Exploration and Primal-dual Methods in Bandits and Reinforcement Learning

Zhihan Xiong

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

Doctor of Philosophy

University of Washington

2025

Reading Committee:

Maryam Fazel, Chair

Kevin Jamieson

Lin Xiao

Pang Wei Koh

Program Authorized to Offer Degree:
Paul G. Allen School of Computer Science & Engineering

University of Washington

Abstract

Exploration and Primal-dual Methods in Bandits and Reinforcement Learning

Zhihan Xiong

Chair of the Supervisory Committee:
Maryam Fazel
Department of Electrical & Computer Engineering

Sequential decision-making, which encompasses both bandit problems and reinforcement learning, forms the foundation of intelligent systems across diverse applications, from adaptive recommendation systems to autonomous robotics. This thesis addresses two fundamental challenges in building reliable, sample-efficient agents that operate robustly in dynamic, complex environments: efficient exploration in non-stationary or structurally complex settings, and the design of appropriate objective functions when multiple approximation layers are inevitable.

Regarding the efficient exploration, we develop the first robust pure exploration algorithm for both stationary and non-stationary linear bandits, achieving strong performance in benign settings while maintaining robustness to environmental changes. For single-step congestion games, we exploit the structure of this special class of games to develop the first algorithms for Nash equilibrium learning under various feedback models. For tabular reinforcement learning, we propose the first near-optimal randomized exploration algorithm that nearly matches the fundamental lower bound.

Regarding the objective design, we analyze learning objectives through the lens of duality between value learning and policy learning. In an online selective sampling problem for linear bandits, we characterize an optimal ellipsoid-based selection rule through primal-dual analysis. For approximate policy optimization, we propose using dual Bregman divergence instead of the common Euclidean norm to measure similarity in dual space, resulting in

the first policy optimization framework with both fast theoretical convergence and superior practical performance.

Collectively, these contributions advance the theoretical frontier of exploration and objective design, close several open complexity gaps, and provide practical algorithms validated on robotic control benchmarks. They offer a principled route towards agents that learn robustly and act reliably in dynamic, high-dimensional environments.

TABLE OF CONTENTS

Page

LIST OF FIGURES

ACKNOWLEDGMENTS

It feels like yesterday that I opened the email with my Ph.D. offer—equal parts joy and fear. Five years later, as I type these words, the journey that once felt like a dream has become real: I am about to receive my Ph.D.

First and foremost, I owe my deepest gratitude to my advisor, Prof. Maryam Fazel, for her unwavering support and guidance. Maryam's wealth of expertise in optimization continually pushed me to seek deeper insight into every problem we studied. Beyond countless hours of discussion, she also connected me to a broader community of experienced researchers—including Prof. Simon S. Du, Lalit Jain, Kevin Jamieson, and Lin Xiao—whose conversations and feedback profoundly shaped my work. I am especially grateful to Lin for his mentorship during the two-year Meta AI Mentorship Program and for our continued collaboration thereafter. That experience strengthened my research and opened doors to cutting-edge industry work, paving the way for my future return to Meta.

Collaboration was the heartbeat of this Ph.D. I was lucky to work with wonderful colleagues: Romain Camilleri, Qiwen Cui, Avinandan Bose, Haozhe Jiang, Ruoqi Shen, Aadirupa Saha, Yuejie Chi, Jieyu Zhang, Yingxiang Yang, Tianyi Liu, Taiqing Wang, and Chong Wang. Thanks a lot for the ideas, patience, and persistence we shared.

Beyond research, my friends made the hard days lighter and the good days unforgettable. To my learning theory crew—Yifang Chen, Qiwen Cui, Daogao Liu, Ruoqi Shen, and Runlong Zhou—thank you for the daily chats over apps and dinners, about everything inside and outside academia. To my tabletop game circle—Tuochao Chen, Benlin Liu, Jingwei Ma, Mengyi Shan, Yitong Shan, Yilun Sheng, Han Wu, Wu Han, Shuangning Li, Shuangping Li, Lingfu Zhang, Daren Chen, and Yuchen Wu—thank you for countless hours of Avalon, One Night Werewolf, and Blood on the Clocktower. To my travel companions—Qiong Wang, Ziying Wang, and Yuchen Xu—thank you for the miles we covered, from Olympic National

Park to Albuquerque and Houston. I am especially grateful to Yifang Chen, Lei Chen, Yilun Sheng, and Qile Zhi for their generous help during my job search; to Yinchen Xu and Xiujun Li for welcoming me into their homes for extended stays and easing my financial burden; and to my long-time friends Yizhi Qiao and Ziying Wang for our day-to-day conversations and shared, "unpopular" hobbies—a space free from the usual worries of reality.

Last, but never least, I thank my parents—for caring more about my well-being than my awards, and for reminding me that a happy life matters more than any achievement. None of this would have been possible without their constant, unconditional love.

This dissertation is not only a record of research; it is the accumulation of countless moments and people who shaped me over these five years—something not measured, but gathered. I dedicate it to all of you. And as one small part of the human current, I will keep carrying our time marching into the future.

Chapter 1

INTRODUCTION

Bandit problems and reinforcement learning–collectively referred to as sequential decision making–form the backbone of intelligent systems across numerous domains, from recommendation systems that adapt to user preferences over time to autonomous robots navigating complex environments. At its core, sequential decision making involves an agent making a series of decisions over time, where each decision influences future states and available actions. Understanding the theoretical principles that govern learning, exploration, and decision-making in such environments, while designing algorithms that are both provably efficient and practically deployable at scale, is therefore a central challenge of contemporary machine learning research.

This thesis confronts that challenge through a unified study that ranges from single-step decision problems (bandits) to multi-step planning problems (reinforcement learning). In particular, we focus on the following two high-level questions.

- **Q1**: *How can an agent learn and explore efficiently in non-stationary or structurally complex environments?*

- **Q2**: *What are the right objective functions—both statistically and computationally—for learning the optimal policy when there are multiple layers of approximation?*

Q1 — Efficient exploration. The prevailing theory of exploration in stationary, "simple" environments relies on optimism-in-the-face-of-uncertainty principles, which encourage agents to visit rarely encountered states and deliver near-optimal guarantees [Auer et al., 2008, Azar et al., 2017]. Extending these ideas to non-stationary or structurally rich environments, however, is far from straightforward. This thesis develops new algorithms that retain strong theoretical performance while coping with temporal drift, abrupt shifts, and

intricate state-action structure.

Q2 — Primal–dual learning objectives. An agent's performance is equally shaped by the objective it optimizes. For this question, the lens through the duality between policy learning and value learning offers a viewpoint that departs from traditional treatments of bandits and RL [Nachum and Dai, 2020]. When multiple approximation layers intervene, errors in the primal and dual spaces interact; therefore, blindly adopting standard L_2 losses can sometimes be detrimental. However, this interaction was often overlooked in prior work, spanning from traditional reinforcement learning [Schulman et al., 2017, Liu et al., 2019, Lan, 2023, Alfano et al., 2024] to the currently popular reinforcement learning from human feedback (RLHF) methods [Ouyang et al., 2022], whose training stability and performance can potentially be improved by implementing more appropriate objectives. To address the gap, this thesis designs and analyzes loss functions grounded in primal–dual methodology across several settings.

In summary, the overarching goal is to develop a systematic methodology for building reliable, sample-efficient agents that operate robustly in dynamic, complex environments.

1.1 Efficient Exploration in Bandits and Reinforcement Learning.

In this part, we study pure exploration problem for non-stationary linear bandits [Xiong et al., 2024a]. In particular, we investigate the fixed-budget best-arm identification (BAI) problem for linear bandits in a potentially non-stationary environment. Given a finite arm set $\mathcal{X} \subset \mathbb{R}^d$, a fixed budget T, and an unpredictable sequence of parameters $\{\theta_t\}_{t=1}^T$, an algorithm will aim to correctly identify the best arm $x^* := \arg\max_{x\in\mathcal{X}} x^\top \sum_{t=1}^T \theta_t$ with probability as high as possible. Prior work has addressed the stationary setting where $\theta_t = \theta_1$ for all t and demonstrated that the error probability decreases as $\exp(-T/\rho^*)$ for a problem-dependent constant ρ^*. But in many real-world $A/B/n$ multivariate testing scenarios that motivate our work, the environment is non-stationary and an algorithm expecting a stationary setting can easily fail. For robust identification, it is well-known that if arms are chosen randomly and non-adaptively from a G-optimal design over $\mathcal{X}$ at each time then the error probability decreases as $\exp(-T\Delta_{(1)}^2/d)$, where $\Delta_{(1)} = \min_{x\neq x^*}(x^* - x)^\top \frac{1}{T}\sum_{t=1}^T \theta_t$.

As there exist environments where $\Delta_{(1)}^2/d \ll 1/\rho^*$, we are motivated to propose a novel algorithm P1-RAGE that aims to obtain the best of both worlds: robustness to non-stationarity and fast rates of identification in benign settings. We characterize the error probability of P1-RAGE and demonstrate empirically that the algorithm indeed never performs worse than G-optimal design but compares favorably to the best algorithms in the stationary setting.

Meanwhile, we design learning and exploration algorithms in single-step congestion games [Cui et al., 2022]. Specifically, we investigate Nash-regret minimization in congestion games, a class of games with benign theoretical structure and broad real-world applications. We first propose a centralized algorithm based on the optimism in the face of uncertainty principle for congestion games with (semi-)bandit feedback, and obtain finite-sample guarantees. Then we propose a decentralized algorithm via a novel combination of the Frank-Wolfe method and G-optimal design. By exploiting the structure of the congestion game, we show the sample complexity of both algorithms depends only polynomially on the number of players and the number of facilities, but not the size of the action set, which can be exponentially large in terms of the number of facilities. We further define a new problem class, Markov congestion games, which allows us to model the non-stationarity in congestion games. We propose a centralized algorithm for Markov congestion games, whose sample complexity again has only polynomial dependence on all relevant problem parameters, but not the size of the action set.

Finally, we propose a near-optimal random exploration strategy in tabular Markov Decision Processes (MDPs) [Xiong et al., 2022]. Concretely, We study algorithms using randomized value functions for exploration in reinforcement learning. This type of algorithms enjoys appealing empirical performance. We show that when we use 1) a single random seed in each episode, and 2) a Bernstein-type magnitude of noise, we obtain a worst-case $\widetilde{O}\left(H\sqrt{SAT}\right)$ regret bound for episodic time-inhomogeneous Markov Decision Process where S is the size of state space, A is the size of action space, H is the planning horizon and T is the number of interactions. This bound polynomially improves all existing bounds for algorithms based on randomized value functions, and for the first time, matches the $\Omega\left(H\sqrt{SAT}\right)$ lower bound up to logarithmic factors. Our result highlights that randomized exploration can be near-optimal, which was previously achieved only by optimistic algorithms. To achieve the

desired result, we develop 1) a new clipping operation to ensure both the probability of being optimistic and the probability of being pessimistic are lower bounded by a constant, and 2) a new recursive formula for the absolute value of estimation errors to analyze the regret.

1.2 Primal-dual Methods in Bandits and Reinforcement Learning.

In this part, we first apply primal-dual methods to address an online selective sampling problem in bandits [Camilleri et al., 2021b]. It considers the problem of *selective-sampling for best-arm identification*. Given a set of potential options $\mathcal{Z} \subset \mathbb{R}^d$, a learner aims to compute with probability greater than $1-\delta$, $\arg\max_{z\in\mathcal{Z}} z^\top \theta_*$ where θ_* is unknown. At each time step, a potential measurement $x_t \in \mathcal{X} \subset \mathbb{R}^d$ is drawn IID and the learner can either choose to take the measurement, in which case they observe a noisy measurement of $x^\top \theta_*$, or to abstain from taking the measurement and wait for a potentially more informative point to arrive in the stream. Hence the learner faces a fundamental trade-off between the number of labeled samples they take and when they have collected enough evidence to declare the best arm and stop sampling. The main results of this work precisely characterize this trade-off between labeled samples and stopping time and provide an algorithm that nearly-optimally achieves the minimal label complexity given a desired stopping time. In addition, by using the primal-dual analysis, we show that the optimal decision rule has a simple geometric form based on deciding whether a point is in an ellipse or not.

Then, we study the approximate policy optimization under primal-dual methods [Xiong et al., 2024b]. As one of the most popular classes of algorithms in reinforcement learning, policy optimization methods have been studied extensively from both empirical and theoretical perspectives. However, while most successful empirical algorithms are developed without thorough theoretical support, most theoretical analyses are presented without empirical evaluations either. To bridge this gap, we propose Dual Approximation Policy Optimization (DAPO), a framework that incorporates general function approximation into policy mirror descent methods. In contrast to the popular approach of using the L_2-norm to measure function approximation errors, DAPO uses the dual Bregman divergence induced by the mirror map for optimization. This duality-consistent framework has both theoretical and practical implications: not only does it achieve fast linear convergence with general

function approximation, but it also includes several well-known practical methods as special cases, immediately providing them with strong convergence guarantees. Furthermore, on MuJoCo benchmarks DAPO successfully achieve performance much better than algorithms with L_2-norm approximation. Finally, we extend our theoretical analysis to continuous state and action spaces, thus completing a bridging between theory and practice.

Part I

EFFICIENT EXPLORATION IN BANDITS AND REINFORCEMENT LEARNING.

Chapter 2

ROBUST PURE-EXPLORATION IN NON-STATIONARY LINEAR BANDITS

This chapter is based on Xiong et al. [2024a], with Romain Camilleri, Maryam Fazel, Lalit Jain and Kevin Jamieson.

2.1 Introduction

Data-driven decision-making and A/B testing enable businesses to evaluate strategies using real-time customer data, offering insights into customer tendencies. As the use of these methods has increased, these technologies are being utilized to determine problems with smaller effect sizes, while also targeting smaller audiences. These two competing trends of smaller effect sizes and smaller sample sizes make it increasingly challenging to obtain statistical significance and correct inference since the absolute number of observations is limited. Consequently, there is a rising trend in using *adaptive* sampling like multi-armed bandits to obtain the same statistical insights using fewer total observations.

However, using adaptive experimentation schemes can come with many pitfalls. Most algorithms that are effective in practice (e.g., Thompson Sampling) are developed with the assumption that the *environment is stationary* and that rewards from treatments are stochastic. However in practice this is far from the case. Non-stationarity can be introduced from a variety of sources such as user populations that change from hour to hour, customer preferences which vary over the course of a year, changes in one part of a platform that lead to latency and higher bounceback, site-wide promotions and sales, interference from competitors, macroeconomic shifts, and many other disruptions. Many of these issues are often totally unobservable, and therefore cannot be controlled, modeled, or accounted for by an experimenter. Under such an environment, it is also possible for the underlying performance of treatments to wildly change, and as a result, the treatment that is best

performing on any given day may change. This makes the concept of "the best-performing arm" poorly defined.

Instead, in time-varying settings, the goal of an experimenter is to identify the "counterfactual best treatment" at the end of the experimentation period. That is, the treatment that would have received the *highest total reward had received all the samples.* However, in the absence of being able to predict or model time-variation, predicting precisely how a treatment would behave at every time point, at which time at most one treatment can be evaluated, is impossible. Fortunately, randomization is a powerful tool to provide the next best thing: unbiased *estimates* of how a treatment would behave as if it had been used at every time in the past. These methods are well-understood in the causal-inference and online learning literature and are commonly known as inverse-propensity score (IPS) estimators. The idea is simple: consider a sequence of evaluations from n treatments at each time $\{x_t\}_{t=1}^T \subset \mathbb{R}^n$. Note that a procedure can only observe at most one treatment per time denoted as $I_t \in [n]$, which is drawn from a distribution p_t over the n treatments. Then $\widehat{X}_i = \frac{1}{T}\sum_{t=1}^T \frac{\mathbb{1}\{I_t=i\}}{p_{t,i}} x_{t,i}$ is an unbiased estimator of the cumulative gain $\frac{1}{T}\sum_{t=1}^T x_{t,i}$ by

$$\mathbb{E}\left[\frac{\mathbb{1}\{I_t=i\}}{p_{t,i}}x_{t,i}\right] = \sum_{j=1}^n \mathbb{P}(I_t=j)\frac{\mathbb{1}\{j=i\}}{p_{t,i}}x_{t,i} = \sum_{j=1}^n p_{t,j}\frac{\mathbb{1}\{j=i\}}{p_{t,i}}x_{t,i} = x_{t,i}, \tag{2.1}$$

as long as $\min_{t,i} p_{t,i} > 0$. Of course, there is no free lunch, and the variance of $\widehat{X}_i$ behaves like $\frac{1}{T^2}\sum_{t=1}^T 1/p_{t,i}$. Intuitively, to maximize efficiency of the samples we do take for inference, we should try to minimize the probabilities on poor performing treatments and prioritize mass for the high performing treatments. However, if the treatment performances vary over time, it can be challenging to determine how one might do this optimally. Fortunately, Abbasi-Yadkori et al. [2018] proposes a novel solution to defining these probabilities in a dynamic way that achieves a "Best of Both Worlds" (BOBW) guarantee, which is an algorithm called P1 that manages to achieve near-optimal rates regardless of whether the environment is stochastic or arbitrarily non-stationary (adversarial). This seminal work is the gold standard for A/B testing in unpredictable non-stationary settings.

If the number of treatments is small (<10 in practice), BOBW provides a robust solution

for practitioners. However, there are many situations that practitioners are interested in for which the number of treatments is very large and intractable for traditional A/B testing. For example, multivariate testing Hill et al. [2017] aims to identify not just a single best item, but a set or bundle of items, such as the best 6 pieces of content to highlight on a home screen. Given n possibilities, this results in $\binom{n}{6}$ total distinct treatments for the A / B test! Given this combinatorial explosion, practitioners have made structural parametric assumptions, such as the expected value of a set of items behaves like

$$\theta^{(0)} + \sum_{i=1}^{n} \theta_i^{(1)} \alpha_i + \sum_{i=2}^{n} \sum_{j<i} \theta_{i,j}^{(2)} \alpha_i \alpha_j,$$

where $\alpha \in \{0,1\}^n$ with $\sum_i \alpha_i = 6$ indicates whether an item was included in the set or not. Note that these sums can be succinctly written as $\langle x, \theta \rangle$ for $\theta = (\theta^{(0)}, \theta^{(1)}, \theta^{(2)})^\top \in \mathbb{R}^{1+n+\binom{n}{2}}$ and an appropriate $x \in \{0,1\}^{1+n+\binom{n}{2}}$. This can reduce the overall number of unknowns, and dimension, to just $O(n^2)$ versus $O(n^6)$. But now the vectors $x \in \mathcal{X}$, each associated with a particular bundle, are overlapping and can share information. A similar situation arises if we have features or covariates that describe each possible treatment. For example, a particular song comes with lots of metadata including artist, genre, beats per minute, etc. which can encode the useful properties about the song.

In these kinds of scenario—whether it be multivariate testing or items with feature descriptions—we would like to perform adaptive experimentation in the presence of time-variation. Recall that without covariates, we have solutions like P1 that are near-optimal for time-variation. And without time-variation, there are many methods that take covariates into account and are known to be near-optimal. This work aims to develop an algorithmic framework for handling covariates with time variation.

The remainder of the paper is organized as follows. We discuss the related work in Section 2.2 and presents detailed problem formulations in Section 2.3. In Section 2.4, we propose a simple algorithm for general non-stationary environments and then in Section 2.5, we propose a robust algorithm that can simultaneously tackle stationary and non-stationary environments. Experiment results are presented in Section 6.7 and our conclusions in Section 4.6.

2.2 Related Work

The problem of identifying the best arm in linear bandits is a well-established and extensively researched problem. [Soare et al., 2014, Karnin, 2016, Xu et al., 2018, Fiez et al., 2019, Katz-Samuels et al., 2020, Degenne et al., 2020, Jedra and Proutiere, 2020, Wagenmaker and Foster, 2023]. Notably, Katz-Samuels et al. [2020], Azizi et al. [2021], Yang and Tan [2021] focus on the fixed-budget setting and are closely related to our paper. One notable limitation of these algorithms is their reliance on (unrealistic) stationary settings, which leads to their critical failure when applied in non-stationary scenarios. This motivated increasing interest in studying models for non-stationarity in bandits problems and algorithms agnostic to non-stationary settings, which we review next.

Models for non-stationarity in bandits. A reasonable approach in bandit problems with distribution shifts is to provide tight models for unknown variations in the reward distribution. Most literature in this setting focuses on minimizing the dynamic regret, which compares the reward obtained against the reward of the best arm in each round t. Garivier and Moulines [2011] demonstrates that existing methods such as Auer et al. [2002] could achieve a dynamic regret of $\widetilde{O}(\sqrt{LT})$ when L, the number of distribution shifts, is known. Then, Auer et al. [2019] makes a significant advancement by introducing an adaptive approach with the same dynamic regret but without the knowledge of L. More recently, Chen et al. [2019], Wei and Luo [2021] establish analogous results in the contextual bandits settings. Measures of non-stationarity other than L are also considered. In particular, Chen et al. [2019] measures the non-stationarity by total variation and Suk and Kpotufe [2022] proposes the novel notion of severe shifts. Note importantly that while this extensive body of work focuses on building tight models of non-stationarity and developing regret minimization algorithms tuned to them, our work is agnostic to such models.

Agnostic non-stationary bandits (Best of both worlds). Bubeck and Slivkins [2012], Seldin and Slivkins [2014], Seldin and Lugosi [2017], Auer and Chiang [2016], Abbasi-Yadkori et al. [2018], Lee et al. [2021] focus on the "best of both worlds" (BOBW) problem: design a bandit algorithm that agnostically achieves optimal performance in both stationary and non-stationary scenarios, even without prior knowledge of the environment. While

most BOBW work focus on regret minimization goals, Abbasi-Yadkori et al. [2018] focuses on BOBW for best-arm identification. In this work, as in Abbasi-Yadkori et al. [2018], we focus on the agnostic setting.

A/B testing. As discussed in the introduction, our work is closely related to non-stationary A/B testing. In settings with non-stationarity and adaptive sample allocations, non-stationarity can lead to Simpson's paradox if the sample means are used to estimate arm means Kohavi and Longbotham [2011]. It is common in large-scale industrial platforms to assume that means vary smoothly Wu et al. [2022], or that the differences between them are constant; i.e., all arms are subject to the same random exogeneous shock Optimizely [2023]. The recent work Qin and Russo [2022] models time-variation as arising from confounding due to a context distribution and aims to find the arm with the best reward on average under this context distribution. Their goal is similar to ours, but, unlike them, we do not assume a context distribution.

2.3 Preliminaries

Notation. Let $[a : b] = \{a, a+1, \ldots, b\}$ for $a, b \in \mathbb{N}$ with $b > a$ and $[a] = \{1, \ldots, a\}$. For a vector $x \in \mathbb{R}^d$ and symmetric positive semi-definite (PSD) matrix $A \in \mathbb{S}_+^d$, we use $\|x\|_A = \sqrt{x^\top A x}$ to denote the Mahalanobis norm. For a finite set $\mathcal{X} \subset \mathbb{R}^d$ and distribution $\lambda \in \triangle_{\mathcal{X}}$ over $\mathcal{X}$, we use $A(\lambda) = \mathbb{E}_{x\sim\lambda}\left[xx^\top\right]$ to denote the covariance matrix under λ.

2.3.1 Linear Bandits Problem Formulation

General stationary/non-stationary environments. In this paper, we assume a standard stationary/non-stationary linear bandits model with fixed horizon T. In particular, let $\mathcal{X} \subset \mathbb{R}^d$ be a finite arm set with $|\mathcal{X}| = K$ such that $\text{span}(\mathcal{X}) = \mathbb{R}^d$. At each time $t = 1, \ldots, T$, the learner will pick some arm $x_t \in \mathcal{X}$ and receive some noisy reward $r_t = x_t^\top \theta_t + \epsilon_t$, where $\epsilon_t \in [-1, 1]$ is some independent zero-mean noise. All parameters $\{\theta_t\}_{t=1}^T$ are chosen and fixed by the environment before the game starts.* The ultimate goal of the learner is to find

*Theoretically, this non-stationary setting has no essential difference with the adversarial setting. We choose this non-stationary setting mainly to keep our presentation concise.

the optimal arm $\arg\max_{x\in\mathcal{X}} x^\top \overline{\theta}_T$, where $\overline{\theta}_T = \frac{1}{T}\sum_{t=1}^{T}\theta_t$ is the average parameter. This protocol is summarized in Figure 2.1.

Input: time horizon, T; arm set, $\mathcal{X} \subset \mathbb{R}^d$
For $t = 1, \ldots, T$
    The learner plays arm $x_t \in \mathcal{X}$
    The learner receives reward $r_t = x_t^\top \theta_t + \epsilon_t$, where ϵ_t is independent zero-mean noise
The learner recommends arm x_{J_T}

Figure 2.1: General protocol of fixed-budget best-arm identification (BAI) for linear bandits.

For simplicity, we further assume that $\forall t \in [T]$, $\forall x \in \mathcal{X}$, $x^\top \theta_t \in [-1, 1]$ and the optimal arm $\arg\max_{x\in\mathcal{X}} x^\top \overline{\theta}_T$ is unique. Meanwhile, similar to Abbasi-Yadkori et al. [2018], we use the subscript (k) to denote the index of k-th best arm in $\mathcal{X}$, which means to have $x_{(1)}^\top \overline{\theta}_T > x_{(2)}^\top \overline{\theta}_T \geq \cdots \geq x_{(K)}^\top \overline{\theta}_T$. For each arm $k \in [K]$, we define its gap Δ_k as

$$\Delta_k = \begin{cases} (x_{(1)} - x_k)^\top \overline{\theta}_T & \text{if } k \neq (1), \\ (x_{(1)} - x_{(2)})^\top \overline{\theta}_T & \text{if } k = (1). \end{cases}$$

That is,we have $\Delta_{(1)} = \Delta_{(2)} \leq \Delta_{(3)} \leq \cdots \leq \Delta_{(K)}$. As a slight abuse of notation, for unindexed arm $x \in \mathcal{X}$, we will use Δ_x to denote the gap of x. The performance of the learner is measured by its error probability $\mathbb{P}_{\overline{\theta}_T}(J_T \neq (1))$, where J_T is the index of the learner's recommendation and the probability measure is taken over the randomness inside the learner and the reward noise. Finally, we note that when the setting is stationary, we simply have $\theta_1 = \cdots = \theta_T = \theta^*$ and everything else is then defined accordingly.

Remark 2.3.1 (Comparison to the adversarial setting). The traditional oblivious adversarial setting can be viewed as a special case of our non-stationary setting, in which we simply pick $\epsilon_t = 0$ for all t [Abbasi-Yadkori et al., 2018].

2.3.2 BAI for Linear Bandits in Stationary Environments

In this section, we briefly review the well-studied best-arm identification problem for linear bandits in stationary settings. This problem's complexity, first proposed in Soare et al.

[2014], is defined as

$$\rho^*(\theta) = H_{\mathsf{LB}}(\theta) = \inf_{\lambda \in \triangle_{\mathcal{X}}} \max_{x \neq x_{(1)}} \frac{\left\|x - x_{(1)}\right\|^2_{A(\lambda)^{-1}}}{\Delta_x^2}, \tag{2.2}$$

where the optimal arm index (1) and gaps Δ_k are defined based on the input parameter θ. As discussed in Soare et al. [2014], this complexity is approximately equal to the number of samples required (up to logarithmic terms) to find the best arm by running an oracle algorithm. Later in Fiez et al. [2019], this complexity is proved to be the optimal sample complexity that a BAI algorithm can possibly achieve in a fixed-confidence setting. Recently, Katz-Samuels et al. [2020] proposes algorithm Peace in fixed-budget setting that achieves error probability $\mathbb{P}_\theta\left(J_T \neq (1)\right) \leq \widetilde{O}\left(\exp\left(-\frac{T}{\rho^*(\theta)\log(d)}\right)\right)$.[†]

2.4 BAI For Linear Bandits in General Non-Stationary Environments

In this section, we present a simple algorithm G-BAI for the general non-stationary environment and analyze its theoretical guarantee. The algorithm is based on the G-optimal design, which refers to the distribution $\lambda^* \in \triangle_{\mathcal{X}}$ such that

$$\lambda^* = \arg\inf_{\lambda \in \triangle_{\mathcal{X}}} \max_{x \in \mathcal{X}} \|x\|^2_{A(\lambda)^{-1}}. \tag{2.3}$$

Intuitively, G-optimal design allows us to estimate unknown parameter θ_t uniformly well over all directions of the arms in $\mathcal{X}$ [Soare et al., 2014]. which is suitable for addressing non-stationarity since θ_t may change arbitrarily and each $x \in \mathcal{X}$ may become the optimal at time t. Meanwhile, to make sure the estimation of θ_t is unbiased in a non-stationary environment, we use an IPS estimator.

Therefore, briefly speaking, at each time t, G-BAI simply samples x_t based on G-optimal design and estimate θ_t through an IPS estimator, whose details are summarized in Algorithm 1.[‡]

[†]Rigorously speaking, the error probability of Peace contains another complexity term called $\gamma^*(\theta)$, which is defined as the minimum of a Gaussian width term. However, as argued in Katz-Samuels et al. [2020], $\gamma^*(\theta)$ is roughly in a same order of $\rho^*(\theta)$.

[‡]We can see $\widehat{\theta}_T$ exactly becomes the more commonly-seen IPS estimator examined in Eq. (2.1) if we

Algorithm 1 G-optimal Best-arm Identification (G-BAI)

Require: budget, $T \in \mathbb{N}$; arm set $\mathcal{X} \subset \mathbb{R}^d$
1: Compute G-optimal design λ^* based on Eq. (2.3)
2: **for** $t = 1, 2, \ldots, T$ **do**
3: Sample $x_t \sim \lambda^*$ and receive reward r_t
4: **end for**
5: Estimate $\widehat{\theta}_T \leftarrow \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}_{x\sim\lambda^*}\left[xx^\top\right]^{-1} x_t r_t$
6: **return** $\arg\max_{x\in\mathcal{X}} x^\top \widehat{\theta}_T$

By the famous Kiefer-Wolfowitz theorem, an important property of the G-optimal design is that $\max_{x\in\mathcal{X}} \|x\|^2_{A(\lambda^*)^{-1}} = d$ [Lattimore and Szepesvári, 2020]. With this property, the variance of estimator $\widehat{\theta}_t$ can be easily controlled. We can then bound the error probability of G-BAI by this fact and the result is summarized in the following theorem.

Theorem 2.4.1 (Error probability of G-BAI)**.** *Fix time horizon T, arm set $\mathcal{X} \subset \mathbb{R}^d$ with $|\mathcal{X}| = K$ and arbitrary unknown parameters $\{\theta_t\}_{t=1}^T$. If we run Algorithm 1 in this non-stationary environment and obtain x_{J_T}, then it holds that*

$$\mathbb{P}_{\bar{\theta}_T}\left(J_T \neq (1)\right) \leq K \exp\left(-\frac{T}{12 H_{\text{G-BAI}}\left(\bar{\theta}_T\right)}\right), \quad \text{where } H_{\text{G-BAI}}\left(\bar{\theta}_T\right) = \frac{d}{\Delta^2_{(1)}}.$$

The proof of Theorem 2.4.1 is deferred to Appendix A.2. Here, we briefly compare this result with the one in multi-armed bandits, which can be treated as a special case of linear bandits by taking $\mathcal{X} = \{\mathbf{e}_1, \ldots, \mathbf{e}_K\}$ to be the canonical vectors (standard basis) with $K = d$.

In particular, Abbasi-Yadkori et al. [2018] shows that in multi-armed bandits setting, a simple uniform sampling algorithm reaches complexity $H_{\text{UNIF}}\left(\bar{\theta}_T\right) = \frac{K}{\Delta^2_{(1)}}$ and it is optimal in non-stationary environments. Meanwhile, based on Theorem 2.4.1, we can see the complexity of G-BAI is $H_{\text{G-BAI}}\left(\bar{\theta}_T\right) = \frac{d}{\Delta^2_{(1)}}$, which is exactly $H_{\text{UNIF}}(\bar{\theta}_T)$ if we treat multi-armed bandits as a special case of linear bandits since $d = K$. Furthermore, if we directly apply G-BAI to multi-armed bandits, meaning to use $\mathcal{X} = \{\mathbf{e}_1, \ldots, \mathbf{e}_K\}$, then λ^* is exactly the uniform distribution over $\mathcal{X}$. That is, in multi-armed bandits, G-BAI exactly recovers the optimal complexity in non-stationary environments, which shows that G-BAI is minimax

apply it to the multi-armed bandits setting, in which we have $K = d$ arms and $\mathcal{X} = \{\mathbf{1}_1, \ldots, \mathbf{e}_d\}$.

optimal for linear bandits.

2.5 A Robust Algorithm For Stationary/Non-Stationary Environments

In this section, we present and analyze a new robust linear bandits BAI algorithm called P1-RAGE, which performs comparable to G-BAI in non-stationary environments but much better than it in stationary environments. We will show that it attains good error probability in both stationary and non-stationary environments simultaneously, without knowing a priori which environment it will encounter. We first discuss some intuitions behind the algorithm design.

Stationary environments. The development of our algorithm P1-RAGE is largely inspired by the high-level idea of the robust algorithm P1, proposed in Abbasi-Yadkori et al. [2018], and the allocation strategy of RAGE, proposed in Fiez et al. [2019]. In particular, as discussed in Abbasi-Yadkori et al. [2018], in multi-armed bandits, to minimize the error probability in stationary environment, we need to control the estimation variance of the optimal arm well enough. Therefore, at each time step, algorithm P1 pulls the current estimated best arm with the highest probability (unnormalized "probability one"), then subsequently the second best arm with second highest probability (unnormalized "probability half") and so on. We can notice that it actually matches the allocation strategy of the successive halving algorithm in Karnin et al. [2013], which is proved to be near-optimal for BAI in stationary multi-armed bandits. Therefore, we design our probability allocation based on the allocation strategy of RAGE, which is proven to be near-optimal for fixed-confidence BAI in stationary linear bandits [Fiez et al., 2019]. In particular, with the estimated parameter $\widehat{\theta}_t$, we first find the estimated best arms $\hat{x}_t^* = \arg\max_{x \in \mathcal{X}} x^\top \widehat{\theta}_t$. Then, we use a subroutine to repeatedly and virtually eliminate arms with estimated gaps larger than certain threshold and compute $\mathcal{XY}$-allocation of the (virtually) remaining arms.[§] Then, we average over the allocation probabilities computed during each iteration.

Non-stationary environments. Finally, to address the potential non-stationarity in environments, we uniformly mix the allocation probability computed above with a G-optimal

[§]The elimination is virtual because no samples are collected during the elimination subroutine.

design. With such a mixture, the variance over all arms can be controlled well and thus the algorithm will be robust for both stationary and non-stationary environments. The details of P1-RAGE are summarized in Algorithm 2 and the subroutine to compute the allocation probability, called RAGE-Elimination, is summarized in Algorithm 3.

Algorithm 2 P1-RAGE

Input: budget, $T \in \mathbb{N}$; arm set $\mathcal{X} \subset \mathbb{R}^d$; maximum number of virtual phases, m
Compute G-optimal design λ^* based on Eq. (2.3) and initialize $\lambda_1 = \lambda^*$
for $t = 1, 2, \ldots, T$ **do**
  Sample $x_t \sim \lambda_t$ and receive reward r_t
  Estimate $\widehat{\theta}_t \leftarrow \frac{1}{t}\sum_{s=1}^{t} \mathbb{E}_{x\sim\lambda_s}\left[xx^\top\right]^{-1} x_s r_s$
  Update $\lambda_{t+1} \leftarrow$RAGE-Elimination$(\widehat{\theta}_t, m)$ // {Call Algorithm 3}
end for
return $\arg\max_{x\in\mathcal{X}} x^\top \widehat{\theta}_T$

Algorithm 3 RAGE-Elimination

Input: arm set $\mathcal{X} \subset \mathbb{R}^d$; current estimate $\widehat{\theta}_t$; maximum number of virtual phases, m
Find $\hat{x}_t^* \leftarrow \arg\max_{x\in\mathcal{X}} x^\top \widehat{\theta}_t$
Initialize $\mathcal{X}_t^{(0)} \leftarrow \mathcal{X}$ and $i \leftarrow 0$
while $|\mathcal{X}_t^{(i)}| > 1$ and $i \le m$ **do**
  $\lambda_t^{(i)} \leftarrow \arg\inf_{\lambda\in\triangle_{\mathcal{X}}} \max_{x,x'\in\mathcal{X}_t^{(i)}} \|x - x'\|^2_{A(\lambda)^{-1}}$
  $\mathcal{X}_t^{(i+1)} \leftarrow \left\{ x \in \mathcal{X}_t^{(i)} \,\middle|\, \widehat{\theta}_t^\top(\hat{x}_t^* - x) \le 2^{-i} \right\}$
  $i \leftarrow i + 1$
end while
return $(\bar{\lambda}_t + \lambda^*)/2$, where $\bar{\lambda}_t = \frac{1}{i}\sum_{i'=0}^{i-1} \lambda_t^{(i')}$

We bound the error probability of P1-RAGE under both stationary and non-stationary settings in the following theorem and its proof is deferred to Appendix A.3.

Theorem 2.5.1 (Error Probability of P1-RAGE). *Fix arm set $\mathcal{X} \subset \mathbb{R}^d$ with $|\mathcal{X}| = K$ and budget T. For a stationary environment with unknown parameter θ, if $m \ge i_0 = \lceil \log_2(1/\Delta_{(1)}) \rceil + 1$, then there exists absolute constant $c > 0$ such that the error probability of P1-RAGE satisfies*

$$\mathbb{P}_\theta\left(J_T \neq (1)\right) \le 2 i_0 K T \exp\left(-\frac{cT}{H_{\text{P1-RAGE}}(\theta)}\right),$$

$$\text{where } H_{\textit{P1-RAGE}}(\theta) = \frac{mi_0}{\Delta_{(1)}} \inf_{\lambda \in \triangle_{\mathcal{X}}} \max_{x \neq x_{(1)}} \frac{\left\|x - x_{(1)}\right\|^2_{A(\lambda)^{-1}}}{\Delta_x} + \frac{m\sqrt{d}}{\Delta_{(1)}} \inf_{\lambda \in \triangle_{\mathcal{X}}} \max_{x \neq x_{(1)}} \left\|x - x_{(1)}\right\|_{A(\lambda)^{-1}}. \tag{2.4}$$

For a non-stationary environment with unknown parameter $\{\theta_t\}_{t=1}^T$*, there exists absolute constant* $c' > 0$ *such that the error probability of P1-RAGE satisfies*

$$\mathbb{P}_{\bar{\theta}_T}\left(J_T \neq (1)\right) \leq K \exp\left(-\frac{c' T \Delta_{(1)}^2}{d}\right).$$

We can immediately see that in non-stationary environments, the error probability of P1-RAGE matches (up to a constant) with G-BAI, showing that P1-RAGE is minimax optimal for linear bandits under non-stationarity. On the other hand, because of the $\frac{1}{\Delta_{(1)}}$ factor, we can see that in stationary environments, $H_{\text{P1-RAGE}}(\theta) \gtrsim H_{\text{LB}}(\theta)$ (defined in Eq. (2.2)), which implies that P1-RAGE is suboptimal in stationary settings. However, this should be expected since even for multi-armed bandits, as proved in Abbasi-Yadkori et al. [2018], it is impossible for an algorithm to achieve $H_{\text{LB}}(\theta)$ while being robust to non-stationarity, let alone linear bandits.

Nevertheless, when applying Theorem 2.5.1 to multi-armed bandits ($\mathcal{X} = \{\mathbf{e}_1, \ldots, \mathbf{e}_K\}$), as long as we choose $m \approx i_0$, we can show that (Corollary A.3.6 in Appendix A.3)

$$H_{\text{P1-RAGE}}(\theta) = \widetilde{O}\left(\frac{1}{\Delta_{(1)}} \max_{k \in [K]} \frac{k}{\Delta_{(k)}}\right) = \widetilde{O}\left(H_{\text{BOB}}(\theta)\right),$$

where $H_{\text{BOB}}(\theta)$ is the best-of-both-worlds complexity proposed in Abbasi-Yadkori et al. [2018]. In particular, Abbasi-Yadkori et al. [2018] proves that $H_{\text{BOB}}(\theta)$ is the best complexity that any algorithm can possibly achieve if it is constrained to be robust to non-stationarity. That is, again, our algorithm P1-RAGE retains the near-optimal complexity for stationary multi-armed bandits if it is constrained to be robust in non-stationary environments.

Remark 2.5.2. Here, we do not elaborate the proof details of Theorem 2.5.1 mainly because we do not recognize them as widely applicable techniques. However, we do want to emphasize that this proof is by no means a simple extension of the analysis of the algorithm P1 in

Abbasi-Yadkori et al. [2018]. In particular, our proof uses a different set of virtual events based on the estimated gaps. Meanwhile, the analysis of subroutine RAGE-Elimination is intricately tailored to the unique characteristics of being a virtual elimination strategy, which is not presented in neither RAGE nor P1 [Abbasi-Yadkori et al., 2018, Fiez et al., 2019].

Theoretical limitations of P1-RAGE. Despite being near-optimal in multi-armed bandits, $H_{\text{P1-RAGE}}(\theta)$ includes an extra low-order term $\frac{m\sqrt{d}}{\Delta_{(1)}} \inf_{\lambda \in \triangle_{\mathcal{X}}} \max_{x \neq x_{(1)}} \left\|x - x_{(1)}\right\|_{A(\lambda)^{-1}}$. This term appears because the Bernstein's inequality requires a bound of the estimator's magnitude, which can be removed if the concentration bound only scales with the estimator's variance. Although this can often be accomplished by using Catoni's robust mean estimator [Wei et al., 2020], it requires a concrete confidence level to be specified before estimation, which is not feasible in our fixed budget setting. Finding an approach to circumvent this difficulty and remove this extra term, or alternatively, demonstrate that it is necessary, is an open question.

Remark 2.5.3. The question of whether the extra term is removable naturally relates the instance-dependent lower bound of this problem. However, proving an instance-dependent lower bound for our setting requires constructing both stationary and non-stationary counterexamples. This task is thereby more challenging compared to proving an instance-dependent lower bound for the fixed-budget best-arm identification problem in linear bandits within a purely stationary setting, an open question that persists (see Yang and Tan [2022] for a minimax lower bound). We thus leave establishing such instance-dependent lower bounds for future work.

Parameter choice of P1-RAGE. Although P1-RAGE requires a user-specified parameter $m \geq \lceil \log(1/\Delta_{(1)}) \rceil + 1$ to bound the total number of virtual phases, it is not difficult to choose a reasonable value for this parameter in a practical implementation. On the one hand, since its dependence on $\Delta_{(1)}^{-1}$ is only logarithmic, taking some moderate value such as $m = 25$ should safely satisfy $m \geq i_0$ for most practical scenarios; on the other hand, in most real-world applications, a sub-optimal arm should always be acceptable as long as its gap is small enough. Indeed, if we take ϵ to be the largest acceptable sub-optimality gap and take $m \geq \lceil \log(1/\epsilon) \rceil + 1$, then P1-RAGE will output arm x_{J_T} that satisfies $\Delta_{J_T} \leq \max\left\{\epsilon, \Delta_{(1)}\right\}$

with high probability in pure stationary environments (Corollary A.3.7 in Appendix A.3). That is, the output arm will either be an optimal arm if $\epsilon \leq \Delta_{(1)}$ or an arm with an acceptable suboptimality gap ϵ otherwise.

2.6 Experiments

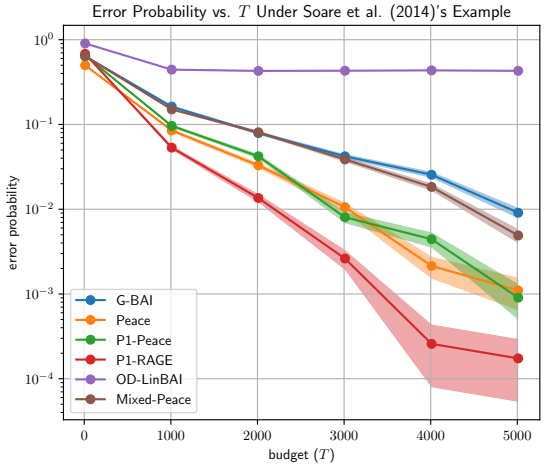
In this section, we present our experiment results on several stationary/non-stationary environments. Since to the best of our knowledge, we are the first to propose best-arm identification algorithms that tackle non-stationarity in linear bandits, the algorithms from other works that we compare with are all specifically designed for stationary environments. In particular, we will compare our algorithms with Peace, which is the first fixed-budget algorithm for linear bandits and also inspires our algorithmic design [Katz-Samuels et al., 2020], and OD-LinBAI, which is the most recent algorithm of this kind and is claimed to be minimax optimal [Yang and Tan, 2022].

Meanwhile, we also examine two additional heuristically designed algorithms for non-stationary environments. The first one is P1-Peace, which has the same design spirit as P1-RAGE but uses a different Peace-based virtual elimination subroutine; the second one is Mixed-Peace, which is a naive mixture of Peace and the G-optimal design. In particular, while P1-RAGE/P1-Peace combines G-optimal design with what RAGE/Peace would sample *in a full run*, Mixed-Peace simply mixes G-optimal design with what Peace in a stationary environment samples *at each time step*. The details of these two additional algorithms are summarized in Algorithm 11 and 13 in Appendix A.1.1, respectively. More implementation details and additional experiments can be found in Appendix A.4.[¶]

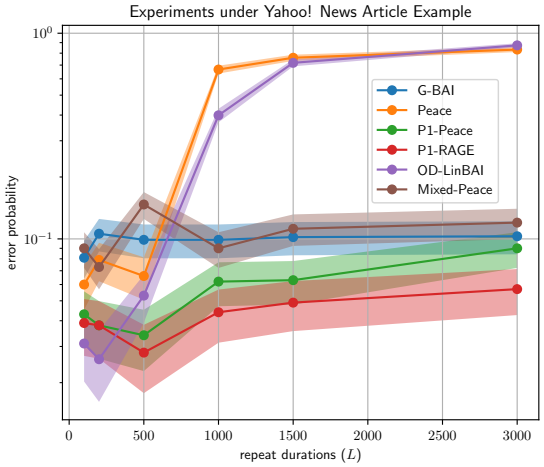
Stationary benchmark example. First, as a sanity check, we consider the famous stationary benchmark example proposed in Soare et al. [2014]. In particular, we have $\mathcal{X} = \{\mathbf{e}_1, \ldots, \mathbf{e}_d, x'\}$, where $x' = \cos(\omega)\mathbf{e}_1 + \sin(\omega)\mathbf{e}_2$ with some small $\omega > 0$, and $\overline{\theta}_T = \theta^* = 2\mathbf{e}_1$ so that $x_{(1)} = \mathbf{e}_1$. An efficient algorithm should pick $\mathbf{e}_2$ frequently to reduce the variance in the direction of $\mathbf{e}_1 - x'$. In this example, we pick $d = 10$ and $\omega = 0.1$.

The results are shown in Figure 2.2a. We can see that both our algorithms, P1-RAGE

[¶]Code repository is available at `https://github.com/FFTypeZero/bobw_linear`.



(a) Each error probability is estimated through at least 2×10^4 independent trials.



(b) Each error probability is estimated through 1000 independent trials.

Figure 2.2: The vertical axis is on log scale and the shaded area represents the 95% confidence interval.

and P1-Peace, perform better than G-BAI and comparably with Peace, showing that our algorithms maintain good performance in stationary environments. Meanwhile, we also notice that Mixed-Peace has performance only comparable to G-BAI, showing that naively mixing the allocation strategy with the G-optimal design can downgrade the performance in stationary environments.

Non-stationary multivariate testing example. We consider a multivariate testing example from Fiez et al. [2019], which is also similar to the one discussed in Introduction. Considering a webpage with D distinct slots and suppose each slot has two content choices, where we represent each layout as an element $w \in \mathcal{W} = \{-1, 1\}^D$. We hope to maximize the click-through rate and we assume it linearly depends on a layout-determined arm $x \in \mathcal{X}$ in a form of

$$x^\top \theta^* = \theta_0^* + \alpha_1 \sum_{j=1}^{D} \theta_j^* w_j + \alpha_2 \sum_{k=1}^{D-1} \sum_{\ell=k+1}^{D} \theta_{k,\ell}^* w_k w_\ell.$$

Here θ_0^* is the common bias, θ_j^* is the weight of j-th slot and $\theta_{k,\ell}^*$ is the weight of the interaction between k-th and ℓ-th slots. Because of the periodic nature of people's life cycle,

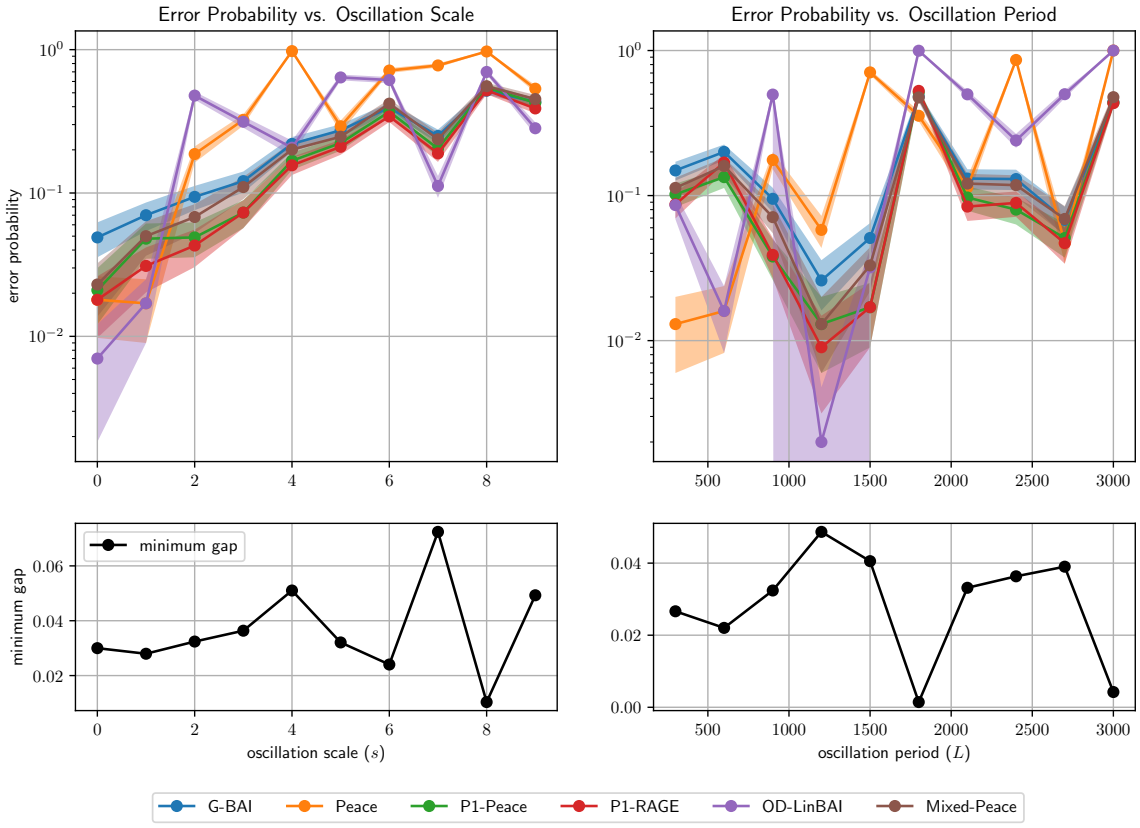


Figure 2.3: Each error probability is estimated through 1000 repeated trials. The bottom two plots give the minimum gap $\Delta_{(1)}$ of each instance as a function of oscillation scale s and oscillation period L.

it is very likely that the real-world weights will periodically change. Therefore, to construct a non-stationary environment, we randomly oscillate the weights with scale s and period L to get

$$\theta_{t,i} = \theta_i^* + sI \left\|\theta^*\right\|_\infty \sin\left(\frac{2\pi t}{L} + \phi_i\right), \quad \text{where } I \sim \mathsf{Unif}(\{0,1\}), \phi_i \sim \mathsf{Unif}([0,2\pi]).$$

Here, in the first series of instances, we fix $L = 900$ and take values $s \in \{0, 1, \ldots, 9\}$, and in the second series of instances, we fix $s = 2$ and take values $L \in \{300, 600, \ldots, 3000\}$. Finally, we take $\alpha_1 = 1$, $\alpha_2 = 0.5$, sample each component of θ^* uniformly in $[-0.1, 0.1]$ and guarantee that $\overline{\theta}_T$ has the same optimal arm as θ^*. We take $T = 10^4$ for all settings and the

results are shown in Figure 2.3.

From the plots, we can see that the error probabilities of Peace and OD-LinBAI, algorithms designed for stationary environments, can range from near 0 to 1 in different non-stationary environments, which is quite unstable. Meanwhile, we can see that the performance of the other four algorithms, which all in certain way contain a G-optimal design, is relatively much more stable.[‖] Furthermore, among these four algorithms, we can see that our algorithms P1-RAGE and P1-Peace consistently outperform (never worse than) G-BAI and Mixed-Peace.

Non-stationary click-through example. To create an instance using real-world data, we use the Yahoo! Webscope Dataset R6A [Yahoo!, 2011].[**] This dataset contains a fraction of user click log of Yahoo!'s news article from May 1st, 2009 to May 10th, 2009. For each click, we take the outer product between user and article features to get a vector in $\mathbb{R}^{36}$ and then we run a principle component analysis to get arm set $\mathcal{Z} \subset \mathbb{R}^{24}$. To create a non-stationary example, we take data from May 1st to May 7th and for each day's data, we fit a ridge regression with regularization 0.01, obtaining $\theta_1^*, \ldots, \theta_7^*$, which can be used to simulate user's weekly periodic behavior. Suppose we receive L visits each day, then, we can define a non-stationary environment where each period consists of $\theta_1^*, \ldots, \theta_1^*, \ldots, \theta_7^*, \ldots, \theta_7^*$ and each θ_i^* repeats for L times. Finally, we form our arm set $\mathcal{X}$ by picking the optimal arm from $\mathcal{Z}$ plus 23 randomly picked arms with gap at least 0.05 so that $\text{span}(\mathcal{X}) = \mathbb{R}^{24}$. We take $T = 2.1 \times 10^4$ and the results are shown in Figure 2.2b. Again, we can see that the performance of Peace and OD-LinBAI is very unstable and the performance of P1-RAGE and P1-Peace consistently outperforms the other two naive G-optimal-design-based algorithms, G-BAI and Mixed-Peace.

2.7 Conclusion and Future Work

To the best of our knowledge, in this paper, we present the first two novel robust linear bandits algorithm for fixed-budget best-arm identification, P1-RAGE and P1-Peace, that tackle stationary and non-stationary environments simultaneously while being agnostic to

[‖] All algorithms fluctuate in the upper right plot mainly because the minimum gaps also have large fluctuation.

[**] `https://webscope.sandbox.yahoo.com/`

the environment. Theoretically, we prove error probability bounds of P1-RAGE in both stationary and non-stationary environments. Empirically, we show that in stationary settings, both P1-RAGE and P1-Peace perform comparably with algorithms designed for such environments, and in non-stationary settings, they consistently outperform naive algorithms based on G-optimal design.

Finally, several questions still remain open. Is the extra term in $H_{\text{P1-RAGE}}(\theta)$, as discussed in Section 2.5, necessary? What is the optimal complexity for this mixed stationary/non-stationary settings? Answering these questions can serve as promising future directions.

Chapter 3

LEARNING AND EXPLORATION IN SINGLE-STEP CONGESTION GAMES

This chapter is based on Cui et al. [2022], with Qiwen Cui, Maryam Fazel and Simon S. Du.

3.1 Introduction

Nash equilibrium (NE) is a widely adopted concept in game theory community, used to describe the behavior of multi-agent systems with selfish players [Roughgarden, 2010]. At the Nash equilibrium, no player has the incentive to change its own strategy unilaterally, which implies it is a steady state of the game dynamics. For a general-sum game, computing the Nash equilibrium is PPAD-hard [Daskalakis, 2013] and the query complexity is exponential in the number of players [Rubinstein, 2016]. To help address these issues, a natural approach is to consider games with special structures. In this paper, we focus on congestion games.

Congestion games are general-sum games with *facilities* (resources) shared among players [Rosenthal, 1973]. During the game, each player will decide what combination of facilities to utilize, and popular facilities will become congested, which results in a possibly higher cost on each user. One example of congestion game is the routing game [Fotakis et al., 2002], where each player needs to travel from a given starting point to a destination point through some shared routes. These routes are represented as a traffic graph and the facilities are the edges. Each player will decide her path to go, and the more players use the same edge, the longer the edge travel time will be. Congestion games also have wide applications in electrical grids [Ibars et al., 2010], internet routing [Al-Kashoash et al., 2017] and rate allocation [Johari and Tsitsiklis, 2004]. In many real-world scenarios, players can only have (semi-)bandit feedback, i.e., players know only the payoff of the facilities they choose. This kind of learning under uncertainty has been widely studied in bandits and in reinforcement

learning for the single-agent setting, while theoretical understanding for the multi-agent case is still largely missing.

There are two types of algorithms in multi-agent systems, namely centralized algorithms and decentralized algorithms. For centralized algorithms, there exists a central authority that can control and receive feedback from all players in the game. As we have global coordination, centralized algorithms usually have favorable performance. On the other hand, such a central authority may not always be available in practice, and thus people turn to decentralized algorithms, i.e., each player makes decisions individually and can only observe her own feedback. However, decentralized algorithms are vulnerable to *nonstationarity* because each player is making decisions in a nonstationary environment as others' strategies are changing [Zhang et al., 2021a]. In this paper, we will study both centralized and decentralized algorithms in congestion games with bandit feedback, and we will provide motivating scenarios for both algorithms in Section 3.1.2.

The main challenge in designing algorithms for m-player congestion games with bandit feedback is the curse of exponential action set, i.e., the number of actions can be exponential in the number of facilities F because every subset of facilities can be an action. As a result, an efficient algorithm should have sample complexity polynomial in m and F and has no dependence on the size of the action space. One closely related type of general-sum game is the potential game, in which each individual's payoff changes, resulting from strategy modification, can be quantified by a common potential function. It is well-known that all congestion games are potential games, and each potential game has an equivalent congestion game formulation [Monderer and Shapley, 1996]. However, existing algorithms designed for potential games all have sample complexity scaling at least linearly in the number of actions [Leonardos et al., 2021, Ding et al., 2022], which is inefficient for congestion games. This motivates the following question:

Can we design provably sample-efficient centralized and decentralized learning algorithms for congestion games with bandit feedback?

We provide an affirmative answer to this question. To be precise, we use Nash-regret minimization (formally defined in Section 3.3) as our objective for learning in congestion games.

Algorithms	Sample complexity	Nash regret	Decentralized
Nash-VI [Liu et al., 2021]	$(\prod_{i=1}^m A_i)F/\epsilon^2$	$\sqrt{(\prod_{i=1}^m A_i)FT}$	No
V-learning [Jin et al., 2021a]	$A_{\max}F/\epsilon^2$ (CCE)	NA	Yes
IPPG [Leonardos et al., 2021]	$A_{\max}mF/\epsilon^6$	NA	Yes
IPGA [Ding et al., 2022]	$A_{\max}^2m^3F^5/\epsilon^5$	$mF^{4/3}\sqrt{A_{\max}}T^{4/5}$	Yes
Nash-UCB I	mF^2/ϵ^2	$F\sqrt{mT}$	No
Nash-UCB II	m^2F^3/ϵ^2	$mF^{3/2}\sqrt{T}$	No
Frank-Wolfe with Exploration I	$m^{12}F^9/\epsilon^6$	$m^2F^{3/2}T^{5/6}$	Yes
Frank-Wolfe with Exploration II	$m^{12}F^{12}/\epsilon^6$	$m^2F^2T^{5/6}$	Yes

Table 3.1: Comparison of algorithms for congestion games in terms of sample complexity and Nash regret, where "IPPG" stands for "independent projected policy gradient", "IPGA" stands for "independent policy gradient ascent", "I" represents the setting of semi-bandit feedback and "II" represents the setting of bandit feedback. Bandit feedback is assumed for algorithms from previous work. Here, A_i is the size of player i's action space, m is the number of players, $A_{\max} = \max_{i\in[m]} A_i$, F is the number of facilities and T is the number of samples collected. Our algorithms are shaded.

This regret-like objective commonly appears in the literature of online learning and reinforcement learning [Orabona, 2019, Ding et al., 2022, Liu et al., 2021], which focuses on finite-time analysis and accumulative rewards throughout the learning process instead of the asymptotic behavior. In general, a sublinear Nash regret implies a best-iterate convergence, meaning that the algorithm has reached the approximate Nash equilibrium at least once, while the converse does not hold.

We highlight our contributions below and compare our results with previous algorithms in Table 3.1. Our algorithms are shaded and we prove sublinear Nash regrets for all of them. In Table 3.1, sample complexity refers to the number of samples required to reach best-iterate convergence to an ϵ-approximate Nash equilibrium and the results are obtained by standard online-to-batch conversion as in Section 3.1 of [Jin et al., 2018].

3.1.1 Main Novelties and Contributions

1. Centralized algorithm for congestion game. We adapt the principle of optimism in the face of uncertainty in stochastic bandits to ensure sufficient exploration in congestion games. We begin with congestion games with semi-bandit feedback, in which each player can

observe the reward of every facility in the action. Instead of estimating the action reward as in stochastic multi-armed bandits, we estimate the facility rewards directly, which *removes the dependence on the size of action space*. Furthermore, we consider congestion games with bandit feedback, in which each player can only observe the overall reward. In this setting, we borrow ideas from linear bandits to estimate the reward function and analyze the algorithm. The algorithm is provably sample efficient in both cases.

2. Decentralized algorithm for congestion game. Our decentralized algorithm is a Frank-Wolfe method with exploration, in which each player only observes her own actions and rewards. To efficiently explore in the congestion game, we utilize G-optimal design allocation for bandit feedback and a specific distribution for semi-bandit feedback. As a result, the sample complexity does not depend on the number of actions. In addition, the L_1 smoothness parameter of the potential function does not depend on the number of actions, which is exploited by the Frank-Wolfe method. With the help of these two specific algorithmic designs for congestion games, we give the first decentralized algorithm for both semi-bandit feedback and bandit feedback that has no dependence on the size of the action space in congestion games.

3. Centralized algorithm for independent Markov congestion game. We extend the formulation of congestion game into a Markov setting and propose the independent Markov congestion game (IMCG), in which each facility has its own internal state and state transition happens independently among all the facilities. In Section 3.1.2, we give some examples that fit in this model. By utilizing techniques from factored MDPs, we extend our centralized algorithms for congestion games to efficiently solve IMCGs, with both semi-bandit and bandit feedback.

3.1.2 Motivating Examples

We provide an exmple here to motivate our proposed models. See Section 3.3 for the formal definition of (semi-)bandit feedback and (Markov) congestion games and Appendix B.1 for additional examples.

Example 3.1.1 (Routing Games). For a routing game, there are multiple players in a

traffic graph travelling from starting points to destination points, and the facilities are the edges (roads). The cost of each edge is the waiting time, which depends on the number of players using that edge.

• **Centralized algorithm for routing games:** Imagine each player is using Google Maps to navigate. Then Google Maps can serve as a center that knows the starting points and the destination points, as well as the real-time feedback of the waiting time on each edge of all the players. Google Maps itself also has the incentive to assign paths according to the Nash equilibrium strategy as then each player will find out that deviating from the navigation has no benefit and thus sticks to the app.

• **Decentralized algorithm for routing games:** Consider the case where players are still using Google Maps but due to privacy concerns or limited bandwidth, they only use the offline version, which has access only to the information of each single user. Then Google Maps needs to use decentralized algorithms so that it can still assign Nash equilibrium strategy to each user after repeated plays.

• **Markov routing games:** For Markov routing games, the time cost on each edge will change between different timesteps, which is a more accurate model of the real-world. For instance, some roads are prone to car accidents, which will result in an increasing cost on the next timestep, and the chance of accidents also depends on the number of players using that edge currently. This is modeled by the Markovian facility state transition in independent Markov congestion games.

3.2 Related Work

Potential Games. Potential games are general-sum games that admit a common potential function to quantify the changes in individual's payoff [Monderer and Shapley, 1996]. Algorithmic game theory community has studied how different dynamics converge to the Nash equilibium, e.g., best response dynamics [Durand, 2018, Swenson et al., 2018] and no-regret dynamics [Heliou et al., 2017, Cheung and Piliouras, 2020], while usually they provide only asymptotic convergence, with either full information setting or bandit feedback setting. Recently, reinforcement learning community studied Markov potential games with bandit feedback, which can be applied to standard potential games. See the Markov Games part

below for more details.

Congestion Games. Congestion games are developed in the seminal work [Rosenthal, 1973], and later Monderer and Shapley [1996] builds a close connection between congestion games and potential games. Congestion games are divided into atomic and non-atomic congestion games depending on whether each player is separable. Many papers consider non-atomic congestion games with non-decreasing cost function, which implies a convex potential function [Roughgarden and Tardos, 2004]. We consider the more difficult atomic congestion game where the potential function can be non-convex. For online non-atomic case, [Krichene et al., 2015] considers partial information setting while they provide convergence in the sense of Cesaro means. [Kleinberg et al., 2009, Krichene et al., 2014] show that some no-regret online learning algorithms asymptotically converges to Nash equilibrium. [Chen and Lu, 2015, 2016] are two closely related works that consider bandit feedback in atomic congestion games and provide non-asymptotic convergence. However, they still assume a convex potential function and the sample complexity has exponential dependence on the number of facilities, which is far from ideal.

Markov Games. Markov games are widely studied since the seminal work [Shapley, 1953]. Recently, the topic has received much attention due to advances in reinforcement learning theory. Liu et al. [2021] provides a centralized algorithm for learning the Nash equilibrium in general-sum Markov games, and [Jin et al., 2021a, Song et al., 2021] provide decentralized algorithms for learning the (coarse) correlated equilibrium. One closely related line of research is on Markov potential games [Leonardos et al., 2021, Zhang et al., 2021b, Fox et al., 2021, Cen et al., 2022a, Ding et al., 2022]. However, applying their algorithms to congestion games leads to explicit dependence on the number of actions, which would be exponentially worse than our algorithms. See Table 3.1 for comparisons. Our independent Markov congestion game is motivated by the state-based potential games studied in Marden [2012] and Macua et al. [2018], and its transition kernel is closely related to the factored MDPs, for which single agent algorithms are studied in [Osband and Van Roy, 2014, Chen et al., 2020, Xu and Tewari, 2020, Tian et al., 2020, Rosenberg and Mansour, 2021].

Learning in Games. Different from our paper, learning in games in traditional literature of game theory mainly considers players' asymptotic behavior [Leslie and Collins, 2005,

Cominetti et al., 2010, Coucheney et al., 2015]. In early literature, Leslie [2004] investigates actor-critic learning and Q-learning algorithms in games with bandit feedback and their connection to best-response dynamics. Leslie and Collins [2005] proposes individual Q-learning algorithm and shows that it converges to the NE almost surely in two-player zero-sum game and Leslie and Collins [2006] studies learning the NE from the perspective of a fictitious play-like process. Later, Cominetti et al. [2010] considers payoff-based learning rules and shows convergence to NE in traffic games, while another payoff-based learning model for continuous games is developed in Bervoets et al. [2020]. Coucheney et al. [2015] derives a new penalty-regulated dynamics and proposes a corresponding learning algorithms that converges to NE in potential games with bandit feedback. Bravo et al. [2018] proposes that in monotone games with bandit feedback, as long as all players are using some no-regret learning algorithm, the dynamics will converge to the NE, and an improved analysis of the same derivative-free algorithm is given in Drusvyatskiy et al. [2022]. In contrast, our learning objective focuses on finite-time cumulative rewards, which is more widely used in current multi-agent reinforcement learning literature [Ding et al., 2022, Liu et al., 2021].

3.3 Preliminaries

General-sum Matrix Games. We consider the model of general-sum matrix games, defined by the tuple $\mathcal{G} = (\{\mathcal{A}_i\}_{i=1}^m, R)$, where m is the number of players, $\mathcal{A}_i$ is the action space of player i and $R(\cdot|\boldsymbol{a})$ is the reward distribution on $[0, r_{\max}]^m$ with mean $\boldsymbol{r}(\boldsymbol{a})$. Let $\mathcal{A} = \mathcal{A}_1 \times \cdots \times \mathcal{A}_m$ be the whole action space and denote an element as $\boldsymbol{a} = (a_1, \ldots, a_m) \in \mathcal{A}$. After all players take actions $\boldsymbol{a} \in \mathcal{A}$, a reward vector is sampled $\boldsymbol{r} \sim R(\cdot|\boldsymbol{a})$ and player i will receive reward $r_i \in [0, r_{\max}]$ with mean $r_i(\boldsymbol{a})$. Each player's objective is to maximize her own reward.

A general policy π is defined as a vector in $\Delta(\mathcal{A})$, the probability simplex over the action space $\mathcal{A}$. A product policy $\pi = (\pi_1, \ldots, \pi_m)$ is defined as a tuple in $\Delta(\mathcal{A}_1) \times \cdots \times \Delta(\mathcal{A}_m)$, in which $\boldsymbol{a} = (a_1, \ldots, a_m) \sim \pi$ represents $a_i \overset{\text{i.i.d.}}{\sim} \pi_i$. The value of policy π for player i is $V_i^\pi = \mathbb{E}_{\boldsymbol{a}\sim\pi}[r_i(\boldsymbol{a})]$.

Nash Equilibrium and Nash Regret. Given a general policy π, let π_{-i} be the marginal joint policy of players $1, \ldots, i-1, i+1, \ldots, m$. Then, the best response of player i under

policy π is $\pi_i^\dagger = \arg\max_{\mu \in \Delta(\mathcal{A}_i)} V_i^{\mu,\pi_{-i}}$ and the corresponding value is $V_i^{\dagger,\pi_{-i}} := V_i^{\pi_i^\dagger,\pi_{-i}}$. Our goal is to find the approximate Nash equilibrium of the matrix game, which is defined below.

Definition 3.3.1. A product policy π is an *ϵ-approximate Nash equilibrium* if $\max_i (V_i^{\dagger,\pi_{-i}} - V_i^\pi) \le \epsilon$.

An ϵ-approximate Nash equilibrium can be obtained by achieving a sublinear Nash regret, which is defined below. See Section 3 in Ding et al. [2022] for a more detailed discussion.

Definition 3.3.2. With π^k being the policy at k-th episode, the *Nash regret* after K episodes is define as

$$\text{Nash-Regret}(K) = \sum_{k=1}^{K} \max_{i \in [m]} \left(V_i^{\dagger,\pi_{-i}^k} - V_i^{\pi^k} \right).$$

Remark 3.3.3. Here, if we replace $\max_{i\in[m]}$ by $\sum_{i=1}^m$ in the definition of Nash regret, the single-step Nash regret at episode k will become the Nikaido-Isoda (NI) function evaluated at π^k, which is a popular objective for equilibrium computation [Nikaidô and Isoda, 1955, Raghunathan et al., 2019]. Replacing $\max_{i\in[m]}$ by $\sum_{i=1}^m$ will multiply our regret bounds by a factor of m, while our conclusion will not be affected.

Potential Games. A potential game is a general-sum game such that there exists a potential function $\Phi : \Delta(\mathcal{A}) \to [0, \Phi_{\max}]$ such that for any player $i \in [m]$ and policies π_i, π_i', π_{-i}, it satisfies

$$\Phi(\pi_i, \pi_{-i}) - \Phi(\pi_i', \pi_{-i}) = V_i^{\pi_i,\pi_{-i}} - V_i^{\pi_i',\pi_{-i}}.$$

We can immediately see that a policy that maximizes the potential function is a Nash equilibrium.

Congestion Games. A congestion game is defined by $\mathcal{G} = (\mathcal{F}, \{\mathcal{A}_i\}_{i=1}^m, \{R^f\}_{f\in\mathcal{F}})$, where $\mathcal{F} = [F]$ is called the facility set and $R^f(\cdot|n) \in [0,1]$ is the reward distribution for facility f with mean $r^f(n)$, where $n \in [m]$. Each action $a_i \in \mathcal{A}_i$ is a subset of $\mathcal{F}$ (i.e., $a_i \subseteq \mathcal{F}$). Suppose the joint action chosen by all the players is $\boldsymbol{a} \in \mathcal{A}$, then a random reward is sampled $r^f \sim R^f(\cdot|n^f(\boldsymbol{a}))$ for each facility f, where $n^f(\boldsymbol{a}) = \sum_{i=1}^m \mathbb{1}\{f \in a_i\}$ is the number

of players using facility f. The reward collected by player i is $r_i = \sum_{f \in a_i} r^f$ with mean $r_i(\boldsymbol{a}) = \sum_{f \in a_i} r^f(n^f(\boldsymbol{a})) \in [0, F]$.

Connection to Potential Games [Monderer and Shapley, 1996]. As a special class of potential game, all congestion games have the potential function: $\Phi(\boldsymbol{a}) = \sum_{f \in \mathcal{F}} \sum_{i=1}^{n^f(\boldsymbol{a})} r^f(i)$. To see this, we can easily verify that $\Phi(a_i, a_{-i}) - \Phi(a_i', a_{-i}) = r_i(a_i, a_{-i}) - r_i(a_i', a_{-i})$ holds. Then, by defining $\Phi(\pi) = \mathbb{E}_{\boldsymbol{a} \sim \pi}[\Phi(\boldsymbol{a})]$, we can have $\Phi(\pi_i, \pi_{-i}) - \Phi(\pi_i', \pi_{-i}) = V_i^{\pi_i, \pi_{-i}} - V_i^{\pi_i', \pi_{-i}}$.

Types of feedback. There are in general two types of reward feedback for the congestion games, semi-bandit feedback and bandit feedback, both of which are reasonable under different scenarios. In semi-bandit feedback, after taking the action, player i will receive reward information r^f for each $f \in a_i$; in bandit feedback, after taking the action, player i will only receive the reward $r_i = \sum_{f \in a_i} r^f$ with no knowledge about each r^f. In this paper, we will address both of them, with more focus on the bandit feedback, which can be directly generalized to semi-bandit feedback.

3.4 Centralized Algorithms for Congestion Games

In this section, we introduce two centralized algorithms for congestion games – one for the semi-bandit feedback and one for the bandit feedback. We will see that both of them can achieve sublinear Nash regret with polynomial dependence on both m and F.

3.4.1 Algorithm for Semi-bandit Feedback

Summarized in Algorithm 4, Nash upper confidence bound (Nash-UCB) for congestion games is developed based on optimism in the face of uncertainty. In particular, the algorithm estimates the reward matrices optimistically in line 4, computes its Nash equilibrium policy in line 6 and then follows this policy.

For convenience, we define the empirical counter $N^{k,f}(n) = \sum_{k'=1}^{k} \mathbb{1}\left\{n^f(\boldsymbol{a}^{k'}) = n\right\}$ and $\tilde{\iota} = 2\log(4(m+1)K/\delta)$. Then, the reward estimator for f and the bonus term are defined

as

$$\hat{r}^{k,f}(n) = \frac{\sum_{k'=1}^{k} r^{k',f} \mathbb{1}\left\{n^f(\boldsymbol{a}^{k'}) = n\right\}}{N^{k,f}(n) \vee 1}, \quad b_i^{k,\mathrm{r}}(\boldsymbol{a}) = \sum_{f \in a_i} \sqrt{\frac{\tilde{\iota}}{N^{k,f}(n^f(\boldsymbol{a})) \vee 1}}, \tag{3.1}$$

where $r^{k,f} \in [0,1]$ is the random reward realization of $r^f(n^f(\boldsymbol{a}^k))$. Naturally, the reward estimator for player i is $\hat{r}_i^k(\boldsymbol{a}) = \sum_{f \in a_i} \hat{r}^{k,f}(n^f(\boldsymbol{a}))$.

Algorithm 4 Nash-UCB for Congestion Games

1: **Input:** ϵ, accuracy parameter for Nash equilibrium computation
2: **for** episode $k = 1, \ldots, K$ **do**
3: **for** player $i = 1, \ldots, m$ **do**
4: $\overline{Q}_i^k(\boldsymbol{a}) \leftarrow \hat{r}_i^k(\boldsymbol{a}) + b_i^{k,\mathrm{r}}(\boldsymbol{a})$ for all $\boldsymbol{a} \in \mathcal{A}$
5: **end for**
6: $\pi^k \leftarrow \epsilon$-Nash$(\overline{Q}_1^k(\cdot), \cdots, \overline{Q}_m^k(\cdot))$ (Algorithm 5)
7: Take action $\boldsymbol{a}^k \sim \pi^k$ and observe reward $r^{k,f}$
8: Update reward estimators $\hat{r}_i^k$ and bonus term $b_i^{k,\mathrm{r}}$
9: **end for**

Algorithm 4 is motivated by the Nash-VI algorithm in [Liu et al., 2021] plus a deliberate utilization of the special reward structure in the congestion games. Moreover, notice that a matrix game with reward functions $\overline{Q}_1^k(\cdot), \ldots, \overline{Q}_m^k(\cdot)$ forms a potential game (see Lemma B.2.1). As a result, in line 6, we can *efficiently compute* the ϵ-approximate Nash equilibrium π^k for that matrix game by utilizing Algorithm 5, (see Lemma B.2.2). It is a simple greedy algorithm such that in each round, it modifies one player's policy whose modification can increase the potential function most. In addition, Algorithm 5 always outputs a deterministic product policy.

3.4.2 Algorithm for Bandit Feedback

When the players can only receive bandit feedback, estimating $\hat{r}^{k,f}$ directly for each $f \in \mathcal{F}$ is no longer feasible. However, notice that the reward function $r_i(\boldsymbol{a}) = \sum_{f \in a_i} r^f(n^f(\boldsymbol{a}))$ can be seen as an inner product between vectors characterized by action $\boldsymbol{a}$ and reward function $r^f(\cdot)$. Therefore, under bandit feedback, we can treat it as a linear bandit and use ridge regression to build the reward estimator $\tilde{r}_i^k$ and corresponding bonus term $\tilde{b}^{k,\mathrm{r}}$, whose index

Algorithm 5 ϵ-approximate Nash Equilibrium for Potential Games

Input: ϵ, accuracy parameter; full information potential game $(\{\mathcal{A}_i\}_{i=1}^m, \{r_i\}_{i=1}^m)$ such that $r_i \in [0, r_{\max}]$ for all $i \in [m]$
Initialize: $\pi^1 = \boldsymbol{a}^1$, arbitrary deterministic product policy
for round $k = 1, \ldots, \left\lceil \frac{mr_{\max}}{\epsilon} \right\rceil$ **do**
  for player $i = 1, \ldots, m$ **do**
    $\Delta_i = \max_{a_i \in \mathcal{A}_i} r_i(a_i, \pi_{-i}^k) - r_i(\pi^k)$
    $a_i^{k+1} = \arg\max_{a \in \mathcal{A}_i} r_i(a_i, \pi_{-i}^k) - r_i(\pi^k)$
  end for
  if $\max_{i \in [m]} \Delta_i \leq \epsilon$ **then**
    return π^k
  end if
  $j = \arg\max_{i \in [m]} \Delta_i$
  $\pi^{k+1}(j) = a_j^{k+1}$, $\pi^{k+1}(i) = \pi^k(i)$, for all $i \neq j$
end for

i is dropped since it is the same for all players. The new algorithm will use these two terms to replace $\hat{r}_i^k$ and $b_i^{k,\mathrm{r}}$ in line 4 of Algorithm 4.

In particular, define $\theta \in [0,1]^{\tilde{d}}$ with $\tilde{d} = mF$ to be the vector such that $r^f(n) = \theta_{n+m(f-1)}$. Meanwhile, for player $i \in [m]$, define $A_i : \mathcal{A} \mapsto \{0,1\}^{\tilde{d}}$ to be the vector-valued function such that

$$[A_i(\boldsymbol{a})]_j = \mathbb{1}\left\{ j = n + m(f-1), f \in a_i, n = n^f(\boldsymbol{a}) \right\}.$$

In other words, $A_i(\boldsymbol{a})$ is a 0-1 vector with element 1 only at indices corresponding to those in θ that represents $r^f(n)$ for $f \in a_i$ and $n = n^f(\boldsymbol{a})$. Now, with these definitions, the reward function can be written as $r_i(\boldsymbol{a}) = \langle A_i(\boldsymbol{a}), \theta \rangle$. Then, we build the reward estimator and the bonus term through ridge regression and corresponding confidence bound, which are defined as the following:

$$\tilde{r}_i^k(\boldsymbol{a}) = \left\langle A_i(\boldsymbol{a}), \widehat{\theta}^k \right\rangle, \quad \tilde{b}^{k,\mathrm{r}}(\boldsymbol{a}) = \max_{i \in [m]} \|A_i(\boldsymbol{a})\|_{(V^k)^{-1}} \sqrt{\tilde{\beta}_k}, \tag{3.2}$$

where $\widehat{\theta}^k = (V^k)^{-1} \sum_{k'=1}^{k-1} \sum_{i=1}^m A_i(\boldsymbol{a}^{k'}) r_i^{k'}$, $V^k = I + \sum_{k'=1}^{k-1} \sum_{i=1}^m A_i(\boldsymbol{a}^{k'}) A_i(\boldsymbol{a}^{k'})^\top$ and $\sqrt{\tilde{\beta}_k} = \sqrt{\tilde{d}} + \sqrt{F\tilde{d} \log\left(1 + \frac{mkF}{\tilde{d}}\right) + F\tilde{\iota}}$. Note that we cannot bound the sum of this bonus

terms by directly applying the elliptical potential lemma. We instead prove its variant in Lemma B.3.2.

3.4.3 Regret Analysis

The Nash regret bounds for the two versions of Algorithm 4 are formally presented in Theorem 3.4.1. The proof details are deferred to Appendix B.3.

Theorem 3.4.1. *Let $\epsilon = 1/K$. For congestion games with semi-bandit feedback, by running Algorithm 4 with reward estimator and bonus term in (3.1), with probability at least $1-\delta$, we can achieve that*

$$\textit{Nash-Regret}(K) \leq \widetilde{\mathcal{O}}\left(F\sqrt{mK}\right).$$

Furthermore, if we only have bandit feedback, then by running Algorithm 4 with reward estimator and bonus term in (3.2), with probability at least $1-\delta$, we can achieve that

$$\textit{Nash-Regret}(K) \leq \widetilde{\mathcal{O}}\left(mF^{3/2}\sqrt{K}\right).$$

Remark 3.4.2. Since each action is a subset of $\mathcal{F}$, the size of each player's action space can be 2^F. As a result, directly applying Nash-VI in [Liu et al., 2021] leads to a regret bound exponential in F.

Remark 3.4.3. Note that we assume $r^f \in [0,1]$, which implies $r_i \in [0,F]$ for each player $i \in [m]$.

3.5 Decentralized Algorithms for Congestion Games

In this section, we present a decentralized algorithm for congestion games. Due to limited space, we only introduce the version of bandit feedback as in Section 3.4.2. The algorithmic details for the semi-bandit feedback setting are deferred into Appendix B.4.3. We will show that under both settings, even though each player can only observe her own actions and rewards, our decentralized algorithm still enjoys sublinear Nash regret with polynomial dependence on m and F.

Algorithm 6 Frank-Wolfe with Exploration for Congestion Game

1: **Input:** γ, ν, mixture weights; π_i^1, initial policy.
2: **Initialize:** ρ_i, the G-optimal design for player i, defined in (3.5).
3: **for** episode $k = 1, \cdots, K$ **do**
4: **for** round $t = 1, \cdots, \tau$ **do**
5: Each player takes action $a_i^{k,t} \sim \pi_i^k$, observes reward $r_i^{k,t}$.
6: **end for**
7: **for** player $i = 1, \cdots, m$ **do**
8: Compute $\widehat{\nabla}_i^k \Phi(a_i)$ by the formula in (3.4) for all $a_i \in \mathcal{A}_i$
9: Compute $\widetilde{\pi}_i^{k+1} \leftarrow \arg\max_{\pi_i \in \Delta(\mathcal{A}_i)} \left\langle \pi_i, \widehat{\nabla}_i^k \Phi \right\rangle$
10: Update $\pi_i^{k+1} \leftarrow (1-\gamma)(\nu \widetilde{\pi}_i^{k+1} + (1-\nu)\pi_i^k) + \gamma \rho_i$
11: **end for**
12: **end for**

We first define the vector-valued function $\phi_i : \mathcal{A}_i \mapsto \{0,1\}^{F_i}$ to be the feature map of player i such that $[\phi_i(a_i)]_f = \mathbb{1}\{f \in a_i\}$ for $a_i \in \mathcal{A}_i$ and $f \in \bigcup_{a_i \in \mathcal{A}_i} a_i$. Here, F_i is the size of $\bigcup_{a_i \in \mathcal{A}_i} a_i \subseteq \mathcal{F}$ and we can immediately see that $F_i \leq F$ for any $i \in [m]$.

The core idea of our algorithm is that the Nash equilibrium can be found by reaching the stationary points of the potential function since all congestion games are potential games. Here, the UCB-like algorithms used in the centralized setting are not applicable because their policy computation requires value functions for all players (e.g., line 6 of Algorithm 4), which are not available in the decentralized setting. Summarized in Algorithm 6, the decentralized algorithm is developed based on the Frank-Wolfe method and has the following three major components.

Gradient Estimator. In line 8, the algorithm builds the estimator $\widehat{\nabla}_i^k \Phi$ defined in (3.4) by using the τ reward samples collected from line 5. Here, $\widehat{\nabla}_i^k \Phi$ estimates the gradient of potential function Φ with respect to the policy π_i^k. Recall that for a congestion game, we have $\Phi(\boldsymbol{a}) = \sum_{f \in \mathcal{F}} \sum_{i=1}^{n^f(\boldsymbol{a})} r^f(i)$ and $\Phi(\pi) = \mathbb{E}_{\boldsymbol{a} \sim \pi}[\Phi(\boldsymbol{a})]$. Then we can define $\nabla_i \Phi := \nabla_{\pi_i} \Phi$ as a vector of dimension $|\mathcal{A}_i|$. For the component indexed by some $a_i \in \mathcal{A}_i$, we can see that $\Phi(\pi) = \pi_i(a_i) \mathbb{E}_{a_{-i} \sim \pi_{-i}}[r_i(a_i, a_{-i})] + \text{const}$, where const does not depend on $\pi_i(a_i)$. Therefore,

we have

$$\nabla_i \Phi(a_i) = \mathbb{E}_{a_{-i}\sim\pi_{-i}}\left[r_i(a_i, a_{-i})\right] = \mathbb{E}_{a_{-i}\sim\pi_{-i}}\left[\sum_{f\in a_i} r^f(n^f(a_i, a_{-i}))\right] = \left\langle \phi_i(a_i), \theta_i(\pi)\right\rangle, \quad (3.3)$$

where $[\theta_i(\pi)]_f = \mathbb{E}_{a_{-i}\sim\pi_{-i}}\left[r^f(n^f(a_{-i})+1)\right]$. Meanwhile, the mean of the t-th reward that player i received at episode k satisfies

$$\mathbb{E}\left[r_i^{k,t} \mid \boldsymbol{a}^{k,t}\right] = r_i(\boldsymbol{a}^{k,t}) = \sum_{f\in a_i^{k,t}} r^f(n^f(\boldsymbol{a}^{k,t})) = \left\langle \phi_i(a_i^{k,t}), \theta_i^{k,t}(a_{-i}^{k,t})\right\rangle,$$

where $[\theta_i^{k,t}(a_{-i}^{k,t})]_f = r^f(n^f(a_{-i}^{k,t})+1)$ and its mean is $[\theta_i(\pi^k)]_f$. Therefore, we can use linear regression to estimate $\theta_i(\pi^k)$. In particular, we have $\widehat{\theta}_i^k(\pi^k) = \frac{1}{\tau}\sum_{t=1}^{\tau}\left(\Sigma_i^k\right)^{-1}\phi_i(a_i^{k,t})r_i^{k,t}$, with the covariance matrix $\Sigma_i^k = \mathbb{E}_{a_i\sim\pi_i^k}\left[\phi_i(a_i)\phi_i(a_i)^\top\right]$. Then, we have the unbiased gradient estimate

$$\widehat{\nabla}_i^k \Phi(a_i) = \left\langle \phi_i(a_i), \widehat{\theta}_i^k(\pi^k)\right\rangle = \frac{1}{\tau}\sum_{t=1}^{\tau}\phi_i(a_i)^\top\left(\Sigma_i^k\right)^{-1}\phi_i(a_i^{k,t})r_i^{k,t}. \quad (3.4)$$

Remark 3.5.1. One difference between Algorithm 6 (decentralized) and Algorithm 4 (centralized) is that in the decentralized algorithm, each player is required to play the same policy for τ times before an update can be applied. An episode is thus defined for convenience as the time period during which the players' policies are fixed. We make this artificial design mainly for controlling the variance of the gradient estimator $\widehat{\nabla}_i^k\Phi(a_i)$. However, we conjecture that with more careful design and analysis, it should be possible to improve Algorithm 6 so that only one sample is required per episode [Zhang et al., 2020a].

G-optimal Design. In line 9 and 10, the algorithm performs standard Frank-Wolfe update and mixes the updated policy with an exploration policy ρ_i, which is defined as the G-optimal allocation for features $\{\phi_i(a_i)\}_{a_i\in\mathcal{A}_i}$. To be specific, we have

$$\rho_i = \arg\min_{\lambda\in\Delta(\mathcal{A}_i)}\max_{a_i\in\mathcal{A}_i}\|\phi_i(a_i)\|^2_{\mathbb{E}_{a_i'\sim\lambda}\left[\phi_i(a_i')\phi_i(a_i')^\top\right]^{-1}}. \quad (3.5)$$

Here ρ_i guarantees that Σ_i^k is invertible and the variance of $\widehat{\nabla}_i^k \Phi(a_i) = \left\langle \phi_i(a_i), \widehat{\theta}_i^k(\pi^k) \right\rangle$ depends only on F instead of the size of action space (Lemma B.4.3) because by the famous Kiefer-Wolfowitz theorem, we have $\max_{a_i \in \mathcal{A}_i} \|\phi_i(a_i)\|^2_{\mathbb{E}_{a_i' \sim \rho_i}[\phi_i(a_i')\phi_i(a_i')^\top]^{-1}} = F_i \leq F$ [Lattimore and Szepesvári, 2020].

Frank-Wolfe Update. Finally, we emphasize that it is crucial to use Frank-Wolfe update because it is compatible with L_1 *norm* and we can show that Φ is mF-smooth with respect to the L_1 norm (Lemma B.4.5). In contrast, its smoothness for L_2 norm will depend on the size of the action space.

Before the game starts, each player i can compute her ρ_i based on her own action set $\mathcal{A}_i$. During the game, all players only have access to their own actions and rewards, which means that Algorithm 6 is fully decentralized. The Nash regret bound for this algorithm is formally stated in Theorem 3.5.2 and the proof details are given in Appendix B.4.1 and B.4.2.

Theorem 3.5.2. *Let $T = K\tau$. For congestion game with bandit feedback, by running Algorithm 6 with gradient estimator $\widehat{\nabla}_i^k \Phi$ in (3.4) and exploration distribution ρ_i in (3.5), if $K \geq \frac{2F}{m}$, then with probability at least $1-\delta$, we have*

$$\textit{Nash-Regret}(T) := \sum_{k=1}^{K} \tau \max_{i \in [m]} \left(V_i^{\dagger, \pi_{-i}^k} - V_i^{\pi^k} \right) \leq \widetilde{\mathcal{O}} \left(m^2 F^2 T^{5/6} + m^3 F^3 T^{2/3} \right).$$

For congestion game with semi-bandit feedback, by running Algorithm 6 with gradient estimator $\widetilde{\nabla}_i^k \Phi(a_i)$ and exploration distribution $\tilde{\rho}_i$ defined in Appendix B.4.3, if $K \geq \frac{2\sqrt{F}}{m}$, then with probability at least $1-\delta$, we have

$$\textit{Nash-Regret}(T) \leq \widetilde{\mathcal{O}} \left(m^2 F^{3/2} T^{5/6} + m^3 F^2 T^{2/3} \right).$$

3.6 Extension to Independent Markov Congestion Games

In this section, we propose and analyze a Markov extension of the congestion games, called the independent Markov congestion games (IMCGs).

3.6.1 Problem Formulation

General-sum Markov Games. A finite-horizon time-inhomogeneous tabular general-sum Markov game is defined by $\mathcal{M} = \{\mathcal{S}, \{\mathcal{A}_i\}_{i=1}^m, H, P, R, s_0\}$, where $\mathcal{S}$ is the state space, m is the number of players, $\mathcal{A}_i$ is the action space of player i, $\mathcal{A} = \mathcal{A}_1 \times \cdots \times \mathcal{A}_m$ is the whole action space, H is the time horizon, s_0 is the initial state*, $P = (P_1, P_2, \cdots, P_H)$ with $P_h \in [0,1]^{S\times A\times S}$ as the transition kernel at timestep h, $R = \{R_h(\cdot|s_h, \boldsymbol{a}_h)\}_{h=1}^H$ with $R_h(\cdot|s_h, \boldsymbol{a}_h)$ as the reward distribution on $[0, r_{\max}]^m$ with mean $\boldsymbol{r}_h(s_h, a_h) \in [0, r_{\max}]^m$ at timestep $h \in [H]$. At timestep h, all players choose their actions simultaneously and a reward vector is sampled $\boldsymbol{r}_h \sim R_h(\cdot|s_h, \boldsymbol{a}_h)$, where s_h is the current state and $\boldsymbol{a}_h = (a_{h,1}, a_{h,2}, \cdots, a_{h,m})$ is the joint action. Each player i receives reward $r_{h,i}$ and the state transits to $s_{h+1} \sim P_h(\cdot|s_h, \boldsymbol{a}_h)$. The objective for each player is to maximize her own total reward. We assume that the initial state s_1 is fixed.

