\pi - \phi_{\pi'})^\top e_{a,c} e_{a,c}^\top (\phi_\pi - \phi_{\pi'}) \\
&= \min_{p_1,\ldots,p_{|\mathcal{C}|}\in\triangle_\mathcal{A}} \sum_{\pi\in\Pi}\sum_{a,c} \frac{\lambda_\pi\gamma_\pi}{\nu_c p_{c,a}} (\phi_\pi - \phi_{\pi'})^\top e_{a,c} e_{a,c}^\top (\phi_\pi - \phi_{\pi'}) \\
&= \sum_c \min_{p_c\in\triangle_\mathcal{A}} \sum_a \sum_{\pi\in\Pi} \frac{\lambda_\pi\gamma_\pi}{\nu_c p_{c,a}} (\phi_\pi - \phi_{\pi'})^\top e_{a,c} e_{a,c}^\top (\phi_\pi - \phi_{\pi'}) \\
&= \sum_c \frac{1}{\nu_c} \min_{p_c\in\triangle_\mathcal{A}} \sum_a \frac{1}{p_{c,a}} \left(\sum_{\pi\in\Pi} \lambda_\pi\gamma_\pi (\phi_\pi - \phi_{\pi'})^\top e_{a,c} e_{a,c}^\top (\phi_\pi - \phi_{\pi'})\right) \\
&= \sum_c \frac{1}{\nu_c} \left(\sum_{a\in\mathcal{A}} \sqrt{\sum_{\pi\in\Pi} \lambda_\pi\gamma_\pi (\phi_\pi - \phi_{\pi'})^\top e_{a,c} e_{a,c}^\top (\phi_\pi - \phi_{\pi'})}\right)^2 \\
&= \sum_c \frac{1}{\nu_c} \left(\sum_{a\in\mathcal{A}} \sqrt{\sum_{\pi\in\Pi} \lambda_\pi\gamma_\pi \nu_c^2 (\mathbf{1}\{\pi'(c)=a, \pi(c)\neq a\} + \mathbf{1}\{\pi'(c)\neq a, \pi(c)=a\})}\right)^2 \\
&= \sum_c \nu_c \left(\sum_{a\in\mathcal{A}} \sqrt{\sum_{\pi\in\Pi} \lambda_\pi\gamma_\pi (\mathbf{1}\{\pi'(c)=a, \pi(c)\neq a\} + \mathbf{1}\{\pi'(c)\neq a, \pi(c)=a\})}\right)^2 \\
&= \mathbb{E}_{c\sim\nu}\left[\left(\sum_{a\in\mathcal{A}} \sqrt{(\lambda\odot\gamma)^\top t_a^{(c)}(\pi')}\right)^2\right].
\end{aligned}$$

Note that the minimizer

$$p_{c,a} = \frac{\sqrt{\sum_{\pi\in\Pi} \lambda_\pi \gamma_\pi (\phi_\pi - \phi_{\pi'})^\top e_{a,c} e_{a,c}^\top (\phi_\pi - \phi_{\pi'})}}{\sum_{a'} \sqrt{\sum_{\pi\in\Pi} \lambda_\pi \gamma_\pi (\phi_\pi - \phi_{\pi'})^\top e_{a',c} e_{a',c}^\top (\phi_\pi - \phi_{\pi'})}}$$
$$\propto \sqrt{\sum_{\pi\in\Pi} \lambda_\pi \gamma_\pi (\mathbf{1}\{\pi'(c) = a, \pi(c) \neq a\} + \mathbf{1}\{\pi'(c) \neq a, \pi(c) = a\})}.$$

□

Lemma A.4.7. *Under $\mathcal{E}$, the choice for n_l in the algorithm satisfies*

$$n_l \lesssim \min_{w\in\Omega} \max_{\pi\in\Pi} \frac{\|\phi_{\pi_*} - \phi_\pi\|^2_{A(w)^{-1}} \log(1/\delta_l)}{\epsilon_l^2 + \Delta(\pi)^2}.$$

Proof.

$$
\begin{aligned}
&h_l(\lambda^l, \gamma^l, n_l) \\
&= \sum_{\pi\in\Pi} [\lambda^l]_\pi \cdot \left(-\widehat{\Delta}_{l-1}^{\gamma^{l-1}}(\pi, \widehat{\pi}_{l-1}) + \frac{\log(1/\delta_l)}{[\gamma^l]_\pi n} \right) + \mathbb{E}_{c\sim\nu_{\mathcal{D}}} \left[\left(\sum_{a\in\mathcal{A}} \sqrt{(\lambda^l \odot \gamma^l)^\top (t_a^{(c)} + \eta_l)} \right)^2 \right] \\
&\le \max_{\lambda\in\triangle_\Pi} \min_{\gamma} \sum_{\pi\in\Pi} \lambda_\pi \cdot \left(-\widehat{\Delta}_{l-1}^{\gamma^{l-1}}(\pi, \widehat{\pi}_{l-1}) + \frac{\log(1/\delta_l)}{\gamma_\pi n} \right) + \mathbb{E}_{c\sim\nu} \left[\left(\sum_{a\in\mathcal{A}} \sqrt{(\lambda \odot \gamma)^\top (t_a^{(c)} + \eta_l)} \right)^2 \right] + \frac{1}{4}\epsilon_l \\
&\qquad\qquad \text{(by Theorem A.5.2, the saddle point argument)} \\
&\le \max_{\lambda\in\triangle_\Pi} \min_{\gamma} \sum_{\pi\in\Pi} \lambda_\pi \cdot \left(-\widehat{\Delta}_{l-1}^{\gamma^{l-1}}(\pi, \widehat{\pi}_{l-1}) + \frac{\log(1/\delta_l)}{\gamma_\pi n} \right) + \mathbb{E}_{c\sim\nu} \left[\left(\sum_{a\in\mathcal{A}} \sqrt{(\lambda \odot \gamma)^\top t_a^{(c)}} \right)^2 \right] + \frac{1}{2}\epsilon_l \\
&\qquad\qquad \text{(by Lemma A.6.3, controlling the bias)} \\
&= \max_{\lambda\in\triangle_\Pi} \min_{w\in\Omega} \min_{\gamma\in\mathbb{R}_+^{|\Pi|}} \sum_{\pi\in\Pi} \lambda_\pi \cdot \left(-\widehat{\Delta}_{l-1}^{\gamma^{l-1}}(\pi, \widehat{\pi}_{l-1}) + \gamma_\pi \left\| \phi_{\widehat{\pi}_{l-1}} - \phi_\pi \right\|_{A(w)^{-1}}^2 + \frac{\log(1/\delta_l)}{\gamma_\pi n} \right) + \frac{1}{2}\epsilon_l \\
&\qquad\qquad \text{(by Lemma A.4.6, the definition of } w) \\
&= \min_{w\in\Omega} \max_{\pi\in\Pi} \min_{\gamma>0} -\frac{1}{8}\widehat{\Delta}_{l-1}^{\gamma^{l-1}}(\pi, \widehat{\pi}_{l-1}) + 8\gamma \left\| \phi_{\widehat{\pi}_{l-1}} - \phi_\pi \right\|_{A(w)^{-1}}^2 + 8\frac{\log(1/\delta_l)}{\gamma n_l} + \frac{1}{2}\epsilon_l \\
&\qquad\qquad \text{(by Lemma A.5.17, the strong duality)} \\
&\le \min_{w\in\Omega} \max_{\pi\in\Pi} \min_{\gamma} \left(-\frac{3}{32}\Delta(\pi, \pi_*) + 8\gamma \left\| \phi_{\widehat{\pi}_{l-1}} - \phi_\pi \right\|_{A(w)^{-1}}^2 + 8\frac{\log(1/\delta_l)}{\gamma n_l} \right) + \frac{3}{4}\epsilon_l \quad \text{(by Lemma A.4.5)} \\
&\le \min_{w\in\Omega} \max_{\pi\in\Pi} \left(-\frac{3}{32}\Delta(\pi, \pi_*) + 16\sqrt{\frac{\left\| \phi_{\widehat{\pi}_{l-1}} - \phi_\pi \right\|_{A(w)^{-1}}^2 \log(1/\delta_l)}{n_l}} \right) + \frac{3}{4}\epsilon_l \\
&\le \min_{w\in\Omega} \max_{\pi\in\Pi} \left(-\frac{3}{32}\Delta(\pi, \pi_*) + 16\sqrt{\frac{\|\phi_{\pi_*} - \phi_\pi\|_{A(w)^{-1}}^2 \log(1/\delta_l)}{n_l}} \right. \\
&\qquad \left. +16\sqrt{\frac{\left\| \phi_{\pi_*} - \phi_{\widehat{\pi}_{l-1}} \right\|_{A(w)^{-1}}^2 \log(1/\delta_l)}{n_l}} \right) + \frac{3}{4}\epsilon_l \\
&\le \min_{w\in\Omega} \max_{\pi\in\Pi} \left(-\frac{3}{32}\Delta(\pi, \pi_*) + 16\sqrt{\frac{\|\phi_{\pi_*} - \phi_\pi\|_{A(w)^{-1}}^2 \log(1/\delta_l)}{n_l}} \right. \\
&\qquad \left. +16\sqrt{\max_{\pi'\in S_{l-1}} \frac{\|\phi_{\pi_*} - \phi_{\pi'}\|_{A(w)^{-1}}^2 \log(1/\delta_l)}{n_l}} \right) + \frac{3}{4}\epsilon_l.
\end{aligned}
$$

which is less than ϵ_l whenever

$$n_l \gtrsim \min_{w\in\Omega}\max_{\pi\in\Pi} \frac{\|\phi_{\pi_*} - \phi_\pi\|^2_{A(w)^{-1}} \log(1/\delta_l)}{\epsilon_l^2 + \Delta(\pi)^2}. \tag{A.8}$$

□

A.5 Convergence analysis of FW-GD

A.5.1 Statement of the convergence results

In this section, we will characterize the performance of Algorithm 8, a.k.a. Algorithm 4. Our goal is to show two results: the duality gap converges to zero, and our algorithm converges to the saddle point. It is known that Frank-Wolfe algorithm directly deals with the duality gap [Pedregosa et al., 2020], so we will define our primal and dual problem in what follows. Since we are computing n_l via binning, in each inner loop n is fixed. Then, we define our dual objective the same as (A.5) with the shorthand notation $h_l(\lambda, \gamma) := h_l(\lambda, \gamma, n)$. We formulate our primal objective as

$$\mathcal{P}_l(w(\lambda,\gamma),\gamma) := \max_{\pi\in\Pi}\left(-\widehat{\Delta}^{\gamma^{l-1}}_{l-1}(\pi, \widehat{\pi}_{l-1}) + \gamma_\pi \left\|\phi_\pi - \phi_{\widehat{\pi}_{l-1}}\right\|^2_{A(w(\lambda,\gamma))^{-1}} + \frac{\log(1/\delta_l)}{\gamma_\pi n}\right), \tag{A.9}$$

where $w(\lambda,\gamma) \in \mathbb{R}^{|\mathcal{A}|\times|\mathcal{C}|}$ such that

$$[w(\lambda,\gamma)]_{a,c} = \nu_c \cdot p_{c,a} = \nu_c \cdot \frac{\sqrt{(\lambda\odot\gamma)^\top (t_a^{(c)} + \eta)}}{\sum_{a'\in\mathcal{A}} \sqrt{(\lambda\odot\gamma)^\top (t_{a'}^{(c)} + \eta)}}. \tag{A.10}$$

Then we will show those two results. First, Theorem A.5.1 bounds the duality gap of the primal and dual objective. Second, Theorem A.5.2 shows that Algorithm 4 converges to a saddle point.

Theorem A.5.1. *For any $l \in \mathbb{N}$, with the number of FW-GD iterations $K_l = O(L^2\epsilon_l^{-2})$ where $L = |\mathcal{A}|^2 \frac{((1+\eta_l)\gamma_{\max})^{5/2}}{\eta_l^{3/2}\gamma_{\min}^2}$, we have*

$$\left|\mathcal{P}_l(w(\lambda^l,\gamma^l),\gamma^l) - h_l(\lambda^l,\gamma^l)\right| \le \epsilon_l.$$

Moreover, K_l depends at most polynomially on $|\mathcal{A}|, \epsilon_l^{-1}, \log(1/\delta_l)$.

Proof. First, Lemma A.6.2 shows that for any λ, γ, and n, $h_l(\lambda, \gamma, n) = \langle \lambda, \nabla_\lambda h_l(\lambda, \gamma, n) \rangle$. Therefore, at some iteration t, the Frank-Wolfe gap

$$g_t = \left\langle \nabla_\lambda h_l(\lambda^t, \gamma^t), \mathbf{e}_{\pi_t} - \lambda^t \right\rangle = \max_{\pi \in \Pi} [\nabla_\lambda h_l(\lambda^t, \gamma^t)]_\pi - h_l(\lambda^t, \gamma^t).$$

Lemma A.5.6 shows that with a small choice of the regularization parameter the primal objective is close to the maximum component of the gradient, i.e. $|\mathcal{P}_l(w(\lambda^l, \gamma^l), \gamma^l) - \max_{\pi \in \Pi} [\nabla_\lambda h_l(\lambda^l, \gamma^l)]_\pi| \leq \frac{\epsilon_l}{2}$. Also, Lemma A.5.5 shows that if $t \geq L^2 \epsilon_l^{-2}$ is large enough, the Frank-Wolfe gap is bounded by ϵ_l. Combining these two lemmas, for $t \geq L^2 \epsilon_l^{-2}$, we have

$$\begin{aligned}
&|\mathcal{P}_l(w(\lambda^l, \gamma^l), \gamma^l) - h_l(\lambda^l, \gamma^l)| \\
&\leq |\mathcal{P}_l(w(\lambda^l, \gamma^l), \gamma^l) - \max_{\pi \in \Pi} [\nabla_\lambda h_l(\lambda^l, \gamma^l)]_\pi| + |h_l(\lambda^l, \gamma^l) - \max_{\pi \in \Pi} [\nabla_\lambda h_l(\lambda^l, \gamma^l)]_\pi| \\
&\leq |\mathcal{P}_l(w(\lambda, \gamma), \gamma) - \max_{\pi \in \Pi} [\nabla_\lambda h_l(\lambda, \gamma)]_\pi| + g_l \\
&\leq \frac{\epsilon_l}{2} + \frac{\epsilon_l}{2} = \epsilon_l.
\end{aligned}$$

Finally, we conclude that $K_l = \text{poly}(|\mathcal{A}|, \epsilon_l^{-1}, \log(1/\delta_l))$ since $\gamma_{\max} = O(|\mathcal{A}|^{-1} \eta_l^{-1/2})$, $\gamma_{\min} = O(\sqrt{\eta_l})$, and $\eta_l = O(|\mathcal{A}|^{-4} \epsilon_l^2)$ all depends polynomially on $|\mathcal{A}|$ and ϵ_l^{-1}. This shows Theorem A.5.1. □

We now have the second main result of this section.

Theorem A.5.2. *For any l, with $K_l = \text{poly}(|\mathcal{A}|, \epsilon_l^{-1}, \log(1/\delta_l))$ and the size of the history $\mathcal{D} \geq \text{poly}(|\mathcal{A}|, \epsilon^{-1}, \log(1/\delta), \log(|\Pi|))$, Algorithm 4 converges to a saddle point, i.e.*

$$\left| \max_{\lambda \in \Delta_\Pi} \min_{\gamma \in [\gamma_{\min}, \gamma_{\max}]^\Pi} h_l(\lambda, \gamma) - h_l(\lambda^l, \gamma^l) \right| \leq \epsilon_l.$$

Proof. Note that

$$
\begin{aligned}
&\mathcal{P}_l(w(\lambda^l,\gamma^l),\gamma^l)\\
&=\max_{\pi\in\Pi}\left[-\widehat{\Delta}_{l-1}^{\gamma^{l-1}}(\pi,\widehat{\pi}_{l-1})+\frac{\log(1/\delta_l)}{[\gamma^l]_\pi n}+[\gamma^l]_\pi\left\|\phi_{\widehat{\pi}_{l-1}}-\phi_\pi\right\|^2_{A(w(\lambda^l,\gamma^l))^{-1}}\right]\\
&\geq\max_{\pi\in\Pi}\min_{\gamma}\left[-\widehat{\Delta}_{l-1}^{\gamma^{l-1}}(\pi,\widehat{\pi}_{l-1})+\frac{\log(1/\delta_l)}{\gamma_\pi n}+\gamma_\pi\left\|\phi_{\widehat{\pi}_{l-1}}-\phi_\pi\right\|^2_{A(w(\lambda^l,\gamma^l))^{-1}}\right]\\
&\geq\min_{w\in\Omega}\max_{\pi\in\Pi}\min_{\gamma}\left[-\widehat{\Delta}_{l-1}^{\gamma^{l-1}}(\pi,\widehat{\pi}_{l-1})+\frac{\log(1/\delta_l)}{\gamma_\pi n}+\gamma_\pi\left\|\phi_{\widehat{\pi}_{l-1}}-\phi_\pi\right\|^2_{A(w)^{-1}}\right]\\
&=\max_{\lambda\in\triangle_\Pi}\min_{w\in\Omega}\min_{\gamma\in[\gamma_{\min},\gamma_{\max}]^\Pi}\sum_{\pi\in\Pi}\lambda_\pi\left(-\widehat{\Delta}_{l-1}^{\gamma^{l-1}}(\pi,\widehat{\pi}_{l-1})+\frac{\log(1/\delta_l)}{\gamma_\pi n}+\gamma_\pi\left\|\phi_{\widehat{\pi}_{l-1}}-\phi_\pi\right\|^2_{A(w)^{-1}}\right)\\
&\qquad\qquad\text{(by Lemma A.5.17, strong duality)}\\
&=\max_{\lambda\in\triangle_\Pi}\min_{\gamma\in[\gamma_{\min},\gamma_{\max}]^\Pi}\sum_{\pi}\lambda_\pi\cdot\left(-\widehat{\Delta}_{l-1}^{\gamma^{l-1}}(\pi,\widehat{\pi}_{l-1})+\frac{\log(1/\delta_l)}{\gamma_\pi n}\right)\\
&\qquad\qquad+\mathbb{E}_{c\sim\nu}\left[\left(\sum_a\sqrt{(\lambda\odot\gamma)^\top t_a^{(c)}}\right)^2\right]\qquad\text{(by Lemma A.4.6)}\\
&\geq\max_{\lambda\in\triangle_\Pi}\min_{\gamma\in[\gamma_{\min},\gamma_{\max}]^\Pi}\sum_{\pi\in\Pi}\lambda_\pi\cdot\left(-\widehat{\Delta}_{l-1}^{\gamma^{l-1}}(\pi,\widehat{\pi}_{l-1})+\frac{\log(1/\delta_l)}{\gamma_\pi n}\right)\\
&\qquad\qquad+\mathbb{E}_{c\sim\nu}\left[\left(\sum_{a\in\mathcal{A}}\sqrt{(\lambda\odot\gamma)^\top(t_a^{(c)}+\eta_l)}\right)^2\right]-\frac{1}{2}\epsilon_l\qquad\text{(by Lemma A.6.3)}\\
&\geq\min_{\gamma\in[\gamma_{\min},\gamma_{\max}]^\Pi}\sum_{\pi}\left[\lambda^l\right]_\pi\cdot\left(-\widehat{\Delta}_{l-1}^{\gamma^{l-1}}(\pi,\widehat{\pi}_{l-1})+\frac{\log(1/\delta_l)}{\gamma_\pi n}\right)\\
&\qquad\qquad+\mathbb{E}_{c\sim\nu}\left[\left(\sum_{a\in\mathcal{A}}\sqrt{(\lambda^l\odot\gamma)^\top(t_a^{(c)}+\eta_l)}\right)^2\right]-\frac{1}{2}\epsilon_l\\
&\geq\min_{\gamma\in[\gamma_{\min},\gamma_{\max}]^\Pi}\sum_{\pi}\left[\lambda^l\right]_\pi\cdot\left(-\widehat{\Delta}_{l-1}^{\gamma^{l-1}}(\pi,\widehat{\pi}_{l-1})+\frac{\log(1/\delta_l)}{\gamma_\pi n}\right)\\
&\qquad\qquad+\mathbb{E}_{c\sim\nu_{\mathcal{D}}}\left[\left(\sum_{a\in\mathcal{A}}\sqrt{(\lambda^l\odot\gamma)^\top(t_a^{(c)}+\eta_l)}\right)^2\right]-\frac{3}{4}\epsilon_l\\
&\qquad\qquad\text{(by Lemma A.3.5, controlling the history)}\\
&\geq\sum_{\pi}\left[\lambda^l\right]_\pi\cdot\left(-\widehat{\Delta}_{l-1}^{\gamma^{l-1}}(\pi,\widehat{\pi}_{l-1})+\frac{\log(1/\delta_l)}{[\gamma^l]_\pi n}\right)\\
&\qquad\qquad+\mathbb{E}_{c\sim\nu_{\mathcal{D}}}\left[\left(\sum_{a\in\mathcal{A}}\sqrt{(\lambda^l\odot\gamma^l)^\top(t_a^{(c)}+\eta_l)}\right)^2\right]-\epsilon_l\\
&\qquad\qquad\text{(by Lemma A.5.7, the GD convergence)}
\end{aligned}
$$

In other words,

$$\mathcal{P}_l(w(\lambda^l,\gamma^l),\gamma^l) \geq \max_{\lambda\in\Delta_\Pi} \min_{\gamma\in[\gamma_{\min},\gamma_{\max}]^\Pi} h_l(\lambda,\gamma) \geq h_l(\lambda^l,\gamma^l) - \epsilon_l.$$

On the other hand, by Theorem A.5.1, we have $\mathcal{P}_l(w(\lambda^l,\gamma^l),\gamma^l) \leq h_l(\lambda^l,\gamma^l) + \epsilon_l$. Therefore, we have

$$\max_{\lambda\in\Delta_\Pi} \min_{\gamma\in[\gamma_{\min},\gamma_{\max}]^\Pi} h_l(\lambda,\gamma) \in \left[h_l(\lambda^l,\gamma^l) - \epsilon_l, h_l(\lambda^l,\gamma^l) + \epsilon_l\right]$$

and so we have our result. □

A.5.2 Technical proofs

Guarantees on γ

We first provides some guarantees of γ and the convergence of the GD subroutine.

Lemma A.5.3. *Consider a fixed n. Let $\gamma^* = \arg\min_\gamma h_l(\lambda,\gamma,n)$. Then we have for all i,*

$$[\gamma^*]_i \in \left[\frac{1}{3}\sqrt{\frac{\eta_l \log(1/\delta_l)}{n}}, \min\left\{\sqrt{\frac{\log(1/\delta_l)}{2n\mathbb{E}_c[\mathbf{1}\{\pi(c)\neq\pi^*(c)\}]}}, \sqrt{\frac{\log(1/\delta_l)}{|\mathcal{A}|^2\eta_l n}}\right\}\right].$$

Proof.

$$\begin{aligned}
&[\nabla_\gamma h_l(\lambda,\gamma)]_\pi \\
&= \mathbb{E}_c\left[\left(\sum_{a\in\mathcal{A}}\sqrt{(\lambda\odot\gamma)^\top(t_a^{(c)}+\eta_l)}\right)\cdot\left(\sum_{a'\in\mathcal{A}}\frac{\lambda_\pi([t_{a'}^{(c)}]_\pi+\eta_l)}{\sqrt{(\lambda\odot\gamma)^\top(t_{a'}^{(c)}+\eta_l)}}\right)\right] - \frac{\lambda_\pi\log(1/\delta_l)}{\gamma_\pi^2 n} \\
&\geq \mathbb{E}_c\left[\left(\sum_{a\in\mathcal{A}}\sqrt{\lambda_\pi([t_a^{(c)}]_\pi+\eta_l)}\right)^2\right] - \frac{\lambda_\pi\log(1/\delta_l)}{\gamma_\pi^2 n} \\
&\geq |\mathcal{A}|^2\eta_l\lambda_\pi + 2\lambda_\pi\mathbb{E}_c[\mathbf{1}\{\pi(c)\neq\pi^*(c)\}] - \frac{\lambda_\pi\log(1/\delta_l)}{\gamma_\pi^2 n},
\end{aligned}$$

where the first to second line follows from Cauchy-Schwartz - $(\sum_a x_a)\sum_a\left(\frac{y_a}{x_a}\right) \geq (\sum_a\sqrt{y_a})^2$. We first solve $\frac{\lambda_\pi\log(1/\delta_l)}{\gamma_\pi^2 n} < |\mathcal{A}|^2\eta_l\lambda_\pi$ and get $\gamma_\pi > \sqrt{\frac{\log(1/\delta_l)}{|\mathcal{A}|^2\eta_l n}}$. We also solve $\frac{\lambda_\pi\log(1/\delta_l)}{\gamma_\pi^2 n} < 2\lambda_\pi\mathbb{E}_c[\mathbf{1}\{\pi(c)\neq\pi^*(c)\}]$ and get $\gamma_\pi < \sqrt{\frac{\log(1/\delta_l)}{2n\mathbb{E}_c[\mathbf{1}\{\pi(c)\neq\pi^*(c)\}]}}$. Therefore, the πth component of

the gradient is always positive whenever $\gamma_\pi > \min\left\{\sqrt{\frac{\log(1/\delta_l)}{2n\mathbb{E}_c[\mathbf{1}\{\pi(c)\neq\pi^*(c)\}]}}, \sqrt{\frac{\log(1/\delta_l)}{|\mathcal{A}|^2\eta_l n}}\right\}$. Therefore, the minimum γ should have $\gamma_\pi \leq \min\left\{\sqrt{\frac{\log(1/\delta_l)}{2n\mathbb{E}_c[\mathbf{1}\{\pi(c)\neq\pi^*(c)\}]}}, \sqrt{\frac{\log(1/\delta_l)}{|\mathcal{A}|^2\eta_l n}}\right\}$. On the other hand, let $s = \arg\min_\pi \gamma_\pi$. Then,

$$\eta_l\gamma_s \leq (\lambda\odot\gamma)^\top(t_a^{(c)}+\eta_l) = \left(\lambda\odot(t_a^{(c)}+\eta_l)\right)^\top\gamma \leq \left\|\lambda\odot(t_a^{(c)}+\eta_l)\right\|_1\cdot\|\gamma\|_\infty.$$

Then

$$\sum_{a\in\mathcal{A}}\sqrt{(\lambda\odot\gamma)^\top(t_a^{(c)}+\eta_l)} \leq \sum_{a\in\mathcal{A}}\sqrt{\left\|\lambda\odot(t_a^{(c)}+\eta_l)\right\|_1}\cdot\sqrt{\|\gamma\|_\infty}.$$

Note that

$$\begin{aligned}\left(\sum_{a\in\mathcal{A}}\sqrt{\left\|\lambda\odot(t_a^{(c)}+\eta_l)\right\|_1}\right)^2 &= \left(\sum_{a\in\mathcal{A}}\sqrt{\lambda^\top(t_a^{(c)}+\eta_l)}\right)^2\\ &\leq \left(\sum_{a\in\mathcal{A}}\lambda^\top(t_a^{(c)}+\eta_l)\right)|\mathcal{A}|\\ &\leq |\mathcal{A}|(1+\eta_l).\end{aligned}$$

Since for any π, $\sum_{a'\in\mathcal{A}}[t_{a'}^{(c)}]_\pi \leq 2$, so

$$[\nabla_\gamma h_l(\lambda,\gamma)]_\pi \leq \sqrt{|\mathcal{A}|(1+\eta_l)\|\gamma\|_\infty}\cdot\frac{(2+\eta_l)\lambda_\pi}{\sqrt{\eta_l\gamma_s}} - \frac{\lambda_\pi\log(1/\delta_l)}{\gamma_\pi^2 n}.$$

Let $\pi = s$, then by the fact that $\|\gamma\|_\infty \leq \sqrt{\frac{\log(1/\delta_l)}{|\mathcal{A}|^2\eta_l n}}$, we have

$$[\nabla_\gamma h_l(\lambda,\gamma)]_s \leq \sqrt{|\mathcal{A}|(1+\eta_l)}\left(\frac{\log(1/\delta_l)}{|\mathcal{A}|^2\eta_l n}\right)^{1/4}\cdot\frac{(2+\eta_l)\lambda_s}{\sqrt{\eta_l\gamma_s}} - \frac{\lambda_s\log(1/\delta_l)}{\gamma_s^2 n}.$$

We solve $\sqrt{|\mathcal{A}|(1+\eta_l)}\left(\frac{\log(1/\delta_l)}{|\mathcal{A}|^2\eta_l n}\right)^{1/4}\cdot\frac{(2+\eta_l)\lambda_s}{\sqrt{\eta_l\gamma_s}} - \frac{\lambda_s\log(1/\delta_l)}{\gamma_s^2 n} < 0$. Then we get

$$\gamma_s < (1+\eta_l)^{-1/3}(2+\eta_l)^{-2/3}\sqrt{\frac{\eta_l\log(1/\delta_l)}{n}}.$$

Since $(1+\eta_l)^{-1/3}(2+\eta_l)^{-2/3} > \frac{1}{3}$ whenever $\eta_l \leq 1$, the sth component of the gradient is negative whenever $\gamma_s < \frac{1}{3}\sqrt{\frac{\eta_l\log(1/\delta_l)}{n}}$. Therefore, $\min_\pi\gamma_\pi \geq \frac{1}{3}\sqrt{\frac{\eta_l\log(1/\delta_l)}{n}}$. □

Convergence of Frank-Wolfe gap

Lemma A.5.4 and A.5.5 shows that the Frank-Wolfe gap is small. The proof technique follows from the general Frank-Wolfe analysis.

Lemma A.5.4. *For any $\xi \in [0,1]$, any t, with $L = |\mathcal{A}|^2 \frac{((1+\eta_l)\gamma_{\max})^{5/2}}{\eta_l^{3/2}\gamma_{\min}^2}$, we have $h_l(\lambda^{t+1}, \gamma^{t+1}) \geq h_l(\lambda^t, \gamma^t) + \xi g_t - \frac{1}{2}\xi^2 L - \kappa_t$.*

Proof. By L-Lipschitz gradient condition of $-h_\ell$ in λ given in Lemma A.5.12 we have

$$-h_l(\lambda^{t+1}, \gamma^{t+1}) \leq -h_l(\lambda^t, \gamma^{t+1}) - \left\langle \nabla_\lambda h_l(\lambda^t, \gamma^{t+1}), \lambda^{t+1} - \lambda^t \right\rangle + \frac{L}{2} \left\| \lambda^{t+1} - \lambda^t \right\|_1^2.$$

Therefore,

$$h_l(\lambda^{t+1}, \gamma^{t+1}) \geq h_l(\lambda^t, \gamma^{t+1}) + \left\langle \nabla_\lambda h_l(\lambda^t, \gamma^{t+1}), \lambda^{t+1} - \lambda^t \right\rangle - \frac{L}{2} \left\| \lambda^{t+1} - \lambda^t \right\|_1^2.$$

Plugging in $\lambda^{t+1} = (1-\beta_t)\lambda^t + \beta_t \mathbf{e}_{\pi_t}$ as in line 8 of Algorithm 3, we have

$$\begin{aligned}
&h_l((1-\beta_t)\lambda^t + \beta_t \mathbf{e}_{\pi_t}, \gamma^{t+1}) \\
&\geq h_l(\lambda^t, \gamma^{t+1}) + \left\langle \nabla_\lambda h_l(\lambda^t, \gamma^{t+1}), (1-\beta_t)\lambda^t + \beta_t \mathbf{e}_{\pi_t} - \lambda^t \right\rangle - \frac{L}{2} \left\| (1+\beta_t)\lambda^t - \beta_t \mathbf{e}_{\pi_t} - \lambda^t \right\|_1^2 \\
&= h_l(\lambda^t, \gamma^{t+1}) + \beta_t \left\langle \nabla_\lambda h_l(\lambda^t, \gamma^{t+1}), \mathbf{e}_{\pi_t} - \lambda^t \right\rangle - \frac{L\beta_t^2}{2} \left\| \mathbf{e}_{\pi_t} - \lambda^t \right\|_1^2 \\
&= h_l(\lambda^t, \gamma^{t+1}) + \beta_t g_t - \frac{L\beta_t^2}{2} \left\| \mathbf{e}_{\pi_t} - \lambda^t \right\|_1^2.
\end{aligned}$$

Choose $\beta_t := \arg\max_{\xi \in [0,1]} \{\xi g_t - \frac{\xi^2 L}{2} \left\| \mathbf{e}_{\pi_t} - \lambda^t \right\|_1^2\}$. Plugging in this expression gives us

$$\begin{aligned}
h_l(\lambda^{t+1}, \gamma^{t+1}) &\geq h_l(\lambda^t, \gamma^{t+1}) + \beta_t \left\langle \nabla_\lambda h_l(\lambda^t, \gamma^{t+1}), \mathbf{e}_{\pi_t} - \lambda^t \right\rangle - \frac{L\beta_t^2}{2} \left\| \mathbf{e}_{\pi_t} - \lambda^t \right\|_1^2 \\
&= h_l(\lambda^t, \gamma^{t+1}) + \max_{\xi \in [0,1]} \{\xi g_t - \frac{\xi^2 L}{2} \left\| \mathbf{e}_{\pi_t} - \lambda^t \right\|_1^2\} \\
&\geq h_l(\lambda^t, \gamma^{t+1}) + \xi g_t - \frac{\xi^2 L}{2}
\end{aligned}$$

for any $\xi \in [0,1]$ since $\left\| \mathbf{e}_{\pi_t} - \lambda^t \right\|_1^2 \leq 1$. Also, by construction of γ^{t+1} and Lemma A.5.7, we have

$$h_l(\lambda^t, \gamma^{t+1}) \geq \min_\gamma h_l(\lambda^t, \gamma) \geq h_l(\lambda^t, \gamma^t) - \kappa_t.$$

Therefore, our result follows. □

Lemma A.5.5. *We have for any t, with $L = |\mathcal{A}|^2 \frac{((1+\eta_l)\gamma_{\max})^{5/2}}{\eta_l^{3/2}\gamma_{\min}^2}$, $\min_{i\in[1,t]} g_i \leq \frac{L}{\sqrt{t+1}}$.*

Proof. With Lemma A.5.4, we have

$$h_l(\lambda^{t+1}, \gamma^{t+1}, n_r) \geq h_l(\lambda^t, \gamma^t, n_r) + \xi g_t - \frac{1}{2}\xi^2 L - \kappa_t.$$

Plugging in the choice $\xi = \min\{\frac{g_t}{L}, 1\}$, we have $h_l(\lambda^{t+1}, \gamma^{t+1}, n_r) \geq h_l(\lambda^t, \gamma^t, n_r) + \frac{g_t}{2}\min\{\frac{g_t}{L}, 1\} - \kappa_t$. Summing this up from 0 to t gives us

$$\begin{aligned} h_l(\lambda^{t+1}, \gamma^{t+1}, n_r) - h_l(\lambda_0, \gamma_0, n_r) &\geq \sum_{i=0}^{t} \frac{g_i}{2}\min\{\frac{g_i}{L}, 1\} - \delta_i \\ &\geq (t+1) g_t^* \min\{\frac{g_t^*}{L}, 1\} - \sum_{i=0}^{t}\delta_i. \end{aligned}$$

where $g_t^* = \min_{i=0,\cdots,t} g_i$. Then, as long as $\sum_{i=0}^{t}\delta_i \leq \epsilon_l$, by the fact that $h_l(\lambda^{t+1}, \gamma^{t+1}) - h_l(\lambda_0, \gamma_0) \leq \max_{\lambda\in\triangle_\Pi}\min_\gamma h_l(\lambda, \gamma) - h_l(\lambda_0, \gamma_0) < \infty$. Therefore, we have $\min_{i\in[1,t]} g_i \leq \frac{L}{\sqrt{t+1}}$. □

Connect the Frank-Wolfe gap to the duality gap

Lemma A.5.6 shows that the primal objective is approximately the maximum component of the gradient of the dual objective, which simplifies our Frank-Wolfe gap expression.

Lemma A.5.6. *Consider some $\lambda \in \triangle_\Pi$, $\gamma \in \mathbb{R}_+^{|\Pi|}$, and $n \in \mathbb{N}$. For $\eta_l < |\mathcal{A}|^{-4}\epsilon_l^2$, we have $|\mathcal{P}_l(w(\lambda^l, \gamma^l), \gamma^l) - \max_{\pi\in\Pi}[\nabla_\lambda h_l(\lambda^l, \gamma^l)]_\pi| \leq \epsilon_l$.*

Proof. Observe that for any $\pi, \pi' \in \Pi$ and any γ,

$$
\begin{aligned}
&\gamma_\pi \left\|\phi_{\pi'} - \phi_\pi\right\|^2_{A(w(\lambda,\gamma))^{-1}} \\
&= \gamma_\pi \sum_{a,c} \frac{\nu_c^2}{[w(\lambda,\gamma)]_{a,c}} \left(\mathbf{1}\{\pi'(c) = a, \pi(c) \neq a\} + \mathbf{1}\{\pi'(c) \neq a, \pi(c) = a\}\right) \\
&= \gamma_\pi \sum_{c} \nu_c \sum_{a} \left(\frac{\nu_c}{[w(\lambda,\gamma)]_{a,c}} \left(\mathbf{1}\{\pi'(c) = a, \pi(c) \neq a\} + \mathbf{1}\{\pi'(c) \neq a, \pi(c) = a\}\right)\right) \\
&= \gamma_\pi \mathbb{E}_{c\sim\nu} \left[\sum_a \frac{\sum_{a' \in \mathcal{A}} \sqrt{(\lambda \odot \gamma)^\top (t_{a'}^{(c)} + \eta_l)}}{\sqrt{(\lambda \odot \gamma)^\top (t_a^{(c)} + \eta_l)}} \left(\mathbf{1}\{\pi'(c) = a, \pi(c) \neq a\} + \mathbf{1}\{\pi'(c) \neq a, \pi(c) = a\}\right)\right] \\
&= \mathbb{E}_{c\sim\nu} \left[\sum_a \frac{\sum_{a' \in \mathcal{A}} \sqrt{(\lambda \odot \gamma)^\top (t_{a'}^{(c)} + \eta_l)}}{\sqrt{(\lambda \odot \gamma)^\top (t_a^{(c)} + \eta_l)}} (\gamma_\pi [t_a^{(c)}]_\pi)\right].
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
&\mathcal{P}_l(w(\lambda^l, \gamma^l), \gamma^l) \\
&= \max_{\pi \in \Pi} \left\{ -\widehat{\Delta}_{l-1}^{\gamma^{l-1}}(\pi) + [\gamma^l]_\pi \left\|\phi_\pi - \phi_{\widehat{\pi}_{l-1}}\right\|^2_{A(w(\lambda^l,\gamma^l))^{-1}} + \frac{\log(1/\delta_l)}{[\gamma^l]_\pi n} \right\} \\
&= \max_{\pi \in \Pi} \left\{ -\widehat{\Delta}_{l-1}^{\gamma^{l-1}}(\pi) + \mathbb{E}_{c\sim\nu} \left[\sum_a \frac{\sum_{a' \in \mathcal{A}} \sqrt{(\lambda^l \odot \gamma^l)^\top (t_{a'}^{(c)} + \eta_l)}}{\sqrt{(\lambda^l \odot \gamma^l)^\top (t_a^{(c)} + \eta_l)}} ([\gamma^l]_\pi [t_a^{(c)}]_\pi)\right] + \frac{\log(1/\delta_l)}{[\gamma^l]_\pi n} \right\}.
\end{aligned}
$$

Lemma A.3.6 guarantees that we could replace the expectation over context to history of contexts $\nu_{\mathcal{D}}$ without incurring much error. In particular, for a sufficiently large history $\mathcal{D}$, it guarantees

$$
\begin{aligned}
&\max_{\pi \in \Pi} \left| \mathbb{E}_{c\sim\nu_{\mathcal{D}}} \left[\sum_a \frac{\sum_{a' \in \mathcal{A}} \sqrt{(\lambda^l \odot \gamma^l)^\top (t_{a'}^{(c)} + \eta_l)}}{\sqrt{(\lambda^l \odot \gamma^l)^\top (t_a^{(c)} + \eta_l)}} ([\gamma^l]_\pi [t_a^{(c)}]_\pi)\right] \right. \\
&\left. - \mathbb{E}_{c\sim\nu} \left[\sum_a \frac{\sum_{a' \in \mathcal{A}} \sqrt{(\lambda^l \odot \gamma^l)^\top (t_{a'}^{(c)} + \eta_l)}}{\sqrt{(\lambda^l \odot \gamma^l)^\top (t_a^{(c)} + \eta_l)}} ([\gamma^l]_\pi [t_a^{(c)}]_\pi)\right] \right| \leq \frac{\epsilon_l}{2}.
\end{aligned}
$$

On the other hand,

$$\max_{\pi\in\Pi}\left\{-\widehat{\Delta}_{l-1}^{\gamma^{l-1}}(\pi)+\mathbb{E}_{c\sim\nu_{\mathcal{D}}}\left[\sum_a \frac{\sum_{a'\in\mathcal{A}}\sqrt{(\lambda^l\odot\gamma^l)^\top(t_{a'}^{(c)}+\eta_l)}}{\sqrt{(\lambda^l\odot\gamma^l)^\top(t_a^{(c)}+\eta_l)}}([\gamma^l]_\pi[t_a^{(c)}]_\pi)\right]+\frac{\log(1/\delta_l)}{[\gamma^l]_\pi n}\right\}$$
$$=\max_{\pi\in\Pi}\left\{[\nabla_\lambda h_l(\lambda^l,\gamma^l)]_\pi-\mathbb{E}_{c\sim\nu_{\mathcal{D}}}\left[\sum_a \frac{\sum_{a'\in\mathcal{A}}\sqrt{(\lambda^l\odot\gamma^l)^\top(t_{a'}^{(c)}+\eta_l)}}{\sqrt{(\lambda^l\odot\gamma^l)^\top(t_a^{(c)}+\eta_l)}}[\gamma^l]_\pi\eta_l\right]\right\}.$$

Note that when $\gamma_\pi\in[\gamma_{\min},\gamma_{\max}]$,

$$\mathbb{E}_{c\sim\nu_{\mathcal{D}}}\left[\sum_a \frac{\sum_{a'\in\mathcal{A}}\sqrt{(\lambda\odot\gamma)^\top(t_{a'}^{(c)}+\eta_l)}}{\sqrt{(\lambda\odot\gamma)^\top(t_a^{(c)}+\eta_l)}}\gamma_\pi\eta_l\right]\in\left[0,|\mathcal{A}|^2\sqrt{\frac{\gamma_{\max}(1+\eta_l)}{\gamma_{\min}\eta_l}}\gamma_{\max}\eta_l\right].$$

Therefore, for $\eta_l<|\mathcal{A}|^{-4}\epsilon_l^2$,

$$\left|\mathbb{E}_{c\sim\nu_{\mathcal{D}}}\left[\sum_a \frac{\sum_{a'\in\mathcal{A}}\sqrt{(\lambda^l\odot\gamma^l)^\top(t_{a'}^{(c)}+\eta_l)}}{\sqrt{(\lambda^l\odot\gamma^l)^\top(t_a^{(c)}+\eta_l)}}[\gamma^l]_\pi\eta_l\right]\right|\le\frac{\epsilon_l}{2}.$$

Therefore, we have our results. □

A.5.3 Convergence of gradient descent

In this subsection we show convergence for gradient descent.