A (Markov) policy π is a collection of H functions $\{\pi_h : \mathcal{S} \mapsto \Delta(\mathcal{A})\}_{h=1}^H$, each of which maps a state to a distribution over the action space. π is a product policy if $\pi_h(\cdot \mid s)$ is a product policy for each $(h, s) \in [H] \times \mathcal{S}$. The value function and Q-value function of player i at timestep h under policy π are defined as

$$V_{h,i}^{\pi}(s) = \mathbb{E}_{\pi}\left[\sum_{h'=h}^{H} r_{h',i}(s_{h'}, \boldsymbol{a}_{h'}) \mid s_h = s\right], \; Q_{h,i}^{\pi}(s, \boldsymbol{a}) = \mathbb{E}_{\pi}\left[\sum_{h'=h}^{H} r_{h',i}(s_{h'}, \boldsymbol{a}_{h'}) \mid s_h = s, \boldsymbol{a}_h = a\right].$$

The best responses and Nash regret can be defined similarly as those for matrix games. In particular, given a policy π, player i's best response policy is $\pi_{h,i}^{\dagger}(\cdot \mid s) = \arg\max_{\mu\in\Delta(\mathcal{A}_i)} V_{h,i}^{\mu,\pi_{-i}}(s)$ and the corresponding value function is denoted as $V_{h,i}^{\dagger,\pi_{-i}}$.

Definition 3.6.1. With π^k being the policy at kth episode, the *Nash regret* after K episodes is define as

$$\text{Nash-Regret}(K) = \sum_{k=1}^{K} \max_{i\in[m]} \left(V_{1,i}^{\dagger,\pi_{-i}^k} - V_{1,i}^{\pi^k}\right)(s_1).$$

Independent Markov Congestion Game. A general-sum Markov game is an independent Markov congestion game (IMCG) if there exists a facility set $\mathcal{F}$ such that $a_i \subseteq \mathcal{F}$ for

*An episode is defined as running H steps from the initial state s_0, which is common for the episodic MDP.

any $a_i \in \mathcal{A}_i$, a state space $\mathcal{S} = \prod_{f\in\mathcal{F}} \mathcal{S}^f$, a set of facility reward distributions $\{R_h^f\}_{h\in[H],f\in\mathcal{F}}$ such that if the joint action at s_h is $\boldsymbol{a}$, we have $r_{h,i} = \sum_{f\in a_i} r_h^f$, where $r_h^f \sim R_h^f(\cdot|s_h, n^f(\boldsymbol{a}))$ with support on $[0,1]$ and mean $r_h^f(s_h, n^f(\boldsymbol{a}))$, and a set of transition matrices $\{P_h^f\}_{h\in[H],f\in\mathcal{F}}$ such that $P_h(s'|s,\boldsymbol{a}) = \prod_{f\in\mathcal{F}} P_h^f(s'^f|s^f, n^f(\boldsymbol{a}))$. In other words, at each timestep h and state $s \in \mathcal{S}$, the players are in a congestion game. Meanwhile, each facility has its own state and independent state transition, which only depends on its current state and number of players using that facility. This transition kernel can be viewed as a special case of that in factored MDPs [Szita and Lőrincz, 2009]. The IMCG also admits two types of feedback, semi-bandit feedback and bandit feedback, just like the congestion game. In this paper, we will consider both types of feedback.

3.6.2 Theoretical Guarantee

Summarized in Algorithm 15, our centralized algorithm for IMCGs is naturally extended from the Nash-UCB (Algorithm 4) by incorporating transition kernel estimators, corresponding bonus terms and Bellman backward update. The key idea is to utilize the independent transition structure to remove the dependence on the exponential size of the state space $S = \prod_{f\in\mathcal{F}} S^f$. We tackle this issue by adapting technique from factored MDP [Chen et al., 2020]. The algorithmic details for both types of feedback are deferred into Appendix B.5. The Nash regret bounds for the two versions of Algorithm 15 are stated in Theorem 3.6.2 and the proof details are deferred to Appendix B.6.

Theorem 3.6.2. *For independent Markov congestion game with semi-bandit feedback, by running the centralized Algorithm 15, with probability at least $1-\delta$, we can achieve that*

$$\textit{Nash-Regret}(K) \le \widetilde{\mathcal{O}}\left(\sum_{f\in\mathcal{F}} FS^f\sqrt{mH^3T}\right) + \widetilde{\mathcal{O}}\left(m^2H^2F\sum_{f\neq f'}\left(S^fS^{f'}\right)^2\right).$$

Furthermore, if we only have bandit feedback, then by running Algorithm 15 with reward estimator and bonus term in (B.7) and (B.8), with probability at least $1-\delta$, we can achieve

that

$$\textit{Nash-Regret}(K) \leq \widetilde{\mathcal{O}}\left(\sum_{f\in\mathcal{F}} FS^f\sqrt{m^2H^3T}\right) + \widetilde{\mathcal{O}}\left(m^2H^2F\sum_{f\neq f'}\left(S^fS^{f'}\right)^2\right).$$

The regret bound in [Liu et al., 2021] is $\widetilde{O}(\sqrt{H^3S^2(\Pi_{i=1}^m A_i)T})$, where both A_i and $S = \prod_{f\in\mathcal{F}} S^f$ can be exponential in F. Our bounds have polynomial dependence on all the parameters.

3.7 Conclusion

In this paper, we study sample-efficient learning in congestion games by utilizing the special reward structure. We propose both centralized and decentralized algorithms for congestion games with two types of feedback, all achieving sample complexities only polynomial in the number of facilities. To the best of our knowledge, each one of them is the first sample-efficient learning algorithm for congestion games in its own setting. We further define the independent Markov congestion game (IMCG) as a natural extension of the congestion game into the Markov setting together with a sample-efficient centralized algorithm for both types of feedback.

One promising future direction is to find a sample-efficient decentralized algorithm such that from each player's own perspective, the algorithm is still no-regret. In other words, diminishing regret is guaranteed for the player by running this algorithm even though other players may use policies from different algorithms. Another important future direction is to find sample-efficient centralized/decentralized algorithms that can explicitly find an approximate Nash equilibrium policy.

Chapter 4

RANDOMIZED EXPLORATION IN REINFORCEMENT LEARNING

This chapter is based on Xiong et al. [2022], with Qiwen Cui, Ruoqi Shen, Maryam Fazel and Simon S. Du.

4.1 Introduction

This paper concerns learning in tabular Markov Decision Processes (MDP), arguably the most fundamental model for reinforcement learning (RL). Existing algorithms that achieve the near-optimal minimax $\widetilde{O}\left(H\sqrt{SAT}\right)$ regret bound are based on the principle of *Optimism in the face of Uncertainty* (OFU), such as upper confidence bound (UCB) [Azar et al., 2017, Zanette and Brunskill, 2019, Dann et al., 2019, Zhang et al., 2020d,b].* Here S is the number of states, A is the number of actions, H is the planning horizon, and T is the total number of interactions between the agent and the environment.

Another broad category is algorithms with randomized exploration such as Thompson Sampling [Osband et al., 2013, Agrawal and Jia, 2017a, Osband et al., 2014]. These algorithms inject (carefully tuned) random noise to value function to encourage exploration. UCB-type algorithms enjoy well-established theoretical guarantees but suffer from difficult implementation since an upper confidence bound is usually infeasible for many practical models like neural networks. Instead, practitioners prefer randomized exploration such as noisy networks in Fortunato et al. [2018], and algorithms with randomized exploration have been widely used in practice [Osband et al., 2017, Chapelle and Li, 2011, Burda et al., 2018, Osband et al., 2018]. However, how to design randomized exploration algorithms in a principled way and perform randomized exploration optimally is far from clear. While randomized exploration can have great performance in practice, theoretically, the best known worst-case regret bound for algorithms with randomized exploration is $\widetilde{O}\left(H^2S\sqrt{AT}\right)$ [Agrawal et al.,

*This bound is for time-inhomogeneous MDP with each reward bounded by 1 and T is sufficiently large.

2021], which is far worse than that of the UCB-type algorithms. In this paper, we introduce a new randomized exploration algorithm and show it enjoys a near-optimal $\widetilde{O}\left(H\sqrt{SAT}\right)$ worst-case regret bound, thus closing the gap. Our work sheds new light on randomized exploration on both the algorithmic side and the theoretical side.

Our Contributions. Our contributions are summarized below:

• We propose a new algorithm, Single Seed Randomization (SSR), which incorporates a crucial algorithmic idea: using a single random seed for the entire episode, in contrast to previous methods of randomized exploration which use one seed for each time step. SSR is able to explore more efficiently than previous methods by avoiding having noise at different time steps canceling with each other. Theoretically, we show, thanks to this new idea, if one uses a Hoeffding-type magnitude of noise, SSR achieves an $\widetilde{O}\left(H^{1.5}\sqrt{SAT}\right)$ regret bound, improving upon the best existing result on randomized exploration algorithm [Agrawal et al., 2021].

• We further design a new Bernstein-type magnitude of noise for our algorithm, and achieve an $\widetilde{O}\left(H\sqrt{SAT}\right)$ regret bound, resolving an open problem raised in Agrawal et al. [2021]. To our knowledge, this is the first time that a Bernstein-type bound is used in randomized exploration. More importantly, our upper bound matches the $\Omega\left(H\sqrt{SAT}\right)$ minimax lower bound up to logarithmic factors.

We note that our goal is not to show randomized exploration is better than optimistic algorithms [Azar et al., 2017] in the tabular setting. Instead, we aim to provide a solid theoretical understanding of a practically relevant algorithm. Indeed, understanding randomized exploration itself is an important theoretical research direction and has attracted much interest in the community [Agrawal and Goyal, 2012, 2017, Agrawal and Jia, 2017b, Russo, 2019, Zanette et al., 2020, Vaswani et al., 2020, Agrawal et al., 2021, Osband et al., 2013, 2014, 2017, 2018].

Main Challenge and Technical Overview. Besides the aforementioned algorithmic ideas (single random seed and Bernstein-type magnitude of noise), we also need additional ideas in analysis to prove the desired regret bound. The main challenge is that unlike UCB-

type algorithms, the estimated value in algorithm with randomized exploration, is not an upper bound of the true optimal value. This leads to the failure of directly utilizing their analysis, which only need to analyze the one-sided error in estimation. We instead work on the *absolute value* of the estimation error, whose analysis is more complicated than that for the one-sided error in UCB-type algorithms. Working with absolute value forces us to ensure that both the probability that the estimated value is optimistic and the probability that the estimated value is pessimistic are lower bounded. However, the clipping strategy in existing algorithm cannot maintain pessimism. To tackle with this issue, we develop a new clipping method. Below we list our technical contributions.

1. First, we propose a new clipping strategy to constrain the estimated value function (cf. Eqn. (4.4)). Previous clipping strategies in [Zanette et al., 2020, Agrawal et al., 2021] are based on uncertainty and can only maintain optimism. Our clipping strategy directly works on the value function, which is similar to those used in UCB-type algorithms [Azar et al., 2017, Jin et al., 2018, Zhang et al., 2020d]. Our clipping strategy can maintain both the optimism and pessimism. In addition, the number of times that the clipping is used can still be bounded.

2. Second, we prove that the single seed randomization ensures that the estimated value function can both be optimistic or pessimistic with constant probability at all states and timesteps. This is stronger than previous randomized exploration algorithms that are only shown to be optimistic at the initial state with constant probability. With this property, we can then bound the difference between the optimal value function and estimated value function from both above and below, which results in a bound on its absolute value. See Section 4.5.1, Appendix C.3 and Appendix C.4.

3. Third, we prove a novel recursion argument on the absolute value of the policy estimation error. As mentioned in [Agrawal et al., 2021], the recursion in UCB-type algorithms can not be directly utilized because our estimated value function is not a high-probability upper bound of the true optimal value function. With the bound of absolute value, we are able to prove new recursion formulas and together we can control the policy estimation error. See Section 4.5.2 and Appendix C.5.

4. At last, we bound the sum of variance in a novel manner. In [Azar et al., 2017], the UCB-type estimation guarantees that the policy estimation error is always positive so the difference of the variance can be directly bounded. We generalize the argument to the absolute value of the estimation error to bound the sum of variance. See Section 4.5.3 and Appendix C.7.

4.2 Related Work

In this section we review existing provably efficient algorithms for tabular MDP. There is a long list of sample complexity guarantees for tabular MDP [Kearns and Singh, 2002, Brafman and Tennenholtz, 2003, Kakade, 2003, Strehl et al., 2006, Strehl and Littman, 2008, Kolter and Ng, 2009, Bartlett and Tewari, 2009, Jaksch et al., 2010, Szita and Szepesvári, 2010, Lattimore and Hutter, 2012, Osband et al., 2013, Dann and Brunskill, 2015, Azar et al., 2017, Dann et al., 2017, Osband and Van Roy, 2017, Agrawal and Jia, 2017a, Jin et al., 2018, Fruit et al., 2018, Talebi and Maillard, 2018, Dann et al., 2019, Dong et al., 2019, Simchowitz and Jamieson, 2019, Russo, 2019, Zhang and Ji, 2019, Cai et al., 2019, Zhang et al., 2020c, Yang et al., 2020, Pacchiano et al., 2020, Neu and Pike-Burke, 2020, Zhang et al., 2020b, Wang et al., 2020, Agrawal et al., 2021, Russo, 2019, Agrawal and Jia, 2017a, Domingues et al., 2021, Menard et al., 2021, Li et al., 2021]. The state-of-the-art methods are based on upper confidence bound (UCB) [Azar et al., 2017, Zanette and Brunskill, 2019, Dann et al., 2019, Zhang et al., 2020d,b, Menard et al., 2021, Li et al., 2021]. For the setting considered in this paper where the transition is time-inhomogeneous and the reward is bounded by 1, one can achieve an $\widetilde{O}\left(H\sqrt{SAT}\right)$ in the regime where T is sufficiently large.

Algorithms with randomized exploration have been proved to enjoy favorable regret bounds in bandit problems [Lai and Robbins, 1985, Agrawal and Goyal, 2012, Kaufmann et al., 2012, Bubeck and Liu, 2014, Agrawal and Goyal, 2017]. In certain settings, randomized exploration can match the worst-case regret bound of UCB-based approaches and achieve nearly minimax optimal regret bounds [Jin et al., 2020, Agrawal and Goyal, 2017]. However, for RL, existing theory for randomized exploration are far from optimal [Agrawal et al., 2021, Russo, 2019, Agrawal and Jia, 2017a, Xu and Tewari, 2019, Zanette et al., 2020]. For the setting considered in this paper, the sharpest existing regret bound among algorithms with

randomized exploration is $\widetilde{O}\left(H^2S\sqrt{AT}\right)$ proved in [Agrawal et al., 2021]. Our paper closes this gap and thus deepens our understanding about randomized exploration.

4.3 Preliminaries

We consider time-inhomogeneous finite-horizon MDP $M = (H, \mathcal{S}, \mathcal{A}, P, R, s_1)$, where $|\mathcal{S}| = S$ and $|\mathcal{A}| = A$. Here, $\mathcal{S} = \{1, ..., S\}$ is the finite state space. $\mathcal{A} = \{1, ..., A\}$ is the finite action space. H is the length of an episode. For convenience, we take s_1 to be the fixed initial state, although a more general initial distribution will not change the conclusion. $P : \mathcal{S} \times \mathcal{A} \times [H] \to \triangle(\mathcal{S})$ is the transition function, where if the agent stays at state s and takes action a at time h, it transits to state s' with probability $P_{h,s,a}(s') \in [0, 1]$. $R : \mathcal{S} \times \mathcal{A} \times [H] \to [0, 1]$ is the reward function, where if the agent stays at s and takes action a at time h, it will receive reward $r_{h,s,a} \in [0, 1]$ such that $\mathbb{E}[r_{h,s,a}] = R_{h,s,a}$.

A deterministic policy for such a MDP is defined as a tuple $\pi = (\pi_1, \ldots, \pi_H)$, where $\pi_h : \mathcal{S} \mapsto \mathcal{A}$. The associated value function at state $s \in \mathcal{S}$ and level $h \in \{1, \ldots, H\}$ is recursively defined as

$$V_h^\pi(s) = R_{h,s,\pi_h(s)} + \sum_{s' \in \mathcal{S}} P_{h,s,\pi_h(s)}(s') V_{h+1}^\pi(s').$$

For convenience, we set $V_{H+1}^\pi = \mathbf{0} \in \mathbb{R}^S$. The corresponding optimal value function is $V_h^*(s) = \max_{\pi \in \Pi} V_h^\pi(s)$, where Π is the set of all possible deterministic policies. For a particular algorithm Alg, let π^k denote the policy that Alg employs during episode k. Then, the regret of running Alg on MDP M for K episodes is defined as

$$\mathrm{Reg}(M, K, \mathsf{Alg}) = \sum_{k=1}^{K} \left(V_1^*(s_1) - V_1^{\pi^k}(s_1)\right). \tag{4.1}$$

Note that the regret, $\mathrm{Reg}(M, K, \mathsf{Alg})$, is a random variable due to randomness in state transition and the algorithm, Alg. In this paper, we show the regret of our proposed algorithm can be upper bounded with high probability, and the upper bound matches the known lower bound up to logarithmic factors.

To facilitate our later analysis, we introduce some notations for empirical estimation. At

episode k, we collect a trajectory $(s_1^k, a_1^k, r_1^k, \cdots, s_H^k, a_H^k, r_H^k)$ as specified in Algorithm 7. Let $n_k(h,s,a) = \sum_{l=1}^{k-1} \mathbb{1}\{(s_h^l, a_h^l) = (s,a)\}$ be the number of times action a is taken at state s and time h before episode k, where $\mathbb{1}\{\cdot\}$ is the indicator function. We define

$$\hat{R}_{h,s,a}^k = \frac{\sum_{l=1}^{k-1} \mathbb{1}\{(s_h^l, a_h^l) = (s,a)\} r_{h,s_h^l,a_h^l}^l}{n_k(h,s,a)+1}, \tag{4.2}$$

$$\hat{P}_{h,s,a}^k(s') = \frac{\sum_{l=1}^{k-1} \mathbb{1}\{(s_h^l, a_h^l, s_{h+1}^l) = (s,a,s')\}}{n_k(h,s,a)+1}. \tag{4.3}$$

Then, define empirical MDP based on our observation and estimation before episode k as the tuple $\hat{M}^k = (H, \mathcal{S}, \mathcal{A}, \hat{P}^k, \hat{R}^k, s_1)$. Since $\hat{P}_{h,s,a}^k$ is not a valid distribution over $\mathcal{S}$, for being rigorous, we can imagine there is an additional virtual absorbing state that every state will transit to with remaining probability.

In addition to the above notations, let $\widetilde{O}(\cdot), \widetilde{\Theta}(\cdot)$ and $\widetilde{\Omega}(\cdot)$ be asymptotic notations ignoring all poly-logarithmic terms. For distribution $D \in \Delta^S$ and value function $V \in \mathbb{R}^S$, let $\mathbb{V}(D,V)$ denote the variance of V under distribution D, which is defined as $\mathbb{V}(D,V) = \sum_{s\in\mathcal{S}} D(s)(V(s) - \langle D, V\rangle)^2$. For constant $a > 0$, we define the corresponding clipping function as $\text{clip}_a(\cdot) = \max\{-a, \min\{a, \cdot\}\}$. Immediately we have $|\text{clip}_a(x)| \le a$ for any $a > 0$. We introduce the definitions of other notations when used. In appendix, we summarize the notations and definitions used in this paper.

4.4 Main Results

4.4.1 Algorithm

The main contribution of this paper is that we show algorithm with randomized value functions can achieve regret that matches the known lower bound $\Omega\left(H\sqrt{SAT}\right)$ [Jaksch et al., 2010, Domingues et al., 2021] up to logarithmic factors in the tabular setting. To facilitate exploration, this type of algorithms uses random value perturbation instead of deterministic bonus. The algorithm we consider is summarized in Algorithm 7. In our algorithm, SSR, the random perturbation ensures that optimism/pessimism can be obtained with constant probability in each episode. Moreover, randomized value function has its origin from posterior sampling for reinforcement learning (Thompson sampling). The randomized

Algorithm 7 Single Seed Randomization (SSR)

Input: ty $\in \{\text{Ho}, \text{Be}\}$, perturbation type
for episode $k = 1, 2, \ldots, K$ **do**
  Sample $\hat{z}_k \sim \mathcal{N}(0, 1)$
  Define terminal value functions $\overline{Q}_{H+1,k} = \mathbf{0} \in \mathbb{R}^{SA}$ and $\overline{V}_{H+1,k} = \mathbf{0} \in \mathbb{R}^{S}$
  for time periods $h = H, \ldots, 1$ **do**
    $\overline{Q}_{h,k}(s,a) \leftarrow \hat{R}^k_{h,s,a} + \left\langle \hat{P}^k_{h,s,a}, \overline{V}_{h+1,k} \right\rangle + \sigma^k_{\text{ty}}(h,s,a)\,\hat{z}_k$
    // $\sigma^k_{\text{ty}}(h,s,a)$ is defined in (4.5) and (4.6).
    Define $\overline{V}_{h,k}(s) = \text{clip}_{2(H-h+1)}\left(\max_{a\in\mathcal{A}} \overline{Q}_{h,k}(s,a)\right)$ for all $s \in \mathcal{S}$
  end for
  Agent takes actions $a^k_h = \arg\max_{a\in\mathcal{A}} \overline{Q}_{h,k}(s^k_h, a)$ throughout the current episode
  Observe data $s^k_1, a^k_1, r^k_1, \ldots, s^k_H, a^k_H, r^k_H$ and compute $\hat{R}^{k+1}_{h,s,a}$, $\hat{P}^{k+1}_{h,s,a}$ and $n_{k+1}(h,s,a)$ for all $(h,s,a) \in [H] \times \mathcal{S} \times \mathcal{A}$
end for

perturbation can be interpreted as approximate sampling from the posterior distribution of the value function on randomized training data [Russo, 2019].

We first give an overview of SSR. In Algorithm 7, the policy used at episode k is computed using the empirical MDP, $\hat{M}^k = (H, \mathcal{S}, \mathcal{A}, \hat{P}^k, \hat{R}^k, s_1)$, which is based on observation and estimation before episode k. However, instead of directly choosing optimal policy for $\hat{M}^k$, we add a small random perturbation when computing the value of each state and action pair. To be more precise, at each episode k, we first estimate the reward and transition function for each state s and action a based on (4.2) and (4.3). Then, we compute the value function for state s and action a,

$$\overline{Q}_{h,k}(s,a) \leftarrow \hat{R}^k_{h,s,a} + \left\langle \hat{P}^k_{h,s,a}, \overline{V}_{h+1,k} \right\rangle + \sigma^k_{\text{ty}}(h,s,a)\,\hat{z}_k.$$

Here, $\hat{z}_k \sim \mathcal{N}(0,1)$ is a standard Gaussian random variable sampled once every episode. The magnitude of the perturbation, σ^k_{ty} depends on how many samples $n_k(h,s,a)$ we have observed and how confident we are on the estimations $\hat{R}^k_{h,s,a}$ and $\hat{P}^k_{h,s,a}$. We will discuss more about the choice of the magnitude later in this section.

In order to prevent estimated value function from behaving badly, we add a clipping to

the value function:

$$\overline{V}_{h,k}(s) = \text{clip}_{2(H-h+1)}\left(\max_{a\in\mathcal{A}} \overline{Q}_{h,k}(s,a)\right) \tag{4.4}$$

As our analysis will show, this kind of clipping can bound the value function, maintain optimism and pessimism and also guarantee that clipping will not happen for a lot of times. The constant 2 (instead of 1) plays a crucial role because it means the value function grows at an additive rate of 2 from $h = H$ to $h = 1$. If we do not consider the added noise, then the value function should at most grow 1 at each timestep because the reward is at most 1. For our clipping technique, if a clip is triggered, there exists a timestep such that the added noise is more than 1, which is equivalent to a small number of visits (cf. Definition C.8 and Lemma C.2.12). As our later analysis will show, the clipping only affects the lower-order term and will not compromise the long-term performance of the algorithm. Finally, after computing the value function and clipping, SSR chooses the action a_h^k that maximizes $\overline{Q}_{h,k}(s_h^k, a)$ at each time step, $h = 1, ..., H$, throughout the episode.

Note that from a Bayesian perspective, when there is no clipping, in Algorithm 7, $\overline{Q}_{h,k}$ follows distribution

$$\overline{Q}_{h,k}(s,a) \mid \overline{V}_{h+1,k} \sim \mathcal{N}\left(\hat{R}_{h,s,a}^k + \left\langle \hat{P}_{h,s,a}^k, \overline{V}_{h+1,k}\right\rangle, \left(\sigma_{\text{ty}}^k(h,s,a)\right)^2\right).$$

This resembles posterior sampling because when estimating some parameter $\theta^* \sim \mathcal{N}\left(0, \beta^2\right)$ based on noisy observations $\theta_1, ..., \theta_n \sim \mathcal{N}(\theta, \beta^2)$, the posterior distribution of θ^* given $\{\theta_i\}_{i=1}^n$ is $\theta^* \mid \{\theta_i\}_{i=1}^n \sim \mathcal{N}\left(\frac{1}{n+1}\sum_{i=1}^n \theta_i, \frac{\beta^2}{n+1}\right)$. Although exact posterior sampling may not be possible in complex reinforcement learning settings, in SSR, $\sigma_{\text{ty}}^k(h,s,a)$ is chosen at scale $\widetilde{\Theta}\left(1/\sqrt{n_k(h,s,a)}\right)$ and therefore can be interpreted as doing approximate posterior sampling. Moreover, SSR can be viewed as a variant of Randomized Least Square Value Iteration (RLSVI). The major differences are at the clipping function and a single random seed used in each episode instead of different random seeds at different tuples (h,s,a). We will discuss more about the choice of the random seed later in this section. We refer to Osband et al. [2017] and Russo [2019] for a more detailed discussion on the relationship

among RLSVI, posterior sampling and randomized value function.

In the following paragraphs, we discuss in more details about the three major algorithmic innovations:

Single Random Seed in Each Episode. SSR is similar to the algorithms analyzed in Russo [2019] and Agrawal et al. [2021]. The major difference is that in the algorithm we propose, we use a single random seed $\hat{z}_k$ to generate the perturbations for all time steps $h = 1, ..., H$ in an episode k.

When using different random seeds in an episode, the algorithm can be optimistic in some time step while being pessimistic in others. Then, the effects of the perturbations at different time steps will cancel with each other. As a result, to ensure sufficient exploration, the magnitude of the perturbation has to large. This issue was also pointed out in Agrawal et al. [2021], Abeille and Lazaric [2017].

A large perturbation magnitude can increase the instability of the algorithm and worsen the algorithm's performance. When a single random seed is used, a small perturbation magnitude is enough to guarantee that the algorithm is optimistic with constant probability in any episode. We are able to show that using a single random seed can significantly increase the stability of the algorithm and therefore enjoy much smaller regret. Coincidentally, Vaswani et al. [2020] also uses a similar single randomization in bandit problems to build a near-optimal randomized exploration algorithm and our work can be treated as its natural extension to RL problems.

Clipping. To obtain a tight regret bound, the estimated value function needs to be well bounded. In [Russo, 2019], no clipping is used and the estimated value function is at the order of $\widetilde{O}(H^{5/2}S)$, which results in a suboptimal regret bound. Generally there are two types of clipping methods. The first one is uncertainty-based, i.e. the value is clipped to $H - h + 1$ at timestep h whenever the uncertainty is large [Zanette et al., 2020, Agrawal et al., 2021]. However this type of clipping cannot maintain pessimism which is critical in our analysis. The other kind of clipping is value-based, mostly in UCB-type algorithms [Jin et al., 2019]. These algorithms truncate estimated value greater than a certain threshold,

i.e. $H-h+1$ at time step h. The problem here is that the number of clippings cannot be bounded because if the true value function is close to $H-h+1$ at timestep h, the clipping will happen with some constant probability.

Our clipping method leverages both type of clipping methods in the existing literature. Though our clipping is based on the value function, we show that whenever the clip is triggered, the estimation error must be large, which implies that the uncertainty at that state is large. This clipping method inherits the desired properties from both uncertainty-based and value-based clipping, i.e. the optimism/pessimism is maintained and the number of clippings can be bounded.

Magnitude of Perturbation. A large magnitude of perturbation can encourage exploration, but at the same time increase instability. In our algorithm, the magnitudes are chosen as the smallest values so that the algorithm can be optimistic with constant probability. Since the value function can roughly be bounded by $O(H)$, a naive choice of the perturbation magnitude can be $\Theta\left(H/\sqrt{n_k(h,s,a)}\right)$. In this way, by Hoeffding's inequality, as long as the random Gaussian variable sampled $\hat{z}_k$ is bigger than a constant, which happens with constant probability, the estimated value function will be optimistic. By similar reasoning, we can see that the estimated value function will also be pessimistic with constant probability.

To make the magnitude even smaller, inspired by [Azar et al., 2017] who showed one can use an (empirical) Bernstein's inequality to derive a sharp exploration bonus for UCB-based algorithms, we propose a new choice of perturbation magnitude based on Bernstein's inequality. The Bernstein-based perturbation uses the empirical variance of the value function, which makes it smaller than the Hoeffiding-based one mostly, but still maintains optimism with constant probability.

In our paper, we study both types of magnitudes. In particular, we show that the regret of SSR based on Bernstein's inequality matches the known lower bound $\Omega\left(H\sqrt{SAT}\right)$. Following are the two choices:

$$\sigma_{\mathrm{Ho}}^{k}(h,s,a)=H\sqrt{\frac{\log\left(2HSAk^{2}\right)}{n_{k}(h,s,a)+1}}+\frac{H}{n_{k}(h,s,a)+1}, \tag{4.5}$$

$$\sigma_{\mathrm{Be}}^{k}(h,s,a)=\sqrt{\frac{16\mathbb{V}\left(\tilde{P}_{h,s,a}^{k},\overline{V}_{k,h+1}\right)\log\left(2HSAk^{2}\right)}{n_{k}(h,s,a)+1}}+\frac{65H\log\left(2HSAk^{2}\right)}{n_{k}(h,s,a)+1}+\sqrt{\frac{\log\left(2HSAk^{2}\right)}{n_{k}(h,s,a)+1}}, \tag{4.6}$$

where subscript "Ho" represents that the perturbation is based on Hoeffding's inequality and "Be" represents Bernstein's inequality, correspondingly. Here, for proof convenience, $\tilde{P}_{h,s,a}^{k}$ is defined by replacing the denominator in $\hat{P}_{h,s,a}^{k}$ by $\max\{n_k(h,s,a),1\}$. To clarify, when subscript "ty" is used, which stands for "type" as a placeholder for "Ho" or "Be", it means that there is no need to write two copies of expressions for Hoeffding-based and Bernstein-based noises separately.

Practical Considerations. Here, we explain why randomized exploration is widely used in practice and why our algorithmic formulation practically has advantage over UCB-type algorithms. In randomized exploration, there are usually two important components: (1) the algorithm (e.g., Algorithm 7) and (2) the noise magnitude (σ_{ty}). In practice, the main advantage of randomized exploration lies in the algorithm component. The generalization from the tabular setting to the function approximation setting is straightforward: one can just add a random regularization term in the value estimation step, whose details can be found in [Osband et al., 2018]. On the other hand, the generalization of optimistic algorithms from the tabular setting to the function approximation setting is more non-trivial because it often requires an explicit construction of the confidence set. For the second component, although generalizing our strategy of tuning noise magnitude to the real-world function approximation setting is indeed not straightforward, it is often set as a hyper-parameter in practice.

4.4.2 Regret Analysis

We analyze the regret, defined in (4.1), of our algorithm SSR using both types of perturbations. Our main theorems are presented in Theorem 4.4.1 and 4.4.2. In particular, Theorem 4.4.2 shows SSR with Bernstein-based perturbation can achieve the regret that matches the known lower bound $\Omega\left(H\sqrt{SAT}\right)$ up to logarithmic factors. We sketch the proof of Theorem 4.4.1 and Theorem 4.4.2 in Section 4.5.

Theorem 4.4.1. *If the Hoeffding-type noise (4.5) is used, then for any MDP $M = (H, \mathcal{S}, \mathcal{A}, P, R, s_1)$, with probability at least $1-\delta$, Algorithm 7 satisfies*

$$\mathrm{Reg}(M, K, \mathsf{SSR}_{\mathrm{Ho}}) \leq \widetilde{O}\left(H^{1.5}\sqrt{SAT} + H^4S^2A\right).$$

In particular, when $T \geq \widetilde{\Omega}\left(H^5S^3A\right)$, it holds that $\mathrm{Reg}(M, K, \mathsf{SSR}_{\mathrm{Ho}}) \leq \widetilde{O}\left(H^{1.5}\sqrt{SAT}\right)$.

Theorem 4.4.2. *If the Bernstein-type noise (4.6) is used, then when $T \geq \widetilde{\Omega}\left(H^5S^2A\right)$, for any MDP $M = (H, \mathcal{S}, \mathcal{A}, P, R, s_1)$, with probability at least $1-\delta$, Algorithm 7 satisfies*

$$\mathrm{Reg}(M, K, \mathsf{SSR}_{\mathrm{Be}}) \leq \widetilde{O}\left(H\sqrt{SAT} + H^4S^2A\right).$$

In particular, if we further have $T \geq \widetilde{\Omega}\left(H^6S^3A\right)$, it then holds that $\mathrm{Reg}(M, K, \mathsf{SSR}_{\mathrm{Be}}) \leq \widetilde{O}\left(H\sqrt{SAT}\right)$.

We give a brief comparison between SSR and other related works. Russo [2019] shows that RLSVI, an algorithm similar to SSR, can achieve $\tilde{O}\left(H^{2.5}S^{1.5}\sqrt{AT}\right)$ regret in expectation over the randomness of MDP and the algorithm. In [Agrawal et al., 2021], an improved high probability regret bound $\widetilde{O}\left(H^2S\sqrt{AT}\right)$ is proposed, which is the sharpest bound for randomized algorithms prior to this work. Our paper closes the gap between those previous bounds and the lower bound in tabular setting.

We also run numerical simulations to empirically compare SSR and RLSVI in the deep-sea environment, which is commonly used as a benchmark to test an algorithm's ability to explore. The results show that SSR significantly outperforms RLSVI as predicted by our regret analysis. More details about our experiment can be found in Appendix C.10.

4.5 Proof Outline

In this section, we present an proof outline of Theorem 4.4.1 and 4.4.2. Since their proofs follow the same framework, we will present an unified outline and explain the individual steps particularly for each case when necessary. The details of complete proof are deferred to the appendix.

Notation For the ease of exposure, we will use a simplified notations during this sketch. Specifically, let $x = (h, s, a)$ and $x_h^k = (h, s_h^k, a_h^k)$.

4.5.1 Concentration and Optimism/Pessimism

We start by introducing a set of MDPs $\mathcal{M}_{\text{ty}}^k$ as a confidence set such that the empirical MDP $\hat{M}^k$ belongs to it with high probability, meaning that we have a good estimation of the true MDP. Specifically, with $M' = (H, \mathcal{S}, \mathcal{A}, P', R', s_1)$, we define

$$\mathcal{M}_{\text{ty}}^k := \left\{ M' : \forall x = (h, s, a), \left| \left(R'_x - R_x\right) + \left\langle P'_x - P_x, V_{h+1}^* \right\rangle \right| \leq \sqrt{e_{\text{ty}}^k(x)} \right\},$$

where $\sqrt{e_{\text{Ho}}^k(x)} = \sigma_{\text{Ho}}^k(x)$ and $\sqrt{e_{\text{Be}}^k(x)} \approx \sigma_{\text{Be}}^k(x)$.

Define the event $\mathcal{C}_{\text{ty}}^k := \left\{ \hat{M}^k \in \mathcal{M}_{\text{ty}}^k \right\}$. Then, by applying Hoeffding's inequality or Bernstein's inequality, for both types of perturbation, it is possible to show that

$$\sum_{k=1}^{\infty} \mathbb{P}\left(\left(\mathcal{C}_{\text{ty}}^k \right)^c \right) = \sum_{k=1}^{\infty} \mathbb{P}\left(\hat{M}^k \notin \mathcal{M}_{\text{ty}}^k \right) \leq \frac{\pi^2}{3}.$$

Since the value function is bounded in $[0, H]$, this inequality tells us that the regret incurred by bad estimation is at most $\widetilde{O}(H)$. To be precise, it holds with high probability that

$$\sum_{k=1}^{K} \mathbb{1}\left\{ \left(\mathcal{C}_{\text{ty}}^k \right)^c \right\} \left(V_1^* - V_{1,k}^{\pi^k} \right)(s_1^k) \leq \widetilde{O}(H). \tag{4.7}$$

Then, to better control the estimated value function, we need it to be bounded, which requires us to clip it. Specifically, we will use two crucial properties of our clipping method. First, if $\overline{Q}_{h,k}(s,a) \geq Q_h^*(s,a)$, $\forall (s,a) \in \mathcal{S} \times \mathcal{A}$, then we have $\overline{V}_{h,k}(s) \geq V_h^*(s), \forall s \in \mathcal{S}$. Similarly if $\overline{Q}_{h,k}(s,a) \leq Q_h^*(s,a)$, $\forall (s,a) \in \mathcal{S} \times \mathcal{A}$, then we have $\overline{V}_{h,k}(s) \leq V_h^*(s), \forall s \in \mathcal{S}$.

In addition, we can prove that whenever a clip is triggered for s_h^k, we have $n_k(h, s_h^k, a_h^k) \leq \alpha_k$ with $\alpha_k = \widetilde{O}(H^2)$. As a result, it is possible to show that the total regret incurred by clipping is at most $\widetilde{O}\left(H^4 SA\right)$, which is a lower-order term when T is sufficiently large. That is, let $\mathcal{E}_{H,k}^{cum}$ denote the event that there is no clipping during episode k. Then, it holds with

high probability that[†]

$$\sum_{k=1}^{K} \mathbb{1}\left\{\mathcal{C}_{\text{ty}}^{k} \cap \left(\mathcal{E}_{H,k}^{cum}\right)^{c}\right\}\left(V_1^* - V_{1,k}^{\pi^k}\right)(s_1^k) \leq \widetilde{O}\left(H^4 S^2 A\right). \tag{4.8}$$

As claimed before, because of the randomness in Gaussian noise, our algorithm SSR will encourage exploration and it takes effect when there is no clipping and the estimation is not too bad. In other words, it can be optimistic. However, also because of this randomness, its optimism only holds in a probabilistic sense. In precise, it is possible to show that

$$\mathbb{P}\left(\overline{V}_{h,k}(s) \geq V_h^*(s), \forall h \in [H], s \in \mathcal{S} \mid \mathcal{C}_{\text{ty}}^{k}\right) \geq C_{\text{ty}}, \tag{4.9}$$

where the value of constant C_{ty} depends on the type of noise we choose. Meanwhile, we can also prove a very similar probabilistic pessimism, which means to have $\overline{V}_{h,k}(s) \leq V_h^*(s), \forall h \in [H], s \in \mathcal{S}$ with constant probability. The property of optimism and pessimism will help us upper bound the absolute value of $V_1^*(s_1^k) - \overline{V}_{1,k}(s_1^k)$, which will be discussed soon.

4.5.2 Regret Decomposition

Now, given equations (4.7) and (4.8), we can see that for each episode k, it only remains to bound $\mathbb{1}\left\{\mathcal{C}_{\text{ty}}^{k} \cap \mathcal{E}_{H,k}^{cum}\right\}\left(V_1^* - V_{1,k}^{\pi^k}\right)(s_1^k)$. Technically, the further defined the good event $\mathcal{G}_k$ will help make $\overline{V}_{h,k}$ better-behaved. Its precise definition will be given in the appendix. Therefore, it is sufficient to bound $\mathbb{1}\{\mathcal{G}_k\}\left(V_1^* - V_{1,k}^{\pi^k}\right)(s_1^k)$, which means to have

$$\text{Reg}\left(M, K, \text{SSR}_{\text{ty}}\right) \leq \sum_{k=1}^{K} \mathbb{1}\{\mathcal{G}_k\}\big(\underbrace{V_1^* - \overline{V}_{1,k}}_{\text{pessimism}} + \underbrace{\overline{V}_{1,k} - V_{1,k}^{\pi^k}}_{\text{estimation error}}\big)(s_1^k) + \widetilde{O}\left(H^4 S A\right). \tag{4.10}$$

To proceed, we need to define two auxiliary value functions $\underline{V}_{h,k}$ and $\overline{\overline{V}}_{h,k}$, which are obtained by virtually running policy π^k on some deliberately perturbed MDPs. In particular, they are designed such that $\underline{V}_{h,k} \leq \overline{V}_{h,k} \leq \overline{\overline{V}}_{h,k}$ holds under the good event $\mathcal{G}_k$.

[†]Technically, this is not precisely how we bound the regret incurred by clipping, but it aligns better with the intuition. Full technical details can be found in Appendix.

Pessimism Term Here, as a technical novelty, we bound the pessimism term's absolute value. Meanwhile, different from Zanette et al. [2020] and Agrawal et al. [2021], by applying both optimism and pessimism, we do not resort to an independent copy of the perturbed MDP to bound the pessimism term and give a conceptually simpler analysis. In particulary, by defining $C_1 = 1/\min\{C_{\text{Ho}}, C_{\text{Be}}\}$. it is possible to show that

$$\mathbb{1}\{\mathcal{G}_k\}\left|V^*_{h,k}(s^k_h) - \overline{V}_{h,k}(s^k_h)\right| \leq \mathbb{1}\{\mathcal{G}_k\} C_1 \left(\left|\overline{\overline{V}}^{\pi^k}_{h,k}(s^k_h) - V^{\pi^k}_{h,k}(s^k_h)\right| + \left|\underline{V}^{\pi^k}_{h,k}(s^k_h) - V^{\pi^k}_{h,k}(s^k_h)\right|\right). \tag{4.11}$$

The full proof is given in Appendix under Lemma C.5.7.

Estimation Error Term The sum of pessimism term and estimation error term can be further bounded via the techniques of recursion used in Azar et al. [2017]. However, we want to emphasize the difference that in their algorithm, the estimated value is optimistic with high probability, which makes $\overline{V}_{h,k}(s^k_h) - V^*_h(s^k_h)$ always positive. Instead, since our optimism only holds with constant probability, we use absolute value to keep the estimation error terms positive. As a result, we show that

$$\left|\overline{V}_{1,k} - V^{\pi^k}_{1,k}\right|(s^k_1) + \left|\overline{\overline{V}}^{\pi^k}_{1,k} - V^{\pi^k}_{1,k}\right|(s^k_1) + \left|\underline{V}^{\pi^k}_{1,k} - V^{\pi^k}_{1,k}\right|(s^k_1) \lesssim e^{3C} \sum_{h=1}^{H} \left(L\sigma^k_{\text{ty}}(x^k_h) + \mathcal{M}_{h,k}\right), \tag{4.12}$$

where L denotes some poly-logarithmic term and $\mathcal{M}_{h,k}$ denotes some martingale difference sequence term at period h, episode k. The full proof is given in Appendix under Lemma C.5.11,

4.5.3 Combining Different Terms

By combining equations (4.10), (4.11) and (4.12) and applying concentration inequalities to MDP $\mathcal{M}_{h,k}$, it is possible to show that

$$\text{Reg}(M, K, \mathsf{SSR}_{\text{ty}}) \leq e^{3C_1} \sum_{k=1}^{K}\sum_{h=1}^{H} \mathbb{1}\{\mathcal{G}_k\} L\sigma^k_{\text{ty}}(x^k_h) + \widetilde{O}\left(H\sqrt{T} + H^4S^2A\right). \tag{4.13}$$

Then, a final high-probability regret bound can be obtained by summing each individual

terms over k, h separately. It is well-known among literature that

$$\sum_{k=1}^{K}\sum_{h=1}^{H}\sqrt{\frac{1}{n_k(x_h^k)+1}} \leq \widetilde{O}\left(\sqrt{HSAT}\right), \qquad \sum_{k=1}^{K}\sum_{h=1}^{H}\frac{1}{n_k(x_h^k)+1} \leq \widetilde{O}\left(HSA\right). \tag{4.14}$$

Recall the definition of σ_{Ho}^k in equation (4.5). By using these two inequalities, the bound in equation (4.13) can be made explicit if we use Hoeffding-type noise. As a result, we have

$$\text{Reg}\left(M, K, \mathsf{SSR}_{\text{Ho}}\right) \leq \widetilde{O}\left(H^{1.5}\sqrt{SAT} + H^4S^2A\right).$$

Bound on Sum of Variance

Analyses become more involved when Bernstein-type noise is used. Specifically, notice that inequalities in (4.14) cannot directly be used to bound $\sum_{k,h}\mathbb{V}\left(\tilde{P}^k_{x_h^k}, \overline{V}_{h+1,k}\right)$. Here, we apply some techniques developed in Azar et al. [2017]. However, since the optimism only holds with constant probability, the details for specific terms are quite different.

For the ease of exposure, we will ignore all constants and define $\hat{\mathbb{V}}^*_{h,k} = \mathbb{V}\left(\tilde{P}^k_{x_h^k}, V^*_h\right)$, $\hat{\overline{\mathbb{V}}}_{h,k} = \mathbb{V}\left(\tilde{P}^k_{x_h^k}, \overline{V}_{h,k}\right)$. Then, by using Cauchy-Schwartz inequality and equation (4.14), we can get

$$U \overset{\text{def}}{=} \sum_{k,h}\mathbb{1}\left\{\mathcal{G}_k\right\}\sqrt{\frac{L}{n_k(x_h^k)+1}}\left(\sqrt{\hat{\mathbb{V}}^*_{h,k}} + \sqrt{\hat{\overline{\mathbb{V}}}_{h,k}}\right) \leq \sqrt{\widetilde{O}\left(HSA\right)\sum_{k,h}\mathbb{1}\left\{\mathcal{G}_k\right\}\left(\hat{\mathbb{V}}^*_{h,k} + \hat{\overline{\mathbb{V}}}_{h,k}\right)} \tag{4.15}$$

Here, note that $U \approx \sum_{k,h}\sigma^k_{\text{Be}}(x_h^k)$. Then, after some steps of algebra, it is possible to show that

$$\sum_{k=1}^{K}\sum_{h=1}^{H}\mathbb{1}\left\{\mathcal{G}_k\right\}\left(\hat{\mathbb{V}}^*_{h,k} + \hat{\overline{\mathbb{V}}}_{h,k}\right) \leq \widetilde{O}\left(HT + H^2U\right) \qquad (\text{When } T \geq \widetilde{\Omega}\left(H^5S^2A\right))$$

$$\implies U \leq \widetilde{O}\left(\sqrt{HSA\left(HT + H^2U\right)}\right) \leq \widetilde{O}\left(H\sqrt{SAT} + H^{1.5}\sqrt{U}\right).$$

(By using equation (4.15))

Now, we can see that $\sum_{k,h}\sigma^k_{\text{Be}}(x_h^k) \approx U \leq \widetilde{O}\left(H\sqrt{SAT}\right)$ satisfies this inequality. Finally,

by plugging this result back into equation (4.13), we can have

$$\mathrm{Reg}\left(M, K, \mathsf{SSR}_{\mathrm{Be}}\right) \leq \widetilde{O}\left(H\sqrt{SAT} + H^4S^2A\right),$$

which matches the known lower bound when $T \geq \widetilde{\Omega}\left(H^6S^3A\right)$.

4.6 Conclusion

We gave a new algorithm with randomized exploration, SSR, for tabular MDP, which enjoys a near-optimal $\widetilde{O}\left(H\sqrt{SAT}\right)$ regret bound in the time-homogeneous model. Previously, near-optimal regret bounds can only be achieved by optimistic algorithms. Our result also highlights the importance of using a single random seed for the entire episode and using the variance information in tuning the magnitude of noise (cf. Bernstein's inequality).

One important open problem is whether randomized exploration can a achieve a horizon-free regret bound in the time-homogeneous model where the transition is the same at different levels [Zanette and Brunskill, 2019, Wang et al., 2020, Zhang et al., 2020b]. Another possible future direction is to consider whether the sub-optimal lower order terms $\widetilde{O}\left(H^4S^2A\right)$ can be further improved to relax the current requirement $T \geq \widetilde{\Omega}\left(H^6S^3A\right)$ for being near-optimal.

Part II

PRIMAL-DUAL METHODS IN BANDITS AND REINFORCEMENT LEARNING

Chapter 5

PRIMAL-DUAL METHODS IN ONLINE SELECTIVE SAMPLING

This chapter is based on Camilleri et al. [2021b], with Romain Camilleri, Maryam Fazel, Lalit Jain and Kevin Jamieson.

5.1 Introduction

In this work we consider *selective sampling for online best-arm identification.* In this setting, at every time step $t = 1, 2, \ldots,$ Nature reveals a potential measurement $x_t \in \mathcal{X} \subset \mathbb{R}^d$ to the learner. The learner can choose to either *query* x_t ($\xi_t = 1$) or abstain ($\xi_t = 0$) and immediately move on to the next time. If the learner chooses to take a query ($\xi_t = 1$), then Nature reveals a noisy linear measurement of an unknown $\theta_* \in \mathbb{R}^d$, i.e. $y_t = \langle x_t, \theta_* \rangle + \epsilon_t$ where ϵ_t is mean zero sub-Gaussian noise. Before the start of the game, the learner has knowledge of a set $\mathcal{Z} \subset \mathbb{R}^d$. The objective of the learner is to identify $z_* := \arg\max_{z \in \mathcal{Z}} \langle z, \theta_* \rangle$ with probability at least $1 - \delta$ at a learner specified stopping time $\mathcal{U}$. It is desirable to minimize both the stopping time $\mathcal{U}$ which counts the total number of unlabeled or labeled queries and the number of labeled queries requested $\mathcal{L} := \sum_{t=1}^{\mathcal{U}} \mathbf{1}\{\xi_t = 1\}$. In this setting, at each time t the learner must make the decision of whether to accept the available measurement x_t, or abstain and wait for an even more informative measurement. While abstention may result in a smaller total labeled sample complexity $\mathcal{L}$, the stopping time $\mathcal{U}$ may be very large. This paper characterizes the set of feasible pairs $(\mathcal{U}, \mathcal{L})$ that are necessary and sufficient to identify z_* with probability at least $1 - \delta$ when x_t are drawn IID at each time t from a distribution ν. Moreover, we propose an algorithm that nearly obtains the minimal information theoretic label sample complexity $\mathcal{L}$ for any desired unlabeled sample complexity $\mathcal{U}$.

While characterizing the sample complexity of selective sampling for online best arm identification is the primary theoretical goal of this work, the study was initially motivated by fundamental questions about how to optimally trade-off the value of information versus

time. Even for this idealized linear setting, it is far from obvious a priori what an optimal decision rule ξ_t looks like and if it can even be succinctly described, or if it is simply the solution to an opaque optimization problem. Remarkably, we show that for every feasible, optimal operating pair $(\mathcal{U}, \mathcal{L})$ there exists a matrix $A \in \mathbb{R}^{d\times d}$ such that the optimal decision rule takes on the form $\xi_t = \mathbf{1}\{x^\top Ax \geq 1\}$ when $x_t \sim \nu$ iid. The fact that for any smooth distribution ν the decision rule is a hard decision equivalent to x_t falling outside a fixed ellipse or not, and not a stochastic rule that varies complementarily with the density of ν over space is perhaps unexpected.

To motivate the problem description, suppose on each day $t = 1, 2, \ldots$ a food blogger posts the *Cocktail of the Day* with a recipe described by a feature vector $x_t \in \mathbb{R}^d$. You have the ingredients (and skills) to make any possible cocktail in the space of all cocktails $\mathcal{Z}$, but you don't know which one you'd like the most, i.e., $z_* := \arg\max_{z\in\mathcal{Z}} \langle z, \theta_* \rangle$, where θ_* captures your preferences over cocktail recipes. You decide to use the *Cocktail of the Day* to inform your search. That is, each day you are presented with the cocktail recipe $x_t \in \mathbb{R}^d$, and if you choose to make it ($\xi_t = 1$) you observe your preference for the cocktail y_t with $\mathbb{E}[y_t] = \langle x_t, \theta_* \rangle$. Of course, making cocktails can get costly, so you don't want to make each day's cocktail, but rather you will only make the cocktail if x_t is informative about θ_* (e.g., uses a new combination of ingredients). At the same time, waiting too many days before making the next cocktail of the day may mean that you never get to learn (and hence drink) the cocktail z_* you like best. The setting above is not limited to cocktails, but rather naturally generalizes to discovering the efficacy of drugs and other therapeutics where blood and tissue samples come to the clinic in a stream and the researcher has to choose whether to take a potentially costly measurement.

Our results hold for arbitrary $\theta_* \in \mathbb{R}^d$, sets $\mathcal{X} \subset \mathbb{R}^d$ and $\mathcal{Z} \subset \mathbb{R}^d$, and measures $\nu \in \triangle_{\mathcal{X}}$* for which we assume $x_t \sim \nu$ is drawn IID. The assumption that each x_t is IID allows us to make very strong statements about optimality. To summarize, our contributions are as follows:

- We present fundamental limits on the trade-off between the amount of unlabelled data

*We denote the set of probability measures over $\mathcal{X}$ as $\triangle_{\mathcal{X}}$.

and labelled data in the form of (the first) information theoretic lower bounds for selective sampling problems that we are aware of. Naturally, they say that there is an absolute minimum amount of unlabelled data that is necessary to solve the problem, but then for any amount of unlabelled data beyond this critical value, the bounds say that the amount of labelled data must exceed some value as a function of the unlabelled data used.

- We propose an algorithm that nearly matches the lower bound at all feasible trade-off points in the sense that given any unlabelled data budget that exceeds the critical threshold, the algorithm takes no more labels than the lower bound suggests. Thus, the upper and lower bounds sketch out a curve of all possible operating points, and the algorithm achieves any point on this curve.

- We characterize the optimal decision rule of whether to take a sample or not, based on any critical point is a simple test: Accept $x_t \in \mathbb{R}^d$ if $x_t^\top A x_t \geq 1$ for some matrix A that depends on the desired operating point and geometry of the task. Geometrically, this is equivalent to x_t falling inside or outside an ellipsoid.

- Our framework is also general enough to capture binary classification, and consequently, we prove results there that improve upon state of the art.

5.1.1 Related Work

Selective Sampling in the Streaming Setting: Online prediction, the setting in which the selective sampling framework was introduced, is a closely related problem to the one studied in this paper and enjoys a much more developed literature Cesa-Bianchi et al. [2009], Dekel et al. [2012], Agarwal [2013], Chen et al. [2021]. In the linear online prediction setting, for $t = 1, 2, \ldots$ Nature reveals $x_t \in \mathbb{R}^d$, the learner predicts $\widehat{y}_t$ and incurs a loss $\ell(\widehat{y}_t, y_t)$, and then the learner decides whether to observe y_t (i.e., $\xi_t = 1$) or not ($\xi_t = 0$), where y_t is a label generated by a composition of a known link function with a linear function of x_t. For example, in the classification setting Agarwal [2013], Cesa-Bianchi et al. [2009], Dekel et al. [2012], one setting assumes $y_t \in \{-1, 1\}$ with $\mathbb{E}[y_t | x_t] = \langle x_t, \theta_* \rangle$ for some unknown

$\theta_* \in \mathbb{R}^d$, and $\ell(\widehat{y}_t, y_t) = \mathbf{1}\{\widehat{y}_t \neq y_t\}$. In the regression setting Chen et al. [2021], one observes $y_t \in [-1, 1]$ with $\mathbb{E}[y_t|x_t] = \langle x_t, \theta_* \rangle$ again, and $\ell(\widehat{y}_t, y_t) = (\widehat{y}_t - y_t)^2$. After any amount of time $\mathcal{U}$, the learner is incentivized to minimize both the amount of requested labels $\sum_{t=1}^{\mathcal{U}} \mathbf{1}\{\xi_t = 1\}$ and the cumulative loss $\sum_{t=1}^{\mathcal{U}} \ell(y_t, \widehat{y}_t)$ (or some measure of regret which compares to predictions using the unknown θ_*). If every label y_t is requested then $\mathcal{L} = \mathcal{U}$ and this is just the classical online learning setting.

These works give a guarantee on the regret and labeled points taken in terms of the hardness of the stream relative to a learner which would see the label at every time. Most do not give the learner the ability to select an operating point that provides a trade-off between the amount of unlabeled versus labeled data taken. Those few works that propose algorithms that do provide this functionality do not provide lower bounds that match their given upper bounds, leaving it unclear whether their algorithm optimally negotiates this trade-off. In contrast, our work fully characterizes the trade-off between the amount of unlabeled and labeled data through an information-theoretic lower bound and a matching upper bound. Specifically, our algorithm includes a tuning parameter, call it τ, that controls the trade-off between the evaluation metric of interest (for us, the quality of the recommended $z \in \mathcal{Z}$), the label complexity $\mathcal{L}$, and the amount of unlabelled data $\mathcal{U}$ that is necessary before the metric of interest can be non-trivial. We prove that each possible setting of τ parametrizes *all* possible trade-offs between unlabeled and labeled data.

Our work is perhaps closest to the streaming setting for agnostic active classification Dasgupta et al. [2008], Huang et al. [2015] where each x_s is drawn i.i.d. from an underlying distribution ν on $\mathcal{X}$, and indeed our results can be specialized to this setting as we discuss in Section 5.3. These papers also evaluate themselves at a single point on the tradeoff curve, namely the number of samples needed in passive supervised learning to obtain a learner with excess risk at most ϵ. They provide minimax guarantees on the amount of labeled data needed in terms of the disagreement coefficient Hanneke et al. [2014]. In contrast, again, our results characterize the full trade-off between the amount of unlabeled data seen, and the amount of labeled data needed to achieve the target excess risk ϵ. We note that using online-to-batch conversion methods, Dekel et al. [2012], Agarwal [2013], Cesa-Bianchi et al. [2009] also provide results on the amount of labeled data needed but they assume a

very specific parametric form to their label distribution unlike our setting which is agnostic. Other works have characterized selective sampling for classification in the realizable setting that assumes there exists a classifer among the set under consideration that perfectly labels every y_t Hanneke and Yang [2021]–our work addresses the agnostic setting where no such assumption is made. Finally, our results apply under the more general setting of *domain adaptation under covariate shift* where we are observing data drawn from the stream ν, but we will evaluate the excess risk of our resulting classifier on a different stream π Rai et al. [2010], Saha et al. [2011], Xiao and Guo [2013].

Best-Arm Identification and Online Experimental Design. Our techniques are based on experimental design methods for best-arm identification in linear bandits, see Soare et al. [2014], Fiez et al. [2019], Camilleri et al. [2021a]. In the setting of these works, there exists a pool of examples $\mathcal{X}$ and at each time any $x \in \mathcal{X}$ can be selected with replacement. The goal is to identify the best arm using as few total selections (labels) as possible. Their algorithms are based on arm-elimination. Specifically, they select examples with probability proportional to an approximate G-optimal design with respect to the current remaining arms. Then, during each round after taking measurements, those arms with high probability of being suboptimal will be eliminated. Remarkably, near-optimal sample complexity has been achieved under this setting. While we apply these techniques of arm-elimination and sampling through G-optimal design, the major difference is that we are facing a stream instead of a pool of examples. Finally, Eghbali et al. [2018] considers a different online experiment design setup where (adversarially chosen) experiments arrive sequentially and a primal-dual algorithm decides whether to choose each, subject to a total budget. Eghbali et al. [2018] studies the competitive ratio of such algorithms (in the manner of online packing algorithms) for problems such as D-optimal experiment design.

5.2 Selective Sampling for Best Arm Identification

Consider the following game: Given known $\mathcal{X}, \mathcal{Z} \subset \mathbb{R}^d$ and unknown $\theta_* \in \mathbb{R}^d$ at each time $t = 1, 2, \ldots$:

1. Nature reveals $x_t \overset{iid}{\sim} \nu$ with $\text{support}(\nu) = \mathcal{X}$

2. Player chooses $Q_t \in \{0,1\}$. If $Q_t = 1$ then nature reveals y_t with $\mathbb{E}[y_t] = \langle x_t, \theta_* \rangle$

3. Player optionally decides to stop at time t and output some $\widehat{z} \in \mathcal{Z}$

If the player stops at time $\mathcal{U}$ after observing $\mathcal{L} = \sum_{t=1}^{\mathcal{U}} Q_t$ labels, the objective is to identify $z_* = \arg\max_{z \in \mathcal{Z}} \langle z, \theta_* \rangle$ with probability at least $1 - \delta$ while minimizing a trade-off of $\mathcal{U}, \mathcal{L}$.

This paper studies the relationship between $\mathcal{U}$ and $\mathcal{L}$ in the context of necessary and sufficient conditions to identify z_* with probability at least $1 - \delta$. Clearly $\mathcal{U}$ must be "large enough" for z_* to be identifiable even if all labels are requested (i.e., $\mathcal{L} = \mathcal{U}$). But if $\mathcal{U}$ is very large, the player can start to become more picky with their decision to observe the label or not. Indeed, one can easily imagine scenarios in which it is advantageous for a player to forgo requesting the label of the current example in favor of waiting for a more informative example to arrive later if they wished to minimize $\mathcal{L}$ alone. Intuitively, $\mathcal{L}$ should decrease as $\mathcal{U}$ increases, but how?

Any selective sampling algorithm for the above protocol at time t is defined by 1) a selection rule $P_t : \mathcal{X} \to [0,1]$ where $Q_t \sim \text{Bernoulli}(P_t(x_t))$, 2) a stopping rule $\mathcal{U}$, and 3) a recommendation rule $\widehat{z} \in \mathcal{Z}$. The algorithm's behavior at time t can use all information collected up to time t

Definition 5.2.1. For any $\delta \in (0,1)$ we say a selective sampling algorithm is δ-PAC for $\nu \in \triangle_{\mathcal{X}}$ if for all $\theta \in \mathbb{R}^d$ the algorithm terminates at time $\mathcal{U}$ which is finite almost surely and outputs $\arg\max_{z \in \mathcal{Z}} \langle z, \theta \rangle$ with probability at least $1 - \delta$.

5.2.1 Optimal design

Before introducing our own algorithm, let us consider a seemingly optimal procedure. For any $\lambda \in \triangle_{\mathcal{X}} = \{p : \sum_{x \in \mathcal{X}} p_x = 1,\ p_x \geq 0\ \forall x \in \mathcal{X}\}$ define

$$\rho(\lambda) := \max_{z \in \mathcal{Z} \setminus \{z_*\}} \frac{\|z - z_*\|^2_{\mathbb{E}_{X \sim \lambda}[XX^\top]^{-1}}}{\langle \theta_*, z_* - z \rangle^2}. \tag{5.1}$$

Intuitively, $\rho(\lambda)$ captures the number of labeled examples drawn from distribution λ to identify z_*. Specifically, for any $\tau \geq \rho(\lambda) \log(|\mathcal{Z}|/\delta)$, if $x_1, \ldots, x_\tau \sim \lambda$ and $y_i = \langle x_i, \theta_* \rangle + \epsilon_i$

where ϵ_i is iid 1 sub-Gaussian noise, then there exists an estimator $\widehat{\theta} := \widehat{\theta}(\{(x_i, y_i)\}_{i=1}^{\tau})$ such that $\langle \widehat{\theta}, z_* \rangle > \max_{z \in \mathcal{Z} \setminus z_*} \langle \widehat{\theta}, z \rangle$ with probability at least $1 - \delta$ Fiez et al. [2019]. In particular, $\tau \geq \rho(\lambda) \log(|\mathcal{Z}|/\delta)$ samples suffice to guarantee that $\arg\max_{z \in \mathcal{Z}} \langle \widehat{\theta}, z \rangle = \arg\max_{z \in \mathcal{Z}} \langle \theta_*, z \rangle =: z_*$.

Thus, if our τ samples are coming from ν, we would expect any reasonable algorithm to require at least $\rho(\nu) \log(|\mathcal{Z}|/\delta)$ examples and labels. However, since we only want to take informative examples, we instead choose to select the tth example $x_t = x$ according to a probability $P(x)$ so that our final labeled samples are coming from the distribution λ where $\lambda(x) \propto P(x)\nu(x)$. In particular, $P(x)$ should be chosen according to the following optimization problem

$$P^* = \underset{P:\mathcal{X} \to [0,1]}{\arg\min}\ \tau \mathbb{E}_{X \sim \nu}[P(X)] \quad \text{subject to} \quad \max_{z \in \mathcal{Z} \setminus \{z_*\}} \frac{\|z_* - z\|^2_{\mathbb{E}_{X \sim \nu}[\tau P(X) X X^\top]^{-1}}}{\langle z_* - z, \theta_* \rangle^2} \beta_\delta \leq 1 \tag{5.2}$$

for $\beta_\delta = \log(|\mathcal{Z}|/\delta)$ where the objective captures the number of samples we select using P^*, and the constraint captures the fact that we have solved the problem. Remarkably, we can reparametrize this result in terms of an optimization problem over $\lambda \in \Delta_{\mathcal{X}}$ instead of $P^* : \mathcal{X} \to [0, 1]$ as

$$\tau \mathbb{E}_{X \sim \nu}[P^*(X)] = \min_{\lambda \in \triangle_{\mathcal{X}}} \rho(\lambda)\beta_\delta \quad \text{subject to} \quad \tau \geq \|\lambda/\nu\|_\infty \rho(\lambda)\beta_\delta$$

where $\|\lambda/\nu\|_\infty = \max_{x \in \mathcal{X}} \lambda(x)/\nu(x)$, as shown in Proposition D.2.6. Note that as $\tau \to \infty$ the constraint becomes inconsequential. Also notice that $\rho(\nu)\beta_\delta$ appears to be a necessary amount of labels to solve the problem even if $P(x) \equiv 1$ (albeit, by arguing about minimizing the upperbound of above).

5.2.2 Main results

In this section we formally justify the sketched argument of the previous section, showing nearly matching upper and lower bounds.

Theorem 5.2.2 (Lower bound)**.** *Fix any $\delta \in (0, 1)$, $\mathcal{X}, \mathcal{Z} \subset \mathbb{R}^d$, and $\theta_* \in \mathbb{R}^d$. Any selective*

sampling algorithm that is δ-PAC for $\nu \in \triangle_{\mathcal{X}}$ and terminates after drawing $\mathcal{U}$ unlabelled examples from ν and requests the labels of just $\mathcal{L}$ of them satisfies

- $\mathbb{E}[\mathcal{U}] \geq \rho(\nu)\log(1/\delta)$, *and*
- $\mathbb{E}[\mathcal{L}] \geq \min_{\lambda \in \triangle_{\mathcal{X}}} \rho(\lambda)\log(1/\delta)$ *subject to* $\mathbb{E}[\mathcal{U}] \geq \|\lambda/\nu\|_\infty \rho(\lambda)\log(1/\delta)$.

The first part of the theorem quantifies the number of rounds or unlabelled draws $\mathcal{U}$ that *any* algorithm must observe before it could hope to stop and output z_* correctly. The second part describes a trade-off between $\mathcal{U}$ and $\mathcal{L}$. One extreme is if $\mathbb{E}[\mathcal{U}] \to \infty$, which effectively removes the constraint so that the number of observed labels must scale like $\min_{\lambda \in \triangle_{\mathcal{X}}} \rho(\lambda)\log(1/\delta)$. Note that this is precisely the number of labels required in the pool-based setting where the agent can choose *any* $x \in \mathcal{X}$ that she desires at each time t (e.g. Fiez et al. [2019]). In the other extreme, $\mathbb{E}[\mathcal{U}] = \rho(\nu)\log(1/\delta)$ so that the constraint in the label complexity $\mathbb{E}[\mathcal{L}]$ is equivalent to $\rho(\nu) \geq \|\lambda/\nu\|_\infty \rho(\lambda)$. This implies that the minimizing λ must either stay very close to ν, or must obtain a substantially smaller value of $\rho(\lambda)$ relative to $\rho(\nu)$ to account for the inflation factor $\|\lambda/\nu\|_\infty$. In some sense, this latter extreme is the most interesting point on the trade-off curve because its asking the algorithm to stop as quickly as the algorithm that observes all labels, but after requesting a minimal number of labels. Note that this lower bound holds even for algorithms that known ν exactly. The proof of Theorem 5.2.2 relies on standard techniques from best arm identification lower bounds (see e.g. Kaufmann et al. [2016], Fiez et al. [2019]).

Remarkably, every point on the trade-off suggested by the lower bound is nearly achievable.

Theorem 5.2.3 (Upper bound)**.** *Fix any $\delta \in (0,1)$, $\mathcal{X}, \mathcal{Z} \subset \mathbb{R}^d$, and $\theta_* \in \mathbb{R}^d$. Let $\Delta = \min_{z \in \mathcal{Z}\setminus\{z_*\}} \langle z_* - z, \theta_* \rangle$ and $\beta_\delta \propto \log(\log(\frac{1}{\Delta})|\mathcal{Z}|/\delta)$ where the precise constant is given in the appendix. For any $\tau \geq \rho(\nu)\beta_\delta$ there exists a δ-PAC selective sampling algorithm that observes $\mathcal{U}$ unlabeled examples and requests just $\mathcal{L}$ labels that satisfies with probability at least $1-\delta$*

- $\mathcal{U} \leq \log_2(\frac{4}{\Delta})\,\tau$, *and*

- $\mathcal{L} \leq 3\log_2(\frac{4}{\Delta}) \min_{\lambda \in \triangle_{\mathcal{X}}} \rho(\lambda)\,\beta_\delta \quad$ *subject to* $\quad \tau \geq \|\lambda/\nu\|_\infty \rho(\lambda)\,\beta_\delta.$

Aside from the $\log(\frac{1}{\Delta})$ factor and the $\log(|\mathcal{Z}|)$ that appears in the β_δ term, this nearly matches the lower bound. Note that the parameter τ parameterizes the algorithm and makes the trade-off between $\mathcal{U}$ and $\mathcal{L}$ explicit. The next section describes the algorithm that achieves this theorem.

5.2.3 Selective Sampling Algorithm

Algorithm 8 contains the pseudo-code of our selective sampling algorithm for best-arm identification. Note that it takes a confidence level $\delta \in (0,1)$ and a parameter τ that controls the unlabeled-labeled budget trade-off as input. The algorithm is effectively an elimination style algorithm and closely mirrors the RAGE algorithm for the pool-based setting of best-arm identification problem Fiez et al. [2019]. The key difference, of course, is that instead of being able to plan over the pool of measurements, this algorithm must plan over the x's that the algorithm may *potentially* see and account for the case that it might not see the x's it wants.

Algorithm 8 Selective Sampling for Best-arm Identification

1: **Input** $\mathcal{Z} \subset \mathbb{R}^d$, $\delta \in (0,1), \tau$
2: **while** $|\mathcal{Z}_\ell| \geq 1$ **do**
3: Let $\widehat{P}_\ell, \widehat{\Sigma}_{\widehat{P}_\ell} \leftarrow$OptimizeDesign$(\mathcal{Z}_\ell, 2^{-\ell}, \tau)$ // $\widehat{\Sigma}_{\widehat{P}_\ell}$ approximates $\mathbb{E}_{X\sim\nu}[\widehat{P}_\ell(X)XX^\top]$
4: **for** $t = (\ell-1)\tau + 1, \ldots, \ell\tau$ **do**
5: Nature reveals x_t drawn iid from ν (with support $\mathbb{R}^d$)
6: Sample $Q_t(x_t) \sim$ Bernoulli$(\widehat{P}_\ell(x_t))$. If $Q_t = 1$ then observe y_t // $\mathbb{E}[y_t|x_t] = \langle \theta_*, x_t \rangle$
7: **end for**
8: Let $\widehat{\theta}_\ell \leftarrow$RIPS$(\{\widehat{\Sigma}_{\widehat{P}_\ell}^{-1} Q_s(x_s) x_s y_s\}_{s=(\ell-1)\tau+1}^{\ell\tau}, \mathcal{Z}\times\mathcal{Z})$ // $\widehat{\theta}_\ell$ approximates θ_*
9: $\mathcal{Z}_{\ell+1} = \mathcal{Z}_\ell \setminus \{z \in \mathcal{Z}_\ell : \max_{z' \in \mathcal{Z}_\ell} \langle z' - z, \widehat{\theta}_\ell \rangle \geq 2^{-\ell}\}$
10: **end while**

In round ℓ, the algorithm maintains an active set $\mathcal{Z}_\ell \subseteq \mathcal{Z}$ with the guarantee that each remaining $z \in \mathcal{Z}_\ell$ satisfies, $\langle z_* - z, \theta_* \rangle \leq 8 \cdot 2^{-\ell}$. In each round, on Line 3 of the algorithm, it calls out to a sub-routine OptimizeDesign$(\mathcal{Z}, \epsilon, \tau)$ that is trying to approximate the ideal

optimal design of (5.2). In particular, the ideal response to OptimizeDesign$(\mathcal{Z}, \epsilon, \tau)$ would return a P^*_ϵ and $\Sigma_{P^*_\epsilon} = \mathbb{E}_{X\sim\nu}[P^*_\epsilon(X)XX^\top]$ where P^*_ϵ is the solution to Equation 5.2 with the one exception that the denominator of the constraint is replaced with $\max\{\epsilon^2, \langle\theta_*, z_* - z\rangle^2\}$. Of course, θ_* is unknown so we cannot solve Equation 5.2 (as well as other outstanding issues that we will address shortly). Consequently, our implementation will aim to *approximate* the optimization problem of Equation 5.2. But assuming our sample complexity is not too far off from this ideal, each round should not request more labels than the number of labels requested by the ideal program with $\epsilon = 0$. Thus, the total number of samples should be bounded by the ideal sample complexity times the number of rounds, which is $O(\log(\Delta^{-1}))$. We will return to implementation issues in the next section.

Assuming we are returned $(\widehat{P}_\ell, \widehat{\Sigma}_{\widehat{P}_\ell})$ that approximate their ideals as just described, the algorithm then proceeds to process the incoming stream of $x_t \sim \nu$. As described above, the decision to request the label of x_t is determined by a coin flip coming up heads with probability $\widehat{P}_\ell(x_t)$–otherwise we do not request the label. Given the collected dataset $\{(x_t, y_t, Q_t, \widehat{P}_\ell(x_t))\}_t$, line 8 then computes an estimate $\widehat{\theta}_\ell$ of θ_* using the RIPS estimator of Camilleri et al. [2021a] which will satisfy

$$|\langle z_* - z, \widehat{\theta}_\ell - \theta_*\rangle| \le O\left(\|z_* - z\|_{\mathbb{E}_{X\sim\nu}[\tau\widehat{P}_\ell(X)XX^\top]^{-1}}\sqrt{\log(2\ell^2|\mathcal{Z}|^2/\delta)}\right) \le 2^{-\ell}$$

for all $z \in \mathcal{Z}_\ell$ simultaneously with probability at least $1 - \delta$. Thus, the final line of the algorithm eliminates any $z \in \mathcal{Z}_\ell$ such that there exists another $z' \in \mathcal{Z}_\ell$ (think z_*) that satisfies $\langle\widehat{\theta}_\ell, z' - z\rangle > 2^{-\ell}$. The process continues until $\mathcal{Z}_\ell = \{z_*\}$.

5.2.4 Implementation of OptimizeDesign

For the subroutine OptimizeDesign passed $(\mathcal{Z}_\ell, \epsilon, \tau)$ the next best thing to computing Equation 5.2 with the denominator of the constraint replaced with $\max\{\epsilon^2, \langle\theta_*, z_* - z\rangle^2\}$, is to compute

$$P_\epsilon = \operatorname*{arg\,min}_{P:\mathcal{X}\to[0,1]} \mathbb{E}_{X\sim\nu}[P(X)] \text{ subject to } \max_{z,z'\in\mathcal{Z}_\ell} \frac{\|z - z'\|^2_{\mathbb{E}_{X\sim\nu}[\tau P(X)XX^\top]^{-1}}}{\epsilon^2}\beta_\delta \le 1 \qquad (5.3)$$

and $\Sigma_{P_\epsilon} = \mathbb{E}_{X\sim\nu}[P_\epsilon(X)XX^\top]$ for an appropriate choice of $\beta_\delta = \Theta(\log(|\mathcal{Z}|/\delta))$. To see this, firstly, any $z \in \mathcal{Z}$ with gap $\langle\theta_*, z_* - z\rangle$ that we could accurately estimate would not be included in $\mathcal{Z}_\ell$, thus we don't need it in the max of the denominator. Secondly, to get rid of z_* in the numerator (which is unknown, of course), we note that for any norm $\max_{z,z'} \|z-z'\| \le \max_z 2\|z-z_*\| \le \max_{z,z'} 2\|z-z'\|$. Assuming we could solve this directly and compute $\Sigma_{P_\epsilon} = \mathbb{E}_{X\sim\nu}[P_\epsilon(X)XX^\top]$, we can obtain the result of Theorem 2 (proven in the Appendix).

However, even if we knew ν exactly, the optimization problem of Equation 5.3 is quite daunting as it is a potentially infinite dimensional optimization problem over $\mathcal{X}$. Fortunately, after forming the Lagrangian with dual variables for each $z - z' \in \mathcal{Z} \times \mathcal{Z}$, optimizing the dual amounts to a finite dimensional optimization problem over the finite number of dual variables. Moreover, this optimization problem is maximizing a simple expectation with respect to ν and thus we can apply standard stochastic gradient ascent and results from stochastic approximation Nemirovski et al.. Given the connection to stochastic approximation, instead of sampling a fresh $\widetilde{x} \sim \nu$ each iteration, it suffices to "replay" a sequence of $\widetilde{x}$'s from historical data. Summing up, this construction allows us to compute a satisfactory P_ϵ and avoid both an infinite-dimensional optimization problem and requiring knowledge of ν (as long as historical data is available).

Meanwhile, with historical data, we can also empirically compute $\mathbb{E}_{X\sim\nu}[P_\epsilon(X)XX^\top]$. Historical data could mean offline samples from ν or just samples from previous rounds. In this setting, Theorem 2 still holds albeit with larger constants. Theorem D.4.1 in the appendix characterizes the necessary amount of historical data needed. Unfortunately (in full disclosure) the theoretical guarantees on the amount of historical data needed is absurdly large, though we suspect this arises from a looseness in our analysis. Similar assumptions and approaches to historical or offline data have been used in other works in the streaming setting e.g. Huang et al. [2015].

5.3 Selective Sampling for Binary Classification

We now review streaming Binary Classification in the agnostic setting Dasgupta et al. [2008], Hanneke et al. [2014], Huang et al. [2015] and show that our approach can be adapted to

this setting. Consider a binary classification problem where $\mathcal{X}$ is the example space and $\mathcal{Y} = \{-1, 1\}$ is the label space. Fix a hypothesis class $\mathcal{H}$ such that each $h \in \mathcal{H}$ is a classifier $h : \mathcal{X} \to \mathcal{Y}$. Assume there exists a fixed regression function $\eta : \mathcal{X} \to [0, 1]$ such that the label of x is Bernoulli with probability $\eta(x) = \mathbb{P}(Y = 1|X = x)$. Being in the agnostic setting, we make no assumption on the relationship between $\mathcal{H}$ and η. Finally, fix any $\nu \in \triangle_{\mathcal{X}}$ and $\pi \in \triangle_{\mathcal{X}}$. Given known $\mathcal{X}, \mathcal{H}$ and unknown regression function η, at each time $t = 1, 2, \ldots$:

1. Nature reveals $x_t \sim \nu$

2. Player chooses $Q_t \in \{0, 1\}$. If $Q_t = 1$ then nature reveals $y_t \sim \text{Bernoulli}(\eta(x_t)) \in \{-1, 1\}$

3. Player optionally decides to stop at time t and output some $\widehat{h} \in \mathcal{H}$.

Define the *risk* of any $h \in \mathcal{H}$ as $R_\pi(h) := \mathbb{P}_{X\sim\pi, Y\sim\eta(X)}(Y \neq h(X))$. If the player stops at time $\mathcal{U}$ after observing $\mathcal{L} = \sum_{t=1}^{\mathcal{U}} Q_t$ labels, the objective is to identify $h_* = \arg\min_{h\in\mathcal{H}} R_\pi(h)$ with probability at least $1 - \delta$ while minimizing a trade-off of $\mathcal{U}, \mathcal{L}$. Note that h_* is the true risk minimizer with respect to distribution π but we observe samples $x_t \sim \nu$; π is not necessarily equal to ν. While we have posed the problem as identifying the potentially unique h^*, our setting naturally generalizes to identifying an ϵ-good h such that $R_\pi(h) - R_\pi(h_*) \leq \epsilon$.

We will now reduce selective sampling for binary classification problem to selective sampling for best arm identification, and thus immediately obtain a result on the sample complexity. For simplicity, assume that $\mathcal{X}$ and $\mathcal{H}$ are finite. Enumerate $\mathcal{X}$ and for each $h \in \mathcal{H}$ define a vector $z^{(h)} \in [0, 1]^{|\mathcal{X}|}$ such that $z_x^{(h)} := \pi(x)\mathbf{1}\{h(x) = 1\}$ for $z^{(h)} = [z_x^{(h)}]_{x\in\mathcal{X}}$. Moreover, define $\theta^* := [\theta_x^*]_{x\in\mathcal{X}}$ where $\theta_x^* := 2\eta(x) - 1$. Then

$$R_\pi(h) = \mathbb{E}_{X\sim\pi, Y\sim\eta(X)}[\mathbf{1}\{Y \neq h(X)\}] = \sum_{x\in\mathcal{X}} \pi(x)(\eta(x)\mathbf{1}\{h(x) \neq 1\} + (1 - \eta(x))\mathbf{1}\{h(x) \neq 0\})$$
$$= \sum_{x\in\mathcal{X}} \pi(x)\eta(x) + \sum_{x\in\mathcal{X}} \pi(x)(1 - 2\eta(x))\mathbf{1}\{h(x) = 1\} = c - \langle z^{(h)}, \theta^*\rangle$$

where $c = \sum_{x\in\mathcal{X}} \pi(x)\eta(x)$ does not depend on h. Thus, if $\mathcal{Z} := \{z^{(h)}\}_{h\in\mathcal{H}}$ then identifying

$h_* = \arg\min_{h\in\mathcal{H}} R_\pi(h)$ is equivalent to identifying $z_* = \arg\max_{z\in\mathcal{Z}}\langle z, \theta^*\rangle$. We can now apply Theorem 5.2.3 to obtain a result describing the sample complexity trade-off. First define,

$$\rho_\pi(\lambda,\varepsilon) := \max_{z\in\mathcal{Z}\setminus\{z_*\}} \frac{\|z-z_*\|^2_{\mathbb{E}_{X\sim\lambda}[XX^\top]^{-1}}}{\max\{\langle\theta_*, z_*-z\rangle^2,\varepsilon^2\}} = \max_{h\in\mathcal{H}\setminus\{h_*\}} \frac{\mathbb{E}_{X\sim\pi}\left[\mathbf{1}\{h(X)\neq h'(X)\}\frac{\pi(X)}{\lambda(X)}\right]}{\max\{(R_\pi(h)-R_\pi(h^*))^2,\varepsilon^2\}}$$

An important case of the above setting is when $X\sim\nu$ and $\pi=\nu$, i.e. we are evaluating the performance of a classifier relative to the same distribution our samples are drawn from. This is the setting of Dasgupta et al. [2008], Huang et al. [2015], Hanneke et al. [2014]. The following theorem shows that the sample complexity obtained by our algorithm is at least as good as the results they present.

Theorem 5.3.1. *Fix any $\delta\in(0,1)$, domain $\mathcal{X}$ with distribution ν, finite hypothesis class $\mathcal{H}$, regression function $\eta:\mathcal{X}\to[0,1]$. Set $\epsilon\geq 0$ and $\beta_\delta = 2048\log(4\log_2^2(4/\epsilon)|\mathcal{H}|/\delta)$. Then for $\tau\geq\rho_\pi(\nu,\epsilon)\beta_\delta$ there exists a selective sampling algorithm that returns $h\in\mathcal{H}$ satisfying $R_\pi(h)-R_\pi(h^*)\leq\epsilon$ by observing $\mathcal{U}$ unlabeled examples and requesting just $\mathcal{L}$ labels such that*

- $\mathcal{U}\leq\log_2(4/\epsilon)\tau$
- $\mathcal{L}\leq 3\log_2(\frac{4}{\varepsilon})\min_{\lambda\in\triangle_\mathcal{X}}\rho_\pi(\lambda,\varepsilon)\beta_\delta \quad s.t. \quad \tau\geq\|\lambda/\nu\|_\infty\rho_\pi(\lambda,\varepsilon)\beta_\delta$

with probability at least $1-\delta$. Furthermore when $\nu=\pi$ and if $\tau\geq 16\rho(\nu,\epsilon)\beta_\delta$ we have that

$$\mathcal{L}\leq 36\log_2(4/\epsilon)\left(\tfrac{R_\nu(h^*)^2}{\epsilon^2}+4\right)\sup_{\xi\geq\epsilon}\theta^*(2R_\nu(h^*)+\xi,\nu)\beta_\delta$$

where $\theta^*(u,\nu)$ is the disagreement coefficient, defined in Appendix D.5.

Note that if τ is sufficiently large then the labeled sample complexity we obtain $\min_{\lambda\in\Delta_\mathcal{X}}\rho(\lambda,\epsilon)$ could be significantly smaller than previous results in the streaming setting, e.g. see Katz-Samuels et al. [2021]. The proof of Theorem 5.3.1 can be found in Appendix D.5.

5.4 Solving the Optimization Problem

Recall that in Algorithm 8, during round ℓ, we need to solve optimization problem (5.3). Solving this optimization problem is not trivial because the number of variables can potentially be infinite if $\mathcal{X}$ is an infinite set. In this section, we will demonstrate how to reduce it to a finite-dimensional problem by considering its dual problem. To simplify the notation, let $\mathcal{Y}_\ell = \{z - z' : z, z' \in \mathcal{Z}_\ell, z \neq z'\}$, and rewrite the problem as follows, where $c_\ell > 0$ is a constant that may depend on round ℓ.

$$\begin{aligned} \min_P \quad & \mathbb{E}_{X\sim\nu}\left[P(X)\right] \\ \text{subject to} \quad & y^\top \mathbb{E}_{X\sim\nu}\left[P(X)XX^\top\right]^{-1} y \leq c_\ell^2, \quad \forall y \in \mathcal{Y}_\ell, \\ & 0 \leq P(x) \leq 1, \quad \forall x \in \mathcal{X}. \end{aligned} \tag{5.4}$$

Using the Schur complement technique, we show in Lemma D.3.10 (Appendix D.3) the following equivalence: $y^\top \mathbb{E}_{X\sim\nu}\left[P(X)XX^\top\right]^{-1} y \leq c_\ell^2 \iff \mathbb{E}_{X\sim\nu}\left[P(X)XX^\top\right] \succeq \frac{1}{c_\ell^2} yy^\top$. This transforms a constraint involving matrix inversion into one with ordering between PSD matrices. Then, we remove the bound constraints $0 \leq P(x) \leq 1$, $\forall x \in \mathcal{X}$ by introducing the barrier function $-\log(1-x) - \log(x)$. That is, instead of working with the objective $\mathbb{E}_{X\sim\nu}\left[P(X)\right]$ directly, we consider the following problem.

$$\begin{aligned} \min_P \quad & \mathbb{E}_{X\sim\nu}[P(X) - \mu_b(\log(1-P(X)) + \log(P(X)))] \\ \text{subject to} \quad & \mathbb{E}_{X\sim\nu}\left[P(X)XX^\top\right] \succeq \tfrac{1}{c_\ell^2} yy^\top, \quad \forall y \in \mathcal{Y}_\ell. \end{aligned} \tag{5.5}$$

Here, $\mu_b \in (0,1)$ is some small constant that controls how strong the barrier is. Intuitively, a smaller μ_b will make problem (5.5) closer to the original problem. We now show that unlike the primal, the dual problem is indeed finite-dimensional. For each constraint of $y \in \mathcal{Y}_\ell$, let the matrix $\Lambda_y \succeq \mathbf{0}$ be its dual variable. Further, let $\Lambda = \sum_{y\in\mathcal{Y}_\ell} \Lambda_y$ and $\mathbf{\Lambda} = (\Lambda_y)_{y\in\mathcal{Y}_\ell}$. The corresponding Lagrangian is

$$\mathcal{L}\left(\mathbf{\Lambda}, P\right) = \mathbb{E}_{X\sim\nu}\left[P(X) - \mu_b\left(\log(1-P(X)) + \log(P(X))\right) - P(X)X^\top \Lambda X\right] + \frac{1}{c_\ell^2}\sum_{y\in\mathcal{Y}_\ell} y^\top \Lambda_y y.$$

The dual problem is $\max_{\Lambda_y \succeq \mathbf{0}, \forall y \in \mathcal{Y}_\ell} \min_P \mathcal{L}(\mathbf{\Lambda}, P)$. Notice that minimization over $P : \mathcal{X} \mapsto [0,1]$ can be done via minimizing $P(x)$ point-wise for each $x \in \mathcal{X}$. To do this, we take the gradient with respect to each $P(x)$ and set it to zero to get

$$1 + \frac{\mu_b}{1 - P(x)} - \frac{\mu_b}{P(x)} - x^\top \Lambda x = 0. \tag{5.6}$$

Solving this equation and defining $q_\Lambda(x) = x^\top \Lambda x - 1$, we get

$$P_\Lambda(x) = \frac{1}{2} - \frac{\mu_b}{q_\Lambda(x)} + \frac{\sqrt{(2\mu_b - q_\Lambda(x))^2 + 4\mu_b q_\Lambda(x)}}{2q_\Lambda(x)}. \tag{5.7}$$

Note that if $\mu_b = 0$ (no barrier), the above reduces to the "threshold" decision rule $P_\Lambda(x) = \frac{1}{2} + \frac{|q_\Lambda(x)|}{2q_\Lambda(x)}$, which gives 0 when $q_\Lambda(x) < 0$ and 1 when $q_\Lambda(x) > 0$.[†] This is exactly the hard elliptical threshold rule mentioned before, in which whether to query the label for x depends on whether it falls inside ($x^\top \Lambda x < 1$) or outside ($x^\top \Lambda x > 1$) of the ellipsoid defined by the positive semidefinite matrix Λ. A visualization of the decision rule P_Λ is given in Figure D.1 in the Appendix.

Now, by plugging in $P_\Lambda(x)$, our dual problem becomes $\max_{\Lambda_y \succeq \mathbf{0}, \forall y} D(\mathbf{\Lambda}) := \mathcal{L}(\mathbf{\Lambda}, P_\Lambda)$. This is a finite-dimensional optimization problem, and can be solved by projected gradient ascent (or projected stochastic gradient ascent when we have only samples from ν). The gradient of $D(\mathbf{\Lambda})$ is

$$\begin{aligned}
\nabla_{\Lambda_y} D(\mathbf{\Lambda}) &= \mathbb{E}_{X \sim \nu}\left[\left(1 + \frac{\mu_b}{1 - P_\Lambda(x)} - \frac{\mu_b}{P_\Lambda(X)} - X^\top \Lambda X\right) \nabla_{\Lambda_y} P_\Lambda(X) - P_\Lambda(X) X X^\top\right] + \frac{yy^\top}{c_\ell^2} \\
&= \frac{yy^\top}{c_\ell^2} - \mathbb{E}_{X \sim \nu}\left[P_\Lambda(X) X X^\top\right]. \qquad \text{(Since } P_\Lambda(X) \text{ solves Eq. (5.6))}
\end{aligned}$$

The algorithm to solve the problem has been summarized in Algorithm 9, in which the gradient during kth iteration is replaced by its unbiased estimator $\frac{yy^\top}{c_\ell^2} - P_{\hat{\Lambda}^{(k)}}(x_k) x_k x_k^\top$. The adaptive learning rate is chosen by following the discussion in chapter 4 of Orabona [2019]. Optimizing the assignment of $\hat{\Lambda}_y$ to each y in line 10 ensures that the re-scaling step

[†]When $q_\Lambda(x) = 0$, $P_\Lambda(x)$ is undetermined from the dual.

in line 11 increases the function value in an optimized way. Finally, the re-scaling step is used to ensure that the output primal objective value $\mathbb{E}_{X\sim\nu}[P(X)]$ is bounded well, which will be explained in more details in Appendix D.3.

Algorithm 9 Projected Stochastic Gradient Ascent to Solve OptimizeDesign

1: **Input:** Number of iterations K; number of samples u; barrier weight $\mu_b \in (0,1)$
2: Initialize $\hat{\Lambda}_y^{(0)} = \mathbf{0}$ for each $y \in \mathcal{Y}_\ell$
3: **for** $k = 0, 1, 2, \ldots, K-1$ **do**
4: Sample $x_k \sim \nu$
5: Set $g_{k,y} = \frac{yy^\top}{c_\ell^2} - P_{\hat{\Lambda}^{(k)}}(x_k)x_k x_k^\top$, where P_Λ is defined in Eq. (5.7)
6: Set $\hat{\Lambda}_y^{(k+1)} \leftarrow \hat{\Lambda}_y^{(k)} + \eta_k g_{k,y}$ for each $y \in \mathcal{Y}_\ell$, where $\eta_k = \frac{1}{\sqrt{2\sum_{s=1}^k \sum_{y\in\mathcal{Y}_\ell}\|g_{s,y}\|_2^2}}$
7: Update $\hat{\Lambda}_y^{(k+1)} \leftarrow \Pi_{\mathbb{S}_+^d}(\hat{\Lambda}_y^{(k+1)})$ for each $y \in \mathcal{Y}_\ell$, a projection to the set of $d \times d$ PSD matrices
8: **end for**
9: Let $\hat{\Lambda}_y = \frac{1}{K}\sum_{k=1}^K \hat{\Lambda}_y^{(k)}$ for each $y \in \mathcal{Y}_\ell$ and $\hat{\Lambda} = \sum_{y\in\mathcal{Y}_\ell}\hat{\Lambda}_y$
10: Update $(\hat{\Lambda}_y)_{y\in\mathcal{Y}_\ell} \leftarrow \arg\max_{\mathbf{\Lambda}} \sum_{y\in\mathcal{Y}_\ell} y^\top \Lambda_y y$, subject to $\sum_{y\in\mathcal{Y}_\ell}\Lambda_y = \hat{\Lambda}, \Lambda_y \succeq \mathbf{0}, \forall y \in \mathcal{Y}_\ell$.
11: Find $s^* \leftarrow \arg\max_{s\in[0,1]} D_E(s\cdot\hat{\mathbf{\Lambda}})$, where D_E empirically evaluates D using u i.i.d. samples
12: **return** $\widetilde{\Lambda} = s^* \cdot \sum_{y\in\mathcal{Y}_\ell}\hat{\Lambda}_y$

Let $\mathbf{\Lambda}^*$ be an optimal solution for $D(\mathbf{\Lambda})$. Intuitively, as long as we run this algorithm with sufficiently large number of iterations K and number of samples u, we can guarantee that $D(\widetilde{\mathbf{\Lambda}})$ and $D(\mathbf{\Lambda}^*)$ are close enough with high probability, which in turn guarantees that the primal constraints are violated by only a tiny amount and $\mathbb{E}_{X\sim\nu}\left[P_{\widetilde{\Lambda}}(X)\right]$ is close enough to the optimal value. Specifically, we can prove the following theorem.