Algorithm 9 GD

Input: λ^t, n, κ_t

1: define $\iota^t=\epsilon_l^3 t^{-3}|\mathcal{A}|^{-6}$

2: clip λ and define $\tilde{\lambda}=\mathrm{clip}(\lambda,\iota_t)$

3: run gradient descent of on γ for $h_l(\tilde{\lambda},\gamma,n)$ over $\mathrm{supp}(\tilde{\lambda})$ and output γ^t

Output: γ^t

We will first state the main result of this section.

Lemma A.5.7. *With the number of iterations* $T = O\big(\frac{L_\gamma}{\iota_t} + \frac{1}{\kappa_t \iota_t}\big)$ *with* $L_\gamma = |\mathcal{A}|^2 \frac{((1+\eta_l)\gamma_{\max})^{3/2}}{\eta_l^{3/2}\gamma_{\min}^2} + \frac{2\log(1/\delta_l)}{n\gamma_{\min}^3}$, *we have* $h_l(\lambda, \gamma^t, n) - \min_\gamma h_l(\lambda, \gamma, n) \leq \kappa_t$.

Proof sketch. Lemma A.5.9 shows that this clipping does not affect the function value that much. Since we do not assume our function to be convex for γ, we will show that the stationary point is unique and the gradient is strictly positive around the stationary point. Lemma A.5.14 first shows that our function is locally strongly convex around any stationary point. In particular, if we are at a point where the L_1 norm of the gradient is less than $\lambda_{\min}$, we are locally strongly convex. Lemma A.5.13 shows our gradient is Lipschitz with respect to the L_1 norm. Then, Lemma A.5.8 then shows that the gradient descent algorithm converges to a stationary point. It is the classical argument for gradient descent algorithm on non-convex objectives [Jin et al., 2021].

Lemma A.5.8. *For any* K, *with* $L_\gamma = |\mathcal{A}|^2 \frac{((1+\eta_l)\gamma_{\max})^{3/2}}{\eta_l^{3/2}\gamma_{\min}^2} + \frac{2\log(1/\delta_l)}{n\gamma_{\min}^3}$,

$$\min_{k \leq K} \|\nabla_\gamma h_l(\lambda, \gamma_k, n)\|_1^2 \leq 2L_\gamma \frac{h_l(\lambda, \gamma_0, n) - \min_\gamma h_l(\lambda, \gamma, n)}{K}.$$

With this lemma, we have for a sufficiently large K, the minimum gradient can be made arbitrarily small. In particular, for $K \geq L_\gamma \lambda_{\min}^{-1}$ we have that the minimum gradient has L_1-norm less than $\lambda_{\min}$, and thus we are in a neighborhood of our stationary point by Lemma A.5.15. After that, it takes $O(\frac{1}{\kappa_t \lambda_{\min}})$ steps to converge to a point whose value is at most κ_t away from the value of the stationary point. The results in [Milnor and Weaver, 1997] coupled with Lemma A.5.14 ensure that our stationary point is unique. Intuitively, if we have two locally strongly convex stationary points, there must be a "hill" between them, which also corresponds to a stationary point, but we have shown that all stationary points must be "holes" due to local strong convexity, so the stationary point has to be unique. Thanks to the clipping, we can lower bound $\lambda_{\min}$ by ι_t, so the total number of steps is $\frac{L}{\lambda_{\min}} + \frac{1}{\kappa_t \lambda_{\min}} = \frac{L}{\iota_t} + \frac{1}{\kappa_t \iota_t}$ which matches the result in Lemma A.5.7.

□

Lemma A.5.9. *For some iterate t, let $\iota_t = \epsilon_l^3 t^{-3}|\mathcal{A}|^{-6}$ and denote $\tilde{\lambda} := \mathrm{clip}(\lambda, \iota_t)$ where $[\mathrm{clip}(\lambda, \epsilon)]_\pi := \lambda_\pi \mathbf{1}\{\lambda_\pi \geq \epsilon\}$. Then, for any γ, we have*

$$\left| h_l(\tilde{\lambda}, \gamma, n) - h_l(\lambda, \gamma, n) \right| \leq \kappa_t.$$

Proof. For the first term in h_l, in the case where $\lambda_\pi \geq \iota_t$, $h_l(\lambda, \gamma, n) = h_l(\tilde{\lambda}, \gamma, n)$. When $0 < \lambda_\pi < \iota_t$. We see that

$$\sum_{\pi \in \Pi, \lambda_\pi < \iota_t} \lambda_\pi \left(-\widehat{\Delta}_{l-1}^{\gamma^{l-1}}(\pi, \widehat{\pi}_{l-1}) + \tfrac{\log(1/\delta_l)}{\gamma_\pi n} \right) < t\epsilon \left(\frac{1}{\gamma_{\min}} + \frac{1}{\gamma_{\min}} \right) = \frac{2t\iota_t}{\gamma_{\min}}.$$

Then we focus on the expectation part of $h_l(\lambda, \gamma, n)$. Note that

$$\begin{aligned}
\sqrt{(\lambda \odot \gamma)^\top (t_a^{(c)} + \eta_l)} &= \sqrt{\sum_{\pi, \lambda_\pi \geq \iota_t} \lambda_\pi \gamma_\pi [t_a^{(c)} + \eta_l]_\pi + \sum_{\pi, \lambda_\pi < \iota_t} \lambda_\pi \gamma_\pi [t_a^{(c)} + \eta_l]_\pi} \\
&= \sqrt{(\tilde{\lambda} \odot \gamma)^\top (t_a^{(c)} + \eta_l) + \sum_{\pi, \lambda_\pi < \iota_t} \lambda_\pi \gamma_\pi [t_a^{(c)} + \eta_l]_\pi} \\
&\leq \sqrt{(\tilde{\lambda} \odot \gamma)^\top (t_a^{(c)} + \eta_l) + t\iota_t \gamma_{\max}} \\
&\leq \sqrt{(\tilde{\lambda} \odot \gamma)^\top (t_a^{(c)} + \eta_l)} + \sqrt{t\iota_t \gamma_{\max}}.
\end{aligned}$$

Therefore,

$$\begin{aligned}
&\mathbb{E}\left[\left(\sum_{a \in \mathcal{A}} \sqrt{(\lambda \odot \gamma)^\top (t_a^{(c)} + \eta_l)} \right)^2 \right] - \mathbb{E}\left[\left(\sum_{a \in \mathcal{A}} \sqrt{(\tilde{\lambda} \odot \gamma)^\top (t_a^{(c)} + \eta_l)} \right)^2 \right] \\
&= \mathbb{E}\left[\left(\sum_{a \in \mathcal{A}} \sqrt{(\lambda \odot \gamma)^\top (t_a^{(c)} + \eta_l)} + \sqrt{(\tilde{\lambda} \odot \gamma)^\top (t_a^{(c)} + \eta_l)} \right) \right. \\
&\qquad \left. \left(\sum_{a \in \mathcal{A}} \sqrt{(\lambda \odot \gamma)^\top (t_a^{(c)} + \eta_l)} - \sqrt{(\tilde{\lambda} \odot \gamma)^\top (t_a^{(c)} + \eta_l)} \right) \right] \\
&\leq |\mathcal{A}| \sqrt{\gamma_{\max}} |\mathcal{A}| \sqrt{t\iota_t \gamma_{\max}} \\
&= |\mathcal{A}|^2 \gamma_{\max} \sqrt{t\iota_t}.
\end{aligned}$$

Combining two displays above and plugging in $\gamma_{\min}$ and $\gamma_{\max}$ gives

$$\begin{aligned}\left|h_l(\tilde{\lambda},\gamma,n)-h_l(\lambda,\gamma,n)\right| &\le \frac{2t\iota_t}{\gamma_{\min}}+|\mathcal{A}|\sqrt{\frac{t\iota_t}{\eta_l}}\\ &= \frac{2t\iota_t|\mathcal{A}|\epsilon_l^{-1}}{\sqrt{\eta_l}}+|\mathcal{A}|\sqrt{\frac{t\iota_t}{\eta_l}}.\end{aligned}$$

Let RHS be κ_t and solve for ι_t we get $\iota_t \le \min\{\frac{\sqrt{\eta_l}\kappa_t\epsilon_l}{2t|\mathcal{A}|}, \frac{\eta_l\kappa_t}{|\mathcal{A}|^2 t}\}$. Plugging in $\eta_l = |\mathcal{A}|^{-4}\epsilon_l^2$ gives the result. □

Lemma A.5.10. *Suppose γ^t satisfies that $h_l(\tilde{\lambda},\gamma^t,n)-\min_\gamma h_l(\tilde{\lambda},\gamma,n)\le\kappa_t$, then we also have $h_l(\lambda,\gamma^t,n)-\min_\gamma h_l(\lambda,\gamma,n)\le\kappa_t$, i.e. γ^t satisfies the desired property.*

Proof. Let $\tilde{\gamma}_* = \arg\min_\gamma h_l(\tilde{\lambda},\gamma,n)$ and $\gamma_* = \arg\min_\gamma h_l(\lambda,\gamma,n)$. The result follows from applying Lemma A.5.9 twice on $h_l(\tilde{\lambda},\gamma^t,n)$ and $h_l(\tilde{\lambda},\gamma_*,n)$. In particular,

$$\begin{aligned}h_l(\lambda,\gamma^t,n) &\le h_l(\tilde{\lambda},\gamma^t,n)+\kappa_t && \text{(Lemma A.5.9)}\\ &\le h_l(\tilde{\lambda},\tilde{\gamma}_*,n)+2\kappa_t && \text{(convergence of GD)}\\ &\le h_l(\tilde{\lambda},\gamma_*,n)+2\kappa_t && \text{(minimality of } \tilde{\gamma}_*)\\ &\le h_l(\lambda,\gamma_*,n)+3\kappa_t && \text{(Lemma A.5.9)}\\ &= \min_\gamma h_l(\lambda,\gamma,n)+3\kappa_t.\end{aligned}$$

□

A.5.4 Guarantees for strong concavity and local strong convexity

The following series of lemmas show that our optimization problem is strongly concave in λ and local strongly convex around the minimum γ, as well as explicitly constructing the Lipschitz constants. These serve as the conditions for convergence of the Frank-Wolfe and gradient descent algorithms.

Lemma A.5.11. *$h_l(\lambda,\gamma,n)$ is a concave function of λ.*

Proof. Note that

$$\mathbb{E}\left[\left(\sum_{a\in\mathcal{A}}\sqrt{(\lambda\odot\gamma)^\top(t_a^{(c)}+\eta_l)}\right)^2\right]=\mathbb{E}\left[\sum_{a\in\mathcal{A}}\sum_{a'\in\mathcal{A}}\sqrt{(t_{a'}^{(c)}+\eta_l)^\top(\lambda\odot\gamma)(\lambda\odot\gamma)^\top(t_a^{(c)}+\eta_l)}\right].$$

we know that $\lambda\mapsto(t_{a'}^{(c)}+\eta_l)^\top(\lambda\odot\gamma)$ and $\lambda\mapsto(\lambda\odot\gamma)^\top(t_a^{(c)}+\eta_l)$ are concave, the square root function is concave and non-decreasing, and sum of concave functions is concave. Therefore, $h_l(\lambda,\gamma,n)$ is concave in λ by property of concave functions. □

Lemma A.5.12. *Consider some λ, γ and n. For any $\lambda_1,\lambda_2\in\triangle_\Pi$, with $L=|\mathcal{A}|^2\frac{((1+\eta_l)\gamma_{\max})^{5/2}}{\eta_l^{3/2}\gamma_{\min}^2}$,*

$$f(\lambda_2,\gamma,n)\le f(\lambda_1,\gamma,n)+\nabla_\lambda f(\lambda_1,\gamma,n)^\top(\lambda_2-\lambda_1)+L\|\lambda_2-\lambda_1\|_1^2,$$

where $f(\lambda,\gamma,n)$ could be either $h_l(\lambda,\gamma,n)$ or $-h_l(\lambda,\gamma,n)$.

Proof. The proof for the negative case is exactly the same as the positive case, so we focus on $f(\lambda,\gamma,n)=h_l(\lambda,\gamma,n)$. We take the gradient of h_l with respect to λ and get

$$\begin{aligned}[\nabla_\lambda h_l(\lambda,\gamma,n)]_\pi=&-\widehat{\Delta}_{l-1}^{\gamma^{l-1}}(\pi,\widehat{\pi}_{l-1})+\frac{\log(1/\delta_l)}{\gamma_\pi n}\\&+\mathbb{E}_{c\sim\nu_\mathcal{D}}\left[\left(\sum_{a\in\mathcal{A}}\sqrt{(\lambda\odot\gamma)^\top(t_a^{(c)}+\eta_l)}\right)\left(\sum_{a'\in\mathcal{A}}\frac{\gamma_\pi(t_{a'}^{(c)}+\eta_l)_\pi}{\sqrt{(\lambda\odot\gamma)^\top(t_{a'}^{(c)}+\eta_l)}}\right)\right].\end{aligned}$$

By Lemma A.6.2, for any $\lambda\in\triangle_\Pi$, we have $\langle\lambda,\nabla_\lambda h_l(\lambda,\gamma,n)\rangle=h_l(\lambda,\gamma,n)$. If we use the shortcut $f(\lambda):=h_l(\lambda,\gamma,n)$, we have

$$f(\lambda_2)-f(\lambda_1)-\nabla_\lambda f(\lambda_1)^\top(\lambda_2-\lambda_1)=f(\lambda_2)-\nabla_\lambda f(\lambda_1)^\top\lambda_2=(\nabla f(\lambda_2)-\nabla f(\lambda_1))^\top\lambda_2.$$

Note that

$$
\begin{aligned}
&(\nabla_\lambda f(\lambda_2) - \nabla_\lambda f(\lambda_1))^\top \lambda_2 \\
&= \sum_{\pi \in \Pi} [\lambda_2]_\pi \mathbb{E}_{c\sim\nu_{\mathcal{D}}} \left[\left(\sum_{a\in\mathcal{A}} \sqrt{(\lambda_2 \odot \gamma)^\top (t_a^{(c)} + \eta_l)} \right) \left(\sum_{a'\in\mathcal{A}} \frac{\gamma_\pi \cdot (t_{a'}^{(c)} + \eta_l)_\pi}{\sqrt{(\lambda_2 \odot \gamma)^\top (t_{a'}^{(c)} + \eta_l)}} \right) \right. \\
&\qquad \left. - \left(\sum_{a\in\mathcal{A}} \sqrt{(\lambda_1 \odot \gamma)^\top (t_a^{(c)} + \eta_l)} \right) \left(\sum_{a'\in\mathcal{A}} \frac{\gamma_\pi \cdot (t_{a'}^{(c)} + \eta_l)_\pi}{\sqrt{(\lambda_1 \odot \gamma)^\top (t_{a'}^{(c)} + \eta_l)}} \right) \right] \\
&= \mathbb{E}_{c\sim\nu_{\mathcal{D}}} \left[\sum_{a'\in\mathcal{A}} (\lambda_2 \odot \gamma)^\top (t_{a'}^{(c)} + \eta_l) \right. \\
&\left. \cdot \sum_{a\in\mathcal{A}} \frac{\sqrt{(\lambda_1 \odot \gamma)^\top (t_{a'}^{(c)} + \eta_l)} \sqrt{(\lambda_2 \odot \gamma)^\top (t_a^{(c)} + \eta_l)} - \sqrt{(\lambda_2 \odot \gamma)^\top (t_{a'}^{(c)} + \eta_l)} \sqrt{(\lambda_1 \odot \gamma)^\top (t_a^{(c)} + \eta_l)}}{\sqrt{(\lambda_2 \odot \gamma)^\top (t_{a'}^{(c)} + \eta_l)} \sqrt{(\lambda_1 \odot \gamma)^\top (t_{a'}^{(c)} + \eta_l)}} \right] \\
&\leq \mathbb{E}_{c\sim\nu_{\mathcal{D}}} \left[\sum_{a'\in\mathcal{A}} (\lambda_2 \odot \gamma)^\top (t_{a'}^{(c)} + \eta_l) \right. \\
&\left. \cdot \sum_{a\in\mathcal{A}} \frac{\left| \sqrt{(\lambda_1 \odot \gamma)^\top (t_{a'}^{(c)} + \eta_l)} \sqrt{(\lambda_2 \odot \gamma)^\top (t_a^{(c)} + \eta_l)} - \sqrt{(\lambda_2 \odot \gamma)^\top (t_{a'}^{(c)} + \eta_l)} \sqrt{(\lambda_1 \odot \gamma)^\top (t_a^{(c)} + \eta_l)} \right|}{\sqrt{(\lambda_2 \odot \gamma)^\top (t_{a'}^{(c)} + \eta_l)} \sqrt{(\lambda_1 \odot \gamma)^\top (t_{a'}^{(c)} + \eta_l)}} \right] \\
&\leq \sum_{a'\in\mathcal{A}} \frac{(1+\eta_l)\gamma_{\max}}{\eta_l \gamma_{\min}} \cdot \mathbb{E}_{c\sim\nu_{\mathcal{D}}} \left[\sum_{a\in\mathcal{A}} \left| \sqrt{(\lambda_2 \odot \gamma)^\top (t_a^{(c)} + \eta_l)} \sqrt{(\lambda_1 \odot \gamma)^\top (t_{a'}^{(c)} + \eta_l)} \right. \right. \\
&\qquad \left. \left. - \sqrt{(\lambda_1 \odot \gamma)^\top (t_a^{(c)} + \eta_l)} \sqrt{(\lambda_2 \odot \gamma)^\top (t_{a'}^{(c)} + \eta_l)} \right| \right] \qquad \text{(A.11)}
\end{aligned}
$$

Note that by triangular inequality

$$
\begin{aligned}
&\left| \sqrt{(\lambda_2 \odot \gamma)^\top (t_a^{(c)} + \eta_l)} \sqrt{(\lambda_1 \odot \gamma)^\top (t_{a'}^{(c)} + \eta_l)} - \sqrt{(\lambda_1 \odot \gamma)^\top (t_a^{(c)} + \eta_l)} \sqrt{(\lambda_2 \odot \gamma)^\top (t_{a'}^{(c)} + \eta_l)} \right| \\
&\leq \left| \sqrt{(\lambda_2 \odot \gamma)^\top (t_a^{(c)} + \eta_l)} - \sqrt{(\lambda_1 \odot \gamma)^\top (t_a^{(c)} + \eta_l)} \right| \sqrt{(\lambda_1 \odot \gamma)^\top (t_{a'}^{(c)} + \eta_l)} \\
&\qquad + \sqrt{(\lambda_1 \odot \gamma)^\top (t_a^{(c)} + \eta_l)} \left| \sqrt{(\lambda_1 \odot \gamma)^\top (t_{a'}^{(c)} + \eta_l)} - \sqrt{(\lambda_2 \odot \gamma)^\top (t_{a'}^{(c)} + \eta_l)} \right|.
\end{aligned}
$$

Also note that

$$\begin{aligned}
&\left|\sqrt{(\lambda_2 \odot \gamma)^\top (t_a^{(c)} + \eta_l)} - \sqrt{(\lambda_1 \odot \gamma)^\top (t_a^{(c)} + \eta_l)}\right| \\
&= \frac{\left|\sum_{\pi \in \Pi} ((\lambda_2)_\pi - (\lambda_1)_\pi)\gamma_\pi (t_a^{(c)} + \eta_l)_\pi\right|}{\sqrt{(\lambda_2 \odot \gamma)^\top (t_a^{(c)} + \eta_l)} + \sqrt{(\lambda_1 \odot \gamma)^\top (t_a^{(c)} + \eta_l)}} \\
&\leq \frac{(1+\eta_l)\gamma_{\max}}{2\sqrt{\eta_l}\gamma_{\min}} \left\|\lambda_2 - \lambda_1\right\|_1,
\end{aligned}$$

so (A.11) is bounded by

$$\begin{aligned}
&\sum_{a' \in \mathcal{A}} \frac{(1+\eta_l)\gamma_{\max}}{\eta_l \gamma_{\min}} \cdot \left(\sum_{a \in \mathcal{A}} 2 \cdot \frac{(1+\eta_l)\gamma_{\max}}{2\sqrt{\eta_l}\gamma_{\min}} \left\|\lambda_2 - \lambda_1\right\|_1 \sqrt{(1+\eta_l)\gamma_{\max}} \right) \\
&= |\mathcal{A}|^2 \frac{((1+\eta_l)\gamma_{\max})^{5/2}}{\eta_l^{3/2}\gamma_{\min}^2} \left\|\lambda_2 - \lambda_1\right\|_1.
\end{aligned}$$

□

Lemma A.5.13. *Consider some λ and n. For any $\gamma_1, \gamma_2 \in \triangle_\Pi$, with $L_\gamma = |\mathcal{A}|^2 \frac{((1+\eta_l)\gamma_{\max})^{3/2}}{\eta_l^{3/2}\gamma_{\min}^2} + \frac{2\log(1/\delta_l)}{n\gamma_{\min}^3}$,*

$$h_l(\lambda, \gamma_2, n) \leq h_l(\lambda, \gamma_1, n) + \nabla_\gamma h_l(\lambda, \gamma_1, n)^\top (\gamma_2 - \gamma_1) + L_\gamma \|\gamma_2 - \gamma_1\|_1^2.$$

Proof.

$$[\nabla_\gamma h_l(\lambda, \gamma)]_\pi = \mathbb{E}_c \left[\left(\sum_{a \in \mathcal{A}} \sqrt{(\lambda \odot \gamma)^\top (t_a^{(c)} + \eta_l)} \right) \cdot \left(\sum_{a' \in \mathcal{A}} \frac{\lambda_\pi ([t_{a'}^{(c)}]_\pi + \eta_l)}{\sqrt{(\lambda \odot \gamma)^\top (t_{a'}^{(c)} + \eta_l)}} \right) \right] - \frac{\lambda_\pi \log(1/\delta_l)}{\gamma_\pi^2 n}.$$

Then we have similar to the proof of Lemma A.5.12, for any γ we have $h_l(\lambda, \gamma, n) - \nabla_\gamma h_l(\lambda, \gamma, n)^\top \gamma = 2\sum_\pi \frac{\lambda_\pi \log(1/\delta_l)}{\gamma_\pi^2 n}$, so

$$\begin{aligned}
&h_l(\lambda, \gamma_2, n) - h_l(\lambda, \gamma_1, n) - \nabla_\gamma h_l(\lambda, \gamma_1, n)^\top (\gamma_2 - \gamma_1) \\
&= 2\sum_\pi \frac{\lambda_\pi \log(1/\delta_l)}{[\gamma_2]_\pi^2 n} - 2\sum_\pi \frac{\lambda_\pi \log(1/\delta_l)}{[\gamma_1]_\pi^2 n} + (\nabla_\gamma h_l(\lambda, \gamma_2, n) - \nabla_\gamma h_l(\lambda, \gamma_1, n))^\top \gamma_2.
\end{aligned}$$

First, we can follow similar techniques in the proof of Lemma A.5.12 to bound the second part and get

$$
\begin{aligned}
&(\nabla_\gamma h_l(\lambda,\gamma_2,n)-\nabla_\gamma h_l(\lambda,\gamma_1,n))^\top\gamma_2\\
&\le \sum_{a'\in\mathcal{A}}(\lambda\odot\gamma_2)^\top(t_{a'}^{(c)}+\eta_l)\\
&\quad\cdot\mathbb{E}_{c\sim\nu_{\mathcal{D}}}\left\{\sum_{a\in\mathcal{A}}\left[\frac{1}{\sqrt{(\lambda\odot\gamma_2)^\top(t_{a'}^{(c)}+\eta_l)}\sqrt{(\lambda\odot\gamma_1)^\top(t_{a'}^{(c)}+\eta_l)}}\right.\right.\\
&\quad\left.\left.\cdot\left|\sqrt{(\lambda\odot\gamma_1)^\top(t_{a'}^{(c)}+\eta_l)}\sqrt{(\lambda\odot\gamma_2)^\top(t_a^{(c)}+\eta_l)}-\sqrt{(\lambda\odot\gamma_2)^\top(t_{a'}^{(c)}+\eta_l)}\sqrt{(\lambda\odot\gamma_1)^\top(t_a^{(c)}+\eta_l)}\right|\right]\right\}\\
&\le \sum_{a'\in\mathcal{A}}\frac{(1+\eta_l)\gamma_{\max}}{\eta_l\gamma_{\min}}\cdot\mathbb{E}_{c\sim\nu_{\mathcal{D}}}\left[\sum_{a\in\mathcal{A}}\left|\sqrt{(\lambda\odot\gamma_2)^\top(t_a^{(c)}+\eta_l)}\sqrt{(\lambda\odot\gamma_1)^\top(t_{a'}^{(c)}+\eta_l)}\right.\right.\\
&\qquad\left.\left.-\sqrt{(\lambda\odot\gamma_1)^\top(t_a^{(c)}+\eta_l)}\sqrt{(\lambda\odot\gamma_2)^\top(t_{a'}^{(c)}+\eta_l)}\right|\right].
\end{aligned}
$$

Also, note that

$$
\begin{aligned}
&\left|\sqrt{(\lambda\odot\gamma_2)^\top(t_a^{(c)}+\eta_l)}-\sqrt{(\lambda\odot\gamma_1)^\top(t_a^{(c)}+\eta_l)}\right|\\
&=\frac{\left|\sum_{\pi\in\Pi}(\lambda_\pi([\gamma_2]_\pi-[\gamma_1]_\pi)(t_a^{(c)})_\pi\right|}{\sqrt{(\lambda\odot\gamma_2)^\top(t_a^{(c)}+\eta_l)}+\sqrt{(\lambda\odot\gamma_1)^\top(t_a^{(c)}+\eta_l)}}\\
&\le\frac{1}{2\sqrt{\eta_l}\gamma_{\min}}\left\|\gamma_2-\gamma_1\right\|_1^2,
\end{aligned}
$$

Therefore, similarly we can bound

$$
\begin{aligned}
&\left|\sqrt{(\lambda\odot\gamma_2)^\top(t_a^{(c)}+\eta_l)}\sqrt{(\lambda\odot\gamma_1)^\top(t_{a'}^{(c)}+\eta_l)}-\sqrt{(\lambda\odot\gamma_1)^\top(t_a^{(c)}+\eta_l)}\sqrt{(\lambda\odot\gamma_2)^\top(t_{a'}^{(c)}+\eta_l)}\right|\\
&\le\frac{\sqrt{(1+\eta_l)\gamma_{\max}}}{2\sqrt{\eta_l}\gamma_{\min}}\|\gamma_2-\gamma_1\|_1^2.
\end{aligned}
$$

For the second term,

$$
\begin{aligned}
&2\sum_\pi\frac{\lambda_\pi\log(1/\delta_l)}{[\gamma_2]_\pi^2 n}-2\sum_\pi\frac{\lambda_\pi\log(1/\delta_l)}{[\gamma_1]_\pi^2 n}\\
&=\frac{2\log(1/\delta_l)}{n}\sum_\pi\lambda_\pi\frac{[\gamma_1]_\pi^2-[\gamma_2]_\pi^2}{[\gamma_1]_\pi^2[\gamma_2]_\pi^2}\\
&\le\frac{2\log(1/\delta_l)}{n\gamma_{\min}^3}\|\gamma_2-\gamma_1\|_1^2.
\end{aligned}
$$

Therefore, we have the result stated above. □

Lemma A.5.14. *Consider some fixed $\lambda \in \triangle_\Pi$ and n. Assume γ_* is a stationary point of $h_l(\lambda, \gamma, n)$, then $h_l(\lambda, \gamma, n)$ is locally strongly convex at γ_*, i.e. for $L_{\text{hess}} = \frac{\lambda_{\min} \log(1/\delta_l)}{\gamma_{\max}^3 n}$, there exists $\epsilon > 0$ such that for all $\gamma \in B_\epsilon(\gamma_*)$, $h_l(\lambda, \gamma, n) \geq h_l(\lambda, \gamma_*, n) + \frac{L_{\text{hess}}}{2} \|\gamma - \gamma_*\|^2$.*

Proof. Since λ and n are fixed, we use the shortcut $g(\gamma) := h_l(\lambda, \gamma, n)$ in the proof. Denote the Hessian of g as M. We aim to show that the Hessian $M \succeq L_{\text{hess}} I$ at γ_*. First, since γ_* is a stationary point, $\nabla_\gamma g(\gamma_*) = 0$, and so for any i,

$$\sum_{c \in \mathcal{D}} \nu_{c_\mathcal{D}} \left(\sum_{a \in \mathcal{A}} \sqrt{(\lambda \odot \gamma)^\top (t_a^{(c)} + \eta_l)} \right) \cdot \left(\sum_{a' \in \mathcal{A}} \frac{\lambda_i ([t_{a'}^{(c)}]_i + \eta_l)}{\sqrt{(\lambda \odot \gamma)^\top (t_{a'}^{(c)} + \eta_l)}} \right) = \frac{\lambda_i \log(1/\delta_l)}{\gamma_i^2 n}. \quad \text{(A.12)}$$

Also, we have for $i \neq j$,

$$\begin{aligned}
\frac{\partial^2 g(\gamma)}{\partial \gamma_i \gamma_j} =& \sum_{c \in \mathcal{D}} \nu_{c_\mathcal{D}} \left(\sum_{a' \in \mathcal{A}} \frac{1}{2} \frac{\lambda_i \left[t_a^{(c)} + \eta_l \right]_i}{\sqrt{(\lambda \odot \gamma)^\top (t_a^{(c)} + \eta_l)}} \right) \cdot \left(\sum_{a \in \mathcal{A}} \frac{\lambda_j \left[t_a^{(c)} + \eta_l \right]_j}{\sqrt{(\lambda \odot \gamma)^\top (t_a^{(c)} + \eta_l)}} \right) \\
&+ \left(\sum_{a \in \mathcal{A}} \sqrt{(\lambda \odot \gamma)^\top (t_a^{(c)} + \eta_l)} \right) \cdot \left(\sum_{a' \in \mathcal{A}} -\frac{1}{2} \cdot \frac{\lambda_i \lambda_j \left[t_{a'}^{(c)} + \eta_l \right]_i \left[t_{a'}^{(c)} + \eta_l \right]_j}{\left((\lambda \odot \gamma)^\top (t_{a'}^{(c)} + \eta_l) \right)^{3/2}} \right).
\end{aligned}$$

And

$$\begin{aligned}
\frac{\partial^2 g(\gamma)}{\partial \gamma_i^2} =& \frac{2\lambda_i \log(1/\delta_l)}{\gamma_i^3 n} + \sum_{c \in \mathcal{D}} \nu_{c_\mathcal{D}} \frac{1}{2} \left(\sum_{a' \in \mathcal{A}} \frac{\lambda_i \left[t_a^{(c)} + \eta_l \right]_i}{\sqrt{(\lambda \odot \gamma)^\top (t_a^{(c)} + \eta_l)}} \right)^2 \\
&- \frac{1}{2} \left(\sum_{a \in \mathcal{A}} \sqrt{(\lambda \odot \gamma)^\top (t_a^{(c)} + \eta_l)} \right) \cdot \left(\sum_{a' \in \mathcal{A}} \frac{\lambda_i^2 \left[t_a^{(c)} + \eta_l \right]_i^2}{\left((\lambda \odot \gamma)^\top (t_a^{(c)} + \eta_l) \right)^{3/2}} \right).
\end{aligned}$$

Then, for any vector $\mu \in \mathbb{R}^{|\Pi|}$ with $\|\mu\| = 1$, we have

$$\begin{aligned}\mu^\top M \mu &= \sum_i \sum_j \mu_i \mu_j M_{ij} = \sum_i \mu_i^2 M_{ii} + \sum_{i \neq j} \mu_i \mu_j M_{ij} \\ &= \sum_i \mu_i^2 \frac{2\lambda_i \log(1/\delta_l)}{\gamma_i^3 n} \qquad (A.13) \\ &+ \sum_c \nu_c \sum_i \sum_j \mu_i \mu_j \frac{1}{2} \left(\sum_{a' \in \mathcal{A}} \frac{\lambda_i \left[t_a^{(c)} + \eta_l \right]_i}{\sqrt{(\lambda \odot \gamma)^\top (t_a^{(c)} + \eta_l)}} \right) \cdot \left(\sum_{a \in \mathcal{A}} \frac{\lambda_j \left[t_a^{(c)} + \eta_l \right]_j}{\sqrt{(\lambda \odot \gamma)^\top (t_a^{(c)} + \eta_l)}} \right) \\ &+ \mu_i \mu_j \left(\sum_{a \in \mathcal{A}} \sqrt{(\lambda \odot \gamma)^\top (t_a^{(c)} + \eta_l)} \right) \cdot \left(\sum_{a' \in \mathcal{A}} -\frac{1}{2} \cdot \frac{\lambda_i \lambda_j \left[t_{a'}^{(c)} + \eta_l \right]_i \left[t_{a'}^{(c)} + \eta_l \right]_j}{\left((\lambda \odot \gamma)^\top (t_{a'}^{(c)} + \eta_l) \right)^{3/2}} \right) . \qquad (A.14)\end{aligned}$$

In what follows, we will first show that

$$\begin{aligned}&\sum_i \mu_i^2 \frac{\lambda_i \log(1/\delta_l)}{\gamma_i^3 n} - \sum_c \nu_c \sum_i \sum_j \mu_i \mu_j \left(\sum_{a \in \mathcal{A}} \sqrt{(\lambda \odot \gamma)^\top (t_a^{(c)} + \eta_l)} \right) \\ &\quad \cdot \left(\sum_{a' \in \mathcal{A}} \cdot \frac{\lambda_i \lambda_j \left[t_{a'}^{(c)} + \eta_l \right]_i \left[t_{a'}^{(c)} + \eta_l \right]_j}{\left((\lambda \odot \gamma)^\top (t_{a'}^{(c)} + \eta_l) \right)^{3/2}} \right) \geq 0. \qquad (A.15)\end{aligned}$$

By equation A.12, the LHS of (A.14) simplifies to

$$\begin{aligned}&\sum_c \nu_c \sum_i \mu_i^2 \frac{1}{\gamma_i} \left(\sum_{a \in \mathcal{A}} \sqrt{(\lambda \odot \gamma)^\top (t_a^{(c)} + \eta_l)} \right) \left(\sum_{a' \in \mathcal{A}} \frac{\lambda_i \left[t_{a'}^{(c)} + \eta_l \right]_i}{\sqrt{(\lambda \odot \gamma)^\top (t_{a'}^{(c)} + \eta_l)}} \right) \\ &- \sum_c \nu_c \sum_i \sum_j \mu_i \mu_j \left(\sum_{a \in \mathcal{A}} \sqrt{(\lambda \odot \gamma)^\top (t_a^{(c)} + \eta_l)} \right) \cdot \left(\sum_{a' \in \mathcal{A}} \cdot \frac{\lambda_i \lambda_j \left[t_{a'}^{(c)} + \eta_l \right]_i \left[t_{a'}^{(c)} + \eta_l \right]_j}{\left((\lambda \odot \gamma)^\top (t_{a'}^{(c)} + \eta_l) \right)^{3/2}} \right) .\end{aligned}$$

Therefore, it is sufficient to show that

$$\sum_i \mu_i^2 \frac{1}{\gamma_i} \left(\sum_{a' \in \mathcal{A}} \frac{\lambda_i \left[t_{a'}^{(c)} + \eta_l \right]_i}{\sqrt{(\lambda \odot \gamma)^\top (t_{a'}^{(c)} + \eta_l)}} \right) - \sum_i \sum_j \mu_i \mu_j \left(\sum_{a' \in \mathcal{A}} \frac{\lambda_i \lambda_j \left[t_{a'}^{(c)} + \eta_l \right]_i \left[t_{a'}^{(c)} + \eta_l \right]_j}{\left((\lambda \odot \gamma)^\top (t_{a'}^{(c)} + \eta_l) \right)^{3/2}} \right) \geq 0.$$

Consider some $a' \in \mathcal{A}$. The LHS of the above simplifies to

$$\begin{aligned}
&\sum_i \mu_i^2 \frac{1}{\gamma_i} \frac{\lambda_i \left[t_{a'}^{(c)} + \eta_l\right]_i}{\sqrt{(\lambda \odot \gamma)^\top (t_{a'}^{(c)} + \eta_l)}} - \sum_i \sum_j \mu_i \mu_j \frac{\lambda_i \lambda_j \left[t_{a'}^{(c)} + \eta_l\right]_i \left[t_{a'}^{(c)} + \eta_l\right]_j}{\left((\lambda \odot \gamma)^\top (t_{a'}^{(c)} + \eta_l)\right)^{3/2}} \\
&= \frac{1}{\left((\lambda \odot \gamma)^\top (t_{a'}^{(c)} + \eta_l)\right)^{3/2}} \left(\sum_i \frac{\mu_i^2}{\gamma_i} \lambda_i \left[t_{a'}^{(c)} + \eta_l\right]_i \left(\sum_j \lambda_j \gamma_j \left[t_{a'}^{(c)} + \eta_l\right]_j \right) \right. \\
&\quad \left. - \sum_i \sum_j \mu_i \mu_j \lambda_i \lambda_j \left[t_{a'}^{(c)} + \eta_l\right]_i \left[t_{a'}^{(c)} + \eta_l\right]_j \right) \\
&= \frac{1}{\left((\lambda \odot \gamma)^\top (t_{a'}^{(c)} + \eta_l)\right)^{3/2}} \left(\sum_i \sum_j \gamma_i^{-1} \left(\mu_i^2 \lambda_i \left[t_{a'}^{(c)} + \eta_l\right]_i \lambda_j \gamma_j \left[t_{a'}^{(c)} + \eta_l\right]_j \right. \right. \\
&\quad \left. \left. - \mu_i \mu_j \lambda_i \lambda_j \gamma_i \left[t_{a'}^{(c)} + \eta_l\right]_i \left[t_{a'}^{(c)} + \eta_l\right]_j \right) \right).
\end{aligned}$$

Each summand is

$$\begin{aligned}
&\gamma_i^{-1} \left(\mu_i^2 \lambda_i \left[t_{a'}^{(c)} + \eta_l\right]_i \lambda_j \gamma_j \left[t_{a'}^{(c)} + \eta_l\right]_j - \mu_i \mu_j \lambda_i \lambda_j \gamma_i \left[t_{a'}^{(c)} + \eta_l\right]_i \left[t_{a'}^{(c)} + \eta_l\right]_j \right) \\
&= \gamma_i^{-1} \mu_i \lambda_i \lambda_j \left[t_{a'}^{(c)} + \eta_l\right]_i \left[t_{a'}^{(c)} + \eta_l\right]_j (\mu_i \gamma_j - \mu_j \gamma_i) \\
&= \gamma_i^{-1} \gamma_j^{-1} \lambda_i \lambda_j \left[t_{a'}^{(c)} + \eta_l\right]_i \left[t_{a'}^{(c)} + \eta_l\right]_j (\mu_i \gamma_j) (\mu_i \gamma_j - \mu_j \gamma_i).
\end{aligned}$$

Exchanging subscripts of i and j, we have

$$\gamma_j^{-1} \gamma_i^{-1} \lambda_j \lambda_i \left[t_{a'}^{(c)} + \eta_l\right]_j \left[t_{a'}^{(c)} + \eta_l\right]_i (\mu_j \gamma_i) (\mu_j \gamma_i - \mu_i \gamma_j).$$

The sum of these two terms is

$$\gamma_i^{-1} \gamma_j^{-1} \lambda_i \lambda_j \left[t_{a'}^{(c)} + \eta_l\right]_i \left[t_{a'}^{(c)} + \eta_l\right]_j (\mu_i \gamma_j - \mu_j \gamma_i)^2 \geq 0.$$

Therefore, we proved equation (A.15). We will show next that

$$\begin{aligned}
&\sum_i \mu_i^2 \frac{\lambda_i \log(1/\delta_l)}{\gamma_i^3 n} + \sum_c \nu_c \sum_i \sum_j \mu_i \mu_j \frac{1}{2} \left(\sum_{a' \in \mathcal{A}} \frac{\lambda_i \left[t_a^{(c)} + \eta_l\right]_i}{\sqrt{(\lambda \odot \gamma)^\top (t_a^{(c)} + \eta_l)}} \right) \\
&\cdot \left(\sum_{a \in \mathcal{A}} \frac{\lambda_j \left[t_a^{(c)} + \eta_l\right]_j}{\sqrt{(\lambda \odot \gamma)^\top (t_a^{(c)} + \eta_l)}} \right) \geq 0.
\end{aligned} \tag{A.16}$$

By similar calculation, we can obtain that the above simplifies to

$$\sum_c \nu_c \sum_i \mu_i \gamma_i^{-1} \left(\sum_{a' \in \mathcal{A}} \frac{\lambda_i \left[t_{a'}^{(c)} + \eta_l \right]_i}{\sqrt{(\lambda \odot \gamma)^\top (t_{a'}^{(c)} + \eta_l)}} \right)$$
$$\cdot \left\{ \mu_i \sum_{a \in \mathcal{A}} \frac{\sum_j \lambda_j \gamma_j [t_a^{(c)} + \eta_l]_j}{\sqrt{(\lambda \odot \gamma)^\top (t_a^{(c)} + \eta_l)}} + \mu_j \gamma_i \sum_{a \in \mathcal{A}} \frac{\sum_j \lambda_j [t_a^{(c)} + \eta_l]_j}{\sqrt{(\lambda \odot \gamma)^\top (t_a^{(c)} + \eta_l)}} \right\}.$$

We can show that the sum of the above is positive by similar techniques for showing (A.15). Plugging equation A.15 and A.16 in equation A.14, we have that

$$\mu^\top M \mu \geq \sum_i \mu_i^2 \frac{\lambda_i \log(1/\delta_l)}{\gamma_i^3 n} \geq \frac{\lambda_{\min} \log(1/\delta_l)}{\gamma_{\max}^3 n},$$

so the Hessian is positive-definite. □

Note that the minimum eigenvalue of the Hessian at the stationary point is $\frac{\lambda_{\min} \log(1/\delta_l)}{\gamma_{\max}^3 n} > 0$, we can extend the result in Lemma A.5.14 to α-stationary points, where $\alpha < \frac{\lambda_{\min} \log(1/\delta_l)}{\gamma_{\max}^3 n}$, and still maintain local strong convexity.