Theorem 5.4.1. *Suppose* $\|x\|_2 \le M$ *for any* $x \in \mathrm{supp}(\nu)$ *and* $\Sigma = \mathbb{E}_{X\sim\nu}\left[XX^\top\right]$ *is invertible. Let* $\mathbf{\Lambda}^* \in \arg\max_{\Lambda_y\succeq\mathbf{0},\forall y\in\mathcal{Y}_\ell} D(\mathbf{\Lambda})$ *and* $\kappa(\Sigma) = \frac{\lambda_{\max}(\Sigma)}{\lambda_{\min}(\Sigma)}$ *be its condition number. Assume* $\|\Lambda^*\|_F > 0$ *and define* $\omega = \min_{\Gamma\in\mathbb{S}^d:\|\Gamma\|_F=1} \mathbb{E}_{X\sim\nu}\left[\left(X^\top\Gamma X\right)^2\right]$, *where* $\mathbb{S}^d$ *is the set of* $d\times d$ *symmetric matrices.*

Then, $\Lambda^* = \sum_{y\in\mathcal{Y}_\ell}\Lambda_y^*$ *is unique. Further, for any* $\epsilon > 0$ *and* $\delta > 0$, *if it holds that*

$\mu_b \leq O\left(\sqrt{\|\Lambda^*\|_F \kappa(\Sigma) M}\right) \cdot \sqrt{(1+\epsilon)/\epsilon}$ *and*

$$K \geq O\left(\frac{|\mathcal{Y}_\ell|^3 \kappa(\Sigma)^2 \|\Lambda^*\|_F^8 M^{16} \log(1/\delta)}{\omega^2 \mu_b^6}\right) \cdot \left(\frac{1+\epsilon}{\epsilon}\right)^2, u \geq O\left(\frac{\kappa(\Sigma)^2 \|\Lambda^*\|_F^6 M^{16} \log(1/\delta)}{\omega^2 \mu_b^6}\right) \cdot \left(\frac{1+\epsilon}{\epsilon}\right)^2,$$

then, with probability at least $1-\delta$*, Algorithm 9 will output* $\widetilde{\Lambda}$ *that satisfies*

- $y^\top \mathbb{E}_{X\sim\nu}\left[P_{\widetilde{\Lambda}}(X) X X^\top\right]^{-1} y \leq (1+\epsilon) c_\ell^2, \quad \forall y \in \mathcal{Y}_\ell.$
- $\mathbb{E}_{X\sim\nu}\left[P_{\widetilde{\Lambda}}(X)\right] \leq \mathbb{E}_{X\sim\nu}\left[\widetilde{P}(X)\right] + 4\sqrt{\mu_b}$*, where* $\widetilde{P}$ *is the optimal solution to problem (5.4) with barrier constraint repaced by* $0 \leq P(x) \leq 1-\mu_b, \forall x \in \mathcal{X}$*.*

The proof is in Appendix D.3. Although $\widetilde{P}$ is not exactly the same as the optimal solution of the original problem (5.4), when μ_b is sufficiently small, they will be very close. Meanwhile, it should be noted that Theorem 5.4.1 mainly reveals that with sufficiently large number of iterations and number of samples, Algorithm 9 can output sufficiently good solution. In future work, we plan to examine how much this bound can be improved via a tighter analysis.

Finally, notice that Algorithm 9 needs to maintain $|\mathcal{Y}_\ell|\, d^2 = O(|\mathcal{Z}_\ell|^2\, d^2)$ variables, which can be large when we have a large set $\mathcal{Z}_\ell$. Therefore, as an alternative, we also propose Algorithm 16 that only needs to maintain d^2 variables but requires more computational power in each iteration. The details are given in Appendix D.3.

5.5 Empirical results

In this section we present a benchmark experiment validating the fundamental trade-offs that are theoretically characterized in Theorem 5.2.2 and Theorem 5.2.3. We take inspiration from Soare et al. [2014] to define our experimental protocol:

- $d = 2$, a two-dimensional problem.
- $\mathcal{Z} = [\mathbf{e}_1, \mathbf{e}_2, (\cos(\omega), \sin(\omega))]$ for $\omega = 0.3$, where $\mathbf{e}_1, \mathbf{e}_2$ are canonical vectors.
- $\theta_* = 2\mathbf{e}_1$ and $y = x^\top \theta_* + \eta$, where $\eta \sim \mathcal{N}(0,1)$.

- The distribution ν for streaming measurements $x_t \overset{i.i.d.}{\sim} \nu$ is such that

$$x_t = (\cos(2I_t\pi/N), \sin(2I_t\pi/N)),$$

where $I_t \in \{0, \ldots, N-1\}$, $\mathbb{P}(I_t = i) \propto \cos(2i\pi/N)^2$, and $N = 30$.

In this problem, the angle ω is small enough that the item $(\cos(\omega), \sin(\omega))$ is hard to discriminate from the best item $\mathbf{e}_1$. As argued in Soare et al. [2014], an efficient sampling strategy for this problem instance would be to pull arms in the direction of $\pm\mathbf{e}_2$ in order to reduce the uncertainty in the direction of interest, $\mathbf{e}_1 - (\cos(\omega), \sin(\omega))$. However, the distribution ν is defined such that it is more likely to receive a vector x_t in the direction of $\pm\mathbf{e}_1$ rather than $\pm\mathbf{e}_2$. Thus, if one seeks a small label complexity, then P should be taken to reject measurements in the direction of $\pm e_1$.

In the benchmark experiment, we compare the following three algorithms which all use Algorithm 8 as a meta-algorithm and just swap out the definition of $\widehat{P}_\ell$. `Naive Algorithm` uses no selective sampling so that $\widehat{P}_\ell(x) = 1$ for all x; the `Oracle Algorithm` uses $\widehat{P}_\ell = P_*$ where P_* is the ideal solution to (5.2), and `Our Algorithm` uses the solution to (5.5) for $\widehat{P}_\ell$, where we take $\mu_b = 2 \times 10^{-5}$. We swept over the values of τ and plotted on the y-axis the amount of labeled data needed before termination, as shown in Figure 5.1.

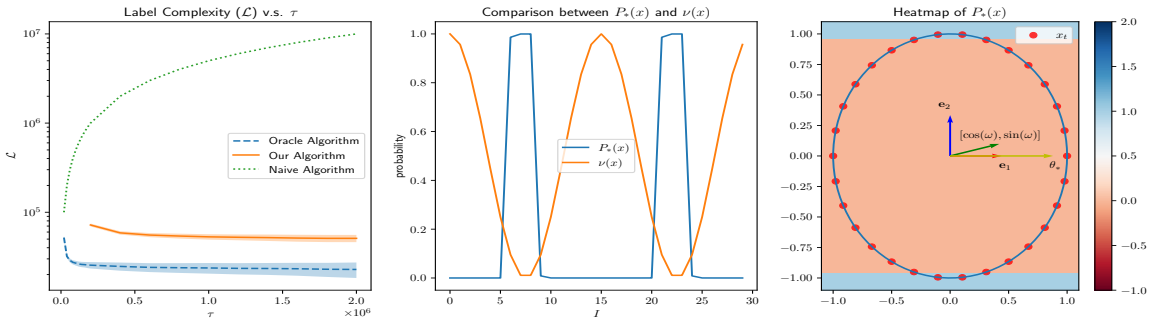


Figure 5.1: (left) For each value of τ, we plot the average label complexity over 50 repeated trials. (middle) Visualization of $P_*(x)$ and $\nu(x)$ v.s. x, where x is indexed by I such that $x_I = (\cos(2I\pi/N), \sin(2I\pi/N))$. Here, P_* is solved with $\tau = 4 \times 10^5$ and distribution ν is not normalized. (right) A heat map of $P_*(x)$ along with the setting of experimental protocol.

We observe in Figure 5.1 that the algorithms using non-naive selection rules require far less label complexity than the naive algorithm for all τ. This reflects the intuition that selection strategies that focus on requesting the more informative streaming measurements are much more efficient than naively observing every streaming measurement. Meanwhile, the trade-off between label complexity $\mathcal{L}$ and sample complexity $\mathcal{U}$ characterized in Theorem 5.2.2 and Theorem 5.2.3 is precisely illustrated in Figure 5.1. Indeed, we see the number of labels queried by the two selective sampling algorithms decrease as the number of unlabeled data seen in each round increases.

5.6 Conclusion

In this paper, we proposed a new approach for the important problem of *selective sampling for best arm identification*. We provide a lower bound that quantifies the trade-off between labeled samples and stopping time and also presented an algorithm that nearly achieves the minimal label complexity given a desired stopping time.

One of the main limitations of this work is that our approach depends on a well-specified model following stationary stochastic assumptions. In practice, dependencies over time and model mismatch are common. Utilizing the proposed algorithm outside of our assumptions may lead to poor performance and unexpected behavior with adverse consequences. While negative results justify some of the most critical assumptions we make (e.g., allowing the stream x_t to be arbitrary, rather than iid, can lead to trivial algorithms, see Theorem 7 of Chen et al. [2021]), exploring what theoretical guarantees are possible under relaxed assumptions is an important topic of future work.

Chapter 6

PRIMAL-DUAL METHODS IN APPROXIMATE POLICY OPTIMIZATION

This chapter is based on Xiong et al. [2024b], with Maryam Fazel and Lin Xiao.

6.1 Introduction

Policy gradient methods represent a paradigm shift in reinforcement learning from value-based methods [Watkins, 1989, Puterman, 1994, Bertsekas, 2015] to a more direct approach of policy optimization [Williams, 1992, Sutton et al., 1999, Konda and Tsitsiklis, 1999]. In particular, the natural policy gradient (NPG) method of Kakade [2001] inspired later development of trust region policy optimization (TRPO) and proximal policy optimization (PPO), both with great empirical success [Schulman et al., 2015, 2017].

Their success also ignited considerable efforts to understand policy gradient methods from a theoretical perspective. Among them, Neu et al. [2017] first connected NPG with the mirror descent (MD) algorithm [Nemirovski and Yudin, 1983, Beck and Teboulle, 2003], which led to a more general class of policy mirror descent (PMD) methods. Convergence guarantees for tabular PMD methods progressed from sublinear convergence [Shani et al., 2020a, Agarwal et al., 2021] to linear convergence [Xiao, 2022, Lan, 2023, Johnson et al., 2023]. Then the linear convergence results were extended to PMD methods with linear function approximation [Yuan et al., 2022], and more recently with general function approximation [Alfano et al., 2024].

However, the progresses of PMD on the empirical and theoretical fronts are more or less disjoint, especially concerning general function approximation. One one hand, Tomar et al. [2020] and Vaswani et al. [2021] derived practical algorithms from the MD principle, but with no or limited convergence guarantees. On the other hand, Alfano et al. [2024] proposed Approximate Mirror Policy Optimization (AMPO), a PMD framework that has

linear convergence guarantee with general function approximation, but has limited empirical success (see our empirical study in Section 6.7).

6.1.1 Contributions and Organization

In this paper, we aim to bridge this gap between theory and practice by proposing Dual Approximation Policy Optimization (DAPO), a new PMD framework that incorporates general function approximation. In contrast to AMPO, which uses the squared L_2-norm to measure the function approximation error and tries to minimize it for policy update, DAPO uses the *dual Bregman divergence* generated by the mirror map used for policy projection. We organize remaining parts of this paper as follows.

In Section 6.3, we first briefly review the basics of Markov decision processes (MDPs) and the MD algorithm. Then, in order to work with negative entropy restricted on the simplex, we further extend the MD algorithm to work with mirror maps whose gradient mapping and conjugate gradient mapping are not inverses of each other.

In Section 6.4, we propose the general framework of DAPO and present several instantiations of DAPO using different mirror maps, including the squared L_2-norm (DAPO-L_2), negative entropy on the positive orthant and negative entropy restricted on the simplex (DAPO-KL). We will elaborate the subtle but important difference between the later two examples and show that DAPO-KL includes two state-of-the-art practical algorithms as special cases: Soft Actor-Critic (SAC) of Haarnoja et al. [2018a] and Mirror Descent Policy Optimization (MDPO) of Tomar et al. [2020].

In Section 6.5, under different choice of step sizes, we prove both $O(1/K)$ and linear convergence rates under general function approximation for two variants, DAPO-L_2 and DAPO-KL, thus immediately providing MDPO with strong convergence guarantees. Then, we in further prove $O(1/K)$ convergence rate for SAC under the framework of entropy-regularized reinforcement learning.

In Section 6.6, we further extend the convergence guarantees for DAPO-KL to MDPs with continuous state and action spaces, including a rigorous development of the MD algorithm in functional space. This is an important result that has been missing from the literature

despite the wide application of PMD methods in continuous state-action spaces.

Finally, in Section 6.7, compare DAPO with SAC and AMPO on several standard MuJoCo benchmark tasks to demonstrate the effectiveness of this duality framework.

6.2 Related Work

PG and PMD in tabular MDPs. Although the proposal of policy gradient theorem and natural policy gradient (NPG) can be traced back to around 2000s or even before [Williams, 1992, Konda and Tsitsiklis, 1999, Sutton et al., 1999, Kakade, 2001], the study of its convergence to the global optimum only started in recent years. On the other hand, mirror descent algorithm [Nemirovski and Yudin, 1983] has been extensively studied for a long time as an online learning algorithm Bubeck et al. [2012]. To connect these two, Neu et al. [2017] first shows that NPG can be viewed as a special case of policy mirror descent (PMD) and most of the following convergence analyses are based on this viewpoint. For tabular MDPs, Shani et al. [2020a] shows that unregularized NPG with a softmax policy has a $O(1/\sqrt{K})$ convergence rate. Agarwal et al. [2021], Vieillard et al. [2020], Xu et al. [2020] then improve it to the $O(1/K)$ convergence rate under different settings. After that, Khodadadian et al. [2021], Bhandari and Russo [2021], Xiao [2022] prove the linear convergence rate for the NPG method. Very recently, Johnson et al. [2023] shows that a linear convergence rate is optimal for NPG in tabular MDPs and Mei et al. [2023] provides a new perspective by proving a necessary and sufficient ordering-based condition for NPG convergence in bandit setting.

PG and PMD in regularized MDPs. Another parallel line of work analyzes applying NPG method to maximum entropy reinforcement learning. Cayci et al. [2021], Cen et al. [2022b] show that NPG with softmax policies can converge linearly in entropy-regularized MDPs while Lan [2023] also shows general PMD method converges linearly. Then, the linear convergence of PMD is extended to MDPs with general convex regularizers by Zhan et al. [2023]. Meanwhile, Li et al. [2022] and Lan et al. [2023] also propose other variants of PMD methods that converge linearly in entropy-regularized MDPs.

PG and PMD with function approximation. Agarwal et al. [2021] shows Q-NPG with log-linear policies achieves $O(1/\sqrt{K})$ convergence rate while Cayci et al. [2021] and Yuan et al. [2022] show that NPG with log-linear policies can converge linearly in entropy-regularized MDPs and unregularized MDPs. Meanwhile, Chen et al. [2022b] and Chen and Maguluri [2022] show similar $O(1/K)$ and linear convergence result under different assumptions, respectively. For more general function approximation setting, Wang et al. [2019] shows that NPG with two-layer neural network has $O(1/\sqrt{K})$ convergence rate and Liu et al. [2019] shows that NPG with multi-layer neural network achieves $O(1/\sqrt{K})$ convergence rate. Recently, Alfano et al. [2024] shows PMD method with general function approximation can converge linearly. The main difference between Alfano et al. [2024] and our work lies on how we define approximation, as discussed in Section 6.4.2.

PG and PMD in continuous MDPs. For continuous MDPs, recent works studying policy optimization mostly focus on control problems with specific structure [Hu et al., 2023]. In particular, Fazel et al. [2018] shows that NPG can converge to the global optimum in LQR problem and related results for LQG problem are discussed in Zheng et al. [2022]. For continuous-space MDPs, Pirotta et al. [2015] show a monotonic improvement of vanilla PG method in Lipschitz MDPs while Bedi et al. [2022] treat the value function as a general non-convex objective and shows convergence to stationary points. For MDPs with both continuous space and time, Lee and Sutton [2021] studies a continuous version of policy iteration; Munos [2006] proposes a continuous-time policy gradient theorem while Zhao et al. [2023] develops a continuous-time TRPO algorithm and studies its monotonic improvement property. Lan [2022] is the only known work that studies convergence of PMD method in general continuous-space MDPs and achieves $O(1/K)$ convergence rate.

Applications of PG. Together with the rise of deep Q-learning [Mnih et al., 2013], PG methods have also inspired many successful practical algorithms for real-world control task, including DDPG in Lillicrap et al. [2015], TRPO in Schulman et al. [2015], PPO in Schulman et al. [2017] and SAC in Haarnoja et al. [2018b,a]. Recently, Tomar et al. [2020] and Vaswani et al. [2021] propose general policy optimization algorithms based on mirror descent that

are similar to ours. However, both of them treat policy parameterization as a black box and neither provides a convergence rate analysis.

Other related work. The capability of policy gradient methods to do exploration in MDPs is also studied in Cai et al. [2020], Agarwal et al. [2020], Shani et al. [2020b], Zanette et al. [2021]. Grudzien et al. [2022] proposes an abstract framework called mirror learning for both tabular and continuous-space MDPs that includes mirror descent as a special case. It provides an asymptotic convergence analysis but does not consider any function approximation setting. Finally, for optimization in functional space, Chu et al. [2019] provides a framework setup that unifies variational inference and reinforcement learning. More recently, Aubin-Frankowski et al. [2022] studies mirror descent in general functional space with a rigorous convergence rate analysis. However, it only focuses on the primal space.

6.3 Preliminaries and New Foundations

We first review the background of Markov decision processes (MDPs) and the general mirror descent method.

6.3.1 Markov Decision Processes

Let $\Delta(\mathcal{X}) = \left\{ p \in \mathbb{R}^{|\mathcal{X}|} \;\middle|\; \sum_{x\in\mathcal{X}} p_x = 1 \text{ and } p_x \geq 0, \forall x \right\}$ denote the probability simplex over an arbitrary finite set $\mathcal{X}$. We consider an infinite-horizon Markov Decision Process (MDP), denoted as $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathcal{P}, c, \gamma)$, where $\mathcal{S}$ is a finite state space, $\mathcal{A}$ is a finite action space, $\mathcal{P} : \mathcal{S} \times \mathcal{A} \mapsto \Delta(\mathcal{S})$ is the transition kernel, $c : \mathcal{S} \times \mathcal{A} \mapsto [0, 1]$ is the single-step cost function and $\gamma \in (0, 1)$ is the discount factor. A stationary policy is defined as a function $\pi : \mathcal{S} \mapsto \Delta(\mathcal{A})$ such that π_s is a probability distribution over $\mathcal{A}$ for each $s \in \mathcal{S}$. At each time t, an agent with policy π takes an action $a_t \sim \pi_{s_t}$, which sends the MDP to the new state $s_{t+1} \sim \mathcal{P}(s_t, a_t)$ and incurs a single-step cost $c(s_t, a_t)$.

The main objective in reinforcement learning is to find a policy that minimizes the accumulated, discounted cost starting from an initial state distribution $\rho \in \Delta(\mathcal{S})$. Formally,

it is defined as $V_\rho^\pi = \mathbb{E}_{s\sim\rho}[V_s^\pi]$, where

$$V_s^\pi = \mathbb{E}_{a_t\sim\pi_{s_t}}\left[\sum_{t=0}^{\infty}\gamma^t c(s_t,a_t)\;\middle|\; s_0 = s\right]. \tag{6.1}$$

The corresponding Q-value function under policy π and state-action pair (s,a) is defined as

$$Q_{s,a}^\pi = \mathbb{E}_{a_t\sim\pi_{s_t}}\left[\sum_{t=0}^{\infty}\gamma^t c(s_t,a_t)\;\middle|\; s_0 = s, a_0 = a\right]. \tag{6.2}$$

We use $Q_s^\pi \in \mathbb{R}^{|\mathcal{A}|}$ to denote the vector $\left[Q_{s,a}^\pi\right]_{a\in\mathcal{A}}$ and we immediately have $V_s^\pi = \langle Q_s^\pi, \pi_s\rangle$.

With initial distribution $\rho \in \Delta(\mathcal{S})$, we define the discounted state-visitation distribution under policy π as

$$d_{\rho,s}^\pi = (1-\gamma)\sum_{t=0}^{\infty}\gamma^t\,\mathbb{P}_{s_0\sim\rho}^\pi(s_t = s), \tag{6.3}$$

where $\mathbb{P}_{s_0\sim\rho}^\pi(s_t = s)$ represents the probability that $s_t = s$ if the agent follows policy π and the initial state s_0 is sampled from distribution ρ. We can easily verify that $\sum_{s\in\mathcal{S}} d_{\rho,s}^\pi = 1$ and thus $d_\rho^\pi \in \mathbb{R}^{|\mathcal{S}|}$ is a valid probability distribution. Meanwhile, by truncating all terms with $t \geq 1$, we obtain $d_{\rho,s}^\pi \geq (1-\gamma)\rho_s$ for any $s \in \mathcal{S}$.

The gradient of V_ρ^π with respect to the policy is given by the famous policy gradient theorem as [Sutton et al., 1999]

$$\nabla_s V_\rho^\pi := \frac{\partial V_\rho^\pi}{\partial \pi_s} = \frac{1}{1-\gamma} d_{\rho,s}^\pi Q_s^\pi. \tag{6.4}$$

Notice that $\nabla_s V_\rho^\pi \in \mathbb{R}^{|\mathcal{A}|}$ and we define $\nabla V_\rho^\pi \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{A}|}$ as the concatenation of $\nabla_s V_\rho^\pi$ for all $s \in \mathcal{S}$.

6.3.2 Mirror Descent

Mirror descent (MD) is a general framework for the construction and analysis of optimization algorithms [Nemirovski and Yudin, 1983]. Its key machinery is a pair of conjugate mirror maps that map the iterations of an optimization algorithm back-and-forth between a primal space and a dual space. We follow the common practice of defining the mirror maps with

the gradient mapping of a convex function of *Legendre-type* [Rockafellar, 1970, Section 26]. Then, we will provide a novel relaxation of it, which serves as the new foundation of our later analysis.

Legendre function and relaxations

Let $\mathcal{X}$ be a normed vector space, possibly of infinite dimension, and $\Phi : \mathcal{X} \to \mathbb{R} \cup \{+\infty\}$ a proper, closed convex function with $\operatorname{dom} \Phi = \{x \in \mathcal{X} \mid \Phi(x) < +\infty\}$.

Definition 6.3.1. The function Φ is of *Legendre type* if

(a) The interior of $\operatorname{dom} \Phi$, denoted by $\mathcal{D}$, is nonempty;

(b) Φ is differentiable and strictly convex on $\mathcal{D}$;

(c) For any sequence $\{x_n\} \subset \mathcal{D}$ which converges to a boundary point of $\mathcal{D}$, it holds that $\lim_{n\to\infty} \|\nabla\Phi(x_n)\| = \infty$.

Let $\mathcal{X}^*$ be the dual vector space of $\mathcal{X}$. The (Legendre) conjugate of Φ is defined as follows: for any $x^\star \in \mathcal{X}^\star$,

$$\Phi^*(x^*) = \sup_{x \in \operatorname{dom} \Phi} \{\langle x, x^* \rangle - \Phi(x)\}. \tag{6.5}$$

Similarly, $\operatorname{dom} \Phi^* = \{x^* \in \mathcal{X}^* \mid \Phi^*(x^*) < +\infty\}$ and $\mathcal{D}^* = \operatorname{int}(\operatorname{dom} \Phi^*)$. If Φ is of Legendre type, then its gradient $\nabla\Phi$ is one-to-one from $\mathcal{D}$ to $\mathcal{D}^*$ and $\nabla\Phi^* = (\nabla\Phi)^{-1}$; in other words, for any $x \in \mathcal{D}$ and $x^* \in \mathcal{D}^*$,

$$\nabla\Phi^*(\nabla\Phi(x)) = x, \qquad \nabla\Phi(\nabla\Phi^*(x^*)) = x^*. \tag{6.6}$$

See Rockafellar [1970, Theorem 26.5] for further details.

However, if Φ is not of Legendre type, then (6.6) may not hold. In particular, this is the case if the $\operatorname{dom} \Phi$ is the simplex $\Delta = \{x \in \mathbb{R}^n_+ \mid \sum_i x_i = 1\}$, which has an empty interior. In fact, such functions are not even differentiable. To see this, let $\Phi(x) = \phi(x) + \delta(x|\Delta)$ where ϕ is convex and differentiable over $\mathbb{R}^n$, and $\delta(\cdot|\Delta)$ is the indicator function of Δ, i.e.,

$\delta(x|\Delta) = 0$ if $x \in \Delta$ and $+\infty$ otherwise. Then Φ is not a differentiable function. However, it is subdifferentiable with subdifferential

$$\partial\Phi(x) = \{\nabla\phi(x) + c\mathbf{1} \,|\, c \in \mathbb{R}\}, \tag{6.7}$$

where $\mathbf{1} = \begin{bmatrix} 1 & \cdots & 1 \end{bmatrix}^\top$.

Admittedly, although being a common assumption, relation (6.6) is not a necessary condition for vanilla mirror descent as presented in Beck and Teboulle [2003]. However, a relation similar to (6.6) will play an important role in our analysis because it helps avoid non-convexity when we consider policy with general parameterization. Therefore, given the importance of simplex in studying MDPs, a relaxation of of (6.6) will be critical to our main results, which is presented in Lemma 6.3.2 as follows.

Lemma 6.3.2. *Suppose* $\Phi(x) = \phi(x) + \delta(x|\mathcal{L})$ *where* ϕ *is a convex function of Legendre type and* $\mathcal{L}$ *is an affine subspace. Assume that* $\operatorname{int}(\operatorname{dom}\phi) \cap \mathcal{L} \neq \emptyset$. *Then we have*

$$\nabla\Phi^*(\nabla\Phi(x)) = x, \qquad \forall\, x \in \operatorname{int}(\operatorname{dom}\phi) \cap \mathcal{L}.$$

And for any $x^* \in \operatorname{int}(\operatorname{dom}\Phi^*)$ *and any* $x, y \in \operatorname{dom}\Phi$,

$$\big\langle \nabla\Phi(\nabla\Phi^*(x^*)), x - y \big\rangle = \langle x^*, x - y \rangle,$$

where $\nabla\Phi(x)$ *denotes any subgradient in* $\partial\Phi(x)$.

Proof. Let $\mathcal{L} = x_0 + \mathcal{V}$ where $\mathcal{V}$ is a subspace, and denote $\mathcal{V}^\perp$ its orthogonal complement. First, it is commonly known that the subdifferential of an indicator function is a normal cone [Bertsekas, 2009]. Thus, we have $\partial\delta(x \mid \mathcal{L}) = \mathcal{N}_{\mathcal{L}}(x) \overset{\text{def}}{=} \{g' \mid \langle g', v + x_0 - x\rangle \leq 0, \forall v \in \mathcal{V}\}$. That is, for any $g' \in \mathcal{N}_{\mathcal{L}}(x)$, we have $\langle g', v\rangle \leq \langle g', x - x_0\rangle$ for any $v \in \mathcal{V}$. Since $\mathcal{V}$ is a subspace, for any $v \in \mathcal{V}$, we have $\alpha v \in \mathcal{V}$ for any $\alpha \in \mathbb{R}$. Therefore, we must have $\langle g', v\rangle = 0$ for any $v \in \mathcal{V}$ and $g' \in \mathcal{N}_{\mathcal{L}}(x)$, which leads to $\mathcal{N}_{\mathcal{L}}(x) = \mathcal{V}^\perp$. (The reverse side is obvious.)

Suppose $x \in \operatorname{int}(\operatorname{dom}\phi) \cap \mathcal{L}$. Then, according to the subdifferential calculus rule, we must have $\nabla\Phi(x) = \nabla\phi(x) + \xi$ for some $\xi \in \mathcal{N}_{\mathcal{L}}(x) = \mathcal{V}^\perp$. Let $x' \triangleq \nabla\Phi^*(\nabla\Phi(x))$. By

definition of Φ^* and strict convexity of ϕ, we have

$$x' = \underset{z \in \mathcal{L} \cap \operatorname{dom} \phi}{\arg\max} \left\{ \langle z, \nabla\phi(x) + \xi \rangle - \phi(z) \right\}.$$

The optimality condition of the above problem is

$$\nabla\phi(x) + \xi - \nabla\phi(x') \in \mathcal{N}_{\mathcal{L} \cap \operatorname{dom} \phi}(x') = \mathcal{V}^\perp + \mathcal{N}_{\operatorname{dom} \phi}(x'),$$

where the last equality above holds because $\mathcal{N}_{\mathcal{L}}(x) = \mathcal{V}^\perp$. Note that $x' = \nabla\Phi^*(\nabla\Phi(x))$ implies $\nabla\Phi(x) \in \partial\Phi(x') = \partial\phi(x') + \partial\delta(x'|\mathcal{L})$. As shown in Rockafellar [1967], $\partial\phi(x) = \emptyset$ for any $x \in \operatorname{bd} \operatorname{dom} \phi$ for a Legendre type function ϕ. Therefore, we must have $x' \in \operatorname{int}(\operatorname{dom} \phi)$, which then implies $\mathcal{N}_{\operatorname{dom} \phi}(x') = \{\mathbf{0}\}$. Thus, we have

$$\nabla\phi(x) + \xi - \nabla\phi(x') \in \mathcal{V}^\perp$$

Since $\xi \in \mathcal{V}^\perp$, we conclude that $\nabla\phi(x) - \nabla\phi(x') \in \mathcal{V}^\perp$. On the other hand, we have $x, x' \in \mathcal{L}$, which implies that $x - x' \in \mathcal{V}$. Therefore,

$$\left\langle \nabla\phi(x) - \nabla\phi(x'),\, x - x' \right\rangle = 0.$$

Since ϕ is strictly convex, we must have $x = x'$, thus proving $\nabla\Phi^*(\nabla\Phi(x)) = x$.

To prove the second statement, let $x' \triangleq \nabla\Phi^*(x^*)$, i.e.,

$$x' = \underset{z \in \mathcal{L} \cap \operatorname{dom} \phi}{\arg\max} \left\{ \langle z, x^* \rangle - \phi(z) \right\}.$$

By similar reasoning, we have $x' \in \operatorname{int}(\operatorname{dom} \phi)$. Thus, the optimality condition is $x^* - \nabla\phi(x') \in \mathcal{V}^\perp$, meaning $\nabla\phi(x') = x^* + \xi$ for some $\xi \in \mathcal{V}^\perp$. Meanwhile,

$$\nabla\Phi(\nabla\Phi^*(x^*)) = \nabla\Phi(x') = \nabla\phi(x') + \xi' = x^* + \xi + \xi',$$

where $\xi' \in \mathcal{V}^\perp$. Since $\xi, \xi' \in \mathcal{V}^\perp$ and $x - y \in \mathcal{V}$, we have

$$\langle \nabla\Phi(\nabla\Phi^*(x^*)), x - y\rangle = \langle x^*, x - y\rangle.$$

This finishes the proof. □

Notice that if $\operatorname{dom}\phi = \mathbb{R}^n_+$ and $\mathcal{L} = \{x \in \mathbb{R}^n | \mathbf{1}^T x = 1\}$, then $\operatorname{dom}\Phi = \operatorname{dom}\phi \cap \mathcal{L} = \Delta$. This is how we will invoke Lemma 6.3.2 with ϕ being the negative entropy function. Here, we call the function Φ defined in Lemma 6.3.2 as the *relaxed Legendre-type function*.

Mirror descent (MD) algorithm

To describe the MD algorithm, we first define *Bregman divergence* and *Bregman projection*. Given a convex function of Legendre type, Φ, the Bregman divergence between any $x \in \operatorname{dom}\Phi$ and $y \in \operatorname{int}(\operatorname{dom}\Phi)$ is defined as

$$D_\Phi(x, y) = \Phi(x) - \Phi(y) - \langle \nabla\Phi(y), x - y\rangle . \tag{6.8}$$

Furthermore, this is also well-defined for relaxed Legendre-type function defined in Lemma 6.3.2, as shown in the following corollary.

Corollary 6.3.3. *In the setting of Lemma 6.3.2, for any $x, y, z \in \operatorname{dom}\Phi$ and $g \in \partial\Phi(z)$, we have*

$$\langle g, x - y\rangle = \langle \nabla\phi(z), x - y\rangle ,$$

which makes expression $\langle \nabla\Phi(z), x - y\rangle$ well-defined for $x, y \in \operatorname{dom}\Phi$.

Proof. Again let $\mathcal{L} = x_0 + \mathcal{V}$. As shown in the proof of Lemma 6.3.2, we have $\partial\Phi(z) = \nabla\phi(z) + \partial\delta(z \mid \mathcal{L})$ and $\partial\delta(z \mid \mathcal{L}) = \mathcal{V}^\perp$. Then, since $\operatorname{dom}\Phi \subseteq \mathcal{L}$, we have $x - y \in \mathcal{V}$, which means to have $\langle g', x - y\rangle = 0$ for any $g' \in \partial\delta(z \mid \mathcal{L})$. Therefore, we have

$$\langle g, x - y\rangle = \langle \nabla\phi(z), x - y\rangle + \langle g', x - y\rangle = \langle \nabla\phi(z), x - y\rangle .$$

□

Let $\mathcal{C}$ be a closed convex set contained in $\operatorname{dom}\Phi$. The *Bregman projection* of any $y \in \operatorname{int}(\operatorname{dom}\Phi)$ onto $\mathcal{C}$ is

$$\operatorname{proj}_{\mathcal{C}}^{\Phi}(y) = \arg\min_{x\in\mathcal{C}} D_{\Phi}(x, y). \tag{6.9}$$

Now consider the problem of minimizing a convex function $f : \mathcal{X} \to \mathbb{R} \cup \{\infty\}$ over a closed convex set $\mathcal{C} \subset \operatorname{dom}\Phi$. We use the presentation of MD given by Bubeck [2015]:

1. Given $x^{(k)}$, find $y^{(k+1)}$ such that

$$\nabla\Phi(y^{(k+1)}) = \nabla\Phi(x^{(k)}) - \eta_k g^{(k)}. \tag{6.10}$$

   where η_k is the step size at iteration k, and $g^{(k)}$ can be the gradient $\nabla f(x^{(k)})$ or a sub-gradient of f at $x^{(k)}$.

2. Compute $x^{(k+1)} = \operatorname{proj}_{\mathcal{C}}^{\Phi}(y^{(k+1)})$.

Using the first identity in (6.6) and the definition in (6.9), we can express the MD algorithm more compactly as

$$x^{(k+1)} = \arg\min_{x\in\mathcal{C}} D_{\Phi}\left(x, \nabla\Phi^*\left(\nabla\Phi(x^{(k)}) - \eta_k g^{(k)}\right)\right) \tag{6.11}$$

It can be further simplified to the following well-known form Beck and Teboulle [2003], Bubeck et al. [2012]

$$x^{(k+1)} = \arg\min_{x\in\mathcal{C}} \left\{\eta_k \langle g^{(k)}, x\rangle + D_{\Phi}(x, x^{(k)})\right\}. \tag{6.12}$$

Next we discuss three examples of the MD algorithm for solving $\min_{x\in\Delta} f(x)$, where Δ is the simplex. Each leads to a variant of the DAPO method we present in Section 6.4.

Example 6.3.4 (Squared L_2-norm)**.** Let $\Phi(x) = \frac{1}{2}\|x\|_2^2$, which is Legendre type with $\operatorname{int}(\operatorname{dom}\Phi) = \operatorname{dom}\Phi = \mathbb{R}^n$. We have $\Phi^*(x^*) = \frac{1}{2}\|x^*\|_2^2$, $\nabla\Phi(x) = x$, $\nabla\Phi^*(x^*) = x^*$, and $\mathcal{D}_{\Phi}(x, y) = \frac{1}{2}\|x - y\|_2^2$. In this case, the MD algorithm (6.11) becomes the classical

projected gradient method

$$x^{(k+1)} = \arg\min_{x\in\Delta} \left\| x - \left(x^{(k)} - \eta_k \nabla f(x^{(k)})\right)\right\|_2^2.$$

Example 6.3.5 (Negative entropy on $\mathbb{R}^n_+$)**.** Consider the (generalized) negative entropy $\Phi(x) = \sum_i (x_i \log(x_i) - x_i)$ with $\operatorname{dom}\Phi = \mathbb{R}^n_+$ (and the convention $0\log 0 = 0$). It is of Legendre type, with $\Phi^*(x^*) = \sum_i \exp(x_i^*)$, $\nabla\Phi(x) = \log(x)$ and $\nabla\Phi^*(x^*) = \exp(x^*)$, where log and exp apply component-wise to vectors. For any $x \in \mathbb{R}^n_+$ and $y \in \mathbb{R}^n_{++}$, their Bregman divergence is the KL-divergence:

$$D_\Phi(x, y) = \sum_i \left(x_i \log \frac{x_i}{y_i} - x_i + y_i \right). \tag{6.13}$$

In this case, the Bregman projection of $y \in \mathbb{R}^n_{++}$ onto Δ is $\operatorname{proj}^\Phi_\Delta(y) = y/\|y\|_1$ and the MD algorithm (6.11) becomes

$$x^{(k+1)} = \frac{x^{(k)} \exp(-\eta_k g^{(k)})}{\|x^{(k)} \exp(-\eta_k g^{(k)})\|_1}. \tag{6.14}$$

Example 6.3.6 (Negative entropy on Δ)**.** Let ϕ be the negative entropy function, that is, $\phi(x) = \sum_i (x_i \log(x_i) - x_i)$ and define $\Phi(x) = \phi(x) + \delta(x|\Delta)$. Apparently $\operatorname{dom}\Phi = \Delta$, which has an empty interior. As discussed in Section 6.3.2, Φ is not of Legendre type and in fact is not differentiable. However, the MD algorithm is still well-defined. Specifically, in (6.10) we interpret $\nabla\Phi(x^{(k)})$ as any subgradient in $\partial\Phi(x^{(k)})$, and find $y^{(k+1)}$ such that there exists some $\nabla\Phi(y^{(k+1)}) \in \partial\Phi(y^{(k+1)})$ to make the equality hold.

Despite $\nabla\Phi(y)$ being multi-valued as in (6.7), the Bregman divergence (6.8) is also well-defined by Corollary 6.3.3. Therefore, D_Φ is the same as (6.13). Using the fact $x, y \in \Delta$, it can be simplified as

$$D_\Phi(x, y) = \sum_i x_i \log \frac{x_i}{y_i}. \tag{6.15}$$

In addition, we have $\Phi^*(x^*) = \log\left(\sum_i \exp(x_i^*)\right)$ with $\operatorname{dom}\Phi^* = \mathbb{R}^n$ [Rockafellar, 1970,

Section 16]. Clearly, Φ^* is a differentiable function throughout $\mathbb{R}^n$ and

$$\nabla\Phi^*(x^*) = \frac{\exp(x^*)}{\|\exp(x^*)\|_1}.$$

In this case, the MD algorithm (6.11) yields the same update as (6.14). However, the projection step is no longer needed because the range of $\nabla\Phi^*$ is the interior of Δ and we can simply express the MD algorithm as

$$x^{(k+1)} = \nabla\Phi^*\big(\nabla\Phi(x^{(k)}) - \eta_k g^{(k)}\big).$$

Although Examples 6.3.5 and 6.3.6 give the same MD update, their subtle difference play a crucial role in our development of policy optimization methods in the next section.

Dual Bregman Divergence

Finally, as a key tool of our duality framework, we introduce the *dual Bregman divergence* D_{Φ^*}, which is induced by the convex conjugate function Φ^*. Its dual relationship with D_Φ is characterized by the following corollary.

Corollary 6.3.7. *Let Φ be a Legendre-type function given in Definition 6.3.1 or relaxed Lgendre-type function given in Lemma 6.3.2. For $x^*, y^* \in \operatorname{int}(\operatorname{dom}\Phi^*)$, we have*

$$D_{\Phi^*}(x^*, y^*) = D_\Phi(\nabla\Phi^*(y^*), \nabla\Phi^*(x^*)).$$

Proof. By definition of the Bregman divergence in Eq. (6.8), we have

$$\begin{aligned}
&D_{\Phi^*}(x^*, y^*) - D_\Phi(\nabla\Phi^*(y^*), \nabla\Phi^*(x^*)) \\
=&\Phi^*(x^*) - \Phi^*(y^*) - \langle\nabla\Phi^*(y^*), x^* - y^*\rangle \\
&\quad - [\Phi(\nabla\Phi^*(y^*)) - \Phi(\nabla\Phi^*(x^*)) - \langle\nabla\Phi(\nabla\Phi^*(x^*)), \nabla\Phi^*(y^*) - \nabla\Phi^*(x^*)\rangle] \\
=&\,[\Phi^*(x^*) + \Phi(\nabla\Phi^*(x^*))] - [\Phi^*(y^*) + \Phi(\nabla\Phi^*(y^*))] \\
&\quad - \langle\nabla\Phi^*(y^*), x^* - y^*\rangle + \langle x^*, \nabla\Phi^*(y^*) - \nabla\Phi^*(x^*)\rangle \\
&\qquad\qquad \text{(By Lemma 6.3.2 and } \nabla\Phi^*(y^*), \nabla\Phi^*(x^*) \in \operatorname{dom}\Phi \text{ or Eq. (6.6).)}
\end{aligned}$$

$$= \langle x^*, \nabla\Phi^*(x^*)\rangle - \langle y^*, \nabla\Phi^*(y^*)\rangle - \langle \nabla\Phi^*(y^*), x^* - y^*\rangle + \langle x^*, \nabla\Phi^*(y^*) - \nabla\Phi^*(x^*)\rangle$$

(By Bertsekas [2009, Proposition 5.4.3].)

$$= \langle \nabla\Phi^*(x^*), x^* - x^*\rangle + \langle \nabla\Phi^*(y^*), -y^* + y^* - x^* + x^*\rangle$$

$$= 0.$$

□

6.4 Dual Approximation Policy Optimization

Notation. Starting from this section, we use $\pi^\star$ to denote arbitrary comparator policy and for policies with index or star such as $\pi^{(k)}$ and $\pi^\star$, we will use $V_\rho^{(k)}$ (respectively $V_\rho^\star$) as a shorthand for $V_\rho^{\pi^{(k)}}$ (respectively $V_\rho^{\pi^\star}$) and similarly for $Q_{s,a}^{(k)}$ and $d_\rho^{(k)}$ (respectively $Q_{s,a}^\star$ and $d_\rho^\star$).

Recall the setting of MDP in Section 6.3.1. In the tabular case, the policy mirror descent (PMD) method Shani et al. [2020a], Lan [2023], Xiao [2022] takes the form of (6.12):

$$\pi_s^{(k+1)} = \operatorname*{arg\,min}_{\pi_s \in \Delta(\mathcal{A})} \Big\{ \eta_k \big\langle \widehat{Q}_s^{(k)}, \pi_s \big\rangle + D_\Phi(\pi_s, \pi_s^{(k)}) \Big\}, \tag{6.16}$$

for each $s \in \mathcal{S}$, where $\widehat{Q}^{(k)}$ is some approximation of $Q^{(k)}$. We note that $Q^{(k)}$ is not the gradient of the value function V_ρ^π at $\pi^{(k)}$, which is given in (6.4); rather, it can be viewed as a preconditioned gradient Kakade [2001].

When the size of the state-action space becomes large (possibly infinite), we have to resort to function approximation. Specifically, let π^θ be a differentiable mapping from the set of parameters $\Theta \subset \mathbb{R}^n$ to the set of stochastic policies. The parameter update step corresponding to (6.16) becomes

$$\theta^{(k+1)} = \operatorname*{arg\,min}_{\theta \in \Theta} \ \mathbb{E}_{s \sim d_\rho^{(k)}} \Big[\mathbb{E}_{a \sim \pi_s^\theta} \big[\widehat{Q}_{s,a}^{(k)} \big] + D_\Phi(\pi_s^\theta, \pi_s^{(k)}) \Big], \tag{6.17}$$

where $\pi^{(k)}$ means $\pi^{\theta^{(k)}}$. This is the approach adopted by, e.g., Tomar et al. [2020] and Vaswani et al. [2021]. However, the above optimization problem with respect to θ is no

Algorithm 10 Dual Approximation Policy Optimization (DAPO)

Input: Initialize policy $\pi^{(0)}$ with parameters $\theta^{(0)}$; mirror map Φ; initial distribution ρ
for $k = 0, \dots, K-1$ **do**
  Find $\widehat{Q}^{(k)}$ that approximates $Q^{(k)}$ (Critic Update)
  Find $\theta^{(k+1)}$ that (approximately) solves the problem

$$\min_{\theta \in \Theta} \mathbb{E}_{s \sim d_\rho^{(k)}} \Big[D_{\Phi^*} \big(\nabla \Phi(\pi_s^{(k)}) - \eta_k \widehat{Q}_s^{(k)},\, f_s^\theta \big) \Big]$$

  Assign $\pi_s^{(k+1)} = \mathrm{proj}_{\Delta(\mathcal{A})}^{\Phi} \big(\nabla \Phi^*(f_s^{\theta^{(k+1)}}) \big), \ \forall\, s \in \mathcal{S}$
end for

longer convex, and convergence analysis becomes more challenging.

Alfano et al. [2024] introduced Approximate Mirror Policy Optimization (AMPO), a framework that incorporates general parametrization into PMD with convergence guarantees. A key instrument they introduced is the *Bregman projected policy* class. The idea is to use a parametrized function $f^\theta : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ to approximate the dual update in (6.10), which in the context of PMD is $\nabla \Phi(\pi^{(k)}) - \eta_k \widehat{Q}^{(k)}$. Then follow the second step in MD to define the policy class

$$\big\{ \pi^\theta : \pi_s^\theta = \mathrm{proj}_{\Delta(\mathcal{A})}^{\Phi} \big(\nabla \Phi^*(f_s^\theta) \big), \ s \in \mathcal{S} \big\}, \quad \theta \in \Theta.$$

For example, with the negative-entropy mirror map (Example 6.3.5 or 6.3.6), it leads to the softmax policy class:

$$\pi_{s,a}^\theta = \frac{\exp(f_{s,a}^\theta)}{\| \exp(f_s^\theta) \|_1}, \quad (s,a) \in \mathcal{S} \times \mathcal{A}. \tag{6.18}$$

While such policy classes are widely used in both theory and practice, recognizing them as the composition of a Bregman projection, a conjugate mirror map and a generic function approximation f^θ (such as neural networks) allows more structured and sharper convergence analysis.

Facilitated with the Bregman projected policy class, extending PMD with function approximation rests upon how we approximate $\nabla \Phi(\pi^{(k)}) - \eta_k \widehat{Q}^{(k)}$ (*existing in the dual space*) using f^θ. AMPO Alfano et al. [2024] proposes to minimize the expected L_2-distance between

them, which is

$$\min_{\theta} \mathbb{E}_{s\sim d_\rho^{(k)}}\Big[\big\|f_s^\theta - \big(\nabla\Phi(\pi_s^{(k)}) - \eta_k \widehat{Q}_s^{(k)}\big)\big\|_2^2\Big], \tag{6.19}$$

On the other hand, Lan [2022] tries to minimize the expected (in state distribution) L_∞-norm of the difference between them.

In contrast, we propose to use the corresponding *dual Bregman divergence* D_{Φ^*} to measure their similarity in the dual space. In particular, Our method finds $\theta^{(k+1)}$ by (approximately) solving

$$\min_{\theta\in\Theta} \mathbb{E}_{s\sim d_\rho^{(k)}}\Big[D_{\Phi^*}\big(\nabla\Phi(\pi_s^{(k)}) - \eta_k \widehat{Q}_s^{(k)},\, f_s^\theta\big)\Big], \tag{6.20}$$

where $d_\rho^{(k)}$ can be replaced with other distributions to accommodate the scenario of off-policy training. Here, we can see that the similarity between the two dual vectors f_s^θ and $\nabla\Phi(\pi_s^{(k)}) - \eta\widehat{Q}_s^{(k)}$ are measured by the Bregman divergence of Φ^*, which naturally lives in the dual space. Together with the Bregman divergence of Φ used in policy projection, they form a complete duality framework.

A complete description of our method is given as Algorithm 10, and we call it Dual Approximation Policy Optimization (DAPO).

Remark 6.4.1. Although Vaswani et al. [2024] also applies the dual Bregman divergence, its usage of D_{Φ^*} is derived from relative smoothness and the algorithm uses it mainly for critic update, which has very different spirit and purpose from DAPO.

6.4.1 Instantiations of DAPO

We give three instantiations of DAPO using the three mirror maps given in Examples 6.3.4-6.3.6. In deriving these instantiations as well as implementing the algorithms, instead of directly using the dual Bregman divergence D_{Φ^*}, it is often more convenient to use the identity:

$$D_{\Phi^*}\big(\nabla\Phi(\pi_s^{(k)}) - \eta_k \widehat{Q}_s^{(k)},\, f_s^\theta\big) = D_\Phi\big(\nabla\Phi^*(f_s^\theta),\, \nabla\Phi^*(\nabla\Phi(\pi_s^{(k)}) - \eta_k \widehat{Q}_s^{(k)})\big), \tag{6.21}$$

which can be obtained by directly applying Corollary 6.3.7. This identity will also facilitate our convergence analysis later.

DAPO-L_2. With Φ being the squared L_2-norm mirror map described in Example 6.3.4, the approximation problem in (6.20) (same as line 4 in Algorithm 10) becomes

$$\min_{\theta} \mathbb{E}_{s\sim d_\rho^{(k)}} \left[\left\| f_s^\theta - \pi_s^{(k)} + \eta_k \widehat{Q}_s^{(k)} \right\|_2^2 \right], \tag{6.22}$$

and Line 5 of Algorithm 10 is the Euclidean projection

$$\pi_s^{(k+1)} = \operatorname*{arg\,min}_{\pi\in\Delta(\mathcal{A})} \left\| \pi - f_s^{(k+1)} \right\|_2^2.$$

Here we have used the simpler notation $f_s^{(k+1)}$ for $f_s^{\theta^{(k+1)}}$. We present its convergence analysis in Section 6.5.1

DAPO-KL*. With Φ being the negative entropy over $\mathbb{R}_+^{|\mathcal{A}|}$ (see Example 6.3.5), we have

$$\nabla\Phi^*(f_s^{(k+1)}) = \exp(f_{s,a}^\theta), \qquad \nabla\Phi^*\big(\nabla\Phi(\pi^{(k)}) - \eta_k \widehat{Q}^{(k)}\big) = \pi_s^{(k)} \exp\big(-\eta_k \widehat{Q}_s^{(k)}\big),$$

and loss in the approximation problem (6.20) becomes

$$\mathbb{E}_{s\sim d_\rho^{(k)}} \left[D_{\mathrm{KL}} \left(\exp(f_s^\theta) \middle\| \pi_s^{(k)} \exp\big(-\eta_k \widehat{Q}_s^{(k)}\big) \right) \right], \tag{6.23}$$

where D_{KL} is given by (6.13). Policy projection as in (6.18) is necessary to obtain $\pi^{(k+1)}$ because $\nabla\Phi^*(f_s^{(k+1)})$ is not in the simplex in general. DAPO-KL* has disadvantages in both theory and practice compared with its close variant DAPO-KL, which we will explain next.

DAPO-KL. With Φ being the negative entropy restricted on $\Delta(\mathcal{A})$ (see Example 6.3.6), the range of $\nabla\Phi^*$ is $\Delta(\mathcal{A})$, thus the projection step (Line 5 of Algorithm 10) becomes redundant. In this case, we have

$$\begin{aligned} \pi_s^{(k+1)} = \nabla\Phi^*(f_s^{(k+1)}) &= \exp(f_{s,a}^\theta)/\|\exp(f_s^\theta)\|_1, \\ \nabla\Phi^*\big(\nabla\Phi(\pi^{(k)}) - \eta_k \widehat{Q}^{(k)}\big) &= \pi_s^{(k)} \exp\big(-\eta_k \widehat{Q}_s^{(k)}\big)/Z_s^{(k)}, \end{aligned}$$

where $Z_s^{(k)} = \left\|\pi_s^{(k)} \exp\left(-\eta_k \widehat{Q}_s^{(k)}\right)\right\|_1$. The loss in the approximation problem (6.20) becomes

$$\mathbb{E}_{s\sim d_\rho^{(k)}}\left[D_{\mathrm{KL}}\left(\pi_s^\theta \,\middle\|\, \pi_s^{(k)} \exp\left(-\eta_k \widehat{Q}_s^{(k)}\right)/Z_s^{(k)}\right)\right], \tag{6.24}$$

where D_{KL} is given by (6.15). There are a couple of distinctions between DAPO-KL and DAPO-KL*.

- The approximation loss in (6.24) is in terms of the full policy parametrization π^θ, matching the implementation of several popular algorithms [Tomar et al., 2020, Vaswani et al., 2021]. In contrast, the loss in (6.23) is in terms of the unnormalized entity $\exp(f^\theta)$, which will suffer additional loss after policy projection.

- In theory,we are able to provide a competitive convergence analysis of DAPO-KL (see section 6.5.2) thanks to the fact that the two arguments in D_{KL} in (6.24) are both on the simplex, which is not the case in (6.23).

For these reasons, we will only consider DAPO-KL from now on. However, we felt it is necessary to expose the subtleties between the two variants, because many work on policy mirror descent methods (including Alfano et al. [2024]) assumes the stronger relation (6.6). We hope to demonstrate that the more nuanced theory based on Lemma 6.3.2 is crucial for developing and analyzing practical algorithms.

6.4.2 Comparison with AMPO, MDPO and FMA-PG

AMPO. As given in Alfano et al. [2024], AMPO replaces the minimization problem in Line 4 of Algorithm 10 by (6.19) regardless of the mirror map used in policy projection. More concretely, let Φ_1 be the negative entropy on $\mathbb{R}_+^n$ and Φ_2 be the squared L_2-norm. Then AMPO's approximation loss can be expressed as

$$\mathbb{E}_{s\sim d_\rho^{(k)}}\left[D_{\Phi_2^*}\left(\nabla\Phi_1(\pi_s^{(k)}) - \eta_k \widehat{Q}_s^{(k)}, f_s^\theta\right)\right], \tag{6.25}$$

and Φ_1 is again used in the policy projection step. While in theory, as long as the approximation error is small, it is possible to establish convergence of the method Alfano et al.

[2024]. However, such a mismatch, or inconsistency, between approximations in primal and dual spaces may cause problems when the approximation error cannot be made sufficiently small. This is precisely the case in practice, where we can only afford to run at most a few steps of the stochastic gradient method to reduce the error. The importance of the compatibility between the two mirror maps has also been pointed out by Tomar et al. [2020]. In Section 6.7, we demonstrate that in standard benchmarks DAPO-KL obtains state-of-the-art performance with only one step of stochastic gradient method in reducing the approximation loss (6.20), comparable to SAC Haarnoja et al. [2018b]. On the other hand, we could not get AMPO competitive with many numbers of stochastic gradient steps.

MDPO. The Mirror Descent Policy Optimization (MDPO) method of Tomar et al. [2020] is based on (6.17). If π^θ belongs to the softmax of (6.18) and D_Φ is the KL-divergence, then it is equivalent to DAPO-KL. Therefore, our convergence analysis in Section 6.5.2 directly applies to MDPO, which is not provided by Tomar et al. [2020].

FMA-PG. The Functional Mirror Ascent (FMA-PG) framework of Vaswani et al. [2021] also takes the form (6.17). However, similar to MDPO, Vaswani et al. [2021] did not exploit any composition structure of the parametrization π^θ or the MDP structure. Rather, they conducted convergence analysis based on the general theory for smooth, non-convex optimization, which leads to considerably weaker results.

6.4.3 SAC as a special case of DAPO-KL

Soft Actor-Critic (SAC) Haarnoja et al. [2018a] is a very popular reinforcement learning algorithm, which was developed under the framework of entropy-regularized reinforcement learning. Tomar et al. [2020] compared SAC's actor update loss function with (6.24) and pointed out that SAC is similar to MDPO (same as DAPO-KL) by replacing the $\pi^{(k)}$ with the uniform distribution. Here, we will prove a much stronger result, showing that by choosing the learning rate η_k appropriately, Eq. (6.24) will exactly become the SAC's policy update rule.

To prove this, we will first briefly introduce the framework of entropy-regularized re-

inforcement learning and then derive the corresponding DAPO-KL algorithm under this framework. More details about the theory of maximum entropy reinforcement learning can be found in Cen et al. [2022b] and Cayci et al. [2021]. In this framework, with regularization parameter $\tau > 0$, the objective is to minimize the entropy-regularized value function, which is defined as

$$\begin{aligned} V_{\tau,\rho}^{\pi} =& \mathbb{E}_{a_t \sim \pi_{s_t}} \left[\sum_{t=0}^{\infty} \gamma^t c(s_t, a_t) \,\middle|\, s_0 \sim \rho \right] + \tau \mathbb{E}_{a_t \sim \pi_{s_t}} \left[\sum_{t=0}^{\infty} \gamma^t \log \pi(a_t \mid s_t) \,\middle|\, s_0 \sim \rho \right] \\ =& \mathbb{E}_{a_t \sim \pi_{s_t}} \left[\sum_{t=0}^{\infty} \gamma^t \left(c(s_t, a_t) + \tau \log \pi(a_t \mid s_t) \right) \,\middle|\, s_0 \sim \rho \right]. \end{aligned}$$

Then, we can similarly define the Q-value function as

$$Q_{\tau}^{\pi}(s,a) = \mathbb{E}_{a_t \sim \pi_{s_t}} \left[\sum_{t=0}^{\infty} \gamma^t \left(c(s_t, a_t) + \tau \log \pi(a_t \mid s_t) \right) \,\middle|\, s_0 = s, a_0 = a \right].$$

Clearly, for any $(s,a) \in \mathcal{S} \times \mathcal{A}$, $Q_{\tau}^{\pi}(s,a)$ and $V_{\tau}^{\pi}(s)$ satisfy

$$\begin{cases} Q_{\tau}^{\pi}(s,a) = c(s,a) + \tau \log \pi(a \mid s) + \gamma \mathbb{E}_{s' \sim \mathcal{P}(s,a)} \left[V_{\tau}^{\pi}(s') \right], \\ V_{\tau}^{\pi}(s) = \mathbb{E}_{a' \sim \pi_s} \left[Q_{\tau}^{\pi}(s,a') \right]. \end{cases} \tag{6.26}$$

As shown in Cayci et al. [2021], its policy gradient is $\nabla_s V_{\tau,\rho}^{\pi} = \frac{1}{1-\gamma} d_{\rho,s}^{\pi} Q_{\tau,s}^{\pi} \in \mathbb{R}^{|\mathcal{A}|}$. Therefore, we can obtain the corresponding DAPO algorithm by using this policy gradient. That is, with mirror map Φ, the policy update rule in line 4 of Algorithm 10 becomes

$$\theta^{(k+1)} \in \arg\min_{\theta} \mathbb{E}_{s \sim d_{\rho}^{(k)}} \left[D_{\Phi^*} \left(\nabla \Phi(\pi_s^{(k)}) - \eta_k Q_{\tau,s}^{(k)}, f_s^{\theta} \right) \right],$$

where we use the exact Q-value function for simplicity. Then, by instantiating Φ as the negative entropy, as derived in Example 6.3.6, we can get the update rule for DAPO-KL as

$$\theta^{(k+1)} \in \arg\min_{\theta} \mathbb{E}_{s \sim d_{\rho}^{(k)}} \left[D_{\mathrm{KL}} \left(\pi_s^{\theta} \,\middle\|\, \pi_s^{(k)} \exp\left(-\eta_k Q_{\tau,s}^{(k)} \right) / Z_s^{(k)} \right) \right], \tag{6.27}$$

where $Z_s^{(k)}$ is the normalization factor.

Now, we turn our attention to the SAC algorithm in Haarnoja et al. [2018a]. Here, the subtlety is that the soft Q-value function used in Haarnoja et al. [2018a] is defined differently from the one in Eq. (6.26) even though the notation is the same. To be clear, we use q_τ^π to denote the soft Q-value function, which for any $(s,a) \in \mathcal{S} \times \mathcal{A}$ is recursively defined as

$$\begin{cases} q_\tau^\pi(s,a) = c(s,a) + \gamma \mathbb{E}_{s' \sim \mathcal{P}(s,a)} \left[V_\tau^\pi(s') \right], \\ V_\tau^\pi(s) = \mathbb{E}_{a' \sim \pi_s} \left[\tau \log \pi(a' \mid s) + q_\tau^\pi(s,a') \right]. \end{cases} \tag{6.28}$$

Note that the definition of V_τ^π remains unaffected by writing V_τ^π explicitly through either Eq. (6.28) or Eq. (6.26). Then, we can immediately obtain the relation $q_\tau^\pi(s,a) = Q_\tau^\pi(s,a) - \tau \log \pi(a \mid s)$ for any $(s,a) \in \mathcal{S} \times \mathcal{A}$. As a result, we can see that the policy update rule in SAC is*

$$\begin{aligned} \theta^{(k+1)} &\in \arg\min_\theta \mathbb{E}_{s \sim d_\rho^{(k)}} \left[D_{\mathrm{KL}} \left(\pi_s^\theta \;\middle\|\; \exp\left(-q_{\tau,s}^{(k)} / \tau \right) / Z_s^{(k)} \right) \right] \\ &= \arg\min_\theta \mathbb{E}_{s \sim d_\rho^{(k)}} \left[D_{\mathrm{KL}} \left(\pi_s^\theta \;\middle\|\; \exp\left(-Q_{\tau,s}^{(k)} / \tau + \log \pi_s^{(k)} \right) / Z_s^{(k)} \right) \right] \\ &= \arg\min_\theta \mathbb{E}_{s \sim d_\rho^{(k)}} \left[D_{\mathrm{KL}} \left(\pi_s^\theta \;\middle\|\; \pi_s^{(k)} \exp\left(-Q_{\tau,s}^{(k)} / \tau \right) / Z_s^{(k)} \right) \right], \end{aligned}$$

which is exactly the same as the update rule in Eq. (6.27) if we take $\eta_k = \frac{1}{\tau}$ for any k. Therefore, we conclude that *SAC's update rule can be obtained by taking* $\eta_k = \frac{1}{\tau}$ *for any* k *in DAPO-KL*. As an immediate consequence, we can have a full convergence rate analysis for SAC, which will be presented in Section 6.5.2.

6.5 Convergence Analysis

In this section, we present the convergence analysis of DAPO-L_2 and DAPO-KL as well as SAC. These results are extensions of similar results for PMD method in the tabular case Xiao [2022] and with the log-linear policy class Yuan et al. [2022].

---

*The original proposition of SAC in Haarnoja et al. [2018a] uses $\exp(q_{\tau,s}^{(k)}/\tau)$ instead of $\exp(-q_{\tau,s}^{(k)}/\tau)$ because it considers reward maximization instead of cost minimization.

6.5.1 Analysis of DAPO-L_2

We make the following two assumptions regarding running Algorithm 10, with an initial state distribution $\rho \in \Delta(\mathcal{S})$.

(A1) There exist constants $\epsilon_{\text{critic}}, \epsilon_{\text{actor}} > 0$ such that for every iteration k, it holds

$$\begin{aligned} &\mathbb{E}_{s\sim d_\rho^{(k)}}\Big[\big\|\widehat{Q}_s^{(k)} - Q_s^{(k)}\big\|_\infty\Big] \le \epsilon_{\text{critic}}, \\ &\mathbb{E}_{s\sim d_\rho^{(k)}}\Big[\big\|f_s^{(k+1)} - \big(\pi_s^{(k)} + \eta_k \widehat{Q}_s^{(k)}\big)\big\|_2^2\Big] \le 2\eta_k^2 \epsilon_{\text{actor}}, \end{aligned} \tag{6.29}$$

(A2) There exists a constant $\vartheta_\rho \ge 1$ such that for any k, it holds

$$\max\left\{\left\|\frac{d_\rho^\star}{d_\rho^{(k+1)}}\right\|_\infty, \left\|\frac{d_\rho^{(k+1)}}{d_\rho^{(k)}}\right\|_\infty, \left\|\frac{d_{d_\rho^\star}^{(k+1)}}{d_\rho^{(k)}}\right\|_\infty, \left\|\frac{d_{d_\rho^\star}^{(k+1)}}{d_\rho^\star}\right\|_\infty\right\} \le \vartheta_\rho.$$

where $\pi^\star$ is some comparator policy.

Here, in (A1), we assume that $\widehat{Q}^{(k)}$ is a good enough approximation of $Q^{(k)}$, which is a problem that has been extensively studied both theoretically and empirically [Li and Lan, 2023, Chen et al., 2022a, Fujimoto et al., 2018]. We also assume that the parameterized function f^θ is powerful enough to approximate the dual vector $\nabla\Phi(\pi^{(k)}) - \eta_k \widehat{Q}^{(k)}$, which is common for studying function approximation [Alfano et al., 2024, Lan, 2022, Agarwal et al., 2021]. The scaling coefficient η_k^2 is consistent with Assumption (A1) in Alfano et al. [2024] as they assume the L_2-error for approximating $\eta_k^{-1}\nabla\Phi(\pi^{(k)}) - \widehat{Q}^{(k)}$ is bounded by ϵ_{critic}.

Then, (A2) assumes that the distribution mismatch coefficient is bounded, which is often needed for analyzing policy gradient methods [Xiao, 2022, Yuan et al., 2022] and can be satisfied if we take $\rho = \mathsf{Unif}(\mathcal{S})$ (see Lemma E.1.8 in Appendix E.1).

Under these assumptions, we have the following theorem.

Theorem 6.5.1 (Sublinear Convergence of DAPO-L_2)**.** *Consider Algorithm 10 with initial policy $\pi^{(0)}$, initial distribution $\rho \in \Delta(\mathcal{S})$ and Φ being the squared L_2-norm. Let $\pi^\star$ be an arbitrary comparator policy. Suppose Assumptions (A1) and (A2) hold and we have constant*

step size $\eta_k = \eta$ *for all* $k \geq 0$. *Then, it holds that*

$$\frac{1}{K}\sum_{k=0}^{K-1}\left(V_\rho^{(k)} - V_\rho^\star\right) \leq \frac{1}{K}\left(\frac{D_0^\star}{(1-\gamma)\eta} + \frac{V_{d_\rho^\star}^{(0)}}{1-\gamma}\right) + \frac{\vartheta_\rho\sqrt{2\epsilon_{\text{actor}}} + (2-\gamma)\vartheta_\rho\epsilon_{\text{critic}}}{(1-\gamma)^2},$$

where $D_0^\star = \mathbb{E}_{s\sim d_\rho^\star}\left[D_\Phi(\pi_s^\star, \pi_s^{(0)})\right]$.

We can see it achieves $O(1/K)$ convergence rate under constant step size. Meanwhile, similar to Xiao [2022], with more algebraic manipulations, it is possible to achieve linear convergence under geometrically increasing step sizes, as stated in the following theorem.

Theorem 6.5.2 (Linear Convergence of DAPO-L_2)**.** *Consider Algorithm 10 with initial policy* $\pi^{(0)}$, *initial distribution* $\rho \in \Delta(\mathcal{S})$ *and* Φ *being the squared* L_2*-norm. Let* $\pi^\star$ *be an arbitrary comparator policy. Suppose Assumptions (A1) and (A2) hold and the step sizes satisfy* $\eta_0 > 1$ *and* $\eta_{k+1} \geq (\vartheta_\rho/(\vartheta_\rho - 1))\,\eta_k$ *for all* $k \geq 0$. *Then, it holds that*

$$V_\rho^{(K)} - V_\rho^\star \leq \left(1 - \frac{1}{\vartheta_\rho}\right)^K \left(V_\rho^{(0)} - V_\rho^\star + \frac{D_0^\star/(\vartheta_\rho - 1)}{(1-\gamma)\eta_0}\right) + \frac{\vartheta_\rho^2\sqrt{2\epsilon_{\text{actor}}} + 2\vartheta_\rho^2\epsilon_{\text{critic}}}{1-\gamma}.$$

where $D_0^\star = \mathbb{E}_{s\sim d_\rho^\star}\left[D_\Phi(\pi_s^\star, \pi_s^{(0)})\right]$.

The proofs of Theorem 6.5.1 and 6.5.2 are given in Appendix E.1. It retains the convergence rate of Alfano et al. [2024] albeit with some different techniques. This is expected since in the L_2 case, DAPO-L_2 is the same as AMPO.

6.5.2 Analysis of DAPO-KL

The analysis of DAPO-KL requires a slightly modified assumption (A1') and an additional assumption (A3).

(A1') Under the same setting as (A1), we instead have

$$\mathbb{E}_{s\sim d_\rho^{(k)}}\left[D_{\text{KL}}\left(\pi_s^{(k+1)} \middle\| \pi_s^{(k)}\exp\left(-\eta_k\widehat{Q}_s^{(k)}\right)/Z_s^{(k)}\right)\right] \leq \eta_k\epsilon_{\text{actor}} \tag{6.30}$$

(A3) There exists constant $C_\rho > 0$ such that for any k,

$$\max_{s\in\mathrm{supp}\left(d_\rho^{(k)}\right)}\left\{\left\|\frac{\pi_s^\star}{\pi_s^{(k+1)}}\right\|_\infty,\ \left\|\frac{\pi_s^{(k)}}{\pi_s^{(k+1)}}\right\|_\infty\right\}\le C_\rho.$$

Notice that in (A1'), the upper bound of the approximation error becomes $\eta_k\epsilon_{\mathrm{actor}}$. This is the result of considering the growth rate of the approximation error in η_k for different Bregman divergences. Specifically, if we keep $f^{(k+1)}$, $f^{(k)}$ and $\widehat{Q}^{(k)}$ fixed, then the L_2-error in (A1) satisfies

$$\mathbb{E}_{s\sim d_\rho^{(k)}}\left[\left\|f_s^{(k+1)}-\left(\pi_s^{(k)}+\eta_k\widehat{Q}_s^{(k)}\right)\right\|_2^2\right]\propto\eta_k^2.$$

However, the KL-divergence in (A1') satisfies

$$\mathbb{E}_{s\sim d_\rho^{(k)}}\left[D_{\mathrm{KL}}\left(\pi_s^{(k+1)}\middle\|\,\pi_s^{(k)}\exp\left(-\eta_k\widehat{Q}_s^{(k)}\right)/Z_s^{(k)}\right)\right]\propto\eta_k.$$

Meanwhile, (A3) is an assumption on the policy evolution. It holds, for example, when we apply DAPO-KL with entropy regularization [Cayci et al., 2021, Cen et al., 2022b].

Under these assumptions, we have the following $O(1/K)$ convergence guarantee.

Theorem 6.5.3 (Sublinear Convergence of DAPO-KL)**.** *Consider Algorithm 10 with initial policy $\pi^{(0)}$, initial distribution $\rho\in\Delta(\mathcal{S})$ and Φ being the negative entropy restricted on $\Delta(\mathcal{A})$. Let $\pi^\star$ be an arbitrary comparator policy. Suppose Assumptions (A1'), (A2) and (A3) hold and we have constant step size $\eta_k=\eta$ for all $k\ge0$. Then, it holds that*

$$\frac{1}{K}\sum_{k=0}^{K-1}\left(V_\rho^{(k)}-V_\rho^\star\right)\le\frac{1}{K}\left(\frac{D_0^\star}{(1-\gamma)\eta}+\frac{V_{d_\rho^\star}^{(0)}}{1-\gamma}\right)+\frac{\vartheta_\rho\psi(\epsilon_{\mathrm{actor}})+(2-\gamma)\vartheta_\rho\epsilon_{\mathrm{critic}}}{(1-\gamma)^2},$$

where $D_0^\star=\mathbb{E}_{s\sim d_\rho^\star}\left[D_\Phi(\pi_s^\star,\pi_s^{(0)})\right]$ and $\psi(x)=(1+C_\rho)\left(x+\sqrt{2x}\right)$ for $x\ge0$.

Similarly, with geometrically increasing step sizes, we can obtain the following linear convergence rate.

Theorem 6.5.4 (Linear Convergence of DAPO-KL)**.** *Consider Algorithm 10 with initial policy $\pi^{(0)}$, initial distribution $\rho\in\Delta(\mathcal{S})$ and Φ being the negative entropy restricted on*

$\Delta(\mathcal{A})$. *Let* $\pi^\star$ *be an arbitrary comparator policy. Suppose Assumptions (A1'), (A2) and (A3) hold and the step sizes satisfy* $\eta_0 > 1$ *and* $\eta_{k+1} \geq (\vartheta_\rho/(\vartheta_\rho - 1))\,\eta_k$ *for all* $k \geq 0$. *Then, we have*

$$V_\rho^{(K)} - V_\rho^\star \leq \left(1 - \frac{1}{\vartheta_\rho}\right)^K \left(V_\rho^{(0)} - V_\rho^\star + \frac{D_0^\star/(\vartheta_\rho - 1)}{(1-\gamma)\eta_0}\right) + \frac{\vartheta_\rho^2 \psi(\epsilon_{\text{actor}}) + 2\vartheta_\rho^2 \epsilon_{\text{critic}}}{1-\gamma},$$

where $D_0^\star = \mathbb{E}_{s\sim d_\rho^\star}\left[D_\Phi(\pi_s^\star, \pi_s^{(0)})\right]$ *and* $\psi(x) = (1 + C_\rho)\left(x + \sqrt{2x}\right)$ *for* $x \geq 0$.

The proofs of Theorem 6.5.3 and 6.5.4 are given in Appendix E.1.

6.5.3 Analysis of SAC

Although we have shown that SAC is a special case of DAPO-KL in Section 6.4.3, its convergence analysis is slightly different, as it is formulated under entropy-regularized reinforcement learning. Specifically, the key difference in analysis lies in the following modified performance difference lemma.

Lemma 6.5.5 (Modified Performance Difference Lemma)**.** *For any two policies* $\pi, \tilde{\pi} : \mathcal{S} \mapsto \Delta(\mathcal{A})$, *initial distribution* $\rho \in \Delta(\mathcal{S})$ *and regularization strength* $\tau > 0$, *it holds that*

$$\begin{aligned} V_{\tau,\rho}^{\pi} - V_{\tau,\rho}^{\tilde{\pi}} &= \frac{1}{1-\gamma}\mathbb{E}_{s\sim d_\rho^\pi}\left[\left\langle Q_{\tau,s}^{\tilde{\pi}}, \pi_s - \tilde{\pi}_s\right\rangle + \tau D_{\mathrm{KL}}(\pi_s \| \tilde{\pi}_s)\right] \\ &= \frac{1}{1-\gamma}\mathbb{E}_{s\sim d_\rho^{\tilde{\pi}}}\left[\left\langle Q_{\tau,s}^{\pi}, \pi_s - \tilde{\pi}_s\right\rangle - \tau D_{\mathrm{KL}}(\tilde{\pi}_s \| \pi_s)\right]. \end{aligned}$$

The proof is given in Appendix E.2. Then, with Lemma 6.5.5, the convergence guarantee of DAPO-KL (and therefore SAC) is summarized in the following theorem.

Theorem 6.5.6 (Sublinear Convergence of SAC)**.** *Consider running Algorithm 10 for entropy-regularized reinforcement learning with initial policy* $\pi^{(0)}$, *regularization strength* τ, *initial distribution* $\rho \in \Delta(\mathcal{S})$ *and* Φ *being the negative entropy restricted on* $\Delta(\mathcal{A})$. *Let* $\pi^\star$ *be an arbitrary comparator policy. Suppose Assumptions (A1'), (A2) and (A3) hold and the step*

sizes satisfy $\eta_k = \eta \leq \frac{1}{\tau\vartheta_\rho}$ *for any* k. *Then, we have*

$$\frac{1}{K}\sum_{k=0}^{K-1}\left(V_{\tau,\rho}^{(k)} - V_{\tau,\rho}^{\star}\right) \leq \frac{1}{K}\left(\frac{D_0^\star}{(1-\gamma)\eta} + \frac{V_{\tau,d_\rho^\star}^{(0)}}{1-\gamma}\right) + \frac{\vartheta_\rho\psi(\epsilon_{\text{actor}}) + (2-\gamma)\vartheta_\rho\epsilon_{\text{critic}}}{(1-\gamma)^2}.$$

where $D_0^\star = \mathbb{E}_{s\sim d_\rho^\star}\left[D_{\text{KL}}\left(\pi_s^\star \,\middle\|\, \pi_s^{(0)}\right)\right]$ *and* $\psi(x) = (1+C_\rho)\left(x + \sqrt{2x}\right)$ *for* $x \geq 0$.

The full proof is given in Appendix E.2. To the best of our knowledge, this is the first convergence rate analysis of the SAC algorithm under general function approximation.

6.6 Extension to Continuous State-Action Space

In this section, we extend previous convergence analysis to MDPs with continuous state and action spaces. First, we briefly introduce the mathematical framework for DAPO in continuous-space MDPs.

6.6.1 Formulation in Continuous-Space MDPs

Notation. Given a compact set $\mathcal{A} \subset \mathbb{R}^n$, let

$$\mathscr{L}_1(\mathcal{A}) = \left\{ f : \mathcal{A} \mapsto \mathbb{R} \;\middle|\; \int_{\mathcal{A}} |f(a)|\,\mathrm{d}a < \infty \right\}$$

denote the vector space of absolutely integrable functions over $\mathcal{A}$. The set of probability measures over $\mathcal{A}$ is

$$\mathscr{P}(\mathcal{A}) = \left\{ f \in \mathscr{L}_1(\mathcal{A}) \;\middle|\; \int_{\mathcal{A}} f(a)\mathrm{d}a = 1 \text{ and } f(a) \geq 0 \right\}.$$

For simplicity, we always use Lebesgue measure as the reference measure for integration. Furthermore, for $f \in \mathscr{L}_1(\mathcal{A})$, we use $\|f\|_\infty \overset{\text{def}}{=} \inf\{\sup_{a\in\mathcal{A}}|g(a)| \mid g = f \text{ a.e.}\}$[†] to denote the *essential supreme* of f [Luenberger, 1997].

We consider an infinite-horizon discounted continuous-space MDP, denoted as the tuple $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathcal{P}, c, \gamma)$, where $\mathcal{S} \subset \mathbb{R}^{n_\mathcal{S}}$ and $\mathcal{A} \subset \mathbb{R}^{n_\mathcal{A}}$ are compact state, action spaces with

[†] $g = f$ a.e. means $\{a \in \mathcal{A} \mid g(a) \neq f(a)\}$ has zero measure.

dimensions $n_{\mathcal{S}}, n_{\mathcal{A}} > 0$, respectively, $\mathcal{P} : \mathcal{S} \times \mathcal{A} \mapsto \mathscr{P}(\mathcal{S})$ is the transition kernel, $c : \mathcal{S} \times \mathcal{A} \mapsto [0, 1]$ is the single-step cost function and γ is the discount factor.

A policy is defined as a measurable function $\pi : \mathcal{S} \times \mathcal{A} \mapsto \mathbb{R}$ such that $\pi_s \in \mathscr{P}(\mathcal{A})$ for any $s \in \mathcal{S}$. Since we assume that both $\mathcal{S}$ and $\mathcal{A}$ are bounded sets, we have $\pi \in \mathscr{L}_1(\mathcal{S} \times \mathcal{A})$. Then, the value function and Q-value function can be similarly defined as in Eq. (6.1) and (6.2), respectively. As a result, we can similarly have $V_s^\pi = \langle Q_s^\pi, \pi_s \rangle := \int_{\mathcal{A}} Q_{s,a}^\pi \pi_{s,a} \mathrm{d}a$.

With initial distribution $\rho \in \mathscr{P}(\mathcal{S})$, let $p_t^\pi(\cdot \mid \rho) \in \mathscr{P}(\mathcal{S})$ denote the probability density of s_t if the agent follows policy π and the initial state s_0 is sampled from ρ. Meanwhile, let $p_0^\pi(s \mid \rho) = \rho_s$ for any $s \in \mathcal{S}$. Then, we can similarly define the state-visitation density as $d_{\rho,s}^\pi = (1-\gamma) \sum_{t=0}^\infty \gamma^t p_t^\pi(s \mid \rho)$. Using the monotone convergence theorem, it is easy to verify that $\int_{s \in \mathcal{S}} d_{\rho,s}^\pi \mathrm{d}s = 1$ and thus $d_\rho^\pi \in \mathscr{P}(\mathcal{S})$. By truncating all terms with $t \geq 1$, we immediately have $d_{\rho,s}^\pi \geq (1-\gamma)\rho_s$, $\forall s \in \mathcal{S}$.

Although it may look straightforward to apply DAPO under these formulations, we will need extra efforts to make sure using DAPO is indeed well-posed for continuous-space MDPs, which will be addressed in the following section.

6.6.2 Well-Posedness of DAPO in Continuous-Space MDPs

In this section, we show that DAPO is a well-posed algorithm also for continuous-space MDPs from a mathematical perspective. In particular, we will prove a policy gradient theorem in continuous state-action space. Furthermore, we will show that requiring Φ to be a Legendre-type functional can make the mirror descent procedure well-defined in an infinite-dimensional normed vector space. As a clarification, during this section, for any $f, g \in \mathcal{X}$, where $\mathcal{X}$ is a normed vector space such as $\mathscr{L}_1(\mathcal{A})$, when we write $f = g$, we always mean "$f = g$ a.e.".

Gâteaux and Fréchet Differentials

We start by defining differentiability, specifically Gâteaux differential, in a normed vector space $\mathcal{X}$ with some general norm $\|\cdot\|$.[‡] which somehow resembles the directional derivative

[‡]The definition of Gâteaux differential itself does not require the vector space to be normed. However, we restrict our attention only on normed vector spaces for our needs. More details about this

in finite-dimensional space [Luenberger, 1997, Section 7].

Definition 6.6.1. Let $\Phi : \mathcal{C} \mapsto \mathbb{R}$ with $\mathcal{C} \subseteq \mathcal{X}$ be some functional and $h \in \mathcal{X}$ be arbitrary. Then, if the limit

$$\delta\Phi(x; h) = \lim_{\lambda \to 0} \frac{\Phi(x + \lambda h) - \Phi(x)}{\lambda}$$

exists, it is called the *Gâteaux differential* of Φ at x with direction h. If this limit exists for any $h \in \mathcal{X}$, then Φ is said to be *Gâteaux differentiable* at x.

Obviously, the above limit only makes sense for sufficiently small λ so that $x + \lambda h \in \mathcal{C}$.

Then, we introduce the Fréchet differential, which is a stronger notion of differentiability than the Gâteaux differential.

Definition 6.6.2. With the setup similar to the Gâteaux differential, if for any sequence $\{h_n\} \subset \mathcal{X}$ with $\lim_{n\to\infty} \|h_n\| = 0$, there exists $\delta\Phi(x; h_n) \in \mathbb{R}$ which is linear and continuous with respect to h_n such that

$$\lim_{n\to\infty} \frac{|\Phi(x + h_n) - \Phi(x) - \delta\Phi(x; h_n)|}{\|h_n\|} = 0,$$

then Φ is said to be *Fréchet differentiable* at x. Furthermore, since $\delta\Phi(x; h)$ is linear in h, it can be written as $\delta\Phi(x; h) = \langle \nabla\Phi(x), h \rangle$, where $\nabla\Phi(x)$ is called the *Fréchet derivative* of Φ at x.

This resembles the usual notion of differentiability in finite-dimensional space. Similarly, when we say a functional is differentiable, we by default means that it is Fréchet differentiable.

Finally, as a stronger notion of differentiability, Fréchet differential has the following basic property, which is presented in Luenberger [1997, Section 7].

Lemma 6.6.3. *If Φ is Fréchet differentiable at x, then it has unique Fréchet derivative $\nabla\Phi(x)$. Furthermore, it is also Gâteaux differentiable at x and its Gâteaux differential at x with direction h is exactly $\langle \nabla\Phi(x), h \rangle$.*

---

topic can be found in Luenberger [1997, Section 7].

Policy Gradient Theorem in Continuous-Space MDPs

To have a well-posed policy gradient theorem in continuous-space MDPs, we need the following additional assumptions:

(A4) There exists a constant $\alpha > 0$ such that $\sup_{(s,a)\in\mathcal{S}\times\mathcal{A}} \|\mathcal{P}(s,a)\|_\infty \leq \alpha$ and $\|\rho\|_\infty \leq \alpha$.

(A5) The Q-value function $Q^\pi \in \mathscr{L}_\infty(\mathcal{S}\times\mathcal{A})$ is continuous in π.

Here, Assumption (A4) basically prevents any positive transition probability mass over a zero-measure set, which helps bound the distribution mismatch coefficient when taking $\rho = \mathsf{Unif}(\mathcal{S})$ (see Lemma E.4.1); Assumption (A4) and (A5) ensure a well-defined policy gradient theorem. As a clarification, we do not allow point mass initial distribution here since it will make the state-visitation distribution pathelogical.

Then, with these additional assumptions, we can now prove the policy gradient theorem for continuous-space MDPs.

Theorem 6.6.4 (Continuous-Space Policy Gradient Theorem)**.** *Let $\rho \in \mathscr{P}(\mathcal{S})$ be some initial distribution. If Assumption (A4) and (A5) holds, then the value function V_ρ^π is Fréchet differentiable with respect to π and its Fréchet derivative is*

$$\nabla V_\rho^\pi = \frac{1}{1-\gamma} d_\rho^\pi Q^\pi \in \mathscr{L}_\infty\left(\mathcal{S}\times\mathcal{A}\right).$$

Proof. Let $\{h_n\} \subset \mathscr{L}_1(\mathcal{S}\times\mathcal{A})$ be a sequence such that $\lim_{n\to\infty}\|h_n\|_1 = 0$ and $\pi + h_n$ is a valid policy for each $n \in \mathbb{N}$. Then, by a continuous-version performance difference lemma (see Lemma E.3.1), we have

$$V_\rho^{\pi+h_n} - V_\rho^\pi = \frac{1}{1-\gamma}\mathbb{E}_{s\sim d_\rho^\pi}\left[\left\langle Q_s^{\pi+h_n}, \pi + h_n - \pi\right\rangle\right] = \left\langle \frac{1}{1-\gamma} d_\rho^\pi Q^{\pi+h_n}, h_n\right\rangle.$$

Using the definition of Fréchet differentiability, we have

$$\lim_{n\to\infty} \frac{\left|V_\rho^{\pi+h_n} - V_\rho^\pi - \frac{1}{1-\gamma}\left\langle d_\rho^\pi Q^\pi, h_n\right\rangle\right|}{\|h_n\|_1} = \frac{1}{1-\gamma}\lim_{n\to\infty}\frac{\left|\left\langle d_\rho^\pi(Q^{\pi+h_n} - Q^\pi), h_n\right\rangle\right|}{\|h_n\|_1}$$

$$\leq \frac{1}{1-\gamma} \lim_{n\to\infty} \left\| d_\rho^\pi (Q^{\pi+h_n} - Q^\pi) \right\|_\infty$$

(By Hölder's inequality.)

$$= 0$$ (By Assumption (A5).)

Now, we only remain to check that the linear functional $h \mapsto \frac{1}{1-\gamma} \left\langle d_\rho^\pi Q^\pi, h \right\rangle$ is continuous in $h \in \mathscr{L}_1(\mathcal{S} \times \mathcal{A})$. To see this, given arbitrary $\epsilon > 0$ and consider $h, h' \in \mathscr{L}_1(\mathcal{S} \times \mathcal{A})$ such that $\|h - h'\|_1 \leq \epsilon$. Then, we have

$$\frac{1}{1-\gamma} \left\langle d_\rho^\pi Q^\pi, h \right\rangle - \frac{1}{1-\gamma} \left\langle d_\rho^\pi Q^\pi, h' \right\rangle \leq \frac{1}{1-\gamma} \left\| d_\rho^\pi Q^\pi \right\|_\infty \|h - h'\|_1 \leq \frac{\epsilon}{1-\gamma} \left\| d_\rho^\pi Q^\pi \right\|_\infty .$$

Clearly, we have $\|Q^\pi\|_\infty \leq \frac{1}{1-\gamma}$ by our bounded cost function assumption. Then, by Lemma E.3.2 together with Assumption (A5), we have $\left\| d_\rho^\pi \right\|_\infty \leq \alpha$. That is, if $\|h - h'\|_1 \leq \epsilon$, we have

$$\frac{1}{1-\gamma} \left\langle d_\rho^\pi Q^\pi, h \right\rangle - \frac{1}{1-\gamma} \left\langle d_\rho^\pi Q^\pi, h' \right\rangle \leq \frac{\alpha\epsilon}{(1-\gamma)^2},$$

which shows the continuity of this linear function and thus the proof is complete. □

Functional Mirror Descent

In this section, we will prove that DAPO is well-posed in a normed vector space if the mirror map $\Phi : \mathcal{C} \mapsto \mathbb{R}$ is a Legendre-type or relaxed Legendre-type functional. First, we recall its definition.

Definition 6.6.5. A closed proper convex functional $\Phi : \mathcal{C} \mapsto \mathbb{R}$ is of *Legendre type*

(a) $\text{int}(\mathcal{C})$, the interior of $\mathcal{C}$, is non-empty.

(b) Φ is Fréchet differentiable and strictly convex on $\text{int}(\mathcal{C})$.

(c) For any sequence $\{x_n\}$ such that $x_n \in \text{int}(\mathcal{C})$ for all n and $\lim_{n\to\infty} x_n \in \text{bd } \mathcal{C}$ (the boundary of $\mathcal{C}$), it holds that $\lim_{n\to\infty} \|\nabla\Phi(x_n)\| = \infty$.

Before proceding, we need to define a generalization of Fréchet derivative, called *subgradient*. In particular, the subgradient of Φ at x is defined as

$$\partial\Phi(x) = \{x^* \in \mathcal{X}^* \mid \langle z - x, x^* \rangle + \Phi(x) \leq \Phi(z), \forall z \in \mathcal{C}\} \subseteq \mathcal{X}^*.$$

By the first-order condition of convex functional, we can immediately see that $\nabla\Phi(x) \in \partial\Phi(x)$ if Φ is Fréchet differentiable at x [Attouch et al., 2014, Section 9]. Furthermore, it has the following property (proof in Appendix E.3) similar to the finite-dimensional case.

Lemma 6.6.6. *If a proper closed convex functional $\Phi : \mathcal{C} \mapsto \mathbb{R}$ is Fréchet differentiable at $x \in \mathcal{C}$, then $\partial\Phi(x) = \{\nabla\Phi(x)\}$. That is, the subgradient is a singleton.*

Then, the following lemma (proof in Appendix E.3) shows that how strict convexity helps and its proof resembles the one in Rockafellar [1967].

Lemma 6.6.7. *Let $\Phi : \mathcal{C} \mapsto \mathbb{R}$ be a proper closed convex functional. If Φ is in addition also strictly convex over $\mathcal{C}$, then for any $x_1, x_2 \in \mathcal{C}$, we have $\partial\Phi(x_1) \cap \partial\Phi(x_2) = \emptyset$.*

Furthermore, the following lemma (proof in Appendix E.3) shows that a Legendre-type functional has no subgradient outside the interior of its domain.

Lemma 6.6.8. *Let $\Phi : \mathcal{C} \mapsto \mathbb{R}$ be a proper closed Legendre-type functional. Then, for any $x \notin \operatorname{int}(\mathcal{C})$, we have $\partial\Phi(x) = \emptyset$.*

Now, we are ready to prove the following key theorem about the Legendre-type functional in infinite-dimensional space.

Theorem 6.6.9. *If a proper closed convex functional $\Phi : \mathcal{C} \mapsto \mathbb{R}$ is also a Legendre-type functional. Then, for any $x \in \operatorname{int}(\mathcal{C})$, we have $\partial\Phi(x) = \{\nabla\Phi(x)\}$ and $\nabla\Phi^* = (\nabla\Phi)^{-1}$ is a well-defined operator..*

Proof. By Lemma 6.6.6 and 6.6.7, we know that for any $x \in \operatorname{int}(\mathcal{C})$, $\partial\Phi(x) = \{\nabla\Phi(x)\}$ and $\nabla\Phi(x_1) \neq \nabla\Phi(x_2)$ for any $x_1 \neq x_2$.

Furthermore, Attouch et al. [2014, Section 9] shows that $x^* \in \partial\Phi(x)$ is equivalent to $x \in \partial\Phi^*(x^*)$ for any $(x, x^*) \in \operatorname{int}(\Phi) \times \operatorname{dom}(\Phi^*)$. That is, for any $x \in \operatorname{int}(\mathcal{C})$, $\partial\Phi^*(\nabla\Phi(x)) \neq \emptyset$

and for any $x^* \in \mathrm{dom}(\Phi^*)$ such that $\partial\Phi^*(x^*) \neq \emptyset$, there exists unique $x \in \mathrm{int}(\mathcal{C})$ such that $x \in \partial\Phi^*(x^*)$.

Therefore, the only thing remained is to show that for any $x \notin \mathrm{int}(\mathcal{C})$, we have $\partial\Phi(x) = \emptyset$, which is given in Lemma 6.6.8 and the proof is complete. □

Finally, we can similarly have the following relaxation for Legendre-type functional.