Lemma A.5.15. *Consider some fixed $\lambda \in \triangle_\Pi$ and n. Assume γ_α is an α-stationary point of $h_l(\lambda, \gamma, n)$, where $\alpha = \frac{\lambda_{\min} \log(1/\delta_l)}{2\gamma_{\max}^3 n}$, then $h_l(\lambda, \gamma, n)$ is locally strongly convex at γ_α, i.e. for $L_{\text{hess}} = \frac{\lambda_{\min} \log(1/\delta_l)}{2\gamma_{\max}^3 n}$, there exists $\epsilon > 0$ such that for all $\gamma \in B_\epsilon(\gamma_\alpha)$, $h_l(\lambda, \gamma, n) \geq h_l(\lambda, \gamma_\alpha, n) + \frac{L_{\text{hess}}}{2} \|\gamma - \gamma_\alpha\|^2$.*

Proof. The proof follows almost identically from that of Lemma A.5.14. Note that the α-stationary point ensures that $\|\nabla_\gamma h_l(\lambda, \gamma)\|_1 \leq \alpha$, so equation A.12 is rewritten as

$$\sum_i \left| \sum_{c \in \mathcal{D}} \nu_{c_\mathcal{D}} \left(\sum_{a \in \mathcal{A}} \sqrt{(\lambda \odot \gamma)^\top (t_a^{(c)} + \eta_l)} \right) \cdot \left(\sum_{a' \in \mathcal{A}} \frac{\lambda_i ([t_{a'}^{(c)}]_i + \eta_l)}{\sqrt{(\lambda \odot \gamma)^\top (t_{a'}^{(c)} + \eta_l)}} \right) - \frac{\lambda_i \log(1/\delta_l)}{\gamma_i^2 n} \right| \leq \alpha. \tag{A.17}$$

Therefore, for any μ we can still use the same trick and get

$$\mu^\top M \mu \geq \sum_i \mu_i^2 \frac{\lambda_i \log(1/\delta_l)}{\gamma_i^3 n} - \alpha \geq \frac{\lambda_{\min} \log(1/\delta_l)}{2\gamma_{\max}^3 n},$$

so our result follows. □

A.5.5 Proof of strong duality

In this section, we would like to show that strong duality holds. We first show that the primal problem is convex for w.

Lemma A.5.16. *The primal problem* (A.5) *is convex for* w.

Proof. Note that the primal problem could be written as

$$\min_{w\in\Omega} c \qquad \text{s.t. } \forall \pi \in \Pi, -\Delta(\pi) + \sqrt{\frac{\|\phi_\pi - \phi_{\pi_*}\|^2_{A(w)^{-1}}}{n}} \le c.$$

Therefore, we consider the function $f(w) := -\Delta(\pi) + \sqrt{\frac{\|\phi_\pi - \phi_{\pi_*}\|^2_{A(w)^{-1}}}{n}}$ for some $\pi \in \Pi$. Note that to show that $f(w) = -\Delta(\pi) + \sqrt{\frac{\|\phi_\pi - \phi_{\pi_*}\|^2_{A(w)^{-1}}}{n}}$ is convex for w, it is equivalent to show that $g(w) := \sqrt{\|\phi_\pi - \phi_{\pi_*}\|^2_{A(w)^{-1}}}$ is convex for w. Note that

$$\begin{aligned} g(w) &= \sqrt{\sum_{a,c} \nu_c^2 w_{a,c}^{-1}(\mathbf{1}\{\pi(c) = a, \pi_*(c) \neq a\} + \mathbf{1}\{\pi(c) \neq a, \pi_*(c) = a\})} \\ &= \sqrt{\sum_{a,c,t_a^{(c)}=1} \nu_c^2 w_{a,c}^{-1}}. \end{aligned}$$

So restricting to a, c such that $t_a^{(c)} = 1$

$$\frac{\partial g(w)}{\partial w_{a,c}} = \frac{1}{2\sqrt{\sum_{a,c,t_a^{(c)}=1} \nu_c^2 w_{a,c}^{-1}}} \cdot (-\nu_c^2 w_{a,c}^{-2}),$$

and

$$\frac{\partial^2 g(w)}{\partial w_{a,c}^2} = -\frac{1}{4\left(\sum_{a,c,t_a^{(c)}=1} \nu_c^2 w_{a,c}^{-1}\right)^{3/2}} \cdot (-\nu_c^2 w_{a,c}^{-2} \cdot -\nu_c^2 w_{a,c}^{-2}) + \frac{1}{\sqrt{\sum_{a,c,t_a^{(c)}=1} \nu_c^2 w_{a,c}^{-1}}} \cdot \nu_c^2 w_{a,c}^{-3}$$

$$\frac{\partial^2 g(w)}{\partial w_{a_1,c_1} \partial w_{a_2,c_2}} = -\frac{1}{4\left(\sum_{a,c,t_a^{(c)}=1} \nu_c^2 w_{a,c}^{-1}\right)^{3/2}} \cdot (-\nu_{c_1}^2 w_{a_1,c_1}^{-2} \cdot -\nu_{c_2}^2 w_{a_2,c_2}^{-2})$$

Denote the Hessian as M. Then, for any vector $\mu \in \mathbb{R}^{|\mathcal{A}|\times|\mathcal{C}|}$ with $\|\mu\|_2 = 1$, we have

$$\mu^\top M\mu = -\frac{1}{4}\sum_{a,c,t_a^{(c)}=1}\sum_{a',c',t_{a'}^{(c')}=1}\mu_{a,c}\mu_{a',c'}\left(\sum_{a,c,t_a^{(c)}=1}\nu_c^2 w_{a,c}^{-1}\right)^{-3/2}\nu_c^2\nu_{c'}^2 w_{a,c}^{-2}w_{a',c'}^{-2}$$
$$+\sum_{a,c,t_a^{(c)}=1}\mu_{a,c}^2\nu_c^2 w_{a,c}^{-3}\left(\sum_{a,c,t_a^{(c)}=1}\nu_c^2 w_{a,c}^{-1}\right)^{-1/2}.$$

To show that this is nonnegative, it is equivalent to show that

$$-\frac{1}{4}\sum_{a,c,t_a^{(c)}=1}\sum_{a',c',t_{a'}^{(c')}=1}\mu_{a,c}\mu_{a',c'}\nu_c^2\nu_{c'}^2 w_{a,c}^{-2}w_{a',c'}^{-2}+\sum_{a,c,t_a^{(c)}=1}\mu_{a,c}^2\nu_c^2 w_{a,c}^{-3}\left(\sum_{a',c',t_{a'}^{(c')}=1}\nu_{c'}^2 w_{a',c'}^{-1}\right)\geq 0,$$

which is equivalent to show that

$$\sum_{a,c,t_a^{(c)}=1}\sum_{a',c',t_{a'}^{(c')}=1}-\mu_{a,c}\mu_{a',c'}\nu_c^2\nu_{c'}^2 w_{a,c}^{-2}w_{a',c'}^{-2}+\mu_{a,c}^2\nu_c^2 w_{a,c}^{-3}\nu_{c'}^2 w_{a',c'}^{-1}\geq 0. \tag{A.18}$$

Note that

$$\begin{aligned}
&-\mu_{a,c}\mu_{a',c'}\nu_c^2\nu_{c'}^2 w_{a,c}^{-2}w_{a',c'}^{-2}+\mu_{a,c}^2\nu_c^2 w_{a,c}^{-3}\nu_{c'}^2 w_{a',c'}^{-1}\\
&=\mu_{a,c}w_{a,c}^{-3}w_{a',c'}^{-2}\nu_c^2\nu_{c'}^2(\mu_{a,c}w_{a',c'}-\mu_{a',c'}w_{a,c})\\
&=w_{a,c}^{-3}w_{a',c'}^{-3}\nu_c^2\nu_{c'}^2(\mu_{a,c}w_{a',c'})(\mu_{a,c}w_{a',c'}-\mu_{a',c'}w_{a,c}).
\end{aligned}$$

Then, exchanging the label of a and a', we also get a term like

$$w_{a',c'}^{-3}w_{a,c}^{-3}\nu_{c'}^2\nu_c^2(\mu_{a',c'}w_{a,c})(\mu_{a',c'}w_{a,c}-\mu_{a,c}w_{a',c'}).$$

The sum of these two terms is

$$\begin{aligned}
&w_{a',c'}^{-3}w_{a,c}^{-3}\nu_{c'}^2\nu_c^2(\mu_{a',c'}w_{a,c})(\mu_{a',c'}w_{a,c}-\mu_{a,c}w_{a',c'})\\
&+w_{a,c}^{-3}w_{a',c'}^{-3}\nu_c^2\nu_{c'}^2(\mu_{a,c}w_{a',c'})(\mu_{a,c}w_{a',c'}-\mu_{a',c'}w_{a,c})\\
&=w_{a',c'}^{-3}w_{a,c}^{-3}\nu_{c'}^2\nu_c^2(\mu_{a',c'}w_{a,c}-\mu_{a,c}w_{a',c'})(\mu_{a',c'}w_{a,c}-\mu_{a,c}w_{a',c'})\\
&=w_{a',c'}^{-3}w_{a,c}^{-3}\nu_{c'}^2\nu_c^2(\mu_{a',c'}w_{a,c}-\mu_{a,c}w_{a',c'})^2\geq 0.
\end{aligned}$$

Therefore, equation A.18 becomes

$$\sum_{a,c,t_a^{(c)}=1} \sum_{\substack{a',c' \\ t_{a'}^{(c')}=1 \\ (a',c')>(a,c)}} (w_{a',c'}^{-3} w_{a,c}^{-3} \nu_{c'}^2 \nu_c^2 (\mu_{a',c'} w_{a,c})(\mu_{a',c'} w_{a,c} - \mu_{a,c} w_{a',c'})$$

$$+ w_{a,c}^{-3} w_{a',c'}^{-3} \nu_c^2 \nu_{c'}^2 (\mu_{a,c} w_{a',c'})(\mu_{a,c} w_{a',c'} - \mu_{a',c'} w_{a,c}))$$

$$= \sum_{a,c,t_a^{(c)}=1} \sum_{\substack{a',c' \\ t_{a'}^{(c')}=1 \\ (a',c')>(a,c)}} w_{a',c'}^{-3} w_{a,c}^{-3} \nu_{c'}^2 \nu_c^2 (\mu_{a',c'} w_{a,c} - \mu_{a,c} w_{a',c'})^2 \geq 0.$$

Since the above holds for any vector μ, the Hessian is positive-semidefinite, and so the function $g(w)$ is convex for w. □

Lemma A.5.17. *In the optimization problem A.5, the strong duality holds, i.e.*

$$\min_{w\in\Omega} \max_{\pi\in\Pi} \left(-\Delta(\pi) + \sqrt{\frac{\|\phi_\pi - \phi_{\pi_*}\|^2_{A(w)^{-1}}}{n}} \right) = \max_{\lambda\in\triangle_\Pi} \min_{w\in\Omega} \sum_{\pi\in\Pi} \lambda_\pi \left(-\Delta(\pi) + \sqrt{\frac{\|\phi_\pi - \phi_{\pi_*}\|^2_{A(w)^{-1}}}{n}} \right).$$

Proof. By Lemma A.5.16, the primal problem is convex for w, so it is left to check the KKT conditions. Note that the lagrangian is

$$\mathcal{L}(w, \lambda, c) = c + \sum_{\pi\in\Pi} \lambda_\pi \cdot \left(-\Delta(\pi) + \sqrt{\frac{\|\phi_\pi - \phi_{\pi_*}\|^2_{A(w)^{-1}}}{n}} - c \right).$$

Let $h_\pi(w) = -\Delta(\pi) + \sqrt{\frac{\|\phi_\pi - \phi_{\pi_*}\|^2_{A(w)^{-1}}}{n}} - c$. At an optimal solution w^* and λ^*, we would like to show that

$$\sum_{\pi\in\Pi} \lambda_\pi^* h_\pi(w^*) = 0.$$

We prove this by contradiction. If there is some π such that $\lambda_\pi > 0$ and $h_\pi(w^*) < 0$. Then we could find another $\lambda' \in \triangle_\Pi$ that places zero mass on this π and thus get a larger objective, so we get a contradiction. The other conditions follow from the optimality of w^* and λ^*. □

A.6 Useful lemmas

In this section, we state several algebraic facts of our function, which serves as the key to derive convergence as well as complexity.

Lemma A.6.1. *For any l,*

$$\min_{w\in\Omega}\max_{\pi\in\Pi}\frac{\left\|\phi_{\widehat{\pi}_{l-1}}-\phi_\pi\right\|^2_{A(w)^{-1}}}{\Delta(\pi)^2}=\min_{p_c\in\triangle_{\mathcal{A}},\forall c\in\mathcal{C}}\max_{\pi\in\Pi}\frac{\mathbb{E}_{c\sim\nu}\left[\left(\frac{1}{p_{c,\widehat{\pi}_{l-1}(c)}}+\frac{1}{p_{c,\pi(c)}}\right)\mathbf{1}\{\widehat{\pi}_{l-1}(c)\neq\pi(c)\}\right]}{\Delta(\pi)^2}.$$

Proof. Let $w_{a,c}=\nu_c p_{c,a}$ for some $p_c\in\triangle_{\mathcal{A}}$. Then, for any $\pi\in\Pi$,

$$\begin{aligned}
&\frac{1}{\Delta(\pi)^2}\left\|\phi_{\widehat{\pi}_{l-1}}-\phi_\pi\right\|^2_{A(w)^{-1}}\\
&=\frac{1}{\Delta(\pi)^2}\sum_{a,c}\frac{\nu_c^2}{w_{a,c}}\left(\mathbf{1}\{\widehat{\pi}_{l-1}(c)=a,\pi(c)\neq a\}+\mathbf{1}\{\widehat{\pi}_{l-1}(c)\neq a,\pi(c)=a\}\right)\\
&=\frac{1}{\Delta(\pi)^2}\sum_{a,c}\frac{\nu_c}{p_{c,a}}\left(\mathbf{1}\{\widehat{\pi}_{l-1}(c)=a,\pi(c)\neq a\}+\mathbf{1}\{\widehat{\pi}_{l-1}(c)\neq a,\pi(c)=a\}\right)\\
&=\frac{1}{\Delta(\pi)^2}\sum_{c}\nu_c\left(\frac{1}{p_{c,\widehat{\pi}_{l-1}(c)}}+\frac{1}{p_{c,\pi(c)}}\right)\mathbf{1}\{\widehat{\pi}_{l-1}(c)\neq\pi(c)\}\\
&=\frac{1}{\Delta(\pi)^2}\mathbb{E}_{c\sim\nu}\left[\left(\frac{1}{p_{c,\widehat{\pi}_{l-1}(c)}}+\frac{1}{p_{c,\pi(c)}}\right)\mathbf{1}\{\widehat{\pi}_{l-1}(c)\neq\pi(c)\}\right].
\end{aligned}$$

Therefore,

$$\min_{w\in\Omega}\max_{\pi\in\Pi}\frac{\left\|\phi_{\widehat{\pi}_{l-1}}-\phi_\pi\right\|^2_{A(w)^{-1}}}{\Delta(\pi)^2}=\min_{p_c\in\triangle_{\mathcal{A}},\forall c\in\mathcal{C}}\max_{\pi\in\Pi}\frac{\mathbb{E}_{c\sim\nu}\left[\left(\frac{1}{p_{c,\widehat{\pi}_{l-1}(c)}}+\frac{1}{p_{c,\pi(c)}}\right)\mathbf{1}\{\widehat{\pi}_{l-1}(c)\neq\pi(c)\}\right]}{\Delta(\pi)^2}.$$

□

Lemma A.6.2. *For any l, any $\lambda\in\triangle_\Pi$, $\gamma>0$, and any n, we have $h_l(\lambda,\gamma,n)=\langle\lambda,\nabla_\lambda h_l(\lambda,\gamma,n)\rangle$.*

Proof. We first compute

$$\begin{aligned}
[\nabla_\lambda h_l(\lambda,\gamma,n)]_\pi=&-\widehat{\Delta}_{l-1}^{\gamma^{l-1}}(\pi,\widehat{\pi}_{l-1})+\frac{\log(1/\delta_l)}{\gamma_\pi n}\\
&+\mathbb{E}_{c\sim\nu_{\mathcal{D}}}\left[\left(\sum_{a\in\mathcal{A}}\sqrt{(\lambda\odot\gamma)^\top(t_a^{(c)}+\eta_l)}\right)\left(\sum_{a'\in\mathcal{A}}\frac{\gamma_\pi(t_{a'}^{(c)}+\eta_l)_\pi}{\sqrt{(\lambda\odot\gamma)^\top(t_{a'}^{(c)}+\eta_l)}}\right)\right].
\end{aligned}$$

Then, by the fact that

$$\begin{aligned}
&\sum_{\pi\in\Pi}\lambda_\pi\cdot\mathbb{E}_{c\sim\nu_{\mathcal{D}}}\left[\left(\sum_{a\in\mathcal{A}}\sqrt{(\lambda\odot\gamma)^\top(t_a^{(c)}+\eta_l)}\right)\left(\sum_{a'\in\mathcal{A}}\frac{\gamma_\pi(t_{a'}^{(c)}+\eta_l)_\pi}{\sqrt{(\lambda\odot\gamma)^\top(t_{a'}^{(c)}+\eta_l)}}\right)\right]\\
&=\mathbb{E}_{c\sim\nu_{\mathcal{D}}}\left[\left(\sum_{a\in\mathcal{A}}\sqrt{(\lambda\odot\gamma)^\top(t_a^{(c)}+\eta_l)}\right)\left(\sum_{a'\in\mathcal{A}}\frac{(\lambda\odot\gamma)^\top(t_{a'}^{(c)}+\eta_l)}{\sqrt{(\lambda\odot\gamma)^\top(t_{a'}^{(c)}+\eta_l)}}\right)\right]\\
&=\mathbb{E}_{c\sim\nu_{\mathcal{D}}}\left[\left(\sum_{a\in\mathcal{A}}\sqrt{(\lambda\odot\gamma)^\top(t_a^{(c)}+\eta_l)}\right)^2\right],
\end{aligned}$$

we have

$$\begin{aligned}
&\langle\lambda,\nabla_\lambda h_l(\lambda,\gamma,n)\rangle\\
&=\sum_{\pi\in\Pi}\lambda_\pi\left[\nabla_\lambda h_l(\lambda,\gamma,n)\right]_\pi\\
&=\sum_{\pi\in\Pi}\lambda_\pi\cdot\left(-\widehat{\Delta}_{l-1}^{\gamma^{l-1}}(\pi,\widehat{\pi}_{l-1})+\frac{\log(1/\delta_l)}{\gamma_\pi n}\right)\\
&\quad+\sum_{\pi\in\Pi}\lambda_\pi\mathbb{E}_{c\sim\nu_{\mathcal{D}}}\left[\left(\sum_{a\in\mathcal{A}}\sqrt{(\lambda\odot\gamma)^\top(t_a^{(c)}+\eta_l)}\right)\left(\sum_{a'\in\mathcal{A}}\frac{\gamma_\pi(t_{a'}^{(c)}+\eta_l)_\pi}{\sqrt{(\lambda\odot\gamma)^\top(t_{a'}^{(c)}+\eta_l)}}\right)\right]\\
&=\sum_{\pi\in\Pi}\lambda_\pi\cdot\left(-\widehat{\Delta}_{l-1}^{\gamma^{l-1}}(\pi,\widehat{\pi}_{l-1})+\frac{\log(1/\delta_l)}{\gamma_\pi n}\right)+\mathbb{E}_{c\sim\nu_{\mathcal{D}}}\left[\left(\sum_{a\in\mathcal{A}}\sqrt{(\lambda\odot\gamma)^\top(t_a^{(c)}+\eta_l)}\right)^2\right]\\
&=h_l(\lambda,\gamma,n).
\end{aligned}$$

□

Lemma A.6.3. *For any* $\lambda\in\triangle_\Pi$ *and* $\gamma\in\left[0,\min\left\{\sqrt{\frac{\log(1/\delta_l)}{2n_l\mathbb{E}_c[\mathbf{1}\{\pi(c)\neq\pi^*(c)\}]}},\sqrt{\frac{\log(1/\delta_l)}{|\mathcal{A}|^2\eta_l n_l}}\right\}\right]^\Pi$, *with* $\eta_l=|\mathcal{A}|^{-4}\epsilon_l^2$, *we have*

$$0\le\mathbb{E}_{c\sim\nu}\left[\left(\sum_{a\in\mathcal{A}}\sqrt{(\lambda\odot\gamma)^\top(t_a^{(c)}+\eta_l)}\right)^2\right]-\mathbb{E}_{c\sim\nu}\left[\left(\sum_a\sqrt{(\lambda\odot\gamma)^\top t_a^{(c)}}\right)^2\right]\le\epsilon_l.$$

Proof. The first inequality is clear since $\eta_l>0$ and $\lambda_\pi,\gamma_\pi\ge0$ for all $\pi\in\Pi$, so we focus on

the upper bound. Note that

$$\begin{aligned}
&\mathbb{E}_{c\sim\nu}\left[\left(\sum_{a\in\mathcal{A}}\sqrt{(\lambda\odot\gamma)^\top(t_a^{(c)}+\eta_l)}\right)^2\right]-\mathbb{E}_c\left[\left(\sum_a\sqrt{(\lambda\odot\gamma)^\top t_a^{(c)}}\right)^2\right]\\
&=\mathbb{E}_{c\sim\nu}\left[\sum_{a\in\mathcal{A}}(\lambda\odot\gamma)^\top(t_a^{(c)}+\eta_l)+\sum_{a_1\in\mathcal{A}}\sum_{a_2\in\mathcal{A}}\sqrt{(\lambda\odot\gamma)^\top(t_{a_1}^{(c)}+\eta_l)(t_{a_2}^{(c)}+\eta_l)^\top(\lambda\odot\gamma)}\right]\\
&\quad-\mathbb{E}_{c\sim\nu}\left[\sum_{a\in\mathcal{A}}(\lambda\odot\gamma)^\top t_a^{(c)}+\sum_{a_1\in\mathcal{A}}\sum_{a_2\in\mathcal{A}}\sqrt{(\lambda\odot\gamma)^\top t_{a_1}^{(c)}{t_{a_2}^{(c)}}^\top(\lambda\odot\gamma)}\right].
\end{aligned}\tag{A.19}$$

Note that

$$\begin{aligned}
&\mathbb{E}_{c\sim\nu}\left[\sum_{a_1\in\mathcal{A}}\sum_{a_2\in\mathcal{A}}\sqrt{(\lambda\odot\gamma)^\top(t_{a_1}^{(c)}+\eta_l)(t_{a_2}^{(c)}+\eta_l)^\top(\lambda\odot\gamma)}\right]\\
&=\mathbb{E}_{c\sim\nu}\left[\sum_{a_1\in\mathcal{A}}\sum_{a_2\in\mathcal{A}}\sqrt{(\lambda\odot\gamma)^\top t_{a_1}^{(c)}(t_{a_2}^{(c)})^\top(\lambda\odot\gamma)+\eta_l\lambda^\top\gamma(\lambda\odot\gamma)^\top(t_{a_1}^{(c)}+t_{a_2}^{(c)})+\eta_l^2(\lambda^\top\gamma)^2}\right]\\
&\le\mathbb{E}_{c\sim\nu}\left[\sum_{a_1\in\mathcal{A}}\sum_{a_2\in\mathcal{A}}\sqrt{(\lambda\odot\gamma)^\top t_{a_1}^{(c)}(t_{a_2}^{(c)})^\top(\lambda\odot\gamma)}\right]\\
&\qquad+2|\mathcal{A}|\mathbb{E}_{c\sim\nu}\left[\sum_{a\in\mathcal{A}}\sqrt{\eta_l\lambda^\top\gamma(\lambda\odot\gamma)^\top t_a^{(c)}}\right]+|\mathcal{A}|^2\eta_l\lambda^\top\gamma.
\end{aligned}$$

Then (A.19) is upper bounded by

$$\begin{aligned}
&\mathbb{E}_{c\sim\nu}\left[\sum_{a\in\mathcal{A}}\eta_l\lambda^\top\gamma\right] + 2|\mathcal{A}|\mathbb{E}_{c\sim\nu}\left[\sum_{a\in\mathcal{A}}\sqrt{\eta_l\lambda^\top\gamma(\lambda\odot\gamma)^\top t_a^{(c)}}\right] + |\mathcal{A}|^2\eta_l\lambda^\top\gamma \\
&= |\mathcal{A}|\eta_l\lambda^\top\gamma + |\mathcal{A}|^2\eta_l\lambda^\top\gamma + 2|\mathcal{A}|\sqrt{\eta_l\lambda^\top\gamma}\mathbb{E}_{c\sim\nu}\left[\sum_{a\in\mathcal{A}}\sqrt{\sum_{\pi\in\Pi}\lambda_\pi\gamma_\pi[t_a^{(c)}]_\pi}\right] \\
&= |\mathcal{A}|\eta_l\lambda^\top\gamma + |\mathcal{A}|^2\eta_l\lambda^\top\gamma + 2|\mathcal{A}|^2\sqrt{\eta_l\lambda^\top\gamma}\mathbb{E}_{c\sim\nu}\left[\sum_{a\in\mathcal{A}}\frac{1}{|\mathcal{A}|}\sqrt{\sum_{\pi\in\Pi}\lambda_\pi\gamma_\pi[t_a^{(c)}]_\pi}\right] \\
&= |\mathcal{A}|\eta_l\lambda^\top\gamma + |\mathcal{A}|^2\eta_l\lambda^\top\gamma + 2|\mathcal{A}|^2\sqrt{\eta_l\lambda^\top\gamma}\mathbb{E}_{c\sim\nu}\left[\mathbb{E}_{a\sim\mu}\left[\sqrt{\sum_{\pi\in\Pi}\lambda_\pi\gamma_\pi[t_a^{(c)}]_\pi}\right]\right] \\
&\leq |\mathcal{A}|\eta_l\lambda^\top\gamma + |\mathcal{A}|^2\eta_l\lambda^\top\gamma + 2|\mathcal{A}|^2\sqrt{\eta_l\lambda^\top\gamma}\sqrt{\sum_{\pi\in\Pi}\lambda_\pi\gamma_\pi\frac{1}{|\mathcal{A}|}\mathbb{E}_{c\sim\nu}\left[\sum_{a\in\mathcal{A}}[t_a^{(c)}]_\pi\right]} \\
&= |\mathcal{A}|\eta_l\lambda^\top\gamma + |\mathcal{A}|^2\eta_l\lambda^\top\gamma + 2|\mathcal{A}|^2\sqrt{\eta_l\lambda^\top\gamma}\sqrt{\sum_{\pi\in\Pi}\lambda_\pi\gamma_\pi\frac{1}{|\mathcal{A}|}2\cdot\mathbb{E}_{c\sim\nu}[\mathbf{1}\{\pi(c)\neq\pi^*(c)\}]}. \quad \text{(A.20)}
\end{aligned}$$

Since $\gamma_\pi \leq \sqrt{\frac{\log(1/\delta_l)}{2n_l\mathbb{E}_c[\mathbf{1}\{\pi(c)\neq\pi^*(c)\}]}}$, $\gamma_\pi\mathbb{E}_{c\sim\nu}[\mathbf{1}\{\pi(c) \neq \pi^*(c)\}] \leq \sqrt{\frac{\mathbb{E}_c[\mathbf{1}\{\pi(c)\neq\pi^*(c)\}]\log(1/\delta_l)}{2n_l}} \leq \sqrt{\frac{\log(1/\delta_l)}{2n_l}}$. We know from the lower bound argument that

$$n_l \gtrsim \min_{w\in\Omega}\max_{\pi\in\Pi}\frac{\|\phi_\pi - \phi_{\pi_*}\|^2_{A(w)^{-1}}}{\Delta(\pi)^2+\epsilon_l^2}\log(1/\delta_l) \geq \epsilon_l^{-1}\log(1/\delta_l),$$

so $\sqrt{\frac{\log(1/\delta_l)}{2n}} \lesssim \sqrt{\epsilon_l}$. Therefore, (A.20) is upper bounded by

$$(|\mathcal{A}| + |\mathcal{A}|^2)\eta_l\lambda^\top\gamma + 2|\mathcal{A}|^{3/2}\sqrt{\epsilon_l\eta_l\lambda^\top\gamma}. \quad \text{(A.21)}$$

Since $\eta_l\lambda^\top\gamma \leq \eta_l\gamma_{\max} = \sqrt{\frac{\eta_l\log(1/\delta_l)}{|\mathcal{A}|^2 n_l}} \leq \sqrt{\eta_l}\frac{1}{|\mathcal{A}|}$. Plugging this as well as $\eta_l \leq |\mathcal{A}|^{-4}\epsilon_l^2$ in equation A.21 gives that the bias is upper bounded by ϵ_l.

□

Appendix B

APPENDIX TO CHAPTER 3

B.1 Notations and General Description

In the following, we let the index t, $1 \leq t \leq T_\ell$ denote the timestep in round ℓ for any ℓ. Throughout this section we will make use of the filtration $\mathcal{F}_t = \{(x_s, \theta_s, y_s)\}_{s=1}^{t-1}$ defined in any round. The table below summarizes the notations used in the proof.

Let $N_{t,x}$ denote the number of times arm x gets pulled at time t. We then define several good events needed to guarantee the performance of PEPS at round ℓ.

$$\mathcal{E}_{1,\ell} = \bigcup_{t=1}^{T_\ell} \left\{ \left\| \widehat{\theta}_t - \theta^* \right\|_{V_{t-1}}^2 \leq \beta(t, \ell^2) \right\},$$

$$\mathcal{E}_{2,\ell} = \bigcup_{t=1}^{T_\ell} \left\{ \max_{x \in \mathcal{X}} |x^\top \widehat{\theta}_t| \leq C_{1,\ell} \right\},$$

$$\mathcal{E}_{3,\ell} = \bigcup_{t \geq T_2}^{T_\ell} \bigcup_{x \in \mathcal{X}} \mathcal{G}_{t,x} \text{ where } \mathcal{G}_{t,x} = \{V_t \geq t^{3/4} A(\lambda^G)\}, \forall t \geq T_2, x \in \mathcal{X}$$

$$\mathcal{E}_{4,\ell} = \cup_{t \geq T_0} \mathbf{1}\{\widehat{z}_t = z_*\}$$

Throughout the proof we also define for some random variable $x \in \mathcal{X}$ with $x \sim p$ and some function $f(x)$,

$$\mathbb{F}_{x \sim p}[f(x)] = \sum_{x \in \mathcal{X}} p_x f(x).$$

The rest of the supplement is organized as follows. In Section B.2, we present a proof of the lower bound stated in Theorem 3.2.1. Section B.6 provides more experimental results.

In Section B.3, we prove the main theorem (Theorem 3.3.4) stated in the paper by combining a saddle-point convergence argument with a guarantee on the likelihood ratio. We

$\bar{p}_{T_\ell} = \frac{1}{T_\ell}\sum_{t=1}^{T_\ell} p_t$	Average of p at the end of round ℓ
$\bar{e}_{T_\ell} = \frac{1}{T_\ell}\sum_{t=1}^{T_\ell} e_{x_t}$	Empirical probability of arms pulled at the end of round ℓ
$\pi_\ell \sim N(\widehat{\theta}_{T_\ell+1}, \eta_p^{-1}V_{T_\ell}^{-1})$ restricted on Θ	The distribution θ is sampled from at the end of round ℓ
$\Delta_{\min} = \min_{x\neq x^*}(x^* - x)^\top\theta^*$	minimum gap
$T_2(\ell) = \max_{x\in\mathcal{X}}\left(\frac{6\sqrt{\log(\lvert\mathcal{X}\rvert T_\ell \ell^2)}}{\lambda_x^G}\right)^4$	a time after which each arm gets sufficiently number of pulls
$T_0(\ell) = \max\left\{\left(\frac{d\beta(t,\ell^2)\max_{z\in\mathcal{Z}}\lVert z\rVert_1}{\Delta_{\min}}\right)^{4/3}, T_2(\ell)+1\right\}$	a time after which we have $\widehat{z}_t = z_*$ with high probability
$\ell_0 := \min\{\ell : T_\ell \geq T_0(\ell)^{3/2}\}$	minimum round number such that we have guarantee of convergence with high probability
L	upper bound on $\max_{x\in\mathcal{X}}\lVert x\rVert_2$
B	upper bound on $\lVert\theta_*\rVert_2$
$B_{\mathcal{X}}$	$\max_{x\in\mathcal{X}}\max_{\theta\in\Theta} x^\top\theta$
$\Delta_{\max}$	$\max_{x\in\mathcal{X}}\max_{\theta,\theta'\in\Theta}\lvert x^\top(\theta-\theta')\rvert$
$\beta(t,1/\delta) = B + \sqrt{2\log(1/\delta) + d\log\left(\frac{d+tL^2}{d}\right)}$	anytime confidence bound for $\left\lVert\widehat{\theta}_t - \theta^*\right\rVert_{V_{t-1}}^2$
$C_{1,\ell} = \Delta_{\max} + L^2\beta(T_\ell, \ell^2)$	an upper bound on $\max_{x\in\mathcal{X}}\max_{t\leq T_\ell}\lvert\langle x, \widehat{\theta}_t\rangle\rvert$
$C_{3,\ell} = B_{\mathcal{X}} + \Delta_{\max} + L^2\beta(T_\ell, \ell^2)$	an upper bound on $\max_{x\in\mathcal{X}}\max_{\theta\in\Theta}\max_{t\leq T_\ell}\lvert\langle x, \theta - \widehat{\theta}_t\rangle\rvert$

Table B.1: Table of constants and upper bounds used in the proof

tackle the latter in Section B.3.1, where we provide we relate the empirical probability of finding the best-arm at the end of a round of PEPS to the likelihood ratio. In Section B.3.2, we show the saddle point approximation and provide a guarantee on how well τ^* is approximated

after one round of PEPS. This argument depends on

- Section B.3.3 and B.3.4 which provide regret guarantees on the max and min learners.
- Section B.3.5 provides lemmas bounding terms related to the approximation error of $\widehat{\theta}_{T_\ell}$ to θ^*.
- Section B.3.6 formally shows that after certain rounds each arm gets enough samples.
- Section B.4 shows that good events needed to guarantee performance of PEPS happen with high probability.

Finally, Section B.5 provides some technical lemmas used in the proof.

B.2 Proof of Theorem 3.2.1

Theorem B.2.1. *Fix $\Theta = \mathbb{R}^d$ and any $\theta_* \in \Theta$. For some λ consider a procedure that draws $x_1, \dots, x_T \sim \lambda$, then observes $y_t = \langle x_t, \theta_* \rangle + \epsilon_t$ with $\epsilon_t \sim \mathcal{N}(0,1)$, and then computes $\widehat{z}_T = \arg\max_{z \in \mathcal{Z}} \langle z, \widehat{\theta}_T \rangle$ where $\widehat{\theta}_T = \arg\min_{\theta \in \Theta} \sum_{t=1}^T \|y_t - \langle \theta, x_t \rangle\|_2^2$. Then for any $\lambda \in \triangle_{\mathcal{X}}$ we have*

$$\limsup_{T \to \infty} -\frac{1}{T} \log\left(\mathbb{P}_{\theta_*, x_t \sim \lambda}(\widehat{z}_T \neq z_*)\right) \leq \tau^*.$$

Proof. Assume that $\{z - z_*\}_{z \in \mathcal{Z}}$ span $\mathbb{R}^d$. Otherwise, discard the components of $\mathcal{X}$ and θ_* that are orthogonal to the span of $\{z - z_*\}_{z \in \mathcal{Z}}$ and reparameterize in the subspace spanned by $\{z - z_*\}_{z \in \mathcal{Z}}$. We can then work in this reparameterized space, so without loss of generality we can assume $\{z - z_*\}_{z \in \mathcal{Z}}$ span $\mathbb{R}^d$.

Furthermore, assume that $\mathcal{X}$ spans $\mathbb{R}^d$. If this were not true, then there could be a component of θ_* that is orthogonal to the span of $\mathcal{X}$ which makes z_* not identifiable since we assumed $\{z - z_*\}_{z \in \mathcal{Z}}$ spans $\mathbb{R}^d$. That is, if $\theta_*^\perp$ is the projection of θ_* onto the subspace orthogonal to the span of $\mathcal{X}$, then $\langle z - z_*, \theta_*^\perp \rangle$ could be arbitrarily large but no measurement could detect $\theta_*^\perp$.

Putting the two assumptions together, we conclude that there exists a $\lambda \in \triangle_{\mathcal{X}}$ such that $A(\lambda) \succ 0$ (equivalently, $\lambda_{\min}(A(\lambda)) > 0$) and $\max_{z \in \mathcal{Z}} \|z - z_*\|_{A(\lambda)^{-1}} < \infty$. Fix any λ satisfying such conditions. Define the event $G_\lambda = \{\sum_{t=1}^T x_t x_t \succeq A(\lambda)T(1 - g_{\lambda,T})\}$ for some $g_{\lambda,T} = o(T)$ sequence to be defined next.

By applying matrix Chernoff to the random matrices $\{\frac{1}{T}A(\lambda)^{-1}x_t x_t^\top\}_t$ we have for any $\epsilon \in [0, 1)$ that

$$\mathbb{P}\Big(\frac{1}{T}\sum_{t=1}^T x_t x_t^\top \succeq A(\lambda)(1-\epsilon)\Big) \geq 1 - d\exp(-\epsilon^2/2R)$$

where $R = \max_t \lambda_{\max}(\frac{1}{T}A(\lambda)^{-1}x_t x_t^\top)$. Observe that

$$\begin{aligned}\lambda_{\max}(\frac{1}{T}A(\lambda)^{-1}x_t x_t^\top) &\leq \|\frac{1}{T}A(\lambda)^{-1}x_t x_t^\top\|_2 \\ &\leq L^2/\lambda_{\min}(A(\lambda))T.\end{aligned}$$

So taking $\epsilon = g_{\lambda,T} = \sqrt{\frac{2L^2\lambda_{\min}(A(\lambda))^{-1}\log(dT)}{T}}$ we have that $\mathbb{P}(G_\lambda) \geq 1 - 1/T$ whenever $g_{\lambda,T} < 1$ which holds for sufficiently large T.