Lemma 6.6.10. *Suppose $\Phi(x) = \phi(x) + \delta(x|\mathcal{L})$ where ϕ is a convex functional of Legendre type and $\mathcal{L} \subset \mathcal{X}$ is an affine subspace. Assume that $\mathrm{int}\,(\mathrm{dom}\,\phi) \cap \mathcal{L} \neq \emptyset$. Then we have*

$$\nabla\Phi^*(\nabla\Phi(x)) = x, \qquad \forall\, x \in \mathrm{int}\,(\mathrm{dom}\,\phi) \cap \mathcal{L}.$$

And for any $x^* \in \mathrm{int}\,(\mathrm{dom}\,\Phi^*)$ and any $x, y \in \mathrm{dom}\,\Phi$,

$$\big\langle \nabla\Phi(\nabla\Phi^*(x^*)), x - y \big\rangle = \langle x^*, x - y \rangle,$$

where $\nabla\Phi(x)$ denotes any subgradient in $\partial\Phi(x)$.

Proof. The proof is similar to Lemma 6.3.2. □

In summary, Theorem 6.6.4 together with Lemma 6.6.10 assures the well-posedness of DAPO in continuous-space MDPs.

6.6.3 Convergence Guarantee for DAPO-KL

We consider the negative differential entropy as the mirror map because it leads to the most widely-used algorithm. In the continuous domain, the negative differential is defined as $\Phi(\pi) = \int_{\mathcal{A}} \pi_a \log \pi_a \mathrm{d}a$ for $\pi \in \mathscr{P}(\mathcal{A}) \subset \mathscr{L}_1(\mathcal{A})$. Its conjugate function $\Phi^*(x^*) = \log\left(\int_{\mathcal{A}} \exp(x_a^*)\mathrm{d}a\right)$ for $x^* \in \mathscr{L}_\infty(\mathcal{A})$ (see Lemma E.4.2). Similarly, it has gradient $\nabla\Phi^*(x^*)(a) \propto \exp(x_a^*)$ and $\nabla\Phi^*(x^*) \in \mathscr{P}(\mathcal{A})$. These lead to the same policy update rule as (6.24) and thus we still call the resulting algorithm DAPO-KL. Then, we have the following theorem.

Theorem 6.6.11 (Linear Convergence for Continuous-space MDPs)**.** *Consider Algorithm 10 with initial policy $\pi^{(0)}$, initial distribution $\rho \in \mathscr{P}(\mathcal{S})$, arbitrary comparator policy $\pi^\star$ and*

Φ being the negative differential entropy. Suppose Assumptions (A1')-(A5) hold and the step sizes satisfy $\eta_0 > 1$ and $\eta_{k+1} \geq (\vartheta_\rho/(\vartheta_\rho - 1))\,\eta_k$ for all $k \geq 0$. Then, we have

$$V_\rho^{(K)} - V_\rho^\star \leq \left(1 - \frac{1}{\vartheta_\rho}\right)^K \left(V_\rho^{(0)} - V_\rho^\star + \frac{D_0^\star/(\vartheta_\rho - 1)}{(1-\gamma)\eta_0}\right) + \frac{\vartheta_\rho^2 \psi(\epsilon_{\text{actor}}) + (1+\vartheta_\rho)\epsilon_{\text{critic}}}{1-\gamma},$$

where $D_0^\star = \mathbb{E}_{s\sim d_\rho^\star}\left[D_\Phi(\pi_s^\star, \pi_s^{(0)})\right]$ and $\psi(x) = (1+C_\rho)\left(x + \sqrt{2x}\right)$ for $x \geq 0$.

It is not surprising that the convergence rate in Theorem 6.6.11 shares the same form as that in Theorem 6.5.4 because it does not explicitly depend on the size of the state-action space. Their differences are captured through the constants ϑ_ρ and C_ρ, which may depends on $|\mathcal{S}|$ and $|\mathcal{A}|$ in the discrete space and $\text{vol}(\mathcal{S})$ and $\text{vol}(\mathcal{A})$ in the continuous space.

Example 6.6.12 (Closed-form Update Rule under Linear Parameterization)**.** As an intriguing special case, we will show that DAPO-KL can be reduced to a closed-form update rule if we use a Gaussian policy with linear parameterization. Specifically, let $A \in \mathbb{R}^{n_\mathcal{A} \times n_\mathcal{S}}$ be a parameter matrix and define $f_{s,a}^A = -\frac{1}{2}\|As - a\|_2^2$, which means the policy is

$$\pi_{s,a}^A \propto \exp(f_{s,a}^A) = \exp\left(-\frac{1}{2}\|As - a\|_2^2\right),$$

which is $\mathcal{N}(As, I)$, a Gaussian distribution with mean $As \in \mathbb{R}^{n_\mathcal{A}}$ and identity covariance.

Now, suppose the Q-value is also approximated as a linear function, meaning to have $\widehat{Q}_{s,a}^B = a^\top B s$ for some parameter matrix $B \in \mathbb{R}^{n_\mathcal{A} \times n_\mathcal{S}}$. Under this setting, the policy optimization problem becomes

$$A^{(k+1)} \in \arg\min_A \mathbb{E}_{s\sim d_\rho^{(k)}}\left[D_{\text{KL}}\left(\pi_s^A \;\middle\|\; \pi_s^{A^{(k)}} \exp\left(-\eta_k \widehat{Q}_s^{B^{(k)}}\right)/Z_s^{(k)}\right)\right], \tag{6.31}$$

Then, we can notice that

$$\begin{aligned}
&-\frac{1}{2}\left\|\left(A^{(k)} - \eta_k B^{(k)}\right)s - a\right\|_2^2 \\
=&-\frac{1}{2}\left\|A^{(k)}s - a\right\|_2^2 + \eta_k\left(A^{(k)}s - a\right)^\top B^{(k)}s - \frac{1}{2}\eta_k^2\left\|B^{(k)}s\right\|_2^2 \\
=&-\frac{1}{2}\left\|A^{(k)}s - a\right\|_2^2 \underbrace{- \eta_k a^\top B^{(k)} s}_{=\eta_k \widehat{Q}_{s,a}^{B^{(k)}}} + \underbrace{\eta_k\left(A^{(k)}s\right)^\top B^{(k)}s - \frac{1}{2}\eta_k^2\left\|B^{(k)}s\right\|_2^2}_{\text{does not depend on } a},
\end{aligned}$$

which implies

$$\pi_{s,a}^{A^{(k)}-\eta_k B^{(k)}} \propto \pi_{s,a}^{(k)} \exp\left(-\eta_k \widehat{Q}_{s,a}^{B^{(k)}}\right).$$

Therefore, Eq. (6.31) can be solved exactly, which gives

$$A^{(k+1)} = A^{(k)} - \eta_k B^{(k)}.$$

6.7 Experiments

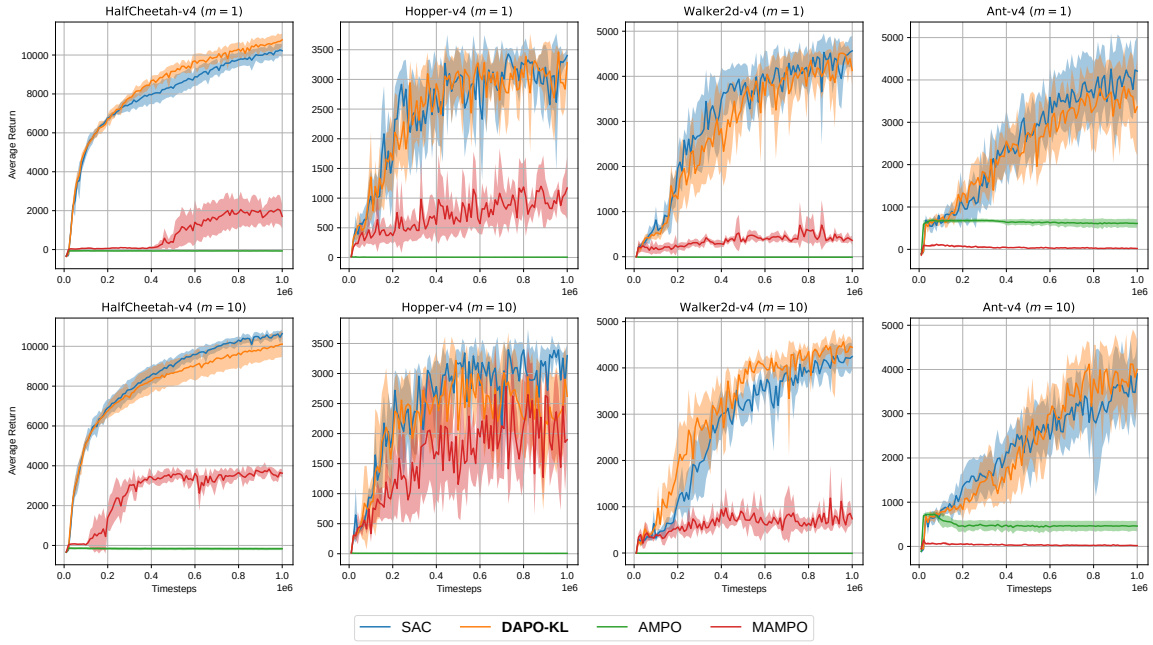


Figure 6.1: Average return curves on MuJoCo benchmarks. Each curve is averaged over 5 random seeds and the shaded area represents the 95% confidence interval. m represents the number of gradient steps in each policy update iteration.

In this section, we present our experiment results on several standard MuJoCo benchmark tasks [Todorov et al., 2012], which indeed have continuous state and action spaces. We compare the performance of DAPO-KL, SAC [Haarnoja et al., 2018b], and AMPO [Alfano et al., 2024].[§]

[§]Code repository is available at `https://github.com/FFTypeZero/DAPO`.

In addition, to demonstrate the importance of having primal-dual consistency, we modify AMPO to enforce this consistency albeit in a naive way. Specifically, as discussed in Section 6.4.2, AMPO's approximation loss can be expressed as (6.25) with Φ_1 being negative entropy and Φ_2 being squared L_2-norm. Therefore, a naive way to enforce compatibility is to replace Φ_1 by Φ_2 in (6.25), which gives us

$$\mathbb{E}_{s\sim d_\rho^{(k)}}\left[\left\|f_s^\theta - \left(\pi_s^{(k)} - \eta_k \widehat{Q}_s^{(k)}\right)\right\|_2^2\right]. \tag{6.32}$$

We call this algorithm Modified AMPO (MAMPO). Note that in MAMPO, Φ_1 (negative entropy) is still used in the policy projection step, which is different from DAPO-L_2.

Although in theory we assume that the policy optimization loss is approximately minimized in each iteration, in practice, it may be only feasible to run a few steps of the stochastic gradient method to reduce the loss. Therefore, the number of stochastic gradient steps per iteration can be an important hyper-parameter for the algorithm. In experiments, all algorithms are evaluated under both $m = 1$ and $m = 10$ stochastic gradient step per iteration. Full implementation details are given in Appendix E.5.

The results are summarized in Fig. C.2. From the plots, we can see that DAPO-KL performs about the same as and sometimes slightly better than SAC on all tasks, which is expected as we have shown that SAC is equivalent to a special case of DAPO-KL. Meanwhile, their performance is not sensitive to number of gradient steps per iteration.

On the other hand, AMPO fails to learn anything non-trivial on all tasks no matter it uses $m = 1$ or $m = 10$ gradient steps. Nevertheless, we retain the possibility that our implementation of AMPO may not be the optimal and provide more details of its hyperparameter tuning in Appendix E.6. In contrast, MAMPO is able to complete non-trivial learning among three tasks and gets better with more gradient steps, indicating the benefit of the primal-dual consistency in (6.32). However, it is still far inferior to DAPO-KL and SAC because it loses this consistency in its policy projection step.

6.8 Conclusions

In summary, DAPO is a mirror-descent based framework for incorporating general function approximation into policy optimization methods. Not only does it enjoy linear convergence rates, but it also naturally incorporates state-of-the-art practical algorithm like SAC as a special case and achieves comparable empirical performance. Furthermore, we give rigorous extension of its convergence analysis to Markov decision problems with continuous state and action spaces, which altogether bridges the theory and practice.

For future directions, DAPO paves the way for exploring new variants of PMD methods based on different mirror maps, e.g., with the negative Tsallis entropy. Another interesting question to investigate is how to characterize the effects of using inconsistent mirror maps in AMPO.

BIBLIOGRAPHY

Yasin Abbasi-Yadkori, Peter Bartlett, Victor Gabillon, Alan Malek, and Michal Valko. Best of both worlds: Stochastic & adversarial best-arm identification. In *Conference on Learning Theory*, pages 918–949. PMLR, 2018.

Marc Abeille and Alessandro Lazaric. Linear thompson sampling revisited. In *Artificial Intelligence and Statistics*, pages 176–184. PMLR, 2017.

Alekh Agarwal. Selective sampling algorithms for cost-sensitive multiclass prediction. In *International Conference on Machine Learning*, pages 1220–1228. PMLR, 2013.

Alekh Agarwal, Mikael Henaff, Sham Kakade, and Wen Sun. Pc-pg: Policy cover directed exploration for provable policy gradient learning. *Advances in neural information processing systems*, 33:13399–13412, 2020.

Alekh Agarwal, Sham M Kakade, Jason D Lee, and Gaurav Mahajan. On the theory of policy gradient methods: Optimality, approximation, and distribution shift. *J. Mach. Learn. Res.*, 22(98):1–76, 2021.

Priyank Agrawal, Jinglin Chen, and Nan Jiang. Improved worst-case regret bounds for randomized least-squares value iteration. *Thirty-fifth AAAI conference on artificial intelligence*, 2021.

Shipra Agrawal and Navin Goyal. Analysis of thompson sampling for the multi-armed bandit problem. In *Conference on learning theory*, pages 39–1. JMLR Workshop and Conference Proceedings, 2012.

Shipra Agrawal and Navin Goyal. Near-optimal regret bounds for thompson sampling. *J. ACM*, 64(5), September 2017. ISSN 0004-5411. doi: 10.1145/3088510. URL `https://doi.org/10.1145/3088510`.

Shipra Agrawal and Randy Jia. Optimistic posterior sampling for reinforcement learning: worst-case regret bounds. In *Advances in Neural Information Processing Systems*, pages 1184–1194, 2017a.

Shipra Agrawal and Randy Jia. Posterior sampling for reinforcement learning: worst-case regret bounds. In *Advances in Neural Information Processing Systems*, pages 1184–1194, 2017b.

Hayder AA Al-Kashoash, Maryam Hafeez, and Andrew H Kemp. Congestion control for 6lowpan networks: A game theoretic framework. *IEEE internet of things journal*, 4(3): 760–771, 2017.

Carlo Alfano, Rui Yuan, and Patrick Rebeschini. A novel framework for policy mirror descent with general parameterization and linear convergence. *Advances in Neural Information Processing Systems*, 36, 2024.

Hedy Attouch, Giuseppe Buttazzo, and Gérard Michaille. *Variational analysis in Sobolev and BV spaces: applications to PDEs and optimization*. SIAM, 2014.

Pierre-Cyril Aubin-Frankowski, Anna Korba, and Flavien Léger. Mirror descent with relative smoothness in measure spaces, with application to sinkhorn and em. *arXiv preprint arXiv:2206.08873*, 2022.

Jean-Yves Audibert, Sébastien Bubeck, and Rémi Munos. Best arm identification in multi-armed bandits. In *COLT*, pages 41–53, 2010.

Peter Auer and Chao-Kai Chiang. An algorithm with nearly optimal pseudo-regret for both stochastic and adversarial bandits. In *Conference on Learning Theory*, pages 116–120. PMLR, 2016.

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.

Peter Auer, Thomas Jaksch, and Ronald Ortner. Near-optimal regret bounds for reinforcement learning. *Advances in neural information processing systems*, 21, 2008.

Peter Auer, Pratik Gajane, and Ronald Ortner. Adaptively tracking the best bandit arm with an unknown number of distribution changes. In *Conference on Learning Theory*, pages 138–158. PMLR, 2019.

Mohammad Gheshlaghi Azar, Ian Osband, and Rémi Munos. Minimax regret bounds for reinforcement learning. In *International conference on machine learning*, pages 263–272. PMLR, 2017.

Mohammad Javad Azizi, Branislav Kveton, and Mohammad Ghavamzadeh. Fixed-budget best-arm identification in structured bandits. *arXiv preprint arXiv:2106.04763*, 2021.

Yu Bai and Chi Jin. Provable self-play algorithms for competitive reinforcement learning. In *International conference on machine learning*, pages 551–560. PMLR, 2020.

Peter L Bartlett and Shahar Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3(Nov):463–482, 2002.

Peter L Bartlett and Ambuj Tewari. Regal: A regularization based algorithm for reinforcement learning in weakly communicating mdps. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, pages 35–42. AUAI Press, 2009.

Amir Beck and Marc Teboulle. Mirror descent and nonlinear projected subgradient methods for convex optimization. *Operations Research Letters*, 31(3):167–175, 2003.

Amrit Singh Bedi, Souradip Chakraborty, Anjaly Parayil, Brian M Sadler, Pratap Tokekar, and Alec Koppel. On the hidden biases of policy mirror ascent in continuous action spaces. In *International Conference on Machine Learning*, pages 1716–1731. PMLR, 2022.

Dimitri Bertsekas. *Convex optimization theory*, volume 1. Athena Scientific, 2009.

Dimitri P Bertsekas. Dynamic programming and optimal control 4th edition, volume ii. *Athena Scientific*, 2015.

Sebastian Bervoets, Mario Bravo, and Mathieu Faure. Learning with minimal information in continuous games. *Theoretical Economics*, 15(4):1471–1508, 2020.

Alina Beygelzimer, John Langford, Lihong Li, Lev Reyzin, and Robert Schapire. Contextual bandit algorithms with supervised learning guarantees. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pages 19–26. JMLR Workshop and Conference Proceedings, 2011.

Jalaj Bhandari and Daniel Russo. On the linear convergence of policy gradient methods for finite mdps. In *International Conference on Artificial Intelligence and Statistics*, pages 2386–2394. PMLR, 2021.

Ronen I. Brafman and Moshe Tennenholtz. R-max - a general polynomial time algorithm for near-optimal reinforcement learning. *J. Mach. Learn. Res.*, 3(Oct):213–231, March 2003. ISSN 1532-4435.

Mario Bravo, David Leslie, and Panayotis Mertikopoulos. Bandit learning in concave n-person games. *Advances in Neural Information Processing Systems*, 31, 2018.

Haim Brezis and Haim Brézis. *Functional analysis, Sobolev spaces and partial differential equations*, volume 2. Springer, 2011.

Sébastien Bubeck. *Convex Optimization: Algorithms and Complexity*. Number 8:3-4 in Foundations and Trends in Machine Learning. now Publishers Inc., 2015.

Sébastien Bubeck and Che-Yu Liu. Prior-free and prior-dependent regret bounds for thompson sampling. In *2014 48th Annual Conference on Information Sciences and Systems (CISS)*, pages 1–9. IEEE, 2014.

Sébastien Bubeck and Aleksandrs Slivkins. The best of both worlds: Stochastic and adversarial bandits. In *Conference on Learning Theory*, pages 42–1. JMLR Workshop and Conference Proceedings, 2012.

Sébastien Bubeck, Nicolo Cesa-Bianchi, et al. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends® in Machine Learning*, 5(1):1–122, 2012.

Yuri Burda, Harrison Edwards, Amos Storkey, and Oleg Klimov. Exploration by random network distillation. *arXiv preprint arXiv:1810.12894*, 2018.

Qi Cai, Zhuoran Yang, Chi Jin, and Zhaoran Wang. Provably efficient exploration in policy optimization. *arXiv preprint arXiv:1912.05830*, 2019.

Qi Cai, Zhuoran Yang, Chi Jin, and Zhaoran Wang. Provably efficient exploration in policy optimization. In *International Conference on Machine Learning*, pages 1283–1294. PMLR, 2020.

Romain Camilleri, Julian Katz-Samuels, and Kevin Jamieson. High-dimensional experimental design and kernel bandits, 2021a.

Romain Camilleri, Zhihan Xiong, Maryam Fazel, Lalit Jain, and Kevin G Jamieson. Selective sampling for online best-arm identification. *Advances in Neural Information Processing Systems*, 34:11071–11082, 2021b.

Semih Cayci, Niao He, and Rayadurgam Srikant. Linear convergence of entropy-regularized natural policy gradient with linear function approximation. *arXiv preprint arXiv:2106.04096*, 2021.

Shicong Cen, Fan Chen, and Yuejie Chi. Independent natural policy gradient methods for potential games: Finite-time global convergence with entropy regularization. *arXiv preprint arXiv:2204.05466*, 2022a.

Shicong Cen, Chen Cheng, Yuxin Chen, Yuting Wei, and Yuejie Chi. Fast global convergence of natural policy gradient methods with entropy regularization. *Operations Research*, 70 (4):2563–2578, 2022b.

Nicolo Cesa-Bianchi, Claudio Gentile, and Francesco Orabona. Robust bounds for classification via selective sampling. In *Proceedings of the 26th annual international conference on machine learning*, pages 121–128, 2009.

Olivier Chapelle and Lihong Li. An empirical evaluation of thompson sampling. In *Advances in neural information processing systems*, pages 2249–2257, 2011.

Po-An Chen and Chi-Jen Lu. Playing congestion games with bandit feedbacks. In *AAMAS*, pages 1721–1722, 2015.

Po-An Chen and Chi-Jen Lu. Generalized mirror descents in congestion games. *Artificial Intelligence*, 241:217–243, 2016.

Xiaoyu Chen, Jiachen Hu, Lihong Li, and Liwei Wang. Efficient reinforcement learning in factored mdps with application to constrained rl. *arXiv preprint arXiv:2008.13319*, 2020.

Yifang Chen, Chung-Wei Lee, Haipeng Luo, and Chen-Yu Wei. A new algorithm for non-stationary contextual bandits: Efficient, optimal and parameter-free. In *Conference on Learning Theory*, pages 696–726. PMLR, 2019.

Yining Chen, Haipeng Luo, Tengyu Ma, and Chicheng Zhang. Active online learning with hidden shifting domains. In *International Conference on Artificial Intelligence and Statistics*, pages 2053–2061. PMLR, 2021.

Zaiwei Chen and Siva Theja Maguluri. Sample complexity of policy-based methods under off-policy sampling and linear function approximation. In *International Conference on Artificial Intelligence and Statistics*, pages 11195–11214. PMLR, 2022.

Zaiwei Chen, John Paul Clarke, and Siva Theja Maguluri. Target network and truncation overcome the deadly triad in q-learning. *arXiv preprint arXiv:2203.02628*, 2022a.

Zaiwei Chen, Sajad Khodadadian, and Siva Theja Maguluri. Finite-sample analysis of off-policy natural actor–critic with linear function approximation. *IEEE Control Systems Letters*, 6:2611–2616, 2022b.

Yun Kuen Cheung and Georgios Piliouras. Chaos, extremism and optimism: Volume analysis of learning in games. *Advances in Neural Information Processing Systems*, 33:9039–9049, 2020.

Casey Chu, Jose Blanchet, and Peter Glynn. Probability functional descent: A unifying perspective on gans, variational inference, and reinforcement learning. In *International Conference on Machine Learning*, pages 1213–1222. PMLR, 2019.

Roberto Cominetti, Emerson Melo, and Sylvain Sorin. A payoff-based learning procedure and its application to traffic games. *Games and Economic Behavior*, 70(1):71–83, 2010.

Pierre Coucheney, Bruno Gaujal, and Panayotis Mertikopoulos. Penalty-regulated dynamics and robust learning procedures in games. *Mathematics of Operations Research*, 40(3):611–633, 2015.

Qiwen Cui, Zhihan Xiong, Maryam Fazel, and Simon S Du. Learning in congestion games with bandit feedback. *Advances in Neural Information Processing Systems*, 35:11009–11022, 2022.

Christoph Dann and Emma Brunskill. Sample complexity of episodic fixed-horizon reinforcement learning. In *Advances in Neural Information Processing Systems*, pages 2818–2826, 2015.

Christoph Dann, Tor Lattimore, and Emma Brunskill. Unifying PAC and regret: Uniform PAC bounds for episodic reinforcement learning. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 5717–5727, Red Hook, NY, USA, 2017. Curran Associates Inc. ISBN 9781510860964.

Christoph Dann, Lihong Li, Wei Wei, and Emma Brunskill. Policy certificates: Towards accountable reinforcement learning. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 1507–1516, Long Beach, California, USA, 09–15 Jun 2019. PMLR.

S. Dasgupta, D. J. Hsu, and C. Monteleoni. A general agnostic active learning algorithm. *Advances in neural information processing systems*, 2008.

Constantinos Daskalakis. On the complexity of approximating a nash equilibrium. *ACM Transactions on Algorithms (TALG)*, 9(3):1–35, 2013.

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.

Ofer Dekel, Claudio Gentile, and Karthik Sridharan. Selective sampling and active learning from single and multiple teachers. *The Journal of Machine Learning Research*, 13(1): 2655–2697, 2012.

Dongsheng Ding, Chen-Yu Wei, Kaiqing Zhang, and Mihailo R. Jovanović. Independent policy gradient for large-scale markov potential games: Sharper rates, function approximation, and game-agnostic convergence, 2022.

Omar Darwiche Domingues, Pierre Ménard, Emilie Kaufmann, and Michal Valko. Episodic reinforcement learning in finite mdps: Minimax lower bounds revisited. In *Algorithmic Learning Theory*, pages 578–598. PMLR, 2021.

Kefan Dong, Yuanhao Wang, Xiaoyu Chen, and Liwei Wang. Q-learning with ucb exploration is sample efficient for infinite-horizon mdp. *arXiv preprint arXiv:1901.09311*, 2019.

Dmitriy Drusvyatskiy, Maryam Fazel, and Lillian J Ratliff. Improved rates for derivative free gradient play in strongly monotone games. In *Proc. IEEE Conference on Decision and Control*, 2022.

Stéphane Durand. *Analysis of Best Response Dynamics in Potential Games*. PhD thesis, Université Grenoble Alpes, 2018.

Reza Eghbali, James Saunderson, and Maryam Fazel. Competitive online algorithms for resource allocation over the positive semidefinite cone. *Mathematical Programming*, 170 (1):267–292, 2018.

Maryam Fazel, Rong Ge, Sham Kakade, and Mehran Mesbahi. Global convergence of policy gradient methods for the linear quadratic regulator. In *International conference on machine learning*, pages 1467–1476. PMLR, 2018.

Tanner Fiez, Lalit Jain, Kevin G Jamieson, and Lillian Ratliff. Sequential experimental design for transductive linear bandits. *Advances in neural information processing systems*, 32, 2019.

Meire Fortunato, Mohammad Gheshlaghi Azar, Bilal Piot, Jacob Menick, Matteo Hessel, Ian Osband, Alex Graves, Volodymyr Mnih, Remi Munos, Demis Hassabis, et al. Noisy networks for exploration. In *International Conference on Learning Representations*, 2018.

Dimitris Fotakis, Spyros Kontogiannis, Elias Koutsoupias, Marios Mavronicolas, and Paul Spirakis. The structure and complexity of nash equilibria for a selfish routing game. In *International Colloquium on Automata, Languages, and Programming*, pages 123–134. Springer, 2002.

Roy Fox, Stephen McAleer, Will Overman, and Ioannis Panageas. Independent natural policy gradient always converges in markov potential games. *arXiv preprint arXiv:2110.10614*, 2021.

David A Freedman. On tail probabilities for martingales. *the Annals of Probability*, pages 100–118, 1975.

Ronan Fruit, Matteo Pirotta, and Alessandro Lazaric. Near optimal exploration-exploitation in non-communicating markov decision processes. In *Advances in Neural Information Processing Systems*, pages 2994–3004, 2018.

Scott Fujimoto, Herke Hoof, and David Meger. Addressing function approximation error in actor-critic methods. In *International conference on machine learning*, pages 1587–1596. PMLR, 2018.

Aurélien Garivier and Eric Moulines. On upper-confidence bound policies for switching bandit problems. In *Algorithmic Learning Theory: 22nd International Conference, ALT 2011, Espoo, Finland, October 5-7, 2011. Proceedings 22*, pages 174–188. Springer, 2011.

Jakub Grudzien, Christian A Schroeder De Witt, and Jakob Foerster. Mirror learning: A unifying framework of policy optimisation. In *International Conference on Machine Learning*, pages 7825–7844. PMLR, 2022.

Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *International conference on machine learning*, pages 1861–1870. PMLR, 2018a.

Tuomas Haarnoja, Aurick Zhou, Kristian Hartikainen, George Tucker, Sehoon Ha, Jie Tan, Vikash Kumar, Henry Zhu, Abhishek Gupta, Pieter Abbeel, et al. Soft actor-critic algorithms and applications. *arXiv preprint arXiv:1812.05905*, 2018b.

Steve Hanneke and Liu Yang. Toward a general theory of online selective sampling: Trading off mistakes and queries. In *International Conference on Artificial Intelligence and Statistics*, pages 3997–4005. PMLR, 2021.

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

Amélie Heliou, Johanne Cohen, and Panayotis Mertikopoulos. Learning with bandit feedback in potential games. *Advances in Neural Information Processing Systems*, 30, 2017.

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.

Roger A Horn and Charles R Johnson. *Matrix analysis*. Cambridge university press, 2012.

Bin Hu, Kaiqing Zhang, Na Li, Mehran Mesbahi, Maryam Fazel, and Tamer Başar. Toward a theoretical foundation of policy optimization for learning control policies. *Annual Review of Control, Robotics, and Autonomous Systems*, 6(1):123–158, 2023.

Tzu-Kuo Huang, Alekh Agarwal, Daniel J Hsu, John Langford, and Robert E Schapire. Efficient and parsimonious agnostic active learning. *arXiv preprint arXiv:1506.08669*, 2015.

Christian Ibars, Monica Navarro, and Lorenza Giupponi. Distributed demand management in smart grid with a congestion game. In *2010 First IEEE International Conference on Smart Grid Communications*, pages 495–500. IEEE, 2010.

Thomas Jaksch, Ronald Ortner, and Peter Auer. Near-optimal regret bounds for reinforcement learning. *Journal of Machine Learning Research*, 11(Apr):1563–1600, 2010.

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

Chi Jin, Zeyuan Allen-Zhu, Sebastien Bubeck, and Michael I Jordan. Is q-learning provably efficient? *Advances in neural information processing systems*, 31, 2018.

Chi Jin, Zhuoran Yang, Zhaoran Wang, and Michael I Jordan. Provably efficient reinforcement learning with linear function approximation. *arXiv preprint arXiv:1907.05388*, 2019.

Chi Jin, Qinghua Liu, Yuanhao Wang, and Tiancheng Yu. V-learning–a simple, efficient, decentralized algorithm for multiagent rl. *arXiv preprint arXiv:2110.14555*, 2021a.

Chi Jin, Qinghua Liu, Yuanhao Wang, and Tiancheng Yu. V-learning – a simple, efficient, decentralized algorithm for multiagent rl, 2021b.

Tianyuan Jin, Pan Xu, Jieming Shi, Xiaokui Xiao, and Quanquan Gu. Mots: Minimax optimal thompson sampling. *arXiv preprint arXiv:2003.01803*, 2020.

Ramesh Johari and John N Tsitsiklis. Efficiency loss in a network resource allocation game. *Mathematics of Operations Research*, 29(3):407–435, 2004.

Emmeran Johnson, Ciara Pike-Burke, and Patrick Rebeschini. Optimal convergence rate for exact policy mirror descent in discounted markov decision processes. *arXiv preprint arXiv:2302.11381*, 2023.

Sham M Kakade. A natural policy gradient. *Advances in neural information processing systems*, 14, 2001.

Sham M Kakade. *On the sample complexity of reinforcement learning*. PhD thesis, University of London London, England, 2003.

Zohar Karnin, Tomer Koren, and Oren Somekh. Almost optimal exploration in multi-armed bandits. In *Proceedings of the 30th International Conference on International Conference on Machine Learning - Volume 28*, ICML'13, page III–1238–III–1246. JMLR.org, 2013.

Zohar S Karnin. Verification based solution for structured mab problems. *Advances in Neural Information Processing Systems*, 29, 2016.

Julian Katz-Samuels, Lalit Jain, Zohar Karnin, and Kevin Jamieson. 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.

Julian Katz-Samuels, Jifan Zhang, Lalit Jain, and Kevin Jamieson. Improved algorithms for agnostic pool-based active classification, 2021.

Emilie Kaufmann, Nathaniel Korda, and Rémi Munos. Thompson sampling: An asymptotically optimal finite-time analysis. In *International conference on algorithmic learning theory*, pages 199–213. Springer, 2012.

Emilie Kaufmann, Olivier Cappé, and Aurélien Garivier. On the complexity of best arm identification in multi-armed bandit models. *Journal of Machine Learning Research*, 17: 1–42, 2016.

Michael Kearns and Satinder Singh. Near-optimal reinforcement learning in polynomial time. *Machine learning*, 49(2-3):209–232, 2002.

Sajad Khodadadian, Prakirt Raj Jhunjhunwala, Sushil Mahavir Varma, and Siva Theja Maguluri. On the linear convergence of natural policy gradient algorithm. In *2021 60th IEEE Conference on Decision and Control (CDC)*, pages 3794–3799. IEEE, 2021.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Robert Kleinberg, Georgios Piliouras, and Éva Tardos. Multiplicative updates outperform generic no-regret learning in congestion games. In *Proceedings of the forty-first annual ACM symposium on Theory of computing*, pages 533–542, 2009.

Ron Kohavi and Roger Longbotham. Unexpected results in online controlled experiments. *ACM SIGKDD Explorations Newsletter*, 12(2):31–35, 2011.

J Zico Kolter and Andrew Y Ng. Near-bayesian exploration in polynomial time. In *Proceedings of the 26th annual international conference on machine learning*, pages 513–520, 2009.

Vijay Konda and John Tsitsiklis. Actor-critic algorithms. *Advances in neural information processing systems*, 12, 1999.

Walid Krichene, Benjamin Drighès, and Alexandre Bayen. On the convergence of no-regret learning in selfish routing. In *International Conference on Machine Learning*, pages 163–171. PMLR, 2014.

Walid Krichene, Benjamin Drighès, and Alexandre M Bayen. Online learning of nash equilibria in congestion games. *SIAM Journal on Control and Optimization*, 53(2):1056–1081, 2015.

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

Guanghui Lan. Policy optimization over general state and action spaces. *arXiv preprint arXiv:2211.16715*, 2022.

Guanghui Lan. Policy mirror descent for reinforcement learning: Linear convergence, new sampling complexity, and generalized problem classes. *Mathematical programming*, 198 (1):1059–1106, 2023.

Guanghui Lan, Yan Li, and Tuo Zhao. Block policy mirror descent. *SIAM Journal on Optimization*, 33(3):2341–2378, 2023.

Tor Lattimore and Marcus Hutter. Pac bounds for discounted mdps. In *International Conference on Algorithmic Learning Theory*, pages 320–334. Springer, 2012.

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

Chung-Wei Lee, Haipeng Luo, Chen-Yu Wei, Mengxiao Zhang, and Xiaojin Zhang. Achieving near instance-optimality and minimax-optimality in stochastic and adversarial linear bandits simultaneously. In *International Conference on Machine Learning*, pages 6142–6151. PMLR, 2021.

Jaeyoung Lee and Richard S Sutton. Policy iterations for reinforcement learning problems in continuous time and space—fundamental theory and methods. *Automatica*, 126:109421, 2021.

Stefanos Leonardos, Will Overman, Ioannis Panageas, and Georgios Piliouras. Global convergence of multi-agent policy gradient in markov potential games, 2021.

David S Leslie. *Reinforcement learning in games*. PhD thesis, University of Bristol, 2004.

David S Leslie and Edmund J Collins. Individual q-learning in normal form games. *SIAM Journal on Control and Optimization*, 44(2):495–514, 2005.

David S Leslie and Edmund J Collins. Generalised weakened fictitious play. *Games and Economic Behavior*, 56(2):285–298, 2006.

Gen Li, Laixi Shi, Yuxin Chen, Yuantao Gu, and Yuejie Chi. Breaking the sample complexity barrier to regret-optimal model-free reinforcement learning. *Advances in Neural Information Processing Systems*, 34, 2021.

Yan Li and Guanghui Lan. Policy mirror descent inherently explores action space. *arXiv preprint arXiv:2303.04386*, 2023.

Yan Li, Guanghui Lan, and Tuo Zhao. Homotopic policy mirror descent: Policy convergence, implicit regularization, and improved sample complexity. *arXiv preprint arXiv:2201.09457*, 2022.

Timothy P Lillicrap, Jonathan J Hunt, Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David Silver, and Daan Wierstra. Continuous control with deep reinforcement learning. *arXiv preprint arXiv:1509.02971*, 2015.

Boyi Liu, Qi Cai, Zhuoran Yang, and Zhaoran Wang. Neural proximal/trust region policy optimization attains globally optimal policy. *arXiv preprint arXiv:1906.10306*, 2019.

Qinghua Liu, Tiancheng Yu, Yu Bai, and Chi Jin. A sharp analysis of model-based reinforcement learning with self-play. In *International Conference on Machine Learning*, pages 7001–7010. PMLR, 2021.

David G Luenberger. *Optimization by vector space methods*. John Wiley & Sons, 1997.

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

Sergio Valcarcel Macua, Javier Zazo, and Santiago Zazo. Learning parametric closed-loop policies for markov potential games. *arXiv preprint arXiv:1802.00899*, 2018.

Jason R Marden. State based potential games. *Automatica*, 48(12):3075–3088, 2012.

Andreas Maurer and Massimiliano Pontil. Empirical Bernstein bounds and sample variance penalization. *arXiv preprint arXiv:0907.3740*, 2009.

Jincheng Mei, Bo Dai, Alekh Agarwal, Mohammad Ghavamzadeh, Csaba Szepesvári, and Dale Schuurmans. Ordering-based conditions for global convergence of policy gradient methods. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

Pierre Menard, Omar Darwiche Domingues, Xuedong Shang, and Michal Valko. Ucb momentum q-learning: Correcting the bias without forgetting. *arXiv preprint arXiv:2103.01312*, 2021.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin Riedmiller. Playing atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602*, 2013.

Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of machine learning*. MIT press, 2018.

Dov Monderer and Lloyd S Shapley. Potential games. *Games and economic behavior*, 14(1): 124–143, 1996.

Rémi Munos. Policy gradient in continuous time. *Journal of Machine Learning Research*, 7: 771–791, 2006.

Ofir Nachum and Bo Dai. Reinforcement learning via fenchel-rockafellar duality. *arXiv preprint arXiv:2001.01866*, 2020.

A Nemirovski, A Juditsky, G Lan, and A Shapiro. Stochastic approximation approach to stochastic programming. In *SIAM J. Optim.* Citeseer.

Arkadi Semenovič Nemirovski and David Borisovich Yudin. Problem complexity and method efficiency in optimization. 1983.

Gergely Neu and Ciara Pike-Burke. A unifying view of optimism in episodic reinforcement learning. *arXiv preprint arXiv:2007.01891*, 2020.

Gergely Neu, Anders Jonsson, and Vicenç Gómez. A unified view of entropy-regularized markov decision processes. *arXiv preprint arXiv:1705.07798*, 2017.

Hukukane Nikaidô and Kazuo Isoda. Note on non-cooperative convex games. *Pacific Journal of Mathematics*, 5(S1):807–815, 1955.

Optimizely. Stats accelerator – acceleration under time-varying signals. `https://support.optimizely.com/hc/en-us/articles/5326213705101-Stats-Accelerator-Acceleration-Under-Time-Varying-Signals`, May 2023.

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

Ian Osband and Benjamin Van Roy. Near-optimal reinforcement learning in factored mdps. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014.

Ian Osband and Benjamin Van Roy. Why is posterior sampling better than optimism for reinforcement learning? In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 2701–2710. JMLR. org, 2017.

Ian Osband, Daniel Russo, and Benjamin Van Roy. (more) efficient reinforcement learning via posterior sampling. In *Advances in Neural Information Processing Systems*, pages 3003–3011, 2013.

Ian Osband, Benjamin Van Roy, and Zheng Wen. Generalization and exploration via randomized value functions. *arXiv preprint arXiv:1402.0635*, 2014.

Ian Osband, Daniel Russo, Zheng Wen, and Benjamin Van Roy. Deep exploration via randomized value functions. *arXiv preprint arXiv:1703.07608*, 2017.

Ian Osband, John Aslanides, and Albin Cassirer. Randomized prior functions for deep reinforcement learning. In *Advances in Neural Information Processing Systems*, pages 8617–8629, 2018.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

Aldo Pacchiano, Philip Ball, Jack Parker-Holder, Krzysztof Choromanski, and Stephen Roberts. On optimism in model-based reinforcement learning. *arXiv preprint arXiv:2006.11911*, 2020.

Matteo Pirotta, Marcello Restelli, and Luca Bascetta. Policy gradient in lipschitz markov decision processes. *Machine Learning*, 100(2):255–283, 2015.

Martin L Puterman. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 1994.

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

Antonin Raffin, Ashley Hill, Adam Gleave, Anssi Kanervisto, Maximilian Ernestus, and Noah Dormann. Stable-baselines3: Reliable reinforcement learning implementations. *The Journal of Machine Learning Research*, 22(1):12348–12355, 2021.

Arvind Raghunathan, Anoop Cherian, and Devesh Jha. Game theoretic optimization via gradient-based nikaido-isoda function. In *International Conference on Machine Learning*, pages 5291–5300. PMLR, 2019.

Piyush Rai, Avishek Saha, Hal Daumé III, and Suresh Venkatasubramanian. Domain adaptation meets active learning. In *Proceedings of the NAACL HLT 2010 Workshop on Active Learning for Natural Language Processing*, pages 27–32, 2010.

R. Tyrrell Rockafellar. *Convex Analysis*. Princeton University Press, 1970.

Ralph Tyrrell Rockafellar. Conjugates and legendre transforms of convex functions. *Canadian Journal of Mathematics*, 19:200–205, 1967.

Aviv Rosenberg and Yishay Mansour. Oracle-efficient regret minimization in factored mdps with unknown structure. *Advances in Neural Information Processing Systems*, 34, 2021.

Robert W Rosenthal. A class of games possessing pure-strategy nash equilibria. *International Journal of Game Theory*, 2(1):65–67, 1973.

Tim Roughgarden. Algorithmic game theory. *Communications of the ACM*, 53(7):78–86, 2010.

Tim Roughgarden and Éva Tardos. Bounding the inefficiency of equilibria in nonatomic congestion games. *Games and economic behavior*, 47(2):389–403, 2004.

Aviad Rubinstein. Settling the complexity of computing approximate two-player nash equilibria. In *2016 IEEE 57th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 258–265. IEEE, 2016.

Daniel Russo. Worst-case regret bounds for exploration via randomized value functions. In *Advances in Neural Information Processing Systems*, pages 14433–14443, 2019.

Avishek Saha, Piyush Rai, Hal Daumé, Suresh Venkatasubramanian, and Scott L DuVall. Active supervised domain adaptation. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 97–112. Springer, 2011.

John Schulman, Sergey Levine, Pieter Abbeel, Michael Jordan, and Philipp Moritz. Trust region policy optimization. In *International conference on machine learning*, pages 1889–1897. PMLR, 2015.

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

Yevgeny Seldin and Gábor Lugosi. An improved parametrization and analysis of the exp3++ algorithm for stochastic and adversarial bandits. In *Conference on Learning Theory*, pages 1743–1759. PMLR, 2017.

Yevgeny Seldin and Aleksandrs Slivkins. One practical algorithm for both stochastic and adversarial bandits. In *International Conference on Machine Learning*, pages 1287–1295. PMLR, 2014.

Lior Shani, Yonathan Efroni, and Shie Mannor. Adaptive trust region policy optimization: Global convergence and faster rates for regularized mdps. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 5668–5675, 2020a.

Lior Shani, Yonathan Efroni, Aviv Rosenberg, and Shie Mannor. Optimistic policy optimization with bandit feedback. In *International Conference on Machine Learning*, pages 8604–8613. PMLR, 2020b.

Lloyd S Shapley. Stochastic games. *Proceedings of the national academy of sciences*, 39(10): 1095–1100, 1953.

Max Simchowitz and Kevin G Jamieson. Non-asymptotic gap-dependent regret bounds for tabular mdps. In *Advances in Neural Information Processing Systems*, pages 1153–1162, 2019.

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

Ziang Song, Song Mei, and Yu Bai. When can we learn general-sum markov games with a large number of players sample-efficiently? *arXiv preprint arXiv:2110.04184*, 2021.

Alexander L Strehl and Michael L Littman. An analysis of model-based interval estimation for markov decision processes. *Journal of Computer and System Sciences*, 74(8):1309–1331, 2008.

Alexander L Strehl, Lihong Li, Eric Wiewiora, John Langford, and Michael L Littman. Pac model-free reinforcement learning. In *Proceedings of the 23rd international conference on Machine learning*, pages 881–888. ACM, 2006.

Joe Suk and Samory Kpotufe. Tracking most significant arm switches in bandits, 2022.

Richard S Sutton, David McAllester, Satinder Singh, and Yishay Mansour. Policy gradient methods for reinforcement learning with function approximation. *Advances in neural information processing systems*, 12, 1999.

Brian Swenson, Ryan Murray, and Soummya Kar. On best-response dynamics in potential games. *SIAM Journal on Control and Optimization*, 56(4):2734–2767, 2018.

István Szita and András Lőrincz. Optimistic initialization and greediness lead to polynomial time learning in factored mdps. In *Proceedings of the 26th annual international conference on machine learning*, pages 1001–1008, 2009.

István Szita and Csaba Szepesvári. Model-based reinforcement learning with nearly tight exploration complexity bounds. In *ICML*, 2010.

Mohammad Sadegh Talebi and Odalric-Ambrym Maillard. Variance-aware regret bounds for undiscounted reinforcement learning in mdps. *arXiv preprint arXiv:1803.01626*, 2018.

Tian Tan, Zhihan Xiong, and Vikranth R Dwaracherla. Parameterized indexed value function for efficient exploration in reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 5948–5955, 2020.

Yi Tian, Jian Qian, and Suvrit Sra. Towards minimax optimal reinforcement learning in factored markov decision processes. *Advances in Neural Information Processing Systems*, 33:19896–19907, 2020.

Emanuel Todorov, Tom Erez, and Yuval Tassa. Mujoco: A physics engine for model-based control. In *2012 IEEE/RSJ international conference on intelligent robots and systems*, pages 5026–5033. IEEE, 2012.

Manan Tomar, Lior Shani, Yonathan Efroni, and Mohammad Ghavamzadeh. Mirror descent policy optimization. *arXiv preprint arXiv:2005.09814*, 2020.

Sharan Vaswani, Abbas Mehrabian, Audrey Durand, and Branislav Kveton. Old dog learns new tricks: Randomized ucb for bandit problems. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, pages 1988–1998, 2020.

Sharan Vaswani, Olivier Bachem, Simone Totaro, Robert Müller, Shivam Garg, Matthieu Geist, Marlos C Machado, Pablo Samuel Castro, and Nicolas Le Roux. A general class of surrogate functions for stable and efficient reinforcement learning. *arXiv preprint arXiv:2108.05828*, 2021.

Sharan Vaswani, Amirreza Kazemi, Reza Babanezhad Harikandeh, and Nicolas Le Roux. Decision-aware actor-critic with function approximation and theoretical guarantees. *Advances in Neural Information Processing Systems*, 36, 2024.

Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices, 2011.

Nino Vieillard, Tadashi Kozuno, Bruno Scherrer, Olivier Pietquin, Rémi Munos, and Matthieu Geist. Leverage the average: an analysis of kl regularization in rl. *arXiv preprint arXiv:2003.14089*, 2020.

Andrew Wagenmaker and Dylan J Foster. Instance-optimality in interactive decision making: Toward a non-asymptotic theory. *arXiv preprint arXiv:2304.12466*, 2023.

Lingxiao Wang, Qi Cai, Zhuoran Yang, and Zhaoran Wang. Neural policy gradient methods: Global optimality and rates of convergence. *arXiv preprint arXiv:1909.01150*, 2019.

Ruosong Wang, Simon S Du, Lin F Yang, and Sham M Kakade. Is long horizon reinforcement learning more difficult than short horizon reinforcement learning? *arXiv preprint arXiv:2005.00527*, 2020.

Christopher John Cornish Hellaby Watkins. Learning from delayed rewards. 1989.

Chen-Yu Wei and Haipeng Luo. Non-stationary reinforcement learning without prior knowledge: An optimal black-box approach. In *Conference on Learning Theory*, pages 4300–4354. PMLR, 2021.

Chen-Yu Wei, Haipeng Luo, and Alekh Agarwal. Taking a hint: How to leverage loss predictors in contextual bandits? In *Conference on Learning Theory*, pages 3583–3634. PMLR, 2020.

Ronald J Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8:229–256, 1992.

Yuhang Wu, Zeyu Zheng, Guangyu Zhang, Zuohua Zhang, and Chu Wang. Non-stationary a/b tests. 2022.

Lin Xiao. On the convergence rates of policy gradient methods. *Journal of Machine Learning Research*, 23(282):1–36, 2022.

Min Xiao and Yuhong Guo. Online active learning for cost sensitive domain adaptation, 2013.

Zhihan Xiong, Ruoqi Shen, Qiwen Cui, Maryam Fazel, and Simon S Du. Near-optimal randomized exploration for tabular markov decision processes. *Advances in neural information processing systems*, 35:6358–6371, 2022.

Zhihan Xiong, Romain Camilleri, Maryam Fazel, Lalit Jain, and Kevin Jamieson. A/b testing and best-arm identification for linear bandits with robustness to non-stationarity. In *International Conference on Artificial Intelligence and Statistics*, pages 1585–1593. PMLR, 2024a.

Zhihan Xiong, Maryam Fazel, and Lin Xiao. Dual approximation policy optimization. *arXiv preprint arXiv:2410.01249*, 2024b.

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.

Tengyu Xu, Zhe Wang, and Yingbin Liang. Improving sample complexity bounds for (natural) actor-critic algorithms. *Advances in Neural Information Processing Systems*, 33: 4358–4369, 2020.

Ziping Xu and Ambuj Tewari. Worst-case regret bound for perturbation based exploration in reinforcement learning. *Ann Arbor*, 1001:48109, 2019.

Ziping Xu and Ambuj Tewari. Reinforcement learning in factored mdps: Oracle-efficient algorithms and tighter regret bounds for the non-episodic setting. *Advances in Neural Information Processing Systems*, 33:18226–18236, 2020.

Yahoo! Yahoo! webscope dataset ydata-frontpage-todaymodule-clicks-v1_0, 2011. URL `https://webscope.sandbox.yahoo.com/catalog.php?datatype=r`.

Junwen Yang and Vincent Tan. Minimax optimal fixed-budget best arm identification in linear bandits. *Advances in Neural Information Processing Systems*, 35:12253–12266, 2022.

Junwen Yang and Vincent YF Tan. Towards minimax optimal best arm identification in linear bandits. *arXiv e-prints*, pages arXiv–2105, 2021.

Kunhe Yang, Lin F Yang, and Simon S Du. Q-learning with logarithmic regret. *arXiv preprint arXiv:2006.09118*, 2020.

Rui Yuan, Simon S Du, Robert M Gower, Alessandro Lazaric, and Lin Xiao. Linear convergence of natural policy gradient methods with log-linear policies. *arXiv preprint arXiv:2210.01400*, 2022.

Andrea Zanette and Emma Brunskill. Tighter problem-dependent regret bounds in reinforcement learning without domain knowledge using value function bounds. In *International Conference on Machine Learning*, pages 7304–7312, 2019.

Andrea Zanette, David Brandfonbrener, Emma Brunskill, Matteo Pirotta, and Alessandro Lazaric. Frequentist regret bounds for randomized least-squares value iteration. In *International Conference on Artificial Intelligence and Statistics*, pages 1954–1964, 2020.

Andrea Zanette, Ching-An Cheng, and Alekh Agarwal. Cautiously optimistic policy optimization and exploration with linear function approximation. In *Conference on Learning Theory*, pages 4473–4525. PMLR, 2021.

Wenhao Zhan, Shicong Cen, Baihe Huang, Yuxin Chen, Jason D Lee, and Yuejie Chi. Policy mirror descent for regularized reinforcement learning: A generalized framework with linear convergence. *SIAM Journal on Optimization*, 33(2):1061–1091, 2023.

Kaiqing Zhang, Zhuoran Yang, and Tamer Basar. Multi-agent reinforcement learning: A selective overview of theories and algorithms. *Handbook of Reinforcement Learning and Control*, pages 321–384, 2021a.

Mingrui Zhang, Zebang Shen, Aryan Mokhtari, Hamed Hassani, and Amin Karbasi. One sample stochastic frank-wolfe. In *International Conference on Artificial Intelligence and Statistics*, pages 4012–4023. PMLR, 2020a.

Runyu Zhang, Zhaolin Ren, and Na Li. Gradient play in stochastic games: stationary points, convergence, and sample complexity. *arXiv preprint arXiv:2106.00198*, 2021b.

Zihan Zhang and Xiangyang Ji. Regret minimization for reinforcement learning by evaluating the optimal bias function. In *Advances in Neural Information Processing Systems*, pages 2823–2832, 2019.

Zihan Zhang, Xiangyang Ji, and Simon S Du. Is reinforcement learning more difficult than bandits? a near-optimal algorithm escaping the curse of horizon. *arXiv preprint arXiv:2009.13503*, 2020b.

Zihan Zhang, Yuan Zhou, and Xiangyang Ji. Almost optimal model-free reinforcement learning via reference-advantage decomposition. *arXiv preprint arXiv:2004.10019*, 2020c.

Zihan Zhang, Yuan Zhou, and Xiangyang Ji. Model-free reinforcement learning: from clipped pseudo-regret to sample complexity. *arXiv preprint arXiv:2006.03864*, 2020d.

Hanyang Zhao, Wenpin Tang, and David D Yao. Policy optimization for continuous reinforcement learning. *arXiv preprint arXiv:2305.18901*, 2023.

Yang Zheng, Yue Sun, Maryam Fazel, and Na Li. Escaping high-order saddles in policy optimization for Linear Quadratic Gaussian (LQG) control. In *Proc. IEEE Conf. on Decision and Control (CDC)*, 2022.

Part III

DEFERRED CONTENTS FROM THE MAIN BODY

Appendix A

OMITTED PROOFS AND EXPERIMENT DETAILS IN CHAPTER 2

A.1 Additional Algorithms in Implementation

A.1.1 A Peace-based Robust Algorithm

In this section, we briefly explain how we design P1-Peace based on intuition similar to P1-RAGE and make it computationally efficient. First, we propose another subroutine, called Peace-Elimination, based on the elimination strategy in Peace Katz-Samuels et al. [2020], which has the same spirit as RAGE. Similar to RAGE-Elimination, Peace-Elimination also repeatedly computes $\mathcal{XY}$-allocation, but (virtually) eliminate arms so that the value of the remaining arms' optimal $\mathcal{XY}$-design is halved. In addition, in P1-Peace, we only update the sampling distribution λ_t after a period of time. The intuition is that if the environment is stationary, then we do not need to update our allocation probability frequently just like RAGE and Peace; if the environment is non-stationary, then the non-stationarity is handled by the mixed G-optimal design λ^*, which is fixed from the very beginning. Therefore, updating λ_t in a low frequency should not severely harm the performance. The new algorithm and elimination subroutine are summarized in Algorithm 11 and 12.

For convenience of presentation, for arm set $\mathcal{Z} \subset \mathbb{R}^d$ and distribution $\lambda \in \triangle_{\mathcal{X}}$, we define

$$\rho(\mathcal{Z}, \lambda) = \max_{x, x' \in \mathcal{Z}} \left\| x - x' \right\|^2_{A(\lambda)^{-1}} . \tag{A.1}$$

A.1.2 A Naive Baseline Mixed Algorithm

In this section, we present a naive mixture of Peace and the G-optimal design, called Mixed-Peace, which eliminates arms and computes design λ_k during each epoch exactly the same as Peace. The only differences are that Mixed-Peace uses IPS estimator and when pulling an arm, it will pull an arm by following $x_t \sim (\lambda_k + \lambda^*)/2$, where λ^* is the G-optimal design

Algorithm 11 P1-Peace

Input: budget, $T \in \mathbb{N}$; arm set $\mathcal{X} \subset \mathbb{R}^d$
Compute epoch length $R \leftarrow \left\lfloor \frac{T}{\log_2(\inf_{\lambda \in \triangle_{\mathcal{X}}} \rho(\mathcal{X}, \lambda))} \right\rfloor$
Compute G-optimal design λ^* based on equation (2.3) and initialize $\lambda_1 = \lambda^*$
for $t = 1, 2, \ldots, T$ **do**
  Sample $x_t \sim \lambda_t$ and receive reward r_t
  Estimate $\widehat{\theta}_t \leftarrow \frac{1}{t} \sum_{s=1}^{t} \mathbb{E}_{x \sim \lambda_s} \left[x x^\top \right]^{-1} x_s r_s$
  $\lambda_{t+1} \leftarrow \lambda_t$
  if $t - 1 = cR$ for some integer c **then**
    Update $\lambda_{t+1} \leftarrow$ Peace-Elimination$(\widehat{\theta}_t)$
  end if
end for
return $\arg\max_{x \in \mathcal{X}} x^\top \widehat{\theta}_T$

defined in equation (2.3). Its details are summarized in Algorithm 13.

A.2 Error Probability of Algorithm 1 In Non-Stationary Environments

Theorem 2.4.1 (Error probability of G-BAI)**.** *Fix time horizon T, arm set $\mathcal{X} \subset \mathbb{R}^d$ with $|\mathcal{X}| = K$ and arbitrary unknown parameters $\{\theta_t\}_{t=1}^T$. If we run Algorithm 1 in this non-stationary environment and obtain x_{J_T}, then it holds that*

$$\mathbb{P}_{\overline{\theta}_T}\left(J_T \neq (1)\right) \leq K \exp\left(-\frac{T}{12 H_{\text{G-BAI}}\left(\overline{\theta}_T\right)}\right), \quad \text{where } H_{\text{G-BAI}}\left(\overline{\theta}_T\right) = \frac{d}{\Delta_{(1)}^2}.$$

Proof. Based on the recommendation rule $x_{J_T} = \arg\max_{x \in \mathcal{X}} x^\top \widehat{\theta}_T$, we have

$$\begin{aligned}
\mathbb{P}\left(J_T \neq (1)\right) =& \mathbb{P}\left(\exists k \in [2:K] \text{ s.t. } x_{(k)}^\top \widehat{\theta}_T \geq x_{(1)}^\top \widehat{\theta}_T\right) \\
\leq& \mathbb{P}\left(\exists k \in [2:K] \text{ s.t. } x_{(k)}^\top \widehat{\theta}_T - x_{(k)}^\top \overline{\theta}_T \geq \frac{\Delta_{(k)}}{2} \text{ or } x_{(1)}^\top \widehat{\theta}_T - x_{(1)}^\top \overline{\theta}_T \leq -\frac{\Delta_{(1)}}{2}\right) \\
\leq& \mathbb{P}\left(x_{(1)}^\top \widehat{\theta}_T - x_{(1)}^\top \overline{\theta}_T \leq -\frac{\Delta_{(1)}}{2}\right) + \sum_{k=2}^{K} \mathbb{P}\left(x_{(k)}^\top \widehat{\theta}_T - x_{(k)}^\top \overline{\theta}_T \geq \frac{\Delta_{(k)}}{2}\right). \quad \text{(A.2)}
\end{aligned}$$

The above terms can be bounded by Bernstein's inequality. In particular, for the first term,

Algorithm 12 Peace-Elimination

Input: arm set $\mathcal{X} \subset \mathbb{R}^d$; current estimate $\widehat{\theta}_t$
Find index $\widehat{(k)}_t$ such that $x_{\widehat{(1)}_t}^\top \widehat{\theta}_t \geq x_{\widehat{(2)}_t}^\top \widehat{\theta}_t \geq \cdots \geq x_{\widehat{(K)}_t}^\top \widehat{\theta}_t$
Initialize $\mathcal{X}_t^{(0)} \leftarrow \mathcal{X}$ and $i \leftarrow 0$
while $|\mathcal{X}_t^{(i)}| > 1$ **do**
  Compute $\lambda_t^{(i)} \leftarrow \arg\inf_{\lambda \in \triangle_\mathcal{X}} \rho(\mathcal{X}_t^{(i)}, \lambda)$
  Find the largest index k_i such that

$$\inf_{\lambda \in \triangle_\mathcal{X}} \rho\left(\{x_{\widehat{(1)}_t}, \ldots, x_{\widehat{(k_i)}_t}\}\right) \leq \frac{1}{2} \cdot \inf_{\lambda \in \triangle_\mathcal{X}} \rho(\mathcal{X}_t^{(i)}, \lambda)$$

  Update $\mathcal{X}_t^{(i+1)} \leftarrow \left\{x_{\widehat{(1)}_t}, \ldots, x_{\widehat{(k_i)}_t}\right\}$
  $i \leftarrow i + 1$
end while
return $(\bar{\lambda}_t + \lambda^*)/2$, where $\bar{\lambda}_t = \frac{1}{i}\sum_{i'=0}^{i-1} \lambda_t^{(i')}$

Algorithm 13 Mixed-Peace

Input: budget, $T \in \mathbb{N}$; arm set $\mathcal{X} \subset \mathbb{R}^d$
Initialize $R \leftarrow \left\lceil \log_2 \left(\inf_{\lambda \in \triangle_\mathcal{X}} \rho(\mathcal{X}, \lambda)\right)\right\rceil$, $N \leftarrow \left\lfloor \frac{T}{R} \right\rfloor$, $\mathcal{X}_0 \leftarrow \mathcal{X}$, $\widehat{\theta}_0 \leftarrow \mathbf{0}$ and $t \leftarrow 1$
Compute G-optimal design λ^* using equation (2.3)
for $r = 0, \ldots, R$ **do**
  Find $\lambda_r \leftarrow (\arg\inf_{\lambda \in \triangle_\mathcal{X}} \rho(\mathcal{X}_r, \lambda) + \lambda^*)/2$
  while $t \leq \min\{T, (r+1)N\}$ **do**
    Sample $x_t \sim \lambda_r$ and receive reward r_t
    Estimate $\widehat{\theta}_t \leftarrow \frac{t-1}{t} \cdot \widehat{\theta}_{t-1} + \frac{1}{t} \cdot \mathbb{E}_{x \sim \lambda_r}\left[xx^\top\right]^{-1} x_t r_t$
    $t \leftarrow t + 1$
  end while
  if $|\mathcal{X}_r| > 1$ **then**
    Reindex $\mathcal{X}_r$ such that $x_1^\top \widehat{\theta}_t \geq x_2^\top \widehat{\theta}_t \geq \cdots \geq x_{n_r}^\top \widehat{\theta}_t$, where $n_r = |\mathcal{X}_r|$
    Find the largest index k_r such that

$$\inf_{\lambda \in \triangle_\mathcal{X}} \rho(\{x_1, \ldots, x_{k_r}\}, \lambda) \leq \frac{1}{2} \cdot \inf_{\lambda \in \triangle_\mathcal{X}} \rho(\mathcal{X}_r, \lambda)$$

    Update $\mathcal{X}_{r+1} \leftarrow \{x_1, \ldots, x_{k_r}\}$
  end if
end forreturn $\arg\max_{x \in \mathcal{X}} x^\top \widehat{\theta}_T$

we have

$$\mathbb{P}\left(x_{(1)}^\top \widehat{\theta}_T - x_{(1)}^\top \overline{\theta}_T \leq -\frac{\Delta_{(1)}}{2}\right) = \mathbb{P}\left(\sum_{t=1}^{T} x_{(1)}^\top \left(A(\lambda^*)^{-1} x_t r_t - \theta_t\right) \leq -\frac{T\Delta_{(1)}}{2}\right).$$

Since IPS estimator is unbiased, $x_{(1)}^\top \left(A(\lambda^*)^{-1} x_t r_t - \theta_t\right)$ is a zero-mean random variable. Based on our bounded reward assumption, we have

$$\left|x_{(1)}^\top \left(A(\lambda^*)^{-1} x_t r_t - \theta_t\right)\right| \leq \left|x_{(1)}^\top A(\lambda^*)^{-1} x_t\right| + 2 \leq \left\|x_{(1)}\right\|_{A(\lambda^*)^{-1}} \|x_t\|_{A(\lambda^*)^{-1}} + 2 \leq d + 2 \leq 3d,$$

where we use the property of G-optimal design $\max_{x\in\mathcal{X}} \|x\|^2_{A(\lambda^*)^{-1}} \leq d$. We can similarly bound its variance by

$$\begin{aligned}
\mathbb{E}\left[\left(x_{(1)}^\top \left(A(\lambda^*)^{-1} x_t r_t - \theta_t\right)\right)^2\right] \leq & \mathbb{E}\left[\left(x_{(1)}^\top A(\lambda^*)^{-1} x_t\right)^2\right] \\
= & x_{(1)}^\top A(\lambda^*)^{-1} \mathbb{E}\left[x_t x_t^\top\right] A(\lambda^*)^{-1} x_{(1)} \\
= & x_{(1)}^\top A(\lambda^*)^{-1} A(\lambda^*) A(\lambda^*)^{-1} x_{(1)} \\
& \qquad \text{(Since } x_t \sim \lambda^* \text{ by algorithm)} \\
= & \left\|x_{(1)}\right\|^2_{A(\lambda^*)^{-1}} \leq d
\end{aligned}$$

Thus, by Bernstein's inequality, we have

$$\mathbb{P}\left(x_{(1)}^\top \widehat{\theta}_T - x_{(1)}^\top \overline{\theta}_T \leq -\frac{\Delta_{(1)}}{2}\right) \leq \exp\left(-\frac{T^2\Delta_{(1)}^2/8}{Td + Td\Delta_{(1)}/2}\right) \leq \exp\left(-\frac{T\Delta_{(1)}^2}{12d}\right),$$

where the last inequality uses the assumption that $\Delta_{(1)} \leq 1$. By similarly applying Bernstein's inequality to other terms in (A.2), we can then have

$$\begin{aligned}
\mathbb{P}\left(J_T \neq x_{(1)}\right) \leq & \mathbb{P}\left(x_{(1)}^\top \widehat{\theta}_T - x_{(1)}^\top \overline{\theta}_T \leq -\frac{\Delta_{(1)}}{2}\right) + \sum_{k=2}^{K} \mathbb{P}\left(x_{(k)}^\top \widehat{\theta}_T - x_{(k)}^\top \overline{\theta}_T \geq \frac{\Delta_{(k)}}{2}\right) \\
\leq & \sum_{k=1}^{K} \exp\left(-\frac{T\Delta_{(k)}^2}{12d}\right)
\end{aligned}$$

$$\leq K \exp\left(-\frac{T\Delta_{(1)}^2}{12d}\right).$$

□

A.3 Error Probability of Algorithm 2

A.3.1 Stationary Environments

We first prove an error probability of Algorithm 2 in stationary environments that contains unspecified parameters from the virtual phases. Without loss of generality, assume that the arms $x_1, \ldots, x_K$ are ordered such that $\theta^\top x_1 > \theta^\top x_2 \geq \cdots \geq \theta^\top x_K$ and $\Delta_1 = \Delta_2 \leq \Delta_3 \leq \cdots \leq \Delta_K$.

Throughout this section, we will the following definitions: $i_0 = \lceil \log_2(1/\Delta_1) \rceil + 1$, $\mathcal{A}_i = \left\{x \in \mathcal{X} \mid \Delta_x \leq 2 \cdot 2^{-i}\right\}$, $\bar{i}(k) = \max\left\{i \in [i_0 - 1] \mid \Delta_k \leq 2^{-i}\right\}$ and

$$f(\mathcal{A}_i) = \min_{\lambda \in \triangle_{\mathcal{X}}} \max_{x, x' \in \mathcal{A}_i} \left\|x - x'\right\|^2_{A(\lambda)^{-1}}.$$

Theorem A.3.1. *Let* $\mathcal{D} = \left\{\mathbf{a} \in [0,1]^{i_0+1} \mid 0 = a_0 < a_1 \leq a_2 \leq \ldots \leq a_{i_0} = 1\right\}$. *Then, if* $m \geq i_0$, *The error probability of Algorithm 2 in a stationary environment with parameter* θ *is bounded as*

$$\mathbb{P}_\theta\left(J_T \neq 1\right) \leq 2 i_0 K T \exp\left(-\frac{T}{\overline{H}_{\textit{P1-RAGE}}(\theta)}\right),$$

$$\overline{H}_{\textit{P1-RAGE}}(\theta) = \min_{\mathbf{a} \in \mathcal{D}} \max_{k \in [K]} \frac{48m \sum_{i'=1}^{\bar{i}(k)} (a_{i'} - a_{i'-1}) f(\mathcal{A}_{i'-2}) + 8(m\sqrt{df(\mathcal{X})} + 1) a_{\bar{i}(k)} \Delta_k}{3 a_{\bar{i}(k)}^2 \Delta_k^2}. \tag{A.3}$$

Proof. With $0 = n_0 < n_1 \leq n_2 \leq \ldots \leq n_{i_0} = T$.* we define the event ξ_i with $i \geq 1$ as follows: after n_i samples all the arms with true gap smaller than $2 \cdot 2^{-i}$ are estimated with precision $2^{-i}/2$, which is

$$\xi_i = \{\forall t \geq n_i, \forall k \in [K] \text{ s.t. } \Delta_k \leq 2 \cdot 2^{-i} \implies |\Delta_k - \widehat{\Delta}_k^{(t)}| < 2^{-i}/2\},$$

*We do not specify the values of $n_1, \ldots, n_{i_0 - 1}$ for now.

where $\widehat{\Delta}_k^{(t)} = (x_1 - x_k)^\top \widehat{\theta}^{(t)}$ for $k > 1$ and $\widehat{\Delta}_1^{(t)} = (x_1 - x_2)^\top \widehat{\theta}^{(t)}$. We first show how these events $\{\xi_i\}_{i=1}^{i_0}$ relate the correctness of Algorithm 2.

Correctness. If $\bigcap_{i=1}^{i_0} \xi_i$ holds then the algorithm successfully identifies the best arm. Indeed, if we assume it does not, then there must exist non-optimal arm k_0 such that $\widehat{\Delta}_{k_0}^{(T)} < 0$. As $\bigcap_{i=1}^{i_0} \xi_i$ holds, for some $i' \le i_0$, it holds that $2^{-i'} < \Delta_{k_0} \le 2 \cdot 2^{-i'}$ and then $|\Delta_{k_0} - \widehat{\Delta}_{k_0}^{(T)}| < 2^{-i'}/2$. Therefore, we have $2^{-i'} < \Delta_{k_0} \le \Delta_{k_0} - \widehat{\Delta}_{k_0}^{(T)} \le |\Delta_{k_0} - \widehat{\Delta}_{k_0}^{(T)}| \le 2^{-i'}/2$, which is a contradiction.

Thus, the error probability is upper bounded by $\mathbb{P}\left(\bigcup_{i=1}^{i_0} \xi_i^c\right)$, which gives us

$$
\begin{aligned}
\mathbb{P}\left(J_T \neq 1\right) \le & \mathbb{P}\left(\bigcup_{i=1}^{i_0} \xi_i^c\right) = \mathbb{P}\left(\bigcup_{i=1}^{i_0} \left(\xi_i^c \setminus \bigcup_{j=1}^{i-1} \xi_j^c\right)\right) \le \sum_{i=1}^{i_0} \mathbb{P}\left(\xi_i^c \setminus \bigcup_{j=1}^{i-1} \xi_j^c\right) \\
= & \sum_{i=1}^{i_0} \mathbb{P}\left(\xi_i^c \cap \left(\bigcup_{j=1}^{i-1} \xi_j^c\right)^c\right) = \sum_{i=1}^{i_0} \mathbb{P}\left(\xi_i^c \cap \left(\bigcap_{j=1}^{i-1} \xi_j\right)\right) \\
\le & \sum_{i=1}^{i_0} \mathbb{P}\left(\xi_i^c \middle| \bigcap_{j=1}^{i-1} \xi_j\right).
\end{aligned}
$$

Bernstein's inequality. Now, we just need to find an upper bound of $\mathbb{P}\left(\xi_i^c \middle| \bigcap_{j=1}^{i-1} \xi_j\right)$. Assume $\exists t \ge n_i, \exists k \in [K]$ s.t. $\Delta_k \le 2 \cdot 2^{-i}$.[†] Then, we have

$$
\begin{aligned}
& \mathbb{P}(|\Delta_k - \widehat{\Delta}_k^{(t)}| \ge 2^{-i}/2) \\
= & \mathbb{P}(|(\theta - \widehat{\theta}_t)^\top (x_1 - x_k)| \ge 2^{-i}/2) \qquad \text{(A.4)} \\
= & \mathbb{P}\left(\left|\sum_{s=1}^{t} \left(\theta - A(\lambda_s)^{-1} x_s r_s\right)^\top (x_1 - x_k)\right| \ge 2^{-i} t/2\right) \\
\overset{\text{(a)}}{\le} & 2\exp\left(-\frac{2^{-2i} t^2/8}{2\sum_{s=1}^{t} \|x_1 - x_k\|^2_{A(\bar{\lambda}_s)^{-1}} + \left(\sqrt{d}\max_{s\in[1:t]} \|x_1 - x_k\|_{A(\bar{\lambda}_s)^{-1}} + 1\right) t 2^{-i}/3}\right) \\
& \qquad \text{(By Bernstein's inequality for martingale differences Freedman [1975])} \\
\le & 2\exp\left(-\frac{2^{-2i} t^2/8}{\text{term I}}\right),
\end{aligned}
$$

[†]Otherwise, ξ_i is vacuously true and $\mathbb{P}\left(\xi_i^c\right) = 0$.

$$\text{where term I} = 2\sum_{i'=1}^{i}\sum_{s=n_{i'-1}+1}^{n_{i'}} \|x_1 - x_k\|^2_{A(\bar{\lambda}_s)^{-1}} + 2\sum_{s=n_i+1}^{t} \|x_1 - x_k\|^2_{A(\bar{\lambda}_s)^{-1}} + \left(\sqrt{d}\max_{s\in[1:t]} \|x_1 - x_k\|_{A(\bar{\lambda}_s)^{-1}} + 1\right)\cdot \frac{t2^{-i}}{3}.$$

Here, to use Bernstein's inequality for martingale differences in the inequality (a) above, we need to bound the variance and magnitude of $\left(\theta - A(\lambda_s)^{-1}x_s r_s\right)^\top (x_1 - x_k)$ condition on λ_s.[‡] In particular, we have

$$\begin{aligned}
\left|\left(\theta - A(\lambda_s)^{-1}x_s r_s\right)^\top (x_1 - x_k)\right| &\le \left|(x_1 - x_k)^\top A(\lambda_s)^{-1}x_s\right| + \Delta_k \\
&\le \|x_1 - x_k\|_{A(\lambda_s)^{-1}} \|x_s\|_{A(\lambda_s)^{-1}} + 2 \\
&\le 2\sqrt{d}\,\|x_1 - x_k\|_{A(\bar{\lambda}_s)^{-1}} + 2. \\
&\qquad \text{(Since } \lambda_s = (\bar{\lambda}_s + \lambda^*)/2 \text{ and } \lambda \mapsto \|x_1 - x_k\|^2_{A(\lambda)^{-1}} \text{ is convex in } \lambda)
\end{aligned}$$

$$\begin{aligned}
&\mathbb{E}\left[\left(\left(\theta - A(\lambda_s)^{-1}x_s r_s\right)^\top (x_1 - x_k)\right)^2 \mid \lambda_s\right] \\
\le &\mathbb{E}\left[\left((x_1 - x_k)^\top A(\lambda_s)^{-1}x_s\right)^2 \mid \lambda_s\right] \\
= &(x_1 - x_k)^\top A(\lambda_s)^{-1}\mathbb{E}\left[x_s x_s^\top \mid \lambda_s\right] A(\lambda_s)^{-1}(x_1 - x_k) \\
= &\|x_1 - x_k\|^2_{A(\lambda_s)^{-1}} \\
\le &2\,\|x_1 - x_k\|^2_{A(\bar{\lambda}_s)^{-1}}. \qquad \text{(Since } \lambda_s = (\bar{\lambda}_s + \lambda^*)/2)
\end{aligned}$$

Single-term error probability. Now, we need to use the property of the subroutine RAGE-Elimination (Line 3 of Algorithm 2) that generates λ_s. That is, by Lemma A.3.4, since $x_k \in \mathcal{A}_i \subseteq \mathcal{A}_{i'}$ for $i' \le i$ and $m \ge i_0$, for $s \in [n_{i'-1}+1, n_{i'}]$, we have $\|x_1 - x_k\|^2_{A(\bar{\lambda}_s)^{-1}} \le m \inf_{\lambda\in\triangle_\mathcal{X}} \max_{x,x'\in\mathcal{A}_{i'-2}} \|x - x'\|^2_{A(\lambda)^{-1}} \overset{\text{def}}{=} m f(\mathcal{A}_{i'-2})$. Thus, we have

$$\mathbb{P}(|\Delta_k - \widehat{\Delta}_k^{(t)}| \ge 2^{-i}/2)$$

[‡]Since IPS estimator is unbiased and λ_s is determined by the history prior to time s, we have $\mathbb{E}\left[\left(\theta - A(\lambda_s)^{-1}x_s r_s\right)^\top (x_1 - x_k) \mid \mathcal{H}_{s-1}\right] = 0$, which implies that it is a martingale difference sequence.

$$\leq 2\exp\left(-\frac{2^{-2i}t^2/8}{2m\sum_{i'=1}^{i}(n_{i'}-n_{i'-1})f(\mathcal{A}_{i'-2})+2m(t-n_i)f(\mathcal{A}_{i-1})+(m\sqrt{df(\mathcal{X})}+1)t2^{-i}/3}\right)$$
$$\leq 2\exp\left(-\frac{2^{-2i}n_i^2/8}{2m\sum_{i'=1}^{i}(n_{i'}-n_{i'-1})f(\mathcal{A}_{i'-2})+(m\sqrt{df(\mathcal{X})}+1)n_i2^{-i}/3}\right),$$

where the last inequality above holds because of $t \geq n_i$ and a simple fact that $t \mapsto \frac{t^2}{at+b}$ is an increasing function when $t \geq 0$ if $a > 0$ and $b > 0$.

Final error probability. Then, with the union bound over all $t \geq n_i$ and $k \in [K]$, it holds for any $0 < n_1 \leq n_2 \ldots \leq n_i \leq T$ that

$$\mathbb{P}\left(\xi_i^c \middle| \bigcap_{j=1}^{i-1}\xi_j\right) \leq 2KT\exp\left(-\frac{2^{-2i}n_i^2/8}{2m\sum_{i'=1}^{i}(n_{i'}-n_{i'-1})f(\mathcal{A}_{i'-2})+(m\sqrt{df(\mathcal{X})}+1)n_i2^{-i}/3}\right)$$
$$\leq 2KT\max_{k\in[K]}\exp\left(-\frac{3n_{\bar{i}(k)}^2\Delta_k^2}{48m\sum_{i'=1}^{\bar{i}(k)}(n_{i'}-n_{i'-1})f(\mathcal{A}_{i'-2})+8(m\sqrt{df(\mathcal{X})}+1)n_{\bar{i}(k)}\Delta_k}\right),$$

where $\bar{i}(k) = \max\left\{i \in [i_0-1] \mid \Delta_k \leq 2^{-i}\right\}$. Here, the last inequality use the same simple fact that $t \mapsto \frac{t^2}{at+b}$ is an increasing function when $t \geq 0$ if $a > 0$ and $b > 0$.

With values of $0 = n_0 < n_1 \leq n_2 \leq \cdots \leq n_{i_0} = T$, we can define $a_i = \frac{n_i}{T}$, which implies $0 = a_0 < a_1 \leq a_2 \leq \cdots \leq a_{i_0} = 1$. Since the choice of values $\mathbf{a} \in \mathcal{D}$ is arbitrary, the final error probability can be bounded as

$$\mathbb{P}\left(J_T \neq 1\right) \leq \sum_{i=1}^{i_0}\mathbb{P}\left(\xi_j^c \middle| \bigcap_{j=1}^{i-1}\xi_j\right)$$
$$\leq 2i_0KT\min_{\mathbf{a}\in\mathcal{D}}\max_{k\in[K]}\exp\left(-\frac{3Ta_{i(k)}^2\Delta_k^2}{48m\sum_{i'=1}^{\bar{i}(k)}(a_{i'}-a_{i'-1})f(\mathcal{A}_{i'-2})+8(m\sqrt{df(\mathcal{X})}+1)a_{i(k)}\Delta_k}\right),$$

which completes the proof □

Properties of RAGE-Elimination

In this section, we prove some properties of the RAGE-Elimination algorithm that will be useful for proving Theorem A.3.1.

Lemma A.3.2. *Assume $t \geq n_i$. Then, under $\bigcap_{j=1}^{i-1} \xi_j$, when running RAGE-Elimination (line 3 in Algorithm 2), it holds that*

$$\mathcal{X}_t^{(i+1)} \subseteq \left\{x \in \mathcal{X} \mid \widehat{\Delta}_x^{(t)} \leq 2^{-i}\right\} \subseteq \mathcal{A}_i.$$

Proof. To show $\mathcal{X}_t^{(i+1)} \subseteq \left\{x \in \mathcal{X} \mid \widehat{\Delta}_x^{(t)} \leq 2^{-i}\right\}$, let $x_{\widehat{(1)}_t} = \arg\max_{x \in \mathcal{X}} \left\langle \widehat{\theta}_t, x \right\rangle$. Then, for some arm x, if we have $\left\langle \widehat{\theta}^{(t)}, x_{\widehat{(1)}_t} - x \right\rangle \leq 2^{-i}$, it holds that

$$\left\langle \widehat{\theta}^{(t)}, x_1 - x \right\rangle = \underbrace{\left\langle \widehat{\theta}^{(t)}, x_1 - x_{\widehat{(1)}_t} \right\rangle}_{\leq 0} + \underbrace{\left\langle \widehat{\theta}^{(t)}, x_{\widehat{(1)}_t} - x \right\rangle}_{\leq 2^{-i}} \leq 2^{-i},$$

which implies $x \in \{x \in \mathcal{X} \mid \widehat{\Delta}_x^{(t)} \leq 2^{-i}\}$.

To show $\left\{x \in \mathcal{X} \mid \widehat{\Delta}_x^{(t)} \leq 2^{-i}\right\} \subseteq \mathcal{A}_i$, let $\widehat{\Delta}_x^{(t)} \leq 2^{-i}$ for some x and assume for the sake of a contradiction that $\Delta_x > 2 \cdot 2^{-i}$. As $\Delta_x > 2 \cdot 2^{-i}$, there must exist $\tilde{i} \leq i-1$ such that $2^{-\tilde{i}} < \Delta_x \leq 2 \cdot 2^{-\tilde{i}}$. Then $|\Delta_x - \widehat{\Delta}_x^{(t)}| < 2^{-\tilde{i}}/2$ since event $\xi_{\tilde{i}}$ holds. Meanwhile, we have $\widehat{\Delta}_x^{(t)} \leq 2^{-i} \leq 2^{-\tilde{i}}/2$ since $\tilde{i} \leq i-1$. Now, this leads to the contradiction

$$2^{-\tilde{i}}/2 = 2^{-\tilde{i}} - 2^{-\tilde{i}}/2 \leq \Delta_x - \widehat{\Delta}_x^{(t)} \leq |\Delta_x - \widehat{\Delta}_j^{(t)}| < 2^{-\tilde{i}}/2.$$

Thus, under $\bigcap_{j=1}^{i-1} \xi_j$, we have

$$\left\{x \in \mathcal{X} \mid \widehat{\Delta}_x^{(t)} \leq 2^{-i}\right\} \subseteq \left\{x \in \mathcal{X} \mid \Delta_x \leq 2 \cdot 2^{-i}\right\} = \mathcal{A}_i.$$

□

Lemma A.3.3. *Assume $t \geq n_i$. Then, under $\bigcap_{j=1}^{i-1} \xi_j$, when running RAGE-Elimination, if $x \in \mathcal{A}_i$, then $x \in \mathcal{X}_t^{(i-1)}$.*

Proof. If $x \in \mathcal{A}_i$, then $\langle \theta, x_1 - x \rangle \leq 2 \cdot 2^{-i}$. Again, let $x_{\widehat{(1)}_t} = \arg\max_{x \in \mathcal{X}} \left\langle \widehat{\theta}_t, x \right\rangle$ and we have

$$\left\langle \widehat{\theta}_t, \hat{x}_1^{(t)} - x \right\rangle = \left\langle \widehat{\theta}_t, x_{\widehat{(1)}_t} - x_1 \right\rangle + \left\langle \widehat{\theta}_t, x_1 - x \right\rangle$$

$$
\begin{aligned}
&= \left\langle \widehat{\theta}_t, x_{\widehat{(1)}_t} - x_1 \right\rangle + \left\langle \widehat{\theta}_t - \theta, x_1 - x \right\rangle + \underbrace{\langle \theta, x_1 - x \rangle}_{\leq 2 \cdot 2^{-i}} \\
&\leq \left\langle \widehat{\theta}_t, x_{\widehat{(1)}_t} - x_1 \right\rangle + |\widehat{\Delta}_x^{(t)} - \Delta_x| + 2 \cdot 2^{-i} \\
&\leq \left\langle \widehat{\theta}_t, x_{\widehat{(1)}_t} - x_1 \right\rangle + 2^{-i} + 2 \cdot 2^{-i} \qquad \text{(Since } \xi_{i-1} \text{ holds)} \\
&= - \widehat{\Delta}_{x_{\widehat{(1)}_t}}^{(t)} + 2^{-i} + 2 \cdot 2^{-i} \\
&\leq 2^{-i} + 2^{-i} + 2 \cdot 2^{-i} \\
&= 4 \cdot 2^{-i}.
\end{aligned}
$$

The last inequality above holds because under $\bigcap_{j=1}^{i-1} \xi_j$, by Lemma A.3.2, we have $x_{\widehat{(1)}_t} \in \mathcal{A}_i$, meaning that $|\widehat{\Delta}_{x_{\widehat{(1)}_t}}^{(t)} - \Delta_{x_{\widehat{(1)}_t}}| < 2^{-i} \implies \widehat{\Delta}_{x_{\widehat{(1)}_t}}^{(t)} > \Delta_{x_{\widehat{(1)}_t}} - 2^{-i} > -2^{-i}$. □

Lemma A.3.4. *Assume $t \geq n_i$ and $\bigcap_{j=1}^{i-1} \xi_j$ holds. When running RAGE-Elimination, If $x_k \in \mathcal{A}_i$, then*

$$
\|x_1 - x_k\|^2_{A(\bar{\lambda}_t)^{-1}} \leq m \min_{\lambda \in \triangle_{\mathcal{X}}} \max_{x, x' \in \mathcal{A}_{i-2}} \|x - x'\|^2_{A(\lambda)^{-1}} .
$$

Proof. By Lemma A.3.3, we have $x_1, x_k \in \mathcal{A}_i \implies x_1, x_k \in \mathcal{X}_t^{(i-1)}$, which means that $\left|\mathcal{X}_t^{(i-1)}\right| \geq 2$ and $\bar{\lambda}_t = \frac{1}{i_t} \sum_{i'=1}^{i_t} \lambda_t^{(i')}$ for some i_t satisfying $i - 1 \leq i_t \leq m$. Thus, We have

$$
\begin{aligned}
\|x_1 - x_k\|^2_{A(\bar{\lambda}_t)^{-1}} &\leq m \|x_1 - x_k\|^2_{A\left(\lambda_t^{(i-1)}\right)^{-1}} \\
&\leq m \max_{x, x' \in \mathcal{X}_t^{(i-1)}} \|x - x'\|^2_{A\left(\lambda_t^{(i-1)}\right)^{-1}} \qquad \text{(Since } x_1, x_k \in \mathcal{X}_{i-1}^{(t)}\text{)} \\
&\overset{\text{(i)}}{\leq} m \min_{\lambda \in \triangle_{\mathcal{X}}} \max_{x, x' \in \mathcal{A}_{i-2}} \|x - x'\|^2_{A(\lambda)^{-1}} .
\end{aligned}
$$

Here, the above inequality (i) holds because by Lemma A.3.2, we have $\mathcal{X}_t^{(i-1)} \subseteq \mathcal{A}_{i-2}$ and by algorithm construction, we have $\lambda_t^{(i-1)} \in \arg\min_{\lambda \in \triangle_{\mathcal{X}}} \max_{x, x' \in \mathcal{X}_t^{(i-1)}} \|x - x'\|^2_{A(\lambda)^{-1}}$. □

Simplified Stationary Complexity and its Relation to Multi-armed Bandits

In this section, we simplify the complexity of Algorithm 2 obtained in Theorem A.3.1 by appropriately choosing values $\mathbf{a} \in \mathcal{D}$. In particular, we have the following theorem.

Theorem A.3.5. *For $\overline{H}_{\text{P1-RAGE}}(\theta)$ defined in equation (A.3), we have*

$$\overline{H}_{\text{P1-RAGE}}(\theta) \le \frac{1024mi_0}{\Delta_1} \inf_{\lambda\in\triangle_{\mathcal{X}}} \max_{x\neq x_1} \frac{\|x-x_1\|^2_{A(\lambda)^{-1}}}{\Delta_x} + \frac{16m\sqrt{d}}{3\Delta_1} \inf_{\lambda\in\triangle_{\mathcal{X}}} \max_{x\neq x_1} \|x-x_1\|_{A(\lambda)^{-1}} + \frac{1}{3\Delta_1}.$$

Proof. For $i \in \{1,\dots,i_0-1\}$, we take $a_i = \frac{\Delta_1}{\Delta_{\bar{k}(i)}}$, where $\bar{k}(i) = \min\left\{k\in[K] \mid \Delta_k \ge \frac{2^{-i}}{2}\right\}$. Then, since $\bar{i}(k) = \max\left\{i\in[i_0-1] \mid \Delta_k \le 2^{-i}\right\}$, for any $k\in[K]$, we have $\frac{2^{-\bar{i}(k)}}{2} \le \Delta_{\bar{k}(\bar{i}(k))} \le \Delta_k$, which further implies

$$a_{\bar{i}(k)}\Delta_k = \frac{\Delta_1}{\Delta_{\bar{k}(\bar{i}(k))}} \cdot \Delta_k \ge \Delta_1.$$

Then, for $\overline{H}_{\text{P1-RAGE}}(\theta)$ (defined in equation (A.3)), we have

$$\begin{aligned}
\overline{H}_{\text{P1-RAGE}}(\theta) &\le \max_{k\in[K]} \left\{ \frac{16m\sum_{i'=1}^{\bar{i}(k)}(a_{i'}-a_{i'-1})f(\mathcal{A}_{i'-2})}{a^2_{\bar{i}(k)}\Delta_k^2} + \frac{8(m\sqrt{df(\mathcal{X})}+1)}{3a_{\bar{i}(k)}\Delta_k} \right\} \\
&\le \frac{16m}{\Delta_1} \max_{k\in[K]} \left\{ \frac{f(\mathcal{A}_{-1})}{\Delta_{\bar{k}(1)}} + \sum_{i'=2}^{\bar{i}(k)} \left(\frac{1}{\Delta_{\bar{k}(i')}} - \frac{1}{\Delta_{\bar{k}(i'-1)}} \right) f(\mathcal{A}_{i'-2}) \right\} + \frac{8(m\sqrt{df(\mathcal{X})}+1)}{3\Delta_1}. \\
&\qquad\qquad \text{(Since } a_0 = 0 \text{ by definition)}
\end{aligned}$$

For the second term, using the definition of $f(\mathcal{X})$, we simply have

$$\begin{aligned}
\frac{8(m\sqrt{df(\mathcal{X})}+1)}{3\Delta_1} &= \frac{8m\sqrt{d}}{3\Delta_1} \inf_{\lambda\in\triangle_{\mathcal{X}}} \max_{x,x'\in\mathcal{X}} \left\|x-x_1+x_1-x'\right\|_{A(\lambda)^{-1}} + \frac{1}{3\Delta_1} \\
&\le \frac{16m\sqrt{d}}{3\Delta_1} \inf_{\lambda\in\triangle_{\mathcal{X}}} \max_{x\neq x_1} \|x-x_1\|_{A(\lambda)^{-1}} + \frac{1}{3\Delta_1}. \qquad \text{(A.5)}
\end{aligned}$$

For the first term, by fixing arm index $k\in[K]$ and defining $j \in \arg\max_{\ell\in[\bar{i}(k)]} \frac{f(\mathcal{A}_{\ell-2})}{\Delta_{\bar{k}(\ell)}}$, we have

$$\begin{aligned}
&\frac{f(\mathcal{A}_{-1})}{\Delta_{\bar{k}(1)}} + \sum_{i'=2}^{\bar{i}(k)} \left(\frac{1}{\Delta_{\bar{k}(i')}} - \frac{1}{\Delta_{\bar{k}(i'-1)}} \right) f(\mathcal{A}_{i'-2}) \\
=&\frac{f(\mathcal{A}_{\bar{i}(k)-2})}{\Delta_{\bar{k}(\bar{i}(k))}} + \sum_{i'=1}^{\bar{i}(k)-1} \frac{f(\mathcal{A}_{i'-2}) - f(\mathcal{A}_{i'-1})}{\Delta_{\bar{k}(i')}}
\end{aligned}$$

$$
\begin{aligned}
&\overset{(a)}{\leq} \frac{f(\mathcal{A}_{j-2})}{\Delta_{\bar{k}(j)}} \left(1 + \sum_{i'=1}^{\bar{i}(k)-1} \frac{f(\mathcal{A}_{i'-2}) - f(\mathcal{A}_{i'-1})}{f(\mathcal{A}_{i'-2})}\right) \\
&\leq \bar{i}(k) \frac{f(\mathcal{A}_{j-2})}{\Delta_{\bar{k}(j)}} && \text{(Since } f(\mathcal{A}_{i'-2}) \geq f(\mathcal{A}_{i'-1})\text{)} \\
&\leq i_0 \max_{\ell \in [\bar{i}(k)]} \frac{f(\mathcal{A}_{\ell-2})}{\Delta_{\bar{k}(\ell)}} && \text{(Since } \bar{i}(k) \leq i_0 \text{ for any } k \in [K]\text{)} \\
&= i_0 \max_{\ell \in [\bar{i}(k)]} \inf_{\lambda \in \triangle_{\mathcal{X}}} \max_{x, x' \in \mathcal{A}_{\ell-2}} \frac{\|x - x'\|^2_{A(\lambda)^{-1}}}{\Delta_{\bar{k}(\ell)}} \\
&\leq i_0 \inf_{\lambda \in \triangle_{\mathcal{X}}} \max_{\ell \in [\bar{i}(k)]} \max_{x, x' \in \mathcal{A}_{\ell-2}} \frac{\|x - x'\|^2_{A(\lambda)^{-1}}}{\Delta_{\bar{k}(\ell)}} && \text{(By the weak duality inequality)} \\
&\leq 64 i_0 \inf_{\lambda \in \triangle_{\mathcal{X}}} \max_{\ell \in [\bar{i}(k)]} \max_{x \in \mathcal{A}_{\ell-2}, x \neq x_1} \frac{\|x - x_1\|^2_{A(\lambda)^{-1}}}{16 \Delta_{\bar{k}(\ell)}} && \text{(By reasoning similar to equation (A.5))} \\
&\overset{(b)}{\leq} 64 i_0 \inf_{\lambda \in \triangle_{\mathcal{X}}} \max_{\ell \in [\bar{i}(k)]} \max_{x \in \mathcal{A}_{\ell-2}, x \neq x_1} \frac{\|x - x_1\|^2_{A(\lambda)^{-1}}}{\Delta_x} \\
&\leq 64 i_0 \inf_{\lambda \in \triangle_{\mathcal{X}}} \max_{x \neq x_1} \frac{\|x - x_1\|^2_{A(\lambda)^{-1}}}{\Delta_x}.
\end{aligned}
$$

Here, the inequality (a) above holds because $f(\mathcal{A}_{i'-2}) \geq f(\mathcal{A}_{i'-1})$ and by definition of j, we have $\frac{f(\mathcal{A}_{\ell-2})}{\Delta_{\bar{k}(\ell)}} \leq \frac{f(\mathcal{A}_{j-2})}{\Delta_{\bar{k}(j)}}$. The inequality (b) above holds because by definitions of $\bar{k}(\ell) = \min\left\{k \in [K] \mid \Delta_k \geq \frac{2^{-i}}{2}\right\}$ and $\mathcal{A}_{\ell-2} = \left\{x \in \mathcal{X} \mid \Delta_x \leq 2 \cdot 2^{-(\ell-2)}\right\}$, we have $16\Delta_{\bar{k}(\ell)} \geq \Delta_x$ for any $x \in \mathcal{A}_{\ell-2}$.