Now, for any $\{x_t\}_{t=1}^T$ that span $\mathbb{R}^d$ (will be guaranteed by event G_λ) we have that

$$\begin{aligned}\widehat{\theta}_T &= \arg\min_{\theta \in \Theta} \sum_{t=1}^T \|y_t - \langle \theta, x_t\rangle\|_2^2 \\ &= \Big(\sum_{t=1}^T x_t x_t^\top\Big)^{-1} \sum_{t=1}^T x_t y_t \\ &= \theta_* + \Big(\sum_{t=1}^T x_t x_t^\top\Big)^{-1} \sum_{t=1}^T x_t \epsilon_t \\ &= \theta_* + \Big(\sum_{t=1}^T x_t x_t^\top\Big)^{-1/2} \eta\end{aligned}$$

where the last line holds with inequality in distribution for $\eta \sim \mathcal{N}(0, I_d)$. We conclude that

for any z that $\langle \widehat{\theta}_T - \theta_*, z - z_* \rangle$ is a zero-mean Gaussian random variable with variance

$$\begin{aligned}\sigma_{z,\lambda}^2 :=& \mathbb{E}[\langle \widehat{\theta}_T - \theta_*, z - z_* \rangle^2] \\ =& \mathbb{E}[\langle \Big(\sum_{t=1}^{T} x_t x_t^\top \Big)^{-1/2} \eta, z - z_* \rangle^2] \\ =& (z - z_*)^\top \Big(\sum_{t=1}^{T} x_t x_t^\top \Big)^{-1} (z - z_*).\end{aligned}$$

Thus, on G_λ we have that $\sigma_{z,\lambda}^2 \leq \frac{1}{T(1-g_{\lambda,T})} \|z - z_*\|_{A(\lambda)^{-1}}^2$.

Consequently,

$$\begin{aligned}\mathbb{P}_{\theta_*}(\widehat{z}_T \neq z_*) &= \mathbb{P}_{\theta_*}\Big(\bigcup_{z \in \mathcal{Z} \setminus z_*} \{\widehat{z}_T = z, z \neq z_*\} \Big) \\ &\geq \max_{z \in \mathcal{Z} \setminus z_*} \mathbb{P}_{\theta_*}(\widehat{z}_T = z, z \neq z_*) \\ &= \max_{z \in \mathcal{Z} \setminus z_*} \mathbb{P}_{\theta_*}(\langle \widehat{\theta}_T, z - z_* \rangle \geq 0) \\ &= \max_{z \in \mathcal{Z} \setminus z_*} \mathbb{P}_{\theta_*}(\langle \widehat{\theta}_T - \theta_*, z - z_* \rangle \geq \langle \theta_*, z - z_* \rangle) \\ &\geq \max_{z \in \mathcal{Z} \setminus z_*} \mathbb{E}_{\{x_t\} \sim \lambda} \mathbb{E}_{\theta_*} \Big[\mathbf{1}\{G_\lambda\} \mathbf{1}\Big\{ \langle \widehat{\theta}_T - \theta_*, z - z_* \rangle \geq \langle \theta_*, z - z_* \rangle \Big\} \Big| \{x_t\} \Big] \Big] \\ &= \max_{z \in \mathcal{Z} \setminus z_*} \mathbb{P}_{\{x_t\} \sim \lambda}(G_\lambda) \mathbb{P}_{\eta_1 \sim \mathcal{N}(0,1)} \big(\eta_1 \sigma_{z,\lambda} \geq \langle \theta_*, z - z_* \rangle \big).\end{aligned}$$

Using the fact that

$$\mathbb{P}_{\eta_1 \sim \mathcal{N}(0,1)} \big(\eta_1 \geq s \big) = \int_{x=s}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx > (\frac{1}{s} - \frac{1}{s^3}) \frac{1}{\sqrt{2\pi}} e^{-s^2/2}$$

for positive s, we conclude that

$$
\begin{aligned}
&\mathbb{P}_{\theta_*}(\widehat{z}_T \neq z_*) \\
&\geq \max_{z\in\mathcal{Z}\setminus z_*} \mathbb{P}_{\{x_t\}\sim\lambda}(G_\lambda)\mathbb{P}_{\eta_1\sim\mathcal{N}(0,1)}\big(\eta_1\sigma_{z,\lambda} \geq \langle\theta_*, z-z_*\rangle\big) \\
&\geq \mathbf{1}\{g_{\lambda,T}<1\}(1-\frac{1}{T}) \max_{z\in\mathcal{Z}\setminus z_*} (\frac{\sigma_{z,\lambda}}{\langle\theta_*, z-z_*\rangle} - \frac{\sigma^3_{z,\lambda}}{\langle\theta_*, z-z_*\rangle^3})\frac{1}{\sqrt{2\pi}} e^{-\frac{\langle\theta_*, z-z_*\rangle^2}{\sigma^2_{z,\lambda}}/2} \\
&\geq \max_{z\in\mathcal{Z}\setminus z_*} \mathbf{1}\{g_{\lambda,T}<1, \tfrac{\langle\theta_*, z-z_*\rangle^2}{\sigma^2_{z,\lambda}} \geq 2\}(1-\frac{1}{T})\frac{\sigma_{z,\lambda}}{\langle\theta_*, z-z_*\rangle}\frac{1}{\sqrt{8\pi}} e^{-\frac{\langle\theta_*, z-z_*\rangle^2}{\sigma^2_{z,\lambda}}/2} \\
&\geq \max_{z\in\mathcal{Z}\setminus z_*} \mathbf{1}\{g_{\lambda,T}<1, \tfrac{T(1-g_{\lambda,T})\langle\theta_*, z-z_*\rangle^2}{\|z-z_*\|^2_{A(\lambda)^{-1}}} \geq 2\}(1-\frac{1}{T})\frac{\|z-z_*\|^2_{A(\lambda)^{-1}}}{T(1-g_{\lambda,T})\langle\theta_*, z-z_*\rangle^2}\frac{1}{\sqrt{8\pi}} e^{-\frac{T(1-g_{\lambda,T})\langle\theta_*, z-z_*\rangle^2}{\|z-z_*\|^2_{A(\lambda)^{-1}}}/2}.
\end{aligned}
$$

Thus, because $g_{\lambda,T} = o(T)$ and $\frac{\|z-z_*\|^2_{A(\lambda)^{-1}}}{\langle\theta_*, z-z_*\rangle^2} < \infty$ we have that

$$
\begin{aligned}
\limsup_{T\to\infty} -\frac{1}{T}\log\Big(\mathbb{P}_{\theta_*, x_t\sim\lambda}(\widehat{z}_T \neq z_*)\Big) &\leq \frac{\langle\theta_*, z-z_*\rangle^2}{\|z-z_*\|^2_{A(\lambda)^{-1}}}/2 \\
&= \min_{\theta\in\Theta^c_{z_*}} \|\theta-\theta_*\|^2_{A(\lambda)}/2 \\
&\leq \max_{\lambda\in\triangle_{\mathcal{X}}} \min_{\theta\in\Theta^c_{z_*}} \|\theta-\theta_*\|^2_{A(\lambda)}/2 = \tau^*
\end{aligned}
$$

where the second line uses the fact that $\Theta = \mathbb{R}^d$. □

B.3 Proof of Theorem 3.3.4

Theorem B.3.1. *Under Algorithm 5 and 6 and Assumption 2, we have the sampling distribution satisfies with probability 1,*

$$\lim_{\ell\to\infty} -\frac{1}{T_\ell}\log\pi_\ell(\Theta^c_{z_*}) = \tau^*.$$

Proof. By Theorem B.3.2, we have that for $\ell \geq \ell_0$, $\mathbb{P}(\mathcal{E}^c_\ell) \leq \frac{5}{\ell^2}$. Also, since $T_\ell = 2^\ell$, and $T_0(\ell)$ only scales logarithmically in ℓ, so $\ell_0 < \infty$. Therefore, $\sum_{\ell=1}^\infty \mathbb{P}(\mathcal{E}_\ell) < \infty$. By Borel-Cantelli, we have

$$\mathbb{P}\left(\limsup_{n\to\infty} \mathcal{E}^c_\ell\right) = 0.$$

Note that $\limsup_{\ell\to\infty}\mathcal{E}_\ell = \bigcap_{\ell=1}^{\infty}\bigcup_{k=\ell}^{\infty}\mathcal{E}_k$, this implies that the probability that infinitely many of them occur is zero, which means that $\mathcal{E}_\ell$ eventually holds for sufficiently large ℓ with probability 1. However, under $\mathcal{E}_\ell$ we have

$$\begin{aligned}
\pi_\ell(\Theta_{z_*}^c) &= \frac{\int_{\Theta_{z_*}^c}\pi_\ell(\theta)d\theta}{\int_\Theta \pi_\ell(\theta)d\theta} = \frac{\int_{\Theta_{z_*}^c}\pi_\ell(\theta)/\pi_\ell(\theta^*)d\theta}{\int_\Theta \pi_\ell(\theta)/\pi_\ell(\theta^*)d\theta} \\
&\doteq \frac{\int_{\Theta_{z_*}^c} e^{-\frac{T_\ell}{2}\|\theta-\theta^*\|^2_{A(\bar{e}_{T_\ell})}}d\theta}{\int_\Theta e^{-\frac{T_\ell}{2}\|\theta-\theta^*\|^2_{A(\bar{e}_{T_\ell})}}d\theta} && \text{(by } \mathcal{E}_\ell\text{)} \\
&\doteq e^{-T_\ell \inf_{\theta\in\Theta_{z_*}^c}\frac{1}{2}\|\theta-\theta^*\|^2_{A(\bar{e}_{T_\ell})}}. && \text{(Lemma B.5.2 and } \inf_{\theta\in\Theta}\|\theta-\theta^*\|^2_{A(\lambda)}=0 \text{ for any } \lambda\text{)}
\end{aligned}$$

This implies that there exists some $\epsilon'_\ell \to 0$ such that

$$\left|-\frac{1}{T_\ell}\log\pi_\ell(\Theta_{z_*}^c) - \inf_{\theta\in\Theta_{z_*}^c}\frac{1}{2}\|\theta-\theta^*\|^2_{A(\bar{e}_{T_\ell})}\right| \le \epsilon'_\ell.$$

Under $\mathcal{E}_{6,\ell}$, there exists some sequence $\epsilon_\ell \to 0$ such that

$$\tau^* - \inf_{\theta\in\Theta_{z_*}^c}\frac{1}{2}\|\theta-\theta^*\|^2_{A(\bar{e}_{T_\ell})} \le \epsilon_\ell.$$

Since

$$\tau^* = \max_{\lambda\in\triangle_{\mathcal{X}}}\inf_{\theta\in\Theta_{z_*}^c}\frac{1}{2}\|\theta-\theta^*\|^2_{A(\lambda)} \ge \inf_{\theta\in\Theta_{z_*}^c}\frac{1}{2}\|\theta-\theta^*\|^2_{A(\bar{e}_{T_\ell})},$$

combining the above three displays, we have under $\mathcal{E}_\ell$,

$$\left|-\frac{1}{T_\ell}\log\pi_\ell(\Theta_{z_*}^c) - \tau^*\right| \le \epsilon_\ell + \epsilon'_\ell,$$

where $\epsilon_\ell + \epsilon'_\ell \to 0$ as $\ell\to\infty$. Combining this with the fact that $\mathbb{P}\left(\limsup_{\ell\to\infty}\mathcal{E}_\ell\right) = 0$, we have with probability 1,

$$\lim_{\ell\to\infty} -\frac{1}{T_\ell}\log\pi_\ell(\Theta_{z_*}^c) = \tau^*.$$

□

Theorem B.3.2. *In round ℓ for $\ell \ge \ell_0$, define*

$$\begin{aligned}
\mathcal{E}_{5,\ell} &= \left\{\sup_{\theta\in\Theta}\frac{1}{T_\ell}\left|\log\frac{\pi_{T_\ell}(\theta^*)}{\pi_{T_\ell}(\theta)} - \frac{T_\ell}{2}\|\theta-\theta^*\|^2_{A(\bar{e}_{T_\ell})}\right| \le \kappa_\ell\right\} \\
\mathcal{E}_{6,\ell} &= \left\{\left|\max_{\lambda\in\triangle_{\mathcal{X}}}\inf_{\theta\in\Theta_{z_*}^c}\frac{1}{2}\|\theta-\theta^*\|^2_{A(\lambda)} - \inf_{\theta\in\Theta_{z_*}^c}\frac{1}{2}\|\theta-\theta^*\|^2_{A(\bar{e}_{T_\ell})}\right| \le \epsilon_\ell\right\}
\end{aligned}$$

with $\epsilon_\ell \to 0$ and $\kappa_\ell \to 0$ as $\ell\to\infty$. Define $\mathcal{E}_\ell = \mathcal{E}_{5,\ell}\cap\mathcal{E}_{6,\ell}$. Then $\mathbb{P}(\mathcal{E}_\ell) \ge 1 - 5/\ell^2$.

Proof. We first summarize the guarantees for the probabilities of events below. For $\ell \geq \ell_0$, we have

- from Lemma B.3.4, we have that $\mathbb{P}(\mathcal{E}_{6,\ell}|\mathcal{E}_{1,\ell} \cap \mathcal{E}_{2,\ell} \cap \mathcal{E}_{3,\ell} \cap \mathcal{E}_{4,\ell}) \geq 1 - 1/\ell^2$ with choice of $\epsilon_\ell = O(T_\ell^{-1/4})$;
- from Lemma B.4.1, $\mathbb{P}(\mathcal{E}_{1,\ell}) \geq 1 - 1/\ell^2$;
- by Lemma B.4.2, $\mathcal{E}_{2,\ell}$ is true under $\mathcal{E}_{3,\ell} \cap \mathcal{E}_{1,\ell}$;
- by Lemma B.3.16, $\mathbb{P}(\mathcal{E}_{4,\ell}|\mathcal{E}_{1,\ell}) \geq 1 - 1/\ell^2$;
- by Lemma B.3.3 with $\kappa_\ell = O(T_\ell^{-1/2})$, $\mathbb{P}(\mathcal{E}_{5,\ell}) \geq 1 - 1/\ell^2$;
- by Lemma B.3.14, $\mathbb{P}(\mathcal{E}_{3,\ell}) \geq 1 - 1/\ell^2$.

Note that $\mathcal{E}_\ell \supset \mathcal{E}_{1,\ell} \cap \mathcal{E}_{2,\ell} \cap \mathcal{E}_{3,\ell} \cap \mathcal{E}_{4,\ell} \cap \mathcal{E}_{5,\ell} \cap \mathcal{E}_{6,\ell}$, and so

$$\begin{aligned}\mathcal{E}_\ell^c &\subset \mathcal{E}_{1,\ell}^c \cup \mathcal{E}_{2,\ell}^c \cup \mathcal{E}_{3,\ell}^c \cup \mathcal{E}_{4,\ell}^c \cup \mathcal{E}_{5,\ell}^c \cup \mathcal{E}_{6,\ell}^c \\ &= \mathcal{E}_{1,\ell}^c \cup (\mathcal{E}_{2,\ell}^c \cap \mathcal{E}_{1,\ell} \cap \mathcal{E}_{3,\ell}) \cup \mathcal{E}_{3,\ell}^c \cup (\mathcal{E}_{4,\ell}^c \cap \mathcal{E}_{1,\ell}) \cup \mathcal{E}_{5,\ell}^c \cup (\mathcal{E}_{6,\ell}^c \cap \mathcal{E}_{1,\ell} \cap \mathcal{E}_{2,\ell} \cap \mathcal{E}_{3,\ell} \cap \mathcal{E}_{4,\ell}).\end{aligned}$$

Therefore, for $\ell \geq \ell_0$,

$$\begin{aligned}&\mathbb{P}(\mathcal{E}_\ell^c) \\ &\leq \mathbb{P}(\mathcal{E}_{1,\ell}^c) + \mathbb{P}(\mathcal{E}_{2,\ell}^c \cap \mathcal{E}_{1,\ell} \cap \mathcal{E}_{3,\ell}) + \mathbb{P}(\mathcal{E}_{3,\ell}^c) + \mathbb{P}(\mathcal{E}_{4,\ell}^c \cap \mathcal{E}_{1,\ell}) + \mathbb{P}(\mathcal{E}_{5,\ell}^c) + \mathbb{P}(\mathcal{E}_{6,\ell}^c \cap \mathcal{E}_{1,\ell} \cap \mathcal{E}_{2,\ell} \cap \mathcal{E}_{3,\ell} \cap \mathcal{E}_{4,\ell}) \\ &\leq \mathbb{P}(\mathcal{E}_{1,\ell}^c) + \mathbb{P}(\mathcal{E}_{2,\ell}^c|\mathcal{E}_{1,\ell} \cap \mathcal{E}_{3,\ell})\mathbb{P}(\mathcal{E}_{1,\ell} \cap \mathcal{E}_{3,\ell}) + \mathbb{P}(\mathcal{E}_{3,\ell}^c) + \mathbb{P}(\mathcal{E}_{4,\ell}^c|\mathcal{E}_{1,\ell})\mathbb{P}(\mathcal{E}_{1,\ell}) \\ &\quad + \mathbb{P}(\mathcal{E}_{5,\ell}^c) + \mathbb{P}(\mathcal{E}_{6,\ell}^c|\mathcal{E}_{1,\ell} \cap \mathcal{E}_{2,\ell} \cap \mathcal{E}_{3,\ell} \cap \mathcal{E}_{4,\ell})\mathbb{P}(\mathcal{E}_{1,\ell} \cap \mathcal{E}_{2,\ell} \cap \mathcal{E}_{3,\ell} \cap \mathcal{E}_{4,\ell}) \\ &\leq \mathbb{P}(\mathcal{E}_{1,\ell}^c) + \mathbb{P}(\mathcal{E}_{2,\ell}^c|\mathcal{E}_{1,\ell} \cap \mathcal{E}_{3,\ell}) + \mathbb{P}(\mathcal{E}_{3,\ell}^c) + \mathbb{P}(\mathcal{E}_{4,\ell}^c|\mathcal{E}_{1,\ell}) + \mathbb{P}(\mathcal{E}_{5,\ell}^c) + \mathbb{P}(\mathcal{E}_{6,\ell}^c|\mathcal{E}_{1,\ell} \cap \mathcal{E}_{2,\ell} \cap \mathcal{E}_{3,\ell} \cap \mathcal{E}_{4,\ell}) \\ &\leq \frac{5}{\ell^2}.\end{aligned}$$

Therefore, $\mathbb{P}(\mathcal{E}_\ell) \geq 1 - \frac{5}{\ell^2}$. □

B.3.1 Guarantees on the Likelihood Ratio

Lemma B.3.3. *We have with probability at least* $1 - 1/\ell^2$,

$$\sup_{\theta \in \Theta} \frac{1}{T_\ell} \left| \log \frac{\pi_\ell(\theta)}{\pi_\ell(\theta^*)} - \frac{T_\ell}{2} \left\| \theta - \theta^* \right\|^2_{A(\bar{e}_{T_\ell})} \right| \leq \Delta_{\max} \sqrt{\frac{2d \log\left(\frac{(d+T_\ell L^2)\ell^2}{d} \right)}{T_\ell}}.$$

Which implies that $\frac{\pi_\ell(\theta)}{\pi_\ell(\theta^*)} \doteq e^{-T_\ell \|\theta - \theta^*\|^2_{A(\bar{e}_{T_\ell})}}$.

Proof. Throughout the following we set $T := T_\ell$. Recall that $\pi_\ell(\theta) = \mathcal{N}(\widehat{\theta}_{T+1}, V_T^{-1})$ restricted on Θ, which means that for each $\theta \in \Theta$,

$$\pi_\ell(\theta) = \frac{\exp\left(-\frac{1}{2} \left\| \theta - \widehat{\theta}_{T+1} \right\|^2_{V_T} \right)}{\int_\Theta \exp\left(-\frac{1}{2} \left\| \theta' - \widehat{\theta}_{T+1} \right\|^2_{V_T} \right) d\theta'}.$$

Since the denominator is independent of θ, this means that

$$\frac{\pi_\ell(\theta)}{\pi_\ell(\theta^*)} = \exp\left(-\frac{1}{2} \left(\left\| \theta - \widehat{\theta}_{T+1} \right\|^2_{V_T} - \left\| \theta^* - \widehat{\theta}_{T+1} \right\|^2_{V_T} \right) \right)$$

where

$$\begin{aligned}
&\left\| \theta^* - \widehat{\theta}_{T+1} \right\|^2_{V_T} - \left\| \theta - \widehat{\theta}_{T+1} \right\|^2_{V_T} \\
&= \|\theta^*\|^2_{V_T} - 2(\theta^*)^\top V_T \widehat{\theta} + \left\| \widehat{\theta}_{T+1} \right\|^2_{V_T} - \left\| \widehat{\theta}_{T+1} \right\|^2_{V_T} + 2(\widehat{\theta}_{T+1})^\top V_T \theta - \|\theta\|^2_{V_T} \\
&= \|\theta^*\|^2_{V_T} - 2\left(\theta^*\right)^\top V_T \left(\theta^* + V_T^{-1} \sum_{s=1}^{T} \epsilon_s x_s \right) + 2\theta^\top V_T \left(\theta^* + V_T^{-1} \sum_{s=1}^{T} \epsilon_s x_s \right) - \|\theta\|^2_{V_T} \\
&= \|\theta^*\|^2_{V_T} - 2\|\theta^*\|^2_{V_T} - 2(\theta^*)^\top \left(\sum_{s=1}^{T} \epsilon_s x_s \right) + 2(\theta^*)^\top V_T \theta + 2\theta^\top \left(\sum_{s=1}^{T} \epsilon_s x_s \right) - \|\theta\|^2_{V_T} \\
&= -\|\theta^* - \theta\|^2_{V_T} - 2\left\langle \theta^* - \theta, \sum_{s=1}^{T} \epsilon_s x_s \right\rangle \\
&= -\|\theta^* - \theta\|^2_{V_T} - 2\sum_{s=1}^{T} \epsilon_s x_s^\top (\theta^* - \theta).
\end{aligned}$$

Note that

$$\sum_{s=1}^{T} \epsilon_s x_s^\top (\theta^* - \theta) = \sum_{s=1}^{T} \epsilon_s x_s^\top V_T^{-1/2} V_T^{1/2} (\theta^* - \theta)$$
$$\leq \left\| \sum_{s=1}^{T} \epsilon_s x_s \right\|_{V_T^{-1}} \|\theta^* - \theta\|_{V_T}.$$

Note that

$$\|\theta^* - \theta\|_{V_T} = \sqrt{\|\theta^* - \theta\|_{V_T}^2} = \sqrt{\sum_{t=1}^{T} (x_t^\top (\theta^* - \theta))^2} \leq \Delta_{\max} \sqrt{T},$$

and since $\mathbb{E}[\epsilon_s x_s | \mathcal{F}_{s-1}] = 0$ for all s, $\epsilon_s x_s$ is a vector-valued martingale. Then by Theorem 1 of Abbasi-Yadkori et al. [2011], with probability greater than $1 - \delta$,

$$\left\| \sum_{s=1}^{T} \epsilon_s x_s \right\|_{V_T^{-1}} \leq \sqrt{2d \log \left(\frac{d + TL^2}{d\delta} \right)}$$

so with probability $1 - \delta$,

$$\left\| \sum_{s=1}^{T} \epsilon_s x_s \right\|_{V_T^{-1}} \|\theta^* - \theta\|_{V_T} \leq \Delta_{\max} \sqrt{T} \sqrt{2d \log \left(\frac{d + TL^2}{d\delta} \right)}.$$

so for any $\theta \in \Theta$,

$$\left| \left(\left\| \theta - \widehat{\theta}_{T+1} \right\|_{V_T}^2 - \left\| \theta^* - \widehat{\theta}_{T+1} \right\|_{V_T}^2 \right) - \|\theta^* - \theta\|_{V_T}^2 \right| \leq \Delta_{\max} \sqrt{T} \sqrt{2d \log \left(\frac{d + TL^2}{d\delta} \right)},$$

which means that

$$\left| \log \frac{\pi_\ell(\theta^*)}{\pi_\ell(\theta)} - \frac{T}{2} \|\theta - \theta^*\|_{A(\bar{e}_T)}^2 \right| \leq \Delta_{\max} \sqrt{T} \sqrt{2d \log \left(\frac{d + TL^2}{d\delta} \right)}.$$

Taking a supremum over $\theta \in \Theta$ on both sides and taking $\delta = \frac{1}{\ell^2}$ gives the result.

□

B.3.2 Guarantee on Saddle-Point Convergence of PEPS in Round ℓ

In this section, we present a key result to this proof, which shows that as round ℓ gets large, the distribution from PEPS achieves the optimal allocation deduced by τ^*. Fix a round ℓ. At iteration t, let $\tilde{\lambda}_t$ denote the sampling distribution of x_t . The result is stated in the following lemma. In the proof, we decompose the difference into several terms and argue about each piece in subsequent sections.

Lemma B.3.4 (Guarantee for PEPS)**.** *On $\mathcal{E}_{1,\ell} \cap \mathcal{E}_{2,\ell} \cap \mathcal{E}_{3,\ell} \cap \mathcal{E}_{4,\ell}$, for $\ell > \ell_0$ then at the end of epoch ℓ, we have with probability at least $1 - \frac{1}{\ell^2}$,*

$$\tau^* - \inf_{\theta \in \Theta^c_{z_*}} \left[\frac{1}{2} \left\| \theta^* - \theta \right\|^2_{A(\bar{e}_{T_\ell})} \right] \leq \epsilon_\ell$$

for a sequence $\epsilon_\ell \to 0$ as $\ell \to \infty$.

Proof. Recall the definition of $\bar{p}_{T_\ell}$ and $\bar{e}_{T_\ell}$ in Section B.1. We first show that there exists some ϵ_ℓ that goes to zero as $\ell \to \infty$ such that under $\mathcal{E}_{1,\ell} \cap \mathcal{E}_{2,\ell} \cap \mathcal{E}_{3,\ell} \cap \mathcal{E}_{4,\ell}$, for $\ell > \ell_0$,

$$\max_{\lambda \in \triangle_{\mathcal{X}}} \mathbb{F}_{\theta \sim \bar{p}_{T_\ell}} \left[\frac{1}{2} \left\| \theta^* - \theta \right\|^2_{A(\lambda)} \right] - \min_{p \in \mathcal{P}(\Theta^c_{z_*})} \mathbb{F}_{\theta \sim p} \left[\frac{1}{2} \left\| \theta^* - \theta \right\|^2_{A(\bar{e}_{T_\ell})} \right] \leq \epsilon_\ell.$$

We have

$$
\begin{aligned}
&\max_{\lambda\in\triangle_{\mathcal{X}}} \mathbb{F}_{\theta\sim\bar{p}_{T_\ell}}\left[\left\|\theta^*-\theta\right\|^2_{A(\lambda)}\right] - \min_{p\in\mathcal{P}(\Theta^c_{z_*})} \mathbb{F}_{\theta\sim p}\left[\left\|\theta^*-\theta\right\|^2_{A(\bar{e}_{T_\ell})}\right] \\
&= \max_{\lambda\in\triangle_{\mathcal{X}}} \mathbb{F}_{\theta\sim\bar{p}_{T_\ell}}\left[\left\|\theta^*-\theta\right\|^2_{A(\lambda)}\right] - \inf_{\theta\in\Theta^c_{z_*}} \left\|\theta^*-\theta\right\|^2_{A(\bar{e}_{T_\ell})} \\
&= \max_{\lambda\in\triangle_{\mathcal{X}}} \mathbb{F}_{\theta\sim\bar{p}_{T_\ell}}\left[\left\|\theta^*-\theta\right\|^2_{A(\lambda)}\right] - \frac{1}{T_\ell}\inf_{\theta\in\Theta^c_{z_*}} \left\|\theta-\widehat{\theta}_{T_\ell+1}\right\|^2_{V_{T_\ell}} + C''_{T_\ell} \\
&= \max_{\lambda\in\triangle_{\mathcal{X}}} \mathbb{F}_{\theta\sim\bar{p}_{T_\ell}}\left[\left\|\theta^*-\theta\right\|^2_{A(\lambda)}\right] - \max_{\lambda\in\triangle_{\mathcal{X}}} \frac{1}{T_\ell}\sum_{t=1}^{T_\ell} \mathbb{F}_{\theta\sim p_t}\left[\left\|\widehat{\theta}_t-\theta\right\|^2_{A(\lambda)}\right] \qquad \text{(S1. } C'_{T_\ell}\text{)} \\
&\quad + \max_{\lambda\in\triangle_{\mathcal{X}}} \frac{1}{T_\ell}\sum_{t=1}^{T_\ell} \mathbb{F}_{\theta\sim p_t}\left[\left\|\widehat{\theta}_t-\theta\right\|^2_{A(\lambda)}\right] - \frac{1}{T_\ell}\sum_{t=1}^{T_\ell} \mathbb{F}_{\theta\sim p_t}\left[\left\|\theta-\widehat{\theta}_t\right\|^2_{A(\tilde{\lambda}_t)}\right] \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{(S2. regret for max learner)} \\
&\quad + \frac{1}{T_\ell}\sum_{t=1}^{T_\ell} \mathbb{F}_{\theta\sim p_t}\left[\left\|\theta-\widehat{\theta}_t\right\|^2_{A(\tilde{\lambda}_t)}\right] - \frac{1}{T_\ell}\sum_{t=1}^{T_\ell} \mathbb{F}_{\theta\sim \tilde{p}_t}\left[\left\|\theta-\widehat{\theta}_t\right\|^2_{A(\tilde{\lambda}_t)}\right] \qquad \text{(S3.)} \\
&\quad + \frac{1}{T_\ell}\sum_{t=1}^{T_\ell} \mathbb{F}_{\theta\sim \tilde{p}_t}\left[\left\|\theta-\widehat{\theta}_t\right\|^2_{A(\tilde{\lambda}_t)}\right] - \frac{1}{T_\ell}\sum_{t=1}^{T_\ell} \mathbb{F}_{\theta\sim \tilde{p}_t}\left[\left\|\theta-\widehat{\theta}_t\right\|^2_{x_t x_t^\top}\right] \qquad \text{(S4.)} \\
&\quad + \frac{1}{T_\ell}\sum_{t=1}^{T_\ell} \mathbb{F}_{\theta\sim \tilde{p}_t}\left[\left\|\theta-\widehat{\theta}_t\right\|^2_{x_t x_t^\top}\right] - \frac{1}{T_\ell}\inf_{\theta\in\Theta^c_{z_*}} \left\|\theta-\widehat{\theta}_{T_\ell+1}\right\|^2_{V_{T_\ell}} \quad \text{(S5. regret for the min learner)} \\
&\quad + C''_{T_\ell},
\end{aligned}
$$

where we define

$$
\begin{aligned}
C'_{T_\ell} &:= \max_{\lambda\in\triangle_{\mathcal{X}}} \mathbb{F}_{\theta\sim\bar{p}_{T_\ell}}\left[\left\|\theta^*-\theta\right\|^2_{A(\lambda)}\right] - \max_{\lambda\in\triangle_{\mathcal{X}}} \frac{1}{T_\ell}\sum_{t=1}^{T_\ell} \mathbb{F}_{\theta\sim p_t}\left[\left\|\widehat{\theta}_t-\theta\right\|^2_{A(\lambda)}\right] \\
C''_{T_\ell} &= \frac{1}{T_\ell}\inf_{\theta\in\Theta^c_{z_*}} \left\|\theta-\widehat{\theta}_{T_\ell+1}\right\|^2_{V_{T_\ell}} - \inf_{\theta\in\Theta^c_{z_*}} \left\|\theta^*-\theta\right\|^2_{A(\bar{e}_{T_\ell})}.
\end{aligned}
$$

We now handle each term separately by referring to the lemma which provides a guarantee.

- **(S1)** By Lemma B.3.10, under $\mathcal{E}_{1,\ell}\cap\mathcal{E}_{2,\ell}$, for $T_\ell \geq T_2(\ell)$,

$$
\begin{aligned}
&\max_{\lambda\in\triangle_{\mathcal{X}}} \mathbb{F}_{\theta\sim\bar{p}_{T_\ell}}\left[\left\|\theta^*-\theta\right\|^2_{A(\lambda)}\right] - \max_{\lambda\in\triangle_{\mathcal{X}}} \frac{1}{T_\ell}\sum_{t=1}^{T_\ell} \mathbb{F}_{\theta\sim p_t}\left[\left\|\widehat{\theta}_t-\theta\right\|^2_{A(\lambda)}\right] \\
&\leq \frac{T_2(\ell)L^2\beta(T_2(\ell),\ell^2)}{T_\ell} + 4d\beta(T_\ell,\ell^2)T_\ell^{-3/4},
\end{aligned}
$$

so for $T_\ell \geq T_2(\ell)^{3/2}$, we have the above is upper bounded by

$$O(L^2\beta(T_2(\ell),\ell^2)T_\ell^{-1/2} + 4d\beta(T_\ell,\ell^2)T_\ell^{-3/4});$$

- **(S2)** By Lemma B.3.5, we have with probability $1-1/(3\ell^2)$ conditioned on $\mathcal{E}_{2,\ell}$

$$\begin{aligned}&\max_{\lambda\in\triangle_{\mathcal{X}}} \sum_{t=1}^{T_\ell} \mathbb{F}_{\theta\sim p_t, x\sim\lambda} \left\|\theta-\widehat{\theta}_t\right\|^2_{xx^\top} - \sum_{t=1}^{T_\ell} \mathbb{F}_{\theta\sim p_t, x\sim\tilde{\lambda}_t} \left\|\theta-\widehat{\theta}_t\right\|^2_{xx^\top} \\ &\leq 2C_{3,\ell}^2\sqrt{\log|\mathcal{X}|T_\ell} + \sqrt{2C_{3,\ell}^2 T_\ell|\mathcal{X}|\log(T_\ell\ell^2)} + 2C_{3,\ell}^2\sum_{t=1}^{T_\ell}\gamma_t,\end{aligned}$$

so with a choice of $\gamma_t = t^{-\alpha}$ with $\alpha = 1/4$,

$$\begin{aligned}&\max_{\lambda\in\triangle_{\mathcal{X}}} \frac{1}{T_\ell}\sum_{t=1}^{T_\ell} \mathbb{F}_{\theta\sim p_t, x\sim\lambda} \left\|\theta-\widehat{\theta}_t\right\|^2_{xx^\top} - \frac{1}{T_\ell}\sum_{t=1}^{T_\ell} \mathbb{F}_{\theta\sim p_t, x\sim\tilde{\lambda}_t} \left\|\theta-\widehat{\theta}_t\right\|^2_{xx^\top} \\ &\leq C_{3,\ell}^2\sqrt{\log|\mathcal{X}|}T_\ell^{-1/2} + \sqrt{2C_{3,\ell}^2\log\ell^2}T_\ell^{-1/2} + \sqrt{2C_{3,\ell}^2|\mathcal{X}|\log(3T\ell^2)}T_\ell^{-1/2} + 2C_{3,\ell}^2T_\ell^{-1/4}\end{aligned}$$

- **(S3)** By Lemma B.3.12, we have conditioned on $\mathcal{E}_{4,\ell}\cap\mathcal{E}_{1,\ell}$ for $\ell\geq\ell_0$,

$$\frac{1}{T_\ell}\sum_{t=1}^{T_\ell}\mathbb{F}_{\theta\sim p_t}\left[\left\|\theta-\widehat{\theta}_t\right\|^2_{A(\tilde{\lambda}_t)}\right] - \frac{1}{T_\ell}\sum_{t=1}^{T_\ell}\mathbb{F}_{\theta\sim\tilde{p}_t}\left[\left\|\theta-\widehat{\theta}_t\right\|^2_{A(\tilde{\lambda}_t)}\right] \leq \frac{2C_{3,\ell}^2T_0(\ell)}{T_\ell}$$

for $T_\ell \geq T_0(\ell)^{3/2}$, we have the above is bounded by $2C_{3,\ell}^2T_\ell^{-1/2}$;

- **(S4)** By Lemma B.3.8, we have with probability $1-1/(3\ell^2)$, conditioned on $\mathcal{E}_{2,\ell}$,

$$\frac{1}{T_\ell}\sum_{t=1}^{T_\ell}\mathbb{F}_{\theta\sim\tilde{p}_t}\left[\left\|\theta-\widehat{\theta}_t\right\|^2_{A(\tilde{\lambda}_t)}\right] - \frac{1}{T_\ell}\sum_{t=1}^{T_\ell}\mathbb{F}_{\theta\sim\tilde{p}_t}\left[\left\|\theta-\widehat{\theta}_t\right\|^2_{x_tx_t^\top}\right] \leq \sqrt{\frac{2C_{1,\ell}\log\ell^2}{T_\ell}}$$

- **(S5)** By Lemma B.3.7, we have with probability $1-1/(3\ell^2)$, conditioned on $\mathcal{E}_{1,\ell}\cap\mathcal{E}_{2,\ell}$,

$$\begin{aligned}&\frac{1}{T_\ell}\left[\sum_{t=1}^{T_\ell}\mathbb{E}_{\theta\sim\tilde{p}_t}\left[\left\|\theta-\widehat{\theta}_t\right\|^2_{x_tx_t^\top}\right] - \inf_{\theta\in\Theta^c_{z_*}}\left\|\theta-\widehat{\theta}_{T_\ell+1}\right\|^2_{V_{T_\ell}}\right] \\ &\leq \sqrt{\frac{C_{3,\ell}^2 d\log(T_\ell C_{1,\ell})}{T_\ell}} + C_{3,\ell}\sqrt{\frac{2d\beta(T_\ell,\ell^2)}{T_\ell}\log\left(\frac{d+T_\ell L^2}{d}\right)} + C_{3,\ell}\sqrt{\frac{2\log(\ell^2)}{T_\ell}}.\end{aligned}$$

- (C''_{T_ℓ}) By Lemma B.3.11, conditioned on $\mathcal{E}_{1,\ell} \cap \mathcal{E}_{2,\ell}$, we have

$$\left| \frac{1}{T_\ell} \inf_{\theta \in \Theta^c_{z_*}} \left\| \theta - \widehat{\theta}_{T_\ell+1} \right\|^2_{V_{T_\ell}} - \frac{1}{T_\ell} \inf_{\theta \in \Theta^c_{z_*}} \|\theta^* - \theta\|^2_{V_{T_\ell}} \right| \leq (C_{3,\ell} + \Delta_{\max}) \sqrt{\frac{\beta(T_\ell, \ell^2)}{T_\ell}}$$

Add them altogether, we get that with probability greater than $1 - 1/\ell^2$ on $\mathcal{E}_{1,\ell} \cap \mathcal{E}_{2,\ell} \cap \mathcal{E}_{4,\ell}$

$$\begin{aligned}
&\max_{\lambda \in \triangle_{\mathcal{X}}} \mathbb{E}_{\theta \sim \bar{p}_{T_\ell}} \left[\|\theta^* - \theta\|^2_{A(\lambda)} \right] - \min_{p \in \mathcal{P}(\Theta^c_{z_*})} \mathbb{E}_{\theta \sim p} \left[\|\theta^* - \theta\|^2_{A(\bar{\lambda}_{T_\ell})} \right] \\
&\leq L^2 \beta(T_2(\ell), \ell^2) T_\ell^{-1/2} + 4d\beta(T_\ell, \ell^2) T_\ell^{-3/4} \\
&\quad + C^2_{3,\ell} \sqrt{\log |\mathcal{X}|} T_\ell^{-1/2} + \sqrt{2C^2_{3,\ell} \log \ell^2} T_\ell^{-1/2} + \sqrt{2C^2_{3,\ell} |\mathcal{X}| \log(3T\ell^2)} T_\ell^{-1/2} + C^2_{3,\ell} T_\ell^{-1/4} \\
&\quad + 2C^2_{3,\ell} T_\ell^{-1/2} + \sqrt{\frac{2C_{1,\ell} \log \ell^2}{T_\ell}} \\
&\quad + \sqrt{\frac{C^2_{3,\ell} d \log(T_\ell C_{1,\ell})}{T_\ell}} + C_{3,\ell} \sqrt{\frac{2d\beta(T_\ell, \ell^2)}{T_\ell} \log\left(\frac{d + T_\ell L^2}{d} \right)} + C_{3,\ell} \sqrt{\frac{2\log(\ell^2)}{T_\ell}} \\
&\quad + (C_{3,\ell} + \Delta_{\max}) \sqrt{\frac{\beta(T_\ell, \ell^2)}{T_\ell}}.
\end{aligned}$$

Note that each term approaches zero as $T_\ell \to \infty$. By the choice of $T_\ell = 2^\ell$ in the algorithm, this implies that there exists some $\epsilon_\ell > 0$ with $\epsilon_\ell \to 0$ as $\ell \to \infty$ such that for each ℓ,

$$\max_{\lambda \in \triangle_{\mathcal{X}}} \mathbb{F}_{\theta \sim \bar{p}_{T_\ell}} \left[\|\theta^* - \theta\|^2_{A(\lambda)} \right] - \min_{p \in \mathcal{P}(\Theta^c_{z_*})} \mathbb{F}_{\theta \sim p} \left[\|\theta^* - \theta\|^2_{A(\bar{e}_{T_\ell})} \right] \leq \epsilon_\ell. \tag{B.1}$$

Now we show how this result leads to the saddle point convergence. Note that

$$\max_{\lambda \in \triangle_{\mathcal{X}}} \mathbb{F}_{\theta \sim \bar{p}_{T_\ell}} \left[\|\theta^* - \theta\|^2_{A(\lambda)} \right] \geq \max_{\lambda \in \triangle_{\mathcal{X}}} \min_{p \in \mathcal{P}(\Theta^c_{z_*})} \mathbb{F}_{\theta \sim p} [\|\theta^* - \theta\|^2_{A(\lambda)}] \geq \min_{p \in \mathcal{P}(\Theta^c_{z_*})} \mathbb{F}_{\theta \sim p} \left[\|\theta^* - \theta\|^2_{A(\bar{e}_{T_\ell})} \right],$$

so using Equation B.1 we have

$$\max_{\lambda \in \triangle_{\mathcal{X}}} \min_{p \in \mathcal{P}(\Theta^c_{z_*})} \mathbb{F}_{\theta \sim p} [\|\theta^* - \theta\|^2_{A(\lambda)}] - \min_{p \in \mathcal{P}(\Theta^c_{z_*})} \mathbb{F}_{\theta \sim p} \left[\|\theta^* - \theta\|^2_{A(\bar{e}_{T_\ell})} \right] \leq \epsilon_\ell.$$

However, note that

$$\min_{p \in \mathcal{P}(\Theta^c_{z_*})} \mathbb{F}_{\theta \sim p} \left[\|\theta^* - \theta\|^2_{A(\bar{e}_{T_\ell})} \right] = \inf_{\theta \in \Theta^c_{z_*}} \|\theta^* - \theta\|^2_{A(\bar{e}_{T_\ell})}$$

and $\max_{\lambda \in \triangle_{\mathcal{X}}} \min_{p \in \mathcal{P}(\Theta^c_{z_*})} \mathbb{F}_{\theta \sim p} [\|\theta^* - \theta\|^2_{A(\lambda)}] = \tau^*$, we have shown that

$$\tau^* - \inf_{\theta \in \Theta^c_{z_*}} \|\theta^* - \theta\|^2_{A(\bar{e}_{T_\ell})} < \epsilon_\ell.$$

□

B.3.3 Guarantees on the max*-learner*

In this section, we show that the max-learner gets sublinear regret as ℓ gets large. The key idea is that we mix a diminishing amount of G-optimal distribution each round, and we show that by its diminishing nature, the mixing of G-optimal distribution keeps the regret sublinear.