Therefore, by plugging the bound of both terms back, we have

$$
\overline{H}_{\text{P1-RAGE}}(\theta) \leq \frac{1024 m i_0}{\Delta_1} \inf_{\lambda \in \triangle_{\mathcal{X}}} \max_{x \neq x_1} \frac{\|x - x_1\|^2_{A(\lambda)^{-1}}}{\Delta_x} + \frac{16 m \sqrt{d}}{3\Delta_1} \inf_{\lambda \in \triangle_{\mathcal{X}}} \max_{x \neq x_1} \|x - x_1\|_{A(\lambda)^{-1}} + \frac{1}{3\Delta_1}.
$$

□

In the following corollary, we show that the above simplified complexity is in a same order (up to logarithmic factors) of H_{BOB} proposed in Abbasi-Yadkori et al. [2018].

Corollary A.3.6. *In multi-armed bandits, meaning $d = K$ and $\mathcal{X} = \{\mathbf{e}_1, \ldots, \mathbf{e}_K\}$, for*

$H_{P1\text{-}RAGE}(\theta)$ (defined in equation (2.4)), if $m = i_0$, we then have

$$H_{P1\text{-}RAGE}(\theta) \leq \frac{2i_0\left(i_0 \log(2K) + 1\right)}{\Delta_{(1)}} \max_{k \in [K]} \frac{k}{\Delta_{(k)}} = 2i_0\left(i_0 \log(2K) + 1\right) H_{\mathrm{BOB}}(\theta).$$

Proof. When in multi-armed bandits, for the first term in $H_{\mathsf{P1\text{-}RAGE}}(\theta)$, we have

$$\inf_{\lambda \in \triangle_{\mathcal{X}}} \max_{x \neq x_{(1)}} \frac{\left\|x - x_{(1)}\right\|^2_{A(\lambda)^{-1}}}{\Delta_x} \leq 2 \sum_{k=1}^{K} \frac{1}{\Delta_k} \leq 2 \log(2K) \max_{k \in [K]} \frac{k}{\Delta_{(k)}},$$

where the first inequality above comes from Soare et al. [2014] and the second inequality comes from Audibert et al. [2010]. For the second term in $H_{\mathsf{P1\text{-}RAGE}}(\theta)$, we have

$$\inf_{\lambda \in \triangle_{\mathcal{X}}} \max_{x \neq x_{(1)}} \left\|x - x_{(1)}\right\|_{A(\lambda)^{-1}} = \inf_{\lambda \in \triangle_{\mathcal{X}}} \max_{k \neq (1)} \sqrt{\frac{1}{\lambda_{(1)}} + \frac{1}{\lambda_k}} = \sqrt{2K},$$

which then gives us $\frac{\sqrt{K} \cdot \sqrt{2K}}{\Delta_{(1)}} \leq \frac{2K}{\Delta_{(1)}\Delta_{(K)}} \leq \frac{2}{\Delta_{(1)}} \max_{k \in [K]} \frac{k}{\Delta_{(k)}}$.

Finally, by plugging these inequalities back into $H_{\mathsf{P1\text{-}RAGE}}(\theta)$ (defined in equation (2.4)), we have

$$\begin{aligned}
H_{\mathsf{P1\text{-}RAGE}}(\theta) =& \frac{mi_0}{\Delta_{(1)}} \inf_{\lambda \in \triangle_{\mathcal{X}}} \max_{x \neq x_{(1)}} \frac{\left\|x - x_{(1)}\right\|^2_{A(\lambda)^{-1}}}{\Delta_x} + \frac{m\sqrt{d}}{\Delta_{(1)}} \inf_{\lambda \in \triangle_{\mathcal{X}}} \max_{x \neq x_{(1)}} \left\|x - x_{(1)}\right\|_{A(\lambda)^{-1}} \\
\leq& \frac{2i_0^2 \log(2K)}{\Delta_{(1)}} \max_{k \in [K]} \frac{k}{\Delta_{(k)}} + \frac{2i_0}{\Delta_{(1)}} \max_{k \in [K]} \frac{k}{\Delta_{(k)}} \\
=& \frac{2i_0\left(i_0 \log(2K) + 1\right)}{\Delta_{(1)}} \max_{k \in [K]} \frac{k}{\Delta_{(k)}}.
\end{aligned}$$

□

Approximate BAI of Algorithm 2

Corollary A.3.7. *Fix arm set $\mathcal{X} \subset \mathbb{R}^d$ with $|\mathcal{X}| = K$ and budget T. For a stationary environment with unknown parameter θ, if $m \geq i_0(\epsilon) = \lceil \log_2(1/\epsilon) \rceil + 1$ for some $\epsilon \geq \Delta_1$,*

then there exists absolute constant $c > 0$ such that the error probability of P1-RAGE *satisfies*

$$\mathbb{P}_\theta\left(J_T \notin \mathcal{A}(\epsilon)\right) \leq 2i_0(\epsilon)KT\exp\left(-\frac{cT}{H_{\text{P1-RAGE}}(\theta,\epsilon)}\right),$$

where $\mathcal{A}(\epsilon) = \{x \in \mathcal{X} \mid \Delta_x \leq \epsilon\}$ and $H_{\text{P1-RAGE}}(\theta,\epsilon)$ is defined as replacing i_0 by $i_0(\epsilon)$ in $H_{\text{P1-RAGE}}(\theta)$ (defined in Eq. (A.3)).

Proof. The proof is the same as Theorem 2.5.1 through simply replacing i_0 by $i_0(\epsilon)$. □

A.3.2 Non-stationary Environments

In this section, we prove the error probability of Algorithm 2 in general non-stationary environments.

Theorem A.3.8. *Fix time horizon T, arm set $\mathcal{X} \subset \mathbb{R}^d$ with $|\mathcal{X}| = K$ and arbitrary unknown parameters $\{\theta_t\}_{t=1}^T$. If we run Algorithm 2 in this non-stationary environment and obtain x_{J_T}, then it holds that*

$$\mathbb{P}_{\overline{\theta}_T}\left(J_T \neq (1)\right) \leq K\exp\left(-\frac{3T\Delta_{(1)}^2}{64d}\right).$$

Proof. The proof will basically resemble the one for Theorem 2.4.1. In particular, by the same reasoning to obtain equation A.2, we have

$$\mathbb{P}\left(J_T \neq (1)\right) \leq \mathbb{P}\left(x_{(1)}^\top\widehat{\theta}_T - x_{(1)}^\top\overline{\theta}_T \leq -\frac{\Delta_{(1)}}{2}\right) + \sum_{k=2}^K \mathbb{P}\left(x_{(k)}^\top\widehat{\theta}_T - x_{(k)}^\top\overline{\theta}_T \geq \frac{\Delta_{(k)}}{2}\right),$$

$$\text{where } \mathbb{P}\left(x_{(1)}^\top\widehat{\theta}_T - x_{(1)}^\top\overline{\theta}_T \leq -\frac{\Delta_{(1)}}{2}\right) = \mathbb{P}\left(\sum_{t=1}^T x_{(1)}^\top\left(A(\lambda_t)^{-1}x_t r_t - \theta_t\right) \leq -\frac{T\Delta_{(1)}}{2}\right).$$

Since $\lambda_t = \frac{\bar{\lambda}_t + \lambda^*}{2}$ and $\lambda \mapsto \|x\|^2_{A(\lambda)^{-1}}$ is convex in λ, to use the Berstein's inequality for martingale differences [Freedman, 1975], we have

$$\left|x_{(1)}^\top\left(A(\lambda_t)^{-1}x_t r_t - \theta_t\right)\right| \leq 2\left\|x_{(1)}\right\|_{A(\lambda^*)^{-1}}\|x_t\|_{A(\lambda^*)^{-1}} + 2 \leq 2d + 2 \leq 4d,$$

$$\mathbb{E}\left[\left(x_{(1)}^\top\left(A(\lambda_t)^{-1}x_t r_t-\theta_t\right)\right)^2 \mid \lambda_t\right]=\left\|x_{(1)}\right\|^2_{A(\lambda_t)^{-1}}\le 2\left\|x_{(1)}\right\|^2_{A(\lambda^*)^{-1}}\le 2d.$$

Therefore, we have

$$\mathbb{P}\left(x_{(1)}^\top\widehat{\theta}_T-x_{(1)}^\top\overline{\theta}_T\le-\frac{\Delta_{(1)}}{2}\right)\le\exp\left(-\frac{T\Delta_{(1)}^2/8}{2d+2d\Delta_{(1)}/3}\right)\le\exp\left(-\frac{3T\Delta_{(1)}^2}{64d}\right).$$

By applying the same inequality to other terms, we have

$$\mathbb{P}\left(J_T\neq(1)\right)\le K\exp\left(-\frac{3T\Delta_{(1)}^2}{64d}\right).$$

□

A.4 Implementation Details and Additional Experiments

In this section, we provide more implementation details and additional experiment results. Experiments are executed through Python 3.10 and paralleled by a Mac M1 Pro chip with 6 cores.

First, we notice that an algorithm for stationary environments usually determines a batch of arms to pull at once during each epoch, while in non-stationary environment, the order of pulling these arms will affect the rewards it will receive. Therefore, when applying stationary algorithms (Peace and OD-LinBAI) into a non-stationary environment, we use a random permutation to determine the order of pulling for each batch of arms.

When implementing P1-RAGE, to be computationally efficient, we update λ_t in the same frequency as P1-Peace, which is summarized in Algorithm 11. We take $m = 15$ for P1-RAGE, which, based on Theorem 2.5.1, is valid as long as $\Delta_{(1)} \ge 2^{-13} \approx 1.22 \times 10^{-4}$. Furthermore, when implementing Peace, for simplicity, we use $\inf_{\lambda\in\triangle_\mathcal{X}}\rho(\mathcal{Z},\lambda)$, defined in equation (A.1), to replace all $\gamma(\mathcal{Z})$ used in Katz-Samuels et al. [2020]. Since the paper of OD-LinBAI does not provide code, we implement it based on the pseudocode in Yang and Tan [2022]. Finally, we use Frank-Wolfe algorithm with stepsize $\frac{1}{2(i+2)}$ in i-th iteration to solve all optimization problems in a form of $\inf_{\lambda\in\triangle_\mathcal{X}}\max_{y\in\mathcal{Y}}\|y\|^2_{A(\lambda)^{-1}}$.

As for code snippets reference, we use part of the code from Katz-Samuels et al. [2020]

to implement the rounding procedure used in Peace[§] and part of the code from Fiez et al. [2019] to generate the base stationary instance for the multivariate testing example.[¶] We also use code from Xu et al. [2018] to preprocess the Yahoo! Webscope dataset.[‖]

A.4.1 Additional Experiments

Here, we provide experiment results on some additional examples to corroborate our theoretical findings.

Malicious non-stationary example Because of the nature of arm elimination, algorithms designed for stationary environment can fail easily in some malicious non-stationary environments. Here, we pick the same $\mathcal{X}$ as Soare et al. [2014]'s stationary benchmark example and set $\omega = 0.5$. Then, we take

$$\theta_t = \begin{cases} \begin{bmatrix} 0 & 1 & 1 & \dots & 1 \end{bmatrix}^\top & \text{for } t = 1, \dots, \frac{T}{3}, \\ \begin{bmatrix} 2 & 0 & 0 & \dots & 0 \end{bmatrix}^\top & \text{for } t = \frac{T}{3} + 1, \dots, T. \end{cases}$$

We can see that the overall best arm is still $x_{(1)} = \mathbf{e}_1$. However, because of the θ_t in the first $1/3$ rounds, algorithms like Peace and OD-LinBAI will eliminate $\mathbf{e}_1$ in its initial phase; on the other hand, our algorithms will be robust to this non-stationarity. Here, we take $T = 10^4$ and the results are shown in right plot of Figure A.1.

Stationary multivariate testing example We also test the performance of these algorithms in multivariate testing example when there is no non-stationarity, i.e. $\theta_t = \theta^*$ for all t. Here, we also take $T = 10^4$ and the results are shown in Figure A.2. We can see that our robust algorithm P1-RAGE again performs better than G-BAI and comparably with Peace.

[§]No license information.

[¶]Under MIT License.

[‖]No license information.

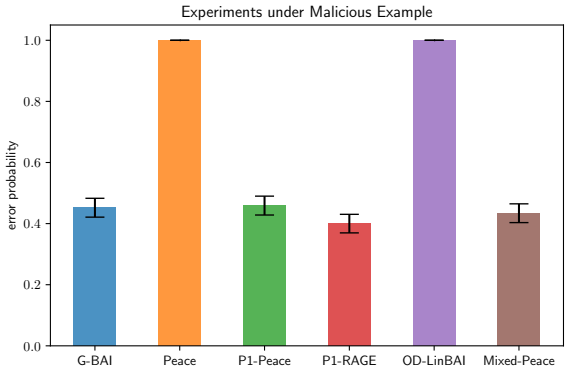


Figure A.1: The error probabilities are estimated through 1000 repeated trials and the error bars represent 95% confidence intervals.

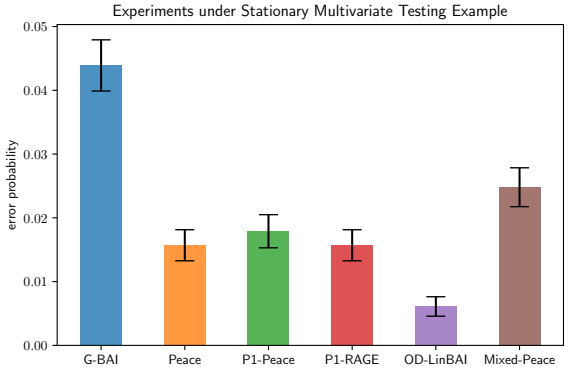


Figure A.2: The error probabilities are estimated through 10^4 repeated trials and the error bars represent 95% confidence intervals.

Non-stationary benchmark example In this example, we add non-stationarity to Soare et al. [2014]'s stationary benchmark example in a more structured instead of malicious way. In particular, we keep the arm set $\mathcal{X}$ the same, take $\omega = 0.5$ and set

$$\theta_t = \left[0.3 \quad 0 \quad 0 \quad \dots \quad -s \sin\left(\tfrac{2\pi t}{L}\right) + 0.5\right]^\top,$$

where s is the oscillation scale and L is the oscillation period, In the first series of instances, we fix $L = 200$ and take values $m \in \{0, 1, \dots, 9\}$; in the second series of instances, we fix $m = 1$ and take values $L \in \{300, 600, \dots, 3000\}$. All non-stationary instances have the same

optimal arm as their stationary counterparts and we take $T = 10^4$ for all of these instances. The results are shown in Figure A.3, from which we can see similar phenomenon as in Figure 2.3. In particular, algorithms designed for stationary environments, Peace and OD-LinBAI, are very unstable in face of non-stationarity. Meanwhile, among the other four relatively robust algorithms, our algorithms P1-RAGE and P1-Peace consistently outperform the other two.

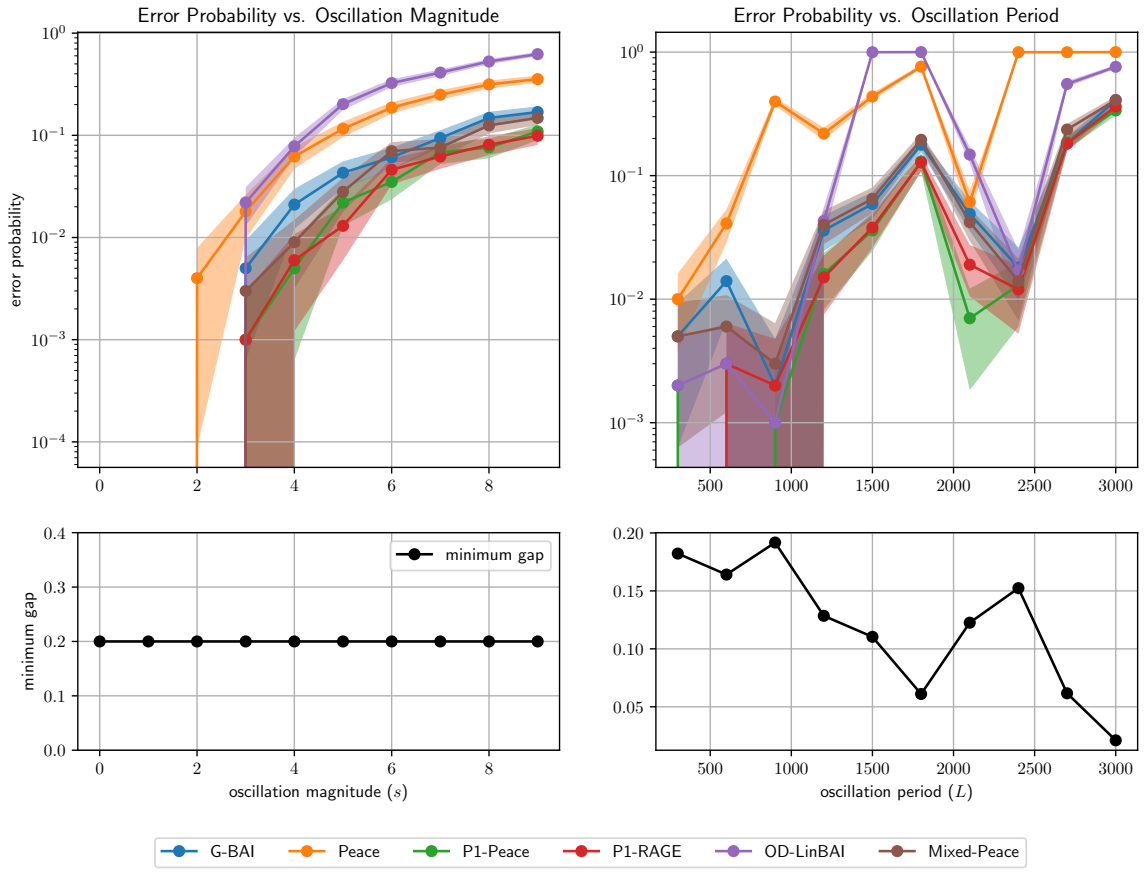


Figure A.3: The vertical axis (error probability) is in log scale. The shaded area represents the 95% confidence interval. Each error probability is estimated through 1000 repeated trials. The bottom two plots give the minimum gap $\Delta_{(1)}$ of each instance that algorithms run over

Appendix B

OMITTED PROOFS IN CHAPTER 3

B.1 Additional Motivating Examples

In this section, we present two additional motivating examples of our proposed models.

Example B.1.1 (Web Advertisements). Consider a set of websites as the facility set and companies who want to advertise their products as the players. Due to budget constraints, each company may only choose some of these websites to put its product ad. For each website, the probability that a user will click on a certain ad (and then buy the product) depends on how many ads are put on the website. If a website receives too many ads, the probability that a user can see a certain ad will decrease, thus making it congested.* The reward each company will receive is measured by the amount of products sold during certain period of time, which is bandit feedback.

Example B.1.2 (Server Usage). Consider a set of servers in a company as the facility set and server users as the players. Each user needs to request several servers to finish her computation task and the cost triggered from each server depends on the number of users requesting that server. Each user will try to minimize the total cost incurred from the servers she requested. As each user can see the cost from all the servers she requested, this is semi-bandit feedback.

B.2 Compute ϵ-approximate Nash Equilibrium in Potential Games

In this section, we show that the ϵ-NASH$(\cdot)$ operation in Algorithm 4 can be computed efficiently by using Algorithm 5.

*Although the website's intelligent recommendation system may more or less mitigate this effect, it can be considered as a part of the reward function's property.

In particular, we first show that the matrix game with reward functions $\overline{Q}_1^k(\cdot), \ldots, \overline{Q}_m^k(\cdot)$ used in Algorithm 4 is a potential game in Lemma B.2.1. Then, we show that Algorithm 5 can efficiently compute an ϵ-approximate Nash equilibrium for potential games and output a product policy as shown in Lemma B.2.2.

Lemma B.2.1. *In line 6 of Algorithm 4, the matrix game with reward functions $\overline{Q}_1^k(\cdot), \ldots, \overline{Q}_m^k(\cdot)$ forms a potential game for both settings of semi-bandit feedback and bandit feedback.*

Proof. In the setting of semi-bandit feedback, since $\overline{Q}_i^k(\boldsymbol{a}) = \sum_{f \in a_i} (\hat{r}^{k,f} + b^{k,f,\mathrm{r}})(\boldsymbol{a})$, the reward functions $\overline{Q}_1^k(\cdot), \ldots, \overline{Q}_m^k(\cdot)$ form a congestion game, which we know is a potential game [Monderer and Shapley, 1996].

In the setting of bandit feedback, notice that by defining $\tilde{r}^{k,f}(i) = \widehat{\theta}^k_{i+m(f-1)}$ for $(i,f) \in [m] \times \mathcal{F}$, we can have $\tilde{r}_i^k(\boldsymbol{a}) = \left\langle A_i(\boldsymbol{a}), \widehat{\theta}^k \right\rangle = \sum_{f \in a_i} \tilde{r}^{k,f}(n^f(\boldsymbol{a}))$. Therefore, we claim that the desired potential function is

$$\Phi^k(\boldsymbol{a}) = \widetilde{\Phi}^k(\boldsymbol{a}) + \tilde{b}^{k,\mathrm{r}}(\boldsymbol{a}), \quad \text{where} \quad \widetilde{\Phi}^k(\boldsymbol{a}) = \sum_{f \in \mathcal{F}} \sum_{i=1}^{n^f(\boldsymbol{a})} \tilde{r}^{k,f}(i).$$

To see this, by referring to the definition of potential function in congestion game [Monderer and Shapley, 1996], since $\tilde{r}_i^k(\boldsymbol{a}) = \sum_{f \in a_i} \tilde{r}^{k,f}(n^f(\boldsymbol{a}))$, we have that

$$\widetilde{\Phi}^k(a_i, a_{-i}) - \widetilde{\Phi}^k(a_i', a_{-i}) = \tilde{r}_i(a_i, a_{-i}) - \tilde{r}_i(a_i', a_{-i}).$$

As a result, we have

$$\begin{aligned}
&\Phi^k(a_i, a_{-i}) - \Phi^k(a_i', a_{-i}) \\
=&\left(\tilde{r}_i(a_i, a_{-i}) + \tilde{b}^{k,\mathrm{r}}(a_i, a_{-i})\right) - \left(\tilde{r}_i(a_i', a_{-i}) + \tilde{b}^{k,\mathrm{r}}(a_i', a_{-i})\right) \\
=&\overline{Q}_i^k(a_i, a_{-i}) - \overline{Q}_i^k(a_i', a_{-i}),
\end{aligned}$$

which means that $\overline{Q}_1^k(\cdot), \ldots, \overline{Q}_m^k(\cdot)$ form a potential game. □

Lemma B.2.2. *Algorithm 5 can output an ϵ-approximate Nash equilibrium.*

Proof. Note that if at round k, we have $\max_{i\in[m]}\Delta_i \le \epsilon$, then π^k is an ϵ-approximate Nash equilibrium. So we only need to prove that $\max_{i\in[m]}\Delta_i \le \epsilon$ is satisfied at some round $k \in \{1,\dots,\lceil\frac{mr_{\max}}{\epsilon}\rceil\}$.

Suppose the potential game $(\{\mathcal{A}_i\}_{i=1}^m, \{r_i\}_{i=1}^m)$ is associated with potential function $\Phi \in [0,\Phi_{\max}]$. Set $\pi^* = \arg\max_{\pi\in\prod_{i\in[m]}\Delta(\mathcal{A}_i)}\Phi(\pi)$. Then for any $\pi \in \prod_{i\in[m]}\Delta(\mathcal{A}_i)$, we have

$$\begin{aligned}\Phi(\pi^*)-\Phi(\pi) =& \sum_{i\in[m]}\big(\Phi(\pi^*_{1:i},\pi_{i+1:m})-\Phi(\pi^*_{1:i-1},\pi_{i:m})\big)\\ =& \sum_{i\in[m]}\left(V_i^{\pi^*_{1:i},\pi_{i+1:m}}-V_i^{\pi^*_{1:i-1},\pi_{i:m}}\right)\\ \le& mr_{\max}.\end{aligned}$$

As a result, we can set $\Phi_{\max} = mr_{\max}$. On the other hand, if $j = \arg\max_{i\in[m]}\Delta_i$ for round k, we have

$$\begin{aligned}\Phi(\pi^{k+1})-\Phi(\pi^k) =& \Phi(\pi_j^{k+1},\pi_{-j}^k)-\Phi(\pi^k)\\ =& V_j^{\pi_j^{k+1},\pi_{-j}^k}-V_j^{\pi^k}\\ =& r_j(a_j^{k+1},\pi_{-j}^k)-r_j(\pi^k) && (\pi^k \text{ is deterministic})\\ =& \Delta_j\\ =& \max_{i\in[m]}\Delta_i.\end{aligned}$$

So there must exist $k \in \{1,\dots,\lceil\frac{mr_{\max}}{\epsilon}\rceil\}$ such that $\max_{i\in[m]}\Delta_i \le \epsilon$, otherwise $\Phi(\pi^k)$ increase at least ϵ at each round, which contradicts $\Phi \in [0, mr_{\max}]$. □

B.3 Analysis for Algorithm 4

Recall that the update rule in Algorithm 4 is $\overline{Q}_i^k(\boldsymbol{a}) = \hat{r}_i^k(\boldsymbol{a}) + b_i^{k,\mathrm{r}}(\boldsymbol{a})$, where we have

$$b_i^{k,\mathrm{r}}(\boldsymbol{a}) = \sum_{f\in a_i} b^{k,f,\mathrm{r}}(\boldsymbol{a}), \quad \text{and} \quad b^{k,f,\mathrm{r}}(\boldsymbol{a}) = \sqrt{\frac{\tilde{\iota}}{N^{k,f}(n^f(\boldsymbol{a}))\vee 1}}.$$

For proof convenience, we define auxiliary value functions

$$\underline{Q}_i^k(\boldsymbol{a}) = \hat{r}_i^k(\boldsymbol{a}) - b_i^{k,\mathrm{r}}(\boldsymbol{a}),$$
$$\overline{V}_i^k = \mathbb{E}_{\boldsymbol{a}\sim\pi^k}[\overline{Q}_i^k(\boldsymbol{a})] \quad \text{and} \quad \underline{V}_i^k = \mathbb{E}_{\boldsymbol{a}\sim\pi^k}[\underline{Q}_i^k(\boldsymbol{a})].$$

With these definitions, we now begin to prove Theorem 3.4.1.

Proof of Theorem 3.4.1. **Semi-bandit Feedback.** By the update rules in Algorithm 4, in the setting of semi-bandit feedback, with probability at least $1-\delta$, simultaneously for all $(k, i, \boldsymbol{a}) \in [K] \times [m] \times \mathcal{A}$, we have

$$\overline{Q}_i^k(\boldsymbol{a}) - r_i(\boldsymbol{a}) = \sum_{f\in a_i} \left[(\hat{r}^{k,f} - r^f)(\boldsymbol{a}) + b^{k,f,\mathrm{r}}(\boldsymbol{a})\right] \geq 0.$$

The second inequality above is obtained by using standard Hoeffding's inequality and union bound, Therefore, we have $\overline{Q}_i^k(\boldsymbol{a}) \geq r_i(\boldsymbol{a})$.

Then, since π^k is the ϵ-approximate Nash equilibrium policy of $\overline{Q}_1^k, \ldots, \overline{Q}_m^k$, we have

$$\begin{aligned}\overline{V}_i^k = \mathbb{E}_{\boldsymbol{a}\sim\pi^k}[\overline{Q}_i^k(\boldsymbol{a})] &= \max_{\nu\in\Delta(\mathcal{A}_i)} \mathbb{E}_{\boldsymbol{a}\sim(\nu,\pi_{-i}^k)}[\overline{Q}_i^k(\boldsymbol{a})] - \epsilon \\ &\geq \max_{\nu\in\Delta(\mathcal{A}_i)} \mathbb{E}_{\boldsymbol{a}\sim(\nu,\pi_{-i}^k)}[r_i(\boldsymbol{a})] - \epsilon = V_i^{\dagger,\pi_{-i}^k} - \epsilon.\end{aligned}$$

Meanwhile, by definition of $\underline{Q}_i^k(\boldsymbol{a})$ and $\underline{V}_i^k$, we can similarly show that $\underline{Q}_i^k(\boldsymbol{a}) \leq r_i(\boldsymbol{a})$ and $\underline{V}_i^k \leq V_i^{\pi^k}$. Therefore, we can have $V_i^{\dagger,\pi_{-i}^k} - V_i^{\pi^k} \leq \overline{V}_i^k - \underline{V}_i^k + \epsilon$.

Now, we define $\widetilde{Q}^k(\boldsymbol{a}) = \max_{i\in[m]} 2b_i^{k,\mathrm{r}}(\boldsymbol{a})$ and $\widetilde{V}^k = \mathbb{E}_{\boldsymbol{a}\sim\pi^k}[\widetilde{Q}^k(\boldsymbol{a})]$. Then, we can notice that

$$\max_{i\in[m]}(\overline{Q}_i^k - \underline{Q}_i^k)(\boldsymbol{a}) \leq \max_{i\in[m]} 2b_i^{k,\mathrm{r}}(\boldsymbol{a}) = \widetilde{Q}^k(\boldsymbol{a}),$$

$$\max_{i\in[m]}(\overline{V}_i^k - \underline{V}_i^k) \leq \mathbb{E}_{\boldsymbol{a}\sim\pi^k}\left[\max_{i\in[m]}(\overline{Q}_i^k - \underline{Q}_i^k)(\boldsymbol{a})\right] \leq \mathbb{E}_{\boldsymbol{a}\sim\pi^k}[\widetilde{Q}^k(\boldsymbol{a})] = \widetilde{V}^k.$$

We further define $\mathcal{M}^k = \mathbb{E}_{\boldsymbol{a}\sim\pi^k}\left[\widetilde{Q}^k(\boldsymbol{a})\right] - \widetilde{Q}^k(\boldsymbol{a}^k) = \widetilde{V}^k - \widetilde{Q}^k(\boldsymbol{a}^k)$. It is not hard to verify that $\mathcal{M}^k$ is a martingale difference sequence with respect to the history from episode 1 to $k-1$. Meanwhile, since $\left|b^{k,\mathrm{r}}(\boldsymbol{a})\right| = \sum_{f\in\mathcal{F}} \sqrt{\frac{\tilde{\iota}}{N^{k,f}(n^f(\boldsymbol{a}))\vee 1}} \leq F\sqrt{\tilde{\iota}}$. Thus, by Azuma-

Hoeffding inequality, we have $\sum_{k=1}^{K} \mathcal{M}^k = \widetilde{\mathcal{O}}\left(F\sqrt{K}\right)$. Therefore, we have

$$
\begin{aligned}
\text{Nash-Regret}(K) &= \sum_{k=1}^{K} \max_{i\in[m]} \left(V_i^{\dagger,\pi_{-i}^k} - V_i^{\pi^k}\right) \\
&= \sum_{k=1}^{K} \min\left\{\max_{i\in[m]} \left(V_i^{\dagger,\pi_{-i}^k} - V_i^{\pi^k}\right), F\right\} \\
&\qquad\qquad \text{(Since the value is always bounded by } F.\text{)} \\
&\le \sum_{k=1}^{K} \min\left\{\max_{i\in[m]} \left(\overline{V}_i^k - \underline{V}_i^k\right), F\right\} + K\epsilon \\
&\le \sum_{k=1}^{K} \min\left\{\widetilde{V}^k, F\right\} + K\epsilon \\
&= \sum_{k=1}^{K} \left(\min\left\{\widetilde{Q}^k(\boldsymbol{a}^k), F\right\} + \mathcal{M}^k\right) + K\epsilon \\
&\le \widetilde{\mathcal{O}}\left(F\sqrt{K}\right) + 2\sum_{k=1}^{K} \left\{\max_{i\in[m]} b_i^{k,\mathrm{r}}(\boldsymbol{a}^k), F\right\} \qquad \text{(By taking } \epsilon = 1/K.\text{)} \\
&\le \widetilde{\mathcal{O}}\left(F\sqrt{K}\right) + 2\sum_{f\in\mathcal{F}}\sum_{k=1}^{K} \sqrt{\frac{\tilde{\iota}}{N^{k,f}(n^f(\boldsymbol{a}^k)) \vee 1}} \\
&\le \widetilde{\mathcal{O}}\left(F\sqrt{mK}\right) \qquad \text{(By Lemma B.3.4.)}
\end{aligned}
$$

Bandit Feedback. By using Lemma B.3.1, which guarantees optimistic estimation, we can similarly show that

$$
\text{Nash-Regret}(K) \le \sum_{k=1}^{K} \mathcal{M}^k + \sum_{k=1}^{K} \min\left\{2\tilde{b}^{k,\mathrm{r}}(\boldsymbol{a}^k), F\right\} + K\epsilon.
$$

To have an upper bound on $\mathcal{M}^k$ here, recall that $\tilde{b}^{k,\mathrm{r}}(\boldsymbol{a}) = \max_{i\in[m]} \|A_i(\boldsymbol{a})\|_{(V^k)^{-1}} \sqrt{\tilde{\beta}_k}$ and $\sqrt{\tilde{\beta}_K} = \widetilde{\mathcal{O}}\left(\sqrt{F\tilde{d}}\right) = \widetilde{\mathcal{O}}\left(F\sqrt{m}\right)$. Meanwhile, we have $\|A_i(\boldsymbol{a})\|_{(V^k)^{-1}} \le \|A_i(\boldsymbol{a})\|_I = \|A_i(\boldsymbol{a})\|_2 \le \sqrt{F}$. Thus, we have $\left|\mathcal{M}^k\right| \le \widetilde{\mathcal{O}}\left(\sqrt{mF^3}\right)$, which by Azuma-Hoeffding inequality implies $\sum_{k=1}^{K} \mathcal{M}^k = \widetilde{\mathcal{O}}\left(\sqrt{mF^3K}\right)$.

Then the sum of the bonus terms can be bounded by using Lemma B.3.2. In particular,

with $\epsilon = 1/K$, we have

$$\begin{aligned}
\text{Nash-Regret}(K) \leq & \widetilde{\mathcal{O}}\left(\sqrt{mF^3K}\right) + 2\sum_{k=1}^{K} \min\left\{\max_{i\in[m]} \left\|A_i(\boldsymbol{a}^k)\right\|_{(V^k)^{-1}} \sqrt{\tilde{\beta}_k}, F\right\} \\
\leq & \widetilde{\mathcal{O}}\left(\sqrt{mF^3K}\right) + 2\sqrt{K\sum_{k=1}^{K} \min\left\{\max_{i\in[m]} \left\|A_i(\boldsymbol{a}^k)\right\|^2_{(V^k)^{-1}} \tilde{\beta}_k, F^2\right\}} \\
\leq & \widetilde{\mathcal{O}}\left(\sqrt{mF^3K}\right) + \sqrt{\widetilde{\mathcal{O}}\left(mF^2K\right)\sum_{k=1}^{K} \min\left\{\max_{i\in[m]} \left\|A_i(\boldsymbol{a}^k)\right\|^2_{(V^k)^{-1}}, 1\right\}} \\
& \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{(Since } \tilde{\beta}_k = \widetilde{\mathcal{O}}\left(mF^2\right).\text{)} \\
\leq & \widetilde{\mathcal{O}}\left(\sqrt{mF^3K}\right) + \widetilde{\mathcal{O}}\left(\sqrt{mF^2K \cdot mF}\right) \qquad\qquad \text{(By Lemma B.3.2.)} \\
\leq & \widetilde{\mathcal{O}}\left(mF^{3/2}\sqrt{K}\right).
\end{aligned}$$

□

B.3.1 Lemmas for Bandit Feedback

The following lemma, as a direct corollary of the confidence bound for least square estimators, shows that the reward estimation error can be bounded by the reward bonus term.

Lemma B.3.1. *With probability at least $1-\delta$, simultaneously for all $(i,k,\boldsymbol{a})$, it holds that $|(\tilde{r}_i^k - r_i)(\boldsymbol{a})| \leq \tilde{b}^{k,\mathrm{r}}(\boldsymbol{a})$, where $\tilde{r}_i^k$ and $\tilde{b}^{k,\mathrm{r}}$ are defined in (3.2).*

Proof. By construction, we have

$$\begin{aligned}
|(\tilde{r}_i^k - r_i)(\boldsymbol{a})| &= \left|\left\langle A_i(\boldsymbol{a}), \widehat{\theta} - \theta \right\rangle\right| \\
&\leq \|A_i(\boldsymbol{a})\|_{(V^k)^{-1}} \left\|\widehat{\theta} - \theta\right\|_{V^k} \\
&\overset{\text{(i)}}{\leq} \|A_i(\boldsymbol{a})\|_{(V^k)^{-1}} \left(\|\theta\|_2 + \sqrt{F\log\left(\det(V^k)\right) + F\tilde{\iota}}\right),
\end{aligned}$$

where the inequality (i) above holds because of Theorem 20.5 in Lattimore and Szepesvári [2020] and the fact that the reward noise is $\sqrt{F}$-subGaussian. Since each element in θ is bounded in $[0,1]$ by construction, we have $\|\theta\|_2 \leq \sqrt{\tilde{d}}$.

Then, by Lemma B.3.2, we have $\det\left(V^k\right) \leq \left(1+\frac{mkF}{\tilde{d}}\right)^{\tilde{d}}$ since by construction $\|A_i(\boldsymbol{a})\|_2^2 \leq F$.

Finally, to make this bound valid for all player $i \in [m]$, we only need to take maximization over $i \in [m]$. Therefore, with probability at least $1-\delta$, we have

$$|(\tilde{r}_i^k - r_i)(\boldsymbol{a})| \leq \max_{i\in[m]} \|A_i(\boldsymbol{a})\|_{(V^k)^{-1}} \sqrt{\tilde{\beta}_k} = \tilde{b}^{k,\mathrm{r}}(\boldsymbol{a}),$$

where $\sqrt{\tilde{\beta}_k} = \sqrt{\tilde{d}} + \sqrt{F\tilde{d}\log\left(1+\frac{mkF}{\tilde{d}}\right) + F\tilde{\iota}}$. □

The following is a variant of the famous elliptical potential lemma, which helps bound the sum of reward bonus under bandit feedback. Here, we apply some techniques from the proof of Lemma 19.4 in Lattimore and Szepesvári [2020].

Lemma B.3.2. *Let $K, m \geq 1$ be integers. Suppose $V^k = I + \sum_{k'=1}^{k-1}\sum_{i=1}^{m} A_i^{k'}\left(A_i^{k'}\right)^\top$, where $A_i^{k'} \in \mathbb{R}^d$ and $\left\|A_i^{k'}\right\|_2^2 \leq F$. Then, it holds that*

$$\det\left(V^k\right) \leq \left(1+\frac{mkF}{d}\right)^d, \quad \textit{and} \quad \sum_{k=1}^{K} \min\left\{\max_{i\in[m]} \left\|A_i^k\right\|_{(V^k)^{-1}}^2, 1\right\} \leq 2d\log\left(1+\frac{mKF}{d}\right).$$

Proof. For the first upper bound about $\det\left(V^k\right)$, we have

$$\begin{aligned}
\det\left(V^k\right) &= \prod_{j=1}^{d} \lambda_j && (\lambda_1, \ldots, \lambda_d \text{ are eigenvalues of } V^k) \\
&\leq \left(\frac{\mathrm{tr}\left(V^k\right)}{d}\right)^d && (\text{By AM-GM inequality}) \\
&= \left(\frac{\mathrm{tr}\left(I\right) + \sum_{k'=1}^{k-1}\sum_{i=1}^{m}\left\|A_i^{k'}\right\|_2^2}{d}\right)^d \\
&\leq \left(1+\frac{mkF}{d}\right)^d. && (\text{Since } \left\|A_i^{k'}\right\|_2^2 \leq F.)
\end{aligned}$$

For the second upper bound. First, we notice that $\min\{1, x\} \leq 2\log(1+x)$ for any

$x \geq 0$. Thus, we have

$$\sum_{k=1}^{K} \min\left\{1, \max_{i\in[m]} \left\|A_i^k\right\|^2_{(V^k)^{-1}}\right\} \leq 2\sum_{k=1}^{K} \log\left(1 + \max_{i\in[m]} \left\|A_i^k\right\|^2_{(V^k)^{-1}}\right).$$

Then, for $k \geq 2$, we can notice that

$$\begin{aligned}
V^k =& V^{k-1} + \sum_{i=1}^{m} A_i^{k-1}\left(A_i^{k-1}\right)^\top \\
=& \left(V^{k-1}\right)^{1/2}\left(I + \left(V^{k-1}\right)^{-1/2}\left(\sum_{i=1}^{m} A_i^{k-1}\left(A_i^{k-1}\right)^\top\right)\left(V^{k-1}\right)^{-1/2}\right)\left(V^{k-1}\right)^{1/2} \\
=& \left(V^{k-1}\right)^{1/2}\left(I + \sum_{i=1}^{m}\left(\left(V^{k-1}\right)^{-1/2} A_i^{k-1}\right)\left(\left(V^{k-1}\right)^{-1/2} A_i^{k-1}\right)^\top\right)\left(V^{k-1}\right)^{1/2}.
\end{aligned}$$

Therefore, we have

$$\begin{aligned}
\det\left(V^k\right) =& \det\left(V^{k-1}\right)\det\left(I + \sum_{i=1}^{m}\left(\left(V^{k-1}\right)^{-1/2} A_i^{k-1}\right)\left(\left(V^{k-1}\right)^{-1/2} A_i^{k-1}\right)^\top\right) \\
\geq& \det\left(V^{k-1}\right)\left(1 + \max_{i\in[m]} \left\|A_i^{k-1}\right\|^2_{(V^{k-1})^{-1}}\right) && \text{(By Lemma B.3.3.)} \\
\geq& \prod_{k'=1}^{k-1}\left(1 + \max_{i\in[m]} \left\|A_i^{k'}\right\|^2_{(V^{k'})^{-1}}\right). && \text{(Since by definition, } V^1 = I.\text{)}
\end{aligned}$$

As a result, we have

$$\begin{aligned}
\sum_{k=1}^{K} \min\left\{\max_{i\in[m]} \left\|A_i^k\right\|^2_{(V^k)^{-1}}, 1\right\} \leq& 2\sum_{k=1}^{K} \log\left(1 + \max_{i\in[m]} \left\|A_i^k\right\|^2_{(V^k)^{-1}}\right) \\
\leq& 2\log\left(\det\left(V^{K+1}\right)\right) \\
\leq& 2d\log\left(1 + \frac{mKF}{d}\right).
\end{aligned}$$

□

B.3.2 Technical Lemmas

Lemma B.3.3. *Let $y_1, \ldots, y_m \in \mathbb{R}^d$ be a set of vectors. Then, it holds that*

$$\det\left(I + \sum_{i=1}^{m} y_i y_i^\top\right) \geq 1 + \max_{i \in [m]} \|y_i\|_2^2 .$$

Proof. Since $I + \sum_{i=1}^m y_i y_i^\top \succeq I + y_i y_i^\top$ for any $i \in [m]$, we have $\det\left(I + \sum_{i=1}^m y_i y_i^\top\right) \geq \det\left(I + y_i y_i^\top\right)$ for any $i \in [m]$. That is, we have

$$\det\left(I + \sum_{i=1}^{m} y_i y_i^\top\right) \geq \max_{i \in [m]} \det\left(I + y_i y_i^\top\right) = 1 + \max_{i \in [m]} \|y_i\|_2^2 .$$

The last line above holds because the matrix $I + y_i y_i^\top$ has eigenvalues $1 + \|y_i\|_2^2$ and 1. □

Lemma B.3.4. *For any $f \in \mathcal{F}$, it holds that*

$$\sum_{k=1}^{K} \sqrt{\frac{1}{\mathbb{N}^{k,f}(n^f(\boldsymbol{a}^k)) \vee 1}} \leq \widetilde{\mathcal{O}}\left(\sqrt{mK}\right).$$

Proof. Here, we have

$$\begin{aligned}
\sum_{k=1}^{K} \sqrt{\frac{1}{N^{k,f}(n^f(\boldsymbol{a}^k)) \vee 1}} =& \sum_{n=0}^{m} \sum_{\ell=1}^{N^{K,f}(n)} \sqrt{\frac{1}{\ell}} \\
\leq& 2 \sum_{n=0}^{m} \sqrt{N^{K,f}(n)} && \text{(By standard technique)} \\
\leq& 2 \sqrt{(m+1) \sum_{n=0}^{m} N^{K,f}(n)} \\
=& \widetilde{\mathcal{O}}\left(\sqrt{mK}\right).
\end{aligned}$$

The last equality above is based on a pigeon-hold principle argument similar to Lemma B.6.5. □

B.4 Analysis for Algorithm 6

B.4.1 Exploration Distribution and Smoothness

We choose the exploration distribution to be the G-optimal design and we have the following properties.

Lemma B.4.1. *(Unbiasedness) For any episode $k \in [K]$, $i \in [m]$ and $a \in \mathcal{A}_i$, we have*

$$\mathbb{E}_k\left[\widehat{\nabla}_i^k \Phi(a)\right] = \nabla_i^k \Phi(a),$$

where $\mathbb{E}_k[\cdot]$ is taken over all the randomness before episode k.

Proof. By the definition of $\widehat{\nabla}_i^k \Phi(a)$, we have

$$\begin{aligned}
\mathbb{E}_k\left[\widehat{\nabla}_i^k \Phi(a)\right] =& \mathbb{E}_k\left\langle \phi_i(a), \widehat{\theta}_i^k(\pi^k)\right\rangle \\
=& \mathbb{E}_k\left[\frac{1}{\tau}\sum_{t=1}^{\tau} \phi_i(a)^\top [\Sigma_i^k]^{-1} \phi_i(a_i^{k,t}) r_i^{k,t}\right] \\
=& \mathbb{E}_k\left[\phi_i(a)^\top [\Sigma_i^k]^{-1} \phi_i(a_i^{k,1}) r_i^{k,1}\right] \\
=& \mathbb{E}_k\left[\phi_i(a)^\top [\Sigma_i^k]^{-1} \phi_i(a_i^{k,1}) \phi_i(a_i^{k,1})^\top \theta_i^{k,1}(\pi^k)\right] \\
=& \sum_{a_i^k \in \mathcal{A}_i} \pi_i^k(a_i^{k,1}) \phi_i^\top(a) [\Sigma_i^k]^{-1} \phi_i(a_i^{k,1}) \phi_i(a_i^{k,1})^\top \theta_i(\pi^k) \\
& \qquad (a_i^{k,1} \text{ only depends on } \pi_i^k \text{ and } \theta_i^{k,1}(\pi^k) \text{ only depends on } \pi_{-i}^k) \\
=& \phi_i^\top(a) [\Sigma_i^k]^{-1} \left[\sum_{a_i^k \in \mathcal{A}_i} \pi_i^k(a_i^{k,1}) \phi_i(a_i^{k,1}) \phi_i(a_i^{k,1})^\top\right] \theta_i(\pi^k) \\
=& \phi_i^\top(a_i) \theta_i(\pi^k) \\
=& \nabla_i^k \Phi(a).
\end{aligned}$$

□

Lemma B.4.2. *For any episode $k \in [K]$, $i \in [m]$ and $a \in \mathcal{A}_i$, we have*

$$\left|\phi_i(a)^\top [\Sigma_i^k]^{-1} \phi_i(a_i^{k,t}) r_i^{k,t}\right| \le \frac{F^2}{\gamma}.$$

Proof. As $\pi_i^k = (1-\gamma)(\nu\widetilde{\pi}_i^k + (1-\gamma)\pi_i^{k-1}) + \gamma\rho_i$, we have

$$\Sigma_i^k = \mathbb{E}_{a_i\sim\pi_i^k}\phi_i(a_i)\phi_i(a_i)^\top \succeq \gamma\mathbb{E}_{a_i\sim\rho_i}\phi_i(a_i)\phi_i(a_i)^\top,$$

and ρ_i is the G-optimal design with respect to $\phi_i(\cdot)$, for any action $a \in \mathcal{A}_i$ we have

$$\|\phi_i(a)\|^2_{[\Sigma_i^k]^{-1}} \leq \frac{1}{\gamma}\|\phi_i(a)\|^2_{[\mathbb{E}_{a_i\sim\rho_i}\phi_i(a_i)\phi_i(a_i)^\top]^{-1}} \leq \frac{F}{\gamma}.$$

Then for any $t \in [\tau]$, since $|r_i^{k,t}| \leq F$, we have

$$\left|r_i^{k,t}\phi_i^\top(a)[\Sigma_i^k]^{-1}\phi_i(a_i^{k,t})\right| \leq \left|r_i^{k,t}\right| \|\phi_i(a)\|_{[\Sigma_i^k]^{-1}} \left\|\phi_i(a_i^{k,t})\right\|_{[\Sigma_i^k]^{-1}} \leq \frac{F^2}{\gamma}.$$

As a result, we have

$$\left|\widehat{\nabla}_i^k\Phi(a)\right| = \left|\frac{1}{\tau}\sum_{t=1}^{\tau}\phi_i(a)^\top[\Sigma_i^k]^{-1}\phi_i(a_i^{k,t})r_i^{k,t}\right| \leq \frac{F^2}{\gamma}$$

□

Lemma B.4.3. *For any episode $k \in [K]$, $i \in [m]$ and $a \in \mathcal{A}_i$, we have*

$$\mathbb{E}_k\left[\left(\phi_i(a)^\top[\Sigma_i^k]^{-1}\phi_i(a_i^{k,t})r_i^{k,t}\right)^2\right] \leq \frac{F^3}{\gamma}.$$

Proof. We first show that for any $t \in [\tau]$, we have

$$\begin{aligned}
&\mathbb{E}_k\left[\left(\phi_i(a)^\top[\Sigma_i^k]^{-1}\phi_i(a_i^{k,t})r_i^{k,t}\right)^2\right] \\
\leq& F^2\mathbb{E}_k\left[\left(\phi_i(a)^\top[\Sigma_i^k]^{-1}\phi_i(a_i^{k,t})\right)^2\right] \\
\leq& F^2\mathbb{E}_k\left[\phi_i(a)^\top[\Sigma_i^k]^{-1}\phi_i(a_i^{k,t})\phi_i(a_i^{k,t})^\top[\Sigma_i^k]^{-1}\phi_i(a)^\top\right] \\
=& F^2\phi_i(a)^\top[\Sigma_i^k]^{-1}\phi_i(a) \\
\leq& \frac{F^3}{\gamma}.
\end{aligned}$$

□

Lemma B.4.4. *With probability $1-\delta$, for all $k \in [K]$, $i \in [m]$ and $a \in \mathcal{A}_i$, we have*

$$\left|\widehat{\nabla}_i^k \Phi(a) - \nabla_i^k \Phi(a)\right| \le c\sqrt{\frac{F^4 \log(mK/\delta)}{\gamma\tau}} + \frac{cF^3 \log(mK/\delta)}{\gamma\tau}$$

Proof. Recall that

$$\widehat{\nabla}_i^k \Phi(a_i) = \frac{1}{\tau}\sum_{t=1}^{\tau} \phi_i^\top(a_i)[\Sigma_i^k]^{-1} r_i^{k,t} \phi_i(a_i^{k,t}),$$

and $(a_i^{k,t}, r_i^{k,t})$ are drawn independently at each $t \in [\tau]$. Lemma B.4.1 shows that $\widehat{\nabla}_i^k \Phi(a_i)$ is an unbiased estimate of $\nabla_i^k \Phi(a_i)$ In addition, Lemma B.4.2 shows that $\phi_i^\top(a_i)[\Sigma_i^k]^{-1} r_i^{k,t} \phi_i(a_i^{k,t})$ is bounded by F^2/γ and Lemma B.4.3 shows that its second moment is bounded by F^3/γ. Then by Bernstein's inequality, for a fixed $k \in [K]$, $i \in [m]$ and $a \in \mathcal{A}_i$, with probability $1-\delta$, we have

$$\left|\widehat{\nabla}_i^k \Phi(a) - \nabla_i^k \Phi(a)\right| \le \sqrt{\frac{2F^3 \log(2/\delta)}{\gamma\tau}} + \frac{3F^2 \log(2/\delta)}{2\gamma\tau}.$$

The argument holds by applying the union bound and the fact that $|\mathcal{A}_i| \le 2^F$.

□

Lemma B.4.5. *$\Phi(\cdot)$ is mF-Lipschitz and mF-smooth with respect to the L1 norm $\|\cdot\|_1$.*

Proof. Recall that $\Phi(\pi) = \mathbb{E}_{\boldsymbol{a}\sim\pi}\Phi(\boldsymbol{a})$ and $\Phi(\boldsymbol{a}) \in [0, mF]$.

$$\begin{aligned}
\Phi(\pi) - \Phi(\pi') =& \mathbb{E}_{\boldsymbol{a}\sim\pi}\Phi(\boldsymbol{a}) - \mathbb{E}_{\boldsymbol{a}\sim\pi'}\Phi(\boldsymbol{a}) \\
=& \sum_{i\in[m]} \mathbb{E}_{a_{1:i-1}\sim\pi'_{1:i-1}, a_{i:m}\sim\pi_{i:m}}\Phi(\boldsymbol{a}) - \mathbb{E}_{a_{1:i}\sim\pi'_{1:i}, a_{i+1:m}\sim\pi_{i+1:m}}\Phi(\boldsymbol{a}) \\
\le& \sum_{i\in[m]} \left\|\pi_i - \pi'_i\right\|_1 \cdot \|\Phi\|_\infty \\
\le& mF \left\|\pi - \pi'\right\|_1.
\end{aligned}$$

Similarly we have $\nabla_\pi \Phi(a_i) = \mathbb{E}_{a_{-i}\sim\pi_{-i}}\Phi(a_i, a_{-i})$. As a result, we have

$$\|\nabla_\pi \Phi - \nabla_{\pi'}\Phi\|_\infty \le mF \left\|\pi - \pi'\right\|_1.$$

□

Definition B.4.6. (Frank Wolfe Gap) The Frank Wolfe gap of a joint strategy π for $\Phi(\cdot)$ is defined as

$$G(\pi) = \max_{\pi'} \left\langle \pi' - \pi, \nabla_\pi \Phi \right\rangle.$$

Lemma B.4.7. *Suppose the Frank Wolfe gap of π is ϵ. Then π is an ϵ-Nash policy.*

Proof. For a fixed player i, suppose player i change her strategy to π_i'.

$$\begin{aligned} V_i^{\pi_i', \pi_{-i}} - V_i^{\pi} &= \Phi(\pi_i', \pi_{-i}) - \Phi(\pi) \\ &= \left\langle \pi_i' - \pi_i, \nabla_{\pi_i} \Phi \right\rangle \\ &\leq \max_{\pi'} \left\langle \pi' - \pi, \nabla_\pi \Phi \right\rangle \\ &\leq \epsilon. \end{aligned}$$

□

B.4.2 Analysis for Frank Wolfe in Bandit Feedback

Theorem B.4.8. *Let $T = K\tau$. For the congestion game with bandit feedback, by running Algorithm 6 with gradient estimator $\widehat{\nabla}_i^k \Phi$ in (3.4) and exploration distribution ρ_i in (3.5), setting parameters $\nu = \frac{F}{m\sqrt{K}}$, $\gamma = \frac{F}{mK}$ and $\tau = K^2$, if $K \geq \frac{2F}{m}$, then with probability $1-\delta$, we have*

$$\textit{Nash-Regret}(T) = \tau \sum_{k=1}^{K} G(\pi^k) = \widetilde{\mathcal{O}}\left(m^2 F^2 T^{5/6} + m^3 F^3 T^{2/3}\right).$$

Proof. Set $\nabla^k \Phi = \nabla \Phi(\Pi^k) \in \mathbb{R}^A$ and $\nabla_i^k \Phi = \nabla^k \Phi(\pi_i) \in \mathbb{R}^{A_i}$. As we have $\Phi(\cdot)$ is mF-smooth w.r.t. $\|\cdot\|_1$, we have

$$\begin{aligned} \Phi(\pi^{k+1}) \geq & \Phi(\pi^k) + \left\langle \nabla \Phi(\pi^k), \pi^{k+1} - \pi^k \right\rangle - \frac{mF}{2} \|\pi^{k+1} - \pi^k\|_1^2 \\ = & \Phi(\pi^k) + (1-\gamma)\nu \left\langle \nabla \Phi(\pi^k), \widetilde{\pi}^{k+1} - \pi^k \right\rangle + \gamma \left\langle \nabla^k \Phi, \rho - \pi^k \right\rangle \\ & - \frac{mF}{2} (2\nu^2 \left\| \widetilde{\pi}^k - \pi^k \right\|_1^2 + 2\gamma^2 \left\| \rho - \pi^k \right\|_1^2) \\ \geq & \Phi(\pi^k) + (1-\gamma)\nu \left\langle \nabla \Phi(\pi^k), \widetilde{\pi}^{k+1} - \pi^k \right\rangle - \gamma \left\| \nabla^k \Phi \right\|_\infty \left\| \rho - \pi^k \right\|_1 \end{aligned}$$

$$
\begin{aligned}
&- \frac{mF}{2}(2\nu^2 \left\|\widetilde{\pi}^k - \pi^k\right\|_1^2 + 2\gamma^2 \left\|\rho - \pi^k\right\|_1^2) \\
\geq &\Phi(\pi^k) + (1-\gamma)\nu \left\langle \nabla\Phi(\pi^k), \widetilde{\pi}^{k+1} - \pi^k \right\rangle - 2\gamma m^2 F - 4m^3 F(\nu^2 + \gamma^2).
\end{aligned}
$$

(By Lemma B.4.5.)

Define the true target policy at episode k

$$
\widehat{\pi}_i^{k+1} = \arg\max_{\pi_i} \left\langle \pi_i, \nabla_i \Phi(\pi_i^k) \right\rangle,
$$

and the Frank Wolfe gap of joint strategy π

$$
G(\pi) = \max_{\pi'} \left\langle \pi' - \pi, \nabla\Phi(\pi) \right\rangle.
$$

Then we have

$$
\begin{aligned}
\left\langle \nabla\Phi(\pi^k), \widetilde{\pi}^{k+1} - \pi^k \right\rangle &= \left\langle \widehat{\nabla}^k\Phi(\pi^k), \widetilde{\pi}^{k+1} - \pi^k \right\rangle + \left\langle \nabla\Phi(\pi^k) - \widehat{\nabla}^k\Phi(\pi^k), \widetilde{\pi}^{k+1} - \pi^k \right\rangle \\
&\geq \left\langle \widehat{\nabla}^k\Phi(\pi^k), \widehat{\pi}^{k+1} - \pi^k \right\rangle + \left\langle \nabla\Phi(\pi^k) - \widehat{\nabla}^k\Phi(\pi^k), \widetilde{\pi}^{k+1} - \pi^k \right\rangle \\
&= \left\langle \nabla\Phi(\pi^k), \widehat{\pi}^{k+1} - \pi^k \right\rangle + \left\langle \nabla\Phi(\pi^k) - \widehat{\nabla}^k\Phi(\pi^k), \widetilde{\pi}^{k+1} - \widehat{\pi}^{k+1} \right\rangle \\
&\geq G(\pi^k) - 2m \left\| \nabla\Phi(\pi^k) - \widehat{\nabla}^k\Phi(\pi^k) \right\|_\infty \\
&\geq G(\pi^k) - c\sqrt{\frac{m^2F^4\log(mK/\delta)}{\gamma\tau}} - \frac{cmF^3\log(mK/\delta)}{\gamma\tau}
\end{aligned}
$$

Apply it to the previous bound and we have

$$
\begin{aligned}
\Phi(\pi^{k+1}) \geq &\Phi(\pi^k) + (1-\gamma)\nu G(\pi^k) - c\frac{(1-\gamma)\nu}{\sqrt{\gamma\tau}}\sqrt{m^2F^4\log(mK/\delta)} \\
&- c\frac{(1-\gamma)\nu}{\gamma\tau} mF^3\log(mK/\delta) - \gamma 2m^2F - 4m^3F(\nu^2+\gamma^2).
\end{aligned}
$$

Summing over $k \in [K]$ and we get

$$
\sum_{k=1}^K G(\pi^k) \leq \frac{\Phi(\pi^{K+1}) - \Phi(\pi^1)}{(1-\gamma)\nu} + c\frac{K}{\sqrt{\gamma\tau}}\sqrt{m^2F^4\log(mK/\delta)} + c\frac{K}{\gamma\tau}mF^3\log(mK/\delta)
$$

$$+\frac{2m^2FK\gamma}{(1-\gamma)\nu}+\frac{4(\nu^2+\gamma^2)m^3FK}{(1-\gamma)\nu}.$$

Set $\nu=\frac{F}{m\sqrt{K}}$, $\gamma=\frac{F}{mK}$, $\tau=K^2$ and notice that when $K\geq\frac{2F}{m}$, we have $1-\gamma\geq\frac{1}{2}$. Since $\Phi(\cdot)$ is bounded in $[0,mF]$, we can have

$$\sum_{k=1}^{K}G(\pi^k)=\widetilde{\mathcal{O}}\left(m^2F^2K^{1/2}+m^3F^3\right).$$

Then by Lemma B.4.7, for $T=K\tau$, we have

$$\text{Nash-Regret}(T)=\tau\sum_{k=1}^{K}G(\pi^k)=\widetilde{\mathcal{O}}\left(m^2F^2T^{5/6}+m^3F^3T^{2/3}\right).$$

□

B.4.3 Algorithm and Analysis for Semi-bandit Feedback

In the setting of semi-bandit feedback, we will need a different gradient estimator $\widetilde{\nabla}_i^k\Phi(a_i)$ and a different exploration distribution $\tilde{\rho}_i$ to utilize the extra reward information from each chosen facility.

Based on the analysis in Section 3.5, using (3.3), we have $\nabla_i^k\Phi(a_i)=\sum_{f\in a_i}[\theta_i(\pi^k)]_f$, where $[\theta_i(\pi^k)]_f=\mathbb{E}_{a_{-i}\sim\pi_{-i}^k}\left[r^f(n^f(a_{-i})+1)\right]$. Meanwhile, in semi-bandit feedback, the mean of t-th reward player i received for facility f at episode k is $r^f(n^f(a_i^{k,t},a_{-i}^{k,t}))$. Therefore, we can use inverse propensity score (IPS) estimator to estimate $[\theta_i(\pi^k)]_f$. In particular, we have

$$[\widetilde{\theta}_i^k(\pi^k)]_f=\frac{1}{\tau}\sum_{t=1}^{\tau}[\widetilde{\theta}_i^{k,t}(\pi^k)]_f,\quad\text{where}\quad[\widetilde{\theta}_i^{k,t}(\pi^k)]_f=\frac{r^{k,t,f}\mathbb{1}\left\{f\in a_i^{k,t}\right\}}{\mathbb{P}_{a_i\sim\pi_i^k}(f\in a_i)}.$$

Then, we can naturally have

$$\widetilde{\nabla}_i^k\Phi(a_i)=\sum_{f\in a_i}[\widetilde{\theta}_i^k(\pi^k)]_f. \tag{B.1}$$

Furthermore, by Lemma B.4.11, we can see that by using $\tilde{\rho}_i$ computed by Algorithm 14, for all players, we have $\mathbb{P}_{a_i\sim\pi_i^k}(f\in a_i)\geq\frac{\gamma}{2F}$ for all $f\in\bigcup_{a_i\in\mathcal{A}_i}a_i$.

Properties of the IPS estimator are summarized in Lemma B.4.12. By using these properties, we can have the following lemma.

Lemma B.4.9. *With probability* $1-\delta$*, for all* $k \in [K]$*,* $i \in [m]$ *and* $a_i \in \mathcal{A}_i$*, we have*

$$\left|\widetilde{\nabla}_i^k \Phi(a_i) - \nabla_i^k \Phi(a_i)\right| \leq \sqrt{\frac{4F^3 \log(2mFK/\delta)}{\gamma\tau}} + \frac{2F^2 \log(2mFK/\delta)}{\gamma\tau}.$$

Proof. By Lemma B.4.12 and Bernstein's inequality, simultaneously for all $(i,k,f) \in [m] \times [K] \times \mathcal{F}$, with probability at least $1-\delta$, we have

$$\left|[\widetilde{\theta}_i^k(\pi^k)]_f - [\theta_i(\pi^k)]_f\right| \leq \sqrt{\frac{4F \log(2mFK/\delta)}{\gamma\tau}} + \frac{2F \log(2mFK/\delta)}{\gamma\tau}.$$

Since $\widetilde{\nabla}_i^k \Phi(a_i) = \sum_{f \in a_i} [\widetilde{\theta}_i^k(\pi^k)]_f$, by triangle inequality, we have

$$\left|\widetilde{\nabla}_i^k \Phi(a_i) - \nabla_i^k \Phi(a_i)\right| \leq \sqrt{\frac{4F^3 \log(2mFK/\delta)}{\gamma\tau}} + \frac{2F^2 \log(2mFK/\delta)}{\gamma\tau}.$$

□

With this more refined gradient estimator, we can now have the following theorem.

Theorem B.4.10. *Let* $T = K\tau$*. For the congestion game with semi-bandit feedback, by running Algorithm 6 with gradient estimator* $\widetilde{\nabla}_i^k \Phi$ *in* (B.1) *and exploration distribution* $\tilde{\rho}_i$ *in Algorithm 14, setting parameters* $\nu = \frac{\sqrt{F}}{m\sqrt{K}}$*,* $\gamma = \frac{\sqrt{F}}{mK}$ *and* $\tau = K^2$*, if* $K \geq \frac{2\sqrt{F}}{m}$*, then with probability* $1-\delta$*, we have*

$$\textit{Nash-Regret}(T) = \tau \sum_{k=1}^{K} G(\pi^k) = \widetilde{\mathcal{O}}\left(m^2 F^{3/2} T^{5/6} + m^3 F^2 T^{2/3}\right).$$

Proof. By following the proof of Theorem B.4.8 and applying the concentration inequality in Lemma B.4.9, we can have

$$\Phi(\pi^{k+1}) \geq \Phi(\pi^k) + (1-\gamma)\nu G(\pi^k) - \frac{(1-\gamma)\nu}{\sqrt{\gamma\tau}} \sqrt{4m^2 F^3 \log(2mK/\delta)}$$

$$- \frac{2(1-\gamma)\nu}{\gamma\tau} mF^2 \log(mK/\delta) - \gamma 2m^2 F - 4m^3 F(\nu^2 + \gamma^2).$$

Summing over $k \in [K]$ and we get

$$\sum_{k=1}^{K} G(\pi^k) \leq \frac{\Phi(\pi^{K+1}) - \Phi(\pi^1)}{(1-\gamma)\nu} + \frac{K}{\sqrt{\gamma\tau}} \sqrt{4m^2 F^3 \log(mK/\delta)} + \frac{2K}{\gamma\tau} mF^2 \log(mK/\delta) + \frac{2m^2 FK\gamma}{(1-\gamma)\nu} + \frac{4(\nu^2+\gamma^2)m^3 FK}{(1-\gamma)\nu}.$$

Set $\nu = \frac{\sqrt{F}}{m\sqrt{K}}$, $\gamma = \frac{\sqrt{F}}{mK}$, $\tau = K^2$ and notice that when $K \geq \frac{2\sqrt{F}}{m}$, we have $1 - \gamma \geq \frac{1}{2}$. Thus, we can have

$$\sum_{k=1}^{K} G(\pi^k) = \widetilde{\mathcal{O}}\left(m^2 F^{3/2} K^{1/2} + m^3 F^2\right).$$

Then by Lemma B.4.7, for $T = K\tau$, we have

$$\text{Nash-Regret}(T) = \tau \sum_{k=1}^{K} G(\pi^k) = \widetilde{\mathcal{O}}\left(m^2 F^{3/2} T^{5/6} + m^3 F^2 T^{2/3}\right).$$

□

B.4.4 Lemmas for Semi-bandit Feedback

Algorithm 14 Compute Exploration Distribution $\tilde{\rho}_i$

Input: $\mathcal{A}_i$, player i-th action set
Initialize $\widetilde{\mathcal{A}}_i \leftarrow \emptyset$
for a_i in $\mathcal{A}_i$ **do**
  if $\exists f \in a_i$ such that $f \notin \bigcup_{a_i' \in \widetilde{\mathcal{A}}_i} a_i'$ **then**
    $\widetilde{\mathcal{A}}_i \leftarrow \widetilde{\mathcal{A}}_i \cup \{a_i\}$
  end if
  if $\mathcal{F}_i = \bigcup_{a_i' \in \widetilde{\mathcal{A}}_i} a_i'$ **then**
    break
  end if
end for
Assign $\tilde{\rho}_i(a_i) \leftarrow \frac{1}{2F}$ for each $a_i \in \widetilde{\mathcal{A}}_i$
Assign remaining probability mass arbitrarily to actions in $\mathcal{A} \setminus \widetilde{\mathcal{A}}_i$
return $\tilde{\rho}_i$

Lemma B.4.11. *Let $\mathcal{F}_i = \bigcup_{a_i \in \mathcal{A}_i} a_i$. For any player i, if $\tilde{\rho}_i$ is the output of Algorithm 14 and π_i^k contains a mixture of $\tilde{\rho}_i$ with weight γ, then we have $\mathbb{P}_{a_i \sim \pi_i^k}(f \in a_i) \geq \frac{\gamma}{2F}$ for any $f \in \mathcal{F}_i$.*

Proof. By Algorithm 14, whenever a new action is added into $\widetilde{\mathcal{A}}_i$, it contains facility not appeared in current $\widetilde{\mathcal{A}}_i$. Then, since there are at most $|\mathcal{F}_i| \leq F$ distinct facilities in the action set $\mathcal{A}_i$, the final $\widetilde{\mathcal{A}}_i$ must satisfy $|\widetilde{\mathcal{A}}_i| \leq F$. Therefore, $\tilde{\rho}_i$ is a valid distribution over $\mathcal{A}_i$.

Since π_i^k contains a mixture of $\tilde{\rho}_i$ with weight γ, for any $a_i \in \mathcal{A}_i$, we have $\pi_i^k(a_i) \geq \gamma \tilde{\rho}_i(a_i)$. Thus, we have

$$\begin{aligned}
\mathbb{P}_{a_i \sim \pi_i^k}(f \in a_i) = &\sum_{a_i \in \mathcal{A}_i} \pi_i^k(a_i) \mathbb{1}\{f \in a_i\} \\
\geq &\gamma \sum_{a_i \in \mathcal{A}_i} \tilde{\rho}_i(a_i) \mathbb{1}\{f \in a_i\} \\
\geq &\gamma \sum_{a_i \in \widetilde{\mathcal{A}}_i} \tilde{\rho}_i(a_i) \mathbb{1}\{f \in a_i\} \\
= &\frac{\gamma}{2F} \sum_{a_i \in \widetilde{\mathcal{A}}_i} \mathbb{1}\{f \in a_i\} \geq \frac{\gamma}{2F}.
\end{aligned}$$

The last inequality above holds since by construction, $\widetilde{\mathcal{A}}_i$ contains all facilities contained in $\mathcal{A}_i$.

□

Lemma B.4.12. *If π_i^k contains a mixture of $\tilde{\rho}_i$ given in Algorithm 14 with weight γ. Then, the IPS estimator $[\widetilde{\theta}_i^k(\pi^k)]_f$ satisfies*

$$\mathbb{E}_k\left[[\widetilde{\theta}_i^{k,t}(\pi^k)]_f\right] = [\theta_i(\pi^k)]_f, \quad |[\widetilde{\theta}_i^{k,t}(\pi^k)]_f| \leq \frac{2F}{\gamma}, \quad \text{and} \quad \mathbb{E}_k\left[[\widetilde{\theta}_i^{k,t}(\pi^k)]_f^2\right] \leq \frac{2F}{\gamma}.$$

Proof. For the first property, since $\mathbb{E}_k\left[r^{k,t,f} \mid \boldsymbol{a}^{k,t}\right] = r^f(n^f(a_i^{k,t}, a_{-i}^{k,t}))$ and $\boldsymbol{a}^{k,t} \sim \pi^k$, We have

$$\mathbb{E}_k\left[[\widetilde{\theta}_i^{k,t}(\pi^k)]_f\right]$$

$$
\begin{aligned}
=&\mathbb{E}_{\boldsymbol{a}\sim\pi^k}\left[\frac{r^f(n^f(a_i,a_{-i}))\mathbb{1}\left\{f\in a_i\right\}}{\mathbb{P}_{a_i'\sim\pi_i^k}(f\in a_i')}\right]\\
=&\frac{1}{\mathbb{P}_{a_i'\sim\pi_i^k}(f\in a_i')}\cdot\mathbb{E}_{a_{-i}\sim\pi_{-i}^k}\left[\mathbb{E}_{a_i\sim\pi_i^k}\left[r^f(n^f(a_i,a_{-i}))\mathbb{1}\left\{f\in a_i\right\}\mid a_{-i}\right]\right]\\
=&\frac{1}{\mathbb{P}_{a_i'\sim\pi_i^k}(f\in a_i')}\cdot\mathbb{E}_{a_{-i}\sim\pi_{-i}^k}\left[\mathbb{E}_{a_i\sim\pi_i^k}\left[r^f(n^f(a_i,a_{-i}))\mid a_{-i},f\in a_i\right]\mathbb{P}_{a_i\sim\pi_i^k}\left(f\in a_i\mid a_{-i}\right)\right]\\
\overset{\text{(i)}}{=}&\frac{\mathbb{P}_{a_i\sim\pi_i^k}\left(f\in a_i\right)}{\mathbb{P}_{a_i'\sim\pi_i^k}(f\in a_i')}\cdot\mathbb{E}_{a_{-i}\sim\pi_{-i}^k}\left[r^f(n^f(a_{-i})+1)\right]\\
=&[\theta_i(\pi^k)]_f.
\end{aligned}
$$

The equality (i) above holds because $\mathbb{E}_{a_i\sim\pi_i^k}\left[r^f(n^f(a_i,a_{-i}))\mid a_{-i},f\in a_i\right]=r^f(n^f(a_{-i})+1)$ and $f\in a_i$ does not depend on a_{-i}.

For the second property, since $\mathbb{P}_{a_i\sim\pi^{k_i}}\left(f\in a_i\right)\geq\frac{\gamma}{2F}$ by Lemma B.4.11 and $r^{k,t,f}\in[0,1]$, we can immediately have $|[\widetilde{\theta}_i^{k,t}(\pi^k)]_f|\leq\frac{2F}{\gamma}$.

For the third property, we have

$$
\begin{aligned}
\mathbb{E}_k\left[[\widetilde{\theta}_i^{k,t}(\pi^k)]_f^2\right]=&\frac{\mathbb{E}_{\boldsymbol{a}\sim\pi^k}\left[r^f(n^f(a_i,a_{-i}))^2\mathbb{1}\left\{f\in a_i\right\}\right]}{\mathbb{P}_{a_i'\sim\pi_i^k}\left(f\in a_i'\right)^2}\\
\leq&\frac{\mathbb{E}_{\boldsymbol{a}\sim\pi^k}\left[\mathbb{1}\left\{f\in a_i\right\}\right]}{\mathbb{P}_{a_i'\sim\pi_i^k}\left(f\in a_i'\right)^2}\\
=&\frac{\mathbb{P}_{a_i\sim\pi_i^k}\left(f\in a_i\right)}{\mathbb{P}_{a_i'\sim\pi_i^k}\left(f\in a_i'\right)^2}\\
\leq&\frac{2F}{\gamma}.
\end{aligned}
$$

□

B.5 Algorithms for Independent Markov Congestion Games

In this section, present missing details of our centralized algorithm for independent Markov congestion games, which is summarized in Algorithm 15. The proof of its theoretical guarantee is given in Appendix B.6.

B.5.1 Algorithm for Semi-bandit Feedback

Under the semi-bandit feedback, the players can receive reward information from all facilities they choose. Therefore, we can similarly define

$$\begin{aligned} N_h^{k,f}(s^f, n) &= \sum_{k'=1}^{k} \mathbb{1}\left\{(s_h^{k',f}, n^f(\boldsymbol{a}_h^{k'})) = (s^f, n)\right\}, \\ \hat{r}_h^{k,f}(s^f, n) &= \frac{\sum_{k'=1}^{k} r_h^{k',f} \mathbb{1}\left\{(s_h^{k',f}, n^f(\boldsymbol{a}_h^{k'})) = (s^f, n)\right\}}{N_h^{k,f}(s^f, n) \vee 1}, \\ \widehat{P}_h^{k,f}(s'^f \mid s^f, n) &= \frac{\sum_{k'=1}^{k} \mathbb{1}\left\{(s_{h+1}^{k',f}, s_h^{k',f}, n^f(\boldsymbol{a}_h^{k'})) = (s'^f, s^f, n)\right\}}{N_h^{k,f}(s^f, n) \vee 1}. \end{aligned}$$

Then, the estimators for the reward function and transition kernel can be defined as

$$\hat{r}_{h,i}^k(s, \boldsymbol{a}) = \sum_{f \in a_i} \hat{r}_h^{k,f}(s^f, n^f(\boldsymbol{a})), \quad \widehat{P}_h^k(s' \mid s, \boldsymbol{a}) = \prod_{f \in \mathcal{F}} \widehat{P}_h^{k,f}(s'^f \mid s^f, n^f(\boldsymbol{a})) \tag{B.2}$$

Then, with $\iota = 2\log(4(m+1)(\sum_{f\in\mathcal{F}} S^f)T/\delta)$, we define the bonus term to be $b_h^k(s, \boldsymbol{a}) = b_h^{k,\mathrm{pv}}(s, \boldsymbol{a}) + b_h^{k,\mathrm{r}}(s, \boldsymbol{a})$, which is a sum of transition bonus and reward bonus. In particular, we have

$$b_h^{k,\mathrm{pv}}(s, \boldsymbol{a}) = \sum_{f \in \mathcal{F}} \sqrt{\frac{4H^2F^2S^f\iota}{N_h^{k,f}(s^f, n^f(\boldsymbol{a})) \vee 1}} + \sum_{f \neq f'} \sqrt{\frac{4H^2F^2\left(S^fS^{f'}\iota\right)^2}{N_h^{k,f}(s^f, n^f(\boldsymbol{a}))N_h^{k,f'}(s^{f'}, n^{f'}(\boldsymbol{a})) \vee 1}}, \tag{B.3}$$

$$b_h^{k,\mathrm{r}}(s, \boldsymbol{a}) = \sum_{f \in \mathcal{F}} \sqrt{\frac{\iota}{N_h^{k,f}(s^f, n^f(\boldsymbol{a})) \vee 1}}. \tag{B.4}$$

For convenience, we define $(\widehat{\mathbb{P}}_h^k V)(s, \boldsymbol{a}) = \mathbb{E}_{s' \sim \widehat{P}_h^k(\cdot|s,\boldsymbol{a})}[V(s')]$ with value function $V : \mathcal{S} \mapsto \mathbb{R}$.

Remark B.5.1. Unlike Algorithm 4 for congestion game, here, $\overline{Q}_{h,1}^k(s, \cdot), \ldots, \overline{Q}_{h,m}^k(s, \cdot)$ in line 6 of Algorithm 15 in general does not form a potential game. Therefore, we cannot use Algorithm 5 and ϵ-NASH is not always computationally efficient.

Algorithm 15 Nash-VI for IMCGs

Input: ϵ, accuracy parameter for Nash equilibrium computation
Initialize: $\overline{V}^k_{H+1,i}(s) = 0$ for all $(i,k,s) \in [m] \times [K] \times \mathcal{S}$
for episode $k = 1, \ldots, K$ **do**
  for step $h = H, H-1, \ldots, 1$ **do**
    for player $i = 1, \ldots, m$ **do**
      $\overline{Q}^k_{h,i}(s, \boldsymbol{a}) \leftarrow \min\left\{(\hat{r}^k_{h,i} + \widehat{\mathbb{P}}^k_h \overline{V}^k_{h+1,i} + b^k_h)(s, \boldsymbol{a}), HF\right\}$ for all $(s, \boldsymbol{a}) \in \mathcal{S} \times \mathcal{A}$
    end for
    for $s \in \mathcal{S}$ **do**
      $\pi^k_h(\cdot \mid s) \leftarrow \epsilon$-NASH$(\overline{Q}^k_{h,1}(s, \cdot), \cdots, \overline{Q}^k_{h,m}(s, \cdot))$
      for player $i = 1, \ldots, m$ **do**
        $\overline{V}^k_{h,i}(s) \leftarrow \mathbb{E}_{\boldsymbol{a} \sim \pi^k_h}[\overline{Q}^k_{h,i}(s, \boldsymbol{a})]$
      end for
    end for
  end for
  for step $h = 1, \ldots, H$ **do**
    Take action $\boldsymbol{a}^k_h \sim \pi^k_h(\cdot \mid s^k_h)$, observe reward $r^{k,f}_h$ and next state s^k_{h+1}
    Update reward estimator $\hat{r}^k_{h,i}$, transition estimator $\widehat{P}^k_h$ and bonus term b^k_h
  end for
end for

B.5.2 Algorithm for Bandit Feedback

In bandit feedback scenario, since players' observation about state transitions remains unaffected, we only need to modify the reward estimator $\hat{r}^k_{h,i}$ defined in (B.2) and reward bonus term $b^{k,\mathrm{r}}_h(s, \boldsymbol{a})$ defined in (B.4).

Similar to the congestion game with bandit feedback introduced in Section 3.4.2, for IMCGs, we can also write its reward function as $r_{h,i}(s, \boldsymbol{a}) = \langle A_i(s, \boldsymbol{a}), \theta_h \rangle$, where θ_h is unknown and $A_i(s, \boldsymbol{a})$ is a 0-1 vector.

In particular, define $\theta_h \in [0,1]^d$ with $d = m \sum_{f \in \mathcal{F}} S^f$ to be the vector such that $\theta_{h,i} = r^f_h(s^f, n)$ for some $f \in \mathcal{F}$ and $(s^f, n) \in \mathcal{S}^f \times [m]$. Then, we can similarly build estimator $\hat{r}^k_{h,i}$ through ridge regression as the following.[†]

[†]For the same reason, we take the regularization parameter in ridge regression to be 1.

$$\text{design matrix:} \quad V_h^k = I + \sum_{k'=1}^{k-1} \sum_{i=1}^{m} A_i(s_h^{k'}, \boldsymbol{a}_h^{k'}) A_i(s_h^{k'}, \boldsymbol{a}_h^{k'})^\top, \tag{B.5}$$

$$\theta_h \text{ estimator:} \quad \widehat{\theta}_h^k = \left(V_h^k\right)^{-1} \sum_{k'=1}^{k-1} \sum_{i=1}^{m} A_i(s_h^{k'}, \boldsymbol{a}_h^{k'}) r_{h,i}^{k'}, \tag{B.6}$$

$$\text{reward estimator:} \quad \tilde{r}_{h,i}^k(s, \boldsymbol{a}) = \left\langle A_i(s, \boldsymbol{a}), \widehat{\theta}_h^k \right\rangle, \tag{B.7}$$

$$\text{reward bonus:} \quad \tilde{b}_h^{k,\mathrm{r}}(s, \boldsymbol{a}) = \max_{i \in [m]} \|A_i(s, \boldsymbol{a})\|_{(V_h^k)^{-1}} \sqrt{\beta_k}, \tag{B.8}$$

where $\sqrt{\beta_k} = \sqrt{d} + \sqrt{Fd \log\left(1 + \frac{mkF}{d}\right) + F\iota}$.

B.6 Analysis for Algorithm 15

B.6.1 Bellman Equations for Genera-sum Markov Games

Before analyzing Algorithm 15, we first give a brief review of the Bellman equations for general-sum Markov games. These equations are well-known among the literature Bai and Jin [2020], Liu et al. [2021], Jin et al. [2021b].

Fixed policies. Given a fixed policy π, for any $(h, i, s, \boldsymbol{a}) \in [H] \times [m] \times \mathcal{S} \times \mathcal{A}$, it holds that

$$Q_{h,i}^\pi(s, \boldsymbol{a}) = (r_{h,i} + \mathbb{P}_h V_{h+1,i}^\pi)(s, \boldsymbol{a}), \quad V_{h,i}^\pi = \mathbb{E}_{\boldsymbol{a}' \sim \pi_h(\cdot|s)} \left[Q_{h,i}^\pi(s, \boldsymbol{a}') \right], \tag{B.9}$$

where $V_{H+1,i}^\pi(s) = 0$ for any $(i, s) \in [m] \times \mathcal{S}$.

Best responses. Given a fixed policy π, define the best response value functions for player i as $Q_{h,i}^{\dagger,\pi_{-i}}(s, \boldsymbol{a}) = \max_{\pi_i \in \Delta(\mathcal{A}_i)} Q_{h,i}^{\pi_i,\pi_{-i}}(s, \boldsymbol{a})$ and $V_{h,i}^{\dagger,\pi_{-i}}(s) = \max_{\pi_i \in \Delta(\mathcal{A}_i)} V_{h,i}^{\pi_i,\pi_{-i}}(s)$. Then, for any $(h, i, s, \boldsymbol{a}) \in [H] \times [m] \times \mathcal{S} \times \mathcal{A}$, it holds that

$$\begin{aligned} Q_{h,i}^{\dagger,\pi_{-i}}(s, \boldsymbol{a}) &= (r_{h,i} + \mathbb{P}_h V_{h+1,i}^{\dagger,\pi_{-i}})(s, \boldsymbol{a}), \\ V_{h,i}^{\dagger,\pi_{-i}}(s) &= \max_{\nu \in \Delta(\mathcal{A}_i)} \mathbb{E}_{\boldsymbol{a}' \sim (\nu, \pi_{h,-i})(\cdot|s)} \left[Q_{h,i}^{\dagger,\pi_{-i}}(s, \boldsymbol{a}') \right], \end{aligned} \tag{B.10}$$

where $V_{H+1,i}^{\dagger,\pi_{-i}}(s) = 0$ for any $(i,s) \in [m] \times \mathcal{S}$.

B.6.2 Proof of Theorem 3.6.2

Recall that the update rule in Algorithm 15 is

$$\overline{Q}_{h,i}^k(s,\boldsymbol{a}) \leftarrow \min\left\{(\hat{r}_{h,i}^k + \widehat{\mathbb{P}}_h^k \overline{V}_{h+1,i}^k + b_h^k)(s,\boldsymbol{a}), HF\right\}, \quad \overline{V}_{h,i}^k(s) \leftarrow \mathbb{E}_{\boldsymbol{a}\sim\pi_h^k}[\overline{Q}_{h,i}^k(s,\boldsymbol{a})].$$

Similar to the proof of Theorem 3.4.1, we define auxiliary value functions

$$\underline{Q}_{h,i}^k(s,\boldsymbol{a}) \leftarrow \max\left\{(\hat{r}_{h,i}^k + \widehat{\mathbb{P}}_h^k \underline{V}_{h+1,i}^k - b_h^k)(s,\boldsymbol{a}), 0\right\}, \quad \underline{V}_{h,i}^k(s) \leftarrow \mathbb{E}_{\boldsymbol{a}\sim\pi_h^k}[\underline{Q}_{h,i}^k(s,\boldsymbol{a})]. \tag{B.11}$$

We now begin to prove the first part of Theorem 3.6.2.

Proof of Theorem 3.6.2. **Step 1.** We first consider the setting of semi-bandit feedback. Assume the result in Lemma B.6.2 holds since it is a high-probability event. Then, for any $(k,s) \in [K] \times \mathcal{S}$, it holds that

$$\max_{i\in[m]} \left(V_{1,i}^{\dagger,\pi_{-i}^k} - V_{1,i}^{\pi^k}\right)(s) \le \max_{i\in[m]} \left(\overline{V}_{1,i}^k - \underline{V}_{1,i}^k\right)(s) + H\epsilon.$$

By the update rules in Algorithm 15, we can notice the following recursive relations

$$\begin{aligned}
(\overline{Q}_{h,i}^k - \underline{Q}_{h,i}^k)(s,\boldsymbol{a}) &\le \min\left\{\widehat{\mathbb{P}}_h^k(\overline{V}_{h+1,i}^k - \underline{V}_{h+1,i}^k)(s,\boldsymbol{a}) + 2b_h^k(s,\boldsymbol{a}), HF\right\}, \\
(\overline{V}_{h,i}^k - \underline{V}_{h,i}^k)(s) &= \mathbb{E}_{\boldsymbol{a}'\sim\pi_h^k(\cdot|s)}\left[(\overline{Q}_{h,i}^k - \underline{Q}_{h,i}^k)(s,\boldsymbol{a}')\right].
\end{aligned}$$

Thus, we define $\widetilde{V}_{H+1}^k(s) = 0$ for any $s \in \mathcal{S}$ and $\widetilde{Q}_h^k$, $\widetilde{V}_h^k$ recursively as

$$\widetilde{Q}_h^k(s,\boldsymbol{a}) = \min\left\{(\widehat{\mathbb{P}}_h^k \widetilde{V}_{h+1}^k)(s,\boldsymbol{a}) + 2b_h^k(s,\boldsymbol{a}), HF\right\}, \quad \widetilde{V}_h^k(s) = \mathbb{E}_{\boldsymbol{a}'\sim\pi_h^k(\cdot|s)}\left[\widetilde{Q}_h^k(s,\boldsymbol{a}')\right]. \tag{B.12}$$

Obviously, we have $\max_{i\in[m]}(\overline{V}_{h,i}^k - \underline{V}_{h,i}^k)(s) \le \widetilde{V}_{H+1}^k$. Then, by inductively assuming the

same relation holds for $h+1$, we can have

$$\begin{aligned}
\max_{i\in[m]}(\overline{Q}_{h,i}^k - \underline{Q}_{h,i}^k)(s,\boldsymbol{a}) =& \min\left\{\max_{i\in[m]} \widehat{\mathbb{P}}_h^k(\overline{V}_{h+1,i}^k - \underline{V}_{h+1,i}^k)(s,\boldsymbol{a}) + 2b_h^k(s,\boldsymbol{a}), HF\right\} \\
\leq& \min\left\{(\widehat{\mathbb{P}}_h^k \widetilde{V}_{h+1}^k)(s,\boldsymbol{a}) + 2b_h^k(s,\boldsymbol{a}), HF\right\} \\
=& \widetilde{Q}_h^k(s,\boldsymbol{a}), \\
\max_{i\in[m]}(\overline{V}_{h,i}^k - \underline{V}_{h,i}^k)(s) \leq& \mathbb{E}_{\boldsymbol{a}'\sim\pi_h^k(\cdot|s)}\left[\max_{i\in[m]}(\overline{Q}_{h,i}^k - \underline{Q}_{h,i}^k)(s,\boldsymbol{a}')\right] \\
\leq& \mathbb{E}_{\boldsymbol{a}'\sim\pi_h^k(\cdot|s)}\left[\widetilde{Q}_h^k(s,\boldsymbol{a}')\right] \\
=& \widetilde{V}_h^k(s).
\end{aligned}$$

Therefore, by induction, for any $h\in[H]$, we have

$$\max_{i\in[m]}(\overline{Q}_{h,i}^k - \underline{Q}_{h,i}^k)(s,\boldsymbol{a}) \leq \widetilde{Q}_h^k(s,\boldsymbol{a}), \quad \max_{i\in[m]}(\overline{V}_{h,i}^k - \underline{V}_{h,i}^k)(s) \leq \widetilde{V}_h^k(s).$$

As a result, we have

$$\text{Nash-Regret}(K) = \sum_{k=1}^K \max_{i\in[m]}\left(V_{1,i}^{\dagger,\pi_{-i}^k} - V_{1,i}^{\pi^k}\right)(s) \leq \sum_{k=1}^K \widetilde{V}_1^k(s_1) + HK\epsilon.$$

Step 2, Semi-bandit Feedback. We define the martingale difference sequences

$$\begin{aligned}
\mathcal{M}_h^k(\widetilde{Q}) &= \mathbb{E}_{\boldsymbol{a}'\sim\pi_h^k(\cdot|s_h^k)}\left[\widetilde{Q}_h^k(s_h^k,\boldsymbol{a}')\right] - \widetilde{Q}_h^k(s_h^k,\boldsymbol{a}_h^k), \\
\mathcal{M}_h^k(\widetilde{V}) &= (\mathbb{P}_h\widetilde{V}_{h+1}^k)(s_h^k,\boldsymbol{a}_h^k) - \widetilde{V}_{h+1}^k(s_{h+1}^k).
\end{aligned}$$

It is not hard to check that $\mathcal{M}_h^k(\widetilde{Q})$ and $\mathcal{M}_h^k(\widetilde{V})$ are both indeed martingale difference sequences with respect to the history till episode k and time step h.

With these definitions, we can now decompose the regret bound as

$$\begin{aligned}
\widetilde{V}_h^k(s_h^k) =& \mathbb{E}_{\boldsymbol{a}'\sim\pi_h^k(\cdot|s_h^k)}\left[\widetilde{Q}_h^k(s_h^k,\boldsymbol{a}')\right] && \text{(By (B.12))} \\
=& \mathcal{M}_h^k(\widetilde{Q}) + \widetilde{Q}_h^k(s_h^k,\boldsymbol{a}_h^k) \\
\leq& \mathcal{M}_h^k(\widetilde{Q}) + 2b_h^k(s_h^k,\boldsymbol{a}_h^k) + (\widehat{\mathbb{P}}_h^k\widetilde{V}_{h+1}^k)(s_h^k,\boldsymbol{a}_h^k) && \text{(By (B.12))}
\end{aligned}$$

$$\begin{aligned}&\overset{\text{(i)}}{\leq}\mathcal{M}_h^k(\widetilde{Q})+3b_h^k(s_h^k,\boldsymbol{a}_h^k)+(\mathbb{P}_h\widetilde{V}_{h+1}^k)(s_h^k,\boldsymbol{a}_h^k)\\&=\mathcal{M}_h^k(\widetilde{Q})+\mathcal{M}_h^k(\widetilde{V})+3b_h^k(s_h^k,\boldsymbol{a}_h^k)+\widetilde{V}_{h+1}^k(s_{h+1}^k)\end{aligned}$$

The above inequality (i) holds by applying Lemma B.6.2 and the fact $\widetilde{V}_h^k(s)\leq HF$, which comes from the definition in (B.12). Then, by unrolling this relation from $h=1$ to $h=H$ and noticing $\widetilde{V}_{H+1}^k=\mathbf{0}$, we can have

$$\begin{aligned}&\text{Nash-Regret}(K)\leq\sum_{k=1}^K\widetilde{V}_1^k(s_1)+HK\epsilon\\&\leq\sum_{k=1}^K\sum_{h=1}^H\left(\mathcal{M}_h^k(\widetilde{Q})+\mathcal{M}_h^k(\widetilde{V})+3b_h^k(s_h^k,\boldsymbol{a}_h^k)\right)+HK\epsilon\qquad\text{(B.13)}\\&\leq\widetilde{\mathcal{O}}\left(HF\sqrt{T}\right)+3\sum_{k=1}^K\sum_{h=1}^H b_h^k(s_h^k,\boldsymbol{a}_h^k)\\&\qquad\qquad\text{(By Azuma-Hoeffding inequality and taking }\epsilon=1/T.\text{)}\\&\leq\widetilde{\mathcal{O}}\left(HF\sqrt{T}\right)+6HF\sum_{f\in\mathcal{F}}\sum_{k=1}^K\sum_{h=1}^H\left(\sqrt{\frac{S^f\iota}{N_h^{k,f}(s_h^{k,f},n^f(\boldsymbol{a}_h^k))\vee1}}+\sqrt{\frac{\iota}{N_h^{k,f}(s_h^{k,f},n^f(\boldsymbol{a}_h^k))\vee1}}\right)\\&\qquad+6HF\sum_{f\neq f'}S^fS^{f'}\sum_{k=1}^K\sum_{h=1}^H\sqrt{\frac{\iota^2}{\left(N_h^{k,f}(s_h^{k,f},n^f(\boldsymbol{a}_h^k))N_h^{k,f'}(s_h^{k,f'},n^{f'}(\boldsymbol{a}_h^{k,f'}))\right)\vee1}}\\&\leq\widetilde{\mathcal{O}}\left(HF\sqrt{T}\right)+\widetilde{\mathcal{O}}\left(\sum_{f\in\mathcal{F}}HFS^f\sqrt{mHT}\right)+\widetilde{\mathcal{O}}\left(m^2H^2F\sum_{f\neq f'}\left(S^fS^{f'}\right)^2\right)\\&\qquad\qquad\text{(By Lemma B.6.5 and B.6.6)}\\&\leq\widetilde{\mathcal{O}}\left(\sum_{f\in\mathcal{F}}FS^f\sqrt{mH^3T}\right)+\widetilde{\mathcal{O}}\left(m^2H^2F\sum_{f\neq f'}\left(S^fS^{f'}\right)^2\right).\end{aligned}$$

Step 3, Bandit Feedback. In the setting of bandit feedback, we only modify the reward estimator $\tilde{r}_{h,i}^k$ and its corresponding bonus term $\tilde{b}_h^{k,\mathrm{r}}$. Thus, by going through the proof of Lemma B.6.2, we can notice that to have the same result for bandit feedback, it suffice to use Lemma B.6.3 to show that the reward estimation error is bounded by the reward bonus term.

Then, by the inequality (B.13), we can notice that to achieve the final Nash-regret bound,

we only need to bound the summation $\sum_{k=1}^{K}\sum_{h=1}^{H}\tilde{b}_h^{k,\mathrm{r}}(s_h^k,\boldsymbol{a}_h^k)$, which is

$$\begin{aligned}
\sum_{k=1}^{K}\sum_{h=1}^{H}\tilde{b}_h^{k,\mathrm{r}}(s_h^k,\boldsymbol{a}_h^k) \leq & \sqrt{\beta_K}\sum_{k=1}^{K}\sum_{h=1}^{H}\max_{i\in[m]}\left\|A_i(s_h^k,\boldsymbol{a}_h^k)\right\|_{(V_h^k)^{-1}} \\
& \qquad \text{(By definition of } \tilde{b}_h^{k,\mathrm{r}} \text{ in (B.8).)} \\
\leq & \left(\sqrt{d}+\sqrt{Fd\log\left(1+\frac{mKF}{d}\right)+F\iota}\right)\widetilde{\mathcal{O}}\left(H\sqrt{dFK}\right) \\
& \qquad \text{(By definition of } \beta_k \text{ and Lemma B.6.4.)} \\
\leq & \widetilde{\mathcal{O}}\left(d\sqrt{HF^2T}\right) \\
= & \widetilde{\mathcal{O}}\left(\sum_{f\in\mathcal{F}} mS^f\sqrt{HF^2T}\right). \qquad \text{(Since } d=m\textstyle\sum_{f\in\mathcal{F}}S^f\text{.)}
\end{aligned}$$

Therefore, by (B.13), with $\epsilon = 1/T$, under bandit feedback, we have

$$\begin{aligned}
& \text{Nash-Regret}(K) \\
\leq & \sum_{k=1}^{K}\sum_{h=1}^{H}\left(\mathcal{M}_h^k(\widetilde{Q})+\mathcal{M}_h^k(\widetilde{V})+3b_h^k(s_h^k,\boldsymbol{a}_h^k)\right) \\
\leq & \widetilde{\mathcal{O}}\left(\sum_{f\in\mathcal{F}} FS^f\sqrt{mH^3T}\right)+\widetilde{\mathcal{O}}\left(m^2H^2F\sum_{f\neq f'}\left(S^fS^{f'}\right)^2\right)+\sum_{k=1}^{K}\sum_{h=1}^{H}\tilde{b}_h^{k,\mathrm{r}}(s_h^k,\boldsymbol{a}_h^k) \\
\leq & \widetilde{\mathcal{O}}\left(\sum_{f\in\mathcal{F}}\left(\sqrt{mH^3}F+m\sqrt{HF^2}\right)S^f\sqrt{T}\right)+\widetilde{\mathcal{O}}\left(m^2H^2F\sum_{f\neq f'}\left(S^fS^{f'}\right)^2\right).
\end{aligned}$$

□

B.6.3 Lemmas for Semi-bandit Feedback

The following two lemmas shows that our value function estimations are indeed optimistic.

Lemma B.6.1. *With probability at least $1-\delta$, simultaneously for arbitrary value function $V\in[0,HF]^{\mathcal{S}}$ and any tuple $(k,h,s,\boldsymbol{a})$, it holds that $|(\widehat{\mathbb{P}}_h^k-\mathbb{P}_h)V(s,\boldsymbol{a})|\leq b_h^{k,\mathrm{pv}}(s,\boldsymbol{a})$, where $b_h^{k,\mathrm{pv}}(s,\boldsymbol{a})$ is defined in (B.3).*

Proof. We define $\mathbb{P}_h^f$ to be the operator such that for some value function $V^f:\mathcal{S}^f\mapsto\mathbb{R}$, we

have $(\mathbb{P}_h^f V^f)(s,\boldsymbol{a}) = \mathbb{E}_{s'^f \sim P_h^f(\cdot|s^f,n^f(\boldsymbol{a}))}\left[V^f(s'^f)\right]$. We also define $\widehat{\mathbb{P}}_h^{k,f}$ similarly. Then, by definition of our transition kernel, for operators $\mathbb{P}_h$ and $\widehat{\mathbb{P}}_h^k$, it holds that

$$\mathbb{P}_h = \prod_{f\in\mathcal{F}} \mathbb{P}_h^f \quad \text{and} \quad \widehat{\mathbb{P}}_h^k = \prod_{f\in\mathcal{F}} \widehat{\mathbb{P}}_h^{k,f}.$$

Therefore, by Lemma E.1 in Chen et al. [2020], since $\|V\|_\infty \le HF$, we have

$$\begin{aligned}|(\widehat{\mathbb{P}}_h^k - \mathbb{P}_h)V(s,\boldsymbol{a})| \le \sum_{f\in\mathcal{F}} \left|(\widehat{\mathbb{P}}_h^{k,f} - \mathbb{P}_h^f)\left(\prod_{f'\neq f}\mathbb{P}_h^{f'}\right)V(s,\boldsymbol{a})\right| \\ + 2HF\sum_{f\neq f'} \mathrm{errp}_h^{k,f}(s,\boldsymbol{a})\cdot \mathrm{errp}_h^{k,f'}(s,\boldsymbol{a}),\end{aligned} \tag{B.14}$$

where $\mathrm{errp}_h^{k,f}(s,\boldsymbol{a}) = \|\widehat{P}_h^{k,f}(\cdot \mid s^f, n^f(\boldsymbol{a})) - P_h^f(\cdot \mid s^f, n^f(\boldsymbol{a}))\|_1$.

Now, notice that $\left(\prod_{f'\neq f}\mathbb{P}_h^{f'}\right)V(s,\boldsymbol{a})$ can be seen as some value function from $\mathcal{S}^f$ to $[0, HF]$. Therefore, by Lemma 12 in Bai and Jin [2020], with probability at least $1-\frac{\delta}{2}$, simultaneously for any V and $(k,h,s,\boldsymbol{a})$, it holds that

$$\left|(\widehat{\mathbb{P}}_h^{k,f} - \mathbb{P}_h^f)\left(\prod_{f'\neq f}\mathbb{P}_h^{f'}\right)V(s,\boldsymbol{a})\right| \le 2HF\sqrt{\frac{S^f\iota}{N_h^{k,f}(s^f,n^f(\boldsymbol{a}))\vee 1}},$$

where $\iota = 2\log(4(m+1)(\sum_{f\in\mathcal{F}}S^f)T/\delta)$. Meanwhile, by standard Hoeffding's inequality and union bound, with probability at least $1-\frac{\delta}{2}$, simultaneously for any $(k,h,s,\boldsymbol{a})$, it holds that

$$\mathrm{errp}_h^{k,f} \le S^f\sqrt{\frac{\iota}{N_h^{k,f}(s^f,n^f(\boldsymbol{a}))\vee 1}}.$$

Finally, by plugging above two concentration inequalities back into (B.14), we can have

$$|(\widehat{\mathbb{P}}_h^k - \mathbb{P}_h)V(s,\boldsymbol{a})| \le b_h^{k,\mathrm{pv}}(s,\boldsymbol{a}).$$

□

Lemma B.6.2. *With probability at least $1-\delta$, for any $(k,h,i,s,\boldsymbol{a}) \in [K]\times[H]\times[m]\times\mathcal{S}\times\mathcal{A}$,*

it holds that

$$\overline{Q}^k_{h,i}(s,\boldsymbol{a}) \geq Q^{\dagger,\pi^k_{-i}}_{h,i}(s,\boldsymbol{a}) - (H-h)\epsilon, \quad \underline{Q}^k_{h,i}(s,\boldsymbol{a}) \leq Q^{\pi^k}_{h,i}(s,\boldsymbol{a}), \tag{B.15}$$

$$\overline{V}^k_{h,i}(s) \geq V^{\dagger,\pi^k_{-i}}_{h,i}(s) - (H-h+1)\epsilon, \quad \underline{V}^k_{h,i}(s) \leq V^{\pi^k}_{h,i}(s.), \tag{B.16}$$

where $\underline{Q}^k_{h,k}$ *and* $\underline{V}^k_{h,i}$ *are defined in* (B.11).

Proof. The proof is adapted from Liu et al. [2021] and goes by induction from $h = H+1$ to $h = 1$. We can see that inequalities (B.16) obviously hold when $h = H+1$ since by definition we have $\overline{V}^k_{H+1,i}(s) = \underline{V}^k_{H+1,i}(s) = 0$ for any (k,i,s). Now, suppose inequalities (B.16) hold for $h+1$. Then, if we have $\overline{Q}^k_{h,i}(s,\boldsymbol{a}) = HF$, it holds trivially that $\overline{Q}^k_{h,i}(s,\boldsymbol{a}) \geq Q^{\dagger,\pi^k_{-i}}_{h,i}(s,\boldsymbol{a})$. Otherwise, by Bellman equations (B.10) and update rule in Algorithm 15, we have

$$\begin{aligned}
&\overline{Q}^k_{h,i}(s,\boldsymbol{a}) - Q^{\dagger,\pi^k_{-i}}_{h,i}(s,\boldsymbol{a}) \\
=&(\hat{r}^k_{h,i} - r_{h,i})(s,\boldsymbol{a}) + (\widehat{\mathbb{P}}^k_h \overline{V}^k_{h+1,i})(s,\boldsymbol{a}) - (\mathbb{P}_h V^{\dagger,\pi^k_{-i}}_{h+1,i})(s,\boldsymbol{a}) + b^k_h(s,\boldsymbol{a}) \\
=&\underbrace{(\hat{r}^k_{h,i} - r_{h,i})(s,\boldsymbol{a})}_{\text{(A)}} + \underbrace{\widehat{\mathbb{P}}^k_h(\overline{V}^k_{h+1,i} - V^{\dagger,\pi^k_{-i}}_{h+1,i})(s,\boldsymbol{a})}_{\text{(B)}} + \underbrace{((\widehat{\mathbb{P}}^k_h - \mathbb{P}_h)V^{\dagger,\pi^k_{-i}}_{h+1,i})(s,\boldsymbol{a})}_{\text{(C)}} + b^k_h(s,\boldsymbol{a}).
\end{aligned}$$

Now, recall that $b^k_h(s,\boldsymbol{a}) = b^{k,\mathrm{pv}}_h(s,\boldsymbol{a}) + b^{k,\mathrm{r}}_h(s,\boldsymbol{a})$. By reward definition in congestion game, we have

$$(\hat{r}^k_{h,i} - r_{h,i})(s,\boldsymbol{a}) = \sum_{f \in a_i} (\hat{r}^{k,f}_{h,i}(s^f, n^f(\boldsymbol{a})) - r^f_{h,i}(s^f, n^f(\boldsymbol{a}))).$$

Thus, by using standard Hoefding's inequality and union bound, we can immediately have $|\text{(A)}| \leq b^{k,\mathrm{r}}_h(s,\boldsymbol{a})$. Then, since $V^{\dagger,\pi^k_{-i}}_{h,i} \in [0,HF]^{\mathcal{S}}$, by Lemma B.6.1, we have $|\text{(C)}| \leq b^{k,\mathrm{pv}}_h(s,\boldsymbol{a})$. That is, we have $\text{(A)} + \text{(C)} + b^k_h(s,\boldsymbol{a}) \geq 0$.

Then, by inductive hypothesis, we know that $\overline{V}^k_{h+1,i} \geq V^{\dagger,\pi^k_{-i}}_{h+1,i} - (H-h)\epsilon$, which implies $\text{(B)} \geq 0$. Therefore, we have $\overline{Q}^k_{h,i}(s,\boldsymbol{a}) - Q^{\dagger,\pi^k_{-i}}_{h,i}(s,\boldsymbol{a}) \geq -(H-h)\epsilon$.

For $\overline{V}^k_{h,i}$ and $V^{\dagger,\pi^k_{-i}}_{h,i}$, we notice that in Algorithm 15, π^k is computed as the ϵ-approximate Nash equilibrium of $(\overline{Q}^k_{h,1}, \ldots, \overline{Q}^k_{h,m})$. Therefore, it holds that

$$\overline{V}^k_{h,i}(s) = \mathbb{E}_{\boldsymbol{a}\sim\pi^k_h(\cdot|s)}\left[\overline{Q}^k_{h,i}(s,\boldsymbol{a})\right] \geq \max_{\nu\in\Delta(\mathcal{A}_i)} \mathbb{E}_{\boldsymbol{a}'\sim(\nu,\pi^k_{h,-i})(\cdot|s)}\left[\overline{Q}^k_{h,i}(s,\boldsymbol{a}')\right] - \epsilon.$$

By Bellman equations (B.10), we also have

$$V_{h,i}^{\dagger,\pi_{-i}^k}(s) = \max_{\nu\in\Delta(\mathcal{A}_i)} \mathbb{E}_{\boldsymbol{a}'\sim(\nu,\pi_{h,-i}^k)(\cdot|s)}\left[Q_{h,i}^{\dagger,\pi_{-i}^k}(s,\boldsymbol{a}')\right].$$

Since $\overline{Q}_{h,i}^k(s,\boldsymbol{a}) - Q_{h,i}^{\dagger,\pi_{-i}^k}(s,\boldsymbol{a}) \geq -(H-h)\epsilon$, we immediately have $\overline{V}_{h,i}^k(s) - V_{h,i}^{\dagger,\pi_{-i}^k}(s) \geq -(H-h+1)\epsilon$. Thus, by induction, we have that $\overline{Q}_{h,i}^k(s,\boldsymbol{a}) \geq Q_{h,i}^{\dagger,\pi_{-i}^k}(s,\boldsymbol{a}) - (H-h)\epsilon$ and $\overline{V}_{h,i}^k(s) \geq V_{h,i}^{\dagger,\pi_{-i}^k}(s) - (H-h+1)\epsilon$ for all $h\in[H]$.

The inequalities for $\underline{V}_{h,i}^k$ and $\underline{Q}_{h,i}^k$ can be proved similarly. □

B.6.4 Additional Lemmas for Bandit Feedback

The following lemma shows that the reward estimation error can be bounded by the reward bonus term.

Lemma B.6.3. *With probability at least $1-\delta$, simultaneously for all $(i,k,h,s,\boldsymbol{a})$, it holds that $|(\tilde{r}_{h,i}^k - r_{h,i})(s,\boldsymbol{a})| \leq \tilde{b}_h^{k,\mathrm{r}}(s,\boldsymbol{a})$, where $\tilde{r}_{h,i}^k$ and $\tilde{b}_h^{k,\mathrm{r}}$ are defined in (B.7) and (B.8).*

Proof. The proof is extremely similar to Lemma B.3.1. By construction, we have

$$\begin{aligned}
|(\tilde{r}_{h,i}^k - r_{h,i})(s,\boldsymbol{a})| &= \left|\left\langle A_i(s,\boldsymbol{a}), \widehat{\theta}_h - \theta_h\right\rangle\right| \\
&\leq \|A_i(s,\boldsymbol{a})\|_{(V_h^k)^{-1}} \left\|\widehat{\theta}_h - \theta_h\right\|_{V_h^k} \\
&\leq \|A_i(s,\boldsymbol{a})\|_{(V_h^k)^{-1}} \left(\|\theta_h\|_2 + \sqrt{F\log\left(\det(V_h^k)\right) + F\iota}\right). \\
&\qquad \text{(By Theorem 20.5 in Lattimore and Szepesvári [2020].)}
\end{aligned}$$

Since each element in θ_h is bounded in $[0,1]$ by construction, we have $\|\theta_h\|_2 \leq \sqrt{d}$.

Then, by Lemma B.3.2, we have $\det\left(V_h^k\right) \leq \left(1 + \frac{mkF}{d}\right)^d$ since by construction $\|A_i(s,\boldsymbol{a})\|_2^2 \leq F$.

Finally, to make this bound valid for all player $i\in[m]$, we only need to take maximization over $i\in[m]$. Therefore, with probability at least $1-\delta$, we have

$$|(\tilde{r}_{h,i}^k - r_{h,i})(s,\boldsymbol{a})| \leq \max_{i\in[m]} \|A_i(s,\boldsymbol{a})\|_{(V_h^k)^{-1}} \sqrt{\beta_k} = \tilde{b}_h^{k,\mathrm{r}}(s,\boldsymbol{a}),$$

where $\sqrt{\beta_k} = \sqrt{d} + \sqrt{Fd\log\left(1+\frac{mkF}{d}\right) + F\iota}$. □

The follow lemma bound the sum of reward bonus under bandit feedback.

Lemma B.6.4. *For any $h \in [H]$, it holds that*

$$\sum_{k=1}^{K} \max_{i\in[m]} \left\|A_i(s_h^k, \boldsymbol{a}_h^k)\right\|_{(V_h^k)^{-1}} \leq \widetilde{\mathcal{O}}\left(\sqrt{dFK}\right),$$

where $d = m\sum_{f\in\mathcal{F}} S^f$.

Proof. First, since $V_h^k = I + \sum_{k'=1}^{k-1}\sum_{i=1}^{m} A_i(s_h^{k'}, \boldsymbol{a}_h^{k'})A_i(s_h^{k'}, \boldsymbol{a}_h^{k'})^\top$, we have $V_h^k \succeq I$ and thus $\left(V_h^k\right)^{-1} \preceq I$. Therefore, we have

$$\left\|A_i(s_h^k, \boldsymbol{a}_h^k)\right\|_{(V_h^k)^{-1}} \leq \left\|A_i(s_h^k, \boldsymbol{a}_h^k)\right\|_I = \left\|A_i(s_h^k, \boldsymbol{a}_h^k)\right\|_2 \leq \sqrt{F}.$$

For simplicity, let $A_{h,i}^k = A_i(s_h^k, \boldsymbol{a}_h^k)$. Then, as a result, we have

$$\begin{aligned}
\sum_{k=1}^{K} \max_{i\in[m]} \left\|A_{h,i}^k\right\|_{(V_h^k)^{-1}} &= \sum_{k=1}^{K} \min\left\{\max_{i\in[m]} \left\|A_{h,i}^k\right\|_{(V_h^k)^{-1}}, \sqrt{F}\right\} \\
&\leq \sqrt{K\sum_{k=1}^{K} \min\left\{\max_{i\in[m]} \left\|A_{h,i}^k\right\|^2_{(V_h^k)^{-1}}, F\right\}} \\
&\leq \sqrt{FK\sum_{k=1}^{K} \min\left\{\max_{i\in[m]} \left\|A_{h,i}^k\right\|^2_{(V_h^k)^{-1}}, 1\right\}} \\
&\leq \sqrt{2FKd\log\left(1+\frac{mKF}{d}\right)} \qquad \text{(By Lemma B.3.2.)} \\
&= \widetilde{\mathcal{O}}\left(\sqrt{dFK}\right).
\end{aligned}$$

□

B.6.5 Technical Lemmas

Lemma B.6.5. *For any $f \in \mathcal{F}$, it holds that*

$$\sum_{k=1}^{K}\sum_{h=1}^{H}\sqrt{\frac{1}{N_h^{k,f}(s_h^{k,f}, n^f(\boldsymbol{a}_h^k)) \vee 1}} \leq \widetilde{\mathcal{O}}\left(\sqrt{mHS^fT}\right).$$

Proof. Here, we have

$$\begin{aligned}
\sum_{k=1}^{K}\sum_{h=1}^{H}\sqrt{\frac{1}{N_h^{k,f}(s_h^{k,f}, n^f(\boldsymbol{a}_h^k)) \vee 1}} =& \sum_{h=1}^{H}\sum_{s^f\in\mathcal{S}^f}\sum_{n=0}^{m}\sum_{\ell=1}^{N_h^{K,f}(s^f,n)}\sqrt{\frac{1}{\ell}} \\
\leq& 2\sum_{h=1}^{H}\sum_{s^f\in\mathcal{S}^f}\sum_{n=0}^{m}\sqrt{N_h^{K,f}(s^f,n)} \\
& \qquad \text{(By standard technique)} \\
\leq& 2\sqrt{(m+1)HS^f\sum_{h=1}^{H}\sum_{s^f\in\mathcal{S}^f}\sum_{n=0}^{m}N_h^{K,f}(s^f,n)} \\
=& \widetilde{\mathcal{O}}\left(\sqrt{mHS^fT}\right).
\end{aligned}$$

The last line above holds because $\sum_{h=1}^{H}\sum_{s^f\in\mathcal{S}^f}\sum_{n=0}^{m}N_h^{K,f}(s^f,n) = T$. This is based on a pigeon-hole principle argument. In particular, whenever the players take one more action, for any $f \in \mathcal{F}$, the count for some tuple (h, s^f, n) will increase exactly by 1. □

Lemma B.6.6 (Chen et al. [2020])**.** *For any $f, f' \in \mathcal{F}$ and $f \neq f'$, it holds that*

$$\sum_{k=1}^{K}\sum_{h=1}^{H}\sqrt{\frac{1}{\left(N_h^{k,f}(s_h^{k,f}, n^f(\boldsymbol{a}_h^k))N_h^{k,f'}(s_h^{k,f'}, n^{f'}(\boldsymbol{a}_h^{k,f'}))\right) \vee 1}} \leq \widetilde{\mathcal{O}}\left(m^2HS^fS^{f'}\right).$$

Proof. We define the joint empirical counter

$$N_h^{k,f,f'}(s^f, s^{f'}, n, n') = \sum_{k'=1}^{k}\mathbb{1}\left\{(s_h^{k',f}, s_h^{k',f'}, n^f(\boldsymbol{a}_h^{k'}), n^{f'}(\boldsymbol{a}_n^{k'})) = (s^f, s^{f'}, n, n')\right\}.$$

Obviously, we have $N_h^{f,f'}(s^f, s^{f'}, n, n') \leq \min\left\{N_h^{k,f}(s^f, n), N_h^{k,f'}(s^{f'}, n')\right\}$, which implies

$$N_h^{k,f,f'}(s, s^{f'}, n, n') \leq \sqrt{N_h^{k,f}(s^f, n) N_h^{k,f'}(s^{f'}, n')}.$$

Therefore, we have

$$\begin{aligned}
&\sum_{k=1}^{K}\sum_{h=1}^{H}\sqrt{\frac{1}{\left(N_h^{k,f}(s_h^{k,f}, n^f(\boldsymbol{a}_h^k))N_h^{k,f'}(s_h^{k,f'}, n^{f'}(\boldsymbol{a}_h^{k,f'}))\right)\vee 1}} \\
\leq &\sum_{k=1}^{K}\sum_{h=1}^{H}\frac{1}{N_h^{k,f,f'}(s_h^{k,f}, s_h^{k,f'}, n^f(\boldsymbol{a}_h^k), n^{f'}(\boldsymbol{a}_h^k))\vee 1} \\
= &\sum_{h=1}^{H}\sum_{s^f\in\mathcal{S}^f}\sum_{s^{f'}\in\mathcal{S}^{f'}}\sum_{n=0}^{m}\sum_{n'=0}^{m}\sum_{\ell=1}^{N_h^{K,f,f'}(s^f,s^{f'},n,n')}\frac{1}{\ell} \\
= &\widetilde{\mathcal{O}}\left(m^2 H S^f S^{f'}\right).
\end{aligned}$$

□

Appendix C

OMITTED PROOFS AND EXPERIMENT RESULTS IN CHAPTER 4

C.1 Table of Notations

Symbol	Meaning
$\mathcal{S}$	The state space
$\mathcal{A}$	The action space
S	Size of state space
A	Size of action space
H	The length of horizon
K	The total number of episodes
T	The total number of steps, $T = HK$
π^k	The greedy policy generated in Algorithm 7 at episode k
$R_{h,s,a}$	Expected reward function at (h, s, a)
$P_{h,s,a}(s')$	Transition probability
M	Underlying true MDP, $M = (H, \mathcal{S}, \mathcal{A}, R, P, s_1)$
$n_k(h, s, a)$	$\sum_{k'=1}^{k-1} \mathbb{1}\{(s_h^{k'}, a_h^{k'}) = (s, a)\}$
$\hat{R}_{h,s,a}^k$	Estimated reward function, $\frac{1}{n_k(h,s,a)+1} \sum_{l=1}^{k-1} \mathbb{1}\{(s_h^l, a_h^l) = (s, a)\} r_{h,s_h^l,a_h^l}^l$
$\hat{P}_{h,s,a}^k(s')$	Estimated transition kernel, $\frac{1}{n_k(h,s,a)+1} \sum_{l=1}^{k-1} \mathbb{1}\{(s_h^l, a_h^l, s_{h+1}^l) = (s, a, s')\}$
$\tilde{P}_{h,s,a}^k(s')$	Estimated transition probability with a slightly different denominator, $\frac{1}{\max\{n_k(h,s,a),1\}} \sum_{l=1}^{k-1} \mathbb{1}\{(s_h^l, a_h^l, s_{h+1}^l) = (s, a, s')\}$
$\hat{M}^k$	Estimated MDP, $\hat{M}^k = (H, \mathcal{S}, \mathcal{A}, \hat{P}, \hat{R}, s_1)$
$\gamma_{\text{ty}}^k(h, s, a)$	$\sigma_{\text{ty}}^k(h, s, a)\sqrt{\log(40k^4)}$
$\hat{z}_k$	Perturbation's single random source during episode k from a standard Gaussian, $\hat{z}_k \sim \mathcal{N}(0, 1)$

$w^k_{\mathrm{ty}}(h,s,a)$	Noise of type "ty", $w^k_{\mathrm{ty}}(h,s,a)=\sigma^k_{\mathrm{ty}}(h,s,a)\hat{z}_k$
$\underline{w}^k_{\mathrm{ty}}(h,s,a)$	$-\gamma^k_{\mathrm{ty}}(h,s,a)$
$\overline{w}^k_{\mathrm{ty}}(h,s,a)$	$\gamma^k_{\mathrm{ty}}(h,s,a)$
$\overline{M}^k_{\mathrm{ty}}$	Perturbed estimated MDP with ty-type noise, $\overline{M}^k=(H,\mathcal{S},\mathcal{A},\hat{P},\hat{R}+w^k_{\mathrm{ty}},s_1)$
$\underline{M}^k_{\mathrm{ty}}$	Negatively perturbed MDP, $\underline{M}^k_{\mathrm{ty}}=(H,\mathcal{S},\mathcal{A},\hat{P},\hat{R}+\underline{w}^k_{\mathrm{ty}},s_1)$
$\overline{\overline{M}}^k_{\mathrm{ty}}$	Positively perturbed MDP, $\underline{M}^k_{\mathrm{ty}}=(H,\mathcal{S},\mathcal{A},\hat{P},\hat{R}+\overline{w}^k_{\mathrm{ty}},s_1)$
V^*_h / $V^*_{h,k}$	Optimal value function at step h for true MDP M
$V^{\pi^k}_h$ / $V^{\pi^k}_{h,k}$	Value function at step h by running policy π^k on true MDP M
$\overline{Q}_{h,k}$	Q-value function obtained by running Algorithm 7
$\overline{V}_{h,k}$	Value function obtained by running policy π^k on $\overline{M}^k$ with a clipping of threshold $2(H-h+1)$
$\underline{V}_{h,k}$	Value function obtained by running policy π^k on $\underline{M}^k$ with a clipping of threshold $2(H-h+1)$
$\overline{\overline{V}}_{h,k}$	Value function obtained by running policy π^k on $\overline{\overline{M}}^k$ with a clipping of threshold $2(H-h+1)$
$\mathcal{R}^k_{h,s,a}$	$\hat{R}^k_{h,s,a}-R_{h,s,a}$
$\mathcal{P}^k_{h,s,a}$	$\left\langle \hat{P}^k_{h,s,a}-P_{h,s,a},V^*_{h+1}\right\rangle$
$\mathcal{H}^k_h$	The historical observations and actions till time h in episode k, $\left\{(s^j_l,a^j_l,r^j_l): j\le k \text{ and } l\le H \text{ if } j<k, \text{ else } l\le h\right\}$
$\overline{\mathcal{H}}^k_h$	The historical observations and actions till time h and episode k, plus the randomness in episode k, $\mathcal{H}^k_h\cup\{\hat{z}_k\}$
$\mathbb{V}(P,V)$	Variance of $V\in\mathbb{R}^S$ under distribution $P\in\Delta^S$, $\sum_{s\in\mathcal{S}}P(s)(V(s)-\langle P,V\rangle)^2$
α_k	$200H^2\log(2HSAk^2)\log(40k^4)$
$\sigma^k_{\mathrm{ty}}(h,s,a)$	Magnitude of perturbation. ty $\in\{\mathrm{Ho},\mathrm{Be}\}$
ty	Reserved subscript for denoting perturbation type, ty $\in\{\mathrm{Ho},\mathrm{Be}\}$, where "Ho" denotes Hoeffding-type and "Be" denotes Bernstein-type
$\sigma^k_{\mathrm{Ho}}(h,s,a)$	$H\sqrt{\frac{\log(2HSAk^2)}{n_k(h,s,a)+1}}+\frac{H}{n_k(h,s,a)+1}$
$\sigma^k_{\mathrm{Be}}(h,s,a)$	$\sqrt{\frac{16\mathbb{V}\left(\tilde{P}^k_{h,s,a},\overline{V}_{k,h+1}\right)\log(2HSAk^2)}{n_k(h,s,a)+1}}+\frac{65H\log(2HSAk^2)}{n_k(h,s,a)+1}+\sqrt{\frac{\log(2HSAk^2)}{n_k(h,s,a)+1}}$
$\sqrt{e^k_{\mathrm{Ho}}(h,s,a)}$	$H\sqrt{\frac{\log(2HSAk^2)}{n_k(h,s,a)+1}}+\frac{H}{n_k(h,s,a)+1}$

$\sqrt{e^k_{\mathrm{Be}}(h,s,a)}$	$\sqrt{\frac{6\mathbb{V}\left(\tilde{P}^k_{h,s,a},V^*_{h+1}\right)\log(2HSAk^2)}{n_k(h,s,a)+1}}+\frac{9H\log(2HSAk^2)}{n_k(h,s,a)+1}+\sqrt{\frac{\log(2HSAk^2)}{n_k(h,s,a)+1}}$
C_1	$\frac{1}{\Phi(1.5)-\Phi(1)}$

C.2 Good Events

Definition C.2.1. Let $M'=(H,\mathcal{S},\mathcal{A},P',R',s_1)$. We define the following confidence sets for both Bernstein-type and Hoeffding-type noise

$$\mathcal{M}^k_{\mathrm{ty}}=\left\{M':\forall\left(h,s,a\right),\left|\left(R'_{h,s,a}-R_{h,s,a}\right)+\left\langle P'_{h,s,a}-P_{h,s,a},V^*_{h+1}\right\rangle\right|\leq\sqrt{e^k_{\mathrm{ty}}\left(h,s,a\right)}\right\},\tag{C.1}$$

where the confidence widths are set as

$$\begin{aligned}\sqrt{e^k_{\mathrm{Be}}\left(h,s,a\right)}=&\sqrt{\frac{6\mathbb{V}\left(\tilde{P}^k_{h,s,a},V^*_{h+1}\right)\log\left(2HSAk^2\right)}{n_k\left(h,s,a\right)+1}}\\&+\frac{9H\log\left(2HSAk^2\right)}{n_k\left(h,s,a\right)+1}+\sqrt{\frac{\log\left(2HSAk^2\right)}{n_k\left(h,s,a\right)+1}},\end{aligned}\tag{C.2}$$

$$\sqrt{e^k_{\mathrm{Ho}}\left(h,s,a\right)}=H\sqrt{\frac{\log\left(2HSAk^2\right)}{n_k\left(h,s,a\right)+1}}+\frac{H}{n_k\left(h,s,a\right)+1}.\tag{C.3}$$

We also define two events $\mathcal{E}^1_k$ and $\mathcal{E}^2_k$ as the following:

$$\mathcal{E}^1_k=\left\{\left|\hat{R}^k_{h,s,a}-R_{h,s,a}\right|\leq\sqrt{\frac{\log\left(2HSAk^2\right)}{n_k\left(h,s,a\right)+1}}+\frac{1}{n_k\left(h,s,a\right)+1},\ \forall\left(h,s,a\right)\right\},\tag{C.4}$$

$$\begin{aligned}\mathcal{E}^2_k=\Bigg\{\left|\left\langle\hat{P}^k_{h,s,a}-P_{h,s,a},V^*_{h+1}\right\rangle\right|\leq&\sqrt{\frac{6\mathbb{V}\left(\tilde{P}^k_{h,s,a},V^*_{h+1}\right)\log\left(2HSAk^2\right)}{n_k\left(h,s,a\right)+1}}\\&+\frac{8H\log\left(2HSAk^2\right)}{n_k\left(h,s,a\right)+1},\forall\left(h,s,a\right)\Bigg\}.\end{aligned}\tag{C.5}$$

We have the following lemmas about concentration of events.

Lemma C.2.2. *For fixed (k,h,s,a), let $n=n_k\left(h,s,a\right)$. Then, if $n\geq1$, for any fixed $\delta>0$,*

we have

$$\mathbb{P}\left(|\hat{R}^k_{h,s,a} - R_{h,s,a}| \geq \sqrt{\frac{\log(2/\delta)}{n+1}} + \frac{1}{n+1}\right) \leq \delta.$$

Proof. Let $\hat{R}^k_{h,s,a} = \frac{1}{n+1}\sum_{i=1}^n r_{(h,s,a),i}$, where $r_{(h,s,a),i} \sim \mathscr{R}_{h,s,a}$ are i.i.d. samples. By definition of the MDP, we have $\mathbb{E}\left[r_{(h,s,a),i}\right] = R_{h,s,a}$. Then, notice that

$$\hat{R}^k_{h,s,a} = \frac{1}{n+1}\sum_{i=1}^n r_{(h,s,a),i} = \frac{1}{n}\sum_{i=1}^n r_{(h,s,a),i} - \frac{1}{n(n+1)}\sum_{i=1}^n r_{(h,s,a),i}.$$

Since the reward is assumed to be bounded in $[0,1]$, we have $\frac{1}{n(n+1)}\sum_{i=1}^n r_{(h,s,a),i} \leq \frac{1}{n+1}$. Then, for fixed $\delta > 0$, we have

$$\begin{aligned}
&\mathbb{P}\left(|\hat{R}^k_{h,s,a} - R_{h,s,a}| \geq \sqrt{\frac{\log(2/\delta)}{n+1}} + \frac{1}{n+1}\right) \\
=&\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^n r_{(h,s,a),i} - R_{h,s,a} - \frac{1}{n(n+1)}\sum_{i=1}^n r_{(h,s,a),i}\right| \geq \sqrt{\frac{\log(2/\delta)}{n+1}} + \frac{1}{n+1}\right) \\
\leq&\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^n r_{(h,s,a),i} - R_{h,s,a}\right| + \frac{1}{n+1} \geq \sqrt{\frac{\log(2/\delta)}{n+1}} + \frac{1}{n+1}\right) \quad \text{(By triangle inequality)} \\
\leq&\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^n r_{(h,s,a),i} - R_{h,s,a}\right| \geq \sqrt{\frac{\log(2/\delta)}{2n}}\right) \qquad \text{(Since } n+1 \leq 2n \text{ for } n \geq 1) \\
\leq&\delta. \qquad \text{(By standard Hoeffding's inequality)}
\end{aligned}$$

□

Lemma C.2.3. *For fixed (k,h,s,a), let $n = n_k(h,s,a)$ and $V \in \mathbb{R}^S$ be some non-negative value function such that $\|V\|_\infty \leq H$. Then, if $n \geq 1$, for any fixed $\delta > 0$, we have*

$$\mathbb{P}\left(\left|\left\langle \hat{P}^k_{h,s,a} - P_{h,s,a}, V\right\rangle\right| \geq H\sqrt{\frac{\log(2/\delta)}{n+1}} + \frac{H}{n+1}\right) \leq \delta, \tag{C.6}$$

$$\mathbb{P}\left(\left|\left\langle \hat{P}^k_{h,s,a} - P_{h,s,a}, V\right\rangle\right| \geq \sqrt{\frac{6\mathbb{V}\left(\tilde{P}^k_{h,s,a}, V\right)\log(2/\delta)}{n+1}} + \frac{8H\log(2/\delta)}{n+1}\right) \leq \delta. \tag{C.7}$$

Proof. For fixed (h,s,a), we generate n i.i.d. samples of $s_{(h,s,a),i} \sim P_{h,s,a}$ and consider $V\left(s_{(h,s,a),i}\right)$. Then, by taking $n_k(h,s,a) = n$, we have

$$\left\langle \hat{P}^k_{h,s,a}, V \right\rangle = \frac{1}{n}\sum_{i=1}^{n} V\left(s_{(h,s,a),i}\right) - \frac{1}{n(n+1)}\sum_{i=1}^{n} V\left(s_{(h,s,a),i}\right).$$

The first result in equation (C.6) can be proved very similarly as Lemma C.2.2 using Hoeffding's inequality by simply replacing the upper bound of 1 in reward by H.

Then, for second result, we first consider $n \geq 2$. For some $\delta > 0$, define

$$b_{(h,s,a),n} = \sqrt{\frac{2\mathbb{V}\left(\tilde{P}^k_{h,s,a}, V\right)\log(2/\delta)}{n-1}} + \frac{7H\log(2/\delta)}{3(n-1)} + \frac{H}{n+1}.$$

By noticing that $F(s) \leq H$ and applying similar technique in proof of Lemma C.2.2, we have

$$\begin{aligned}
&\mathbb{P}\left(\left|\left\langle \hat{P}^k_{h,s,a} - P_{h,s,a}, V\right\rangle\right| \geq b_{(h,s,a),n}\right) \\
\leq &\mathbb{P}\left(\left|\left\langle \tilde{P}^k_{h,s,a} - P_{h,s,a}, V\right\rangle\right| \geq \sqrt{\frac{2\mathbb{V}\left(\tilde{P}^k_{h,s,a}, V\right)\log(2/\delta)}{n-1}} + \frac{7H\log(2/\delta)}{3(n-1)}\right) \\
\leq &\delta. \qquad \text{(By Lemma C.9.2, the empirical Bernstein's inequality)}
\end{aligned}$$

Then, since $3(n-1) \geq n+1$ when $n \geq 2$, we can easily check that

$$b_{(h,s,a),n} \leq \sqrt{\frac{6\mathbb{V}\left(\tilde{P}^k_{h,s,a}, V\right)\log(2/\delta)}{n+1}} + \frac{8H\log(2/\delta)}{n+1}.$$

Finally, since $\|V\|_\infty \leq H$, when $n = 1$, we trivially have

$$\left|\left\langle \hat{P}^k_{h,s,a} - P_{h,s,a}, V\right\rangle\right| \leq H \leq \sqrt{\frac{6\mathbb{V}\left(\tilde{P}^k_{h,s,a}, V\right)\log(2/\delta)}{n+1}} + \frac{8H\log(2/\delta)}{n+1}.$$

Therefore, we can conclude that

$$\mathbb{P}\left(\left|\left\langle \hat{P}^k_{h,s,a} - P_{h,s,a}, V\right\rangle\right| \geq \sqrt{\frac{6\mathbb{V}\left(\tilde{P}^k_{h,s,a}, V\right)\log\left(2/\delta\right)}{n+1}} + \frac{8H\log\left(2/\delta\right)}{n+1}\right) \leq \delta.$$

□

Lemma C.2.4. $\sum_{k=1}^{\infty}\mathbb{P}\left(\left(\mathcal{E}^1_k\right)^c\right) \leq \frac{\pi^2}{6}$.

Proof. Let $n = n_k\left(h,s,a\right)$. Then, for some fixed (h,s,a), $n \geq 1$ and $\delta_n > 0$, by Lemma C.2.2, we have

$$\mathbb{P}\left(|\hat{R}^k_{h,s,a} - R_{h,s,a}| \geq \sqrt{\frac{\log\left(2/\delta_n\right)}{n+1}} + \frac{1}{n+1}\right) \leq \delta_n.$$

Therefore, by taking $\delta_n = \frac{1}{HSAn^2}$, a union bound will give us

$$\sum_{n=1}^{\infty}\sum_{h,s,a}\mathbb{P}\left(\left\{|\hat{R}^k_{h,s,a} - R_{h,s,a}| \geq \sqrt{\frac{\log\left(2HSAn^2\right)}{n+1}} + \frac{1}{n+1}\right\}\right) \leq \sum_{n=1}^{\infty}\frac{1}{n^2} = \frac{\pi^2}{6}.$$

Therefore, we have

$$\sum_{k=1}^{\infty}\mathbb{P}\left(\exists\left(h,s,a\right) : n_k\left(h,s,a\right) > 0, |\hat{R}^k_{h,s,a} - R_{h,s,a}| \geq \sqrt{\frac{\log\left(2HSAn_k\left(h,s,a\right)^2\right)}{n_k\left(h,s,a\right)+1}} + \frac{1}{n_k\left(h,s,a\right)+1}\right)$$
$$\leq \frac{\pi^2}{6}.$$

Since the MDP is time-inhomogeneous, each (h,s,a) can only be visited at most once during one episode, which implies $n_k\left(h,s,a\right) \leq k$. Therefore, we have

$$\sqrt{\frac{\log\left(2HSAn_k\left(h,s,a\right)\right)}{n_k\left(h,s,a\right)+1}} + \frac{1}{n_k\left(h,s,a\right)+1} \leq \sqrt{\frac{\log\left(2HSAk^2\right)}{n_k\left(h,s,a\right)+1}} + \frac{1}{n_k\left(h,s,a\right)+1}$$

and thus the proof is complete. □

Lemma C.2.5. $\sum_{k=1}^{\infty}\mathbb{P}\left(\left(\mathcal{E}^2_k\right)^c\right) \leq \frac{\pi^2}{6}$.

Proof. This proof will be very similar to proof of Lemma C.2.4. In specific, for fixed (h, s, a), let $n = n_k(h, s, a) \geq 1$. Then, for any $\delta_n > 0$, since $\left\|V^*_{h+1}\right\|_\infty \leq H$, by Lemma C.2.3, we have

$$\mathbb{P}\left(\left|\left\langle \hat{P}^k_{h,s,a} - P_{h,s,a}, V^*_{h+1}\right\rangle\right| \geq \sqrt{\frac{6\mathbb{V}\left(\tilde{P}^k_{h,s,a}, V^*_{h+1}\right)\log\left(2/\delta_n\right)}{n+1}} + \frac{8H\log\left(2/\delta_n\right)}{n+1}\right) \leq \delta_n.$$

Therefore, by taking $\delta_n = \frac{1}{HSAn^2}$ and applying a similar union bound argument used in the proof of Lemma C.2.4, we can conclude $\sum_{k=1}^{\infty}\mathbb{P}\left(\left(\mathcal{E}^2_k\right)^c\right) \leq \frac{\pi^2}{6}$. □

We further define the event $\mathcal{C}^k_{\text{ty}} = \left\{\hat{M}^k \in \mathcal{M}^k_{\text{ty}}\right\}$. With what we have proved above, it will be straightforward to show the following results about $\mathcal{C}^k_{\text{ty}}$.

Lemma C.2.6. $\sum_{k=1}^{\infty}\mathbb{P}\left(\left(\mathcal{C}^k_{\text{Be}}\right)^c\right) = \sum_{k=1}^{\infty}\mathbb{P}\left(\hat{M}^k \notin \mathcal{M}^k_{\text{Be}}\right) \leq \frac{\pi^2}{3}$

Proof. We can easily notice $\mathcal{E}^1_k \cap \mathcal{E}^2_k \implies \hat{M}^k \in \mathcal{M}^k_{\text{Be}}$, which implies $\hat{M}^k \notin \mathcal{M}^k_{\text{Be}} \implies \left(\mathcal{E}^1_k\right)^c \cup \left(\mathcal{E}^2_k\right)^c$. The first result then follows straightforwardly by applying Lemma C.2.4 and Lemma C.2.5. □

Lemma C.2.7. $\sum_{k=1}^{\infty}\mathbb{P}\left(\left(\mathcal{C}^k_{\text{Ho}}\right)^c\right) = \sum_{k=1}^{\infty}\mathbb{P}\left(\hat{M}^k \notin \mathcal{M}^k_{\text{Ho}}\right) \leq \frac{\pi^2}{3}$.

Proof. Similarly, for fixed (h, s, a), we generate n i.i.d. samples $s_{(h,s,a),i} \sim P_{h,s,a}$ and $r_{(h,s,a),i} \sim \mathscr{R}_{h,s,a}$ for $i = 1, \ldots, n$ respectively. Define $Y_{(h,s,a),i} = r_{(h,s,a),i} + V^*_{h+1}(s_{(h,s,a),i})$ and we have $\mathbb{E}\left[Y_{(h,s,a),i}\right] = R_{h,s,a} + \left\langle P_{h,s,a}, V^*_{h+1}\right\rangle$.

By definition of MDP, we know that $Y_{(h,s,a),i} \leq H$. Thus, we can use an argument similar to the proof of Lemma C.2.2. In specific, let $n = n_k(h, s, a)$ and for $\delta_n > 0$, we have

$$\begin{aligned}
&\mathbb{P}\left(\left|\frac{1}{n+1}\sum_{i=1}^{n} Y_{(h,s,a),i} - \mathbb{E}\left[Y_{(h,s,a),i}\right]\right| \geq H\sqrt{\frac{\log\left(2/\delta_n\right)}{n+1}} + \frac{H}{n+1}\right) \\
=&\mathbb{P}\left(\left|\left(\hat{R}^k_{h,s,a} - R_{h,s,a}\right) + \left\langle \hat{P}^k_{h,s,a} - P_{h,s,a}, V^*_{h+1}\right\rangle\right| \geq H\sqrt{\frac{\log\left(2/\delta_n\right)}{n+1}} + \frac{H}{n+1}\right) \\
\leq&\delta_n.
\end{aligned}$$

Then, we can take $\delta_n = \frac{1}{HSAn^2}$ and apply a similar union bound argument in used in the proof of Lemma C.2.4. As a result, we can obtain

$$\sum_{k=1}^{\infty} \mathbb{P}\left(\hat{M}^k \notin \mathcal{M}_{\text{Ho}}^k\right) \leq \frac{\pi^2}{6} \leq \frac{\pi^2}{3}.$$

□

We can also have well-behaved bounds on magnitude of noise and estimated value functions.

Definition C.2.8. We define $w_{\text{ty}}^k(h,s,a) = \sigma_{\text{ty}}^k(h,s,a)\hat{z}_k$ and $\gamma_{\text{ty}}^k(h,s,a) = \sigma_{\text{ty}}^k(h,s,a)\sqrt{\log(40k^4)}$. We define the event $\mathcal{E}_k^w$ as

$$\mathcal{E}_k^w = \left\{\forall(h,s,a), |w_{\text{ty}}^k(h,s,a)| \leq \gamma_{\text{ty}}^k(h,s,a)\right\}.$$

Lemma C.2.9. $\sum_{k=1}^K \mathbb{P}\left(\left(\mathcal{E}_k^w\right)^c\right) \leq \frac{\pi^2}{3}$ *regardless the type of noise we choose.*

Proof. For any $k \in [K]$, by the tail bound of Gaussian distribution,

$$\mathbb{P}\left(|\hat{z}_k| \geq \sqrt{\log\left(40k^4\right)}\right) \leq 2\exp\left(-\frac{\log\left(40k^4\right)}{2}\right) \leq \frac{2}{k^2}.$$

Summing over $k \in [K]$,

$$\sum_{k=1}^K \mathbb{P}\left(\left(\mathcal{E}_w^k\right)^c\right) = \sum_{k=1}^K \mathbb{P}\left(|\hat{z}_k| \geq \sqrt{\log\left(40k^4\right)}\right) \leq \sum_{k=1}^{\infty} \frac{2}{k^2} \leq \frac{\pi^2}{3}.$$

Note that this result does not depend on the type of noise we choose. □

Now, we define the following good events that hold with high probability and will be used throughout the whole proof.

Definition C.2.10 (Good events $\mathcal{G}_k$)**.** Let $\mathcal{G}_{k,\text{ty}} = \left\{\mathcal{C}_{\text{ty}}^k \cap \mathcal{E}_k^w\right\}$.

The subscript "ty" will be ignored later since it is clear from the context.

Definition C.2.11. With $\alpha_k = 200H^2 \log(2HSAk^2)\log(40k^4)$, we define events $\mathcal{E}_{h,k}^{th}$ and $\mathcal{E}_{h,k}^{cum}$ as

$$\mathcal{E}_{h,k}^{th} = \left\{ n_k(h, s_h^k, a_h^k) \geq \alpha_k \right\}, \quad \mathcal{E}_{h,k}^{cum} = \bigcap_{i=1}^{h} \mathcal{E}_{i,k}^{th}. \tag{C.8}$$

We will show that under events $\mathcal{E}_k^w$, $\mathcal{E}_{h,k}^{th}$ and $\hat{M}^k \in \mathcal{M}_{\text{ty}}^k$, no clipping happens on s_h^k.

Lemma C.2.12. *Assume that $\mathcal{E}_k^w$, $\mathcal{E}_{h,k}^{th}$ and $\hat{M}^k \in \mathcal{M}_{\text{ty}}^k$ hold. Then, regardless the type of noise we choose, it holds that*

$$|\overline{Q}_{h,k}(s_h^k, a_h^k)| \leq 2(H - h + 1),$$

which immediately tells us that no clipping is triggered for any (s_h^k, a_h^k).

Proof. We have that

$$\overline{Q}_{h,k}(s_h^k, a_h^k) = \hat{R}_{h,s_h^k,a_h^k}^k + \left\langle \hat{P}_{h,s_h^k,a_h^k}^k, \overline{V}_{h+1,k} \right\rangle + \sigma_{\text{ty}}^k(h, s_h^k, a_h^k)\hat{z}_k.$$

As we have $\left|\overline{V}_{h+1,k}\right| \leq 2(H-h)$ by clipping and $\hat{R}_{h,s_h^k,a_h^k}^k \in [0,1]$, we only need to show that $\sigma_{\text{ty}}^k(h, s_h^k, a_h^k)\hat{z}_k \leq 1$. Under event $\mathcal{E}_w^k$, we have $\left|\sigma_{\text{ty}}^k(h, s_h^k, a_h^k)\hat{z}_k\right|$ is bounded by $\gamma_{\text{ty}}^k(h, s_h^k, a_h^k) = \sigma_{\text{ty}}^k(h, s_h^k, a_h^k)\sqrt{\log(40k^4)}$. Note that we have $\overline{V}_{h+1,k}(s) \in [-2H, 2H]$ by clipping for any $s \in \mathcal{S}$. Thus, by Lemma C.9.3, we have $\mathbb{V}\left(\tilde{P}_{h,s,a}^k, \overline{V}_{h+1,k}\right) \leq 4H^2$ for any (h, s, a).

By taking $\alpha_k = 200H^2 \log(2HSAk^2)\log(40k^4)$ and referring to the definitions of $\sigma_{\text{Be}}^k(h, s, a)$ in Equation (4.6), we can check that

$$\begin{aligned}
&\gamma_{\text{Be}}^k(h, s_h^k, a_h^k) \\
=&\sigma_{\text{Be}}^k(h, s_h^k, a_h^k)\sqrt{\log(40k^4)} \\
=&\left(\sqrt{\frac{16\mathbb{V}\left(\tilde{P}_{h,s_h^k,a_h^k}^k, \overline{V}_{k,h+1}\right)\log\left(2HSAk^2\right)}{n_k\left(h, s_h^k, a_h^k\right)+1}} + \frac{65H\log\left(2HSAk^2\right)}{n_k\left(h, s_h^k, a_h^k\right)+1} \right)\sqrt{\log(40k^4)} \\
&\quad + \sqrt{\frac{\log\left(2HSAk^2\right)}{n_k\left(h, s_h^k, a_h^k\right)+1}} \cdot \sqrt{\log(40k^4)}
\end{aligned}$$

$$\leq \left(\sqrt{\frac{64H^2 \log\left(2HSAk^2\right)}{\alpha_k}} + \frac{65H \log\left(2HSAk^2\right)}{\alpha_k} + \sqrt{\frac{\log\left(2HSAk^2\right)}{\alpha_k}} \right) \sqrt{\log(40k^4)}$$

(Event $\mathcal{E}_k^{th}$ implies $n_k(h, s_h^k, a_h^k) \geq \alpha_k$)

$$\leq \sqrt{\frac{64}{200}} + \frac{65}{200H} + \sqrt{\frac{1}{200H^2}} \leq 1.$$

Thus, we have $\gamma_{\mathrm{Be}}^k(h, s, a) \leq 1$ and we can similarly check that $\gamma_{\mathrm{Ho}}^k(h, s, a) \leq 1$. As a result, we have

$$\left| \overline{Q}_{h,k}(s_h^k, a_h^k) \right| \leq 2(H - h + 1),$$

which completes the proof. □

C.3 Optimism

Let $\mathcal{H}_h^k$ denote the history trajectory, which is defined as

$$\mathcal{H}_h^k = \left\{ (s_l^j, a_l^j, r_l^j) : j \leq k \text{ and } l \leq H \text{ if } j < k, \text{ else } l \leq h \right\}. \tag{C.9}$$

We will prove that for both types of noise, $\overline{V}_{h,k}$ is optimistic with constant probability under certain conditions.

C.3.1 Hoeffding-type Noise

Lemma C.3.1. *Condition on history $\mathcal{H}_H^{k-1}$, if $\mathcal{G}_{k,\mathrm{Ho}}$ holds and Hoeffding-based noise is applied, then $\overline{V}_{h,k}$ is optimisitic with constant probability for any $h \in [H]$. Specifically, we have*

$$\mathbb{P}\left(\overline{V}_{h,k}(s) \geq V_h^*(s), \forall h \in [H], s \in \mathcal{S} \mid \mathcal{H}_H^{k-1}, \mathcal{G}_{k,\mathrm{Ho}} \right) \geq \Phi(1.9) - \Phi(1) := C_{\mathrm{Ho}}.$$

Proof. We will show that if $\hat{z}_k \geq 1$, then for all $h \in [H]$ and $s \in \mathcal{S}$, we have $\overline{V}_{h,k}(s) \geq V_h^*(s)$. The proof will use induction and the argument is true for $h = H + 1$ as $\overline{V}_{H+1,k}(s) = V_{H+1}^*(s) = 0$. Suppose the argument is true for timestep $h + 1$ and for timestep h we have

$$
\begin{aligned}
\overline{V}_{h,k}(s) =& \mathrm{clip}_{2(H-h+1)}\left(\max_{a\in\mathcal{A}} \overline{Q}_{h,k}(s,a)\right) \\
\geq& \min\left\{2(H-h+1), \max_{a\in\mathcal{A}} \overline{Q}_{h,k}(s,a)\right\} \\
\geq& \min\left\{(H-h+1), \overline{Q}_{h,k}(s,\pi_h^*(s))\right\} \\
\geq& \min\left\{(H-h+1), \hat{R}^k_{h,s,\pi_h^*(s)} + \left\langle \hat{P}^k_{h,s,\pi_h^*(s)}, \overline{V}_{h+1,k}\right\rangle + \sigma^k_{\mathrm{ty}}(h,s,\pi_h^*(s))\hat{z}_k\right\} \\
\geq& \min\left\{(H-h+1), \hat{R}^k_{h,s,\pi_h^*(s)} + \left\langle \hat{P}^k_{h,s,\pi_h^*(s)}, V^*_{h+1,k}\right\rangle + \sigma^k_{\mathrm{ty}}(h,s,\pi_h^*(s))\hat{z}_k\right\} \\
& \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{(Inductive hypothesis)} \\
\geq& \min\left\{(H-h+1), R^k_{h,s,\pi_h^*(s)} + \left\langle P^k_{h,s,\pi_h^*(s)}, V^*_{h+1,k}\right\rangle\right\} \\
& \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{(Since } \hat{M}^k \in \mathcal{M}^k_{\mathrm{Ho}} \text{ inferred by } \mathcal{G}_{k,\mathrm{Ho}} \text{ and } \hat{z}_k \geq 1) \\
\geq& \min\left\{(H-h+1), Q^*_h(s,\pi_h^*(s))\right\} \\
\geq& V^*_h(s).
\end{aligned}
$$

Then by induction we have that the optimism is achieved for all $h \in [H]$ and $s \in \mathcal{S}$ simultaneously. Meanwhile, as stated in Definition C.2.8, we have $\hat{z}_k \leq \sqrt{\log(40k^4)}$ under event $\mathcal{E}^w_k$ and numerically, $\sqrt{\log(40k^4)} \geq 1.9$. Therefore, the probability that $\hat{z}_k \geq 1$ under $\mathcal{E}^w_k$, inferred by $\mathcal{G}_{k,\mathrm{Ho}}$, is at least

$$
\mathbb{P}\left(\hat{z}_k \geq 1 \mid \mathcal{H}^{k-1}_H, \mathcal{G}_{k,\mathrm{Ho}}\right) = \frac{\Phi(1.9)-\Phi(1)}{\Phi(1.9)-\Phi(-1.9)} \geq \Phi(1.9)-\Phi(1) := C_{\mathrm{Ho}}.
$$

Thus, we can conclude that

$$
\mathbb{P}\left(\overline{V}_{h,k}(s) \geq V^*_h(s), \forall h \in [H], s \in \mathcal{S} \mid \mathcal{H}^{k-1}_H, \mathcal{G}_{k,\mathrm{Ho}}\right) \geq C_{\mathrm{Ho}}.
$$

□

C.3.2 Bernstein-type Noise

The following proof of optimism applies some techniques used in Zhang et al. [2020b]. We first present a technical lemma.

Lemma C.3.2. *Let* $f_z : \Delta^S \times \mathbb{R}_+^S \times \mathbb{R} \times \mathbb{R} \mapsto \mathbb{R}$ *with* $f_z(p,v,n,L) = \frac{n}{n+1}\langle p,v\rangle + \max\left\{4\sqrt{\frac{\mathbb{V}(p,v)L}{n+1}}, \frac{64HL}{n+1}\right\}\cdot z$ *for some constant* $H>0$ *and* $z\in\mathbb{R}$. *Then,* f_z *satisfies*

(i) $f_z(p,v,n,L)$ *is non-decreasing in* $v(s)$ *for all* $p\in\Delta^S$, $\|v\|_\infty \le 2H$, $L>0$, $n\ge 3$ *and* $z\in[-1.5,1.5]$

(ii) $f_z(p,v,n,L) \ge \frac{n}{n+1}\langle p,v\rangle + \left(3\sqrt{\frac{\mathbb{V}(p,v)L}{n+1}} + \frac{8HL}{n+1}\right)\cdot z$ *for* $z\in[1,1.5]$.

(iii) $f_z(p,v,n,L) \le \frac{n}{n+1}\langle p,v\rangle + \left(3\sqrt{\frac{\mathbb{V}(p,v)L}{n+1}} + \frac{8HL}{n+1}\right)\cdot z$ *for* $z\in[-1.5,-1]$.

Proof. It is obvious that $f_z(p,v,n,L)$ is continuous in $v(s)$ and not differentiable at only one point where $4\sqrt{\frac{\mathbb{V}(p,v)L}{n+1}} = \frac{64HL}{n+1}$. Therefore, to prove statement (i), we only need to show that $\frac{\partial f_z(p,v,n,L)}{\partial v(s)} \ge 0$. Specifically, we have

$$\begin{aligned}
\frac{\partial f_z(p,v,n,L)}{\partial v(s)} &= \frac{n}{n+1}\cdot p(s) + \mathbb{1}\left\{4\sqrt{\frac{\mathbb{V}(p,v)L}{n+1}} \ge \frac{64HL}{n+1}\right\}\frac{4p(s)(v(s)-\langle p,v\rangle)L}{\sqrt{(n+1)\mathbb{V}(p,v)L}}\cdot z \\
&\overset{(a)}{\ge} \frac{n}{n+1}\cdot p(s) + \mathbb{1}\left\{4\sqrt{\frac{\mathbb{V}(p,v)L}{n+1}} \ge \frac{64HL}{n+1}\right\}\frac{-8HL}{\sqrt{(n+1)\mathbb{V}(p,v)L}}\cdot z \\
&\overset{(b)}{\ge} p(s)\left(\frac{n}{n+1} - \frac{z}{2}\right) \\
&\ge 0.
\end{aligned}$$

Here, The inequality (a) holds because $\|v\|_\infty \le 2H$ and v is non-negative, which means to have $v(s) - \langle p,v\rangle \ge -2H$. The inequality (b) above holds because when the condition inside indicator $\mathbb{1}\{\cdot\}$ holds, we will have $\sqrt{(n+1)\mathbb{V}(p,v)L} \ge 16HL$. The last inequality holds because we have $n\ge 3$ and $z\le 1.5$. Therefore, $f_z(p,v,n,L)$ is non-decreasing in $v(s)$.

For the statement (ii), we consider two cases. First, when $4\sqrt{\frac{\mathbb{V}(p,v)L}{n+1}} \ge \frac{64HL}{n+1}$ holds, we have $\frac{8HL}{n+1} \le \frac{1}{2}\sqrt{\frac{\mathbb{V}(p,v)L}{n+1}}$, which means to have

$$\frac{n}{n+1}\langle p,v\rangle + \left(3\sqrt{\frac{\mathbb{V}(p,v)L}{n+1}} + \frac{8HL}{n+1}\right)\cdot z \le \frac{n}{n+1}\langle p,v\rangle + \frac{7z}{2}\sqrt{\frac{\mathbb{V}(p,v)L}{n+1}} \le f_z(p,v,n,L).$$

When $4\sqrt{\frac{\mathbb{V}(p,v)L}{n+1}} \leq \frac{64HL}{n+1}$ holds, we have $3\sqrt{\frac{\mathbb{V}(p,v)L}{n+1}} \leq \frac{48HL}{n+1}$, which similarly leads to

$$\frac{n}{n+1}\langle p, v\rangle + \left(3\sqrt{\frac{\mathbb{V}(p,v)L}{n+1}} + \frac{8HL}{n+1}\right)\cdot z \leq f_z(p,v,n,L).$$

The state (iii) can be shown similarly and thus the proof is complete. □

Lemma C.3.3. *Condition on history $\mathcal{H}_H^{k-1}$, if $\mathcal{G}_{k,\text{Ho}}$ holds and Bernstein-based noise is applied, then $\overline{V}_{h,k}$ is optimisitic with constant probability for any $h \in [H]$. Specifically, we have*

$$\mathbb{P}\left(\overline{V}_{h,k}(s) \geq V_h^*(s), \forall h \in [H], s \in \mathcal{S} \mid \mathcal{H}_H^{k-1}, \mathcal{G}_{k,\text{Be}}\right) \geq \Phi(1.5) - \Phi(1) := C_{\text{Be}}$$

Proof. Similar to what we have discussed in the proof of Lemma C.3.1, under event $\mathcal{E}_k^w$, we have $\hat{z}_k \in [1, 1.5]$ with probability at least $\Phi(1.5) - \Phi(1) = C_{\text{Be}}$. Then, we will show that $\overline{Q}_{h,k}(s,a) \geq Q_h^*(s,a)$ for any h with arbitrary s, a and $\hat{z}_k \in [1, 1.5]$. The proof will use induction. For simplicity, let $L = \log(2HSAk^2)$.

For $h = H+1$, the inequality holds trivially because both sides are 0. Then, by assuming $\overline{Q}_{h+1,k}(s,a) \geq Q_h^*(s,a)$ for any (s,a) such that $n_k(h,s,a) \geq 3$, we have

$$\begin{aligned}
\overline{Q}_{h,k}(s,a) &= \hat{R}_{h,s,a}^k + \left\langle \hat{P}_{h,s,a}^k, \overline{V}_{h+1,k}\right\rangle + \sigma_{\text{Be}}^k(h,s,a)\,\hat{z}_k \\
&\geq R_{h,s,a} + \left\langle \hat{P}_{h,s,a}^k, \overline{V}_{h+1,k}\right\rangle + \left(4\sqrt{\frac{\mathbb{V}\left(\tilde{P}_{h,s,a}^k, \overline{V}_{h+1,k}\right)L}{n_k(h,s,a)+1}} + \frac{64HL}{n_k(h,s,a)+1}\right)\cdot \hat{z}_k \\
&\qquad \text{(Replace } \hat{R}_{h,s,a} \text{ by } R_{h,s,a} \text{ through applying event } \mathcal{E}_k^1 \text{ defined in (C.4))} \\
&\geq R_{h,s,a} + \left\langle \hat{P}_{h,s,a}^k, \overline{V}_{h+1,k}\right\rangle + \max\left\{4\sqrt{\frac{\mathbb{V}\left(\tilde{P}_{h,s,a}^k, \overline{V}_{h+1,k}\right)L}{n_k(h,s,a)+1}}, \frac{64HL}{n_k(h,s,a)+1}\right\}\cdot \hat{z}_k \\
&\overset{\text{(a)}}{\geq} R_{h,s,a} + \left\langle \hat{P}_{h,s,a}^k, V_{h+1}^*\right\rangle + \max\left\{4\sqrt{\frac{\mathbb{V}\left(\tilde{P}_{h,s,a}^k, V_{h+1}^*\right)L}{n_k(h,s,a)+1}}, \frac{64HL}{n_k(h,s,a)+1}\right\}\cdot \hat{z}_k \\
&\geq R_{h,s,a} + \left\langle \hat{P}_{h,s,a}^k, V_{h+1}^*\right\rangle + \left(3\sqrt{\frac{\mathbb{V}\left(\tilde{P}_{h,s,a}^k, V_{h+1}^*\right)L}{n_k(h,s,a)+1}} + \frac{8HL}{n_k(h,s,a)+1}\right)\cdot \hat{z}_k \\
&\qquad \text{(By applying statement (ii) of Lemma C.3.2)}
\end{aligned}$$

$$
\begin{aligned}
&\geq R_{h,s,a}+\left\langle \hat{P}^k_{h,s,a}, V^*_{h+1}\right\rangle + \sqrt{\frac{6\mathbb{V}\left(\tilde{P}^k_{h,s,a}, V^*_{h+1}\right)}{n_k(h,s,a)+1}} + \frac{8HL}{n_k(h,s,a)+1} && \text{(Since } \hat{z}_k \geq 1\text{)}\\
&\geq R_{h,s,a}+\langle P_{h,s,a}, V^*_{h+1}\rangle && \text{(By applying event } \mathcal{E}^2_k \text{ defined in (C.5))}\\
&= Q^*_h(s,a).
\end{aligned}
$$

Here, the above inequality (a) holds by applying inductive hypothesis and statement (i) in Lemma C.3.2. It is applicable because when $\mathcal{E}^w_k$ holds, and by the clipping function, $\left\|\overline{V}_{h+1,k}\right\|_\infty \leq 2H$. When $n_k(h,s,a) < 3$, $\overline{Q}_{h,k}(s,a) \geq Q^*_h(s,a)$ holds trivially because $Q^*_h(s,a) \leq H$ by definition. Therefore, the induction is complete.

Now, for arbitrary (k,h,s), set $a = \arg\max_{a\in\mathcal{A}} \overline{Q}_{h,k}(s,a)$ and we have

$$
\begin{aligned}
\overline{V}_{h,k}(s) =& \text{clip}_{2(H-h+1)}\left(\max_{a\in\mathcal{A}} \overline{Q}_{h,k}(s,a)\right)\\
\geq& \min\left\{2(H-h+1), \max_{a\in\mathcal{A}} \overline{Q}_{h,k}(s,a)\right\}\\
\geq& \min\left\{(H-h+1), \overline{Q}_{h,k}(s,\pi^*_h(s))\right\}\\
\geq& \min\left\{(H-h+1), Q^*_{h,k}(s,\pi^*_h(s))\right\}\\
\geq& V^*_{h,k}(s)
\end{aligned}
$$

□

C.4 Pessimism

Similar to what we have proved in Section C.3, in this section we will prove that for both types of noise, $\overline{V}_{h,k}$ is pessimistic with constant probability under certain conditions.

C.4.1 Hoeffding-type Noise

Lemma C.4.1. *Condition on history $\mathcal{H}^{k-1}_H$, if $\mathcal{G}_{k,\text{Ho}}$ holds and Hoeffding-based noise is applied, then $\overline{V}_{h,k}$ is optimisitic with constant probability for any $h \in [H]$. Specifically, we have*

$$
\mathbb{P}\left(\overline{V}_{h,k}(s) \leq V^*_h(s), \forall h\in[H], s\in\mathcal{S} \mid \mathcal{H}^{k-1}_H, \mathcal{G}_{k,\text{Ho}}\right) \geq \Phi(1.9) - \Phi(1) := C_{\text{Ho}}.
$$

Proof. We will show that if $\hat{z}_k \leq -1$, then for all $h \in [H]$ and $s \in \mathcal{S}$, we have $\overline{V}_{h,k}(s) \leq V_h^*(s)$. The proof will use induction and the argument is true for $h = H+1$ as $\overline{V}_{H+1,k}(s) = V_{H+1}^*(s) = 0$. Suppose the argument is true for timestep $h+1$ and we consider timestep h. Set $a = \arg\max_{a \in \mathcal{A}} \overline{Q}_{h,k}(s,a)$.

$$
\begin{aligned}
\overline{V}_{h,k}(s) =& \text{clip}_{2(H-h+1)}\left(\overline{Q}_{h,k}(s,a)\right) \\
\leq& \max\left\{-2(H-h+1), \overline{Q}_{h,k}(s,a)\right\} \\
\leq& \max\left\{-(H-h+1), \overline{Q}_{h,k}(s,a)\right\} \\
\leq& \max\left\{-(H-h+1), \hat{R}_{h,s,a}^k + \left\langle \hat{P}_{h,s,a}^k, \overline{V}_{h+1,k} \right\rangle + \sigma_{\text{ty}}^k(h,s,a)\hat{z}_k \right\} \\
\leq& \max\left\{-(H-h+1), \hat{R}_{h,s,a}^k + \left\langle \hat{P}_{h,s,a}^k, V_{h+1,k}^* \right\rangle + \sigma_{\text{ty}}^k(h,s,a)\hat{z}_k \right\} \\
& \qquad \text{(Induction Hypothesis)} \\
\leq& \max\left\{-(H-h+1), R_{h,s,a}^k + \left\langle P_{h,s,a}^k, V_{h+1,k}^* \right\rangle \right\} \\
& \qquad \text{(Since } \hat{M}^k \in \mathcal{M}_{\text{Ho}}^k \text{ and } \hat{z}_k \leq -1) \\
\leq& \max\left\{-(H-h+1), Q_h^*(s,a)\right\} \\
\leq& \max\left\{-(H-h+1), \max_{a \in \mathcal{A}} Q_h^*(s,a) \right\} \\
\leq& V_h^*(s).
\end{aligned}
$$

Then by induction we have that the optimism is achieved for all $h \in [H]$ and $s \in \mathcal{S}$ simultaneously. By using argument similar to the proof of Lemma C.3.1, we can see that when $\hat{z}_k \leq -1$, we have $\overline{V}_{h,k}(s) \leq V_h^*(s)$ an this hold simultaneously for any $h \in [H]$, $s \in \mathcal{S}$. Furtherm as stated in Definition C.2.8, we have $|\hat{z}_k| \leq \sqrt{\log(40k^4)}$ under event $\mathcal{E}_k^w$ and numerically, $\sqrt{\log(40k^4)} \geq 1.9$. Therefore, the probability that $\hat{z}_k \leq -1$ under $\mathcal{E}_k^w$ is at least

$$
\mathbb{P}\left(\hat{z}_k \leq -1 \mid \mathcal{H}_H^{k-1}, \mathcal{G}_{k,\text{Ho}}\right) = \frac{\Phi(1.9) - \Phi(1)}{\Phi(1.9) - \Phi(-1.9)} \geq \Phi(1.9) - \Phi(1) = C_{\text{Ho}}.
$$

Thus, we can conclude that

$$
\mathbb{P}\left(\overline{V}_{h,k}(s) \leq V_h^*(s), \forall h \in [H], s \in \mathcal{S} \mid \mathcal{H}_H^{k-1}, \mathcal{G}_{k,\text{Ho}}\right) \geq C_{\text{Ho}}.
$$

□

C.4.2 Bernstein-type Noise

Lemma C.4.2. *Condition on history $\mathcal{H}_H^{k-1}$, if $\mathcal{G}_{k,\mathrm{Ho}}$ holds and Bernstein-based noise is applied, then $\overline{V}_{h,k}$ is pessimistic with constant probability for any $h \in [H]$. Specifically, we have*

$$\mathbb{P}\left(\overline{V}_{h,k}(s) \leq V_h^*(s), \forall h \in [H], s \in \mathcal{S} \mid \mathcal{H}_H^{k-1}, \mathcal{G}_{k,\mathrm{Be}}\right) \geq C_{\mathrm{Be}}$$

Proof. Similar to what we have discussed in the proof of Lemma C.4.1, under event $\mathcal{E}_k^w$, we have $\hat{z}_k \in [-1.5, -1]$ with probability at least $\Phi(1.5) - \Phi(1) = C_{\mathrm{Be}}$. Then, we will show that $\overline{Q}_{h,k}(s,a) \leq Q_h^*(s,a)$ for any h with arbitrary s, a and $\hat{z}_k \in [-1.5, -1]$. The proof will go by induction. For simplicity, let $L = \log\left(2HSAk^2\right)$.

For $h = H+1$, the inequality holds trivially because both sides are 0. Then, by assuming $\overline{Q}_{h+1,k}(s,a) \leq Q_h^*(s,a)$ for any (s,a) such that $n_k(h,s,a) \geq 3$, we have

$$\begin{aligned}
\overline{Q}_{h,k}(s,a) &= \hat{R}_{h,s,a}^k + \left\langle \hat{P}_{h,s,a}^k, \overline{V}_{h+1,k} \right\rangle + \sigma_{\mathrm{Be}}^k(h,s,a)\,\hat{z}_k \\
&\leq R_{h,s,a} + \left\langle \hat{P}_{h,s,a}^k, \overline{V}_{h+1,k} \right\rangle - \left(4\sqrt{\frac{\mathbb{V}\left(\tilde{P}_{h,s,a}^k, \overline{V}_{h+1,k}\right) L}{n_k(h,s,a)+1}} + \frac{64HL}{n_k(h,s,a)+1} \right) \\
&\qquad \text{(Replace } \hat{R}_{h,s,a} \text{ by } R_{h,s,a} \text{ through applying event } \mathcal{E}_k^1 \text{ defined in (C.4))} \\
&\overset{(a)}{\leq} R_{h,s,a} + \left\langle \hat{P}_{h,s,a}^k, V_{h+1}^* \right\rangle - \max\left\{ 4\sqrt{\frac{\mathbb{V}\left(\tilde{P}_{h,s,a}^k, V_{h+1}^*\right) L}{n_k(h,s,a)+1}}, \frac{64HL}{n_k(h,s,a)+1} \right\} \\
&\leq R_{h,s,a} + \left\langle \hat{P}_{h,s,a}^k, V_{h+1}^* \right\rangle - \left(3\sqrt{\frac{\mathbb{V}\left(\tilde{P}_{h,s,a}^k, V_{h+1}^*\right) L}{n_k(h,s,a)+1}} + \frac{8HL}{n_k(h,s,a)+1} \right) \\
&\qquad \text{(By applying statement (iii) of Lemma C.3.2)} \\
&\leq R_{h,s,a} + \left\langle \hat{P}_{h,s,a}^k, V_{h+1}^* \right\rangle - \sqrt{\frac{6\mathbb{V}\left(\tilde{P}_{h,s,a}^k, V_{h+1}^*\right)}{n_k(h,s,a)+1}} - \frac{8HL}{n_k(h,s,a)+1} \\
&\leq R_{h,s,a} + \left\langle P_{h,s,a}, V_{h+1}^* \right\rangle \qquad \text{(By applying event } \mathcal{E}_k^2 \text{ defined in (C.5))} \\
&= Q_h^*(s,a).
\end{aligned}$$

Here, the above inequality (a) holds by applying inductive hypothesis and statement (i) in Lemma C.3.2. It is applicable because when $\mathcal{E}_k^w$ holds, by the clipping function, $\left\|\overline{V}_{h+1,k}\right\|_\infty \leq 2H$. When $n_k(h,s,a) < 3$, $\overline{Q}_{h,k}(s,a) \leq Q_h^*(s,a)$ holds trivially because $0 \leq Q_h^*(s,a) \leq H$ by definition. Therefore, the induction is complete.

Now, for arbitrary (k,h,s), set $a = \arg\max_{a\in\mathcal{A}} \overline{Q}_{h,k}(s,a)$ and we have

$$\begin{aligned}
\overline{V}_{h,k}(s) =& \text{clip}_{2(H-h+1)}\left(\overline{Q}_{h,k}(s,a)\right) \\
\leq& \max\left\{-2(H-h+1), \overline{Q}_{h,k}(s,a)\right\} \\
\leq& \max\left\{-(H-h+1), \overline{Q}_{h,k}(s,a)\right\} \\
\leq& \max\left\{-(H-h+1), Q_{h,k}^*(s,a)\right\} \\
\leq& \max\left\{-(H-h+1), Q_{h,k}^*(s,\pi_h^*(s))\right\} \\
\leq& V_{h,k}^*(s).
\end{aligned}$$

□

C.5 Regret Decomposition

In this section, we prove the multiple lemmas necessary for bounding the regret. The regret is mainly composed of two terms, the pessimism term and the estimation error term. The pessimism term, $V_{1,k}^*(s_1^k) - \overline{V}_{1,k}(s_1^k)$, measures how much regret is due to the value the algorithm uses, $\overline{V}_{1,k}$, is smaller than the true value, $V_{1,k}^*(s_1^k)$. The estimation error term, $\overline{V}_{1,k}(s_1^k) - V_{1,k}^{\pi^k}(s_1^k)$ measure how much regret is due to the value, $\overline{V}_{1,k}$, does not estimate $V_{1,k}^{\pi^k}(s_1^k)$, the true value of the policy π^k accurately.

We first introduce a few definitions key to this section. In this section, we omit k if it is clear from the context. Let $a_h^k = \pi_h^k(s_h^k)$ unless specified otherwise.

Definition C.5.1. Let $\mathcal{P}_{h,s,a}^k = \left\langle \hat{P}_{h,s,a}^k - P_{h,s,a}, V_{h+1}^* \right\rangle$ and $\mathcal{R}_{h,s,a}^k = \hat{R}_{h,s,a}^k - R_{h,s,a}$.

Definition C.5.2 ($\underline{M}_{\text{ty}}^k$ and $\underline{V}_{h,k}$)**.** Given history $\mathcal{H}_H^{k-1}$ (defined in equation (C.9)), $\hat{P}^k$ and $\hat{R}^k$, we define $\underline{w}_{\text{ty}}^k(h,s,a) = -\gamma_{\text{ty}}^k(h,s,a)$ and $\underline{V}_{h,k}$ be the value function obtained by running policy π^k on the MDP $\underline{M}_{\text{ty}}^k = (H, \mathcal{S}, \mathcal{A}, \hat{P}^k, \hat{R}^k + \underline{w}_{\text{ty}}^k, s_1^k)$ plus a magnitude clipping with threshold $2(H-h+1)$.

Definition C.5.3 ($\overline{\overline{M}}^k_{\text{ty}}$ and $\overline{\overline{V}}_{h,k}$)**.** Given history $\mathcal{H}_H^{k-1}$ (defined in equation (C.9)), $\hat{P}^k$ and $\hat{R}^k$, we define $\overline{w}^k_{\text{ty}}(h,s,a) = \gamma^k_{\text{ty}}(h,s,a)$ and $\overline{\overline{V}}_{h,k}$ be the value function obtained by running policy π^k on the MDP $\overline{\overline{M}}^k_{\text{ty}} = (H, \mathcal{S}, \mathcal{A}, \hat{P}^k, \hat{R}^k + \overline{w}^k_{\text{ty}}, s_1^k)$ plus a magnitude clipping with threshold $2(H-h+1)$.

Similar to Lemma C.2.12, we can also show that under good event $\mathcal{G}_k$ and $\mathcal{E}^{th}_{h,k}$, no clipping happens on s_h^k for $\underline{V}_{h,k}(s_h^k)$ and $\overline{\overline{V}}_{h,k}(s_h^k)$.

Lemma C.5.4. *Under the good event $\mathcal{G}_k$, we have $\underline{V}_{h,k}(s) \le \overline{V}_{h,k}(s) \le \overline{\overline{V}}_{h,k}(s)$ for all $h \in [H]$, $s \in \mathcal{S}$.*

Proof. This is an immediate result by noticing that under good event $\mathcal{G}_k$, we have $\underline{w}^k_{\text{ty}}(h,s,a) \le w^k_{\text{ty}}(h,s,a) \le \overline{w}^k_{\text{ty}}(h,s,a)$ for all $h \in [H]$ and $s \in \mathcal{S}$. □

Definition C.5.5. Define $\underline{\delta}^\pi_h(s_h)$, $\overline{\delta}^\pi_h(s_h)$, $\overline{\overline{\delta}}^\pi_h(s_h)$, $\delta^\pi_h(s_h)$,$\underline{\delta}_h(s_h)$, $\overline{\delta}_h(s_h)$ and $\overline{\overline{\delta}}_h(s_h)$ as

$$
\begin{aligned}
\underline{\delta}^\pi_h(s_h) &= \underline{V}_h(s_h) - V^\pi_h(s_h),\\
\overline{\delta}^\pi_h(s_h) &= \overline{V}_h(s_h) - V^\pi_h(s_h),\\
\overline{\overline{\delta}}^\pi_h(s_h) &= \overline{\overline{V}}(s_h) - V^\pi_h(s_h),\\
\delta^\pi_h(s_h) &= V^*_h(s_h) - V^\pi_h(s_h),\\
\underline{\delta}_h(s_h) &= \underline{V}_h(s_h) - V^*_h(s_h),\\
\overline{\delta}_h(s_h) &= \overline{V}_h(s_h) - V^*_h(s_h),\\
\overline{\overline{\delta}}_h(s_h) &= \overline{\overline{V}}_h(s_h) - V^*_h(s_h).
\end{aligned}
$$

Definition C.5.6. We denote the history trajectory $\overline{\mathcal{H}}^k_h = \mathcal{H}^k_h \cup \{\hat{z}_k\}$. With filtration sets $\left\{\overline{\mathcal{H}}^k_h\right\}_{h,k}$, we define the following sequences:

$$
\mathcal{M}_{\delta_h(s_h)} = \mathbb{1}\{\mathcal{G}_k \cap \mathcal{E}^{cum}_{h,k}\}\left[\left\langle P_{h,s_h,a_h}, \delta_{h+1}\right\rangle - \delta_{h+1}(s_{h+1})\right],
$$

where $\delta \in \{\underline{\delta}^\pi, \overline{\delta}^\pi, \overline{\overline{\delta}}^\pi, \delta^\pi, \underline{\delta}, \overline{\delta}, \overline{\overline{\delta}}\}$. We will show the sequences are martingales in Lemma C.6.1.

Finally, the regret can be decomposed as

$$
\begin{aligned}
&\operatorname{Regret}\left(M, K, \mathsf{SSR}_{\mathrm{ty}}\right) \\
=&\sum_{k=1}^{K}\left(V_1^*(s_1^k)-V_{1,k}^{\pi^k}(s_1^k)\right) \\
=&\sum_{k=1}^{K}\mathbb{1}\left(\mathcal{C}_{\mathrm{ty}}^k\right)\left(V_1^*(s_1^k)-V_{1,k}^{\pi^k}(s_1^k)\right)+\sum_{k=1}^{K}\mathbb{1}\left(\left(\mathcal{C}_{\mathrm{ty}}^k\right)^c\right)\left(V_1^*(s_1^k)-V_{1,k}^{\pi^k}(s_1^k)\right) \\
=&\sum_{k=1}^{K}\mathbb{1}\left\{\mathcal{C}_{\mathrm{ty}}^k\right\}\Bigg(\underbrace{V_{1,k}^*(s_1^k)-\overline{V}_{1,k}(s_1^k)}_{\text{pessimism term}=-\overline{\delta}_{1,k}(s_1^k)}+\underbrace{\overline{V}_{1,k}(s_1^k)-V_{1,k}^{\pi^k}(s_1^k)}_{\text{estimation error term}=\overline{\delta}_{1,k}^{\pi^k}(s_1^k)}\Bigg) \\
&\quad\underbrace{+\sum_{k=1}^{K}\mathbb{1}\left(\left(\mathcal{C}_{\mathrm{ty}}^k\right)^c\right)\left(V_1^*(s_1^k)-V_{1,k}^{\pi^k}(s_1^k)\right)}_{\text{(a)}}.
\end{aligned}
$$

By Lemma C.2.6 and C.2.7, we know that

$$
\mathbb{E}\left[\sum_{k=1}^{K}\mathbb{1}\left(\left(\mathcal{C}_{\mathrm{ty}}^k\right)^c\right)\right]=\sum_{k=1}^{K}\mathbb{P}\left(\left(\mathcal{C}_{\mathrm{ty}}^k\right)^c\right)\leq\sum_{k=1}^{\infty}\mathbb{P}\left(\left(\mathcal{C}_{\mathrm{ty}}^k\right)^c\right)\leq\frac{\pi^2}{3}.
$$

Therefore, by standard Hoeffding's inequality, it holds with probability at least $1-\delta$ that

$$
\sum_{k=1}^{K}\mathbb{1}\left(\left(\mathcal{C}_{\mathrm{ty}}^k\right)^c\right)\leq\frac{\pi^2}{3}+\sqrt{\frac{\log(1/\delta)}{2K}}.
$$

Since the value functions of true MDP is bounded in $[0,H]$, with probability at least $1-\delta$, we have

$$
\text{(a)}\leq H\sum_{k=1}^{K}\mathbb{1}\left(\left(\mathcal{C}_{\mathrm{ty}}^k\right)^c\right)\leq\frac{\pi^2 H}{3}+H\sqrt{\frac{\log(1/\delta)}{2K}}=\widetilde{O}\left(H\right).
$$

Further, notice that the good event $\mathcal{G}_k=\mathcal{C}_{\mathrm{ty}}^k\cap\mathcal{E}_k^w$ and by Lemma C.2.9, we have $\sum_{k=1}^{\infty}\mathbb{P}\left(\left(\mathcal{E}_k^w\right)^c\right)\leq\frac{\pi^2}{3}$. Therefore, we can similarly address the regret incurred by $\left(\mathcal{E}_k^w\right)^c$ as the bound for term (a). As a result, it will be sufficient to only consider $\mathbb{1}\left\{\mathcal{G}_k\right\}\left(V_{1,k}^*(s_1^k)-V_{1,k}^{\pi^k}(s_1^k)\right)$ when bounding pessimism and estimation error terms. That is, with probability

at least $1-\delta$, it holds that

$$\text{Regret}\left(M, K, \mathsf{SSR}_{\text{ty}}\right) \leq \sum_{k=1}^{K} \mathbb{1}\left\{\mathcal{G}_k\right\}\left(\left|\overline{\delta}_{1,k}(s_1^k)\right| + \left|\overline{\delta}^{\pi^k}_{1,k}(s_1^k)\right|\right) + \widetilde{O}\left(H\right). \tag{C.10}$$

Then, we decompose the estimation error term in Section C.5.2. We decompose the pessimism term in Section C.5.1. We combine the decomposition of the pessimism term and the estimation error term in Section C.5.3.

C.5.1 Pessimism Term

Lemma C.5.7. *Let* $C_1 = \max\left\{\frac{1}{\Phi(1.9)-\Phi(1)}, \frac{1}{\Phi(1.5)-\Phi(1)}\right\} = \frac{1}{\Phi(1.5)-\Phi(1)} \approx 10.9$. *Then, for any* h, k, s_h^k *and the type of noise we used, under the good event* $\mathcal{G}_k$*, the following bound holds,*

$$\mathbb{1}\left\{\mathcal{G}_k\right\}\left|\overline{\delta}_{h,k}(s_h^k)\right| \leq \mathbb{1}\left\{\mathcal{G}_k\right\} C_1\left(\left|\overline{\overline{\delta}}^{\pi^k}_{h,k}(s_h^k)\right| + \left|\underline{\delta}^{\pi^k}_{h,k}(s_h^k)\right|\right). \tag{C.11}$$

Proof. Let $\mathcal{O}_k$ be the event that $\overline{V}_{h,k}(s) \geq V_h^*(s)$ simultaneously for all $s \in \mathcal{S}$ and $h \in [H]$. By Lemma C.3.1 and C.3.3, we know that $\mathbb{P}\left(\mathcal{O}_k \mid \mathcal{H}_H^{k-1}, \mathcal{G}_k\right) \geq \min\left\{\Phi(1.9)-\Phi(1), \Phi(1.5)-\Phi(1)\right\} = \Phi(1.5)-\Phi(1)$, which means $\frac{1}{\mathbb{P}(\mathcal{O}_k)} \leq C_1$ regardless the type of noise used.