Lemma B.3.5. *Under $\mathcal{E}_{\ell,2}$, with the choice of $\eta_\lambda = \sqrt{\frac{\log|\mathcal{X}|}{C_{3,\ell}^4 T}}$, we have with probability greater than $1 - 1/\ell^2$,*

$$\begin{aligned}
&\max_{\lambda \in \triangle_{\mathcal{X}}} \sum_{t=1}^{T_\ell} \mathbb{F}_{\theta\sim p_t, x\sim\lambda} \left\|\theta - \widehat{\theta}_t\right\|_{xx^\top}^2 - \sum_{t=1}^{T_\ell} \mathbb{F}_{\theta\sim p_t, x\sim\tilde{\lambda}_t} \left\|\theta - \widehat{\theta}_t\right\|_{xx^\top}^2 \\
&\leq 2C_{3,\ell}^2 \sqrt{\log|\mathcal{X}|T_\ell} + \sqrt{2C_{3,\ell}^2 T_\ell |\mathcal{X}| \log(T_\ell \ell^2)} + 2C_{3,\ell}^2 \sum_{t=1}^{T_\ell} \gamma_t.
\end{aligned}$$

Proof. We first show that the statement is true for some fixed λ, i.e. we would like to show that with probability $1-\delta$,

$$\begin{aligned}
&\sum_{t=1}^{T_\ell} \mathbb{F}_{\theta\sim p_t, x\sim\lambda} \left[\left\|\theta - \widehat{\theta}_t\right\|_{xx^\top}^2\right] - \sum_{t=1}^{T_\ell} \mathbb{F}_{\theta\sim p_t, x\sim\tilde{\lambda}_t} \left[\left\|\theta - \widehat{\theta}_t\right\|_{xx^\top}^2\right] \\
&\leq C_{3,\ell}^2 \sqrt{\log|\mathcal{X}|T_\ell} + \sqrt{2C_{3,\ell}^2 T_\ell \log(1/\delta)} + 2C_{3,\ell}^2 \sum_{t=1}^{T_\ell} \gamma_t.
\end{aligned}$$

Let $\mathcal{F}_{t-1}$ be the history up to time t. Then for any fixed λ,

$$\mathbb{E}_{\theta_t}[\mathbb{F}_{x\sim\lambda}[\|\widehat{\theta}_t - \theta_t\|_{xx^\top}^2]|\mathcal{F}_{t-1}] = \mathbb{F}_{\theta\sim p_t, x\sim\lambda}[\|\widehat{\theta}_t - \theta\|_{xx^\top}^2].$$

Thus, setting

$$\begin{aligned}
X_t = &\mathbb{F}_{x\sim\tilde{\lambda}_t} \left[\left\|\theta_t - \widehat{\theta}_t\right\|_{xx^\top}^2\right] - \mathbb{F}_{x\sim\tilde{\lambda}_t, \theta\sim p_t} \left\|\theta - \widehat{\theta}_t\right\|_{xx^\top}^2 \\
&- \left[\mathbb{F}_{x\sim\lambda} \left[\left\|\theta_t - \widehat{\theta}_t\right\|_{xx^\top}^2\right] - \mathbb{F}_{x\sim\lambda, \theta\sim p_t} \left\|\theta - \widehat{\theta}_t\right\|_{xx^\top}^2\right]
\end{aligned}$$

we see that the X_t form a Martingale difference sequence, i.e. $\mathbb{E}[X_t|\mathcal{F}_{t-1}] = 0$. Note that for any $\theta \in \Theta$,

$$\begin{aligned}
&\mathbb{F}_{x\sim\lambda_t}\left[\left\|\theta-\widehat{\theta}_t\right\|^2_{xx^\top}\right] \\
&= \mathbb{F}_{x\sim\tilde{\lambda}_t}\left[\left\|\theta-\widehat{\theta}_t\right\|^2_{xx^\top}\right] + \gamma_t\left(\mathbb{F}_{x\sim\lambda_t}\left[\left\|\theta-\widehat{\theta}_t\right\|^2_{xx^\top}\right] - \mathbb{F}_{x\sim\lambda^G}\left[\left\|\theta-\widehat{\theta}_t\right\|^2_{xx^\top}\right]\right),
\end{aligned}$$

Since under $\mathcal{E}_{2,\ell}$, we have for any $x \in \mathcal{X}$, $\theta \in \Theta$, any $t \le T_\ell$, $\left\|\theta-\widehat{\theta}_t\right\|^2_{xx^\top} \le C^2_{3,\ell}$, we have for any $\theta \in \Theta$,

$$\mathbb{F}_{x\sim\lambda_t}\left[\left\|\theta-\widehat{\theta}_t\right\|^2_{xx^\top}\right] \le \mathbb{F}_{x\sim\tilde{\lambda}_t}\left[\left\|\theta-\widehat{\theta}_t\right\|^2_{xx^\top}\right] + 2C^2_{3,\ell}\gamma_t.$$

Then we have

$$\begin{aligned}
&\sum_{t=1}^{T_\ell}\mathbb{F}_{\theta\sim p_t,x\sim\lambda}\left[\left\|\theta-\widehat{\theta}_t\right\|^2_{xx^\top}\right] - \sum_{t=1}^{T_\ell}\mathbb{F}_{\theta\sim p_t,x\sim\tilde{\lambda}_t}\left[\left\|\theta-\widehat{\theta}_t\right\|^2_{xx^\top}\right] \\
&= \sum_{t=1}^{T_\ell}\mathbb{F}_{x\sim\lambda}\left[\left\|\theta_t-\widehat{\theta}_t\right\|^2_{xx^\top}\right] - \sum_{t=1}^{T_\ell}\mathbb{F}_{x\sim\tilde{\lambda}_t}\left[\left\|\theta_t-\widehat{\theta}_t\right\|^2_{xx^\top}\right] \\
&\quad - \left[\sum_{t=1}^{T_\ell}\mathbb{F}_{x\sim\lambda}\left[\left\|\theta_t-\widehat{\theta}_t\right\|^2_{xx^\top}\right] - \sum_{t=1}^{T_\ell}\mathbb{F}_{x\sim\tilde{\lambda}_t}\left[\left\|\theta_t-\widehat{\theta}_t\right\|^2_{xx^\top}\right]\right] \\
&\quad - \left[\sum_{t=1}^{T_\ell}\mathbb{F}_{\theta\sim p_t,x\sim\lambda}\left\|\theta-\widehat{\theta}_t\right\|^2_{xx^\top} - \sum_{t=1}^{T_\ell}\mathbb{F}_{x\sim\tilde{\lambda}_t,\theta\sim p_t}\left\|\theta-\widehat{\theta}_t\right\|^2_{xx^\top}\right]\Bigg] \\
&\le \sum_{t=1}^{T_\ell}\mathbb{F}_{x\sim\lambda}\left[\left\|\theta_t-\widehat{\theta}_t\right\|^2_{xx^\top}\right] - \sum_{t=1}^{T_\ell}\mathbb{F}_{x\sim\lambda_t}\left[\left\|\theta_t-\widehat{\theta}_t\right\|^2_{xx^\top}\right] \\
&\quad - \left[\sum_{t=1}^{T_\ell}\mathbb{F}_{x\sim\lambda}\left[\left\|\theta_t-\widehat{\theta}_t\right\|^2_{xx^\top}\right] - \sum_{t=1}^{T_\ell}\mathbb{F}_{x\sim\tilde{\lambda}_t}\left[\left\|\theta_t-\widehat{\theta}_t\right\|^2_{xx^\top}\right]\right] \\
&\quad - \left[\sum_{t=1}^{T_\ell}\mathbb{F}_{\theta\sim p_t,x\sim\lambda}\left\|\theta-\widehat{\theta}_t\right\|^2_{xx^\top} - \sum_{t=1}^{T_\ell}\mathbb{F}_{x\sim\tilde{\lambda}_t,\theta\sim p_t}\left\|\theta-\widehat{\theta}_t\right\|^2_{xx^\top}\right]\Bigg] + 2C^2_{3,\ell}\sum_{t=1}^{T_\ell}\gamma_t \qquad \text{(B.2)}
\end{aligned}$$

Note that

$$\sum_{t=1}^{T_\ell} \mathbb{F}_{x\sim\lambda}\left[\left\|\theta_t - \widehat{\theta}_t\right\|^2_{xx^\top}\right] - \sum_{t=1}^{T_\ell} \mathbb{F}_{x\sim\tilde{\lambda}_t}\left[\left\|\theta_t - \widehat{\theta}_t\right\|^2_{xx^\top}\right] - \left[\sum_{t=1}^{T_\ell} \mathbb{F}_{\theta\sim p_t, x\sim\lambda}\left\|\theta - \widehat{\theta}_t\right\|^2_{xx^\top} - \sum_{t=1}^{T_\ell} \mathbb{F}_{x\sim\tilde{\lambda}_t,\theta\sim p_t}\left\|\theta - \widehat{\theta}_t\right\|^2_{xx^\top}\right] = \sum_{t=1}^{T_\ell} X_t.$$

We know that under $\mathcal{E}_{2,\ell}$, we have for any $x \in \mathcal{X}$, $\theta \in \Theta$, any $t \le T_\ell$, $\left\|\theta - \widehat{\theta}_t\right\|^2_{xx^\top} \le C^2_{3,\ell}$. Then, for any t, $|X_t| \le 4C^2_{3,\ell}$, so by Azuma-Hoeffding, with probability $1-\delta$, $\sum_{t=1}^{T_\ell} X_t \le \sqrt{8C^2_{3,\ell} T_\ell \log(1/\delta)}$. Plugging the above and Lemma B.3.6 in Equation B.2 gives us

$$\sum_{t=1}^{T_\ell} \mathbb{F}_{\theta\sim p_t, x\sim\lambda}\left[\left\|\theta - \widehat{\theta}_t\right\|^2_{xx^\top}\right] - \sum_{t=1}^{T_\ell} \mathbb{F}_{\theta\sim p_t, x\sim\tilde{\lambda}_t}\left[\left\|\theta - \widehat{\theta}_t\right\|^2_{xx^\top}\right] \le C^2_{3,\ell}\sqrt{\log|\mathcal{X}|T_\ell} + \sqrt{2C^2_{3,\ell}T_\ell\log(1/\delta)} + 2C^2_{3,\ell}\sum_{t=1}^{T_\ell}\gamma_t.$$

This result holds for any λ, but in particular we want it to hold for the λ which maximizes the reward, so we perform a covering argument on λ.

We take an ϵ-cover $\mathcal{S}_\epsilon$ of $\triangle_\mathcal{X}$ in $\|\cdot\|_1$. Then, we know that for any $\lambda \in \triangle_\mathcal{X}$, there is some $\lambda' \in \mathcal{S}_\epsilon$ such that $\|\lambda - \lambda'\|_1 \le \epsilon$. Let $w_t(\lambda) := \mathbb{F}_{\theta\sim p_t, x\sim\lambda}\left\|\theta - \widehat{\theta}_t\right\|^2_{xx^\top}$. Then, note that for any t and $\lambda_1, \lambda_2 \in \triangle_\mathcal{X}$,

$$\begin{aligned} w(\lambda_1) - w(\lambda_2) &= \mathbb{F}_{\theta\sim p_t, x\sim\lambda_1}\left\|\theta - \widehat{\theta}_t\right\|^2_{xx^\top} - \mathbb{F}_{\theta\sim p_t, x\sim\lambda_2}\left\|\theta - \widehat{\theta}_t\right\|^2_{xx^\top} \\ &= \mathbb{F}_{\theta\sim p_t}\sum_x([\lambda_1]_x - [\lambda_2]_x)(x^\top(\theta - \widehat{\theta}_t))^2 \\ &\le C^2_{3,\ell}\mathbb{F}_{\theta\sim p_t}\sum_x([\lambda_1]_x - [\lambda_2]_x) \\ &= C^2_{3,\ell}\|\lambda_1 - \lambda_2\|_1, \end{aligned}$$

so $w_t(\lambda)$ is $C^2_{3,\ell}$-Lipschitz for any t. Then, assuming that $\bar{\lambda} \in \triangle_\mathcal{X}$ satisfies that

$$\sum_{t=1}^{T_\ell} \mathbb{F}_{\theta\sim p_t, x\sim\bar{\lambda}}\left\|\theta - \widehat{\theta}_t\right\|^2_{xx^\top} = \max_{\lambda\in\triangle_\mathcal{X}} \sum_{t=1}^{T_\ell} \mathbb{F}_{\theta\sim p_t, x\sim\lambda}\left\|\theta - \widehat{\theta}_t\right\|^2_{xx^\top},$$

we can find some $\lambda_0 \in \mathcal{S}_\epsilon$ such that $\left\|\lambda_0 - \bar{\lambda}\right\| \leq \epsilon$, so by Lipschitzness of w_t for any t, we have

$$\begin{aligned}
&\max_{\lambda \in \triangle_{\mathcal{X}}} \sum_{t=1}^{T_\ell} \mathbb{F}_{\theta \sim p_t, x \sim \lambda} \left\|\theta - \widehat{\theta}_t\right\|_{xx^\top}^2 - \max_{\lambda \in \mathcal{S}_\epsilon} \sum_{t=1}^{T_\ell} \mathbb{F}_{\theta \sim p_t, x \sim \lambda} \left\|\theta - \widehat{\theta}_t\right\|_{xx^\top}^2 \\
&= \sum_{t=1}^{T_\ell} \mathbb{F}_{\theta \sim p_t, x \sim \bar{\lambda}} \left\|\theta - \widehat{\theta}_t\right\|_{xx^\top}^2 - \max_{\lambda \in \mathcal{S}_\epsilon} \sum_{t=1}^{T_\ell} \mathbb{F}_{\theta \sim p_t, x \sim \lambda} \left\|\theta - \widehat{\theta}_t\right\|_{xx^\top}^2 \\
&\leq \sum_{t=1}^{T_\ell} \mathbb{F}_{\theta \sim p_t, x \sim \bar{\lambda}} \left\|\theta - \widehat{\theta}_t\right\|_{xx^\top}^2 - \sum_{t=1}^{T_\ell} \mathbb{F}_{\theta \sim p_t, x \sim \lambda_0} \left\|\theta - \widehat{\theta}_t\right\|_{xx^\top}^2 \\
&\leq C_{3,\ell}^2 T_\ell \epsilon.
\end{aligned}$$

Also, let $K = |\mathcal{X}|$. Denote B_1^K as the l_1 ball with dimension K. We know that for $\epsilon \leq 1$, $N(B_1^K, \|\cdot\|_1, \epsilon) \leq \left(\frac{3}{\epsilon}\right)^K$. Since $\triangle_{\mathcal{X}} \subset B_1^K$, we have the covering number

$$N(\triangle_{\mathcal{X}}, \|\cdot\|_1, \epsilon) \leq N(B_1^k, \|\cdot\|_1, \epsilon) \leq \left(\frac{3}{\epsilon}\right)^K.$$

Therefore, $|\mathcal{S}_\epsilon| \leq \left(\frac{3}{\epsilon}\right)^K$. By union bounding over all $\lambda \in \mathcal{S}_\epsilon$, we have with probability at least $1 - \delta$,

$$\begin{aligned}
&\max_{\lambda \in \mathcal{S}_\epsilon} \sum_{t=1}^{T_\ell} \mathbb{F}_{\theta \sim p_t, x \sim \lambda} \left\|\theta - \widehat{\theta}_t\right\|_{xx^\top}^2 - \sum_{t=1}^{T_\ell} \mathbb{F}_{\theta \sim p_t, x \sim \lambda_t} \left\|\theta - \widehat{\theta}_t\right\|_{xx^\top}^2 \\
&\leq C_{3,\ell}^2 \sqrt{\log |\mathcal{X}| T_\ell} + \sqrt{2 C_{3,\ell}^2 T_\ell \log(1/(\delta |\mathcal{S}_\epsilon|))} + 2 C_{3,\ell}^2 \sum_{t=1}^{T_\ell} \gamma_t \\
&\leq C_{3,\ell}^2 \sqrt{\log |\mathcal{X}| T_\ell} + \sqrt{2 C_{3,\ell}^2 T_\ell |\mathcal{X}| \log(3/(\epsilon\delta))} + 2 C_{3,\ell}^2 \sum_{t=1}^{T_\ell} \gamma_t.
\end{aligned}$$

Combining two displays gives us

$$\begin{aligned}
&\max_{\lambda \in \triangle_{\mathcal{X}}} \sum_{t=1}^{T_\ell} \mathbb{F}_{\theta \sim p_t, x \sim \lambda} \left\|\theta - \widehat{\theta}_t\right\|_{xx^\top}^2 - \sum_{t=1}^{T_\ell} \mathbb{F}_{\theta \sim p_t, x \sim \lambda_t} \left\|\theta - \widehat{\theta}_t\right\|_{xx^\top}^2 \\
&\leq C_{3,\ell}^2 \sqrt{\log |\mathcal{X}| T_\ell} + \sqrt{2 C_{3,\ell}^2 T_\ell |\mathcal{X}| \log(3/(\delta\epsilon))} + 2 C_{3,\ell}^2 \sum_{t=1}^{T_\ell} \gamma_t + C_{3,\ell}^2 T_\ell \epsilon.
\end{aligned}$$

Taking $\epsilon = 1/\sqrt{T_\ell}$ and $\delta = 1/\ell^2$ gives us the result. □

Lemma B.3.6. *Under $\mathcal{E}_{2,\ell}$, with the choice of $\eta = \sqrt{\frac{\log|\mathcal{X}|}{C_{3,\ell}^4 T_\ell}}$, we have for any λ,*

$$\sum_{t=1}^{T_\ell} \left\|\theta_t - \widehat{\theta}_t\right\|_{A(\lambda)}^2 - \sum_{t=1}^{T_\ell} \left\|\theta_t - \widehat{\theta}_t\right\|_{A(\lambda_t)}^2 \leq C_{3,\ell}^2 \sqrt{\log|\mathcal{X}| T_\ell}.$$

Proof. Let $\ell_t(\lambda) = -\left\|\theta_t - \widehat{\theta}_t\right\|_{A(\lambda)}^2$. Then we have

$$[\nabla_\lambda \ell_t(\lambda_t)]_x = -\left\|\theta_t - \widehat{\theta}_t\right\|_{xx^\top}^2 = \tilde{g}_{t,x}.$$

Since

$$\max_{t\in[T_\ell]} \|\tilde{g}_t\|_\infty = \max_{t\in[T_\ell], x\in\mathcal{X}} \left\|\theta_t - \widehat{\theta}_t\right\|_{xx^\top}^2 \leq C_{3,\ell}^2,$$

by the guarantee of exponentiated gradient algorithm Orabona [2019], we have that for any λ,

$$\sum_{t=1}^{T_\ell} [\ell_t(\lambda_t) - \ell_t(\lambda)] \leq \frac{\log|\mathcal{X}|}{\eta} + \frac{\eta T_\ell}{2} C_{3,\ell}^4.$$

Plugging in the definition of $\ell_t(\lambda)$, we have

$$\sum_{t=1}^{T_\ell} \left\|\theta_t - \widehat{\theta}_t\right\|_{A(\lambda)}^2 - \sum_{t=1}^{T_\ell} \left\|\theta_t - \widehat{\theta}_t\right\|_{A(\lambda_t)}^2 \leq \frac{\log|\mathcal{X}|}{\eta} + \frac{\eta T_\ell}{2} C_{3,\ell}^4.$$

Choosing $\eta = \sqrt{\frac{\log|\mathcal{X}|}{C_{3,\ell}^4 T_\ell}}$, we have

$$\sum_{t=1}^{T_\ell} \left\|\theta_t - \widehat{\theta}_t\right\|_{A(\lambda)}^2 - \sum_{t=1}^{T_\ell} \left\|\theta_t - \widehat{\theta}_t\right\|_{A(\lambda_t)}^2 \leq C_{3,\ell}^2 \sqrt{\log|\mathcal{X}| T_\ell}.$$

□

B.3.4 Guarantees on the min-*learner*

In this section, we show that the min-learner gets sublinear regret as ℓ gets large. For the min learner, we see that the update for the sampling distribution is very similar to the continuous exponential weights updates Bubeck [2011]. The difference between our setting and continuous exponential weights is that the space $\Theta_{\widehat{z}_t}^c$ is changing each time, so

we potentially have a changing action space each time. To overcome this challenge, we first analyze the regret guarantee when we assume access to the true alternative in Lemma B.3.7, and use Lemma B.3.16 to argue that the estimate $\Theta^c_{\widehat{z}_t}$ is good enough. We state the following guarantee for the min-learner.

Lemma B.3.7. *On event $\mathcal{E}_{\ell,1} \cap \mathcal{E}_{\ell,2}$, with probability $1 - 1/\ell^2$,*

$$\frac{1}{T_\ell}\left[\sum_{t=1}^{T_\ell} \mathbb{E}_{\theta\sim\tilde{p}_t}\left[\left\|\theta - \widehat{\theta}_t\right\|^2_{x_t x_t^\top}\right] - \inf_{\theta\in\Theta^c_{z_*}} \left\|\theta - \widehat{\theta}_{T_\ell+1}\right\|^2_{V_{T_\ell}}\right]$$
$$\leq \sqrt{\frac{C_{3,\ell}^2 d\log(T_\ell C_{1,\ell})}{T_\ell}} + C_{3,\ell}\sqrt{\frac{2d\beta(T_\ell,\ell^2)}{T_\ell}\log\left(\frac{d+T_\ell L^2}{d}\right)} + C_{3,\ell}\sqrt{\frac{2\log(\ell^2)}{T_\ell}}.$$

Proof. We begin by a bound that will be useful in our exponential weights analogy. At iteration t, we apply Hoeffding's lemma with the following upper bound given $\mathcal{E}_{\ell,1} \cap \mathcal{E}_{\ell,2}$ and Lemma B.5.1,

$$\begin{aligned}
&\mathbb{E}_{\theta\sim\tilde{p}_t}\left[\left\|\theta - \widehat{\theta}_t\right\|^2_{x_t x_t^\top} + \left\|\theta - \widehat{\theta}_{t+1}\right\|^2_{V_t} - \left\|\theta - \widehat{\theta}_t\right\|^2_{V_t}\right] \\
&\leq C_{3,\ell}^2 + \mathbb{E}_{\theta\sim\tilde{p}_t}\left[\left\|\theta - \widehat{\theta}_{t+1}\right\|^2_{V_{t-1}} - \left\|\theta - \widehat{\theta}_t\right\|^2_{V_{t-1}}\right] && (\mathcal{E}_{\ell,2}) \\
&\leq C_{3,\ell}^2 + 2C_{3,\ell}(C_{1,\ell}+1) && \text{(Lemma B.5.1)} \\
&\leq 4C_{3,\ell}^2.
\end{aligned}$$

At round $t > 1$, we define $W_t = \int_{\theta\in\Theta^c_{z_*}} \exp\left(-\eta_p \left\|\theta - \widehat{\theta}_t\right\|^2_{V_{t-1}}\right) d\theta$ and W_1 being a uniform

distribution on $\Theta^c_{z_*}$. Then

$$
\begin{aligned}
&\log \frac{W_{t+1}}{W_t} \\
&= \log \frac{\int_{\theta\in\Theta^c_{z_*}} \exp\left(-\eta_p \left\|\theta-\widehat{\theta}_{t+1}\right\|^2_{V_t}\right) d\theta}{\int_{\theta\in\Theta^c_{z_*}} \exp\left(-\eta_p \left\|\theta-\widehat{\theta}_{t}\right\|^2_{V_{t-1}}\right) d\theta} \\
&= \log \frac{\int_{\theta\in\Theta^c_{z_*}} \exp\left(-\eta_p \left\|\theta-\widehat{\theta}_{t+1}\right\|^2_{V_t} - \eta_p \left\|\theta-\widehat{\theta}_{t}\right\|^2_{x_tx_t^\top} + \eta_p \left\|\theta-\widehat{\theta}_{t}\right\|^2_{x_tx_t^\top} + \eta_p \left\|\theta-\widehat{\theta}_{t}\right\|^2_{V_{t-1}} - \eta_p \left\|\theta-\widehat{\theta}_{t}\right\|^2_{V_{t-1}}\right) d\theta}{\int_{\theta\in\Theta^c_{z_*}} \exp\left(-\eta_p \left\|\theta-\widehat{\theta}_{t}\right\|^2_{V_{t-1}}\right) d\theta} \\
&= \log \frac{\int_{\theta\in\Theta^c_{z_*}} \exp\left(-\eta_p \left\|\theta-\widehat{\theta}_{t}\right\|^2_{x_tx_t^\top} - \eta_p \left\|\theta-\widehat{\theta}_{t+1}\right\|^2_{V_t} + \eta_p \left\|\theta-\widehat{\theta}_{t}\right\|^2_{V_t} - \eta_p \left\|\theta-\widehat{\theta}_{t}\right\|^2_{V_{t-1}}\right) d\theta}{\int_{\theta\in\Theta^c_{z_*}} \exp\left(-\eta_p \left\|\theta-\widehat{\theta}_{t}\right\|^2_{V_{t-1}}\right) d\theta} \\
&\le -\eta_p \mathbb{E}_{\theta\sim p_t(\Theta^c_{z_*})}\left[\left\|\theta-\widehat{\theta}_{t}\right\|^2_{x_tx_t^\top} + \left\|\theta-\widehat{\theta}_{t+1}\right\|^2_{V_t} - \left\|\theta-\widehat{\theta}_{t}\right\|^2_{V_t}\right] + \frac{\eta_p^2\cdot 4C_{3,\ell}^2}{8}
\end{aligned}
$$

where the inequality follows from the Hoeffding inequality $\ln \mathbb{E}e^{sX} \le s\mathbb{E}X + \frac{s^2(a-b)^2}{8}$. By telescoping, we have

$$
\begin{aligned}
\log \frac{W_{T_\ell+1}}{W_1} &= \ln \frac{W_{T_\ell+1}}{W_{T_\ell}} + \ln \frac{W_{T_\ell}}{W_{T_\ell-1}} + \cdots + \ln \frac{W_2}{W_1} \\
&\le -\eta_p \sum_{t=1}^{T_\ell} \mathbb{E}_{\theta\sim p_t(\Theta^c_{z_*})}\left[\left\|\theta-\widehat{\theta}_{t}\right\|^2_{x_tx_t^\top} + \left\|\theta-\widehat{\theta}_{t+1}\right\|^2_{V_t} - \left\|\theta-\widehat{\theta}_{t}\right\|^2_{V_t}\right] + \frac{T_\ell\eta_p^2 C_{3,\ell}^2}{2}.
\end{aligned}
$$

On the other hand, let $\tilde{\theta} = \arg\inf_{\theta\in\Theta^c_{z_*}} \left\|\theta-\widehat{\theta}_{T_\ell+1}\right\|^2_{V_{T_\ell}}$. Let $w_t(\theta) = \exp\left(-\eta_p \left\|\theta-\widehat{\theta}_{t}\right\|^2_{V_{t-1}}\right)$.

Let $\mathcal{N}_\gamma := \{(1-\gamma)\tilde{\theta} + \gamma\theta, \theta \in \Theta^c_{z_*}\}$ for $\gamma > 0$ that we choose later. We have

$$
\begin{aligned}
\log \frac{W_{T_\ell+1}}{W_1} &= \log \left(\frac{\int_{\theta \in \Theta^c_{z_*}} \exp\left(-\eta_p \left\|\theta - \widehat{\theta}_{T_\ell+1}\right\|^2_{V_{T_\ell}}\right) d\theta}{\int_{\theta \in \Theta^c_{z_*}} 1 d\theta} \right) \\
&\geq \log \left(\frac{\int_{\theta \in \mathcal{N}_\gamma} \exp\left(-\eta_p \left\|\theta - \widehat{\theta}_{T_\ell+1}\right\|^2_{V_{T_\ell}}\right) d\theta}{\int_{\theta \in \Theta^c_{z_*}} 1 d\theta} \right) \\
&\geq \log \left(\frac{\int_{\theta \in \gamma\Theta^c_{z_*}} \exp\left(-\eta_p \left\|(1-\gamma)\tilde{\theta} + \theta - \widehat{\theta}_{T_\ell+1}\right\|^2_{V_{T_\ell}}\right) d\theta}{\int_{\theta \in \Theta^c_{z_*}} 1 d\theta} \right) \\
&= \log \left(\frac{\int_{\theta \in \Theta^c_{z_*}} \gamma^d \exp\left(-\eta_p \left\|(1-\gamma)\tilde{\theta} + \gamma\theta - \widehat{\theta}_{T_\ell+1}\right\|^2_{V_{T_\ell}}\right) d\theta}{\int_{\theta \in \Theta^c_{z_*}} 1 d\theta} \right) \\
&= \log \left(\frac{\int_{\theta \in \Theta^c_{z_*}} \gamma^d \exp\left(-\eta_p \left\|(1-\gamma)\tilde{\theta} + \gamma\theta - \widehat{\theta}_{T_\ell+1}\right\|^2_{V_{T_\ell}}\right) d\theta}{\int_{\theta \in \Theta^c_{z_*}} 1 d\theta} \right) \\
&\geq \log \left(\frac{\int_{\theta \in \Theta^c_{z_*}} \gamma^d \exp\left(-\eta_p \left((1-\gamma)\left\|\tilde{\theta} - \widehat{\theta}_{T_\ell+1}\right\|^2_{V_{T_\ell}} + \gamma \left\|\theta - \widehat{\theta}_{T_\ell+1}\right\|^2_{V_{T_\ell}}\right)\right) d\theta}{\int_{\theta \in \Theta^c_{z_*}} 1 d\theta} \right) \\
&\geq \log \left(\frac{\int_{\theta \in \Theta^c_{z_*}} \gamma^d \exp\left(-\eta_p \left((1-\gamma)\left\|\tilde{\theta} - \widehat{\theta}_{T_\ell+1}\right\|^2_{V_{T_\ell}} + \gamma \left\|\theta - \widehat{\theta}_{T_\ell+1}\right\|^2_{V_{T_\ell}}\right)\right) d\theta}{\int_{\theta \in \Theta^c_{z_*}} 1 d\theta} \right) \\
&\geq \log \left(\frac{\int_{\theta \in \Theta^c_{z_*}} \gamma^d \exp\left(-\eta_p \left(\left\|\tilde{\theta} - \widehat{\theta}_{T_\ell+1}\right\|^2_{V_{T_\ell}} + \gamma T_\ell C_{1,\ell}\right)\right) d\theta}{\int_{\theta \in \Theta^c_{z_*}} 1 d\theta} \right) \\
&= -\eta_p \inf_{\theta \in \Theta^c_{z_*}} \left\|\theta - \widehat{\theta}_{T_\ell+1}\right\|^2_{V_{T_\ell}} + d \log \gamma - \eta_p \gamma T_\ell C_{1,\ell}.
\end{aligned}
$$

where the last inequality follows from the fact that for any $\theta \in \Theta$,

$$\left\|\theta - \widehat{\theta}_{T_\ell+1}\right\|_{V_{T_\ell}}^2 = \sum_{t=1}^{T_\ell} (x_t^\top(\theta - \widehat{\theta}_{T_\ell+1})^2 \le T_\ell C_{3,\ell}^2$$

under $\mathcal{E}_{2,\ell}$. Combining the two displays gives us

$$\begin{aligned}
&- \eta_p \inf_{\theta \in \Theta_{z_*}^c} \left\|\theta - \widehat{\theta}_{T_\ell+1}\right\|_{V_{T_\ell}}^2 + d\log\gamma - \eta_p \gamma T_\ell C_{1,\ell} \\
&\le -\eta_p \sum_{t=1}^{T_\ell} \mathbb{E}_{\theta \sim p_t(\Theta_{z_*}^c)} \left[\left\|\theta - \widehat{\theta}_t\right\|_{x_t x_t^\top}^2 + \left\|\theta - \widehat{\theta}_{t+1}\right\|_{V_t}^2 - \left\|\theta - \widehat{\theta}_t\right\|_{V_t}^2 \right] + \frac{T_\ell \eta_p^2 C_{3,\ell}^2}{2}.
\end{aligned}$$

Rearranging, we have

$$\begin{aligned}
&\sum_{t=1}^{T_\ell} \mathbb{E}_{\theta \sim p_t(\Theta_{z_*}^c)} \left[\left\|\theta - \widehat{\theta}_t\right\|_{x_t x_t^\top}^2 + \left\|\theta - \widehat{\theta}_{t+1}\right\|_{V_t}^2 - \left\|\theta - \widehat{\theta}_t\right\|_{V_t}^2 \right] - \inf_{\theta \in \Theta_{z_*}^c} \left\|\theta - \widehat{\theta}_{T_\ell+1}\right\|_{V_{T_\ell}}^2 \\
&\le \frac{\eta_p C_{3,\ell}^2 T_\ell}{2} + \frac{d\log(1/\gamma)}{\eta_p} + \gamma T_\ell C_{1,\ell}.
\end{aligned}$$

By choosing $\gamma = \frac{1}{T_\ell C_{1,\ell}}$ and $\eta_p = \sqrt{\frac{d\log(T_\ell C_{1,\ell})}{C_{3,\ell}^2 T_\ell}}$, we have

$$\sum_{t=1}^{T_\ell} \mathbb{E}_{\theta \sim p_t(\Theta_{z_*}^c)} \left[\left\|\theta - \widehat{\theta}_t\right\|_{x_t x_t^\top}^2 + \left\|\theta - \widehat{\theta}_{t+1}\right\|_{V_t}^2 - \left\|\theta - \widehat{\theta}_t\right\|_{V_t}^2 \right] - \inf_{\theta \in \Theta_{z_*}^c} \left\|\theta - \widehat{\theta}_{T_\ell+1}\right\|_{V_{T_\ell}}^2 \le \sqrt{T_\ell C_{3,\ell}^2 d\log(T_\ell C_{1,\ell})},$$

so

$$\frac{1}{T_\ell}\left[\sum_{t=1}^{T_\ell} \mathbb{E}_{\theta \sim p_t(\Theta_{z_*}^c)} \left[\left\|\theta - \widehat{\theta}_t\right\|_{x_t x_t^\top}^2 + \left\|\theta - \widehat{\theta}_{t+1}\right\|_{V_t}^2 - \left\|\theta - \widehat{\theta}_t\right\|_{V_t}^2 \right] - \inf_{\theta \in \Theta_{z_*}^c} \left\|\theta - \widehat{\theta}_{T_\ell+1}\right\|_{V_{T_\ell}}^2\right] \le \sqrt{\frac{C_{3,\ell}^2 d\log(T_\ell C_{1,\ell})}{T_\ell}}.$$

In other words,

$$\begin{aligned}
&\frac{1}{T_\ell}\left[\sum_{t=1}^{T_\ell} \mathbb{E}_{\theta \sim \tilde{p}_t} \left[\left\|\theta - \widehat{\theta}_t\right\|_{x_t x_t^\top}^2 \right] - \inf_{\theta \in \Theta_{z_*}^c} \left\|\theta - \widehat{\theta}_{T_\ell+1}\right\|_{V_{T_\ell}}^2\right] \\
&\le \sqrt{\frac{C_{3,\ell}^2 d\log(T_\ell C_{1,\ell})}{T_\ell}} + \frac{1}{T_\ell}\sum_{t=1}^{T_\ell} \mathbb{E}_{\theta \sim \tilde{p}_t} \left[\left\|\theta - \widehat{\theta}_{t+1}\right\|_{V_t}^2 - \left\|\theta - \widehat{\theta}_t\right\|_{V_t}^2 \right].
\end{aligned}$$

By Lemma B.3.9, we have with probability $1 - 1/\ell^2$,

$$\frac{1}{T_\ell}\sum_{t=1}^{T_\ell} \mathbb{F}_{\theta \sim \tilde{p}_t} \left[\left\|\theta - \widehat{\theta}_{t+1}\right\|_{V_t}^2 - \left\|\theta - \widehat{\theta}_t\right\|_{V_t}^2 \right] \le C_{3,\ell}\sqrt{\frac{2d\beta(T_\ell, \ell^2)}{T_\ell}\log\left(\frac{d + T_\ell L^2}{d}\right)} + C_{3,\ell}\sqrt{\frac{2\log(\ell^2)}{T_\ell}}.$$

Combining the above two displays gives us with probability $1 - 1/\ell^2$,

$$\frac{1}{T_\ell}\left[\sum_{t=1}^{T_\ell}\mathbb{E}_{\theta\sim\tilde{p}_t}\left[\left\|\theta-\widehat{\theta}_t\right\|^2_{x_t x_t^\top}\right] - \inf_{\theta\in\Theta^c_{z_*}}\left\|\theta-\widehat{\theta}_{T_\ell+1}\right\|^2_{V_{T_\ell}}\right]$$
$$\leq \sqrt{\frac{C_{3,\ell}^2 d\log(T_\ell C_{1,\ell})}{T_\ell}} + C_{3,\ell}\sqrt{\frac{2d\beta(T_\ell,\ell^2)}{T_\ell}\log\left(\frac{d+T_\ell L^2}{d}\right)} + C_{3,\ell}\sqrt{\frac{2\log(\ell^2)}{T_\ell}}.$$

□

B.3.5 Approximation Guarantees

In this section, we present several technical lemmas bounding the terms related to the approximation error of $\widehat{\theta}_t$ to θ^* in each iteration t. More specifically, these lemmas show upper bound on the terms in the decomposition in the proof of lemma B.3.4.