The definition of $\mathcal{O}_k$ implies $V_h^* \leq \mathbb{E}\left[\overline{V}_{h,k} \mid \mathcal{O}_k, \mathcal{H}_H^{k-1}, \mathcal{G}_k\right]$. Meanwhile, notice that

$$\begin{aligned}
&\mathbb{1}\left\{\mathcal{G}_k\right\}\left(\mathbb{E}\left[\overline{V}_{h,k} \mid \mathcal{H}_H^{k-1}, \mathcal{G}_k\right] - \underline{V}_{h,k}\right) \\
=&\mathbb{1}\left\{\mathcal{G}_k\right\}\mathbb{P}\left(\mathcal{O}_k \mid \mathcal{H}_H^{k-1}, \mathcal{G}_k\right)\left(\mathbb{E}\left[\overline{V}_{h,k} \mid \mathcal{O}_k, \mathcal{H}_H^{k-1}, \mathcal{G}_k\right] - \underline{V}_{h,k}\right) \\
&\quad + \mathbb{1}\left\{\mathcal{G}_k\right\}\mathbb{P}\left((\mathcal{O}_k)^c \mid \mathcal{H}_H^{k-1}, \mathcal{G}_k\right)\underbrace{\left(\mathbb{E}\left[\overline{V}_{h,k} \mid (\mathcal{O}_k)^c, \mathcal{H}_H^{k-1}, \mathcal{G}_k\right] - \underline{V}_{h,k}\right)}_{\text{(a)} \geq \mathbf{0}} \\
\geq&\mathbb{1}\left\{\mathcal{G}_k\right\}\mathbb{P}\left(\mathcal{O}_k \mid \mathcal{H}_H^{k-1}, \mathcal{G}_k\right)\left(\mathbb{E}\left[\overline{V}_{h,k} \mid \mathcal{O}_k, \mathcal{H}_H^{k-1}, \mathcal{G}_k\right] - \underline{V}_{h,k}\right)
\end{aligned}$$

$$\implies \mathbb{1}\left\{\mathcal{G}_k\right\}\left(\mathbb{E}\left[\overline{V}_{h,k} \mid \mathcal{O}_k, \mathcal{H}_H^{k-1}, \mathcal{G}_k\right] - \underline{V}_{h,k}\right) \leq \mathbb{1}\left\{\mathcal{G}_k\right\} C_1\left(\mathbb{E}\left[\overline{V}_{h,k} \mid \mathcal{H}_H^{k-1}, \mathcal{G}_k\right] - \underline{V}_{h,k}\right).$$

Here, we have term (a) $\geq \mathbf{0}$ since $\underline{V}_{h,k} \leq \overline{V}_{h,k}$ under event $\mathcal{G}_k$, by Lemma C.5.4.

Therefore, we have

$$\begin{aligned}\mathbb{1}\{\mathcal{G}_k\}\left(V_h^*(s_h^k)-\overline{V}_{h,k}(s_h^k)\right) &\le \mathbb{1}\{\mathcal{G}_k\}\left(\mathbb{E}\left[\overline{V}_{h,k}\mid \mathcal{O}_k,\mathcal{H}_H^{k-1},\mathcal{G}_k\right](s_h^k)-\overline{V}_{h,k}(s_h^k)\right)\\ &\le \mathbb{1}\{\mathcal{G}_k\}\left(\mathbb{E}\left[\overline{V}_{h,k}\mid \mathcal{O}_k,\mathcal{H}_H^{k-1},\mathcal{G}_k\right](s_h^k)-\underline{V}_{h,k}(s_h^k)\right)\\ &\le \mathbb{1}\{\mathcal{G}_k\}C_1\left(\mathbb{E}\left[\overline{V}_{h,k}\mid \mathcal{H}_H^{k-1},\mathcal{G}_k\right](s_h^k)-\underline{V}_{h,k}(s_h^k)\right). \quad \text{(C.12)}\end{aligned}$$

We can similarly use constant probability pessimism shown in Lemma C.4.1 and C.4.2. In particular, let $\mathcal{N}_k$ be the event that $\overline{V}_{h,k}(s)\le V_h^*(s)$ for all $s\in\mathcal{S}$ and $h\in[H]$. Then, we have

$$\begin{aligned}&\mathbb{1}\{\mathcal{G}_k\}\left(\mathbb{E}\left[\overline{V}_{h,k}\mid \mathcal{H}_H^{k-1},\mathcal{G}_k\right]-\overline{\overline{V}}_{h,k}\right)\\ =&\mathbb{1}\{\mathcal{G}_k\}\mathbb{P}\left(\mathcal{N}_k\mid \mathcal{H}_H^{k-1},\mathcal{G}_k\right)\left(\mathbb{E}\left[\overline{V}_{h,k}\mid \mathcal{N}_k,\mathcal{H}_H^{k-1},\mathcal{G}_k\right]-\overline{\overline{V}}_{h,k}\right)\\ &+\mathbb{1}\{\mathcal{G}_k\}\mathbb{P}\left((\mathcal{N}_k)^c\mid \mathcal{H}_H^{k-1},\mathcal{G}_k\right)\underbrace{\left(\mathbb{E}\left[\overline{V}_{h,k}\mid (\mathcal{N}_k)^c,\mathcal{H}_H^{k-1},\mathcal{G}_k\right]-\overline{\overline{V}}_{h,k}\right)}_{\text{(b)}\le\mathbf{0}}\\ \le&\mathbb{1}\{\mathcal{G}_k\}\mathbb{P}\left(\mathcal{N}_k\mid \mathcal{H}_H^{k-1},\mathcal{G}_k\right)\left(\mathbb{E}\left[\overline{V}_{h,k}\mid \mathcal{N}_k,\mathcal{H}_H^{k-1},\mathcal{G}_k\right]-\overline{\overline{V}}_{h,k}\right)\end{aligned}$$

$$\implies \mathbb{1}\{\mathcal{G}_k\}\left(\mathbb{E}\left[\overline{V}_{h,k}\mid \mathcal{N}_k,\mathcal{H}_H^{k-1},\mathcal{G}_k\right]-\overline{\overline{V}}_{h,k}\right)\ge \mathbb{1}\{\mathcal{G}_k\}C_1\left(\mathbb{E}\left[\overline{V}_{h,k}\mid \mathcal{H}_H^{k-1},\mathcal{G}_k\right]-\overline{\overline{V}}_{h,k}\right).$$

Thus, we have

$$\begin{aligned}\mathbb{1}\{\mathcal{G}_k\}\left(V_h^*(s_h^k)-\overline{V}_{h,k}(s_h^k)\right) &\ge \mathbb{1}\{\mathcal{G}_k\}\left(\mathbb{E}\left[\overline{V}_{h,k}\mid \mathcal{N}_k,\mathcal{H}_H^{k-1},\mathcal{G}_k\right](s_h^k)-\overline{\overline{V}}_{h,k}(s_h^k)\right)\\ &\ge \mathbb{1}\{\mathcal{G}_k\}C_1\left(\mathbb{E}\left[\overline{V}_{h,k}\mid \mathcal{H}_H^{k-1},\mathcal{G}_k\right](s_h^k)-\overline{\overline{V}}_{h,k}(s_h^k)\right). \quad \text{(C.13)}\end{aligned}$$

Since good event $\mathcal{G}_k$ implies $\underline{V}_{h,k}\le\overline{V}_{h,k}\le\overline{\overline{V}}_{h,k}$ by Lemma C.5.4, the RHS of (C.12) is non-negative and the RHS of (C.13) is non-positive. Therefore, we can then conclude

$$\begin{aligned}&\mathbb{1}\{\mathcal{G}_k\}\left|V_h^*(s_h^k)-\overline{V}_{h,k}(s_h^k)\right|\\ \le&\mathbb{1}\{\mathcal{G}_k\}C_1\left(\left(\mathbb{E}\left[\overline{V}_{h,k}\mid \mathcal{H}_H^{k-1},\mathcal{G}_k\right](s_h^k)-\underline{V}_{h,k}(s_h^k)\right)-\left(\mathbb{E}\left[\overline{V}_{h,k}\mid \mathcal{H}_H^{k-1},\mathcal{G}_k\right](s_h^k)-\overline{\overline{V}}_{h,k}(s_h^k)\right)\right)\\ =&\mathbb{1}\{\mathcal{G}_k\}C_1\left(\overline{\overline{V}}_{h,k}(s_h^k)-\underline{V}_{h,k}(s_h^k)\right)\end{aligned}$$

$$\leq \mathbb{1}\{\mathcal{G}_k\} C_1 \left(\left| \overline{\overline{\delta}}^{\pi^k}_{h,k}(s^k_h) \right| + \left| \underline{\delta}^{\pi^k}_{h,k}(s^k_h) \right| \right).$$

□

C.5.2 Estimation Error Term

We first bound the estimation error of $\overline{\overline{V}}$, which can be regarded as the optimistic estimate used in UCB-type algorithms. For convenience, we will ignore notation $\mathbb{1}\{\mathcal{G}_k\}$ in this section since all statements are proved under the good event $\mathcal{G}_k$.

Lemma C.5.8. *With probability at least $1-\delta$, for all (k,h,s^k_i), under the good event $\mathcal{G}_k$ it holds that*

$$\begin{aligned}
&\mathbb{1}\{\mathcal{E}^{cum}_{h,k}\} \left| \overline{\overline{\delta}}^{\pi^k}_{h,k}(s^k_h) \right| \\
&\leq \mathbb{1}\{\mathcal{E}^{cum}_{h,k}\} \left(\left| \mathcal{P}^k_{h,s^k_h,a^k_h} + \mathcal{R}^k_{h,s^k_h,a^k_h} + \overline{w}^k_{\mathrm{ty}}(h,s^k_h,a^k_h) \right| + \mathcal{M}_{\left|\overline{\overline{\delta}}^{\pi^k}_{h,k}(s^k_h)\right|} + \mathcal{M}_{\left|\overline{\overline{\delta}}_{h,k}(s^k_h)\right|} + \frac{2SH^2L}{n_k(h,s^k_h,a^k_h)} \right) \\
&+ \mathbb{1}\{\mathcal{E}^{cum}_{h+1,k}\} \left(\frac{C_1}{H} \left| \underline{\delta}^{\pi^k}_{h+1,k}(s^k_{h+1}) \right| + \frac{H+1+C_1}{H} \left| \overline{\overline{\delta}}^{\pi^k}_{h+1,k}(s^k_{h+1}) \right| + \frac{1}{H} \left| \overline{\delta}^{\pi^k}_{h+1,k}(s^k_{h+1}) \right| \right) \\
&+ \mathbb{1}\left\{ \mathcal{E}^{cum}_{h,k} \cap \left(\mathcal{E}^{th}_{h+1,k}\right)^c \right\} \left(\frac{C_1}{H} \left| \underline{\delta}^{\pi^k}_{h+1,k}(s^k_{h+1}) \right| + \frac{H+1+C_1}{H} \left| \overline{\overline{\delta}}^{\pi^k}_{h+1,k}(s^k_{h+1}) \right| + \frac{1}{H} \left| \overline{\delta}^{\pi^k}_{h+1,k}(s^k_{h+1}) \right| \right),
\end{aligned}$$

where $L = \log(2HS^2AK/\delta)$.

Proof. Since both $\overline{\overline{V}}$ and V^{π^k} are obtained by choosing actions based on policy π^k under event $\mathcal{G}_k$, we have

$$\begin{aligned}
&\mathbb{1}\left\{\mathcal{E}^{cum}_{h,k}\right\} \left| \overline{\overline{\delta}}^{\pi^k}_{h,k}(s^k_h) \right| \\
=&\mathbb{1}\left\{\mathcal{E}^{cum}_{h,k}\right\} \left| \overline{\overline{V}}_{h,k}(s^k_h) - V^{\pi^k}_{h,k}(s^k_h) \right| \\
=&\mathbb{1}\left\{\mathcal{E}^{cum}_{h,k}\right\} \left| \overline{\overline{Q}}_{h,k}(s^k_h,a^k_h) - Q^{\pi^k}_{h,k}(s^k_h,a^k_h) \right| \qquad \text{(Since no clipping under } \mathcal{E}^{cum}_{h,k} \text{ for } \overline{\overline{V}}_{h,k}(s^k_h)) \\
=&\mathbb{1}\left\{\mathcal{E}^{cum}_{h,k}\right\} \left| \hat{R}^k_{h,s^k_h,a^k_h} - R_{h,s^k_h,a^k_h} + \overline{w}^k_{\mathrm{ty}}(h,s^k_h,a^k_h) + \left\langle \hat{P}^k_{h,s^k_h,a^k_h}, \overline{\overline{V}}_{h+1,k} \right\rangle - \left\langle P^k_{h,s^k_h,a^k_h}, V^{\pi^k}_{h+1,k} \right\rangle \right| \\
=&\mathbb{1}\left\{\mathcal{E}^{cum}_{h,k}\right\} \left| \hat{R}^k_{h,s^k_h,a^k_h} - R_{h,s^k_h,a^k_h} + \overline{w}^k_{\mathrm{ty}}(h,s^k_h,a^k_h) + \left\langle \hat{P}^k_{h,s^k_h,a^k_h}, \overline{\overline{V}}_{h+1,k} \right\rangle - \left\langle P_{h,s^k_h,a^k_h}, V^{\pi^k}_{h+1,k} \right\rangle \right. \\
&\qquad \left. + \left\langle \hat{P}^k_{h,s^k_h,a^k_h} - P_{h,s^k_h,a^k_h}, V^*_{h+1} \right\rangle - \left\langle \hat{P}^k_{h,s^k_h,a^k_h} - P_{h,s^k_h,a^k_h}, V^*_{h+1} \right\rangle \right| \\
\leq&\mathbb{1}\left\{\mathcal{E}^{cum}_{h,k}\right\} \left(\left| \mathcal{P}^k_{h,s^k_h,a^k_h} + \mathcal{R}^k_{h,s^k_h,a^k_h} + \overline{w}^k_{\mathrm{ty}}(h,s^k_h,a^k_h) \right| + \left\langle P_{h,s^k_h,a^k_h}, \left| \overline{\overline{V}}_{h+1,k} - V^{\pi^k}_{h+1,k} \right| \right\rangle \right)
\end{aligned}$$

$$
\begin{aligned}
&+\mathbb{1}\left\{\mathcal{E}_{h,k}^{cum}\right\}\left|\left\langle \hat{P}_{h,s_h^k,a_h^k}^k-P_{h,s_h^k,a_h^k},\overline{\overline{V}}_{h+1,k}-V_{h+1}^*\right\rangle\right| \\
=&\mathbb{1}\left\{\mathcal{E}_{h,k}^{cum}\right\}\left(\left|\mathcal{P}_{h,s_h^k,a_h^k}^k+\mathcal{R}_{h,s_h^k,a_h^k}^k+\overline{w}_{\mathrm{ty}}^k(h,s_h^k,a_h^k)\right|+\left|\overline{\overline{\delta}}_{h+1,k}^{\pi^k}(s_{h+1}^k)\right|+\mathcal{M}_{\left|\overline{\overline{\delta}}_{h,k}^{\pi^k}(s_h^k)\right|}\right) \\
&+\mathbb{1}\left\{\mathcal{E}_{h,k}^{cum}\right\}\left|\left\langle \hat{P}_{h,s_h^k,a_h^k}^k-P_{h,s_h^k,a_h^k},\overline{\overline{V}}_{h+1,k}-V_{h+1}^*\right\rangle\right|.
\end{aligned}
$$

For the last term, we use Lemma C.9.4 and then for $L=\log(2HS^2AK/\delta)$, with probability at least $1-\delta$, we have

$$
\begin{aligned}
&\left|\left\langle \hat{P}_{h,s_h^k,a_h^k}^k-P_{h,s_h^k,a_h^k},\overline{\overline{V}}_{h+1,k}-V_{h+1}^*\right\rangle\right| \\
\le&\sum_{s_{h+1}\in\mathcal{S}}\left|\hat{P}_{h,s_h^k,a_h^k}^k(s_{h+1})-P_{h,s_h^k,a_h^k}(s_{h+1})\right|\left|\overline{\overline{V}}_{h+1,k}(s_h+1)-V_{h+1}^*(s_{h+1})\right| \\
\le&\sum_{s_{h+1}\in\mathcal{S}}\left(2\sqrt{\frac{P_{h,s_h^k,a_h^k}(s_{h+1})L}{n_k(h,s_h^k,a_h^k)}}+\frac{4L}{3n_k(h,s_h^k,a_h^k)}\right)\left|\overline{\overline{\delta}}_{h+1,k}(s_{h+1})\right| \\
=&\sum_{s_{h+1}:P_{h,s_h^k,a_h^k}(s_{h+1})n_k(h,s_h^k,a_h^k)\ge 4LH^2}2P_{h,s_h^k,a_h^k}(s_{h+1})\sqrt{\frac{L}{P_{h,s_h^k,a_h^k}(s_{h+1})n_k(h,s_h^k,a_h^k)}}\left|\overline{\overline{\delta}}_{h+1,k}(s_{h+1})\right| \\
&+\sum_{s_{h+1}:P_{h,s_h^k,a_h^k}(s_{h+1})n_k(h,s_h^k,a_h^k)<4LH^2}2\sqrt{\frac{LP_{h,s_h^k,a_h^k}(s_{h+1})n_k(h,s_h^k,a_h^k)}{n_k(h,s_h^k,a_h^k)^2}}\left|\overline{\overline{\delta}}_{h+1,k}(s_{h+1})\right| \\
&+\frac{4SHL}{3n_k(h,s_h^k,a_h^k)} \\
\le&\sum_{s_{h+1}\in\mathcal{S}}P_{h,s_h^k,a_h^k}(s_{h+1})\frac{1}{H}\left|\overline{\overline{\delta}}_{h+1,k}(s_{h+1})\right|+\frac{4SHL+2SH^2\sqrt{L}}{3n_k(h,s_h^k,a_h^k)} \\
\le&\frac{1}{H}\left|\overline{\overline{\delta}}_{h+1,k}(s_{h+1}^k)\right|+\mathcal{M}_{\left|\overline{\overline{\delta}}_{h,k}(s_h^k)\right|}+\frac{2SH^2L}{n_k(h,s_h^k,a_h^k)} \\
\le&\frac{1}{H}\left|\overline{\overline{\delta}}_{h+1,k}^{\pi^k}(s_{h+1}^k)\right|+\frac{1}{H}\left|\overline{\delta}_{h+1,k}^{\pi^k}(s_{h+1}^k)\right|+\frac{1}{H}\left|\overline{\delta}_{h+1,k}(s_{h+1}^k)\right|+\mathcal{M}_{\left|\overline{\overline{\delta}}_{h,k}(s_h^k)\right|}+\frac{2SH^2L}{n_k(h,s_h^k,a_h^k)} \\
&\qquad\qquad\text{(By triangle inequality)} \\
\le&\frac{1+C_1}{H}\left|\overline{\overline{\delta}}_{h+1,k}^{\pi^k}(s_{h+1}^k)\right|+\frac{1}{H}\left|\overline{\delta}_{h+1,k}^{\pi^k}(s_{h+1}^k)\right|+\frac{C_1}{H}\left|\underline{\delta}_{h+1,k}^{\pi^k}(s_{h+1}^k)\right|+\mathcal{M}_{\left|\overline{\overline{\delta}}_{h,k}(s_h^k)\right|}+\frac{2SH^2L}{n_k(h,s_h^k,a_h^k)}. \\
&\qquad\qquad\text{(By using Lemma C.5.7)}
\end{aligned}
$$

Combining the above two arguments, we can prove the argument:

$$
\mathbb{1}\left\{\mathcal{E}_{h,k}^{cum}\right\}\left|\overline{\overline{\delta}}_{h,k}^{\pi^k}(s_h^k)\right|
$$

$$\leq \mathbb{1}\{\mathcal{E}_{h,k}^{cum}\}\left(\left|\mathcal{P}^k_{h,s_h^k,a_h^k}+\mathcal{R}^k_{h,s_h^k,a_h^k}+\overline{w}^k_{\mathrm{ty}}(h,s_h^k,a_h^k)\right|+\mathcal{M}_{\left|\overline{\underline{\delta}}^{\pi^k}_{h,k}(s_h^k)\right|}+\mathcal{M}_{\left|\overline{\underline{\delta}}_{h,k}(s_h^k)\right|}+\frac{2SH^2L}{n_k(h,s_h^k,a_h^k)}\right)$$
$$+\mathbb{1}\{\mathcal{E}_{h,k}^{cum}\}\left(\frac{H+1+C_1}{H}\left|\overline{\underline{\delta}}^{\pi^k}_{h+1,k}(s_{h+1}^k)\right|+\frac{1}{H}\left|\overline{\delta}^{\pi^k}_{h+1,k}(s_{h+1}^k)\right|+\frac{C_1}{H}\left|\underline{\delta}^{\pi^k}_{h+1,k}(s_{h+1}^k)\right|\right).$$

Then, the proof is complete by noticing that $\mathcal{E}_{h+1,k}^{cum}=\mathcal{E}_{h,k}^{cum}\cap\mathcal{E}_{h+1,k}^{th}$. □

Lemma C.5.9. *With probability at least $1-\delta$, for all (k,h,s_h^k), under good event $\mathcal{G}_k$ it holds that*

$$\mathbb{1}\{\mathcal{E}_{h,k}^{cum}\}\left|\underline{\delta}^{\pi^k}_{h,k}(s_h^k)\right|$$
$$\leq \mathbb{1}\{\mathcal{E}_{h,k}^{cum}\}\left(\left|\mathcal{P}^k_{h,s_h^k,a_h^k}+\mathcal{R}^k_{h,s_h^k,a_h^k}+\underline{w}^k_{\mathrm{ty}}(h,s_h^k,a_h^k)\right|+\mathcal{M}_{\left|\underline{\delta}^{\pi^k}_{h,k}(s_h^k)\right|}+\mathcal{M}_{\left|\underline{\delta}_{h,k}(s_h^k)\right|}+\frac{2SH^2L}{n_k(h,s_h^k,a_h^k)}\right)$$
$$+\mathbb{1}\{\mathcal{E}_{h+1,k}^{cum}\}\left(\frac{C_1}{H}\left|\overline{\underline{\delta}}^{\pi^k}_{h+1,k}(s_{h+1}^k)\right|+\frac{H+1+C_1}{H}\left|\underline{\delta}^{\pi^k}_{h+1,k}(s_{h+1}^k)\right|+\frac{1}{H}\left|\overline{\delta}^{\pi^k}_{h+1,k}(s_{h+1}^k)\right|\right)$$
$$+\mathbb{1}\left\{\mathcal{E}_{h,k}^{cum}\cap\left(\mathcal{E}_{h+1,k}^{th}\right)^c\right\}\left(\frac{C_1}{H}\left|\overline{\underline{\delta}}^{\pi^k}_{h+1,k}(s_{h+1}^k)\right|+\frac{H+1+C_1}{H}\left|\underline{\delta}^{\pi^k}_{h+1,k}(s_{h+1}^k)\right|+\frac{1}{H}\left|\overline{\delta}^{\pi^k}_{h+1,k}(s_{h+1}^k)\right|\right).$$

Proof. The proof exactly follows the proof of Lemma C.5.8. □

Lemma C.5.10. *With probability at least $1-\delta$, for all (k,h,s_h^k), under good event $\mathcal{G}_k$ it holds that*

$$\mathbb{1}\{\mathcal{E}_{h,k}^{cum}\}\left|\overline{\delta}^{\pi^k}_{h,k}(s_h^k)\right|$$
$$\leq \mathbb{1}\{\mathcal{E}_{h,k}^{cum}\}\left(\left|\mathcal{P}^k_{h,s_h^k,a_h^k}+\mathcal{R}^k_{h,s_h^k,a_h^k}+w^k_{\mathrm{ty}}(h,s_h^k,a_h^k)\right|+\mathcal{M}_{\left|\overline{\delta}^{\pi^k}_{h,k}(s_h^k)\right|}+\mathcal{M}_{\left|\overline{\delta}_{h,k}(s_h^k)\right|}+\frac{2SH^2L}{n_k(h,s_h^k,a_h^k)}\right)$$
$$+\mathbb{1}\{\mathcal{E}_{h+1,k}^{cum}\}\left(\frac{C_1}{H}\left|\overline{\underline{\delta}}^{\pi^k}_{h+1,k}(s_{h+1}^k)\right|+\frac{C_1}{H}\left|\underline{\delta}^{\pi^k}_{h+1,k}(s_{h+1}^k)\right|+\frac{H+1}{H}\left|\overline{\delta}^{\pi^k}_{h+1,k}(s_{h+1}^k)\right|\right)$$
$$+\mathbb{1}\left\{\mathcal{E}_{h,k}^{cum}\cap\left(\mathcal{E}_{h+1,k}^{th}\right)^c\right\}\left(\frac{C_1}{H}\left|\overline{\underline{\delta}}^{\pi^k}_{h+1,k}(s_{h+1}^k)\right|+\frac{C_1}{H}\left|\underline{\delta}^{\pi^k}_{h+1,k}(s_{h+1}^k)\right|+\frac{H+1}{H}\left|\overline{\delta}^{\pi^k}_{h+1,k}(s_{h+1}^k)\right|\right).$$

Proof. The proof exactly follows the proof of Lemma C.5.8. □

Lemma C.5.11. *With probability at least $1-\delta$, for all (k,i,s_i^k), under good event $\mathcal{G}_k$ it holds that*

$$\mathbb{1}\left\{\mathcal{E}_{i,k}^{cum}\right\}\left(\left|\overline{\delta}^{\pi^k}_{i,k}(s_i^k)\right|+\left|\overline{\underline{\delta}}^{\pi^k}_{i,k}(s_i^k)\right|+\left|\underline{\delta}^{\pi^k}_{i,k}(s_i^k)\right|\right)$$

$$
\begin{aligned}
\leq & 3e^{3C_1}\left(\sum_{h=i}^{H}\sqrt{e^k_{\mathrm{ty}}(h,s^k_h,a^k_h)}+\sum_{h=i}^{H}\gamma^k_{\mathrm{ty}}(h,s^k_h,a^k_h)+\sum_{h=i}^{H}\frac{SH^2L}{n_k(h,s^k_h,a^k_h)}\right)\\
&+e^{3C_1}\sum_{h=i+1}^{H}\mathbb{1}\left\{\left(\mathcal{E}^{th}_{i+1,k}\right)^c\right\}\left(\left|\overline{\delta}^{\pi^k}_{i+1,k}(s^k_{i+1})\right|+\left|\overline{\overline{\delta}}^{\pi^k}_{i+1,k}(s^k_{i+1})\right|+\left|\underline{\delta}^{\pi^k}_{i+1,k}(s^k_{i+1})\right|\right)\\
&+\sum_{h=i}^{H}(1+\frac{3C_1}{H})^{h-1}\mathbb{1}\left\{\mathcal{E}^{cum}_{h,k}\right\}\mathcal{M}_{h,k},
\end{aligned}
$$

$$
\begin{aligned}
\textit{where } \mathcal{M}_{h,k} =& \mathcal{M}_{\left|\overline{\overline{\delta}}^{\pi^k}_{h,k}(s^k_h)\right|}+\mathcal{M}_{\left|\overline{\overline{\delta}}_{h,k}(s^k_h)\right|}+\mathcal{M}_{\left|\underline{\delta}^{\pi^k}_{h,k}(s^k_h)\right|}\\
&+\mathcal{M}_{\left|\underline{\delta}_{h,k}(s^k_h)\right|}+\mathcal{M}_{\left|\overline{\delta}^{\pi^k}_{h,k}(s^k_h)\right|}+\mathcal{M}_{\left|\overline{\delta}_{h,k}(s^k_h)\right|}.
\end{aligned}
$$

Proof. By summing results in Lemma C.5.8, Lemma C.5.9 and Lemma C.5.10, we have

$$
\begin{aligned}
&\mathbb{1}\left\{\mathcal{E}^{cum}_{h,k}\right\}\left(\left|\overline{\delta}^{\pi^k}_{h,k}(s^k_h)\right|+\left|\overline{\overline{\delta}}^{\pi^k}_{h,k}(s^k_h)\right|+\left|\underline{\delta}^{\pi^k}_{h,k}(s^k_h)\right|\right)\\
\leq&\mathbb{1}\left\{\mathcal{E}^{cum}_{h+1,k}\right\}\left(1+\frac{3C_1}{H}\right)\left(\left|\overline{\delta}^{\pi^k}_{h+1,k}(s^k_{h+1})\right|+\left|\overline{\overline{\delta}}^{\pi^k}_{h+1,k}(s^k_{h+1})\right|+\left|\underline{\delta}^{\pi^k}_{h+1,k}(s^k_{h+1})\right|\right)\\
&+\left|w^k_{\mathrm{ty}}(h,s^k_h,a^k_h)\right|+\left|\overline{w}^k_{\mathrm{ty}}(h,s^k_h,a^k_h)\right|+\left|\underline{w}^k_{\mathrm{ty}}(h,s^k_h,a^k_h)\right|+\frac{6SH^2L}{n_k(h,s^k_h,a^k_h)}+\mathbb{1}\left\{\mathcal{E}^{cum}_{h,k}\right\}\mathcal{M}_{h,k}\\
&+\mathbb{1}\left\{\mathcal{E}^{cum}_{h,k}\cap\left(\mathcal{E}^{th}_{h+1,k}\right)^c\right\}\left(1+\frac{3C_1}{H}\right)\left(\left|\overline{\delta}^{\pi^k}_{h+1,k}(s^k_{h+1})\right|+\left|\overline{\overline{\delta}}^{\pi^k}_{h+1,k}(s^k_{h+1})\right|+\left|\underline{\delta}^{\pi^k}_{h+1,k}(s^k_{h+1})\right|\right)\\
&+3\left|\mathcal{P}^k_{h,s^k_h,a^k_h}+\mathcal{R}^k_{h,s^k_h,a^k_h}\right|\\
\overset{\text{(i)}}{\leq}&\mathbb{1}\left\{\mathcal{E}^{cum}_{h+1,k}\right\}\left(1+\frac{3C_1}{H}\right)\left(\left|\overline{\delta}^{\pi^k}_{h+1,k}(s^k_{h+1})\right|+\left|\overline{\overline{\delta}}^{\pi^k}_{h+1,k}(s^k_{h+1})\right|+\left|\underline{\delta}^{\pi^k}_{h+1,k}(s^k_{h+1})\right|\right)\\
&+3\sqrt{e^k_{\mathrm{ty}}(h,s^k_h,a^k_h)}+3\gamma^k_{\mathrm{ty}}(h,s^k_h,a^k_h)+\frac{6SH^2L}{n_k(h,s^k_h,a^k_h)}+\mathbb{1}\left\{\mathcal{E}^{cum}_{h,k}\right\}\mathcal{M}_{h,k}\\
&+\mathbb{1}\left\{\left(\mathcal{E}^{th}_{h+1,k}\right)^c\right\}\left(1+\frac{3C_1}{H}\right)\left(\left|\overline{\delta}^{\pi^k}_{h+1,k}(s^k_{h+1})\right|+\left|\overline{\overline{\delta}}^{\pi^k}_{h+1,k}(s^k_{h+1})\right|+\left|\underline{\delta}^{\pi^k}_{h+1,k}(s^k_{h+1})\right|\right)
\end{aligned}
$$

Here, the inequality (i) above holds because of two reasons. Firstly, under event $\mathcal{G}_k$, we have

$|w^k_{\text{ty}}(h, s^k_h, a^k_h)| \le |\underline{w}^k_{\text{ty}}(h, s^k_h, a^k_h)|$ and

$$\left|\mathcal{P}^k_{h,s^k_h,a^k_h} + \mathcal{R}^k_{h,s^k_h,a^k_h}\right| = \left|\left\langle \hat{P}^k_{h,s^k_h,a^k_h} - P_{h,s^k_h,a^k_h}, V^*_{h+1}\right\rangle + \left(\hat{R}^k_{h,s^k_h,a^k_h} - R_{h,s^k_h,a^k_h}\right)\right|$$

(By Definition C.5.1)

$$\le \sqrt{e^k_{\text{ty}}(h, s^k_h, a^k_h)}. \qquad \text{(Under event } \mathcal{G}_k,\ \hat{M} \in \mathcal{M}^k_{\text{ty}})$$

Then, the proof is complete by using this recursion from $h = i$ to $h = H$ and utilizing the fact that $(1 + \frac{3C_1}{H})^H \le e^{3C_1}$. □

C.5.3 Combining Estimation and Pessimism Terms

Lemma C.5.12. *With probability at least $1 - \delta$, it holds that*

$$\text{Regret}\left(M, K, \mathsf{SSR}_{\text{ty}}\right) \le \mathbb{1}\left\{\mathcal{G}_k\right\} 3C_1 e^{3C_1} \sum_{k=1}^{K}\sum_{h=i}^{H}\left(\sqrt{e^k_{\text{ty}}(h, s^k_h, a^k_h)} + \gamma^k_{\text{ty}}(h, s^k_h, a^k_h)\right) + \widetilde{O}\left(H^4 S^2 A + H\sqrt{T}\right).$$

Proof. Recall in equation (C.10), with probability at least $1 - \delta$, we have

$$\begin{aligned}
&\text{Regret}\left(M, K, \mathsf{SSR}_{\text{ty}}\right)\\
\le&\sum_{k=1}^{K}\mathbb{1}\left\{\mathcal{G}_k\right\}\left(\left|\overline{\delta}_{1,k}(s^k_1)\right| + \left|\overline{\delta}^{\pi^k}_{1,k}(s^k_1)\right|\right) + \widetilde{O}\left(H\right)\\
\le&\sum_{k=1}^{K}\mathbb{1}\left\{\mathcal{G}_k\right\} C_1\left(\left|\overline{\delta}^{\pi^k}_{1,k}(s^k_1)\right| + \left|\overline{\overline{\delta}}^{\pi^k}_{1,k}(s^k_1)\right| + \left|\underline{\delta}^{\pi^k}_{1,k}(s^k_1)\right|\right) + \widetilde{O}\left(H\right) \qquad \text{(By using Lemma C.5.7)}\\
=&\sum_{k=1}^{K}\mathbb{1}\left\{\mathcal{G}_k \cap \mathcal{E}^{cum}_{1,k}\right\} C_1\left(\left|\overline{\delta}^{\pi^k}_{1,k}(s^k_1)\right| + \left|\overline{\overline{\delta}}^{\pi^k}_{1,k}(s^k_1)\right| + \left|\underline{\delta}^{\pi^k}_{1,k}(s^k_1)\right|\right) + \widetilde{O}\left(H\right)\\
&+\sum_{k=1}^{K}\mathbb{1}\left\{\left(\mathcal{E}^{th}_{1,k}\right)^c\right\}\left(\left|\overline{\delta}^{\pi^k}_{1,k}(s^k_1)\right| + \left|\overline{\overline{\delta}}^{\pi^k}_{1,k}(s^k_1)\right| + \left|\underline{\delta}^{\pi^k}_{1,k}(s^k_1)\right|\right)\\
\le&\mathbb{1}\left\{\mathcal{G}_k\right\} 3C_1 e^{3C_1}\left(\sum_{k=1}^{K}\sum_{h=i}^{H}\sqrt{e^k_{\text{ty}}(h, s^k_h, a^k_h)} + \sum_{k=1}^{K}\sum_{h=i}^{H}\gamma^k_{\text{ty}}(h, s^k_h, a^k_h) + \sum_{k=1}^{K}\sum_{h=i}^{H}\frac{SH^2L}{n_k(h, s^k_h, a^k_h)}\right)\\
&+\sum_{k=1}^{K}\sum_{h=i}^{H}(1 + \frac{3C_1}{H})^{h-1}\mathbb{1}\left\{\mathcal{G}_k \cap \mathcal{E}^{cum}_{h,k}\right\}\mathcal{M}_{h,k} + \widetilde{O}\left(H\right)
\end{aligned}$$

$$+\sum_{k=1}^{K}\sum_{h=1}^{H}\mathbb{1}\left\{\left(\mathcal{E}_{h,k}^{th}\right)^{c}\right\}\left(\left|\overline{\delta}_{h,k}^{\pi^k}(s_1^k)\right|+\left|\overline{\overline{\delta}}_{h,k}^{\pi^k}(s_h^k)\right|+\left|\underline{\delta}_{h,k}^{\pi^k}(s_h^k)\right|\right)$$

(By using Lemma C.5.11)

$$\overset{\text{(i)}}{\leq}\mathbb{1}\left\{\mathcal{G}_k\right\}3C_1e^{3C_1}\sum_{k=1}^{K}\sum_{h=i}^{H}\left(\sqrt{e_{\text{ty}}^k(h,s_h^k,a_h^k)}+\gamma_{\text{ty}}^k(h,s_h^k,a_h^k)\right)+\widetilde{O}\left(H^3S^2A+H\sqrt{T}\right)$$
$$+\widetilde{O}\left(H\right)\sum_{k=1}^{K}\sum_{h=1}^{H}\mathbb{1}\left\{\left(\mathcal{E}_{h,k}^{th}\right)^{c}\right\}$$
$$\leq\mathbb{1}\left\{\mathcal{G}_k\right\}3C_1e^{3C_1}\sum_{k=1}^{K}\sum_{h=i}^{H}\left(\sqrt{e_{\text{ty}}^k(h,s_h^k,a_h^k)}+\gamma_{\text{ty}}^k(h,s_h^k,a_h^k)\right)+\widetilde{O}\left(H^4S^2A+H\sqrt{T}\right).$$

(By using Lemma C.5.13)

The inequality (i) above holds for two reasons. First, it uses Lemma C.6.1 and C.6.3. Second, by our clipping threshold, we know that $\left(\left|\overline{\delta}_{h,k}^{\pi^k}(s_1^k)\right|+\left|\overline{\overline{\delta}}_{h,k}^{\pi^k}(s_h^k)\right|+\left|\underline{\delta}_{h,k}^{\pi^k}(s_h^k)\right|\right)\leq\widehat{O}\left(H\right)$. □

Lemma C.5.13 (Lemma 20 in Agrawal et al. [2021])**.**

$$\sum_{k=1}^{K}\sum_{h=1}^{H}\mathbb{1}\left\{\left(\mathcal{E}_{h,k}^{th}\right)^{c}\right\}\leq\widetilde{O}\left(H^3SA\right).$$

Proof. It holds that

$$\begin{aligned}\sum_{k=1}^{K}\sum_{h=1}^{H}\mathbb{1}\left\{\left(\mathcal{E}_{h,k}^{th}\right)^{c}\right\}&=\sum_{k=1}^{K}\sum_{h=1}^{H}\mathbb{1}\left\{n_k(h,s_h^k,a_h^k)\leq\alpha_k\right\}\\&\leq\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}}\sum_{h=1}^{H}\alpha_k\\&\leq 200H^3SA\log\left(2HSAK^2\right)\log\left(40K^4\right)\qquad\text{(By our choice of }\alpha_k\text{)}\\&=\widetilde{O}\left(H^3SA\right).\end{aligned}$$

□

C.6 Bounds on Individual Terms

C.6.1 Bounds on Martingale Difference

Lemma C.6.1. *For $i \in [H]$, the sequences starting from 0 and with difference between two consecutive terms given by $\mathbb{1}\{\mathcal{G}_k\}\mathcal{M}_{h,k}$ for $h = i, ..., H$, $k = 1, ...K$ are martingales with respect to filtration $\left\{\overline{\mathcal{H}}_h^k\right\}_{\substack{h=i,...,H,\\k=1,...,K}}$. Moreover, for any $\delta' > 0$, with probability at least $1-\delta'$, for any $i \in [H]$, the following hold,*

$$\left|\sum_{k=1}^{K}\sum_{h=i}^{H}\left(1+\frac{3C_1}{H}\right)^h \mathbb{1}\{\mathcal{G}_k \cap \mathcal{E}_{h,k}^{cum}\}\mathcal{M}_{h,k}\right| = \tilde{O}\left(H\sqrt{T}\right).$$

Proof. We first show the sequence starting from 0 and with difference between two consecutive terms given by $\mathbb{1}\{\mathcal{G}_k \cap \mathcal{E}_{h,k}^{cum}\}\left(1+\frac{3C_1}{H}\right)^h \mathcal{M}_{|\overline{\delta}_{h,k}^{\pi^k}(s_h^k)|}$ is a martingale sequence. For any $h \in \{i, ..., H\}$ and $k \in [K]$,

$$\begin{aligned}
&\mathbb{E}\left[\mathbb{1}\{\mathcal{G}_k \cap \mathcal{E}_{h,k}^{cum}\}\mathcal{M}_{|\overline{\delta}_{h,k}^{\pi^k}(s_h^k)|} \mid \overline{\mathcal{H}}_h^k\right] \\
=&\mathbb{E}\left[\mathbb{1}\{\mathcal{G}_k \cap \mathcal{E}_{h,k}^{cum}\}\left(\left\langle P_{h,s_h^k,a_h^k}, \left|\overline{\delta}_{h+1,k}^{\pi^k}(s_{h+1}^k)\right|\right\rangle - \left|\overline{\delta}_{h+1,k}^{\pi^k}(s_{h+1}^k)\right|\right) \mid \overline{\mathcal{H}}_h^k\right] = 0.
\end{aligned}$$

Similarly, we have $\mathbb{1}\{\mathcal{G}_k \cap \mathcal{E}_{h,k}^{cum}\}\mathcal{M}_{|\overline{\overline{\delta}}_{h,k}^{\pi^k}(s_h^k)|}$, $\mathbb{1}\{\mathcal{G}_k \cap \mathcal{E}_{h,k}^{cum}\}\mathcal{M}_{|\underline{\delta}_{h,k}^{\pi^k}(s_h^k)|}$, $\mathbb{1}\{\mathcal{G}_k \cap \mathcal{E}_{h,k}^{cum}\}\mathcal{M}_{|\overline{\delta}_{h,k}(s_h^k)|}$, $\mathbb{1}\{\mathcal{G}_k \cap \mathcal{E}_{h,k}^{cum}\}\mathcal{M}_{|\overline{\overline{\delta}}_{h,k}(s_h^k)|}$, $\mathbb{1}\{\mathcal{G}_k \cap \mathcal{E}_{h,k}^{cum}\}\mathcal{M}_{|\underline{\delta}_{h,k}(s_h^k)|}$ are martingale difference sequences. As $\mathbb{1}\{\mathcal{G}_k \cap \mathcal{E}_{h,k}^{cum}\}\mathcal{M}_{h,k}$ is the sum of several martingale difference sequences, it is a martingale difference sequence.

Next, we bound $\left|\mathbb{1}\{\mathcal{G}_k \cap \mathcal{E}_{h,k}^{cum}\}\mathcal{M}_{|\overline{\delta}_{h,k}^{\pi^k}(s_h^k)|}\right|$. When $h = H$, $\left|\mathbb{1}\{\mathcal{G}_k \cap \mathcal{E}_{h,k}^{cum}\}\mathcal{M}_{|\overline{\delta}_{h,k}^{\pi^k}(s_h^k)|}\right| = 0$. When $\mathcal{G}_k$ holds, for $h < H$ and any state x,

$$\begin{aligned}
\left|\overline{\delta}_{h+1,k}^{\pi^k}(x)\right| = \left|\overline{V}_{h+1}(x) - V_{h+1}^{\pi}(x)\right| &= \left|\left\langle P_{h+2,x,\pi(x)}, \overline{V}_{h+2} - V_{h+2}^{\pi}\right\rangle + w_{\text{ty}}^k(h+1,x,\pi(x))\right| \\
&\leq \left\langle P_{h+2,x,\pi(x)}, \left|\overline{V}_{h+2} - V_{h+2}^{\pi}\right|\right\rangle + \left|w_{\text{ty}}^k(h+1,x,\pi(x))\right|
\end{aligned}$$

By our choice of α_k, when $\mathcal{G}_k$ holds, $\left|w_{\text{ty}}^k(h,s,a)\right| \leq \gamma_{\text{ty}}^k(h,s,a) \leq 1$ for all k, h, s, a as shown in Lemma C.2.12. Then, by expanding $\left|\overline{\delta}_{h+1,k}^{\pi^k}(x)\right| = \left|\overline{V}_{h+1}(x) - V_{h+1}^{\pi}(x)\right|$ recursively from

$h+1$ to H, we have

$$\left|\mathbb{1}\{\mathcal{G}_k \cap \mathcal{E}_{h,k}^{cum}\}\mathcal{M}_{|\overline{\delta}_{h,k}^{\pi^k}(s_h^k)|}\right| \leq 2H\gamma_{\text{ty}}^k(h,s,a) \leq 2H.$$

Similarly, we have the bound on $\mathbb{1}\{\mathcal{G}_k \cap \mathcal{E}_{h,k}^{cum}\}\mathcal{M}_{|\overline{\overline{\delta}}_{h,k}^{\pi^k}(s_h^k)|}$, $\mathbb{1}\{\mathcal{G}_k \cap \mathcal{E}_{h,k}^{cum}\}\mathcal{M}_{|\underline{\delta}_{h,k}^{\pi^k}(s_h^k)|}$, $\mathbb{1}\{\mathcal{G}_k \cap \mathcal{E}_{h,k}^{cum}\}\mathcal{M}_{|\overline{\delta}_{h,k}(s_h^k)|}$, $\mathbb{1}\{\mathcal{G}_k \cap \mathcal{E}_{h,k}^{cum}\}\mathcal{M}_{|\overline{\overline{\delta}}_{h,k}(s_h^k)|}$ and $\mathbb{1}\{\mathcal{G}_k \cap \mathcal{E}_{h,k}^{cum}\}\mathcal{M}_{|\underline{\delta}_{h,k}(s_h^k)|}$.

As a result, $\left|\mathbb{1}\{\mathcal{G}_k \cap \mathcal{E}_{h,k}^{cum}\}\left(1+\frac{3C_1}{H}\right)^h \mathcal{M}_{h,k}\right|$ is bounded by $12e^{3C_1}H$. By Azuma-Hoeffding inequality, with probability at least $1-\delta'$, we have

$$\left|\sum_{k=1}^{K}\sum_{h=i}^{H}\left(1+\frac{3C_1}{H}\right)^h \mathbb{1}\{\mathcal{G}_k \cap \mathcal{E}_{h,k}^{cum}\}\mathcal{M}_{h,k}\right| \leq \widetilde{O}\left(\sqrt{\sum_{k=1}^{K}\sum_{h=i}^{H}H^2}\right) = \widetilde{O}\left(H\sqrt{T}\right).$$

□

C.6.2 Bounds on Lower-order Terms

The following two lemmas are standard results in literature and we present their proofs here for completeness.

Lemma C.6.2.

$$\sum_{k=1}^{K}\sum_{h=1}^{H}\sqrt{\frac{\log\left(2HSAk^2\right)}{n_k\left(h,s_h^k,a_h^k\right)+1}} \leq \widetilde{O}\left(\sqrt{HSAT}\right).$$

Proof. Let $L = \log\left(2HSAK^2\right)$. Then, it can be bounded as

$$\begin{aligned}
\sum_{k=1}^{K}\sum_{h=1}^{H}\sqrt{\frac{\log\left(2HSAk^2\right)}{n_k\left(h,s_h^k,a_h^k\right)+1}} &\leq \sqrt{L}\sum_{k=1}^{K}\sum_{h=1}^{H}\sqrt{\frac{1}{n_k(h,s_h^k,a_h^k)+1}} \\
&= \sqrt{L}\sum_{h,s,a}\sum_{n=1}^{n_K(h,s,a)}\sqrt{\frac{1}{n+1}} \\
&\leq \sqrt{L}\sum_{h,s,a}\int_0^{n_K(h,s,a)}\sqrt{\frac{1}{x}}\mathrm{d}x \\
&\leq 2\sqrt{L}\sum_{h,s,a}\sqrt{n_K\left(h,s,a\right)}
\end{aligned}$$

$$\leq 2\sqrt{L} \cdot \sqrt{HSA \sum_{h,s,a} n_K(h,s,a)}$$

(By Cauchy-Schwartz inequality)

$$= \widetilde{O}\left(\sqrt{HSAT}\right). \qquad \text{(Since } \textstyle\sum_{h,s,a} n_K(h,s,a) = T\text{)}$$

□

Lemma C.6.3.

$$\sum_{k=1}^{K}\sum_{h=1}^{H} \frac{\log\left(2HSAk^2\right)}{n_k\left(h, s_h^k, a_h^k\right) + 1} \leq \widetilde{O}\left(HSA\right).$$

Proof. Let $L = \log\left(2HSAK^2\right)$. Then, it can be bounded as

$$\begin{aligned}
\sum_{k=1}^{K}\sum_{h=1}^{H} \frac{\log\left(2HSAk^2\right)}{n_k\left(h, s_h^k, a_h^k\right) + 1} &\leq L \sum_{k=1}^{K}\sum_{h=1}^{H} \frac{1}{n_k(h, s_h^k, a_h^k) + 1} \\
&= L \sum_{h,s,a} \sum_{n=1}^{n_K(h,s,a)} \frac{1}{n+1} \\
&\leq L \sum_{h,s,a} \log\left(n_K(h,s,a)\right) \qquad \text{(Since } \textstyle\sum_{n=1}^{N} \frac{1}{n} \leq \log(N) + 1\text{)} \\
&\leq LHSA \cdot \max_{h,s,a} \log(n_K(h,s,a)) \\
&\leq LHSA \log(T) \\
&= \widetilde{O}\left(HSA\right).
\end{aligned}$$

□

C.7 Bounds on Sum of Variance

When we use the Bernstein-type noise, the regret analysis needs to bound the sum of variance. This proof applies some techniques developed in Azar et al. [2017]. However, since our optimism only holds with constant probability instead of deterministically, the details are quite different. For simplicity, we first define

$$\hat{\mathbb{V}}^*_{h+1,k} = \mathbb{V}\left(\tilde{P}^k_{h,s_h^k,a_h^k}, V^*_{h+1}\right), \quad \mathbb{V}^*_{h+1,k} = \mathbb{V}\left(P_{h,s_h^k,a_h^k}, V^*_{h+1}\right),$$

$$\hat{\overline{\mathbb{V}}}_{h+1,k} = \mathbb{V}\left(\tilde{P}^k_{h,s^k_h,a^k_h}, \overline{V}_{h+1,k}\right), \quad \overline{\mathbb{V}}_{h+1,k} = \mathbb{V}\left(P_{h,s^k_h,a^k_h}, \overline{V}_{h+1,k}\right),$$
$$\mathbb{V}^{\pi^k}_{h+1,k} = \mathbb{V}\left(P_{h,s^k_h,a^k_h}, V^{\pi^k}_{h+1,k}\right),$$
$$U_{h,k,1} = \sqrt{\frac{\hat{\mathbb{V}}^*_{h+1,k}\log\left(2HSAK^2\right)}{n_k(h,s^k_h,a^k_h)+1}}, \quad U_{h,k,2} = \sqrt{\frac{\hat{\overline{\mathbb{V}}}_{h+1,k}\log\left(2HSAK^2\right)}{n_k(h,s^k_h,a^k_h)+1}},$$

We will first give a full proof of the bound on sum of variance and then present all the auxiliary lemmas in Section C.7.1.

Lemma C.7.1. *Let $U_{h,k} = U_{h,k,1} + U_{h,k,2}$. For any $\delta > 0$, with probability at least $1-\delta$, when $T \geq \Omega\left(H^5S^2A\right)$, it holds that*

$$\sum_{k=1}^{K}\sum_{h=1}^{H-1} \mathbb{1}\left\{\mathcal{G}_k\right\} U_{h,k} \leq \widetilde{O}\left(H\sqrt{SAT}\right).$$

Proof. First, we have

$$\begin{aligned}
\sum_{k=1}^{K}\sum_{h=1}^{H-1} \mathbb{1}\left\{\mathcal{G}_k\right\} U_{h,k} &\leq \sum_{k=1}^{K}\sum_{h=1}^{H-1} \mathbb{1}\left\{\mathcal{G}_k\right\} \sqrt{\frac{\log\left(2HSAK^2\right)}{n_k(h,s^k_h,a^k_h)+1}}\left(\sqrt{\hat{\mathbb{V}}^*_{h+1,k}} + \sqrt{\hat{\overline{\mathbb{V}}}_{h+1,k}}\right) \\
&\leq \sum_{k=1}^{K}\sum_{h=1}^{H-1} \mathbb{1}\left\{\mathcal{G}_k\right\} \sqrt{\frac{\log\left(2HSAK^2\right)}{n_k(h,s^k_h,a^k_h)+1}} \cdot \sqrt{2}\sqrt{\hat{\mathbb{V}}^*_{h+1,k} + \hat{\overline{\mathbb{V}}}_{h+1,k}} \\
&\qquad \text{(Since } \sqrt{a}+\sqrt{b} \leq \sqrt{2(a+b)} \text{ for } a,b \geq 0\text{)} \\
&\leq \sqrt{2}\sqrt{\left(\sum_{k=1}^{K}\sum_{h=1}^{H-1} \frac{\log\left(2HSAK^2\right)}{n_k(h,s^k_h,a^k_h)+1}\right)\left(\sum_{k=1}^{K}\sum_{h=1}^{H-1} \mathbb{1}\left\{\mathcal{G}_k\right\}\left(\hat{\mathbb{V}}^*_{h+1,k} + \hat{\overline{\mathbb{V}}}_{h+1,k}\right)\right)} \\
&\qquad \text{(By Cauchy-Schwartz inequality)} \\
&\leq \sqrt{\widetilde{O}\left(HSA\right)\left(\sum_{k=1}^{K}\sum_{h=1}^{H-1} \mathbb{1}\left\{\mathcal{G}_k\right\}\left(\hat{\mathbb{V}}^*_{h+1,k} + \hat{\overline{\mathbb{V}}}_{h+1,k}\right)\right)}, \qquad \text{(C.14)}
\end{aligned}$$

where the last inequality above applies Lemma C.6.3.

We will then bound the two sums of variance separately. Specifically, by applying Lemma C.7.2 and Lemma C.7.4, we have with probability at least $1-\delta/3$,

$$\sum_{k=1}^{K}\sum_{h=1}^{H-1} \mathbb{1}\left\{\mathcal{G}_k\right\} \hat{\mathbb{V}}^*_{h+1,k} \qquad \text{(C.15)}$$

$$
\begin{aligned}
=&\frac{3}{2}\sum_{k=1}^{K}\sum_{h=1}^{H-1}\mathbb{1}\left\{\mathcal{G}_k\right\}\mathbb{V}_{h+1,k}^{\pi^k}+\sum_{k=1}^{K}\sum_{h=1}^{H-1}\mathbb{1}\left\{\mathcal{G}_k\right\}\left(\hat{\mathbb{V}}_{h+1,k}^{*}-\frac{3}{2}\mathbb{V}_{h+1,k}^{\pi^k}\right)\\
\leq&\widetilde{O}\left(HT+H^2\sqrt{T}+H^3+H^3S^2A+H\sum_{k=1}^{K}\sum_{h=1}^{H-1}\mathbb{1}\left\{\mathcal{G}_k\right\}\delta_{h+1,k}^{\pi^k}(s_{h+1}^k)\right). \qquad \text{(C.16)}
\end{aligned}
$$

By similarly applying Lemma C.7.2 and Lemma C.7.5, we have with probability at least $1-\delta/3$,

$$
\begin{aligned}
&\sum_{k=1}^{K}\sum_{h=1}^{H-1}\mathbb{1}\left\{\mathcal{G}_k\right\}\hat{\overline{\mathbb{V}}}_{h+1,k} \qquad \text{(C.17)}\\
=&\frac{3}{2}\sum_{k=1}^{K}\sum_{h=1}^{H-1}\mathbb{1}\left\{\mathcal{G}_k\right\}\mathbb{V}_{h+1,k}^{\pi^k}+\sum_{k=1}^{K}\sum_{h=1}^{H-1}\mathbb{1}\left\{\mathcal{G}_k\right\}\left(\hat{\overline{\mathbb{V}}}_{h+1,k}-\frac{3}{2}\mathbb{V}_{h+1,k}^{\pi^k}\right)\\
\leq&\widetilde{O}\left(HT+H^2\sqrt{T}+H^3+H^3S^2A+H\sum_{k=1}^{K}\sum_{h=1}^{H-1}\mathbb{1}\left\{\mathcal{G}_k\right\}\left|\overline{\delta}_{h+1,k}^{\pi^k}(s_{h+1}^k)\right|\right), \qquad \text{(C.18)}
\end{aligned}
$$

By combining equations (C.16) and (C.18), we have

$$
\begin{aligned}
&\sum_{k=1}^{K}\sum_{h=1}^{H-1}\mathbb{1}\left\{\mathcal{G}_k\right\}\left(\hat{\mathbb{V}}_{h+1,k}^{*}+\hat{\overline{\mathbb{V}}}_{h+1,k}\right)\\
\leq&\widetilde{O}\left(HT+H^2\sqrt{T}+H^3S^2A+H\sum_{k=1}^{K}\sum_{h=1}^{H-1}\mathbb{1}\left\{\mathcal{G}_k\right\}\left(\delta_{h+1,k}^{\pi^k}(s_{h+1}^k)+\left|\overline{\delta}_{h+1,k}^{\pi^k}(s_{h+1}^k)\right|\right)\right). \qquad \text{(C.19)}
\end{aligned}
$$

Then, by referring to definitions of $\sqrt{e_{\mathrm{Be}}^k(h,s_h^k,a_h^k)}$ and $\gamma_{\mathrm{Be}}^k(h,s_h^k,a_h^k)$, with probability at least $1-\delta/3$, we have

$$
\begin{aligned}
&\sum_{k=1}^{K}\sum_{h=1}^{H-1}\mathbb{1}\left\{\mathcal{G}_k\right\}\left(\delta_{h+1,k}^{\pi^k}(s_{h+1}^k)+\left|\overline{\delta}_{h+1,k}^{\pi^k}(s_{h+1}^k)\right|\right)\\
\leq&\sum_{h=1}^{H-1}\left(\sum_{k=1}^{K}\mathbb{1}\left\{\mathcal{G}_k\right\}\left(\left|\delta_{h+1,k}^{\pi^k}(s_{h+1}^k)\right|+\left|\overline{\delta}_{h+1,k}^{\pi^k}(s_{h+1}^k)\right|\right)\right)\\
\leq&\mathbb{1}\left\{\mathcal{G}_k\right\}3C_1e^{3C_1}H\sum_{k=1}^{K}\sum_{h=1}^{H}\left(\sqrt{e_{\mathrm{ty}}^k(h,s_h^k,a_h^k)}+\gamma_{\mathrm{ty}}^k(h,s_h^k,a_h^k)\right)+\widetilde{O}\left(H^5S^2A+H^2\sqrt{T}\right)
\end{aligned}
$$

(By referring to the proof of Lemma C.5.12)

$$\leq \widetilde{O}\left(H^5S^2A + H^2\sqrt{T} + \sqrt{H^3SAT} + H\sum_{k=1}^{K}\sum_{h=1}^{H-1}\mathbb{1}\left\{\mathcal{G}_k\right\}U_{h,k}\right)$$

(By Lemma C.6.2 and C.6.3)

$$\leq \widetilde{O}\left(H^5S^2A + H^2\sqrt{SAT} + H\sum_{k=1}^{K}\sum_{h=1}^{H-1}\mathbb{1}\left\{\mathcal{G}_k\right\}U_{h,k}\right). \tag{C.20}$$

By plugging equation (C.20) into equation (C.19), we can have

$$\begin{aligned}
&\sum_{k=1}^{K}\sum_{h=1}^{H-1}\mathbb{1}\left\{\mathcal{G}_k\right\}\left(\hat{\mathbb{V}}^*_{h+1,k} + \hat{\bar{\mathbb{V}}}_{h+1,k}\right)\\
\leq&\widetilde{O}\left(HT + H^2\sqrt{T} + H^3S^2A + H^6S^2A + H^3\sqrt{SAT} + H^2\sum_{k=1}^{K}\sum_{h=1}^{H-1}\mathbb{1}\left\{\mathcal{G}_k\right\}U_{h,k}\right)\\
\leq&\widetilde{O}\left(HT + H^3\sqrt{SAT} + H^6S^2A + H^2\sum_{k=1}^{K}\sum_{h=1}^{H-1}\mathbb{1}\left\{\mathcal{G}_k\right\}U_{h,k}\right)\\
\leq&\widetilde{O}\left(HT + H^2\sum_{k=1}^{K}\sum_{h=1}^{H-1}\mathbb{1}\left\{\mathcal{G}_k\right\}U_{h,k}\right) \qquad \text{(When } T \geq \Omega\left(H^5S^2A\right)\text{)}
\end{aligned}$$

Now, by plugging the above result into equation (C.14), when $T \geq \Omega\left(H^5S^2A\right)$, it holds that

$$\begin{aligned}
\sum_{k=1}^{K}\sum_{h=1}^{H-1}\mathbb{1}\left\{\mathcal{G}_k\right\}U_{h,k} &\leq \sqrt{\widetilde{O}\left(HSA\left(HT + H^2\sum_{k=1}^{K}\sum_{h=1}^{H-1}\mathbb{1}\left\{\mathcal{G}_k\right\}U_{h,k}\right)\right)}\\
&\leq \widetilde{O}\left(H\sqrt{SAT} + H^{1.5}\sqrt{\sum_{k=1}^{K}\sum_{h=1}^{H-1}\mathbb{1}\left\{\mathcal{G}_k\right\}U_{h,k}}\right).
\end{aligned}$$

It is easy to check that the above inequality implies $\sum_{k=1}^{K}\sum_{h=1}^{H-1}\mathbb{1}\left\{\mathcal{G}_k\right\}U_{h,k} \leq \widetilde{O}\left(H\sqrt{SAT}\right)$ and thus the proof is complete. □

C.7.1 Auxiliary Lemmas

The lemmas used for proving Lemma C.7.1 are presented as the following.

Lemma C.7.2 (Lemma 8 in Azar et al. [2017]). *For any $\delta > 0$, with probability at least*

$1-\delta$, *it holds that*

$$\sum_{k=1}^{K}\sum_{h=1}^{H-1}\mathbb{1}\left(\mathcal{G}_k\right)\mathbb{V}_{h+1,k}^{\pi^k}\leq\tilde{O}\left(HT+H^2\sqrt{T}+H^3\right).$$

Lemma C.7.3. *For any $\delta>0$, with probability at least $1-\delta$, for any $k\in[K]$, $h\in[H]$, it holds that*

$$\hat{\mathbb{V}}_{h+1,k}^{*}\leq\frac{3}{2}\mathbb{V}_{h+1,k}^{*}+\frac{2H^2S\log\left(2HS^2AK/\delta\right)}{n_k\left(h,s_h^k,a_h^k\right)},$$

$$\hat{\overline{\mathbb{V}}}_{h+1,k}\leq\frac{3}{2}\overline{\mathbb{V}}_{h+1,k}+\frac{2H^2S\log\left(2HS^2AK/\delta\right)}{n_k\left(h,s_h^k,a_h^k\right)}.$$

Proof. The proof apply some techniques in Zhang et al. [2020b]. Fix some $\delta>0$ and let $L=\log\left(2HS^2AK/\delta\right)$ for simplicity. First, by Lemma C.9.1, for some tuple (k,h,s,a,s'), we have

$$\begin{aligned}
&\mathbb{P}\left(\tilde{P}_{h,s,a}^{k}(s')\geq\frac{3}{2}P_{h,s,a}(s')+\frac{2L}{n_k\left(h,s,a\right)}\right)\\
\leq&\mathbb{P}\left(\tilde{P}_{h,s,a}^{k}(s')-P_{h,s,a}(s')\geq\sqrt{\frac{2P_{h,s,a}(s')L}{n_k\left(h,s,a\right)}}+\frac{L}{n_k\left(h,s,a\right)}\right)\\
&\qquad\qquad\text{(Since } a+b\geq2\sqrt{ab} \text{ for } a,b\geq0)\\
\leq&\frac{\delta}{HS^2AK}.
\end{aligned}$$

Then, a union bound says that its complement holds for any (k,h,s,a,s') with probability at least $1-\delta$. Thus, we have

$$\begin{aligned}
\hat{\mathbb{V}}_{h+1,k}^{*}&=\sum_{s'\in\mathcal{S}}\tilde{P}_{h,s_h^k,a_h^k}^{k}(s')\left(V_{h+1}^{*}(s')-\left\langle\tilde{P}_{h,s_h^k,a_h^k}^{k},V_{h+1}^{*}\right\rangle\right)^2\\
&\leq\sum_{s'\in\mathcal{S}}\tilde{P}_{h,s_h^k,a_h^k}^{k}(s')\left(V_{h+1}^{*}(s')-\left\langle P_{h,s_h^k,a_h^k},V_{h+1}^{*}\right\rangle\right)^2\\
&\qquad\qquad\text{(Since } \mathbb{E}\left[X\right] \text{ is the minimizer of } \min_x\mathbb{E}\left[(X-x)^2\right])\\
&\leq\sum_{s'\in\mathcal{S}}\left(\frac{3}{2}P_{h,s,a}(s')+\frac{2L}{n_k(h,s,a)}\right)\left(V_{h+1}^{*}(s')-\left\langle P_{h,s_h^k,a_h^k},V_{h+1}^{*}\right\rangle\right)^2\\
&\leq\frac{3}{2}\mathbb{V}_{h+1,k}^{*}+\frac{2H^2S\log\left(2HS^2AK/\delta\right)}{n_k(h,s_h^k,a_h^k)}.
\end{aligned}$$

For $\hat{\overline{\mathbb{V}}}_{h+1,k}$, we just need to follow a similar argument and thus the proof is complete. □

Lemma C.7.4. *For any $\delta > 0$, with probability at least $1-\delta$, it holds that*

$$\sum_{k=1}^{K}\sum_{h=1}^{H-1}\mathbb{1}\left\{\mathcal{G}_k\right\}\left(\hat{\mathbb{V}}^{*}_{h+1,k}-\frac{3}{2}\mathbb{V}^{\pi^k}_{h+1,k}\right)\leq\widetilde{O}\left(H\sum_{k=1}^{K}\sum_{h=1}^{H-1}\mathbb{1}\left\{\mathcal{G}_k\right\}\delta^{\pi^k}_{h+1,k}(s^k_{h+1})+H^2\sqrt{T}+H^3S^2A\right).$$

Proof. We begin by applying Lemma C.7.3. Thus, with probability at least $1-\frac{\delta}{2}$, we have

$$\begin{aligned}
&\sum_{k=1}^{K}\sum_{h=1}^{H-1}\mathbb{1}\left\{\mathcal{G}_k\right\}\left(\hat{\mathbb{V}}^{*}_{h+1,k}-\frac{3}{2}\mathbb{V}^{\pi^k}_{h+1,k}\right)\\
\leq&\sum_{k=1}^{K}\sum_{h=1}^{H-1}\mathbb{1}\left\{\mathcal{G}_k\right\}\left(\frac{3}{2}\mathbb{V}^{*}_{h+1,k}-\frac{3}{2}\mathbb{V}^{\pi^k}_{h+1,k}\right)+\frac{4H^2S\log\left(4HS^2AK/\delta\right)}{3n_k(h,s^k_h,a^k_h)} && \text{(By Lemma C.7.3)}\\
\leq&\sum_{k=1}^{K}\sum_{h=1}^{H-1}\mathbb{1}\left\{\mathcal{G}_k\right\}\left(\frac{3}{2}\mathbb{V}^{*}_{h+1,k}-\frac{3}{2}\mathbb{V}^{\pi^k}_{h+1,k}\right)+\widetilde{O}\left(H^3S^2A\right) && \text{(By Lemma C.6.3)}\\
\leq&\sum_{k=1}^{K}\sum_{h=1}^{H-1}\mathbb{1}\left\{\mathcal{G}_k\right\}\frac{3}{2}\mathbb{E}_{s'\sim P_{h,s^k_h,a^k_h}}\left[\left(V^{*}_{h+1}(s')\right)^2-\left(V^{\pi^k}_{h+1,k}(s')\right)^2\right]+\widetilde{O}\left(H^3S^2A\right)\\
& && \text{(Since } V^{\pi^k}_{h+1,k}\leq V^{*}_{h+1}\text{)}\\
\leq&3H\sum_{k=1}^{K}\sum_{h=1}^{H-1}\mathbb{1}\left\{\mathcal{G}_k\right\}\mathbb{E}_{s'\sim P_{h,s^k_h,a^k_h}}\left[\delta^{\pi^k}_{h+1,k}(s')\right]+\widetilde{O}\left(H^3S^2A\right)\\
& && \text{(Since } V^{\pi^k}_{h+1,k}\leq V^{*}_{h+1}\leq H \text{ and } a^2-b^2=(a+b)(a-b)\text{)}\\
\leq&\widetilde{O}\left(H\sum_{k=1}^{K}\sum_{h=1}^{H-1}\mathbb{1}\left\{\mathcal{G}_k\right\}\delta^{\pi^k}_{h+1,k}(s^k_{h+1})+H^2\sqrt{T}+H^3S^2A\right). && \text{(By Lemma C.6.1)}
\end{aligned}$$

The last line above holds because by Lemma C.6.1, with probability at least $1-\frac{\delta}{2}$, we have

$$\left|\sum_{k=1}^{K}\sum_{h=1}^{H-1}\mathbb{1}\left\{\mathcal{G}_k\right\}\left(\mathbb{E}_{s'\sim P_{h,s^k_h,a^k_h}}\left[\delta^{\pi^k}_{h+1,k}(s')\right]-\delta^{\pi^k}_{h+1,k}(s^k_{h+1})\right)\right|\leq\widetilde{O}\left(H\sqrt{T}\right).$$

□

Lemma C.7.5. *For any $\delta > 0$, w ith probability at least $1-\delta$, it hold that*

$$\sum_{k=1}^{K}\sum_{h=1}^{H-1}\mathbb{1}\left\{\mathcal{G}_k\right\}\left(\hat{\overline{\mathbb{V}}}_{h+1,k}-\frac{3}{2}\mathbb{V}^{\pi^k}_{h+1,k}\right)\leq\widetilde{O}\left(H\sum_{k=1}^{K}\sum_{h=1}^{H-1}\mathbb{1}\left\{\mathcal{G}_k\right\}\left|\overline{\delta}^{\pi^k}_{h+1,k}(s^k_{h+1})\right|+H^2\sqrt{T}+H^3S^2A\right).$$

Proof. Similarly, we begin by applying Lemma C.7.3 and with probability at least $1-\frac{\delta}{3}$, we have

$$
\begin{aligned}
&\sum_{k=1}^{K}\sum_{h=1}^{H-1}\mathbb{1}\left\{\mathcal{G}_k\right\}\left(\hat{\overline{\mathbb{V}}}_{h+1,k}-\frac{3}{2}\mathbb{V}^{\pi^k}_{h+1,k}\right)\\
\leq&\sum_{k=1}^{K}\sum_{h=1}^{H-1}\mathbb{1}\left\{\mathcal{G}_k\right\}\left(\frac{3}{2}\overline{\mathbb{V}}_{h+1,k}-\frac{3}{2}\mathbb{V}^{\pi^k}_{h+1,k}\right)+\frac{4H^2S\log\left(6HS^2AK/\delta\right)}{3n_k(h,s_h^k,a_h^k)}\\
\leq&\sum_{k=1}^{K}\sum_{h=1}^{H-1}\mathbb{1}\left\{\mathcal{G}_k\right\}\left(\frac{3}{2}\overline{\mathbb{V}}_{h+1,k}-\frac{3}{2}\mathbb{V}^{\pi^k}_{h+1,k}\right)+\widetilde{O}\left(H^3S^2A\right) && \text{(By Lemma C.6.3)}\\
=&\underbrace{\frac{3}{2}\sum_{k=1}^{K}\sum_{h=1}^{H-1}\mathbb{1}\left\{\mathcal{G}_k\right\}\left(\left\langle P_{h,s_h^k,a_h^k},\left(\overline{V}_{h+1,k}\right)^2\right\rangle-\left\langle P_{h,s_h^k,a_h^k},\left(V^{\pi^k}_{h+1,k}\right)^2\right\rangle\right)}_{\text{(a)}}\\
&+\underbrace{\frac{3}{2}\sum_{k=1}^{K}\sum_{h=1}^{H-1}\mathbb{1}\left\{\mathcal{G}_k\right\}\left(\left\langle P_{h,s_h^k,a_h^k},V^{\pi^k}_{h+1,k}\right\rangle^2-\left\langle P_{h,s_h^k,a_h^k},\overline{V}_{h+1,k}\right\rangle^2\right)}_{\text{(b)}}+\widetilde{O}\left(H^3S^2A\right).\\
& && \text{(By definition of variance)}
\end{aligned}
$$

We will bound (a) and (b) separately. For term (a), with probability at least $1-\frac{\delta}{3}$, we have

$$
\begin{aligned}
\text{(a)}=&\sum_{k=1}^{K}\sum_{h=1}^{H-1}\mathbb{1}\left\{\mathcal{G}_k\right\}\left\langle P_{h,s_h^k,a_h^k},\left(\overline{V}_{h+1,k}\right)^2-\left(V^{\pi^k}_{h+1,k}\right)^2\right\rangle\\
\leq&\sum_{k=1}^{K}\sum_{h=1}^{H-1}\mathbb{1}\left\{\mathcal{G}_k\right\}\left\langle P_{h,s_h^k,a_h^k},\left|\overline{V}_{h+1,k}-V^{\pi^k}_{h+1,k}\right|\left|\overline{V}_{h+1,k}+V^{\pi^k}_{h+1,k}\right|\right\rangle\\
& && \text{(Since } a^2-b^2=(a+b)(a-b)\text{)}\\
\leq&\,3H\sum_{k=1}^{K}\sum_{h=1}^{H-1}\mathbb{1}\left\{\mathcal{G}_k\right\}\left\langle P_{h,s_h^k,a_h^k},\left|\overline{V}_{h+1,k}-V^{\pi^k}_{h+1,k}\right|\right\rangle\\
& && \text{(Since } \left\|\overline{V}_{h+1,k}\right\|_\infty\leq 2H \text{ under } \mathcal{G}_k\text{)}\\
\leq&\,3H\sum_{k=1}^{K}\sum_{h=1}^{H-1}\mathbb{1}\left\{\mathcal{G}_k\right\}\left|\overline{\delta}^{\pi^k}_{h+1,k}(s_{h+1}^k)\right|+\widetilde{O}\left(H^2\sqrt{T}\right). && \text{(By Lemma C.6.1)}
\end{aligned}
$$

For term (b), with probability at least $1-\frac{\delta}{3}$, we have

$$\begin{aligned}
\text{(b)} &= \sum_{k=1}^{K}\sum_{h=1}^{H-1} \mathbb{1}\left\{\mathcal{G}_k\right\} \left\langle P_{h,s_h^k,a_h^k}, V_{h+1,k}^{\pi^k} + \overline{V}_{h+1,k} \right\rangle \left\langle P_{h,s_h^k,a_h^k}, V_{h+1,k}^{\pi^k} - \overline{V}_{h+1,k} \right\rangle \\
&\leq 3H \sum_{k=1}^{K}\sum_{h=1}^{H-1} \mathbb{1}\left\{\mathcal{G}_k\right\} \left\langle P_{h,s_h^k,a_h^k}, \left| V_{h+1,k}^{\pi^k} - \overline{V}_{h+1,k} \right| \right\rangle \\
&\leq 3H \sum_{k=1}^{K}\sum_{h=1}^{H-1} \mathbb{1}\left\{\mathcal{G}_k\right\} \left| \overline{\delta}_{h+1,k}^{\pi^k}(s_{h+1}^k) \right| + \widetilde{O}\left(H^2\sqrt{T}\right). \qquad \text{(By Lemma C.6.1)}
\end{aligned}$$

Therefore, in summary, we have with probability at least $1-\delta/2$,

$$\sum_{k=1}^{K}\sum_{h=1}^{H-1} \mathbb{1}\left\{\mathcal{G}_k\right\} \left(\hat{\overline{\mathbb{V}}}_{h+1,k} - \frac{3}{2}\mathbb{V}_{h+1,k}^{\pi^k} \right) \leq \widetilde{O}\left(H \sum_{k=1}^{K}\sum_{h=1}^{H-1} \mathbb{1}\left\{\mathcal{G}_k\right\} \left| \overline{\delta}_{h+1,k}^{\pi^k}(s_{h+1}^k) \right| + H^2\sqrt{T} + H^3S^2A \right).$$

□

C.8 Proof of the Main Theorems

In this section, we state and prove our two main theorems.

Theorem C.8.1. *If the Hoeffding-type noise is used, then for any MDP $M = (H, \mathcal{S}, \mathcal{A}, P, R, s_1)$, for any $\delta > 0$, with probability at least $1-\delta$, Algorithm 7 satisfies*

$$\mathrm{Reg}(M, K, \mathsf{SSR}_{\mathrm{Ho}}) \leq \widetilde{O}\left(H^{1.5}\sqrt{SAT} + H^4S^2A\right).$$

In particular, when $T \geq \widetilde{\Omega}\left(H^5S^3A\right)$, it holds that $\mathrm{Reg}(M, K, \mathsf{SSR}_{\mathrm{Ho}}) \leq \widetilde{O}\left(H^{1.5}\sqrt{SAT}\right)$.

Proof. By using the result of Lemma C.5.12, under Hoeffding-type noise, with probability at least $1-\delta$, we have

$$\begin{aligned}
&\mathrm{Reg}\left(M, K, \mathsf{SSR}_{\mathrm{Ho}}\right) \\
\leq& \mathbb{1}\left\{\mathcal{G}_k\right\} 3C_1 e^{3C_1} \sum_{k=1}^{K}\sum_{h=i}^{H} \left(\sqrt{e_{\mathrm{Ho}}^k(h, s_h^k, a_h^k)} + \gamma_{\mathrm{Ho}}^k(h, s_h^k, a_h^k) \right) + \widetilde{O}\left(H^4S^2A + H\sqrt{T}\right) \\
\leq& 6C_1 e^{3C_1} \sum_{k=1}^{K}\sum_{h=1}^{H-1} \left(H\sqrt{\frac{\log(2HSAk^2)}{n_k(h,s,a)+1}} + \frac{H}{n_k(h,s,a)+1} \right) + \widetilde{O}\left(H^4S^2A + H\sqrt{T}\right)
\end{aligned}$$

$$=\widetilde{O}(H^{1.5}\sqrt{SAT}+H^4S^2A).$$

Here, the second inequality is from the definitions of $\sqrt{e^k_{\mathrm{Ho}}(h,s^k_h,a^k_h)}$ and $\gamma^k_{\mathrm{Ho}}(h,s^k_h,a^k_h)$, and the last step is from Lemma C.6.2 and C.6.3.

□

Theorem C.8.2. *For Bernstein-type noise and $T \geq \widetilde{\Omega}\left(H^5S^2A\right)$, then for any MDP $M = (H,\mathcal{S},\mathcal{A},P,R,s_1)$, for any $\delta>0$, with probability at least $1-\delta$, Algorithm 7 satisfies*

$$\mathrm{Reg}(M,K,\mathsf{SSR}_{\mathrm{Be}}) \leq \widetilde{O}\left(H\sqrt{SAT}+H^4S^2A\right).$$

In particular, if we further have $T \geq \widetilde{\Omega}\left(H^6S^3A\right)$, it then holds that $\mathrm{Reg}(M,K,\mathsf{SSR}_{\mathrm{Be}}) \leq \widetilde{O}\left(H\sqrt{SAT}\right)$.

Proof. Similar to the proof of Theorem C.8.1, under Bernstein-type noise, it holds with probability at least $1-\frac{\delta}{2}$ that

$$\begin{aligned}
&\mathrm{Reg}\left(M,K,\mathsf{SSR}_{\mathrm{Be}}\right)\\
\leq&\mathbb{1}\left\{\mathcal{G}_k\right\}3C_1e^{3C_1}\sum_{k=1}^{K}\sum_{h=i}^{H}\left(\sqrt{e^k_{\mathrm{Be}}(h,s^k_h,a^k_h)}+\gamma^k_{\mathrm{Be}}(h,s^k_h,a^k_h)\right)+\widetilde{O}\left(H^4S^2A+H\sqrt{T}\right)\\
\leq&\widetilde{O}\left(\sum_{k=1}^{K}\sum_{h=1}^{H-1}\left(\mathbb{1}\left\{\mathcal{G}_k\right\}U_{h,k}+\sqrt{\frac{\log(2HSAk^2)}{n_k(h,s,a)+1}}+\frac{H}{n_k(h,s,a)+1}\right)\right)+\widetilde{O}(H^4S^2A+H\sqrt{T})\\
=&\widetilde{O}(H\sqrt{SAT}+H^4S^2A),
\end{aligned}$$

where the last step is from Lemma C.7.1.

□

C.9 Technical Lemmas

Lemma C.9.1 (Bennet's Inequality)**.** *Let* $Z_1, \dots, Z_n$ *be i.i.d. random variables bounded in* $[0,1]$*. Then, for any* $\delta > 0$*, we have*

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n} Z_i - \mathbb{E}\left[Z\right]\right| \geq \sqrt{\frac{2\mathrm{Var}\left(Z\right)\log\left(2/\delta\right)}{n}} + \frac{\log\left(2/\delta\right)}{n}\right) \leq \delta.$$

Lemma C.9.2 (from Maurer and Pontil [2009])**.** *Let* $Z_1, \dots, Z_n$ *with* $n \geq 2$ *be i.i.d. random variables bounded in* $[0, H]$*. Define* $\bar{Z} = \frac{1}{n}\sum_{i=1}^{n} Z_i$ *and* $\hat{V}_n = \frac{1}{n}\sum_{i=1}^{n}\left(Z_i - \bar{Z}\right)^2$*. Then, for any* $\delta > 0$*, we have*

$$\mathbb{P}\left(\left|\mathbb{E}\left[\bar{Z}\right] - \sum_{i=1}^{n} Z_i\right| \geq \sqrt{\frac{2\hat{V}_n\log\left(2/\delta\right)}{n-1}} + \frac{7\log\left(2/\delta\right)}{3\left(n-1\right)}\right) \leq \delta.$$

Lemma C.9.3. *Let* X *be arbitrary random variable bounded in* $[a,b]$ *for some* $a, b \in \mathbb{R}$*. Then, we have* $\mathrm{Var}(X) \leq \frac{(b-a)^2}{4}$*.*

Lemma C.9.4. *For any* $\delta > 0$*, with probability at least* $1 - \delta$*, it holds for all* k, h, s, a, s' *that*

$$\left|\hat{P}^k_{h,s,a}(s') - P_{h,s,a}(s')\right| \leq \sqrt{\frac{4P_{h,s,a}(s')(1-P_{h,s,a}(s'))\log(2HS^2AK/\delta)}{n_k(h,s,a)+1}} + \frac{3\log(2HS^2AK/\delta)}{n_k(h,s,a)+1}.$$

Proof. Let $\delta' = \frac{\delta}{HS^2AK}$ and fix (k,h,s,a,s') such that $n_k(h,s,a) \geq 1$. Then, we have

$$\begin{aligned}
&\mathbb{P}\left(\left|\hat{P}^k_{h,s,a}(s') - P_{h,s,a}(s')\right| \geq \sqrt{\frac{4P_{h,s,a}(s')(1-P_{h,s,a}(s'))\log(2/\delta')}{n_k(h,s,a)+1}} + \frac{3\log(2/\delta')}{n_k(h,s,a)+1}\right)\\
\leq&\mathbb{P}\left(\left|\tilde{P}^k_{h,s,a}(s') - P_{h,s,a}(s')\right| \geq \sqrt{\frac{4P_{h,s,a}(s')(1-P_{h,s,a}(s'))\log(2/\delta')}{n_k(h,s,a)+1}} + \frac{3\log(2/\delta')-1}{n_k(h,s,a)+1}\right)\\
\leq&\mathbb{P}\left(\left|\tilde{P}^k_{h,s,a}(s') - P_{h,s,a}(s')\right| \geq \sqrt{\frac{4P_{h,s,a}(s')(1-P_{h,s,a}(s'))\log(2/\delta')}{n_k(h,s,a)+1}} + \frac{2\log(2/\delta')}{n_k(h,s,a)+1}\right)\\
\leq&\mathbb{P}\left(\left|\tilde{P}^k_{h,s,a}(s') - P_{h,s,a}(s')\right| \geq \sqrt{\frac{2P_{h,s,a}(s')(1-P_{h,s,a}(s'))\log(2/\delta')}{n_k(h,s,a)}} + \frac{\log(2/\delta')}{n_k(h,s,a)}\right)
\end{aligned}$$

(Since $n+1 \leq 2n$ for $n \geq 1$)

$$\leq \delta' = \frac{\delta}{HS^2AK}. \qquad \text{(By Lemma C.9.1, the Bennet's inequality)}$$

Then, the proof is complete by taking a union bound over all possible (k, h, s, a, s'). □

C.10 Numeric Simulations

In this section, we empirically compare RLSVI Russo [2019], UCBVI Azar et al. [2017] and our algorithm SSR on the famous deep sea environment, which is a tabular environment frequently used to test an algorithm's ability to do efficient exploration Osband et al. [2018, 2017], Tan et al. [2020].

Deep sea, as shown in Figure C.1, is a grid-like deterministic environment with $N \times N$ cell states, action space $\{0, 1\}$ and action mask $M_{ij} \sim \mathsf{Bernoulli}\,(0.5)$, $(i, j) \in \mathcal{S}$, whose values are sampled when initializing the environment. At each cell (i, j). Action M_{ij} represents going "right", which leads the agent to the lower right cell, and $1 - M_{ij}$ represents going "left", which leads the agent to the lower left cell. An episode of this environment will end after N steps. When going "left" or going "right" at the off-diagonal, the agent will receive 0 reward; when going "right" along the diagonal before reaching the lower right corner, the agent will receive negative reward $-\frac{0.01}{N}$. Finally, when reaching the lower right corner, depending on the environment initialization, the agent will either receive reward $+1$ or -1. In our experiment, we set this to $+1$, which results in an obvious optimal policy "always going right" with total reward 0.99 per episode.

The experiment results are shown in Figure C.2.* From the plots, we can see that in both settings, SSR performs significantly better than RLSVI as predicted by our theory. Specifically, because of the instability incurred by the independent random seeds and large perturbation magnitude, RLSVI almost never reaches the lower right corner in both settings and thus incurs linear regret. On the other hand, SSR obtains a much lower sub-linear regret because it can explore consistently with the single random seed.

Meanwhile, in both settings, SSR performs comparably with the UCBVI, which is expected since both algorithms achieve the minimax lower bound and our analysis does not

*Bonuses for all three algorithms are scaled down from the theoretical values by a factor of 7×10^4 since without scaling, none of them can learn anything even in the deep sea with $N = 5$.

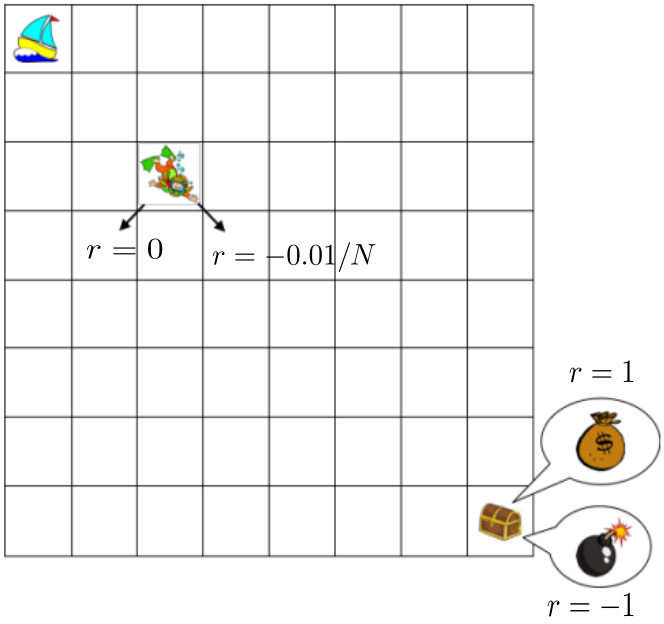


Figure C.1: An example deep sea environment with $N = 8$ Osband et al. [2017].

indicate that one is better than the other.

Finally, we also do an ablation study to show that the better performance of SSR over the RLSVI indeed comes from the single seed randomization instead of smaller noise magnitude. In particular, we run both algorithms in a deep sea environment with $N = 25$ and apply the same noise magnitude, whose results are shown in Figure C.3. We can see that although using the same noise magnitude, SSR still significantly outperforms RLSVI.

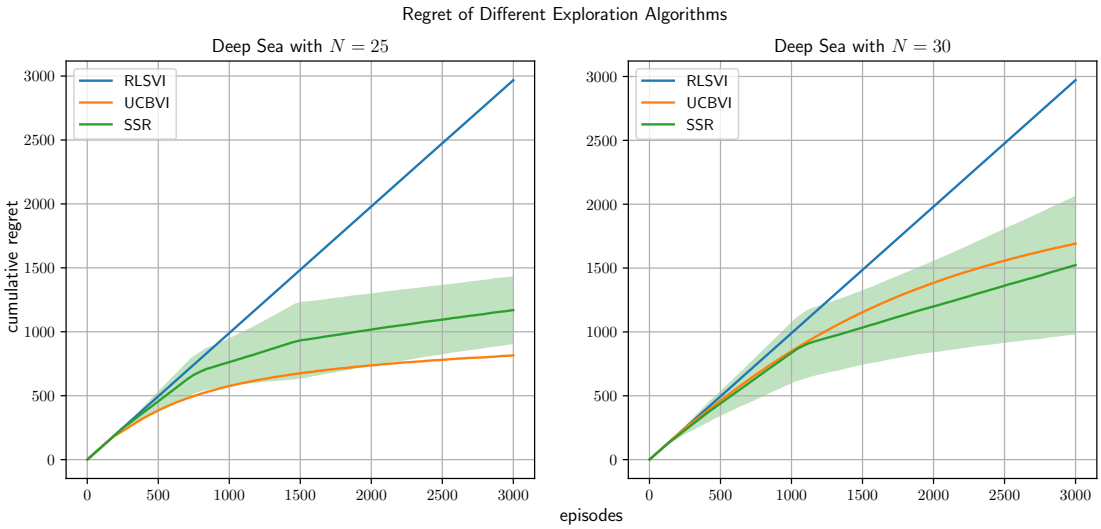


Figure C.2: Empirical evaluation of RLSVI, UCBVI and SSR in deep sea environments with $N = 25$ and $N = 30$. The results are averaged over 10 repeated trials and the shaded area represents the standard deviation. For simplicity, we use Hoeffding-type bonus for both UCBVI and SSR.

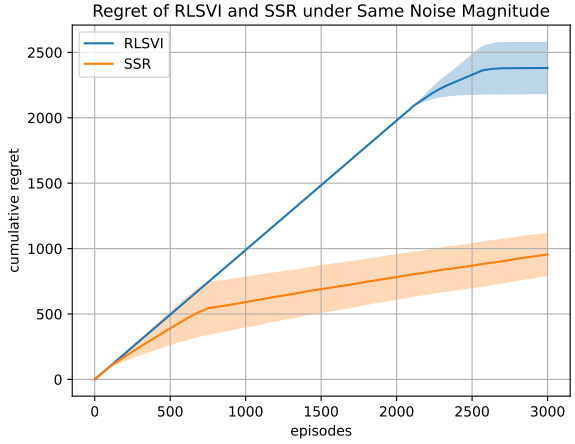


Figure C.3: Empirical evaluation of RLSVI and SSR in deep sea environments with $N = 25$, where both algorithms use the same noise magnitude.

Appendix D

OMITTED PROOFS IN CHAPTER 5

D.1 Selective Sampling Lower Bound

First, we review the standard argument for best-arm identification lower bounds applied to linear bandits. Fix $\theta_* \in \mathbb{R}^d$ and let $z_* = \arg\max_{z \in \mathcal{Z}} \langle z, \theta_* \rangle$. Define the set $\mathcal{C} = \{\theta \in \mathbb{R}^d : \exists z \in \mathcal{Z} \text{ s.t. } \langle \theta, z - z_* \rangle \geq 0\}$ as those θ in which z_* is note the best arm under θ. We now recall the transportation lemma of Kaufmann et al. [2016]. Under a δ-PAC strategy for finding the best arm for the bandit instance $(\mathcal{X}, \mathcal{Z}, \theta_*)$, let T_x denote the random variable which is the number of times arm x is pulled. In addition let $\mathcal{N}_{\theta,x}$ denote the reward distribution of the arm x of $\mathcal{X}$, i.e. $\mathcal{N}_{\theta,x} = \mathcal{N}(x^\top \theta, 1)$. Then for any δ-PAC algorithm

$$\begin{aligned}
\log(1/2.4\delta) &\leq \min_{\theta \in \mathcal{C}} \sum_{x \in \mathcal{X}} \mathbb{E}[T_x] \mathrm{KL}(\mathcal{N}_{\theta_*,x}, \mathcal{N}_{\theta,x}) \\
&= \min_{\theta \in \mathcal{C}} \sum_{x \in \mathcal{X}} \mathbb{E}[T_x] \, \tfrac{1}{2} \|\theta_* - \theta\|^2_{xx^\top} \\
&= \min_{\theta \in \mathcal{C}} \tfrac{1}{2} \|\theta_* - \theta\|^2_{(\sum_{x \in \mathcal{X}} \mathbb{E}[T_x] \, xx^\top)} \\
&\leq \min_{z \in \mathcal{Z} \setminus z_*} \tfrac{1}{2} \|\theta_* - \theta_z(\epsilon)\|^2_{(\sum_{x \in \mathcal{X}} \mathbb{E}[T_x] \, xx^\top)}
\end{aligned}$$

where

$$\theta_z(\varepsilon) = \theta_* - \frac{((z_* - z)^\top \theta_* + \varepsilon)(\sum_{x \in \mathcal{X}} \mathbb{E}[T_x] \, xx^\top)^{-1}(z_* - z)^\top}{(z_* - z)^\top (\sum_{x \in \mathcal{X}} \mathbb{E}[T_x] \, xx^\top)^{-1}(z_* - z)}$$

for some small ϵ. This is a valid choice since for all $z \in \mathcal{Z} \setminus z_*$ we have $(z_* - z)^\top \theta_z(\varepsilon) = -\varepsilon < 0$ and thus $\theta_z(\varepsilon) \in \mathcal{C}$. A straightforward calculation shows that

$$\|\theta_* - \theta_z(\epsilon)\|^2_{(\sum_{x \in \mathcal{X}} \mathbb{E}[T_x] \, xx^\top)} = \frac{(\langle z_* - z, \theta_* \rangle + \varepsilon)^2}{\|z_* - z\|^2_{(\sum_{x \in \mathcal{X}} \mathbb{E}[T_x] \, xx^\top)^{-1}}}$$

so that after rearranging and lettering $\epsilon \to 0$ we have that any δ-PAC algorithm satisfies

$$\max_{z \in \mathcal{Z} \setminus z_*} \frac{2\|z_* - z\|^2_{(\sum_{x \in \mathcal{X}} \mathbb{E}[T_x]\, xx^\top)^{-1}}}{\langle z_* - z, \theta_* \rangle^2} \log(1/2.4\delta) \leq 1. \tag{D.1}$$

This series of steps will be applied for each bullet point of the theorem.