Lemma B.3.8 (S4)**.** *Under $\mathcal{E}_{2,\ell}$, with probability $1 - 1/\ell^2$,*

$$\frac{1}{T_\ell}\sum_{t=1}^{T_\ell}\mathbb{F}_{\theta\sim\tilde{p}_t}\left[\left\|\theta-\widehat{\theta}_t\right\|^2_{A(\tilde{\lambda}_t)}\right] - \frac{1}{T_\ell}\sum_{t=1}^{T_\ell}\mathbb{F}_{\theta\sim\tilde{p}_t}\left[\left\|\theta-\widehat{\theta}_t\right\|^2_{x_t x_t^\top}\right] \leq \sqrt{\frac{2C_{1,\ell}\log\ell^2}{T_\ell}}.$$

Proof. Define $M_t = \mathbb{F}_{\theta\sim\tilde{p}_t}\left[\left\|\theta-\widehat{\theta}_t\right\|^2_{x_t x_t^\top}\right]$. Note that

$$\mathbb{E}_{x_t}[M_t|\mathcal{F}_{t-1}] = \mathbb{F}_{\theta\sim\tilde{p}_t}\left[\left\|\theta-\widehat{\theta}_t\right\|^2_{A(\tilde{\lambda}_t)}\right],$$

so $\tilde{M}_t = M_t - \mathbb{F}_{\theta\sim\tilde{p}_t}\left[\left\|\theta-\widehat{\theta}_t\right\|^2_{A(\tilde{\lambda}_t)}\right]$ is a mean-zero martingale. Also, under $\mathcal{E}_{2,\ell}$, $|M_t| \leq C_{1,\ell}$, then by Azuma-Hoeffding, we have with probability at least $1-\frac{1}{\ell^2}$, $\sum_{t=1}^{T_\ell}\tilde{M}_t \leq \sqrt{2C_{1,\ell}T_\ell\log\ell^2}$, so

$$\frac{1}{T_\ell}\sum_{t=1}^{T_\ell}\mathbb{F}_{\theta\sim\tilde{p}_t}\left[\left\|\theta-\widehat{\theta}_t\right\|^2_{A(\tilde{\lambda}_t)}\right] - \frac{1}{T_\ell}\sum_{t=1}^{T_\ell}\mathbb{F}_{\theta\sim\tilde{p}_t}\left[\left\|\theta-\widehat{\theta}_t\right\|^2_{x_t x_t^\top}\right] \leq \sqrt{\frac{2C_{1,\ell}\log\ell^2}{T_\ell}}.$$

□

Lemma B.3.9 (C_{T_ℓ})**.** *Under $\mathcal{E}_{1,\ell}\cap\mathcal{E}_{2,\ell}$, with probability $1 - 1/\ell^2$,*

$$\frac{1}{T_\ell}\sum_{t=1}^{T_\ell}\mathbb{F}_{\theta\sim\tilde{p}_t}\left[\left\|\theta-\widehat{\theta}_{t+1}\right\|^2_{V_t} - \left\|\theta-\widehat{\theta}_t\right\|^2_{V_t}\right] \leq C_{3,\ell}\sqrt{\frac{2d\beta(T_\ell,\ell^2)}{T_\ell}\log\left(\frac{d+T_\ell L^2}{d}\right)} + C_{3,\ell}\sqrt{\frac{2\log(\ell^2)}{T_\ell}}.$$

Proof. We first consider some round t and some θ. By Lemma B.5.1,

$$\left\|\theta - \widehat{\theta}_t\right\|_{V_{t-1}}^2 - \left\|\theta - \widehat{\theta}_{t+1}\right\|_{V_{t-1}}^2 \leq 2C_{3,\ell}(y_t - x_t^\top \widehat{\theta}_t).$$

Therefore,

$$\frac{1}{T_\ell}\sum_{t=1}^{T_\ell} \mathbb{F}_{\theta\sim\tilde{p}_t}\left[\left\|\theta - \widehat{\theta}_t\right\|_{V_{t-1}}^2 - \left\|\theta - \widehat{\theta}_{t+1}\right\|_{V_{t-1}}^2\right] \leq \frac{2C_{3,\ell}}{T_\ell}\sum_{t=1}^{T_\ell}(y_t - x_t^\top \widehat{\theta}_t). \tag{B.3}$$

Now, note that

$$\begin{aligned} y_t - x_t^\top \widehat{\theta}_t &= x_t^\top(\theta^* - \widehat{\theta}_t) + \epsilon_t \\ &\leq \|x_t\|_{V_{t-1}^{-1}} \left\|\theta^* - \widehat{\theta}_t\right\|_{V_{t-1}} + \epsilon_t \\ &\leq \|x_t\|_{V_{t-1}^{-1}} \sqrt{\beta(t,\ell^2)} + \epsilon_t. && \text{(by } \mathcal{E}_{1,\ell}\text{)} \end{aligned}$$

Note that since $\epsilon_t \sim N(0,1)$ is 1-subGaussian, by Azuma-Hoeffding, we have with probability $1 - 1/\ell^2$,

$$\sum_{t=1}^{T_\ell} \epsilon_t \leq \sqrt{2T_\ell \log(\ell^2)}.$$

By summing it from 1 to T_ℓ, we have under $\mathcal{E}_{1,\ell}$, with probability $1 - 1/\ell^2$,

$$\begin{aligned} \sum_{t=1}^{T_\ell}(y_t - x_t^\top \widehat{\theta}_t) &\leq \sum_{t=1}^{T_\ell} \sqrt{\beta(t,\ell^2)}\,\|x_t\|_{V_{t-1}^{-1}} + \sum_{t=1}^{T_\ell}\epsilon_t \\ &\leq \sum_{t=1}^{T_\ell} \sqrt{\beta(t,\ell^2)}\,\|x_t\|_{V_{t-1}^{-1}} + \sqrt{2T_\ell \log(\ell^2)} \\ &\leq \sqrt{T_\ell \sum_{t=1}^{T_\ell} \beta(t,\ell^2)\,\|x_t\|_{V_{t-1}^{-1}}^2} + \sqrt{2T_\ell \log(\ell^2)} && \text{(by Cauchy-Schwarz)} \\ &\leq \sqrt{T_\ell \beta(T_\ell,\ell^2) \sum_{t=1}^{T_\ell} \|x_t\|_{V_{t-1}^{-1}}^2} + \sqrt{2T_\ell \log(\ell^2)} && \text{(by Cauchy-Schwarz)} \\ &\leq \sqrt{T_\ell \beta(T_\ell,\ell^2) 2d \log\left(\frac{d + T_\ell L^2}{d}\right)} + \sqrt{2T_\ell \log(\ell^2)}. \end{aligned}$$

(by Elliptical potential lemma [Abbasi-Yadkori et al., 2011])

Plugging this in Equation B.3 gives the result. □

Lemma B.3.10 (C'_{T_ℓ})**.** *Under $\mathcal{E}_{1,\ell} \cap \mathcal{E}_{2,\ell}$, we have*

$$\max_{\lambda \in \triangle_{\mathcal{X}}} \frac{1}{T_\ell} \sum_{t=1}^{T_\ell} \mathbb{F}_{\theta \sim p_t} \left[\|\theta^* - \theta\|_{A(\lambda)}^2 \right] - \max_{\lambda \in \triangle_{\mathcal{X}}} \frac{1}{T_\ell} \sum_{t=1}^{T_\ell} \mathbb{F}_{\theta \sim p_t} \left[\left\| \widehat{\theta}_t - \theta \right\|_{A(\lambda)}^2 \right]$$
$$\leq \frac{T_2(\ell) L^2 \beta(T_2(\ell), \ell^2)}{T_\ell} + 4d\beta(T_\ell, \ell^2) T_\ell^{-3/4}.$$

Proof. We have

$$\max_{\lambda \in \triangle_{\mathcal{X}}} \frac{1}{T_\ell} \sum_{t=1}^{T_\ell} \mathbb{F}_{\theta \sim p_t} \left[\|\theta^* - \theta\|_{A(\lambda)}^2 \right] - \max_{\lambda \in \triangle_{\mathcal{X}}} \frac{1}{T_\ell} \sum_{t=1}^{T_\ell} \mathbb{F}_{\theta \sim p_t} \left[\left\| \widehat{\theta}_t - \theta \right\|_{A(\lambda)}^2 \right]$$
$$\leq \max_{\lambda \in \triangle_{\mathcal{X}}} \frac{1}{T_\ell} \sum_{t=1}^{T_\ell} \mathbb{F}_{\theta \sim p_t} \left[\|\theta^* - \theta\|_{A(\lambda)}^2 - \left\| \widehat{\theta}_t - \theta \right\|_{A(\lambda)}^2 \right].$$

We fix some θ and λ. Note that

$$\begin{aligned}
&\|\theta^* - \theta\|_{A(\lambda)}^2 - \left\| \widehat{\theta}_t - \theta \right\|_{A(\lambda)}^2 \\
&= (\theta^* + \widehat{\theta}_t - 2\theta)^\top A(\lambda)(\theta^* - \widehat{\theta}_t) \\
&= \sum_{x \in \mathcal{X}} \lambda_x (\theta^* + \widehat{\theta}_t - 2\theta)^\top x x^\top (\theta^* - \widehat{\theta}_t) \\
&\leq \max_{x \in \mathcal{X}} (\theta^* + \widehat{\theta}_t - 2\theta)^\top x x^\top (\theta^* - \widehat{\theta}_t) \\
&\leq (C_{3,\ell} + \Delta_{\max}) \max_{x \in \mathcal{X}} x^\top (\theta^* - \widehat{\theta}_t).
\end{aligned}$$

Therefore,

$$\max_{\lambda \in \triangle_{\mathcal{X}}} \frac{1}{T_\ell} \sum_{t=1}^{T_\ell} \mathbb{F}_{\theta \sim p_t} \left[\|\theta^* - \theta\|_{A(\lambda)}^2 - \left\| \widehat{\theta}_t - \theta \right\|_{A(\lambda)}^2 \right] \leq (C_{3,\ell} + \Delta_{\max}) \max_{x \in \mathcal{X}} \frac{1}{T_\ell} \sum_{t=1}^{T_\ell} \left\langle \widehat{\theta}_t - \theta^*, x \right\rangle. \tag{B.4}$$

By Lemma B.3.15, under $\mathcal{E}_{3,\ell} \cap \mathcal{E}_{1,\ell}$, for any $t \geq T_2(\ell) + 1$, we have for any $x \in \mathcal{X}$,

$$\left\langle x, \widehat{\theta}_t - \theta^* \right\rangle \leq \frac{d}{t^{3/4}} \beta(t, \ell^2).$$

Also, by Lemma B.4.2, under $\mathcal{E}_{1,\ell}$, we have for any $t \geq 1$,

$$\left\langle x, \widehat{\theta}_t - \theta^* \right\rangle \leq L^2 \beta(t, \ell^2).$$

Therefore,

$$
\begin{aligned}
&\max_{x\in\mathcal{X}} \frac{1}{T_\ell}\sum_{t=1}^{T_\ell}\left\langle \widehat{\theta}_t - \theta^*, x\right\rangle \\
&\le \max_{x\in\mathcal{X}} \frac{1}{T_\ell}\left[\sum_{t=1}^{T_2(\ell)}\left\langle \widehat{\theta}_t - \theta^*, x\right\rangle + \sum_{t=T_2(\ell)+1}^{T_\ell}\left\langle \widehat{\theta}_t - \theta^*, x\right\rangle\right] \\
&\le \frac{1}{T_\ell}\left[T_2(\ell)L^2\beta(T_2(\ell),\ell^2) + \sum_{t=T_2(\ell)+1}^{T_\ell}\frac{d}{t^{3/4}}\beta(t,\ell^2)\right] \qquad \text{(by Lemma B.4.2 and B.3.15)} \\
&\le \frac{1}{T_\ell}\left[T_2(\ell)L^2\beta(T_2(\ell),\ell^2) + d\beta(T_\ell,\ell^2)\int_{t=T_2(\ell)}^{T_\ell} t^{-3/4}dt\right] \\
&= \frac{1}{T_\ell}\left[T_2(\ell)L^2\beta(T_2(\ell),\ell^2) + d\beta(T_\ell,\ell^2)(4T_\ell^{1/4} - 4T_2(\ell)^{1/4})\right] \\
&\le \frac{T_2(\ell)L^2\beta(T_2(\ell),\ell^2)}{T_\ell} + 4d\beta(T_\ell,\ell^2){T_\ell}^{-3/4}.
\end{aligned}
$$

Plugging this in Equation B.4 gives us

$$
\max_{\lambda\in\triangle_{\mathcal{X}}} \frac{1}{T_\ell}\sum_{t=1}^{T_\ell}\mathbb{F}_{\theta\sim p_t}\left[\|\theta^*-\theta\|_{A(\lambda)}^2 - \left\|\widehat{\theta}_t-\theta\right\|_{A(\lambda)}^2\right] \le \frac{T_2(\ell)L^2\beta(T_2(\ell),\ell^2)}{T_\ell} + 4d\beta(T_\ell,\ell^2)T_\ell^{-3/4}.
$$

□

Lemma B.3.11 (C''_{T_ℓ})**.** *Assume that Θ is closed. Then, we have under $\mathcal{E}_{1,\ell}\cap\mathcal{E}_{2,\ell}$,*

$$
\left|\frac{1}{T_\ell}\inf_{\theta\in\Theta_{z_*}^c}\left\|\theta-\widehat{\theta}_{T_\ell+1}\right\|_{V_{T_\ell}}^2 - \frac{1}{T_\ell}\inf_{\theta\in\Theta_{z_*}^c}\|\theta^*-\theta\|_{V_{T_\ell}}^2\right| \le (C_{3,\ell}+\Delta_{\max})\sqrt{\frac{\beta(T_\ell,\ell^2)}{T_\ell}}.
$$

Proof. Let $\theta_1 := \arg\inf_{\theta\in\Theta_{z_*}^c}\left\|\theta-\widehat{\theta}_{T_\ell+1}\right\|_{V_{T_\ell}}^2$ and $\theta_2 := \arg\inf_{\theta\in\Theta_{z_*}^c}\|\theta-\theta^*\|_{V_{T_\ell}}^2$. We have

$$
\begin{aligned}
&\inf_{\theta\in\Theta_{z_*}^c}\left\|\theta-\widehat{\theta}_{T_\ell+1}\right\|_{V_{T_\ell}}^2 - \inf_{\theta\in\Theta_{z_*}^c}\|\theta^*-\theta\|_{V_{T_\ell}}^2 \\
&\le \left\|\widehat{\theta}_{T_\ell+1}-\theta_2\right\|_{V_{T_\ell}}^2 - \|\theta^*-\theta_2\|_{V_{T_\ell}}^2 \\
&= \left(\left\|\widehat{\theta}_{T_\ell+1}-\theta_2\right\|_{V_{T_\ell}} - \|\theta^*-\theta_2\|_{V_{T_\ell}}\right)\left(\left\|\widehat{\theta}_{T_\ell+1}-\theta_2\right\|_{V_{T_\ell}} + \|\theta^*-\theta_2\|_{V_{T_\ell}}\right). \\
&\le \left\|\widehat{\theta}_{T_\ell+1}-\theta_*\right\|_{V_{T_\ell}}\left(\left\|\widehat{\theta}_{T_\ell+1}-\theta_2\right\|_{V_{T_\ell}} + \|\theta^*-\theta_2\|_{V_{T_\ell}}\right).
\end{aligned}
$$

Note that under $\mathcal{E}_{2,\ell}$,

$$\left\|\widehat{\theta}_{T_\ell+1}-\theta_1\right\|_{V_{T_\ell}}=\sqrt{\sum_{t=1}^{T_\ell}(x_t^\top(\widehat{\theta}_{T_\ell+1}-\theta_1))^2}\le C_{3,\ell}\sqrt{T_\ell};$$
$$\left\|\theta^*-\theta_2\right\|_{V_{T_\ell}}=\sqrt{\sum_{t=1}^{T_\ell}(x_t^\top(\theta^*-\theta_2))^2}\le \Delta_{\max}\sqrt{T_\ell}.$$

Therefore,

$$\begin{aligned}
&\inf_{\theta\in\Theta^c_{z_*}}\left\|\theta-\widehat{\theta}_{T_\ell+1}\right\|^2_{V_{T_\ell}}-\inf_{\theta\in\Theta^c_{z_*}}\|\theta^*-\theta\|^2_{V_{T_\ell}}\\
&\le (C_{3,\ell}+\Delta_{\max})\sqrt{T_\ell}\left\|\widehat{\theta}_{T_\ell+1}-\theta^*\right\|_{V_{T_\ell}}\\
&\le (C_{3,\ell}+\Delta_{\max})\sqrt{T_\ell\beta(T_\ell,\ell^2)}. && \text{(by } \mathcal{E}_{1,\ell})
\end{aligned}$$

□

We use the above lemma to bound the term that relates $\tilde{p}_t$ to p_t.

Lemma B.3.12 ($\tilde{p}_t$ to p_t)**.** *Under $\mathcal{E}_{2,\ell}\cap\mathcal{E}_{4,\ell}$ for $T_\ell\ge T_0$,*

$$\frac{1}{T_\ell}\sum_{t=1}^{T_\ell}\mathbb{F}_{\theta\sim p_t}\left[\left\|\theta-\widehat{\theta}_t\right\|^2_{A(\tilde{\lambda}_t)}\right]-\frac{1}{T_\ell}\sum_{t=1}^{T_\ell}\mathbb{F}_{\theta\sim\tilde{p}_t}\left[\left\|\theta-\widehat{\theta}_t\right\|^2_{A(\tilde{\lambda}_t)}\right]\le\frac{2C^2_{3,\ell}T_0(\ell)}{T_\ell}.$$

Proof. Note that $\tilde{p}_t=p_t$ under $\mathcal{E}_{4,\ell}$,

$$\begin{aligned}
&\frac{1}{T_\ell}\sum_{t=1}^{T_\ell}\left(\mathbb{F}_{\theta\sim p_t}\left[\left\|\theta-\widehat{\theta}_t\right\|^2_{A(\tilde{\lambda}_t)}\right]-\mathbb{F}_{\theta\sim\tilde{p}_t}\left[\left\|\theta-\widehat{\theta}_t\right\|^2_{A(\tilde{\lambda}_t)}\right]\right)\\
&=\frac{1}{T_\ell}\sum_{t=1}^{T_0(\ell)}\left(\mathbb{F}_{\theta\sim p_t}\left[\left\|\theta-\widehat{\theta}_t\right\|^2_{A(\tilde{\lambda}_t)}\right]-\mathbb{F}_{\theta\sim\tilde{p}_t}\left[\left\|\theta-\widehat{\theta}_t\right\|^2_{A(\tilde{\lambda}_t)}\right]\right)\\
&\quad+\frac{1}{T_\ell}\sum_{t=T_0(\ell)+1}^{T_\ell}\left(\mathbb{F}_{\theta\sim p_t}\left[\left\|\theta-\widehat{\theta}_t\right\|^2_{A(\tilde{\lambda}_t)}\right]-\mathbb{F}_{\theta\sim\tilde{p}_t}\left[\left\|\theta-\widehat{\theta}_t\right\|^2_{A(\tilde{\lambda}_t)}\right]\right)\\
&=\frac{1}{T_\ell}\sum_{t=1}^{T_0(\ell)}\left(\mathbb{F}_{\theta\sim p_t}\left[\left\|\theta-\widehat{\theta}_t\right\|^2_{A(\tilde{\lambda}_t)}\right]-\mathbb{F}_{\theta\sim\tilde{p}_t}\left[\left\|\theta-\widehat{\theta}_t\right\|^2_{A(\tilde{\lambda}_t)}\right]\right).
\end{aligned}$$

Since for any $\theta \in \Theta$, under $\mathcal{E}_{2,\ell}$,

$$\left\|\theta - \widehat{\theta}_t\right\|^2_{A(\tilde{\lambda}_t)} = \sum_{x\in\mathcal{X}} \tilde{\lambda}_{t,x} \left\|\theta - \widehat{\theta}_t\right\|^2_{xx^\top} \leq \max_{x\in\mathcal{X}} \left\|\theta - \widehat{\theta}_t\right\|^2_{xx^\top} \leq C^2_{3,\ell},$$

we have

$$\frac{1}{T_\ell} \sum_{t=1}^{T_0(\ell)} \left(\mathbb{F}_{\theta\sim p_t} \left[\left\|\theta - \widehat{\theta}_t\right\|^2_{A(\tilde{\lambda}_t)} \right] - \mathbb{F}_{\theta\sim \tilde{p}_t} \left[\left\|\theta - \widehat{\theta}_t\right\|^2_{A(\tilde{\lambda}_t)} \right] \right) \leq \frac{2C^2_{3,\ell} T_0(\ell)}{T_\ell}.$$

□

B.3.6 Guarantees on sampling and learning the estimate

In this section we provide some general guarantees on sampling together with a threshold after which each arm gets enough samples and . Consider a setting where at each time we receive a distribution $\tilde{\lambda} = (1-\gamma_t)\lambda_t + \gamma_t P$ for a fixed distribution P.

Lemma B.3.13. *Fix a distribution P on $\mathcal{X}$ with full support. On an event that is true with probability greater than $1-\delta$, for any $0 < \alpha < 1/2$ there exists a $T_1 := T_1(\alpha, \delta, T)$ such that for any $t \geq T_1$,*

$$V_t \geq \frac{c}{1-\alpha} A(P) t^{1-\alpha}.$$

Proof. Fix $x \in \mathcal{X}$, let $N_{t,x} = \sum_{s=1}^t Z_s$ where $Z_s = 1$ if $x_s = x$ else 0. Then, $V_t = \sum_{x\in\mathcal{X}} \sum_{s=1}^t Z_s x x^\top$. We assume that $\gamma_s = 1/s^\alpha, s \geq 1$.

Note that $\mathbb{P}(Z_s = 1|\mathcal{F}_{s-1}) = (1-\gamma_s)\lambda_{s,x} + \gamma_s P_x$. So for $t > 1$,

$$\begin{aligned}
\mathbb{P}\left(\sum_{s=1}^{t} Z_s \le cP_x \sum_{s=1}^{t}\gamma_s\right) &= \mathbb{P}\left(\sum_{s=1}^{t} Z_s - (1-\gamma_s)\lambda_{s,x} - \gamma_s P_x \le \sum_{s=1}^{t} cP_x\gamma_s - (1-\gamma_s)\lambda_{s,x} - \gamma_s P_x\right) \\
&= \mathbb{P}\left(\sum_{s=1}^{t} Z_s - (1-\gamma_s)\lambda_{s,x} - \gamma_s P_x \le \sum_{s=1}^{t} (c-1)P_x\gamma_s - (1-\gamma_s)\lambda_{s,x}\right) \\
&\le \mathbb{P}\left(\sum_{s=1}^{t} Z_s - (1-\gamma_s)\lambda_{s,x} - \gamma_s P_x \le \sum_{s=1}^{t} (c-1)P_x\gamma_s\right) \\
&\le \mathbb{P}\left(\sum_{s=1}^{t} Z_s - (1-\gamma_s)\lambda_{s,x} - \gamma_s P_x \le -\sum_{s=1}^{t} (1-c)P_x\gamma_s\right) \\
&\le \exp\left(-\frac{1}{t}\left(\sum_{s=1}^{t}(1-c)P_x\gamma_s\right)^2\right) \qquad \text{(Azuma-Hoeffding)} \\
&= \exp\left(-\left(\frac{(1-c)P_x}{\sqrt{t}}\sum_{s=1}^{t}\gamma_s\right)^2\right) \\
&\le \exp\left(-\left(\frac{(1-c)P_x}{\sqrt{t}}\frac{t^{1-\alpha}-1}{1-\alpha}\right)^2\right) \qquad \left(\textstyle\sum_{s=1}^{t}\frac{1}{s^\alpha} \ge \frac{t^{1-\alpha}-1}{1-\alpha}\right) \\
&\le \exp\left(-\left((1-c)P_x\frac{t^{1/2-\alpha}-t^{-1/2}}{1-\alpha}\right)^2\right) \\
&\le \exp\left(-\left(\frac{(1-c)P_x}{2(1-\alpha)}t^{1/2-\alpha}\right)^2\right) \quad \left(t^{1/2-\alpha}-t^{-1/2} > \tfrac{1}{2}t^{1/2-\alpha},\, t \ge 2\right) \\
&\le \exp\left(-\left(\frac{(1-c)P_x}{2(1-\alpha)}\right)^2 t^{1-2\alpha}\right)
\end{aligned}$$

This implies that with the sequence $\gamma_s = 1/s^\alpha, \alpha < 1/2$ (to ensure $1-2\alpha > 0$), with probability greater than $1-\delta$ we have

$$N_{t,x} = \sum_{s=1}^{t} Z_s \ge cP_x\sum_{s=1}^{t}\gamma_s \ge \frac{cP_x}{1-\alpha}(t^{1-\alpha}-1) \quad \text{whenever} \quad t \ge \left(\frac{2(1-\alpha)\sqrt{\log(1/\delta)}}{(1-c)P_x}\right)^{\frac{2}{1-2\alpha}}.$$

□

The lemma below states that there exists some time T_2 such that all the arms get enough samples.

Lemma B.3.14. *For* $T_2(\ell) = \max_{x\in\mathcal{X}} \left(\frac{6\sqrt{\log(|\mathcal{X}|T_\ell \ell^2)}}{\lambda_x^G}\right)^4$, *we have*

$$\mathbb{P}\left(\mathcal{E}_{3,\ell}\right) \geq 1 - 1/\ell^2.$$

Proof. By Lemma B.3.13 with a choice of $c = 1-\alpha$, $\alpha = \frac{1}{4}$, $\delta = \frac{1}{|\mathcal{X}|T_\ell \ell^2}$, and $P = \lambda^G$, we have for any $t \geq \left(\frac{2(1-\alpha)\sqrt{\log(1/\delta)}}{(1-c)P_x}\right)^{\frac{2}{1-2\alpha}} = \left(\frac{6\sqrt{\log(|\mathcal{X}|T_\ell\ell^2)}}{\lambda_x^G}\right)^4$, we have $\mathbb{P}(V_t \geq t^{3/4}A(\lambda^G)) \geq 1 - \frac{1}{|\mathcal{X}|T_\ell\ell^2}$. Let $T_2(\ell) := \max_{x\in\mathcal{X}} \left(\frac{6\sqrt{\log(|\mathcal{X}|T_\ell \ell^2)}}{\lambda_x^G}\right)^4$, union bounding for $t \in [T_2, T_\ell]$ and $x \in \mathcal{X}$ gives the result. □

Lemma B.3.15. *Under* $\mathcal{E}_{3,\ell} \cap \mathcal{E}_{1,\ell}$, *for any* $t \geq T_2(\ell) + 1$, *we have for any* $x \in \mathcal{X}$,

$$\left\langle x, \widehat{\theta}_t - \theta^* \right\rangle \leq \frac{d}{t^{3/4}}\beta(t, \ell^2).$$

Proof. Let $N_{t,x}$ be the number of times arm x gets pulled at round t. By Lemma B.3.14, for $t \geq T_2(\ell) + 1$, under $\mathcal{E}_{3,\ell}$, we have

$$V_{t-1} = \sum_{x\in\mathcal{X}} N_{t-1,x} x x^\top \geq t^{3/4} A(\lambda^G).$$

Therefore, for any $x \in \mathcal{X}$,

$$\|x\|^2_{V_{t-1}^{-1}} \leq \frac{1}{t^{3/4}} \|x\|^2_{A(\lambda^G)^{-1}} \leq \frac{d}{t^{3/4}}$$

by Kiefer-Wolfowitz. Therefore, under $\mathcal{E}_{1,\ell}$, for any $x \in \mathcal{X}$,

$$\begin{aligned}
\left\langle x, \widehat{\theta}_t - \theta^* \right\rangle &\leq \|x\|^2_{V_{t-1}^{-1}} \left\|\widehat{\theta}_t - \theta^*\right\|^2_{V_{t-1}} \\
&\leq \frac{d}{t^{3/4}} \left\|\widehat{\theta}_t - \theta^*\right\|^2_{V_{t-1}} \\
&\leq \frac{d}{t^{3/4}}\beta(t, \ell^2).
\end{aligned}$$

□

The following lemma provides a guarantee that we eventually finds z_*.

Lemma B.3.16. *For* $T_0(\ell) = \max\left\{\left(\frac{d\beta(T_\ell,\ell^2)\max_{z\in\mathcal{Z}}\|z\|_1}{\Delta_{\min}}\right)^{4/3}, T_2(\ell)+1\right\}$, *we have* $\mathbb{P}(\mathcal{E}_{4,\ell}|\mathcal{E}_{1,\ell}\cap\mathcal{E}_{3,\ell}) \geq 1-1/\ell^2$.

Proof. By Lemma B.3.15, we know that for any $t \geq T_2(\ell)+1$, under $\mathcal{E}_{1,\ell}\cap\mathcal{E}_{3,\ell}$ we have for any $x \in \mathcal{X}$,

$$\left\langle x, \widehat{\theta}_t - \theta^* \right\rangle \leq \frac{d}{t^{3/4}}\beta(t,\ell^2).$$

Since the span of $\mathcal{Z}$ is in the subset of $\mathcal{X}$, for any $z \in \mathcal{Z}$, we write $z_* - z = \sum_{x\in\mathcal{X}} \alpha_{z,x} x$. Then

$$\begin{aligned}(z_*-z)^\top(\theta_*-\widehat{\theta}_t) &= \sum_{x\in\mathcal{X}} \alpha_{z,x} x^\top(\theta_*-\widehat{\theta}_t)\\ &\leq \sum_{x\in\mathcal{X}} \alpha_{z,x}\frac{d}{t^{3/4}}\beta(t,\ell^2)\\ &\leq \max_{z\in\mathcal{Z}}\|z\|_1\frac{d}{t^{3/4}}\beta(t,\ell^2).\end{aligned}$$

Then, for any $t > \left(\frac{d\beta(t,\ell^2)\max_{z\in\mathcal{Z}}\|z\|_1}{\Delta_{\min}}\right)^{4/3}$, we have

$$\max_{z\in\mathcal{Z}}\|z\|_1\frac{d}{t^{3/4}}\beta(t,\ell^2) < \Delta_{\min},$$

which implies that for any z,

$$\begin{aligned}&(z_*-z)^\top(\theta_*-\widehat{\theta}_t) < \Delta_{\min}\\ \Rightarrow &(z_*-z)^\top(\widehat{\theta}_t-\theta_*) > -\Delta_{\min}\\ \Rightarrow &(z_*-z)^\top\widehat{\theta}_t > 0,\end{aligned}$$

which implies that $\widehat{z}_t = z_*$. □

B.4 Bounds and Events that Hold True Each Round

The following lemma states an anytime confidence bound for the least-squares estimator. It is a restatement of Theorem 20.5 of Lattimore and Szepesvári [2020] in our setting.

Lemma B.4.1 ($\mathcal{E}_{1,\ell}$)**.** *With probability* $1-1/\ell^2$, *for all* t, *we have*

$$\left\|\widehat{\theta}_t-\theta^*\right\|_{V_{t-1}}^2 \leq B + \sqrt{2\log(\ell^2)+d\log\left(\frac{d+tL^2}{d}\right)}.$$

Proof. Follows from Theorem 20.5 of Lattimore and Szepesvári [2020]. □

Lemma B.4.2 ($\mathcal{E}_{2,\ell}$)**.** *Under $\mathcal{E}_{1,\ell}$, we have for any $x \in \mathcal{X}$ and any $t \in [1, T_\ell]$, $\left\langle x, \widehat{\theta}_t \right\rangle \leq \Delta_{\max} + L^2\beta(T_\ell, \ell^2)$.*

Proof. For any $x \in \mathcal{X}$,

$$\begin{aligned}\left\langle x, \widehat{\theta}_t \right\rangle &= \langle x, \theta^* \rangle + \left\langle x, \widehat{\theta}_t - \theta^* \right\rangle \\ &\leq \Delta_{\max} + \|x\|^2_{V_{t-1}^{-1}} \left\|\widehat{\theta}_t - \theta^*\right\|^2_{V_{t-1}} \\ &\leq \Delta_{\max} + \|x\|^2_{V_{t-1}^{-1}} \beta(t, \ell^2). \qquad \text{(under } \mathcal{E}_{1,\ell}\text{)}\end{aligned}$$

Since we have

$$V_{t-1} = V_0 + \sum_{s=1}^{t-1} x_s x_s^\top,$$

for $V_0 = I$, we have the minimum eigenvalue $\sigma_{\min}(V_{t-1}) \geq \sigma_{\min}(V_0) + \sigma_{\min}\left(\sum_{s=1}^{t-1} x_s x_s^\top\right) \geq 1$, so

$$\sigma_{\max}(V_{t-1}^{-1}) = \frac{1}{\sigma_{\min}(V_{t-1})} \leq 1,$$

which implies that

$$\max_{x \in \mathcal{X}} \|x\|^2_{V_{t-1}^{-1}} \leq \sigma_{\max}(V_{t-1}^{-1}) \max_{x \in \mathcal{X}} \|x\|_2^2 \leq L^2.$$

Therefore,

$$\left\langle x, \widehat{\theta}_t \right\rangle \leq \Delta_{\max} + L^2\beta(t, \ell^2) \leq \Delta_{\max} + L^2\beta(T_\ell, \ell^2).$$

□

B.5 Technical Lemmas

Lemma B.5.1 (Recursive Least Squares Guarantee)**.** *In any round ℓ, conditional on event $\mathcal{E}_{1,\ell} \cap \mathcal{E}_{2,\ell}$, for any $\theta \in \Theta$ and any $t \in [1, T_\ell]$ we have*

$$\left\|\theta - \widehat{\theta}_{t+1}\right\|^2_{V_t} - \left\|\theta - \widehat{\theta}_t\right\|^2_{V_t} \leq 2C_{3,\ell}(y_t - x_t^\top \widehat{\theta}_t) \leq 2C_{3,\ell}(C_{1,\ell} + 1),$$

assuming that all rewards are bounded in $[-1, 1]$.

Proof. We first consider some round t and some θ. Note that $\widehat{\theta}_t = V_t^{-1} X_t^\top Y_t$. Then

$$\begin{aligned}
\widehat{\theta}_{t+1} &= (V_{t-1} + x_t x_t^\top)^{-1}(X_{t-1}^\top Y_{t-1} + x_t y_t) \\
&= \left(V_{t-1}^{-1} - \frac{V_{t-1}^{-1} x_t x_t^\top V_{t-1}^{-1}}{1 + x_t^\top V_{t-1}^{-1} x_t}\right)(X_{t-1}^\top Y_{t-1} + x_t y_t) \\
&= \widehat{\theta}_t - \frac{V_{t-1}^{-1} x_t x_t^\top \widehat{\theta}_t}{1 + x_t^\top V_{t-1}^{-1} x_t} + V_{t-1}^{-1} x_t y_t - \frac{V_{t-1}^{-1} x_t x_t^\top V_{t-1}^{-1} x_t y_t}{1 + x_t^\top V_{t-1}^{-1} x_t} \\
&= \widehat{\theta}_t - \frac{V_{t-1}^{-1} x_t x_t^\top \widehat{\theta}_t}{1 + x_t^\top V_{t-1}^{-1} x_t} + \frac{V_{t-1}^{-1} x_t y_t (1 + x_t^\top V_{t-1}^{-1} x_t) - x_t^\top V_{t-1}^{-1} x_t V_{t-1}^{-1} x_t y_t}{(1 + x_t^\top V_{t-1}^{-1} x_t)} \\
&= \widehat{\theta}_t - \frac{V_{t-1}^{-1} x_t x_t^\top \widehat{\theta}_t}{1 + x_t^\top V_{t-1}^{-1} x_t} + \frac{V_{t-1}^{-1} x_t y_t}{(1 + x_t^\top V_{t-1}^{-1} x_t)} \\
&= \widehat{\theta}_t + \frac{V_{t-1}^{-1} x_t (y_t - x_t^\top \widehat{\theta}_t)}{1 + x_t^\top V_{t-1}^{-1} x_t}
\end{aligned}$$

Hence

$$\widehat{\theta}_{t+1} - \widehat{\theta}_t = \frac{V_{t-1}^{-1} x_t}{1 + x_t^\top V_{t-1}^{-1} x_t}(y_t - x_t^\top \widehat{\theta}_t)$$

and

$$\begin{aligned}
V_t(\widehat{\theta}_{t+1} - \widehat{\theta}_t) &= \frac{V_t V_{t-1}^{-1} x_t}{1 + x_t^\top V_{t-1}^{-1} x_t}(y_t - x_t^\top \widehat{\theta}_t) \\
&= \frac{(I + x_t x_t^\top V_{t-1}^{-1}) x_t}{1 + x_t^\top V_{t-1}^{-1} x_t}(y_t - x_t^\top \widehat{\theta}_t) \\
&= \frac{x_t (1 + x_t^\top V_{t-1}^{-1} x_t)}{1 + x_t^\top V_{t-1}^{-1} x_t}(y_t - x_t^\top \widehat{\theta}_t) \\
&= (y_t - x_t^\top \widehat{\theta}_t) x_t
\end{aligned}$$

Then

$$\begin{aligned}
&\left\|\theta - \widehat{\theta}_{t+1}\right\|_{V_t}^2 - \left\|\theta - \widehat{\theta}_t\right\|_{V_t}^2 \\
&= (\widehat{\theta}_{t+1} - \widehat{\theta}_t)^\top V_t(\widehat{\theta}_{t+1} + \widehat{\theta}_t - 2\theta) \\
&= (y_t - x_t^\top \widehat{\theta}_t) x_t^\top (\widehat{\theta}_{t+1} + \widehat{\theta}_t - 2\theta) \\
&\leq 2C_{3,\ell}(y_t - x_t^\top \widehat{\theta}_t) \\
&\leq 2C_{3,\ell}(C_{1,\ell} + 1)
\end{aligned}$$

assuming all rewards are bounded by 1. □

Lemma B.5.2. *For any open set $\tilde{\Theta} \subset \Theta$, we have*

$$\int_{\tilde{\Theta}} \exp\left(-\frac{T_\ell}{2}\left(\|\theta^* - \theta\|^2_{A(\bar{e}_{T_\ell})}\right)\right) d\theta \doteq \exp\left(-\frac{T_\ell}{2} \inf_{\theta \in \tilde{\Theta}} \|\theta^* - \theta\|^2_{A(\bar{e}_{T_\ell})}\right).$$

Proof. The following argument is inspired by an analogous one in Lemma 11 of Russo [2016]. Let $\iota_\ell := \int_{\tilde{\Theta}} \exp\left(-\frac{T_\ell}{2} \|\theta^* - \theta\|^2_{A(\bar{e}_{T_\ell})}\right) d\theta$ and $W_{T_\ell}(\theta) := \frac{1}{2} \|\theta^* - \theta\|^2_{A(\bar{e}_{T_\ell})}$. Also, let $\tilde{\theta}_\ell \in \text{closure}(\tilde{\Theta})$ be a point that attains the infimum, i.e.

$$\tilde{\theta}_\ell := \arg\inf_{\theta \in \tilde{\Theta}} \|\theta^* - \theta\|^2_{A(\bar{e}_{T_\ell})}.$$

Such a point must exist by the continuity of $W_{T_\ell}(\theta)$ and closure$(\tilde{\Theta})$ being compact. Then, we first observe that

$$\int_{\tilde{\Theta}} \exp\left(-\frac{T_\ell}{2} \|\theta^* - \theta\|^2_{A(\bar{e}_{T_\ell})}\right) d\theta \leq \text{Vol}(\tilde{\Theta}) \exp\left(-\frac{T_\ell}{2} \left\|\theta^* - \tilde{\theta}_\ell\right\|^2_{A(\bar{e}_{T_\ell})}\right),$$

so

$$\limsup_{\ell \to \infty} \frac{1}{T_\ell} \log(\iota_\ell) + W_{T_\ell}(\tilde{\theta}_\ell) \leq 0.$$

Second, we fix some arbitrary $\epsilon > 0$. Note that for any $\theta, \theta' \in \Theta$,

$$\begin{aligned}
|W_{T_\ell}(\theta) - W_{T_\ell}(\theta')| &= \frac{1}{2}\left(\|\theta^* - \theta\|^2_{A(\bar{e}_{T_\ell})} - \|\theta^* - \theta'\|^2_{A(\bar{e}_{T_\ell})}\right) \\
&= \frac{1}{2}\left((2\theta^* - \theta - \theta')^\top A(\bar{e}_{T_\ell})(\theta - \theta')\right) \\
&= \frac{1}{2T_\ell}\sum_{t=1}^{T_\ell}\left((2\theta^* - \theta - \theta')^\top x_t x_t^\top (\theta - \theta')\right) \\
&\leq \Delta_{\max} \max_{x \in \mathcal{X}} x^\top (\theta - \theta') \\
&\leq \Delta_{\max} \max_{x \in \mathcal{X}} \|x\|_2 \|\theta - \theta'\|_2 \\
&\leq L\Delta_{\max} \|\theta - \theta'\|_2 .
\end{aligned}$$

Then, there exists $\delta > 0$ such that

$$\|\theta - \theta'\|_2 < \delta \Rightarrow |W_{T_\ell}(\theta) - W_{T_\ell}(\theta')| < \epsilon.$$

Then, we take a δ-cover of Θ with $\|\cdot\|_2$, and intersect them with $\tilde{\Theta}$, and denote the resulting cover as $\mathcal{O}$. Then, $\tilde{\theta}_\ell \in O$ for some $O \in \mathcal{O}$. Since we know that $\mathrm{Vol}(O) > 0$ for any $O \in \mathcal{O}$, we have

$$\iota_\ell \geq \int_O \exp\left(-T_\ell W_{T_\ell}(\theta)\right) d\theta \geq \mathrm{Vol}(O) \exp\left(-T_\ell\left(W_{T_\ell}\left(\tilde{\theta}_\ell\right) - \epsilon\right)\right).$$

Taking logarithm on both sides implies that

$$\frac{1}{T_\ell}\log\left(\iota_\ell\right) + W_{T_\ell}\left(\tilde{\theta}_\ell\right) \geq \frac{\mathrm{Vol}(O)}{T_\ell} - \epsilon \to -\epsilon.$$

Since we choose $\epsilon > 0$ arbitrarily, we have

$$\liminf_{\ell \to \infty} \frac{1}{T_\ell}\log\left(\iota_\ell\right) + W_{T_\ell}\left(\tilde{\theta}_\ell\right) \geq 0.$$

Therefore, $\lim_{\ell\to\infty} \frac{1}{T_\ell}\log\left(\iota_\ell\right) + W_{T_\ell}\left(\tilde{\theta}_\ell\right) = 0$ and the statement follows. □

B.6 Supplementary Plots

In this section, we present more supplementary plots. All experiments in the main text and supplement are run on a computing cluster with 64 AMD EPYC 7302 16-Core Processor

(1500 MHz) with 1TB of RAM. For LinGame, LinGapE, and Oracle algorithms, we directly use the existing implementation from Tirinzoni and Degenne [2022] with the open-source GitHub link: https://github.com/AndreaTirinzoni/bandit-elimination.

We demonstrate that the computational cost of our algorithm is not heavy. We first plot the average number of rejection samples taken to get some $\theta \in \Theta^c_{\widehat{z}_t}$ in the alternative and the running time for our algorithm to demonstrate the computation cost rejection sampling takes. Figures B.1 and B.2 show the result. By comparing Figure B.1 with Figure 3.1, we see that the number of rejection samples needed to get some $\theta \in \Theta^c_{\widehat{z}_t}$ is generally less than 30 until $\delta < 0.01$. This shows that the computational burden for rejection sampling is generally not large unless we have basically solved the problem. Also, we can see from Figure B.2 that the running time per iteration is generally very small, which means our algorithm runs very fast.

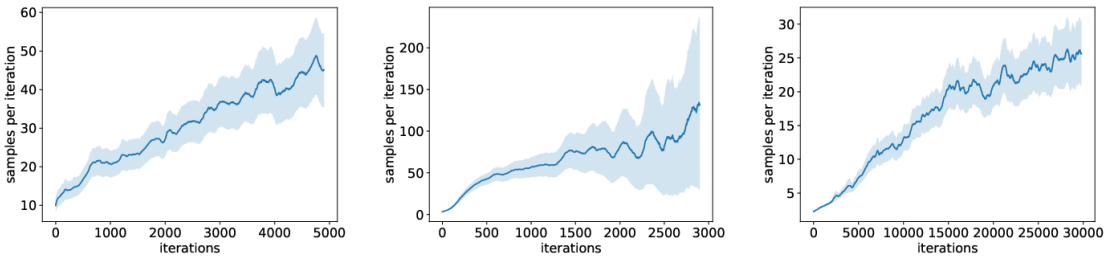

Figure B.1: Average number of rejection samples taken until finding some $\theta \in \Theta^c_{\widehat{z}_t}$

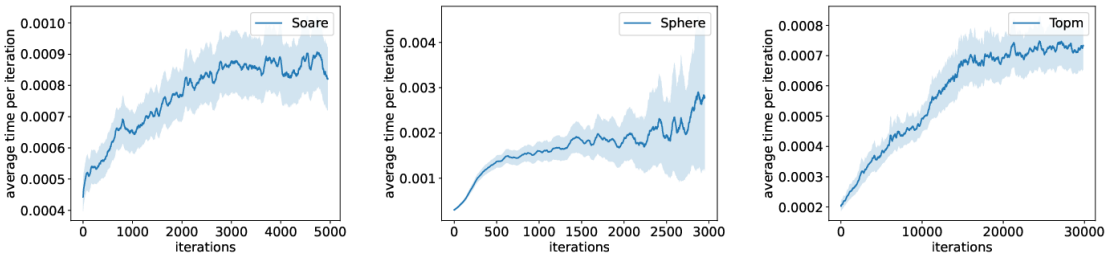

Figure B.2: Average clock time per iteration for PEPS under three scenarios

To make a clear comparison of the sampling part in our method with computing the best alternative step in LinGame, we implemented our algorithm, PEPS, in Julia and compared its clock time to existing LinGame implementations on a sphere instance with varying arm numbers, denoted as K. We run both algorithms for a fixed budget of 1000 iterations across 100 trials and compute the average clock time per iteration. We assessed both methods for $K = 50, 200, 1000, 5000, 10000, 20000$, with results presented in milliseconds. Table B.2 shows the results. We can see that our method consistently running faster than the benchmark LinGame, particularly as the number of arms increases. This distinction becomes especially significant when $K = 10000$ and $K = 20000$, which corresponds to the case that calculating the best alternative is expensive. Therefore, our method maintains efficiency even in scenarios when computing the alternative is really expensive.

	$K = 50$	$K = 200$	$K = 1000$	$K = 5000$	$K = 10000$	$K = 20000$
PEPS	0.132	0.484	0.681	3.770	6.710	17.110
LinGame	0.152	0.596	3.265	18.610	46.762	126.683

Table B.2: Average clock time per iteration for PEPS and LinGame under the sphere instance with $d = 6$ and various number of arms K. Numbers are displayed in milliseconds.

Appendix C

APPENDIX TO CHAPTER 4

C.1 Proofs for Section 4.2

Let $Q_{X,0}$ be the marginal distribution of X under P_0, and let $Q_{Y^*,0}$ and $Q_{Y^\dagger,0}$ be respectively the conditional distribution of Y^* and $Y^\dagger$ given A, X under P_0. Let $\{P_\epsilon : \epsilon \in \mathbb{R}\} \subset \mathcal{M}$ be a parametric submodel that is such that $P_\epsilon = P_0$ when $\epsilon = 0$. This submodel is defined so that the marginal distribution of X and the conditional distributions of $Y^\dagger$ and Y^* given (A, X) satisfy

$$dQ_{X,\epsilon}(x) = (1 + \epsilon S_X(x)) dQ_{X,0}(x), \text{ where } \mathbb{E}_0 [S_X(x)] = 0 \text{ and } \sup_x |S_X(x)| \leq m < \infty, \tag{C.1}$$

$$dQ_{Y^\dagger,\epsilon}(z \mid a, x) = (1 + \epsilon S_{Y^\dagger}(z \mid a, x))\, dQ_{Y^\dagger,0}(z \mid a, x), \tag{C.2}$$
$$\text{where } \mathbb{E}_0 [S_{Y^\dagger} \mid A, X] = 0 \ P_0\text{-a.s. and } \sup_{x,a,z} |S_{Y^\dagger}(z \mid a, x)| < \infty, \text{ and}$$

$$dQ_{Y^*,\epsilon}(y \mid a, x) = (1 + \epsilon S_{Y^*}(y \mid a, x))\, dQ_{Y^*,0}(y \mid a, x) \tag{C.3}$$
$$\text{where } \mathbb{E}_0 [S_{Y^*} \mid A, X] = 0 \ P_0\text{-a.s. and } \sup_{x,a,y} |S_{Y^*}(y \mid a, x)| < \infty.$$

We let $q_{b,\epsilon}(x) = q_b(P_\epsilon)(x)$ and $s_{b,\epsilon}(x) = s_b(P_\epsilon)(x)$.

Proof of Lemma 4.2.1. Note that $\pi^*_P(x) = \mathbb{I}\{q_b(P)(x) > 0\}$ for all $x \in \mathcal{X}$. Following the idea of the proof of Theorem 3 in Luedtke and Van Der Laan [2016], we observe that

$$\Psi^*(P) - \mathbb{E}_P \mathbb{E}_P [Y^\dagger \mid A = 0, X] = \mathbb{E}_P [\pi^*_P(X) s_b(P)(X)] .$$

By a telescoping argument,

$$\begin{aligned}\Psi^*(P_\epsilon) - \Psi^*(P_0) &= \mathbb{E}_{P_\epsilon}\mathbb{E}_{P_\epsilon}[Y^\dagger | A = \pi^*_{P_\epsilon}(X), X] - \mathbb{E}_{P_0}\mathbb{E}_{P_0}[Y^\dagger | A = \pi^*(X), X] \\ &= \mathbb{E}_{P_\epsilon}\mathbb{E}_{P_\epsilon}[Y^\dagger | A = \pi^*_{P_\epsilon}(X), X] - \mathbb{E}_{P_\epsilon}\mathbb{E}_{P_\epsilon}[Y^\dagger | A = \pi^*(X), X] \\ &\quad + \mathbb{E}_{P_\epsilon}\mathbb{E}_{P_\epsilon}[Y^\dagger | A = \pi^*(X), X] - \mathbb{E}_{P_0}\mathbb{E}_{P_0}[Y^\dagger | A = \pi^*(X), X] \\ &= \mathbb{E}_{P_\epsilon}[(\mathbb{I}(q_{b,\epsilon} > 0) - \mathbb{I}(q_{b,0} > 0)) \cdot s_{b,\epsilon}] + \Psi_{\pi^*}(P_\epsilon) - \Psi_{\pi^*}(P_0). \qquad \text{(C.4)}\end{aligned}$$

It is known that for a fixed π, Ψ_π is pathwise differentiable with gradient $D(\pi, P_0)$. We shall now show that the first term is $o(\epsilon)$. Letting $B_1 := \{x \in \mathcal{X} : q_{b,0}(x) = 0\}$, we have

$$\begin{aligned}&\mathbb{E}_{P_\epsilon}\left[\left(I\left(q_{b,\epsilon} > 0\right) - I\left(q_{b,0} > 0\right)\right) s_{b,\epsilon}\right] \\ &= \int_{\mathcal{X}\setminus B_1} \left(I\left(q_{b,\epsilon} > 0\right) - I\left(q_{b,0} > 0\right)\right) s_{b,\epsilon} dQ_{X,\epsilon} + \int_{B_1} \left(I\left(q_{b,\epsilon} > 0\right) - I\left(q_{b,0} > 0\right)\right) s_{b,\epsilon} dQ_{X,\epsilon}.\end{aligned}$$

Under Condition 1, we know that $\Pr_0(q_{b,0}(X) \neq 0) = 1$, so the second term is zero. Then we aim to show that the first term is $o(|\epsilon|)$. Note that

$$\begin{aligned}\left|\int_{\mathcal{X}\setminus B_1} \left(I\left(q_{b,\epsilon} > 0\right) - I\left(q_{b,0} > 0\right)\right) s_{b,\epsilon} dQ_{X,\epsilon}\right| &\leq \int_{\mathcal{X}\setminus B_1} \left|\left(I\left(q_{b,\epsilon} > 0\right) - I\left(q_{b,0} > 0\right)\right) s_{b,\epsilon}\right| dQ_{X,\epsilon} \\ &\leq \int_{\mathcal{X}\setminus B_1} I\left(|q_{b,0}| < |q_{b,\epsilon} - q_{b,0}|\right) |s_{b,\epsilon}| \, dQ_{X,\epsilon}\end{aligned}$$

by looking at the sign of $q_{b,\epsilon}$ and $q_{b,0}$. Also,

$$\begin{aligned}q_{b,\epsilon}(x) &= \int y \left(dQ_{Y^*,\epsilon}(y \mid A = 1, X = x) - dQ_{Y^*,\epsilon}(y \mid A = 0, X = x)\right) \\ &= q_{b,0}(x) + \epsilon \left(\mathbb{E}_0\left[Y^* S_{Y^*}(Y^* \mid 1, X) \mid A = 1, X = x\right] - \mathbb{E}_0\left[Y^* S_{Y^*}(Y^* \mid 0, X) \mid A = 0, X = x\right]\right) \\ &= q_{b,0}(x) + \epsilon \bar{h}(x)\end{aligned}$$

where

$$\bar{h}(x) = \mathbb{E}_0[Y^* S_{Y^*}(Y^*|1, X)|A = 1, X = x] - \mathbb{E}_0[Y^* S_{Y^*}(Y^*|0, X)|A = 0, X = x].$$

Similarly, $s_{b,\epsilon}(x) = s_{b,0}(x) + \epsilon \cdot \tilde{h}(x)$ where

$$\tilde{h}(x) = \mathbb{E}_0[Y^\dagger S_{Y^\dagger}(Y^\dagger|1, X)|A = 1, X = x] - \mathbb{E}_0[Y^\dagger S_{Y^\dagger}(Y^\dagger|0, X)|A = 0, X = x].$$

Note that $\tilde{h}$ and $\bar{h}$ are uniformly bounded since Y^*, $Y^\dagger$, S_{Y^*}, and $S_{Y^\dagger}$ are bounded. Let $H = \max\{\sup_x |\bar{h}(x)|,$
$\sup_x |\tilde{h}(x)|\}$. Therefore,

$$\begin{aligned}\int_{\mathcal{X}\setminus B_1} I\left(|q_{b,0}| < |q_{b,\epsilon} - q_{b,0}|\right)|s_{b,\epsilon}|\, dQ_{X,\epsilon} &\le \int_{\mathcal{X}\setminus B_1} I\left(|q_{b,0}| < H|\epsilon|\right)\left(|s_{b,0}| + H|\epsilon|\right) dQ_{X,\epsilon} \\ &\le (1 + m|\epsilon|)\int_{\mathcal{X}\setminus B_1} I\left(|q_{b,0}| < H|\epsilon|\right)\left(|s_{b,0}| + H|\epsilon|\right) dQ_{X,0} \\ &= (1 + m|\epsilon|)\int_{\mathcal{X}\setminus B_1} I\left(0 < |q_{b,0}| < H|\epsilon|\right)\left(|s_{b,0}| + H|\epsilon|\right) dQ_{X,0}.\end{aligned}$$

Denote $\tilde{\mathcal{X}} = \mathcal{X} \setminus B_1$. Under the first condition, define the set

$$B_{2,t} = \{x \in \tilde{\mathcal{X}} : |s_{b,0}(x)| < Ct^{-1}|q_{b,0}(x)|\}.$$

Then

$$\begin{aligned}&\int_{\mathcal{X}\setminus B_1} I\left(|q_{b,0}| < H|\epsilon|\right)\left(|s_{b,0}| + H|\epsilon|\right) dQ_{X,0} \\ &= \int_{\tilde{\mathcal{X}}} I\left(0 < |q_{b,0}| < H|\epsilon|\right)\left(|s_{b,0}| + H|\epsilon|\right) dQ_{X,0} \\ &= \int_{B_{2,t}} I\left(0 < |q_{b,0}| < H|\epsilon|\right)\left(|s_{b,0}| + H|\epsilon|\right) dQ_{X,0} + \int_{\tilde{\mathcal{X}}\setminus B_{2,t}} I\left(0 < |q_{b,0}| < H|\epsilon|\right)\left(|s_{b,0}| + H|\epsilon|\right) dQ_{X,0}.\end{aligned}$$

On one hand, note that for $x \in B_{2,t}$ and under the fact that $|q_{b,0}(x)| \le H|\epsilon|$ we have $|s_b(x)| \le CHt^{-1}|\epsilon|$. define C_2 such that $P_0(0 < |q_{b,0}(X)| < t) \le C_2 t^\gamma$ for any $t > 0$, the first term

$$\begin{aligned}\int_{B_{2,t}} I\left(0 < |q_{b,0}| < H|\epsilon|\right)\left(|s_{b,0}| + H|\epsilon|\right) dQ_{X,0} &\le \int_{B_{2,t}} I\left(0 < |q_{b,0}| < H|\epsilon|\right)\left(CHt^{-1}|\epsilon| + H|\epsilon|\right) dQ_{X,0} \\ &\le \left(CHt^{-1}|\epsilon| + H|\epsilon|\right) P_0\left(0 < |q_{b,0}(X)| < H|\epsilon|\right) \\ &\le \left(Ct^{-1}|\epsilon| + H|\epsilon|\right) C_2 (H|\epsilon|)^\gamma \qquad \text{(C.5)}\end{aligned}$$

for $t < 1$. For the second term, let $C_3 := \sup_x |s_{b,0}(x)|$, we have

$$\begin{aligned}&\int_{\tilde{\mathcal{X}}\setminus B_{2,t}} I\left(0 < |q_{b,0}| < H|\epsilon|\right)\left(|s_{b,0}| + H|\epsilon|\right) dQ_{X,0} \\ &\le (C_3 + H|\epsilon|) P_0(0 < |s_{b,0}(X)| > Ct^{-1}|q_{b,0}(X)|)) \\ &\le (C_3 + H|\epsilon|) t^\zeta\end{aligned}$$

where the last inequality follows from Condition 1. Therefore, the sum is bounded by

$$\left(Ct^{-1}|\epsilon| + H|\epsilon|\right) C_2(H|\epsilon|)^\gamma + (C_3 + H|\epsilon|)t^\zeta.$$

Taking $t = |\epsilon|^{\frac{1+\gamma}{\zeta+1}}$ gives that this is $O(|\epsilon|^{1+\gamma-\frac{1+\gamma}{\zeta+1}})$, which is $o(|\epsilon|)$ given that $\gamma > \frac{1}{\zeta}$. Combining all of the results above gives

$$\lim_{\epsilon\to 0} \frac{1}{\epsilon}\mathbb{E}_{P_\epsilon}\left[\left(I\left(q_{b,\epsilon} > 0\right) - I\left(q_{b,0} > 0\right)\right) s_{b,\epsilon}\right] = 0.$$

Therefore, Ψ^* is pathwise differentiable, and, per (C.4), has the same canonical gradient as the parameter Ψ_{π^*}, namely $D(\pi^*, P_0)$. □

Proof of Theorem 4.2.2. We would first like to show that $\psi_{OS,n}$ is an asymptotically linear estimator of ψ_0. For simplicity of notation, we let $\pi_n^* := \pi^*_{\widehat{P}_n}$ and drop the dependence of π in the definition of Ψ_π in this proof. Note that $\psi_{OS,n} - \psi_0 = (P_n - P_0)D(P_0) + (P_n - P_0)[D(\widehat{P}_n) - D(P_0)] + R(\widehat{P}_n, P_0)$. Note that the first term $(P_n - P_0)D(P_0)$ is the linear term and $(P_n - P_0)[D(\widehat{P}_n) - D(P_0)] = o_{P_0}(n^{-1/2})$ under the Donsker condition and the fact that $\|D(\widehat{P}_n) - D(P_0)\|_2 \xrightarrow{p} 0$ (Lemma 19.24 of Van der Vaart [2000]). To show that $\psi_{OS,n}$ is asymptotically linear, we only need to argue that the remainder term $R(\widehat{P}_n, P_0)$ is $o_{P_0}(n^{-1/2})$. Note that

$$\begin{aligned} P_0 D(\widehat{P}_n) &= \mathbb{E}_0\left[\frac{\mathbb{I}\{A = \pi_n^*(X)\}}{p_n(A|X)}(Y^\dagger - s(A,X)) + s(\pi_n^*(X), X) - \Psi(\widehat{P}_n)\right] \\ &= \mathbb{E}_0\left[\frac{\mathbb{I}\{A = \pi_n^*(X)\}}{p_n(A|X)}(s_0(A,X) - s(A,X)) + s(\pi_n^*(X), X) - \Psi(\widehat{P}_n)\right], \end{aligned}$$

by the law of total expectation. Therefore,

$$
\begin{aligned}
R(\widehat{P}_n, P_0) &= \Psi(\widehat{P}_n) - \Psi(P_0) + P_0 D(\widehat{P}_n) \\
&= \int \left\{ \frac{\mathbb{I}\{a = \pi_n^*(x)\}}{p_n(a|x)} (s_0(a,x) - s_n(a,x)) + s_n(\pi_n^*(x), x) - s_0(\pi^*(x), x) \right\} dP_0(a,x) \\
&= \int \left(\frac{\mathbb{I}\{a = \pi_n^*(x)\}}{p_n(a|x)} - 1 \right) [s_0(\pi_n^*(x), x) - s_n(\pi_n^*(x), x)] dP_0(a,x) + \Psi_{\pi_n^*}(P_0) - \Psi_{\pi^*}(P_0) \\
&= \iint \left(\frac{\mathbb{I}\{a = \pi_n^*(x)\}}{p_n(a|x)} - 1 \right) [s_0(\pi_n^*(x), x) - s_n(\pi_n^*(x), x)] p_0(a|x) da \, dP_0(x) \\
&\quad + \Psi_{\pi_n^*}(P_0) - \Psi_{\pi^*}(P_0) \\
&= \int \left(\frac{p_0(\pi_n^*(x)|x)}{p_n(\pi_n^*(x)|x)} - 1 \right) [s_0(\pi_n^*(x), x) - s_n(\pi_n^*(x), x)] dP_0(x) \\
&\quad + \Psi_{\pi_n^*}(P_0) - \Psi_{\pi^*}(P_0) \\
&=: R_{1n} + R_{2n}.
\end{aligned}
$$

The first term R_{1n} is $o_{P_0}(n^{-1/2})$ under under Condition 4 — see Proposition C.1.1. As for the second term R_{2n}, Proposition C.1.2 shows that it is $o_{P_0}(n^{-1/2})$ under the margin condition. □

Proposition C.1.1. *Under Condition 4,* $R_{1n} = o_{P_0}(n^{-1/2})$.

Proof. By Jensen's inequality, the fact that $\pi_n^*(x) \in \{0, 1\}$ for all x, the fact that $(b + c) \leq$

$2\max\{b, c\}$ for $b, c \in \mathbb{R}$, and Cauchy-Schwarz, we have that

$$\begin{aligned}
|R_{1n}| &= \left| \int \left(\frac{p_0(\pi_n^*(x)|x)}{p_n(\pi_n^*(x)|x)} - 1 \right) [s_0(\pi_n^*(x), x) - s_n(\pi_n^*(x), x)] dP_0(x) \right| \\
&\leq \int \left| \left(\frac{p_0(\pi_n^*(x)|x)}{p_n(\pi_n^*(x)|x)} - 1 \right) [s_0(\pi_n^*(x), x) - s_n(\pi_n^*(x), x)] \right| dP_0(x) \\
&\leq \int \sum_{a=0}^{1} \left| \left(\frac{p_0(a|x)}{p_n(a|x)} - 1 \right) [s_0(a, x) - s_n(a, x)] \right| dP_0(x) \\
&= \sum_{a=0}^{1} \int \left| \left(\frac{p_0(a|x)}{p_n(a|x)} - 1 \right) [s_0(a, x) - s_n(a, x)] \right| dP_0(x) \\
&\leq 2 \max_{a \in \{0,1\}} \int \left| \left(\frac{p_0(a|x)}{p_n(a|x)} - 1 \right) [s_0(a, x) - s_n(a, x)] \right| dP_0(x) \\
&\leq 2 \max_{a \in \{0,1\}} \left\{ \left\| \frac{p_0(a \mid X)}{p_n(a \mid X)} - 1 \right\|_{2,P_0} \|s_n(a, X) - s_0(a, X)\|_{2,P_0} \right\}.
\end{aligned}$$

□

The following proposition shows that the second term R_{2n} is $o_{P_0}(n^{-1/2})$ under our margin condition.

Proposition C.1.2. *Assume Conditions 1, 2, and 3 hold. Then, for any $\epsilon > 0$, $|R_{2n}| = o_{P_0}(n^{-1/2})$.*

Proof. We adopt the idea in proof of Theorem 8 of Luedtke and Van Der Laan [2016]. Let $B'_{3,u} = \{x \in \mathcal{X} : |s_{b,0}(x)| < C_1 u |q_{b,0}(x)|\}$ and $A_u = \{x \in \mathcal{X} : C_1 u |q_{b,0}(x)| \leq |s_{b,0}(x)| < C_1(u+1)\,|q_{b,0}(x)|\}$. Then for any $t > 0$,

$$\begin{aligned}
|\Psi_{\pi_n^*}(P_0) - \Psi_{\pi^*}(P_0)\,| &= \mathbb{E}_{P_0}\left[s_{b,0}(X)(\pi_n^*(X) - \pi^*(X))\right] \\
&\leq \mathbb{E}_0\left[|s_{b,0}(X)|\, I\left(\pi^*(X) \neq \pi_n^*(X)\right)\right] \\
&= \sum_{u=0}^{\infty} \mathbb{E}_0[|s_{b,0}(X)| I(\pi^*(X) \neq \pi_n^*(X)) I(A_u)] \\
&\leq \sum_{u=0}^{\infty} \mathbb{E}_0[|s_{b,0}(X)| I(|q_{b,0}(X)| \leq |q_{b,n}(X) - q_{b,0}(X)|) I(A_u)].
\end{aligned}$$

where the last inequality follows from the fact that for any $x \in \mathcal{X}$, $\pi^*(x) \neq \pi_n^*(x)$ implies that $|q_{b,n}(x) - q_{b,0}(x)| \geq |q_{b,0}(x)|$. From Condition 1 we know that $q_{b,0}(X) \neq 0$ with P_0-probability 1, so

$$\begin{aligned}
&\sum_{u=0}^{\infty} \mathbb{E}_0[|s_{b,0}(X)| I(|q_{b,0}(X)| \leq |q_{b,n}(X) - q_{b,0}(X)|) I(A_u)] \\
&= \sum_{u=0}^{\infty} \mathbb{E}_0[|s_{b,0}(X)| I(0 < |q_{b,0}(X)| \leq |q_{b,n}(X) - q_{b,0}(X)|) I(A_u)].
\end{aligned}$$

For any $x \in A_u$, $|s_{b,0}(x)| \leq C_1(u+1)|q_{b,0}(x)|$, so for each u,

$$\begin{aligned}
&\mathbb{E}_0 \left[|s_{b,0}(X)| \, I \left(0 < |q_{b,0}(X)| \leq |q_{b,n}(X) - q_{b,0}(X)|\right) I(A_u)\right] \\
&\leq C_1 \mathbb{E}_0 \left[(u+1) \, |q_{b,0}(X)| \, I \left(0 < |q_{b,0}(X)| \leq |q_{b,n}(X) - q_{b,0}(X)|\right) I(A_u)\right] \\
&\leq C_1 \mathbb{E}_0 \left[(u+1) \, |q_{b,n}(X) - q_{b,0}(X)| \, I \left(0 < |q_{b,0}(X)| \leq |q_{b,n}(X) - q_{b,0}(X)|\right) I(A_u)\right] \\
&\leq C_1 \mathbb{E}_0 \left[(u+1) \max_{x \in \mathcal{X}} \|q_{b,n}(x) - q_{b,0}(x)\| \, I \left(0 < |q_{b,0}(X)| \leq \max_{x \in \mathcal{X}} \|q_{b,n}(x) - q_{b,0}(x)\|\right) I(A_u)\right] \\
&= C_1(u+1) \, \|q_{b,n} - q_{b,0}\|_{\infty,P_0} \, \mathbb{E}_0 \left[I \left(0 < |q_{b,0}(X)| \leq \max_{x \in \mathcal{X}} \|q_{b,n}(x) - q_{b,0}(x)\|\right) I(A_u)\right] \\
&= C_1(u+1) \, \|q_{b,n} - q_{b,0}\|_{\infty,P_0} \, P_0(0 < |q_{b,0}(X)| \leq \|q_{b,n} - q_{b,0}\|_{\infty,P_0}, A_u).
\end{aligned}$$

For an event $\mathcal{E} \subseteq \mathcal{X}$, let $\mathbb{P}^{\infty}(\mathcal{E}) := P_0(0 < |q_{b,0}(X)| \leq \|q_{b,n} - q_{b,0}\|_{\infty,P_0}, \mathcal{E})$. Then, for any $k \in \mathbb{N}$,

$$
\begin{aligned}
&\sum_{u=0}^{k} \mathbb{E}_0[|s_{b,0}(X)|I(0<|q_{b,0}(X)|\leq|q_{b,n}(X)-q_{b,0}(X)|)I(A_u)] \\
&\leq \sum_{u=0}^{k} C_1(u+1)\left\|q_{b,n}-q_{b,0}\right\|_{\infty,P_0}\mathbb{P}^{\infty}(A_u) \\
&= \sum_{u=0}^{k} C_1(u+1)\left\|q_{b,n}-q_{b,0}\right\|_{\infty,P_0}[\mathbb{P}^{\infty}(B'_{3,u+1})-\mathbb{P}^{\infty}(B'_{3,u})] \\
&= \sum_{u=0}^{k} C_1(u+1)\left\|q_{b,n}-q_{b,0}\right\|_{\infty,P_0}\mathbb{P}^{\infty}(B'_{3,u+1}) - \sum_{u=0}^{k} C_1(u+1)\left\|q_{b,n}-q_{b,0}\right\|_{\infty,P_0}\mathbb{P}^{\infty}(B'_{3,u}) \\
&= \sum_{u=1}^{k+1} C_1 u\left\|q_{b,n}-q_{b,0}\right\|_{\infty,P_0}\mathbb{P}^{\infty}(B'_{3,u}) - \sum_{u=0}^{k} C_1(u+1)\left\|q_{b,n}-q_{b,0}\right\|_{\infty,P_0}\mathbb{P}^{\infty}(B'_{3,u}) \\
&= C_1(k+1)\left\|q_{b,n}-q_{b,0}\right\|_{\infty,P_0}\mathbb{P}^{\infty}(B'_{3,k+1}) - \sum_{u=0}^{k} C_1\left\|q_{b,n}-q_{b,0}\right\|_{\infty,P_0}\mathbb{P}^{\infty}(B'_{3,u}) \\
&= \sum_{u=0}^{k} C_1\left\|q_{b,n}-q_{b,0}\right\|_{\infty,P_0}[\mathbb{P}^{\infty}(B'_{3,k+1})-\mathbb{P}^{\infty}(B'_{3,u})] \\
&\leq \sum_{u=0}^{k} C_1\left\|q_{b,n}-q_{b,0}\right\|_{\infty,P_0}[\mathbb{P}^{\infty}(\mathcal{X})-\mathbb{P}^{\infty}(B'_{3,u})] \\
&= \sum_{u=0}^{k} C_1\left\|q_{b,n}-q_{b,0}\right\|_{\infty,P_0}[\mathbb{P}^{\infty}(B'^{c}_{3,u})] \\
&= \sum_{u=0}^{k} C_1\left\|q_{b,n}-q_{b,0}\right\|_{\infty,P_0}[\mathbb{P}_0(0<|q_{b,0}(X)|\leq\left\|q_{b,n}-q_{b,0}\right\|_{\infty,P_0}, B'^{c}_{3,u})] \\
&\leq \sum_{u=0}^{k} C_1\left\|q_{b,n}-q_{b,0}\right\|_{\infty,P_0}^{1+\gamma/2} u^{-\zeta/2}.
\end{aligned}
$$

where the last step follows from Holder's inequality. Since $\zeta > 2$, let $k \to \infty$ and the infinite sum converges. Therefore,

$$
\begin{aligned}
|\Psi_{\pi_n^*}(P_0) - \Psi_{\pi^*}(P_0)| &= \sum_{u=1}^{\infty} \mathbb{E}_0[|s_{b,0}(X)|I(\pi^*(X)\neq\pi_n^*(X))|A_u]\mathbb{P}(A_u) \\
&= \lim_{k\to\infty}\sum_{u=1}^{k} \mathbb{E}_0[|s_{b,0}(X)|I(\pi^*(X)\neq\pi_n^*(X))|A_u]\mathbb{P}(A_u) \lesssim \left\|q_{b,n}-q_{b,0}\right\|_{p,P_0}^{1+\gamma/2}.
\end{aligned}
$$

Note that under Condition 3, we have $\|q_{b,n} - q_{b,0}\|_{\infty,P_0}^{1+\gamma/2} = o_{P_0}(n^{-1/2})$ for any $\gamma > 0$, so $|R_{2n}| = o_{P_0}(n^{-1/2})$. □

C.2 Proofs for Section 4.3

For notational simplicity, throughout this section and later we denote $\psi_\pi := \Psi_\pi(P_0)$ for some policy $\pi \in \Pi$.

Lemma C.2.1. *If* $\inf_{\pi\in\Pi} \sigma_\pi(P_0) > 0$, *and* $\widehat{\sigma}_\pi$ *is a consistent estimator of* $\sigma_\pi(P_0)$ *for each* $\pi \in \Pi$, *an asymptotically valid uniform* β*-level confidence band is given by* $\left\{\widehat{\omega}_\pi \pm \frac{\widehat{\sigma}_\pi t_\beta}{n^{1/2}} : \pi \in \Pi\right\}$.

Proof of Lemma C.2.1. To see that this is the case, note that t_β is the $1 - \beta/2$ quantile of $\sup_{f\in\mathcal{F}} \mathbb{G}f$, and also

$$\begin{aligned}
&P \cap_{\pi\in\Pi} \left\{\widehat{\omega}_\pi - \frac{\widehat{\sigma}_\pi t_\beta}{n^{1/2}} \le \omega_\pi \le \widehat{\omega}_\pi + \frac{\widehat{\sigma}_\pi t_\beta}{n^{1/2}}\right\} \\
&= P \cap_{\pi\in\Pi} \left\{-t_\beta \le n^{1/2}\frac{\widehat{\omega}_\pi - \omega_\pi}{\widehat{\sigma}_\pi} \le t_\beta\right\} \\
&\to P \cap_{\pi\in\Pi} \{-t_\beta \le \mathbb{G}f \le t_\beta\} \\
&= P \cap_{\pi\in\Pi} \left[\left\{-t_\beta \le \inf_{f\in\mathcal{F}} \mathbb{G}f\right\} \cap \left\{\sup_{f\in\mathcal{F}} \mathbb{G}f \le t_\beta\right\}\right] \\
&= 1 - \beta,
\end{aligned}$$

where the convergence follows from the fact that $n^{1/2}\frac{\widehat{\omega}_\pi - \omega_\pi}{\widehat{\sigma}_\pi} \rightsquigarrow \mathbb{G}f$ by Lemma C.2.3 and Slutsky's Theorem. □

Proof of Lemma 4.3.1. We have that

$$\left\{\Pi^* \subseteq \widehat{\Pi}_\beta\right\} = \left\{\omega_{\pi'} < \sup_{\pi\in\Pi} \omega_\pi, \forall \pi' \in \widehat{\Pi}_\beta^C\right\}.$$

Therefore,

$$\begin{aligned}
&\left\{\Pi^* \subseteq \widehat{\Pi}_\beta\right\}^C \\
&= \left\{\exists \pi' \in \widehat{\Pi}_\beta^C : \omega_{\pi'} = \sup_{\pi\in\Pi} \omega_\pi\right\} \\
&\subseteq \left\{\exists \pi' \in \widehat{\Pi}_\beta^C : \left[\omega_{\pi'} - \widehat{\omega}_{\pi'} - \frac{\widehat{\sigma}_{\pi'} t_\beta}{n^{1/2}} + L_n\right] > \sup_{\pi\in\Pi} \omega_\pi, \right\} \\
&= \left\{\exists \pi' \in \widehat{\Pi}_\beta^C : \left[\omega_{\pi'} - \widehat{\omega}_{\pi'} - \frac{\widehat{\sigma}_{\pi'} t_\beta}{n^{1/2}}\right] > \sup_{\pi\in\Pi} \omega_\pi - L_n\right\}, \qquad \text{(C.6)}
\end{aligned}$$

where the inclusion follows from the definition of $\widehat{\Pi}_\beta$. Let $\mathcal{A}$ denote the event $\{L_n \leq \sup_{\pi\in\Pi} \omega_\pi\} \cap \left[\cap_{\pi\in\Pi} \left\{\omega_\pi \leq \widehat{\omega}_\pi + \frac{\widehat{\sigma}_\pi t_\beta}{n^{1/2}}\right\}\right]$. Hence, (C.6) shows that

$$\begin{aligned}
&\left\{\Pi^* \not\subseteq \widehat{\Pi}_\beta\right\}^C \\
&\subseteq \left[\left\{\exists \pi' \in \widehat{\Pi}_\beta^C : \left[\omega_{\pi'} - \widehat{\omega}_{\pi'} - \frac{\widehat{\sigma}_{\pi'} t_\beta}{n^{1/2}}\right] > \sup_{\pi\in\Pi} \omega_\pi - L_n\right\} \cap \mathcal{A}\right] \cup \mathcal{A}^C \\
&\subseteq \left[\left\{\exists \pi' \in \widehat{\Pi}_\beta^C : \omega_{\pi'} - \widehat{\omega}_{\pi'} - \frac{\widehat{\sigma}_{\pi'} t_\beta}{n^{1/2}} > 0\right\} \cap \mathcal{A}\right] \cup \mathcal{A}^C \\
&= \mathcal{A}^C,
\end{aligned}$$

where the final equality used that the leading event in the union above is equal to the null set since under $\mathcal{A}$, we have $\omega_{\pi'} - \widehat{\omega}_{\pi'} - \frac{\widehat{\sigma}_{\pi'} t_\beta}{n^{1/2}} \leq 0$ for each $\pi \in \Pi$. Also, note that by Lemma C.2.1, $\Pr\left(\cap_{\pi\in\Pi} \left\{\omega_\pi \leq \widehat{\omega}_\pi + \frac{\widehat{\sigma}_\pi t_\beta}{n^{1/2}}\right\}\right) \to 1 - \beta/2$, and by definition of L_n, $\limsup_n \Pr\left(\{L_n < \sup_{\pi\in\Pi} \omega_\pi\}\right) \geq 1 - \beta/2$. Hence, by a union bound,

$$\limsup_n P\left\{\Pi^* \not\subseteq \widehat{\Pi}_\beta\right\} \leq \beta.$$

□

Lemma C.2.2. *For any $\beta > 0$, $\liminf_{n\to\infty} \mathbb{P}\left(\omega_{\pi^*} - \inf_{\pi\in\widehat{\Pi}_\beta} \omega_\pi \leq \frac{4t_\beta}{n^{1/2}} \sup_{\pi\in\Pi} \widehat{\sigma}_\pi\right) \geq 1 - \beta$.*

Proof of Lemma C.2.2. Note that by the definition of $\widehat{\Pi}_\beta$, we have

$$\inf_{\pi\in\widehat{\Pi}_\beta} \left[\widehat{\omega}_\pi + \frac{\widehat{\sigma}_\pi t_\beta}{n^{1/2}}\right] \geq \sup_{\pi\in\Pi} \left[\widehat{\omega}_\pi - \frac{\widehat{\sigma}_\pi t_\beta}{n^{1/2}}\right],$$

so

$$\omega_{\pi^*} - \inf_{\pi \in \widehat{\Pi}_\beta} \omega_\pi \leq \omega_{\pi^*} - \sup_{\pi \in \Pi} \left[\widehat{\omega}_\pi - \frac{\widehat{\sigma}_\pi t_\beta}{n^{1/2}} \right] + \inf_{\pi \in \widehat{\Pi}_\beta} \left[\widehat{\omega}_\pi + \frac{\widehat{\sigma}_\pi t_\beta}{n^{1/2}} \right] - \inf_{\pi \in \widehat{\Pi}_\beta} \omega_\pi.$$

Hence,

$$\begin{aligned}
&\left\{ \omega_{\pi^*} - \inf_{\pi \in \widehat{\Pi}_\beta} \omega_\pi > \frac{4t_\beta}{n^{1/2}} \sup_{\pi \in \Pi} \widehat{\sigma}_\pi \right\} \\
&\subseteq \left\{ \omega_{\pi^*} - \sup_{\pi \in \Pi} \left\{ \widehat{\omega}_\pi - \frac{\widehat{\sigma}_\pi t_\beta}{n^{1/2}} \right\} + \inf_{\pi \in \widehat{\Pi}_\beta} \left\{ \widehat{\omega}_\pi + \frac{\widehat{\sigma}_\pi t_\beta}{n^{1/2}} \right\} - \inf_{\pi \in \widehat{\Pi}_\beta} \omega_\pi > \frac{4t_\beta}{n^{1/2}} \sup_{\pi \in \Pi} \widehat{\sigma}_\pi \right\} \\
&\subseteq \left\{ \omega_{\pi^*} > \sup_{\pi \in \Pi} \left\{ \widehat{\omega}_\pi - \frac{\widehat{\sigma}_\pi t_\beta}{n^{1/2}} \right\} + 2 \sup_{\pi \in \Pi} \frac{\widehat{\sigma}_\pi t_\beta}{n^{1/2}} \right\} \\
&\qquad \cup \left\{ \inf_{\pi \in \widehat{\Pi}_\beta} \omega_\pi < \inf_{\pi \in \widehat{\Pi}_\beta} \left\{ \widehat{\omega}_\pi + \frac{\widehat{\sigma}_\pi t_\beta}{n^{1/2}} \right\} - 2 \sup_{\pi \in \Pi} \frac{\widehat{\sigma}_\pi t_\beta}{n^{1/2}} \right\}. \qquad \text{(C.7)}
\end{aligned}$$

In the remainder of this proof, we will show that the two events on the right-hand side each occur with probability no more than $\beta/2$. The result then follows by a union bound. Note that

$$\begin{aligned}
\bigcap_{\pi \in \Pi} \left\{ \omega_\pi \leq \widehat{\omega}_\pi + \frac{\widehat{\sigma}_\pi t_\beta}{n^{1/2}} \right\} &\subseteq \left\{ \omega_{\pi^*} \leq \sup_{\pi \in \Pi} \left\{ \widehat{\omega}_\pi + \frac{\widehat{\sigma}_\pi t_\beta}{n^{1/2}} \right\} \right\} \\
&\subseteq \left\{ \omega_{\pi^*} \leq \sup_{\pi \in \Pi} \left\{ \widehat{\omega}_\pi - \frac{\widehat{\sigma}_\pi t_\beta}{n^{1/2}} \right\} + 2 \sup_{\pi \in \Pi} \frac{\widehat{\sigma}_\pi t_\beta}{n^{1/2}} \right\},
\end{aligned}$$

where the latter inclusion holds because $\sup[f + g] \leq \sup f + \sup g$. So

$$\begin{aligned}
&\liminf_{n \to \infty} \mathbb{P} \left(\omega_{\pi^*} - \sup_{\pi \in \Pi} \left\{ \widehat{\omega}_\pi - \frac{\widehat{\sigma}_\pi t_\beta}{n^{1/2}} \right\} \leq 2 \sup_{\pi \in \Pi} \frac{\widehat{\sigma}_\pi t_\beta}{n^{1/2}} \right) \\
&\geq \liminf_{n \to \infty} \mathbb{P} \left(\bigcap_{\pi \in \Pi} \left\{ \omega_\pi \leq \widehat{\omega}_\pi + \frac{\widehat{\sigma}_\pi t_\beta}{n^{1/2}} \right\} \right) \geq 1 - \frac{\beta}{2},
\end{aligned}$$

where the last step follows from Lemma C.2.1. Hence, the first event on the right-hand side

of (C.7) occurs with probability no more than probability $\beta/2$. We also have that

$$\begin{aligned}
&\bigcap_{\pi\in\Pi}\left\{\omega_\pi \geq \widehat{\omega}_\pi - \frac{\widehat{\sigma}_\pi t_\beta}{n^{1/2}}\right\} \subseteq \bigcap_{\pi\in\widehat{\Pi}_\beta}\left\{\omega_\pi \geq \widehat{\omega}_\pi - \frac{\widehat{\sigma}_\pi t_\beta}{n^{1/2}}\right\} \\
&\subseteq \left\{\inf_{\pi\in\widehat{\Pi}_\beta}\omega_\pi \geq \inf_{\pi\in\widehat{\Pi}_\beta}\left\{\widehat{\omega}_\pi - \frac{\widehat{\sigma}_\pi t_\beta}{n^{1/2}}\right\}\right\} \\
&\subseteq \left\{\inf_{\pi\in\widehat{\Pi}_\beta}\omega_\pi \geq \inf_{\pi\in\widehat{\Pi}_\beta}\left\{\widehat{\omega}_\pi + \frac{\widehat{\sigma}_\pi t_\beta}{n^{1/2}}\right\} - 2\sup_{\pi\in\widehat{\Pi}_\beta}\frac{\widehat{\sigma}_\pi t_\beta}{n^{1/2}}\right\},
\end{aligned}$$

since $\inf_{[} f - g] \geq \inf_f - \sup g$. So

$$\begin{aligned}
&\liminf_{n\to\infty}\mathbb{P}\left(\inf_{\pi\in\widehat{\Pi}_\beta}\left\{\widehat{\omega}_\pi + \frac{\widehat{\sigma}_\pi t_\beta}{n^{1/2}}\right\} - 2\sup_{\pi\in\widehat{\Pi}_\beta}\frac{\widehat{\sigma}_\pi t_\beta}{n^{1/2}} \leq \inf_{\pi\in\widehat{\Pi}_\beta}\omega_\pi\right) \\
&\geq \liminf_{n\to\infty}\mathbb{P}\left(\bigcap_{\pi\in\Pi}\left\{\widehat{\omega}_\pi - \frac{\widehat{\sigma}_\pi t_\beta}{n^{1/2}} \leq \omega_\pi\right\}\right) \geq 1 - \frac{\beta}{2},
\end{aligned}$$

where the last step follows from Lemma C.2.1. Hence, the second event on the right-hand side of (C.7) occurs with probability no more than probability $\beta/2$. □

In the following lemma, for some subset $\mathcal{G}$ of a space $L^2(Q)$, define the covering number $N(\epsilon, \mathcal{G}, L^2(Q))$ to be the minimal cardinality of an ϵ-cover of $\mathcal{G}$ with respect to the $L^2(Q)$ metric Van Der Vaart and Wellner [2013]. Before stating the lemma, we recall that $\mathcal{F} := \{D_\pi(P_0)/\sigma_\pi(P_0) : \pi \in \Pi\}$.

Lemma C.2.3 ($\mathcal{F}$ is P_0-Donsker)**.** *Assume that Conditions 6 and 7 hold and also that*

(i) Π *satisfies the uniform entropy bound, that is,* $\int_0^\infty \sup_{Q_X}\sqrt{\log N(\varepsilon, \Pi, L^2(Q_X))}d\varepsilon < \infty$, *where the supremum is over all finitely supported measures on* $\mathcal{X}$*;*

(ii) there exists $L > 0$ *such that, for all finitely supported distributions* Q *of* (X, A, Y) *with support on* $\mathcal{X}\times\{0,1\}\times\mathcal{Y}$*, the gradient map* $\pi \mapsto D_\pi$ *is* L*-Lipschitz, in the sense that, for any* $\pi, \pi' \in \Pi$, $\|D_\pi - D_{\pi'}\|_{L^2(Q)} \leq L\|\pi - \pi'\|_{L^2(Q_X)}$, *where* Q_X *is the marginal distribution of* X *under* Q*;*

(iii) $\sup_{\pi\in\Pi} \operatorname{ess\,sup}_{x\in\mathcal{X},a\in\{0,1\},y\in\mathcal{Y}} |D_\pi(P_0)(x,a,y)| < \infty$.

Then, the set $\mathcal{F} := \{D_\pi(P_0)/\sigma_\pi(P_0) : \pi \in \Pi\}$ *is* P_0*-Donsker.*

Proof of Lemma C.2.3. We would like to use Theorem 2.5.2 of Van Der Vaart and Wellner [2013]. First, by (iii) and Condition 7,

$$C := \frac{\sup_{\pi\in\Pi} \operatorname{ess\,sup}_{x\in\mathcal{X},a\in\{0,1\},y\in\mathcal{Y}} |D_\pi(P_0)(x,a,y)|}{\inf_{\pi\in\Pi} \sigma_\pi(P_0)} < \infty.$$

Hence, an envelope function for $\mathcal{F}$ is given by the constant function $F(x,a,y) = C$. By (ii) and properties of covering numbers, for any Q as stated in (ii) and implied marginal distribution Q_X, we have that $N(C\varepsilon, \mathcal{F}, L^2(Q)) \leq N(C\varepsilon/L, \Pi, L^2(Q_X))$. Combining this with (i) shows that $\mathcal{F}$ satisfies the uniform entropy bound in the sense that $\int_0^\infty \sup_Q \sqrt{\log N(\varepsilon, \mathcal{F}, L^2(Q))}d\varepsilon < \infty$, where the supremum is over all finitely supported measures on $\mathcal{X}\times\{0,1\}\times\mathcal{Y}$. Hence, $\mathcal{F}$ is P_0-Donsker by Theorem 2.5.2 of Van Der Vaart and Wellner [2013]. □

Lemma C.2.4. Π^* *is a closed subset of* $L^2(P_0)$.

Proof. Let $(\pi_k)_{k=1}^\infty$ be a Π^*-valued sequence that converges to some π^* in $L^2(P)$. Since $\pi \mapsto \omega_\pi$ is a continuous map from $\{0,1\}^\mathcal{X}$ to $\mathbb{R}$ when the domain is equipped with the $L^2(P)$-topology, $\omega_{\pi_k} \to \omega_{\pi^*}$. As $\pi_k \in \Pi^*$ for all k, $\omega_{\pi_k} = \sup_{\pi\in\Pi} \omega_\pi$ for all k. Hence, $\omega_{\pi^*} = \sup_{\pi\in\Pi} \omega_\pi$. As Π is closed, this shows that $\pi^* \in \Pi^*$. Hence, Π^* is a closed subset of $L^2(P)$. □

Lemma C.2.5. *If* Π^* *is closed in* $L^2(P_0)$ *and* Π^* *is* P_0*-Donsker,* Π^* *is compact.*

Proof of Lemma C.2.5. Since Π^* is P_0-Donsker following from Π being P_0-Donsker, then Π^* is totally bounded in $L^2(P_0)$. Also, since $L^2(P_0)$ is complete, Π^* being closed implies that Π^* is complete. And totally bounded and complete subsets of a metric space are compact, so Π^* is compact. □

Proof of Theorem 4.3.2. We have that

$$\begin{aligned}
&\left\{[\inf_{\pi\in\Pi^*}\psi_\pi, \sup_{\pi\in\Pi^*}\psi_\pi] \not\subseteq \mathrm{CI}_n\right\} \\
&= \left\{\inf_{\pi\in\Pi^*}\psi_\pi < \inf_{\pi\in\widehat{\Pi}_\beta}\left[\widehat{\psi}_\pi - \frac{\widehat{\kappa}_\pi z_{\alpha,\beta}}{n^{1/2}}\right]\right\} \cup \left\{\sup_{\pi\in\Pi^*}\psi_\pi > \sup_{\pi\in\widehat{\Pi}_\beta}\left[\widehat{\psi}_\pi + \frac{\widehat{\kappa}_\pi z_{\alpha,\beta}}{n^{1/2}}\right]\right\} \\
&\subseteq \left\{\inf_{\pi\in\Pi^*}\psi_\pi < \inf_{\pi\in\widehat{\Pi}_\beta}\left[\widehat{\psi}_\pi - \frac{\widehat{\kappa}_\pi z_{\alpha,\beta}}{n^{1/2}}\right], \Pi^* \subseteq \widehat{\Pi}_\beta\right\} \\
&\quad \cup \left\{\sup_{\pi\in\Pi^*}\psi_\pi > \sup_{\pi\in\widehat{\Pi}_\beta}\left[\widehat{\psi}_\pi + \frac{\widehat{\kappa}_\pi z_{\alpha,\beta}}{n^{1/2}}\right], \Pi^* \subseteq \widehat{\Pi}_\beta\right\} \cup \left\{\Pi^* \not\subseteq \widehat{\Pi}_\beta\right\}.
\end{aligned}$$

Hence, by a union bound and the fact that $\limsup_n(a_n+b_n+c_n) \leq \limsup_n a_n + \limsup_n b_n + \limsup_n c_n$, we see that

$$\begin{aligned}
&\limsup_n P\left\{\mathrm{CI}_n \not\subseteq [\inf_{\pi\in\Pi^*}\psi_\pi, \sup_{\pi\in\Pi^*}\psi_\pi]\right\} \\
&\leq \limsup_n P\left\{\inf_{\pi\in\Pi^*}\psi_\pi < \inf_{\pi\in\widehat{\Pi}_\beta}\left[\widehat{\psi}_\pi - \frac{\widehat{\kappa}_\pi z_{\alpha,\beta}}{n^{1/2}}\right], \Pi^* \subseteq \widehat{\Pi}_\beta\right\} \\
&\quad + \limsup_n P\left\{\sup_{\pi\in\Pi^*}\psi_\pi > \sup_{\pi\in\widehat{\Pi}_\beta}\left[\widehat{\psi}_\pi + \frac{\widehat{\kappa}_\pi z_{\alpha,\beta}}{n^{1/2}}\right], \Pi^* \subseteq \widehat{\Pi}_\beta\right\} \\
&\quad + \limsup_n P\left\{\Pi^* \not\subseteq \widehat{\Pi}_\beta\right\}.
\end{aligned}$$

The third term is upper bounded by β by Lemma 4.3.1. In what follows we will show that the first term on the right-hand side is no more than $(\alpha-\beta)/2$. Similar arguments can be used to show that the second term is also no more than $(\alpha-\beta)/2$. By a union bound argument, the sum of three terms is upper bounded by α, which completes the proof.

We begin by noting that, for any $n \in \mathbb{N}$,

$$\begin{aligned}
&\left\{\inf_{\pi\in\Pi^*}\psi_\pi < \inf_{\pi\in\widehat{\Pi}_\beta}\left[\widehat{\psi}_\pi - \frac{\widehat{\kappa}_\pi z_{\alpha,\beta}}{n^{1/2}}\right], \Pi^* \subseteq \widehat{\Pi}_\beta\right\} \\
&\subseteq \left\{\inf_{\pi\in\Pi^*}\psi_\pi < \inf_{\pi\in\Pi^*}\left[\widehat{\psi}_\pi - \frac{\widehat{\kappa}_\pi z_{\alpha,\beta}}{n^{1/2}}\right], \Pi^* \subseteq \widehat{\Pi}_\beta\right\} \\
&\subseteq \left\{\inf_{\pi\in\Pi^*}\psi_\pi < \inf_{\pi\in\Pi^*}\left[\widehat{\psi}_\pi - \frac{\widehat{\kappa}_\pi z_{\alpha,\beta}}{n^{1/2}}\right]\right\}.
\end{aligned}$$

By Lemma C.2.5 and $\pi \mapsto \psi_\pi$ is continuous, there exists a π^ℓ such that $\psi_{\pi^\ell} = \inf_{\pi \in \Pi^*} \psi_\pi$. Combining this with the above, we see that

$$\left\{ \inf_{\pi \in \Pi^*} \psi_\pi < \inf_{\pi \in \widehat{\Pi}_\beta} \left[\widehat{\psi}_\pi - \frac{\widehat{\kappa}_\pi z_{\alpha,\beta}}{n^{1/2}} \right], \Pi^* \subseteq \widehat{\Pi}_\beta \right\} \subseteq \left\{ \psi_{\pi^\ell} < \inf_{\pi \in \Pi^*} \left[\widehat{\psi}_\pi - \frac{\widehat{\kappa}_\pi z_{\alpha,\beta}}{n^{1/2}} \right] \right\}$$
$$\subseteq \left\{ \psi_{\pi^\ell} < \widehat{\psi}_{\pi^\ell} - \frac{\widehat{\kappa}_{\pi^\ell} z_{\alpha,\beta}}{n^{1/2}} \right\}.$$

Then

$$P\left(\psi_{\pi^\ell} < \widehat{\psi}_{\pi^\ell} - \frac{\widehat{\kappa}_{\pi^\ell} z_{\alpha,\beta}}{n^{1/2}} \right) = P\left(n^{1/2} \frac{\widehat{\psi}_{\pi^\ell} - \psi_{\pi^\ell}}{\widehat{\kappa}_{\pi^\ell}} > z_{\alpha,\beta} \right).$$

By Condition 7, $\widehat{\kappa}_{\pi^\ell}$ is a consistent estimator for $\kappa_{\pi^\ell}(P_0)$. Then with Slutsky's Theorem, $n^{1/2} \frac{\widehat{\psi}_{\pi^\ell} - \psi_{\pi^\ell}}{\widehat{\kappa}_{\pi^\ell}} \rightsquigarrow \mathbb{G} f_{\pi^\ell}$, so by definition of $z_{\alpha,\beta}$, $P\left(n^{1/2} \frac{\widehat{\psi}_{\pi^\ell} - \psi_{\pi^\ell}}{\widehat{\kappa}_{\pi^\ell}} > z_{\alpha,\beta} \right) \leq (\alpha - \beta)/2$, and so

$$\limsup_{n \to \infty} \mathbb{P}\left\{ \inf_{\pi \in \Pi^*} \psi_\pi < \inf_{\pi \in \widehat{\Pi}_\beta} \left[\widehat{\psi}_\pi - \frac{\widehat{\kappa}_\pi z_{\alpha,\beta}}{n^{1/2}} \right], \Pi^* \subseteq \widehat{\Pi}_\beta \right\} \leq (\alpha - \beta)/2.$$

By a symmetric argument, we also have $\limsup_{n \to \infty} P\left\{ \sup_{\pi \in \Pi^*} \psi_\pi > \sup_{\pi \in \widehat{\Pi}_\beta} \left[\widehat{\psi}_\pi + \frac{\widehat{\kappa}_\pi z_{\alpha,\beta}}{n^{1/2}} \right], \Pi^* \subseteq \widehat{\Pi}_\beta \right\} \leq (\alpha - \beta)/2$. Therefore, an asymptotic $1 - \alpha$ confidence interval for $[\psi_0^l, \psi_0^u]$ is

$$\left[\inf_{\pi \in \widehat{\Pi}_\beta} \left\{ \widehat{\psi}_\pi - \frac{\widehat{\kappa}_\pi z_{\alpha,\beta}}{n^{1/2}} \right\}, \sup_{\pi \in \widehat{\Pi}_\beta} \left\{ \widehat{\psi}_\pi + \frac{\widehat{\kappa}_\pi z_{\alpha,\beta}}{n^{1/2}} \right\} \right].$$

□

Proof of Lemma 4.3.3. To show this lemma, we first define two events $\{\Pi^* \subseteq \widehat{\Pi}_\beta\}$ and $\{\omega_{\pi^*} - \inf_{\pi \in \widehat{\Pi}_\beta} \omega_\pi \leq \frac{4 t_\beta}{n^{1/2}} \sup_{\pi \in \Pi} \widehat{\sigma}_\pi\}$. These events ensure that all Ω-optimal policies are contained in $\widehat{\Pi}_\beta$, and $\widehat{\Pi}_\beta$ only contains nearly optimal policies. Lemma 4.3.1 and C.2.2 ensure that both events happen with probability at least $1 - \beta$ asymptotically. The lemma below ensures that our confidence interval shrinks at an $n^{-1/2}$ rate under these events. □

Lemma C.2.6. *In the setting of Lemma 4.3.3, under the event* $\{\Pi^* \subseteq \widehat{\Pi}_\beta\}$ *and* $\{\omega_{\pi^*} - \inf_{\pi \in \widehat{\Pi}_\beta} \omega_\pi \leq \frac{4 t_\beta}{n^{1/2}} \sup_{\pi \in \Pi} \widehat{\sigma}_\pi\}$*, the width of the confidence interval for* ψ_0 *is* $O_p(n^{-1/2})$.

Proof. We first show that $\sup_{\pi\in\widehat{\Pi}_\beta}\left[\widehat{\psi}_\pi+\frac{\widehat{\kappa}_\pi z_{\alpha,\beta}}{n^{1/2}}\right]=\psi_0+O_p(n^{-1/2})$. We know that

$$\sup_{\pi\in\widehat{\Pi}_\beta}\left[\widehat{\psi}_\pi+\frac{\widehat{\kappa}_\pi z_{\alpha,\beta}}{n^{1/2}}\right]\leq\sup_{\pi\in\widehat{\Pi}_\beta}\psi_\pi+\sup_{\pi\in\widehat{\Pi}_\beta}\left[\widehat{\psi}_\pi-\psi_\pi+\frac{\widehat{\kappa}_\pi z_{\alpha,\beta}}{n^{1/2}}\right].$$

We then show that $\sup_{\pi\in\Pi^*}\psi_\pi-\sup_{\pi\in\widehat{\Pi}_\beta}\psi_\pi=O_p(n^{-1/2})$. Consider some $\pi_1\in\Pi^*$ and $\pi_2\in\widehat{\Pi}_\beta$. Let $B_{1,0}=\{x\in\mathcal{X}:\pi_1(x)=1,\pi_2(x)=0\}$ and $B_{0,1}=\{x\in\mathcal{X}:\pi_1(x)=0,\pi_2(x)=1\}$. By the definition of Π^* we know that $\omega_{\pi_1}\geq\omega_{\pi_2}$, and

$$\begin{aligned}\omega_{\pi_1}-\omega_{\pi_2}&=\int\mathbb{E}[Y^*|A=\pi_1(x),x]dP_0(x)-\int\mathbb{E}[Y^*|A=\pi_2(x),x]dP_0(x)\\&=\int_{B_{1,0}}q_{b,0}(x)dP_0(x)-\int_{B_{0,1}}q_{b,0}(x)dP_0(x).\end{aligned}$$

Since $\pi_1\in\Pi^*$ and Π^* contains unrestricted optimal policies by assumption, ω_{π_1} is largest among all $\pi\in\Pi$, which implies that for $x\in B_{1,0}$, $q_{b,0}(x)\geq 0$ and for $x\in B_{0,1}$, $q_{b,0}(x)\leq 0$. This gives us

$$\omega_{\pi_1}-\omega_{\pi_2}=\int_{B_{1,0}}|q_{b,0}(x)|dP_0(x)+\int_{B_{0,1}}|q_{b,0}(x)|dP_0(x).$$

On the other hand, on the event $\{\Pi^*\subseteq\widehat{\Pi}_\beta\}$, we have $\sup_{\pi\in\widehat{\Pi}_\beta}\psi_\pi\geq\sup_{\pi\in\Pi^*}\psi_\pi$, and

$$\begin{aligned}|\psi_{\pi_2}-\psi_{\pi_1}|&=\left|\int\mathbb{E}[Y^\dagger|A=\pi_2(x),x]dP_0(x)-\int\mathbb{E}[Y^\dagger|A=\pi_1(x),x]dP_0(x)\right|\\&=\left|\int_{B_{0,1}}s_{b,0}(x)dP_0(x)-\int_{B_{1,0}}s_{b,0}(x)dP_0(x)\right|\\&\leq\int_{B_{1,0}}|s_{b,0}(x)|dP_0(x)+\int_{B_{0,1}}|s_{b,0}(x)|dP_0(x)\\&\leq C\int_{B_{1,0}}|q_{b,0}(x)|dP_0(x)+C\int_{B_{0,1}}|q_{b,0}(x)|dP_0(x).\end{aligned}$$

Therefore, $|\psi_{\pi_2}-\psi_{\pi_1}|\leq C(\omega_{\pi_1}-\omega_{\pi_2})$ for some $C<\infty$. Since this holds for any $\pi_1\in\Pi^*$ and $\pi_2\in\widehat{\Pi}_\beta$, we have $\sup_{\pi\in\widehat{\Pi}_\beta}\psi_\pi-\inf_{\pi\in\Pi^*}\psi_\pi\leq C(\sup_{\pi\in\Pi^*}\omega_\pi-\inf_{\pi\in\widehat{\Pi}_\beta}\omega_\pi)$. Under the event $\{\omega_{\pi^*}-\inf_{\pi\in\widehat{\Pi}_\beta}\omega_\pi\leq\frac{4t_\beta}{n^{1/2}}\sup_{\pi\in\Pi}\widehat{\sigma}_\pi\}$, we have that $\sup_{\pi\in\widehat{\Pi}_\beta}\psi_\pi-\inf_{\pi\in\Pi^*}\psi_\pi\leq C\frac{4t_\beta}{n^{1/2}}\sup_{\pi\in\Pi}\widehat{\sigma}_\pi$. Under Condition 7, we know that $\sup_{\pi\in\Pi}\widehat{\sigma}_\pi-\sup_{\pi\in\Pi}\sigma_\pi(P_0)=o_p(1)$, so

$$\sup_{\pi\in\widehat{\Pi}_\beta}\psi_\pi-\inf_{\pi\in\Pi^*}\psi_\pi\leq C\frac{4t_\beta}{n^{1/2}}\sup_{\pi\in\Pi}\widehat{\sigma}_\pi=C\frac{4t_\beta}{n^{1/2}}\left(\sup_{\pi\in\Pi}\sigma_\pi(P_0)+o_p(n^{-1/2})\right)=O_p(n^{-1/2}).\quad\text{(C.8)}$$

Under the event $\{\Pi^* \subseteq \widehat{\Pi}_\beta\}$, we have $\sup_{\pi\in\widehat{\Pi}_\beta} \psi_\pi \geq \sup_{\pi\in\Pi^*} \psi_\pi \geq \inf_{\pi\in\Pi^*} \psi_\pi$, so we have $\sup_{\pi\in\Pi^*} \psi_\pi - \sup_{\pi\in\widehat{\Pi}_\beta} \psi_\pi = O_p(n^{-1/2})$. Also,

$$\sup_{\pi\in\widehat{\Pi}_\beta} \left[\widehat{\psi}_\pi - \psi_\pi + \frac{\widehat{\kappa}_\pi z_{\alpha,\beta}}{n^{1/2}}\right] \leq \sup_{\pi\in\Pi} \left[\widehat{\psi}_\pi - \psi_\pi + \frac{\widehat{\kappa}_\pi z_{\alpha,\beta}}{n^{1/2}}\right] \leq \sup_{\pi\in\Pi} \left[\widehat{\psi}_\pi - \psi_\pi\right] + \sup_{\pi\in\Pi} \frac{\widehat{\kappa}_\pi z_{\alpha,\beta}}{n^{1/2}}.$$

The first term is $O_p(n^{-1/2})$ under Condition 5. As for the second term, under Condition 7,

$$\sup_{\pi\in\Pi} \frac{\widehat{\kappa}_\pi z_{\alpha,\beta}}{n^{1/2}} = \sup_{\pi\in\Pi} \frac{\kappa_\pi(P_0) z_{\alpha,\beta}}{n^{1/2}} + o_p(n^{-1/2}) = O_p(n^{-1/2}).$$

Therefore, $\sup_{\pi\in\widehat{\Pi}_\beta} \left[\widehat{\psi}_\pi - \psi_\pi + \frac{\widehat{\kappa}_\pi z_{\alpha,\beta}}{n^{1/2}}\right] = O_p(n^{-1/2})$ and so $\sup_{\pi\in\widehat{\Pi}_\beta} \left[\widehat{\psi}_\pi + \frac{\widehat{\kappa}_\pi z_{\alpha,\beta}}{n^{1/2}}\right] = \psi_0^u + O_p(n^{-1/2})$ as desired. By symmetry, $\inf_{\pi\in\widehat{\Pi}_\beta} \left[\widehat{\psi}_\pi - \frac{\widehat{\kappa}_\pi z_{\alpha,\beta}}{n^{1/2}}\right] = \psi_0 - O_p(n^{-1/2})$ as well. □

Proof of Theorem 4.3.4. To establish this theorem, we show that

$$\liminf_n \mathbb{P}\left(\sup_{\pi\in\Pi^*} \psi_\pi \leq \sup_{\pi\in\widehat{\Pi}^\dagger} \left[\widehat{\psi}_\pi + \frac{\widehat{\kappa}_\pi u_\alpha^\dagger}{n^{1/2}}\right]\right) \geq 1 - \alpha/2.$$

We can similarly get

$$\liminf_n \mathbb{P}\left(\inf_{\pi\in\Pi^*} \psi_\pi \geq \inf_{\pi\in\widehat{\Pi}^\dagger} \left[\widehat{\psi}_\pi - \frac{\widehat{\kappa}_\pi u_\alpha^\dagger}{n^{1/2}}\right]\right) \geq 1 - \alpha/2.$$

Combining the two displays gives us the theorem statement. Note that

$$\begin{aligned}
&\left\{\sup_{\pi\in\Pi^*} \psi_\pi \leq \sup_{\pi\in\widehat{\Pi}^\dagger} \left[\widehat{\psi}_\pi + \frac{\widehat{\kappa}_\pi u_\alpha^\dagger}{n^{1/2}}\right]\right\} \\
&\supseteq \left\{\sup_{\pi\in\Pi^*} \psi_\pi \leq \sup_{\pi\in\widehat{\Pi}^\dagger} \left[\widehat{\psi}_\pi + \frac{\widehat{\kappa}_\pi u_\alpha^\dagger}{n^{1/2}}\right], \Pi^* \subseteq \widehat{\Pi}^\dagger\right\}.
\end{aligned}$$

Since Π^* is P_0-Donsker following from Π being P_0-Donsker, Π^* is totally bounded in $L^2(P_0)$ Luedtke and Van Der Laan [2016]. Also, since $L^2(P_0)$ is complete, Π^* being closed in $L^2(P_0)$ implies that Π^* is complete in $L^2(P_0)$. So Π^* is compact in $L^2(P_0)$. Combining this with the fact that $\pi \mapsto \psi_\pi$ is continuous implies that there exists a $\pi^u \in \Pi^*$ such that $\psi_{\pi^u} = \sup_{\pi\in\Pi^*} \psi_\pi$.

Combining this with the above, we see that

$$\begin{aligned}
&\left\{\sup_{\pi\in\Pi^*}\psi_\pi \le \sup_{\pi\in\widehat{\Pi}^\dagger}\left[\widehat{\psi}_\pi + \frac{\widehat{\kappa}_\pi u_\alpha^\dagger}{n^{1/2}}\right]\right\} \\
&\supseteq \left\{\psi_{\pi^u} \le \sup_{\pi\in\widehat{\Pi}^\dagger}\left[\widehat{\psi}_\pi + \frac{\widehat{\kappa}_\pi u_\alpha^\dagger}{n^{1/2}}\right], \Pi^* \subseteq \widehat{\Pi}^\dagger\right\} \\
&\supseteq \left\{\psi_{\pi^u} \le \widehat{\psi}_{\pi^u} + \frac{\widehat{\kappa}_{\pi^u} u_\alpha^\dagger}{n^{1/2}}, \Pi^* \subseteq \widehat{\Pi}^\dagger\right\} \\
&= \left\{\psi_{\pi^u} \le \widehat{\psi}_{\pi^u} + \frac{\widehat{\kappa}_{\pi^u} u_\alpha^\dagger}{n^{1/2}}, \omega_{\pi'} < \sup_{\pi\in\Pi}\omega_\pi, \forall \pi' \in (\widehat{\Pi}^\dagger)^C\right\}. \qquad \text{(C.9)}
\end{aligned}$$

Note that

$$\begin{aligned}
&\left\{\omega_{\pi'} < \sup_{\pi\in\Pi}\omega_\pi, \forall \pi' \in (\widehat{\Pi}^\dagger)^C\right\}^C = \left\{\exists \pi' \in (\widehat{\Pi}^\dagger)^C : \omega_{\pi'} = \sup_{\pi\in\Pi}\omega_\pi\right\} \\
&\subseteq \left\{\exists \pi' \in (\widehat{\Pi}^\dagger)^C : \left[\omega_{\pi'} - \widehat{\omega}_{\pi'} - \frac{\widehat{\sigma}_{\pi'} t_\alpha^\dagger}{n^{1/2}} + \sup_{\pi\in\Pi}\left[\widehat{\omega}_\pi - \frac{\widehat{\sigma}_\pi s_\alpha^\dagger}{n^{1/2}}\right]\right] > \sup_{\pi\in\Pi}\omega_\pi\right\} \\
&= \left\{\exists \pi' \in (\widehat{\Pi}^\dagger)^C : \left[\omega_{\pi'} - \widehat{\omega}_{\pi'} - \frac{\widehat{\sigma}_{\pi'} t_\alpha^\dagger}{n^{1/2}}\right] > \sup_{\pi\in\Pi}\omega_\pi - \sup_{\pi\in\Pi}\left[\widehat{\omega}_\pi - \frac{\widehat{\sigma}_\pi s_\alpha^\dagger}{n^{1/2}}\right]\right\}, \qquad \text{(C.10)}
\end{aligned}$$

where the inclusion follows from the definition of $\widehat{\Pi}^\dagger$. Let $\mathcal{A}'$ denote the event

$$\left\{\sup_{\pi\in\Pi}\left[\widehat{\omega}_\pi - \frac{\widehat{\sigma}_\pi s_\alpha^\dagger}{n^{1/2}}\right] \le \sup_{\pi\in\Pi}\omega_\pi\right\} \bigcap \left[\bigcap_{\pi\in\Pi}\left\{\omega_\pi \le \widehat{\omega}_\pi + \frac{\widehat{\sigma}_\pi t_\alpha^\dagger}{n^{1/2}}\right\}\right].$$

Hence, (C.10) shows that

$$\begin{aligned}
&\left\{\exists \pi' \in (\widehat{\Pi}^\dagger)^C : \omega_{\pi'} = \sup_{\pi\in\Pi}\omega_\pi\right\} \\
&\subseteq \left[\left\{\exists \pi' \in (\widehat{\Pi}^\dagger)^C : \omega_{\pi'} - \widehat{\omega}_{\pi'} - \frac{\widehat{\sigma}_{\pi'} t_\alpha^\dagger}{n^{1/2}} > \sup_{\pi\in\Pi}\omega_\pi - \sup_{\pi\in\Pi}\left[\widehat{\omega}_\pi - \frac{\widehat{\sigma}_\pi s_\alpha^\dagger}{n^{1/2}}\right]\right\} \cap \mathcal{A}'\right] \cup \mathcal{A}'^C \\
&\subseteq \left[\left\{\exists \pi' \in (\widehat{\Pi}^\dagger)^C : \omega_{\pi'} - \widehat{\omega}_{\pi'} - \frac{\widehat{\sigma}_{\pi'} t_\alpha^\dagger}{n^{1/2}} > 0\right\} \cap \mathcal{A}'\right] \cup \mathcal{A}'^C \\
&= \mathcal{A}'^C.
\end{aligned}$$

For each $\pi \in \Pi$, we define $\widehat{B}_{n,\pi} := n^{1/2}\frac{\widehat{\omega}_\pi - \omega_\pi}{\widehat{\sigma}_\pi}$ and $\widetilde{B}_{n,\pi} := n^{1/2}\frac{\widehat{\psi}_\pi - \psi_\pi}{\widehat{\kappa}_\pi}$. Then starting from

(C.10), we have

$$\left\{\omega_{\pi'} < \sup_{\pi\in\Pi}\omega_\pi, \forall\pi' \in (\widehat{\Pi}^\dagger)^C\right\} \supseteq \mathcal{A}' = \left\{\sup_{\pi\in\Pi}\left[\widehat{\omega}_\pi - \frac{\widehat{\sigma}_\pi s_\alpha^\dagger}{n^{1/2}}\right] < \sup_{\pi\in\Pi}\omega_\pi\right\} \bigcap \left[\bigcap_{\pi\in\Pi}\left\{\omega_\pi \le \widehat{\omega}_\pi + \frac{\widehat{\sigma}_\pi t_\alpha^\dagger}{n^{1/2}}\right\}\right]$$

$$\begin{aligned}
&\supseteq \bigcap_{\pi\in\Pi}\left\{\widehat{\omega}_\pi - \frac{\widehat{\sigma}_\pi s_\alpha^\dagger}{n^{1/2}} < \omega_\pi < \widehat{\omega}_\pi + \frac{\widehat{\sigma}_\pi t_\alpha^\dagger}{n^{1/2}}\right\} \\
&= \bigcap_{\pi\in\Pi}\left\{-t_\alpha^\dagger < n^{1/2}\frac{\widehat{\omega}_\pi - \omega_\pi}{\widehat{\sigma}_\pi} < s_\alpha^\dagger\right\} \\
&= \bigcap_{\pi\in\Pi}\left\{-t_\alpha^\dagger < B_{n,\pi} < s_\alpha^\dagger\right\} \\
&= \{-t_\alpha^\dagger \le \inf_{\pi\in\Pi} B_{n,\pi}\} \cap \{\sup_{\pi\in\Pi} B_{n,\pi} \le s_\alpha^\dagger\}.
\end{aligned}$$

Using the above to study the event on the right-hand side of (C.9) shows that

$$\begin{aligned}
&\left\{\psi_{\pi^u} \le \widehat{\psi}_{\pi^u} + \frac{\widehat{\kappa}_{\pi^u} u_\alpha^\dagger}{n^{1/2}}, \omega_{\pi'} < \sup_{\pi\in\Pi}\omega_\pi, \forall\pi' \in (\widehat{\Pi}^\dagger)^C\right\} \\
&\supseteq \left\{\psi_{\pi^u} < \widehat{\psi}_{\pi^u} + \frac{\widehat{\kappa}_{\pi^u} u_\alpha^\dagger}{n^{1/2}}, -t_\alpha^\dagger \le \inf_{\pi\in\Pi} B_{n,\pi}, \sup_{\pi\in\Pi} B_{n,\pi} \le s_\alpha^\dagger\right\} \\
&= \left\{\widetilde{B}_{n,\pi^u} > -u_\alpha^\dagger, -t_\alpha^\dagger \le \inf_{\pi\in\Pi} B_{n,\pi}, \sup_{\pi\in\Pi} B_{n,\pi} \le s_\alpha^\dagger\right\}. \qquad \text{(C.11)}
\end{aligned}$$

We know that the choices $(s_\alpha^\dagger, t_\alpha^\dagger, u_\alpha^\dagger)$ satisfy that

$$\inf_{\pi\in\Pi}\mathbb{P}\left\{\inf_{f\in\mathcal{F}}\mathbb{G}f \ge -t_\alpha^\dagger, \sup_{f\in\mathcal{F}}\mathbb{G}f \le s_\alpha^\dagger, \mathbb{G}\tilde{f}_\pi \ge -u_\alpha^\dagger\right\} \ge 1 - \alpha/2. \qquad \text{(C.12)}$$

Note that by Condition 5, we have $\sup_{\pi\in\Pi}\left[n^{1/2}\frac{\widehat{\omega}_\pi - \omega_\pi}{\widehat{\sigma}_\pi} - \mathbb{G}_n f_\pi\right] = o_p(1)$ and also $\frac{\widehat{\psi}_{\pi^u} - \psi_{\pi^u}}{\widehat{\kappa}_{\pi^u}} - \mathbb{G}_n\tilde{f}_{\pi^u} = o_p(1)$. Since $\sup_{f\in\mathcal{F}}\mathbb{G}_n f \rightsquigarrow \sup_{f\in\mathcal{F}}\mathbb{G}f$, $\inf_{f\in\mathcal{F}}\mathbb{G}_n f \rightsquigarrow \inf_{f\in\mathcal{F}}\mathbb{G}f$, and for each $\pi \in \Pi$, $\widehat{\sigma}_\pi$ is a consistent estimator of σ_π, by Slutsky Theorem, we have $\sup_{\pi\in\Pi} B_{n,\pi} \rightsquigarrow \sup_{f\in\mathcal{F}}\mathbb{G}f$ and $\inf_{\pi\in\Pi} B_{n,\pi} \rightsquigarrow \inf_{f\in\mathcal{F}}\mathbb{G}f$. Also, since for each $\tilde{f} \in \tilde{\mathcal{F}}$, $\mathbb{G}_n\tilde{f} \rightsquigarrow \mathbb{G}\tilde{f}$ and $\widehat{\sigma}_\pi$ is a consistent estimator

of σ_π, we similarly have $\widetilde{B}_{n,\pi^u} \rightsquigarrow \mathbb{G}\tilde{f}_{\pi^u}$. Combining (C.9), (C.11), and (C.12), we have

$$\begin{aligned}
&\liminf_n \mathbb{P}\left(\sup_{\pi\in\Pi^*}\psi_\pi < \sup_{\pi\in\widehat{\Pi}}\left[\widehat{\psi}_\pi + \frac{\widehat{\kappa}_\pi u_\alpha^\dagger}{n^{1/2}}\right]\right) \\
&\geq \liminf_n \mathbb{P}\left(\widetilde{B}_{n,\pi^u} < u_\alpha^\dagger, -s_\alpha^\dagger < \inf_{\pi\in\Pi} B_{n,\pi}, \sup_{\pi\in\Pi} B_{n,\pi} < t_\alpha^\dagger\right) \\
&\to \mathbb{P}\left(\mathbb{G}\tilde{f}_{\pi^u} < u_\alpha^\dagger, -s_\alpha^\dagger < \inf_{f\in\mathcal{F}}\mathbb{G}f, \sup_{f\in\mathcal{F}}\mathbb{G}f < t_\alpha^\dagger\right) \\
&\geq \inf_{\pi\in\Pi}\mathbb{P}\left(\mathbb{G}\tilde{f}_\pi < u_\alpha^\dagger, -s_\alpha^\dagger < \inf_{f\in\mathcal{F}}\mathbb{G}f, \sup_{f\in\mathcal{F}}\mathbb{G}f < t_\alpha^\dagger\right) = 1-\alpha/2.
\end{aligned}$$

□

C.3 Multiplier bootstrap

In practice, we use multiplier bootstrap Chernozhukov et al. [2013] to estimate the quantiles described in Section 4.3 and we provide the pseudocodes of the algorithms below. Algorithm 10 estimates t_β defined just above Lemma 4.3.1. Algorithm 11 estimates the quantiles described in (4.6). In this algorithm, we take $s_\alpha^\dagger = t_\alpha^\dagger$ for simplicity and estimate the best $(t_\alpha^\dagger, u_\alpha^\dagger)$ given samples. Both algorithms approximate suprema and infima over sets indexed by $\pi \in \Pi$ by maxima and minima over π belonging to a grid approximation of Π.

Algorithm 10 Multiplier bootstrap

Input: samples $\{(x_i, a_i, y_i)\}_{i=1}^n$, policy set Π, bootstrap sample size B, confidence level β

Take a grid estimate $\{\pi_1, \cdots, \pi_K\}$ of Π

for each $k \in [K]$, compute normalized one-step estimates $\{o_i^{(\pi_k)}\}_{i=1}^n$ using collected samples $\{(x_i, a_i, y_i)\}_{i=1}^n$

for $j = 1, \cdots, B$ **do**

    get multiplier bootstrap samples ϵ_{ij} for $i = 1, \cdots, n$ and $k = 1, \cdots, K$

    compute $n^{-1/2} \sum_{i=1}^n \epsilon_{ij} o_i^{(\pi_k)}$ and denote the result as $f_{\pi_k}^{(j)}$

end for

Apply quantile normalization to $f_{\pi_k}^{(j)}$ across j for each policy π_k

compute $\max_{k \in [K]} f_{\pi_k}^{(j)}$ for each j and denote the resulting dataset as $\{t_i\}_{i=1}^B$

Output: $(1-\beta)$-th quantile of $\{t_i\}_{i=1}^B$

Algorithm 11 Multiplier bootstrap for joint probability

Input: samples $\{(x_i, a_i, y_i, z_i)\}_{i=1}^n$, policy set Π, bootstrap sample size B, confidence level α

Take a grid estimate $\{\pi_1, \cdots, \pi_K\}$ of Π

for $k \in [K]$ **do**

    compute normalized one-step estimates $\{o_i^{(\pi_k)}\}_{i=1}^n$ using collected samples $\{(x_i, a_i, y_i)\}_{i=1}^n$

    compute normalized one-step estimates $\{\tilde{o}_i^{(\pi_k)}\}_{i=1}^n$ using collected samples $\{(x_i, a_i, z_i)\}_{i=1}^n$

end for

for $j = 1, \cdots, B$ **do**

    get multiplier bootstrap samples $\epsilon_{ik}^{(j)}$ for $i = 1, \cdots, n$ and $k = 1, \cdots, K$

    compute $n^{-1/2} \sum_{i=1}^n \epsilon_{ik}^{(j)} o_i^{(\pi_k)}$ and denote the result as $f_{\pi_k}^{(j)}$

    compute $n^{-1/2} \sum_{i=1}^n \epsilon_{ik}^{(j)} \tilde{o}_i^{(\pi_k)}$ and denote the result as $\tilde{f}_{\pi_k}^{(j)}$

end for

Apply quantile normalization to $f_{\pi_k}^{(j)}$ and $\tilde{f}_{\pi_k}^{(j)}$ across j for each policy π_k

compute $\max_{k \in [K]} f_{\pi_k}^{(j)}$ for each j and denote the results as $\{s_j\}_{j=1}^B$

compute probability $\mathbb{P}(\max_{k \in [K]} f_{\pi_k} \le t, \tilde{f}_{\pi_k} \le u)$ for each $k = 1, \cdots, K$ using the B samples

Output: pairs (t, u) such that $\min_{k \in [K]} \mathbb{P}(\max_{k \in [K]} f_{\pi_k} \le t, \tilde{f}_{\pi_k} \le u) = 1 - \alpha$.