D.1.1 Proof of Theorem 5.2.2, part I

We use the consequence of Lemma 19 of Kaufmann et al. [2016]. Consider a δ-PAC algorithm that sets $P(x) = 1$ for all $x \in \mathcal{X}$ for all time until it exits at time $\mathcal{U}$ after this many unlabelled examples have been observed. If T_x denotes the number of times $x \in \mathcal{X}$ was observed before stopping time $\mathcal{U}$, then by Wald's identity we have that

$$\mathbb{E}[T_x] = \mathbb{E}\left[\sum_{t=1}^{\mathcal{U}} \mathbf{1}\{x_t = x\}\right] = \nu(x)\mathbb{E}[\mathcal{U}].$$

Plugging this into Equation D.1 and rearranging we conclude that

$$\mathbb{E}[\mathcal{U}] \geq \max_{z \in \mathcal{Z} \setminus z_*} \frac{2\|z_* - z\|^2_{(\sum_{x \in \mathcal{X}} \nu(x)\, xx^\top)^{-1}}}{\langle z_* - z, \theta_* \rangle^2} \log(1/2.4\delta) =: \rho(\nu) \log(1/2.4\delta)$$

which concludes the proof of the first bullet.

D.1.2 Proof of Theorem 5.2.2, part II

By definition, the (random) number of times we measure x is

$$\mathcal{L}_x = \sum_{s=1}^{\mathcal{U}} \mathbf{1}\{x_s = x, Q_s(x) = 1\}$$

and we want to show that $\mathbb{E}[\mathcal{L}_x] = \nu(x)\mathbb{E}\left[\sum_{\ell=1}^{\mathcal{U}} P_\ell(x)\right]$. To do so, we define

$$M_t = \sum_{s=1}^{t} \left(\mathbf{1}\{x_s = x, Q_s(x) = 1\} - \nu(x) P_s(x)\right)$$

It is easy to check that $P_{t+1} \in \mathcal{F}_t := \{(x_s, y_s, Q_s)\}_{s=1}^t$ and that

$$\mathbb{E}[M_{t+1}|\mathcal{F}_t] = M_t + \mathbb{E}[\mathbf{1}\{x_s = x, Q_s(x) = 1\} - \nu(x)P_s(x)|\mathcal{F}_t] = M_t$$

Applying Doob's equality $\mathbb{E}[M_{\mathcal{U}}] = \mathbb{E}[M_0] = 0$. Consequence:

$$\mathbb{E}[\mathcal{L}_x] = \mathbb{E}\left[\sum_{s=1}^{\mathcal{U}} \mathbf{1}\{x_s = x, Q_s(x) = 1\}\right] = \nu(v)\mathbb{E}\left[\sum_{s=1}^{\mathcal{U}} P_s(x)\right]$$

Define $\alpha(x) := \frac{\mathbb{E}[\sum_{s=1}^{\mathcal{U}} P_s(x)]}{\mathbb{E}[\mathcal{U}]}$ and note that each $\alpha_x \in [0, 1]$. Then $\mathbb{E}[\mathcal{L}_x] = \mathbb{E}[\mathcal{U}]\alpha(x)\nu(x)$ so applying equation (18) of Kaufmann et al. [2016] again, we have

$$\begin{aligned}
\log(1/2.4\delta) &\leq \min_{\theta \in \mathcal{C}} \sum_{x \in \mathcal{X}} \mathbb{E}[\mathcal{L}_x] \mathrm{KL}(\mathcal{N}_{\theta_*, x}, \mathcal{N}_{\theta, x}) \\
&= \min_{\theta \in \mathcal{C}} \sum_{x \in \mathcal{X}} \mathbb{E}[\mathcal{L}_x] \, \|\theta - \theta_*\|^2_{xx^\top}/2 \\
&= \min_{z \in \mathcal{Z} \setminus z_*} \frac{\langle \theta_*, z_* - z \rangle^2}{2\|z - z_*\|^2_{(\sum_{x \in \mathcal{X}} \mathbb{E}[\mathcal{L}_x] xx^\top)^{-1}}} \\
&= \min_{z \in \mathcal{Z} \setminus z_*} \frac{\langle \theta_*, z_* - z \rangle^2}{2\|z - z_*\|^2_{(\sum_{x \in \mathcal{X}} \nu(x)\alpha(x) xx^\top)^{-1}}} \mathbb{E}[\mathcal{U}].
\end{aligned}$$

Rearranging, and applying the identity $\mathbb{E}_{X \sim \nu}[\alpha(X) X X^\top] = \sum_{x \in \mathcal{X}} \nu(x)\alpha(x) xx^\top$, the above implies that

$$\mathbb{E}[\mathcal{U}] \geq \max_{z \in \mathcal{Z} \setminus z_*} \frac{2\|z - z_*\|^2_{\mathbb{E}_{X \sim \nu}[\alpha(X) X X^\top]^{-1}}}{\langle \theta_*, z_* - z \rangle^2} \log(1/2.4\delta).$$

Noting that the total expected number of labels is equal to

$$\mathbb{E}[\mathcal{L}] = \sum_{x \in \mathcal{X}} \mathbb{E}[\mathcal{L}_x] = \sum_{x \in \mathcal{X}} \mathbb{E}[\mathcal{U}]\alpha(x)\nu(x) = \mathbb{E}[\mathcal{U}] \, \mathbb{E}_{X \sim \nu}[\alpha(X)]$$

we conclude that

$$\mathbb{E}[\mathcal{L}] \geq \min_{\alpha : \mathcal{X} \to [0,1]} \quad \mathbb{E}[\mathcal{U}] \; \mathbb{E}_{X \sim \nu}[\alpha(X)]$$

$$\text{subject to} \quad \mathbb{E}[\mathcal{U}] \geq \max_{z \in \mathcal{Z} \setminus \{z_*\}} \frac{2\|z - z_*\|^2_{\mathbb{E}_{X\sim\nu}[\alpha(X)XX^\top]^{-1}}}{\langle \theta_*, z_* - z\rangle^2} \log(1/2.4\delta).$$

The second bullet point result follows by denoting α as P and applying Proposition D.2.6.

D.2 Selective Sampling Algorithm for Known Distribution ν

D.2.1 Proof of Theorem 5.2.3, upper bound

At each round ℓ we assume an implementation such that $\widehat{P}_\ell, \widehat{\Sigma}_{\widehat{P}_\ell} \leftarrow$OptimizeDesign$(\mathcal{Z}_\ell, 2^{-\ell}, \tau)$ returns the solution of Equation 5.3 with $\epsilon = 2^{-\ell}$, essentially. More explicitly, let $\epsilon_\ell := 2^{-\ell}$, $B < \infty$ such that $\max_{x\in\mathcal{X}} |\langle x, \theta_*\rangle| \leq B$, and $\sigma < \infty$ such that $\mathbb{E}[(y_s - \langle \theta_*, x_s\rangle)^2 | x_s] \leq \sigma^2$. If

$$\beta_{\delta,\ell} := 16(B^2 + \sigma^2) \log(2\ell^2 |\mathcal{Z}|^2/\delta)$$

then $\widehat{P}_\ell = P_\ell$ where

$$P_\ell := \operatorname*{arg\,min}_{P:\mathcal{X}\to[0,1]} \mathbb{E}_{X\sim\nu}[P(X)] \text{ subject to } \max_{z,z'\in\mathcal{Z}_\ell} \frac{\|z - z'\|^2_{\mathbb{E}_{X\sim\nu}[\tau P(X)XX^\top]^{-1}}}{\epsilon_\ell^2} \beta_{\delta,\ell} \leq 1$$

and $\widehat{\Sigma}_{\widehat{P}_\ell} := \mathbb{E}_{X\sim\nu}[P_\ell(X)XX^\top]$

We first provide an intermediate lemma on the correctness of Algorithm 8 that relies on the feasibility of P_ℓ which we will show shortly.

Lemma D.2.1. *With probability at least $1 - \delta$ we have for all stages $\ell \in \mathbb{N}$ such that P_ℓ is feasible, that $z_* \in \mathcal{Z}_\ell$ and $\max_{z\in\mathcal{Z}_\ell} \langle z_* - z, \theta_*\rangle \leq 4\epsilon_\ell$.*

Proof. Define the event $\mathcal{E}$ as

$$\mathcal{E} := \bigcap_{\ell=1}^{\infty} \bigcap_{z,z'\in\mathcal{Z}_\ell} \left\{ |\langle z - z', \widehat{\theta}_\ell - \theta_*\rangle| \leq \epsilon_\ell \right\}$$

By Lemma D.2.2, we know that $\mathbb{P}(\mathcal{E}) \geq 1 - \delta$. Then, the rest of the proof is the same as the one in Fiez et al. [2019], but we include it here for completeness. Assume that $\mathcal{E}$ holds.

Then for any $z' \in \mathcal{Z}_\ell$

$$\begin{aligned}\langle z' - z^*, \widehat{\theta}_\ell \rangle &= \langle z' - z^*, \widehat{\theta}_\ell - \theta^* \rangle + \langle z' - z^*, \theta^* \rangle \\ &= \langle z' - z^*, \widehat{\theta}_\ell - \theta^* \rangle \\ &\leq \epsilon_\ell\end{aligned}$$

so that z^* would survive to round $\mathcal{Z}_{\ell+1}$. And for any $z \in \mathcal{Z}_\ell$ such that $\langle z^* - z, \theta^* \rangle > 2\epsilon_\ell$, we have

$$\begin{aligned}\max_{z' \in \mathcal{Z}_\ell} \langle z' - z, \widehat{\theta}_\ell \rangle &\geq \langle z^* - z, \widehat{\theta}_\ell \rangle \\ &= \langle z^* - z, \widehat{\theta}_\ell - \theta^* \rangle + \langle z^* - z, \theta^* \rangle \\ &> -\epsilon_\ell + 2\epsilon_\ell \\ &= \epsilon_\ell\end{aligned}$$

which implies this z would be kicked out. Note that this implies that $\max_{z \in \mathcal{Z}_{\ell+1}} \langle z^* - z, \theta^* \rangle \leq 2\epsilon_\ell = 4\epsilon_{\ell+1}$. □

We can now prove Theorem 5.2.3. After $L := \lceil \log_2(\frac{4}{\Delta}) \rceil$ rounds $\mathcal{Z}_\ell = \{z_*\}$ by the above lemma. Thus, the total number of labels requested after L rounds is equal to $\mathcal{L} := \sum_{\ell=1}^{L} \sum_{t=(\ell-1)\tau+1}^{\ell\tau} Q_\ell(x_t)$. By Freedman's inequality (c.f., Theorem 1 of Beygelzimer et al. [2011]) we have that

$$\sum_{\ell=1}^{L} \sum_{t=(\ell-1)\tau+1}^{\ell\tau} Q_\ell(x_t) \leq 2 \sum_{\ell=1}^{L} \tau \mathbb{E}_{X \sim \nu}[P_\ell(X) | \mathcal{Z}_\ell] + \log(1/\delta)$$

We can now bound the expected sample complexity of this algorithm.

$$\begin{aligned}&\sum_{\ell=1}^{L} \tau \mathbb{E}_{X \sim \nu}[P_\ell(X) | \mathcal{Z}_\ell] \\ &= \sum_{\ell=1}^{L} \left[\min_{P : \mathcal{X} \to [0,1]} \tau \mathbb{E}_{X \sim \nu}[P(X)] \quad \text{subject to} \quad \max_{z, z' \in \mathcal{Z}_\ell} \frac{\|z - z'\|^2_{\mathbb{E}_{X \sim \nu}[\tau P(X) X X^\top]^{-1}}}{\epsilon_\ell^2} \beta_{\delta,\ell} \leq 1 \right].\end{aligned}$$

Using Lemma D.2.5, we have

$$\begin{aligned}\max_{z,z'\in\mathcal{Z}_\ell}\frac{\|z-z'\|^2_{\mathbb{E}_{X\sim\nu}[\tau P(X)XX^\top]^{-1}}}{\epsilon_\ell^2}\beta_{\delta,\ell} &\le \beta_{\delta,L}\max_{z,z'\in\mathcal{Z}_\ell}\frac{\|z-z'\|^2_{\mathbb{E}_{X\sim\nu}[\tau P(X)XX^\top]^{-1}}}{\epsilon_\ell^2}\\ &\le 64\beta_{\delta,L}\max_{z\in\mathcal{Z}\setminus z_*}\frac{\|z-z_*\|^2_{\mathbb{E}_{X\sim\nu}[\tau P(X)XX^\top]^{-1}}}{\langle z-z_*,\theta_*\rangle^2}\\ &=:\max_{z\in\mathcal{Z}\setminus z_*}\frac{\|z-z_*\|^2_{\mathbb{E}_{X\sim\nu}[\tau P(X)XX^\top]^{-1}}}{\langle z-z_*,\theta_*\rangle^2}\beta_\delta\end{aligned}$$

Note that the last line also describes a condition for which a P_ℓ is feasible. Indeed, at round ℓ, a sufficient condition for a feasible P_ℓ (i.e., the RHS ≤ 1) is if τ exceeds $\rho(\nu)\beta_\delta$ with $\beta_\delta := 1024(B^2+\sigma^2)\log(2L^2|\mathcal{Z}|^2/\delta)$ and $\rho(\nu) = \max_{z\in\mathcal{Z}\setminus z_*}\frac{\|z-z_*\|^2_{\mathbb{E}_{X\sim\nu}[XX^\top]^{-1}}}{\langle z-z_*,\theta_*\rangle^2}$, which holds by assumption in the theorem.

Plugging this constraint back into above we have

$$\begin{aligned}&\sum_{\ell=1}^L \tau\mathbb{E}_{X\sim\nu}[P_\ell(X)|\mathcal{Z}_\ell]\\ &\le \sum_{\ell=1}^L\left[\min_{P:\mathcal{X}\to[0,1]}\tau\mathbb{E}_{X\sim\nu}[P(X)] \quad\text{subject to}\quad \max_{z\in\mathcal{Z}\setminus z_*}\frac{\|z-z_*\|^2_{\mathbb{E}_{X\sim\nu}[\tau P(X)XX^\top]^{-1}}}{\langle z-z_*,\theta_*\rangle^2}\beta_\delta\le 1\right]\\ &\le L\min_{\lambda\in\triangle_\mathcal{X}}\rho(\lambda)\beta_\delta \quad\text{subject to}\quad \|\lambda/\nu\|_\infty\rho(\lambda)\beta_\delta\le\tau\end{aligned}$$

where the last line follows by applying the reparameterization of Proposition D.2.6.

High-probability Events

Lemma D.2.2. *We have* $\mathbb{P}(\mathcal{E})\ge 1-\delta$.

Proof. For any $\mathcal{V}\subseteq\mathcal{Z}$ and $z,z'\in\mathcal{V}$ define

$$\mathcal{E}_{z,z',\ell}(\mathcal{V}) = \{|\langle z-z', \widehat{\theta}_\ell(\mathcal{V})-\theta_*\rangle|\le\epsilon_\ell\}$$

where $\widehat{\theta}_\ell(\mathcal{V})$ is the estimator that would be constructed by the algorithm at stage ℓ with $\mathcal{Z}_\ell=\mathcal{V}$. For fixed $\mathcal{V}\subset\mathcal{Z}$ and $\ell\in\mathbb{N}$ we apply Proposition D.2.4 so that with probability at

least $1 - \frac{\delta}{\ell^2|\mathcal{Z}|^2}$ we have that for any $z, z' \in \mathcal{V}$

$$\begin{aligned}|\langle z - z', \widehat{\theta}_\ell(\mathcal{V}) - \theta_* \rangle| &\leq \|z - z'\|_{\mathbb{E}_{X\sim\nu}[\tau P_\ell(X) X X^\top]^{-1}} \sqrt{16(B^2 + \sigma^2)\log(2\ell^2|\mathcal{Z}|^2/\delta)} \\ &\leq \epsilon_\ell\end{aligned}$$

Noting that $\mathcal{E} := \bigcap_{\ell=1}^{\infty} \bigcap_{z,z' \in \mathcal{Z}_\ell} \mathcal{E}_{z,z',\ell}(\mathcal{Z}_\ell)$ we have

$$\begin{aligned}\mathbb{P}\left(\bigcup_{\ell=1}^{\infty} \bigcup_{z,z'\in\mathcal{Z}_\ell} \{\mathcal{E}^c_{z,z',\ell}(\mathcal{Z}_\ell)\}\right) &\leq \sum_{\ell=1}^{\infty} \mathbb{P}\left(\bigcup_{z,z'\in\mathcal{Z}_\ell} \{\mathcal{E}^c_{z,z',\ell}(\mathcal{Z}_\ell)\}\right) \\ &= \sum_{\ell=1}^{\infty} \sum_{\mathcal{V}\subseteq\mathcal{Z}} \mathbb{P}\left(\bigcup_{z,z'\in\mathcal{V}} \{\mathcal{E}^c_{z,z',\ell}(\mathcal{V})\}, \mathcal{Z}_\ell = \mathcal{V}\right) \\ &= \sum_{\ell=1}^{\infty} \sum_{\mathcal{V}\subseteq\mathcal{Z}} \mathbb{P}\left(\bigcup_{z,z'\in\mathcal{V}} \{\mathcal{E}^c_{z,z',\ell}(\mathcal{V})\}\right) \mathbb{P}(\mathcal{Z}_\ell = \mathcal{V}) \\ &\leq \sum_{\ell=1}^{\infty} \sum_{\mathcal{V}\subseteq\mathcal{Z}} \tfrac{\delta}{\ell^2|\mathcal{Z}|^2} \binom{|\mathcal{V}|}{2} \mathbb{P}(\mathcal{Z}_\ell = \mathcal{V}) \\ &\leq \sum_{\ell=1}^{\infty} \sum_{\mathcal{V}\subseteq\mathcal{Z}} \tfrac{\delta}{2\ell^2} \mathbb{P}(\mathcal{Z}_\ell = \mathcal{V}) \leq \delta\end{aligned}$$

□

D.2.2 Technical Lemmas

The following definition characterizes the RIPS estimator we used in Algorithm 8.

Definition D.2.3. Let $X_1, \ldots, X_n$ be i.i.d. random variables with mean $\bar{x}$ and variance ν^2. Let $\delta \in (0, 1)$. We say that $\widehat{\mu}(X_1, \ldots, X_n)$ is a *δ-robust estimator* if there exist universal constants $c_1, c_0 > 0$ such that if $n \geq c_1 \log(1/\delta)$, then with probability at least $1 - \delta$

$$|\widehat{\mu}(\{X_t\}_{t=1}^n) - \bar{x}| \leq c_0 \sqrt{\frac{\nu^2 \log(1/\delta)}{n}}.$$

Examples of δ-robust estimators include the median-of-means estimator and Catoni's estimator Lugosi and Mendelson [2019]. This work employs the use of the Catoni estimator

which satisfies $|\widehat{\mu}(\{X_t\}_{t=1}^n) - \bar{x}| \leq \sqrt{\frac{2\nu^2 \log(1/\delta)}{n - 2\log(1/\delta)}}$ for $n > 2\log(1/\delta)$ which leads to an optimal leading constant as $n \to \infty$. See Camilleri et al. [2021a] or Lugosi and Mendelson [2019] for more details.

Proposition D.2.4. *Let $x_1, \ldots, x_n$ be drawn IID from a distribution ν. Assume that $|\langle \theta, x_s \rangle| \leq B$ and $\mathbb{E}[|\langle \theta, x_s \rangle - y_s|^2] \leq \sigma^2$. Let $P : \mathcal{X} \to [0,1]$ be arbitrary. Let $Q(x_s) \sim$ Bernoulli$(P(x_s))$ independently for all $s \in [n]$. For a given finite set $\mathcal{V} \subset \mathbb{R}^d$ define for any $v \in \mathcal{V}$*

$$w_v = \text{Catoni}(\{\langle v, \mathbb{E}_{X\sim\nu}[P(X)XX^\top]^{-1} Q(x_s) x_s y_s \rangle\}_{s=1}^n).$$

If $\widehat{\theta} = \arg\min_\theta \max_v \frac{|w_v - \langle \theta, v \rangle|}{\|v\|_{\mathbb{E}_{X\sim\nu}[P(X)XX^\top]^{-1}}}$ and $n \geq 4\log(2|\mathcal{V}|/\delta)$, then with probability at least $1-\delta$, it holds that

$$|\langle v, \widehat{\theta} - \theta \rangle| \leq \|v\|_{\mathbb{E}_{X\sim\nu}[nP(X)XX^\top]^{-1}} \sqrt{16(B^2 + \sigma^2)\log(2|\mathcal{V}|/\delta)}$$

Proof. Inspired by Camilleri et al. [2021a], we note that

$$\begin{aligned}
\max_{v\in\mathcal{V}} \frac{|\langle \widehat{\theta}, v \rangle - \langle \theta, v \rangle|}{\|v\|_{\mathbb{E}_{X\sim\nu}[nP(X)XX^\top]^{-1}}} &= \max_{v\in\mathcal{V}} \frac{|\langle \widehat{\theta}, v \rangle - w_v + w_v - \langle \theta, v \rangle|}{\|v\|_{\mathbb{E}_{X\sim\nu}[nP(X)XX^\top]^{-1}}} \\
&\leq \max_{v\in\mathcal{V}} \frac{|\langle \widehat{\theta}, v \rangle - w_v|}{\|v\|_{\mathbb{E}_{X\sim\nu}[nP(X)XX^\top]^{-1}}} + \max_{v\in\mathcal{V}} \frac{|w_v - \langle \theta, v \rangle|}{\|v\|_{\mathbb{E}_{X\sim\nu}[nP(X)XX^\top]^{-1}}} \\
&= \min_\theta \max_{v\in\mathcal{V}} \frac{|\langle \theta, v \rangle - w_v|}{\|v\|_{\mathbb{E}_{X\sim\nu}[nP(X)XX^\top]^{-1}}} + \max_{v\in\mathcal{V}} \frac{|w_v - \langle \theta, v \rangle|}{\|v\|_{\mathbb{E}_{X\sim\nu}[nP(X)XX^\top]^{-1}}} \\
&\leq 2 \max_{v\in\mathcal{V}} \frac{|\langle \theta, v \rangle - w_v|}{\|v\|_{\mathbb{E}_{X\sim\nu}[nP(X)XX^\top]^{-1}}}
\end{aligned}$$

So it suffices to show that each $|\langle \theta, v \rangle - w_v|$ is small. We begin by fixing some $v \in \mathcal{V}$ and bounding the variance of $v^\top \mathbb{E}_{X\sim\nu}[P(X)XX^\top]^{-1} Q(x_s) x_s y_s$ for any $s \leq n$ which is necessary to use the robust estimator. For readability purposes, we shorten $\mathbb{E}_{x_s\sim\nu, Q(x_s)\sim P(x_s)}$ as $\mathbb{E}_{x_s,Q}$ in the rest of this proof. Note that

$$\mathbb{V}\text{ar}_{x_s\sim\nu, Q(x_s)\sim P(x_s)}(v^\top \mathbb{E}_{X\sim\nu}[P(X)XX^\top]^{-1} Q(x_s) x_s y_s)$$

$$
\begin{aligned}
=&\mathbb{E}_{x_s,Q}[(v^\top \mathbb{E}_{X\sim\nu}[P(X)XX^\top]^{-1}Q(x_s)x_s y_s)^2] \\
&\quad - \mathbb{E}_{x_s,Q}[v^\top \mathbb{E}_{X\sim\nu}[P(X)XX^\top]^{-1}Q(x_s)x_s y_s]^2
\end{aligned}
$$

which means we can drop the second term to bound the variance by

$$
\begin{aligned}
&\mathbb{E}_{x_s,Q}[\big((v^\top \mathbb{E}_{X\sim\nu}[P(X)XX^\top]^{-1}Q(x_s)x_s y_s\big)^2] \\
&= \mathbb{E}_{x_s,Q}[\big(v^\top \mathbb{E}_{X\sim\nu}[P(X)XX^\top]^{-1}Q(x_s)x_s(x_s^\top\theta + \xi_s)\big)^2] \\
&= \mathbb{E}_{x_s,Q}[\big(v^\top \mathbb{E}_{X\sim\nu}[P(X)XX^\top]^{-1}Q(x_s)x_s(x_s^\top\theta)\big)^2] \\
&\qquad + \mathbb{E}_{x_s,Q}[\big(v^\top \mathbb{E}_{X\sim\nu}[P(X)XX^\top]^{-1}Q(x_s)x_s\big)^2 \xi_t^2] \\
&\le B^2\mathbb{E}_{x_s,Q}[\big(v^\top \mathbb{E}_{X\sim\nu}[P(X)XX^\top]^{-1}Q(x_s)x_s\big)^2] \\
&\qquad + \sigma^2\mathbb{E}_{x_s,Q}[\big(v^\top \mathbb{E}_{X\sim\nu}[P(X)XX^\top]^{-1}Q(x_s)x_s\big)^2] \\
&= \mathbb{E}_{x_s\sim\nu}\left[(B^2+\sigma^2)\mathbb{E}_{Q(x_s)\sim P(x_s)}[v^\top \mathbb{E}_{X\sim\nu}[P(X)XX^\top]^{-1}Q(x_s)x_s x_s^\top Q(x_s)\mathbb{E}_{X\sim\nu}[P(X)XX^\top]^{-1}v]\right] \\
&\overset{\text{(i)}}{=} \mathbb{E}_{x_s\sim\nu}\left[(B^2+\sigma^2)\mathbb{E}_{Q(x_s)\sim P(x_s)}[v^\top \mathbb{E}_{X\sim\nu}[P(X)XX^\top]^{-1}Q(x_s)x_s x_s^\top \mathbb{E}_{X\sim\nu}[P(X)XX^\top]^{-1}v]\right] \\
&\le \mathbb{E}_{x_s\sim\nu}\left[(B^2+\sigma^2)v^\top \mathbb{E}_{X\sim\nu}[P(X)XX^\top]^{-1}P(x_s)x_s x_s^\top \mathbb{E}_{X\sim\nu}[P(X)XX^\top]^{-1}v]\right],
\end{aligned}
$$

where we used that $Q(x_s)^2 = Q(x_s)$ in equality (i) above. Thus, we have

$$
\begin{aligned}
&\mathbb{V}\text{ar}(v^\top \mathbb{E}_{X\sim\nu}[P(X)XX^\top]^{-1}Q(x_s)x_s y_s) \\
\le&(B^2+\sigma^2)v^\top(\mathbb{E}_{X\sim\nu}[P(X)XX^\top]^{-1}\mathbb{E}_{x_s\sim\nu}[P(x_s)x_s x_s^\top](\mathbb{E}_{X\sim\nu}[P(X)XX^\top]^{-1})v \\
=&(B^2+\sigma^2)\|v\|^2_{(\mathbb{E}_{X\sim\nu}[P(X)XX^\top]^{-1}}
\end{aligned}
$$

By using the property of Catoni estimator stated in Definition D.2.3, we have $c_0 = \sqrt{2}$ and

$$
\begin{aligned}
&|\langle\theta_*, v\rangle - w_v| \\
=&|\text{Catoni}(\{\langle v, \mathbb{E}_{X\sim\nu}[P(X)XX^\top]^{-1}Q(x_s)x_s y_s\rangle\}_{s=1}^n) - \mathbb{E}[\langle v, \mathbb{E}_{X\sim\nu}[P(X)XX^\top]^{-1}Q(x_s)x_s y_s\rangle]| \\
\le&\sqrt{2}\sqrt{\left(\mathbb{V}\text{ar}(\langle v, \mathbb{E}_{X\sim\nu}[P(X)XX^\top]^{-1}Q(x_s)x_s y_s\rangle)\right)\frac{\log(\frac{2}{\delta})}{n/2}} \\
&\qquad\qquad\qquad\qquad\text{(with probability at least } 1-\delta \text{ if } n \ge 4\log(2/\delta))
\end{aligned}
$$

$$\leq \|v\|_{(\mathbb{E}_{X\sim\nu}[P(X)XX^\top]^{-1}} \sqrt{\frac{4(B^2+\sigma^2)\log(\frac{2}{\delta})}{n}}$$
$$= \|v\|_{\mathbb{E}_{X\sim\nu}[nP(X)XX^\top]^{-1}} \sqrt{4(B^2+\sigma^2)\log(2/\delta)}.$$

Finally, the proof is complete by taking union bounding over all $v \in \mathcal{V}$. □

Lemma D.2.5. *Holds*

$$\max_{z,z'\in\mathcal{Z}_\ell} \frac{\|z-z'\|^2_{\mathbb{E}_{X\sim\nu}[\tau P(X)XX^\top]^{-1}}}{\epsilon_\ell^2} \leq 64 \max_{z\in\mathcal{Z}\setminus z_*} \frac{\|z-z_*\|^2_{\mathbb{E}_{X\sim\nu}[\tau P(X)XX^\top]^{-1}}}{\langle z-z_*,\theta_*\rangle^2}$$

Proof. Let $\mathcal{S}_\ell = \{z \in \mathcal{Z} : \langle z_* - z, \theta_*\rangle \leq 4\epsilon_\ell\}$. We have

$$\begin{aligned}
\max_{z,z'\in\mathcal{Z}_\ell} \frac{\|z-z'\|^2_{\mathbb{E}_{X\sim\nu}[\tau P(X)XX^\top]^{-1}}}{\epsilon_\ell^2} &\leq \max_{z,z'\in\mathcal{S}_\ell} \frac{\|z-z'\|^2_{\mathbb{E}_{X\sim\nu}[\tau P(X)XX^\top]^{-1}}}{\epsilon_\ell^2} \\
&= 16 \max_{z,z'\in\mathcal{S}_\ell} \frac{\|z-z'\|^2_{\mathbb{E}_{X\sim\nu}[\tau P(X)XX^\top]^{-1}}}{(4\epsilon_\ell)^2} \\
&\leq 64 \max_{z\in\mathcal{S}_\ell} \frac{\|z-z_*\|^2_{\mathbb{E}_{X\sim\nu}[\tau P(X)XX^\top]^{-1}}}{(4\epsilon_\ell)^2} \\
&= 64 \max_{z\in\mathcal{S}_\ell\setminus z_*} \frac{\|z-z_*\|^2_{\mathbb{E}_{X\sim\nu}[\tau P(X)XX^\top]^{-1}}}{\max\{(4\epsilon_\ell)^2, \langle z-z_*,\theta_*\rangle^2\}} \\
&\leq 64 \max_{z\in\mathcal{Z}\setminus z_*} \frac{\|z-z_*\|^2_{\mathbb{E}_{X\sim\nu}[\tau P(X)XX^\top]^{-1}}}{\langle z-z_*,\theta_*\rangle^2}.
\end{aligned}$$

□

Reparameterization

Proposition D.2.6. *Fix $\nu \in \triangle_\mathcal{X}$ and any $\lambda \in \triangle_\mathcal{X}$. Define $\|\lambda/\nu\|_\infty = \sup_{x\in\mathcal{X}} \lambda(x)/\nu(x)$ and $\rho(\lambda) = \max_{z\neq z_*} \frac{\|z-z_*\|^2_{\mathbb{E}_{X\sim\lambda}[XX^\top]^{-1}}}{\langle z_*-z,\theta_*\rangle^2}$. For any $t, \beta \in \mathbb{R}_+$ the following optimization problems achieve the same value*

- $\min_{P:\mathcal{X}\to[0,1]} t\,\mathbb{E}_{X\sim\nu}[P(X)]$ *subject to* $\max_{z\neq z_*} \frac{\|z-z_*\|^2_{\mathbb{E}_{X\sim\nu}[P(X)XX^\top]^{-1}}}{\langle z_*-z,\theta_*\rangle^2}\beta \leq t$
- $\min_{\lambda\in\triangle_\mathcal{X}} \rho(\lambda)\beta \quad$ *subject to* $\quad \|\lambda/\nu\|_\infty \rho(\lambda)\beta \leq t$

Let us first prove a simple lemma.

Lemma D.2.7. *Let $\mathcal{P}$ denote the set of all functions $P : \mathcal{X} \to [0,1]$. And for any $\nu \in \triangle_{\mathcal{X}}$ with support $\mathcal{X}$ let $\mathcal{P}' = \{\kappa\lambda_x/\nu_x : \lambda \in \triangle_{\mathcal{X}}, \kappa \geq 0 : \kappa\lambda_x/\nu_x \in [0,1]\}$. Then $\mathcal{P} = \mathcal{P}'$.*

Proof. Fix any $P \in \mathcal{P}$. If $\lambda_x = P_x\nu_x/\|P \circ \nu\|_1$ and $\kappa = \|P \circ \nu\|_1$ then $\kappa\lambda/\nu \in \mathcal{P}'$ and is equal to P. This implies $\mathcal{P} \subseteq \mathcal{P}'$.

For the other direction, fix any $\lambda \in \triangle_{\mathcal{X}}$ and $\kappa \geq 0$ such that $\kappa\lambda_x/\nu_x \in [0,1]$ for all x. If $P = \kappa\lambda/\nu$ then $P \in \mathcal{P}$ which implies $\mathcal{P}' \subseteq \mathcal{P}$ and concludes the proof. □

Proof of Proposition D.2.6. Using the above lemma we have that

$$\min_{P:\mathcal{X}\to[0,1]} t\,\mathbb{E}_{X\sim\nu}[P(X)] \quad \text{subject to} \quad \max_{z\neq z_*} \frac{\|z - z_*\|^2_{\mathbb{E}_{X\sim\nu}[P(X)XX^\top]^{-1}}}{\langle z_* - z, \theta_*\rangle^2}\beta \leq t$$

is equivalent to

$$\min_{\kappa\geq 0,\lambda\in\triangle_{\mathcal{X}}} t\,\mathbb{E}_{X\sim\nu}[\kappa\lambda(X)/\nu(X)] \quad \text{subject to} \quad \max_{z\neq z_*} \frac{\|z - z_*\|^2_{\mathbb{E}_{X\sim\nu}[\kappa\lambda(X)/\nu(X)XX^\top]^{-1}}}{\langle z_* - z, \theta_*\rangle^2}\beta \leq t$$
$$\kappa\lambda(x)/\nu(x) \leq 1 \quad \forall x \in \mathcal{X}$$

which is equal to, after simplifying,

$$\min_{\kappa\geq 0,\lambda\in\triangle_{\mathcal{X}}} t\,\kappa \quad \text{subject to} \quad \max_{z\neq z_*} \frac{\|z - z_*\|^2_{\mathbb{E}_{X\sim\lambda}[XX^\top]^{-1}}}{\langle z_* - z, \theta_*\rangle^2}\beta \leq t\kappa$$
$$\kappa\lambda(x)/\nu(x) \leq 1 \quad \forall x \in \mathcal{X}$$

which is equal to

$$\min_{u\geq 0,\lambda\in\triangle_{\mathcal{X}}} u \quad \text{subject to} \quad \rho(\lambda)\beta \leq u$$
$$\|\lambda/\nu\|_\infty \leq \frac{t}{u}.$$

Note, there exists a feasible (λ, u) precisely when there exists a $\lambda \in \triangle_{\mathcal{X}}$ such that

$\|\lambda/\nu\|_\infty \rho(\lambda) \le t$, in which case the optimization problem is equal to

$$\min_{\lambda \in \triangle_{\mathcal{X}}} \rho(\lambda)\beta \quad \text{subject to} \quad \|\lambda/\nu\|_\infty \rho(\lambda)\beta \le t$$

□

D.3 Analysis of the Optimization Problem

D.3.1 Proof of Theorem 5.4.1

For simplicity, we will use μ instead of μ_b to denote the number that controls the intensity of barrier function.

The proof relies on analyzing another function $\overline{D} : \mathbb{R}^{d\times d}_{\succeq \mathbf{0}} \mapsto \mathbb{R}$. For simplicity, first, we define

$$h_\Lambda(x) = P_\Lambda(x) - \mu\left(\log(1 - P_\Lambda(x)) + \log(P_\Lambda(x))\right) - P_\Lambda(x) x^\top \Lambda x. \tag{D.2}$$

Recall that our dual objective is $D(\mathbf{\Lambda}) = \mathbb{E}_{X\sim\nu}\left[h_\Lambda(X)\right] + \frac{1}{c_\ell^2}\sum_{y\in\mathcal{Y}_\ell} y^\top \Lambda_y y$. Since the first term in $\mathbb{E}_{X\sim\nu}\left[h_\Lambda(X)\right]$ only depends on $\Lambda = \sum_{y\in\mathcal{Y}_\ell}\Lambda_y$, we can consider the following optimization problem.

$$\begin{aligned} f(\Lambda) = \max_{\Lambda_y} \quad & \textstyle\sum_{y\in\mathcal{Y}_\ell} y^\top \Lambda_y y \\ \text{subject to} \quad & \textstyle\sum_{y\in\mathcal{Y}_\ell} \Lambda_y = \Lambda \\ & \Lambda_y \succeq \mathbf{0}, \quad \forall y \in \mathcal{Y}_\ell. \end{aligned} \tag{D.3}$$

Then, the alternative dual objective $\overline{D}(\Lambda)$ is defined as $\overline{D}(\Lambda) = \mathbb{E}_{X\sim\nu}\left[h_\Lambda(X)\right] + \frac{1}{c_\ell^2} f(\Lambda)$. We can immediately see that maximizing $\overline{D}(\cdot)$ is equivalent to maximizing $D(\cdot)$. In particular, let $\Lambda^* \in \arg\max_{\Lambda\succeq\mathbf{0}} \overline{D}(\Lambda)$ and $\left(\Lambda_y^*\right)_{y\in\mathcal{Y}_\ell}$ be the set of PSD matrices that solve problem (D.3) and evaluate $f(\Lambda^*)$. We can see that $\left(\Lambda_y^*\right)_{y\in\mathcal{Y}_\ell}$ also maximizes $D(\cdot)$. Conversely, for $\mathbf{\Lambda}^* = \left(\Lambda_y^*\right)_{y\in\mathcal{Y}_\ell} \in \arg\max_{\Lambda_y\succeq\mathbf{0},\forall y} D(\mathbf{\Lambda})$, we also have $\sum_{y\in\mathcal{Y}_\ell}\Lambda_y^* \in \arg\max_{\Lambda\succeq\mathbf{0}} \overline{D}(\Lambda)$.

Further, we also define their empirical version D_E and $\overline{D}_E$ with extra i.i.d. samples $x_1, \ldots, x_u$ as

$$D_E(\mathbf{\Lambda}) = \frac{1}{u}\sum_{i=1}^u h_\Lambda(x_i) + \frac{1}{c_\ell^2}\sum_{y\in\mathcal{Y}_\ell} y^\top\Lambda_y y \quad \text{and} \quad \overline{D}_E(\Lambda) = \frac{1}{u}\sum_{i=1}^u h_\Lambda(x_i) + \frac{1}{c_\ell^2} f(\Lambda). \tag{D.4}$$

Recall that the problem Algorithm 9 tries to solve is

$$\begin{aligned} \min_P \quad & \mathbb{E}_{X\sim\nu}[P(X) - \mu(\log(1-P(X)) + \log(P(X)))] \\ \text{subject to} \quad & \mathbb{E}_{X\sim\nu}\left[P(X)XX^\top\right] \succeq \frac{1}{c_\ell^2} yy^\top, \quad \forall y \in \mathcal{Y}_\ell. \end{aligned} \tag{D.5}$$

We will restate a more precise version of Theorem 5.4.1 and then prove it.

Theorem D.3.1. *Suppose* $\|x\|_2 \le M$ *for any* $x \in \operatorname{supp}(\nu)$ *and* $\Sigma = \mathbb{E}_{X\sim\nu}\left[XX^\top\right]$ *is invertible. Let* $\mathbf{\Lambda}^* \in \arg\max_{\Lambda_y \succeq \mathbf{0}, \forall y} D(\mathbf{\Lambda})$ *and* $\kappa(\Sigma) = \frac{\lambda_{\max}(\Sigma)}{\lambda_{\min}(\Sigma)}$ *be condition number. Assume* $\|\Lambda^*\|_F > 0$ *and define* $\omega = \min_{\Gamma\in\mathbb{S}^d:\|\Gamma\|_F=1} \mathbb{E}_{X\sim\nu}\left[\left(X^\top\Gamma X\right)^2\right]$, *where* $\mathbb{S}^d$ *is the set of* $d\times d$ *symmetric matrices. Let* $|\mathcal{Y}_\ell|\, C_\ell^2 = \frac{1}{c_\ell^2}\sum_{y\in\mathcal{Y}_\ell}\|y\|_2^4$.

Then, $\Lambda^* = \sum_{y\in\mathcal{Y}_\ell}\Lambda_y^*$ *is unique. Further, for any* $\epsilon > 0$ *and* $\delta > 0$, *suppose it holds that*

$$\begin{aligned} \mu &\le \min\left\{\sqrt{\frac{3\kappa(\Sigma)\,\|\Lambda^*\|_F\,M^2}{8}\cdot\frac{1+\epsilon}{\epsilon}}, \frac{4}{9}\|\Lambda^*\|_F^2\,M^4, \frac{1}{2\sqrt{3}}\right\} \\ K &\ge \frac{288\kappa(\Sigma)^2\,|\mathcal{Y}_\ell|^3\,\|\Lambda^*\|_F^4\,M^4(M^4+C_\ell^2)\cdot\left(2\,\|\Lambda^*\|_F\,M^2+1\right)^4\log(6/\delta)}{\omega^2\mu^6}\cdot\left(\frac{1+\epsilon}{\epsilon}\right)^2 \\ u &\ge \frac{576\kappa(\Sigma)^2\,\|\Lambda^*\|_F^2\,M^8\cdot\left(2\,\|\Lambda^*\|_F\,M^2+1\right)^4\log(6/\delta)}{\omega^2\mu^6}\cdot\left(\frac{1+\epsilon}{\epsilon}\right)^2. \end{aligned}$$

Then, with probability at least $1-\delta$, *Algorithm 9 will output* $\widetilde{\Lambda}$ *that satisfies*

- $y^\top\mathbb{E}_{X\sim\nu}\left[P_{\widetilde{\Lambda}}(X)XX^\top\right]^{-1}y \le (1+\epsilon)c_\ell^2, \quad \forall y\in\mathcal{Y}_\ell.$
- $\mathbb{E}_{X\sim\nu}\left[P_{\widetilde{\Lambda}}(X)\right] \le \mathbb{E}_{X\sim\nu}\left[\widetilde{P}(X)\right] + 4\sqrt{\mu}$, *where* $\widetilde{P}$ *is the optimal solution to problem* (D.13).

Proof. **First Bullet Point.** Fix some $\epsilon > 0$. Let $\hat{\mathbf{\Lambda}}$ and corresponding $\hat{\Lambda} = \sum_{y\in\mathcal{Y}_\ell}\hat{\Lambda}_y$ be the parameters obtained by Algorithm 9 just before the re-scaling step, which means that at line 10 of Algorithm 9, the assignment of $\hat{\Lambda}_y$ to each $y\in\mathcal{Y}_\ell$ has been optimized by solving problem (D.3). That is, we have $D(\hat{\mathbf{\Lambda}}) = \overline{D}(\hat{\Lambda})$ and $D_E(\hat{\mathbf{\Lambda}}) = \overline{D}_E(\hat{\Lambda})$. Let $\widetilde{\mathbf{\Lambda}}$ and $\widetilde{\Lambda}$ be the ones after the re-scaling step. Then, by Theorem 3.13 of Orabona [2019], with probability

at least $1-\frac{\delta}{3}$, it holds that

$$\overline{D}(\Lambda^*) - \overline{D}(\hat{\Lambda}) = D(\mathbf{\Lambda}^*) - D(\hat{\mathbf{\Lambda}}) \leq \frac{\mathrm{Reg}(K) + 2\sqrt{2K\log(6/\delta)}}{K},$$

where $\mathrm{Reg}(K)$ is the regret of running projected stochastic gradient ascent for K steps with η_k specified in Algorithm 9. Meanwhile, by Theorem 4.14 of Orabona [2019] also, we have $\mathrm{Reg}(K) = \sqrt{2}B^2\sqrt{\sum_{k=1}^{K}\sum_{y\in\mathcal{Y}_\ell}\left\|g_{k,y}\right\|_2^2}$, where $B = \sqrt{|\mathcal{Y}_\ell|}\left\|\Lambda^*\right\|_F$ bound the norm of $\mathbf{\Lambda}^* = \left(\Lambda_y^*\right)_{y\in\mathcal{Y}_\ell}$. Since $g_{k,y} = \frac{yy^\top}{c_\ell^2} - P_{\hat{\Lambda}^{(k)}}(x_k)x_k x_k^\top$, we can easily get $\sum_{y\in\mathcal{Y}_\ell}\left\|g_{k,y}\right\|_2^2 \leq 2\left|\mathcal{Y}_\ell\right|M^4 + \frac{2}{c_\ell^2}\sum_{y\in\mathcal{Y}_\ell}\left\|y\right\|_2^4 = 2\left|\mathcal{Y}_\ell\right|M^4 + 2\left|\mathcal{Y}_\ell\right|C_\ell^2$. Thus, we have

$$\mathrm{Reg}(K) \leq 2\left|\mathcal{Y}_\ell\right|\left\|\Lambda^*\right\|_F^2\sqrt{\left|\mathcal{Y}_\ell\right|M^4 + \left|\mathcal{Y}_\ell\right|C_\ell^2}\cdot\sqrt{K} := C_{\mathrm{Reg}}\sqrt{K} \tag{D.6}$$

$$\Longrightarrow\ \overline{D}(\Lambda^*) - \overline{D}(\hat{\Lambda}) \leq \frac{C_{\mathrm{Reg}} + 2\sqrt{2\log(6/\delta)}}{\sqrt{K}}, \tag{D.7}$$

We now consider the effect of using u i.i.d. samples in the re-scaling step. First, since re-scaling always increases the function value, we must have $D_E(\hat{\mathbf{\Lambda}}) \leq D_E(\widetilde{\mathbf{\Lambda}})$. Meanwhile, since $D_E(\hat{\mathbf{\Lambda}}) = \overline{D}_E(\hat{\Lambda})$, by Lemma D.3.7, we have $D_E(\hat{\mathbf{\Lambda}}) = \overline{D}_E(\hat{\Lambda})$, which together implies $\overline{D}_E(\hat{\Lambda}) \leq \overline{D}_E(\widetilde{\Lambda})$.

By Lemma D.3.2, we know that Λ^* is unique and as long as $\mu \leq \frac{1}{2\sqrt{3}}$, $\overline{D}(\Lambda)$ is G-strongly concave with respect to ℓ_2 norm over $\mathcal{S} = \{\Lambda \succeq \mathbf{0} : \left\|\Lambda\right\|_F \leq 2\left\|\Lambda^*\right\|_F\}$, where G is defined in Eq. (D.14). Thus, by Lemma D.3.8, if K is large enough such that

$$\overline{D}(\Lambda^*) - \overline{D}(\hat{\Lambda}) \leq \frac{C_{\mathrm{Reg}} + 2\sqrt{2\log(6/\delta)}}{\sqrt{K}} \leq \frac{G\left\|\Lambda^*\right\|_F}{2},$$

then $\left\|\hat{\Lambda} - \Lambda^*\right\|_F \leq \left\|\Lambda^*\right\|_F$, which implies $\left\|\hat{\Lambda}\right\|_F \leq 2\left\|\Lambda^*\right\|_F$. That is, $\hat{\Lambda} \in \mathcal{S}$. Then, under this condition, by using Lemma D.3.5, when $\mu \leq \frac{4}{9}\left\|\Lambda^*\right\|_F M^4$ and

$$u \geq \left(\frac{6\kappa(\Sigma)\left\|\Lambda^*\right\|_F M^4\left(2+\sqrt{2\log(6/\delta)}\right)}{G\mu^2}\cdot\frac{1+\epsilon}{\epsilon}\right)^2, \tag{D.8}$$

for $\widetilde{\Lambda}$ after re-scaling, with probability at least $1-\frac{\delta}{3}$, it holds simultaneously that

$$\left|\overline{D}_E(\hat{\Lambda})-\overline{D}(\hat{\Lambda})\right| \le \frac{G\mu^2}{3M^2\kappa(\Sigma)}\cdot\frac{\epsilon}{1+\epsilon} \quad \text{and} \quad \left|\overline{D}_E(\widetilde{\Lambda})-\overline{D}(\widetilde{\Lambda})\right| \le \frac{G\mu^2}{3M^2\kappa(\Sigma)}\cdot\frac{\epsilon}{1+\epsilon} \tag{D.9}$$

$$\begin{aligned}
\implies \overline{D}(\Lambda^*)-\overline{D}(\widetilde{\Lambda}) &\le \overline{D}(\Lambda^*)-\overline{D}(\hat{\Lambda})+\overline{D}(\hat{\Lambda})-\overline{D}(\widetilde{\Lambda}) \\
&\le \overline{D}(\Lambda^*)-\overline{D}(\hat{\Lambda})+\overline{D}(\hat{\Lambda})-\overline{D}_E(\hat{\Lambda})+\overline{D}_E(\widetilde{\Lambda})-\overline{D}(\widetilde{\Lambda}) \\
&\qquad\qquad \text{(Since } \overline{D}_E(\hat{\Lambda}) \le \overline{D}_E(\widetilde{\Lambda})\text{)} \\
&\le \frac{C_{\mathrm{Reg}}+2\sqrt{2\log(6/\delta)}}{\sqrt{K}}+\frac{2G\mu^2}{3M^2\kappa(\Sigma)}\cdot\frac{\epsilon}{1+\epsilon}. \\
&\qquad\qquad \text{(By Eq. (D.7) and (D.9))}
\end{aligned}$$

Since $\widetilde{\Lambda}$ is a smaller re-scaling of $\hat{\Lambda}$, we have $\widetilde{\Lambda} \in \mathcal{S}$, which implies $\frac{G}{2}\left\|\Lambda^*-\widetilde{\Lambda}\right\|_F \le \overline{D}(\Lambda^*)-\overline{D}(\widetilde{\Lambda})$ by property of strongly concave function Bertsekas [2009]. Therefore, by Lemma D.3.9, to guarantee an at most ϵ multiplicative constraint violation, it is sufficient to choose K such that

$$\begin{aligned}
\frac{G}{2}\left\|\Lambda^*-\widetilde{\Lambda}\right\|_F &\le \overline{D}(\Lambda^*)-\overline{D}(\widetilde{\Lambda}) \\
&\le \frac{C_{\mathrm{Reg}}+2\sqrt{2\log(6/\delta)}}{\sqrt{K}}+\frac{2G\mu^2}{3M^2\kappa(\Sigma)}\cdot\frac{\epsilon}{1+\epsilon} \\
&\le \min\left\{\frac{4G\mu^2}{3M^2\kappa(\Sigma)}\cdot\frac{\epsilon}{1+\epsilon}, \frac{G\left\|\Lambda^*\right\|_F}{2}\right\} \\
&= \frac{4G\mu^2}{3M^2\kappa(\Sigma)}\cdot\frac{\epsilon}{1+\epsilon}. \qquad \text{(If } \mu \le \sqrt{\tfrac{3\kappa(\Sigma)\|\Lambda^*\|_F M^2}{8}\cdot\tfrac{1+\epsilon}{\epsilon}}\text{)}
\end{aligned}$$

An algebraic rearrangement gives us

$$K \ge \left(\frac{3\kappa(\Sigma)M^2\left(C_{\mathrm{Reg}}+2\sqrt{2\log(6/\delta)}\right)}{2G\mu^2}\cdot\frac{1+\epsilon}{\epsilon}\right)^2. \tag{D.10}$$

Second Bullet Point. We then prove the upper bound for primal objective value $\mathbb{E}_{X\sim\nu}\left[P_{\widetilde{\Lambda}}(X)\right]$, which explains the reason why an extra re-scaling step is needed. Define $g(s) = D_E(s\cdot\widetilde{\boldsymbol{\Lambda}})$. By construction, we know that $g(s)$ is maximized at $s=1$ because

$\widetilde{\mathbf{\Lambda}} = s^* \cdot \hat{\mathbf{\Lambda}}$, where $s^* = \arg\max_{s\in[0,1]} D_E(s \cdot \hat{\mathbf{\Lambda}})$. Therefore, we have $g'(1) \geq 0$, which in turn gives us

$$g'(1) = \frac{1}{c_\ell^2} \sum_{y\in\mathcal{Y}_\ell} y^\top \widetilde{\Lambda}_y y - \frac{1}{u} \sum_{i=1}^{u} P_{\widetilde{\Lambda}}(x_i) x_i^\top \widetilde{\Lambda} x_i \geq 0.$$

By the concentration inequality in Lemma D.3.5, we know that when

$$u \geq \left(\frac{2 \left\|\Lambda^*\right\|_F M^2 \left(\left\|\Lambda^*\right\|_F M^2 + \mu\sqrt{2\log(6/\delta)} \right)}{\mu^{3/2}} \right)^2, \tag{D.11}$$

with probability at least $1 - \frac{\delta}{3}$, it holds that

$$\begin{aligned} &\left| \mathbb{E}_{X\sim\nu} \left[P_\Lambda(X) X^\top \Lambda X \right] - \frac{1}{u} \sum_{i=1}^{u} P_\Lambda(x_i) x_i^\top \Lambda x_i \right| \leq \sqrt{\mu} \\ \implies &\frac{1}{c_\ell^2} \sum_{y\in\mathcal{Y}_\ell} y^\top \widetilde{\Lambda}_y y - \mathbb{E}_{X\sim\nu} \left[P_{\widetilde{\Lambda}}(X) X^\top \widetilde{\Lambda} X \right] + \sqrt{\mu} \geq 0. \end{aligned} \tag{D.12}$$

Now, let $\widetilde{P}$ be the optimal solution of problem (D.13) and $\hat{P}$ be the optimal solution of the same problem with bound constraint $\mu \leq P(x) \leq 1 - \mu$.

$$\begin{aligned} \min_P \quad & \mathbb{E}_{X\sim\nu} \left[P(X) \right] \\ \text{subject to} \quad & y^\top \mathbb{E}_{X\sim\nu} \left[P(X) X X^\top \right]^{-1} y \leq c_\ell^2, \quad \forall y \in \mathcal{Y}_\ell, \\ & 0 \leq P(x) \leq 1 - \mu, \quad \forall x \in \mathcal{X}. \end{aligned} \tag{D.13}$$

Then, we can notice that

$$\begin{aligned} &\mathbb{E}_{X\sim\nu} \left[P_{\widetilde{\Lambda}}(X) \right] \\ \leq &\mathbb{E}_{X\sim\nu} \left[P_{\widetilde{\Lambda}}(X) - \mu(\log(1 - P_{\widetilde{\Lambda}}(X)) + \log(P_{\widetilde{\Lambda}}(X))) \right] \\ \leq &\mathbb{E}_{X\sim\nu} \left[P_{\widetilde{\Lambda}}(X) - \mu(\log(1 - P_{\widetilde{\Lambda}}(X)) + \log(P_{\widetilde{\Lambda}}(X))) \right] \\ &\quad + \frac{1}{c_\ell^2} \sum_{y\in\mathcal{Y}_\ell} y^\top \widetilde{\Lambda}_y y - \mathbb{E}_{X\sim\nu} \left[P_{\widetilde{\Lambda}}(X) X^\top \widetilde{\Lambda} X \right] + \sqrt{\mu} \qquad \text{(By Eq. (D.12))} \\ = &\inf_P \mathcal{L}(P, \widetilde{\mathbf{\Lambda}}) + \sqrt{\mu} \qquad \text{(By definition of Lagrangian function and how we solve for } P_\Lambda) \\ \leq &\max_{\Lambda_y \succeq \mathbf{0}, \forall y\in\mathcal{Y}_\ell} \inf_P \mathcal{L}\left(P, \mathbf{\Lambda}\right) + \sqrt{\mu} \end{aligned}$$

$$
\begin{aligned}
=&\mathbb{E}_{X\sim\nu}\left[P_{\Lambda^*}(X)-\mu(\log(1-P_{\Lambda^*}(X))+\log(P_{\Lambda^*}(X)))\right]+\sqrt{\mu}\\
\leq&\mathbb{E}_{X\sim\nu}\left[\hat{P}(X)-\mu\log(1-\hat{P}(X))\right]-\mu\log\left(\hat{P}(X)\right)+\sqrt{\mu}\\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\text{(Since } \hat{P} \text{ is feasible to problem (D.5))}\\
\leq&\mathbb{E}_{X\sim\nu}\left[\hat{P}(X)\right]+3\sqrt{\mu}, \qquad\qquad\qquad\qquad\qquad\qquad\text{(Since } -a\log(a)\leq\sqrt{a} \text{ for } a\in(0,1))\\
\leq&\mathbb{E}_{X\sim\nu}\left[\widetilde{P}(X)\right]+4\sqrt{\mu}. \qquad\text{(Since } \hat{P}(x) \text{ can have at most } \mu \text{ more contribution than } \widetilde{P})
\end{aligned}
$$

Therefore, in summary, Suppose K and u satisfy conditions specified in Eq. (D.10), (D.8) and (D.11) and $\mu\leq\min\left\{\sqrt{\frac{3\kappa(\Sigma)\|\Lambda^*\|_F M^2}{8}\cdot\frac{1+\epsilon}{\epsilon}},\frac{4}{9}\|\Lambda^*\|_F^2 M^4,\frac{1}{2\sqrt{3}}\right\}$, where C_{Reg} and G are defined in Eq. (D.6) and (D.14), respectively. Then. by applying a simple union bound, with probability at least $1-\delta$, the output of Algorithm 9 $\widetilde{\Lambda}$ satisfies $y^\top\mathbb{E}_{X\sim\nu}\left[P(X)XX^\top\right]^{-1}y\leq(1+\epsilon)c_\ell^2,\forall y\in\mathcal{Y}_\ell$ and $\mathbb{E}_{X\sim\nu}\left[P_{\widetilde{\Lambda}}(X)\right]\leq\mathbb{E}_{X\sim\nu}\left[\widetilde{P}(X)\right]+4\sqrt{\mu}$. □

D.3.2 Relevant Lemmas

Strong Concavity of $\overline{D}(\Lambda)$

Lemma D.3.2. *As long as* $\mu\leq\frac{1}{2\sqrt{3}}$, $\overline{D}(\Lambda)$ *is* G*-strongly concave with respect to* ℓ_2*-norm on the bounded region* $\mathcal{S}=\{\Lambda\succeq\mathbf{0}:\|\Lambda\|_F\leq2\|\Lambda^*\|_F\}$ *with coefficient*

$$
G=\frac{\mu}{2\left(2\|\Lambda^*\|_F M^2+1\right)^2}\cdot\min_{\Gamma\in\mathbb{S}^d:\|\Gamma\|_F=1}\mathbb{E}_{X\sim\nu}\left[\left(X^\top\Gamma X\right)^2\right]. \tag{D.14}
$$

Because of this, as a corollary, Λ^* *will be unique.*

Proof. By Lemma D.3.3, since $f(\Lambda)$ is concave in Λ, it is sufficient to prove that $\mathbb{E}_{X\sim\nu}\left[h_\Lambda(X)\right]$ is G-strongly concave on $\mathcal{S}$, where $h_\Lambda(x)$ is defined in Eq. (D.2). Then, we have

$$
-\nabla_\Lambda^2\mathbb{E}_{X\sim\nu}\left[h_\Lambda(X)\right]=\mathbb{E}_{X\sim\nu}\left[\frac{\mathrm{d}P_\Lambda}{\mathrm{d}q_\Lambda}(X)\mathrm{vec}\left(XX^\top\right)\mathrm{vec}\left(XX^\top\right)^\top\right].
$$

Since $\|x\|_2\leq M$, for any $\Lambda\in\mathcal{S}$, we have $q_\Lambda(x)=x^\top\Lambda x-1\leq2\|\Lambda^*\|_F M^2+1$. By Lemma, D.3.11, we know that if $12\mu^2\leq\left(2\|\Lambda^*\|_F M^2+1\right)^2$, which can be done by choosing

$\mu \leq \frac{1}{2\sqrt{3}}$, we have $\frac{\mathrm{d}P_\Lambda}{\mathrm{d}q_\Lambda}(x) \geq \frac{\mu}{2\left(2\|\Lambda^*\|_F M^2+1\right)^2}$ for any $x \in \mathcal{X}$ and $\Lambda \in \mathcal{S}$. Therefore, we have

$$-\nabla^2_\Lambda \mathbb{E}_{X\sim\nu}\left[h_\Lambda(X)\right] \succeq \gamma \cdot \mathbb{E}_{X\sim\nu}\left[\mathrm{vec}\left(XX^\top\right)\mathrm{vec}\left(XX^\top\right)^\top\right]$$

Now, let $\mathbb{S}$ be the set of all $d \times d$ symmetric matrices. It is obvious that $\mathbb{S}$ is a subspace of the vector space of all $d \times d$ matrices and $\mathcal{S} \subseteq \mathbb{S}$. Thus, by applying Lemma D.3.4, we can conclude that $\mathbb{E}_{X\sim\nu}\left[h_\Lambda(X)\right]$ is G-strongly concave on $\mathcal{S}$ with respect to ℓ_2 norm and

$$\begin{aligned} G &= \frac{\mu}{2\left(2\left\|\Lambda^*\right\|_F M^2+1\right)^2} \cdot \min_{\Gamma\in\mathbb{S}^d:\|\Gamma\|_F=1} \mathrm{vec}(\Gamma)^\top \mathbb{E}_{X\sim\nu}\left[\mathrm{vec}\left(XX^\top\right)\mathrm{vec}\left(XX^\top\right)^\top\right]\mathrm{vec}(\Gamma) \\ &= \frac{\mu}{2\left(2\left\|\Lambda^*\right\|_F M^2+1\right)^2} \cdot \min_{\Gamma\in\mathbb{S}^d:\|\Gamma\|_F=1} \mathbb{E}_{X\sim\nu}\left[\left(X^\top\Gamma X\right)^2\right]. \end{aligned}$$

Thus the proof is complete. □

Lemma D.3.3. *$f(\Lambda)$ defined in Eq. (D.3) is concave in Λ.*

Proof. To show its concavity, consider $\Lambda^{(1)} \succeq \mathbf{0}$, $\Lambda^{(2)} \succeq \mathbf{0}$ and some $\gamma \in (0,1)$. Let $(\Lambda_y^{(i)})_{y\in\mathcal{Y}_\ell}$ be the optimal solution obtained by evaluating $f(\Lambda^{(i)})$ for $i \in \{1,2\}$. Then, we can notice that

$$\begin{aligned} \gamma f(\Lambda^{(1)}) + (1-\gamma) f(\Lambda^{(2)}) &= \gamma \sum_{y\in\mathcal{Y}_\ell} y^\top \Lambda_y^{(1)} y + (1-\gamma) \sum_{y\in\mathcal{Y}_\ell} y^\top \Lambda_y^{(2)} y \\ &= \sum_{y\in\mathcal{Y}_\ell} y^\top (\gamma\Lambda_y^{(1)} + (1-\gamma)\Lambda_y^{(2)}) y \\ &\leq f(\gamma\Lambda^{(1)} + (1-\gamma)\Lambda^{(2)}). \end{aligned}$$

The last inequality above holds because $\sum_{y\in\mathcal{Y}_\ell} \Lambda_y^{(i)} = \Lambda^{(i)}$ for $i \in \{1,2\}$ and thus $\sum_{y\in\mathcal{Y}_\ell}\left(\gamma\Lambda_y^{(1)} + (1-\gamma)\Lambda_y^{(2)}\right) = \gamma\Lambda^{(1)} + (1-\gamma)\Lambda^{(2)}$, which means that $(\gamma\Lambda_y^{(1)} + (1-\gamma)\Lambda_y^{(2)})_{y\in\mathcal{Y}_\ell}$ is a feasible solution for problem (D.3) with parameter $\gamma\Lambda^{(1)} + (1-\gamma)\Lambda^{(2)}$. Therefore, we can conclude that $f(\Lambda)$ is concave in Λ. □

Lemma D.3.4. *Let $f : \mathbb{R}^d \mapsto \mathbb{R}$ be a convex and twice differentiable function in $\mathbb{R}^d$. If for some subspace $S \subseteq \mathbb{R}^d$, we have $\min_{w\in S:\|w\|_2=1} w^\top \nabla^2 f(x) w \geq \sigma > 0$, $\forall x \in S$, then f is*

σ-strongly convex with respect to ℓ_2-norm on S.

Proof. Suppose S has dimension m and let $v_1, \ldots, v_m$ be a set of orthonormal basis that span S. Then, for each $x \in S$, there exists unique $z \in \mathbb{R}^m$ such that $x = Vz$, where $V = \begin{bmatrix} v_1 & \ldots & v_m \end{bmatrix}$. That is, there is one-to-one correspondence between S and $\mathbb{R}^m$.

Now, we define $g : \mathbb{R}^m \mapsto \mathbb{R}$ as $g(z) = f(Vz)$. It is easy to compute $\nabla^2 g(z) = V^\top \nabla^2 f(Vz) V$. Then, notice that for any $w' \in \mathbb{R}^m$ such that $\|w'\|_2 = 1$, we have $Vw' \in S$ and $\|Vw'\|_2 = \sqrt{w'^\top V^\top V w'} = \sqrt{w'^\top w'} = 1$. Thus, we have

$$\begin{aligned}\min_{w' \in \mathbb{R}^m : \|w'\|_2 = 1} w'^\top \nabla^2 g(z) w' &= \min_{w' \in \mathbb{R}^m : \|w'\|_2 = 1} w'^\top V^\top \nabla^2 f(Vz) V w' \\ &= \min_{w \in S : \|w\|_2 = 1} w^\top \nabla^2 f(Vz) w \geq \sigma.\end{aligned}$$

Therefore, g is σ-strongly convex with respect to ℓ_2 norm. Then, for any $x_1, x_2 \in S$, there exists unique $z_1, z_2 \in \mathbb{R}^m$ such that $x_1 = Vz_1$ and $x_2 = Vz_2$. Notice that $\|z_1 - z_2\|_2 = \|x_1 - x_2\|_2$ since V preserves the norm. Further, by definition of strong convexity, for any $\alpha \in [0, 1]$, we have

$$\begin{aligned} & g(\alpha z_1 + (1-\alpha) z_2) + \frac{\sigma}{2} \alpha (1-\alpha) \|z_1 - z_2\|_2^2 \leq \alpha g(z_1) + (1-\alpha) g(z_2) \\ \implies & f(\alpha V z_1 + (1-\alpha) V z_2) + \frac{\sigma}{2} \alpha (1-\alpha) \|x_1 - x_2\|_2^2 \leq \alpha f(V z_1) + (1-\alpha) f(V z_2) \\ \implies & f(\alpha x_1 + (1-\alpha) x_2) + \frac{\sigma}{2} \alpha (1-\alpha) \|x_1 - x_2\|_2^2 \leq \alpha f(x_1) + (1-\alpha) f(x_2).\end{aligned}$$

Thus, f is also σ-strongly convex with respect to ℓ_2 norm on S. □

Concentration Inequalities

Lemma D.3.5. *Let $x_1, \ldots, x_u \sim \nu$ be i.i.d. samples. If $\left\|\hat{\Lambda}\right\|_F \leq 2 \|\Lambda^*\|_F$, $\|x\|_2 \leq M$ for any $x \in \mathcal{X}$ and $\mu \leq \frac{4}{9} \|\Lambda^*\|_F^2 M^4$, then with probability at least $1 - \frac{2\delta}{3}$, it holds for any $\Lambda \in \Theta = \left\{ s \cdot \hat{\Lambda} : s \in [0, 1] \right\}$ simultaneously that*

$$\left| \mathbb{E}_{X \sim \nu} \left[h_\Lambda(X) \right] - \frac{1}{u} \sum_{i=1}^{u} h_\Lambda(x_i) \right| \leq \frac{2 \|\Lambda^*\|_F M^2 \left(2 + \sqrt{2 \log(6/\delta)} \right)}{\sqrt{u}}$$

$$\left|\mathbb{E}_{X\sim\nu}\left[P_\Lambda(X)X^\top\Lambda X\right]-\frac{1}{u}\sum_{i=1}^{u}P_\Lambda(x_i)x_i^\top\Lambda x_i\right|\le\frac{2\left\|\Lambda^*\right\|_F M^2\left(\left\|\Lambda^*\right\|_F M^2+\mu\sqrt{2\log(6/\delta)}\right)}{\mu\sqrt{u}}.$$

Proof. To prove the first inequality, first, notice that we have $h_\Lambda(x) = -P_\Lambda(x)q_\Lambda(x) - \mu\left(\log(1-P_\Lambda(x))+\log(P_\Lambda(x))\right)$, where $q_\Lambda(x) = x^\top\Lambda x - 1$. Since $P_\Lambda(x)$, defined in Eq. (5.7), explicitly only depends on $q_\Lambda(x)$ instead of x directly, we can treat h_Λ as a function of q_Λ and define a function class $\mathcal{F} = \left\{x\mapsto x^\top(s\cdot\hat{\Lambda})x : s\in[0,1]\right\}$. It is well-known that if h_Λ is L_1-Lipschitz in q_Λ and $|h_\Lambda(x)|\le R_1$ for any $\Lambda\in\Theta$ and $x\sim\nu$, then, with probability at least $1-\frac{\delta}{3}$, it holds simultaneously for all $\Lambda\in\Theta$ that Bartlett and Mendelson [2002], Mohri et al. [2018]

$$\left|\mathbb{E}_{X\sim\nu}\left[h_\Lambda(X)\right]-\frac{1}{u}\sum_{i=1}^{u}h_\Lambda(x_i)\right|\le 2L_1\cdot\mathcal{R}_u(\mathcal{F})+R_1\sqrt{\frac{2\log(6/\delta)}{u}},\tag{D.15}$$

where $\mathcal{R}_u(\mathcal{F})$ is the Rademacher complexity of $\mathcal{F}$.

To find L_1, we can compute

$$\begin{aligned}\frac{\mathrm{d}h_\Lambda}{\mathrm{d}q_\Lambda} &= -\frac{\mathrm{d}P_\Lambda}{\mathrm{d}q_\Lambda}q_\Lambda - P_\Lambda + \frac{\mathrm{d}P_\Lambda}{\mathrm{d}q_\Lambda}\left(\frac{\mu}{1-P_\Lambda}-\frac{\mu}{P_\Lambda}\right)\\ &= -\frac{\mathrm{d}P_\Lambda}{\mathrm{d}\cdot q_\Lambda}q_\Lambda - P_\Lambda + \frac{\mathrm{d}P_\Lambda}{\mathrm{d}q_\Lambda}\cdot q_\Lambda \qquad \text{(Since } P_\Lambda \text{ satisfies Eq. (5.6))}\\ &= -P_\Lambda\end{aligned}$$

Therefore, we have $\frac{\mathrm{d}h_\Lambda}{\mathrm{d}q_\Lambda}\in\left[-1,-\frac{\mu}{3}\right]$ by Lemma D.3.11. Therefore, we can set $L_1 = 1$.

Let h_0 be the value of h_Λ when $q_\Lambda = -1$, which means $x^\top\Lambda x = 0$. To find R_1, notice that since $\frac{\mathrm{d}h_\Lambda}{\mathrm{d}q_\Lambda}\in\left[-1,-\frac{\mu}{3}\right]$, we must have $-q_\Lambda + h_0 \le h_\Lambda \le -\frac{\mu}{3}q_\Lambda + h_0$. By Lemma D.3.11, we know that $h_0\in\left[0,2\sqrt{\mu}\right]$. Therefore, we have $-x^\top\Lambda x\le h_\Lambda(x)\le -\frac{\mu}{3}x^\top\Lambda x + 3\sqrt{\mu}$ for any $x\in\mathcal{X}$ and $\Lambda\in\Theta$. Since $\|\Lambda\|_F\le\left\|\hat{\Lambda}\right\|_F\le 2\left\|\Lambda^*\right\|_F$, we have $|h_\Lambda(x)|\le 2\left\|\Lambda^*\right\|_F M^2 := R_1$, which holds when $\mu\le\frac{4}{9}\left\|\Lambda^*\right\|_F^2 M^4$. Then, by Lemma D.3.6, we know that $\mathcal{R}_u(\mathcal{F})\le\frac{2\|\Lambda^*\|_F M^2}{\sqrt{u}}$. Thus, plugging in values of L_1, R_1 and $\mathcal{R}_u(\mathcal{F})$ into Eq. (D.15) gives our first concentration inequality.

We can basically follow exactly the same strategy to prove the second concentration inequality. In particular, define $\tilde{h}_\Lambda(x) = P_\Lambda(x)x^\top\Lambda x = P_\Lambda(x)q_\Lambda(x) + P_\Lambda(x)$. Then, with

probability at least $1 - \frac{\delta}{3}$, it holds simultaneously for any $\Lambda \in \Theta$ that

$$\left| \mathbb{E}_{X \sim \nu} \left[\tilde{h}_\Lambda(X) \right] - \frac{1}{u} \sum_{i=1}^{u} \tilde{h}_\Lambda(x_i) \right| \leq 2L_2 \cdot \mathcal{R}_u(\mathcal{F}) + R_2 \sqrt{\frac{2 \log(6/\delta)}{u}}, \tag{D.16}$$

where $\left| \tilde{h}_\Lambda(x) \right| \leq R_2$ for any $x \in \mathcal{X}$, $\Lambda \in \Theta$ and $\tilde{h}_\Lambda$ is L_2-Lipschitz in q_Λ.

To find L_2, we can compute

$$\frac{\mathrm{d}\tilde{h}_\Lambda}{\mathrm{d}q_\Lambda} = P_\Lambda + \frac{\mathrm{d}P_\Lambda}{\mathrm{d}q_\Lambda} \cdot x^\top \Lambda x.$$

By Lemma D.3.11, we know that $\frac{\mathrm{d}P_\Lambda}{\mathrm{d}q_\Lambda} \in \left[0, \frac{1}{8\mu}\right]$. Thus, we have $\left| \frac{\mathrm{d}\tilde{h}_\Lambda}{\mathrm{d}q_\Lambda} \right| \leq 1 + \frac{\|\Lambda^*\|_F M^2}{4\mu} := L_2$. It is obvious that $\tilde{h}_\Lambda(x) \leq 2 \|\Lambda^*\|_F M^2 := R_2$. Thus, by plugging the values of L_2, R_2 and $\mathcal{R}_u(\mathcal{F})$ into Eq. (D.16), we can obtain the second concentration inequality.

Finally, both concentration inequalities hold simultaneously with probability at least $1 - \frac{2\delta}{3}$ by a simple union bound. □

Lemma D.3.6. *If* $\left\| \hat{\Lambda} \right\|_F \leq 2 \|\Lambda^*\|_F$, *then, we have* $\mathcal{R}_u(\mathcal{F}) \leq \sqrt{\frac{\mathbb{E}_{X \sim \nu}\left[(X^\top \hat{\Lambda} X)^2\right]}{u}} \leq \frac{2\|\Lambda^*\|_F M^2}{\sqrt{u}}$, *where* $\mathcal{F} = \left\{ x \mapsto x^\top (s \cdot \hat{\Lambda}) x : s \in [0,1] \right\}$.

Proof. Let $\sigma_1, \ldots, \sigma_u$ be i.i.d. Rademacher random variables, which are uniform over $\{-1, +1\}$. Let $x_1, \ldots, x_u \sim \nu$ be i.i.d. samples. Then, by definition of Rademacher complexity, we have

$$\begin{aligned}
\mathcal{R}_u(\mathcal{F}) &= \mathbb{E}\left[\sup_{q \in \mathcal{F}} \frac{1}{u} \sum_{i=1}^{u} \sigma_i q(x_i) \right] \\
&= \mathbb{E}\left[\sup_{s \in [0,1]} \frac{1}{u} \sum_{i=1}^{u} \sigma_i x_i^\top (s\hat{\Lambda}) x_i \right] && \text{(By definition of } \mathcal{F}) \\
&\overset{\text{(i)}}{=} \frac{1}{u} \mathbb{E}\left[\mathbb{1}\left\{ \sum_{i=1}^{n} \sigma_i x_i^\top \hat{\Lambda} x_i \geq 0 \right\} \sum_{i=1}^{n} \sigma_i x_i^\top \hat{\Lambda} x_i \right]. \\
&\leq \frac{1}{u} \mathbb{E}\left[\left| \sum_{i=1}^{u} \sigma_i x_i^\top \hat{\Lambda} x_i \right| \right] \\
&\leq \frac{1}{u} \sqrt{\mathbb{E}\left[\left(\sum_{i=1}^{u} \sigma_i x_i^\top \hat{\Lambda} x_i \right)^2 \right]} && \text{(By Jensen's inequality)}
\end{aligned}$$

$$= \frac{1}{u}\sqrt{\mathbb{E}\left[\sum_{i=1}^{u}\left(x_i^\top \hat{\Lambda} x_i\right)^2\right]} \qquad \text{(Since } \sigma_i\text{'s are i.i.d. and } \mathbb{E}[\sigma_i] = 0\text{)}$$

$$= \sqrt{\frac{\mathbb{E}_{X\sim\nu}\left[\left(X^\top \hat{\Lambda} X\right)^2\right]}{u}} \leq \frac{2\left\|\Lambda^*\right\|_F M^2}{\sqrt{u}}.$$

Here, the equality (i) holds because when $\sum_{i=1}^{n}\sigma_i x_i^\top \hat{\Lambda} x_i < 0$, the supremum over $s \in [0,1]$ will be obtained by taking $s = 0$; otherwise, it will be obtained by taking $s = 1$. □

Other Lemmas

The following lemma basically shows that $f(\Lambda)$ is linear in scalar multiplication.

Lemma D.3.7. *If $D_E(\hat{\mathbf{\Lambda}}) = \overline{D}_E(\hat{\Lambda})$, with $\hat{\Lambda} = \sum_{y\in\mathcal{Y}_\ell}\hat{\Lambda}_y$, then, for any $s \geq 0$, it holds that $D_E(s\cdot\hat{\mathbf{\Lambda}}) = \overline{D}_E(s\cdot\hat{\Lambda})$, where D_E and $\overline{D}_E$ are defined in Eq. (D.4).*

Proof. It suffices to show that if $\sum_{y\in\mathcal{Y}_\ell} y^\top \hat{\Lambda}_y y = f(\hat{\Lambda})$, then $\sum_{y\in\mathcal{Y}_\ell} y^\top (s\cdot\hat{\Lambda}_y) y = f(s\cdot\hat{\Lambda})$ for any $s > 0$. By definition, we have

$$\begin{aligned} f(s\cdot\hat{\Lambda}) = \max_{\Lambda_y} \quad & \textstyle\sum_{y\in\mathcal{Y}_\ell} y^\top \Lambda_y y \\ \text{subject to} \quad & \textstyle\sum_{y\in\mathcal{Y}_\ell}\Lambda_y = s\cdot\hat{\Lambda} \\ & \Lambda_y \succeq \mathbf{0}, \quad \forall y \in \mathcal{Y}_\ell. \end{aligned}$$

For the above optimization problem, we can do a change of variable by setting $\Lambda'_y = \frac{1}{s}\cdot\Lambda_y \implies \Lambda_y = s\cdot\Lambda'_y$. Then, we have

$$\begin{aligned} f(s\cdot\hat{\Lambda}) = \max_{\Lambda_y} \quad & \textstyle\sum_{y\in\mathcal{Y}_\ell} y^\top (s\cdot\Lambda'_y) y \\ \text{subject to} \quad & \textstyle\sum_{y\in\mathcal{Y}_\ell} s\cdot\Lambda'_y = s\cdot\hat{\Lambda} \\ & s\cdot\Lambda'_y \succeq \mathbf{0}, \quad \forall y \in \mathcal{Y}_\ell. \end{aligned}$$

$$\begin{aligned} \implies f(s\cdot\hat{\Lambda}) = \max_{\Lambda_y} \quad & s\textstyle\sum_{y\in\mathcal{Y}_\ell} y^\top \Lambda'_y y \\ \text{subject to} \quad & \textstyle\sum_{y\in\mathcal{Y}_\ell}\Lambda'_y = \hat{\Lambda} \\ & \Lambda'_y \succeq \mathbf{0}, \quad \forall y \in \mathcal{Y}_\ell. \end{aligned}$$

$$\implies f(s \cdot \hat{\Lambda}) = s \cdot f(\hat{\Lambda}) = s \cdot \sum_{y \in \mathcal{Y}_\ell} y^\top \Lambda_y y = \sum_{y \in \mathcal{Y}_\ell} y^\top (s \cdot \hat{\Lambda}_y) y.$$

Thus, the proof is complete. □

Lemma D.3.8. *Let $f : \mathbb{R}^d \mapsto \mathbb{R}$ be a concave function with maximizer x^* over the convex set $\mathcal{C}$. Further, assume that f is G-strongly concave with respect to ℓ_2 norm in region $\mathcal{S} \cap \mathcal{C}$, where $\mathcal{S} = \{x : \|x - x^*\|_2 \leq A\}$. If $f(x^*) - f(x) \leq \frac{AG}{2}$ and $c \in \mathcal{C}$, then $x \in \mathcal{S}$.*

Proof. By property of strong concavity, we know that, $f(x^*) - f(x) \geq \frac{G}{2} \|x - x^*\|_2$ for any $x \in \mathcal{S} \cap \mathcal{C}$. Now, suppose x' satisfies $f(x^*) - f(x') \leq \frac{AG}{2}$, $x' \in \mathcal{C}$ and $x' \notin \mathcal{S}$. Then, we must have $\|x' - x^*\|_2 > A$.

Let $\gamma \in (0, 1)$ be some number such that $z = \gamma x' + (1 - \gamma) x^*$ lies on the boundary of $\mathcal{S}$. By convexity, we also have $z \in \mathcal{C}$. Then, since f is concave, we have $f(z) \geq \gamma f(x') + (1 - \gamma) f(x^*) > f(x')$, where the second inequality is strict because f is strongly concave in a region around x^*. Since $f(x^*) - f(x') \leq \frac{AG}{2}$, f is G-strongly concave on $\mathcal{S}$ and z lies on the boundary of $\mathcal{S}$, we have

$$\frac{AG}{2} = \frac{G}{2} \|z - x^*\|_2 \leq f(x^*) - f(z) < f(x^*) - f(x') \leq \frac{AG}{2}.$$

This is a contradiction and thus we must have $x' \in \mathcal{S}$. □

The following lemma quantitatively describes how close $\widetilde{\Lambda}$ and Λ^* needs to be to ensure an at most ϵ multiplicative constraint violation.

Lemma D.3.9. *Assume $\|x\|_2 \leq M$ for any $x \in \mathcal{X}$. Let $\Sigma = \mathbb{E}_{X \sim \nu} \left[X X^\top \right] \succ \mathbf{0}$ and $\Lambda^* = \arg\max_{\Lambda \succeq \mathbf{0}} \overline{D}(\Lambda)$. Then, for any $\epsilon > 0$, if we have*

$$\left\| \widetilde{\Lambda} - \Lambda^* \right\|_F \leq \frac{8 \mu^2 \lambda_{\min}(\Sigma)}{3 M^2 \lambda_{\max}(\Sigma)} \cdot \frac{\epsilon}{1 + \epsilon},$$

then it holds that $y^\top \mathbb{E}_{X \sim \nu} \left[P_{\widetilde{\Lambda}}(X) X X^\top \right]^{-1} y \leq (1 + \epsilon) c_\ell^2$ for any $y \in \mathcal{Y}_\ell$.

Proof. Fix some $\epsilon > 0$. First, notice that if we regard P_Λ as a function of $q_\Lambda(x) = x^\top \Lambda x - 1$,

it then holds that

$$\|\nabla_\Lambda P_\Lambda(x)\|_2 = \left\|\frac{\mathrm{d}P_\Lambda}{\mathrm{d}q_\Lambda}\nabla_\Lambda q_\Lambda(x)\right\|_2 \le \left|\frac{\mathrm{d}P_\Lambda}{\mathrm{d}q_\Lambda}\right| \left\|xx^\top\right\|_2 \le \left|\frac{\mathrm{d}P_\Lambda}{\mathrm{d}q_\Lambda}\right| M^2 \le \frac{M^2}{8\mu},$$

where we obtain the last inequality by using Lemma D.3.11. Therefore, for any $x \in \mathcal{X}$ and $\widetilde{\Lambda} \succeq \mathbf{0}$, we have $\left|P_{\widetilde{\Lambda}}(x) - P_{\Lambda^*}(x)\right| \le \frac{M^2}{8\mu} \cdot \left\|\widetilde{\Lambda} - \Lambda^*\right\|_F$ by mean value theorem and Cauchy-Schwartz. inequality.

Therefore, if we have $\left\|\widetilde{\Lambda} - \Lambda^*\right\|_F \le \delta$, then

$$\left|P_{\widetilde{\Lambda}}(x) - P_{\Lambda^*}(x)\right| \le \frac{M^2\delta}{8\mu} \implies P_{\widetilde{\Lambda}}(x) \ge P_{\Lambda^*}(x) - \frac{M^2\delta}{8\mu}$$

$$\implies \mathbb{E}_{X\sim\nu}\left[P_{\widetilde{\Lambda}}(X)XX^\top\right] \succeq \mathbb{E}_{X\sim\nu}\left[P_{\Lambda^*}(X)XX^\top\right] - \frac{M^2\delta}{8\mu}\mathbb{E}_{X\sim\nu}\left[XX^\top\right].$$

By Lemma D.3.10, we know that

$$y^\top \mathbb{E}_{X\sim\nu}\left[P_{\widetilde{\Lambda}}(X)XX^\top\right]^{-1} y \le c_\ell^2(1+\epsilon) \iff \mathbb{E}_{X\sim\nu}\left[P_{\widetilde{\Lambda}}(X)XX^\top\right] \succeq \frac{yy^\top}{(1+\epsilon)c_\ell^2}. \tag{D.17}$$

Let $\Sigma^* = \mathbb{E}_{X\sim\nu}\left[P_{\Lambda^*}(X)XX^\top\right]$. Therefore, to guarantee the condition in Eq. (D.17), it is sufficient to guarantee that $\Sigma^* - \frac{M^2\delta}{8\mu}\Sigma \succeq \frac{yy^\top}{(1+\epsilon)c_\ell^2}$, which is equivalent to

$$w^\top\Sigma^* w - \frac{M^2\delta}{8\mu}w^\top\Sigma w \ge \frac{(w^\top y)^2}{c_\ell^2(1+\epsilon)}, \quad \forall \text{unit vector } w \in \mathbb{R}^d$$

$$\iff \frac{1}{w^\top\Sigma w}\cdot w^\top\left(\Sigma^* - \frac{yy^\top}{(1+\epsilon)c_\ell^2}\right)w \ge \frac{M^2\delta}{8\mu}, \quad \forall \text{unit vector } w \in \mathbb{R}^d.$$

Therefore, it is sufficient to choose δ such that

$$\frac{M^2\delta}{8\mu} \le \frac{1}{\lambda_{\max}(\Sigma)}\cdot\lambda_{\min}\left(\Sigma^* - \frac{yy^\top}{c_\ell^2(1+\epsilon)}\right) \le \min_{w:\|w\|_2=1}\frac{1}{w^\top\Sigma w}\cdot w^\top\left(\Sigma^* - \frac{yy^\top}{(1+\epsilon)c_\ell^2}\right)w.$$

Since P_{Λ^*} satisfies the constraint defined in problem (D.5), we have $\Sigma^* \succeq \frac{yy^\top}{c_\ell^2}$. Meanwhile, by Lemma D.3.11, we know that $P_{\Lambda^*}(x) \ge \frac{\mu}{3}$ for any $x \in \mathcal{X}$, which means that $\Sigma^* \succeq \frac{\mu}{3}\cdot\Sigma$.

That is, for any unit vector $w \in \mathbb{R}^d$, we have

$$w^\top \Sigma^* w \geq \frac{\left(w^\top y\right)^2}{c_\ell^2} \quad \text{and} \quad w^\top \Sigma^* w \geq \frac{\mu}{3}\lambda_{\min}\left(\Sigma\right),$$

which together implies $w^\top \Sigma^* w \geq \max\left\{\frac{\mu}{3}\cdot\lambda_{\min}(\Sigma), \frac{\left(w^\top y\right)^2}{c_\ell^2}\right\}$. Therefore, it holds that

$$\begin{aligned} w^\top \Sigma w - \frac{\left(w^\top y\right)^2}{(1+\epsilon)c_\ell^2} &\geq \max\left\{\frac{\mu}{3}\cdot\lambda_{\min}(\Sigma), \frac{\left(w^\top y\right)^2}{c_\ell^2}\right\} - \frac{\left(w^\top y\right)^2}{(1+\epsilon)c_\ell^2} \\ &= \max\left\{\frac{\mu}{3}\cdot\lambda_{\min}(\Sigma) - \frac{\left(w^\top y\right)^2}{(1+\epsilon)c_\ell^2}, \frac{\epsilon\left(w^\top y\right)^2}{(1+\epsilon)c_\ell^2}\right\} \\ &\geq \frac{\epsilon\mu}{3(1+\epsilon)}\cdot\lambda_{\min}\left(\Sigma\right) \\ \implies \lambda_{\min}\left(\Sigma^* - \frac{yy^\top}{c_\ell^2(1+\epsilon)}\right) &\geq \frac{\epsilon\mu}{3(1+\epsilon)}\cdot\lambda_{\min}\left(\Sigma\right). \end{aligned}$$

Therefore, to guarantee the condition in Eq. (D.17), it is sufficient to have

$$\frac{M^2\delta}{8\mu} = \frac{\epsilon\mu\lambda_{\min}(\Sigma)}{3(1+\epsilon)\lambda_{\max}(\Sigma)} \implies \mu = \frac{8\mu^2\lambda_{\min}(\Sigma)}{3M^2\lambda_{\max}(\Sigma)}\cdot\frac{\epsilon}{1+\epsilon},$$

Thus, the proof is complete. □

The following lemma is a result of standard Schur complement technique.

Lemma D.3.10. *If $\mathbb{E}_{X\sim\nu}\left[P(X)XX^\top\right]$ is invertible and $c_\ell > 0$, then*

$$y^\top \mathbb{E}_{X\sim\nu}\left[P(X)XX^\top\right]^{-1} y \leq c_\ell^2 \iff \mathbb{E}_{X\sim\nu}\left[P(X)XX^\top\right] \succeq \frac{yy^\top}{c_\ell^2}.$$

Proof. For simplicity, let $A = E_{X\sim\nu}\left[P(X)XX^\top\right] \succ \mathbf{0}$. Then, we consider the block matrix $\begin{bmatrix} A & y \\ y^\top & c_\ell^2 \end{bmatrix} \in \mathbb{R}^{(d+1)\times(d+1)}$. Let $\begin{bmatrix} u & a \end{bmatrix}^\top \in \mathbb{R}^{d+1}$ with $u \in \mathbb{R}^d$ be some vector.

Now, for one direction, suppose $y^\top A^{-1} y \le c_\ell^2$ holds. Consider

$$\begin{bmatrix} u & a \end{bmatrix} \begin{bmatrix} A & y \\ y^\top & c_\ell^2 \end{bmatrix} \begin{bmatrix} u \\ a \end{bmatrix} = u^\top A u + 2 a u^\top y + 2 c_\ell^2 a^2 := r(u, a).$$

If we minimize $r(u, a)$ over u, which means to treat a as fixed, we can get (by taking gradient and setting it to zero)

$$u^* = -a A^{-1} y \implies r(u^*, a) = a^2 (c_\ell^2 - y^\top A^{-1} y).$$

Since $y^\top A^{-1} y \le c_\ell^2$, we know that $r(u^*, a) \ge 0$, which means $r(u, a) \ge 0$ for any $\begin{bmatrix} u & a \end{bmatrix}^\top \in \mathbb{R}^{d+1}$.

Then, if we minimize $r(u, a)$ over a, we can get

$$a^* = -\frac{u^\top y}{c_\ell^2} \implies r(u, a^*) = u^\top A u - \frac{\left(u^\top y\right)^2}{c_\ell^2}.$$

Since $r(u, a) \ge 0$ for any $\begin{bmatrix} u & a \end{bmatrix}^\top \in \mathbb{R}^{d+1}$, we know that $u^\top A u - \frac{\left(u^\top y\right)^2}{c_\ell^2} \ge 0$ for any $u \in \mathbb{R}^d$. That is, we have $A \succeq \frac{y y^\top}{c_\ell^2}$.

The other direction simply takes the above calculation in a reversed way and thus the proof is complete. □

Properties of P_Λ

A visualization of P_Λ is given in Figure D.1.

Lemma D.3.11. *The function $P_\Lambda(x)$ defined in (5.7), if regarding as a function of $q_\Lambda(x) = x^\top \Lambda x - 1 \ge -1$, satisfies*

- $\lim_{q_\Lambda \to 0} P_\Lambda = \frac{1}{2}$ *for any* $\mu \in (0, 1)$
- *When* $q_\Lambda = -1$, $P_\Lambda = \frac{1}{2} + \mu - \frac{\sqrt{1+4\mu^2}}{2} \ge \frac{\mu}{3}$ *and* $P_\Lambda - \mu(\log(1 - P_\Lambda) + \log(P_\Lambda)) \le 2\sqrt{\mu}$ *for any* $\mu \in (0, 1)$.

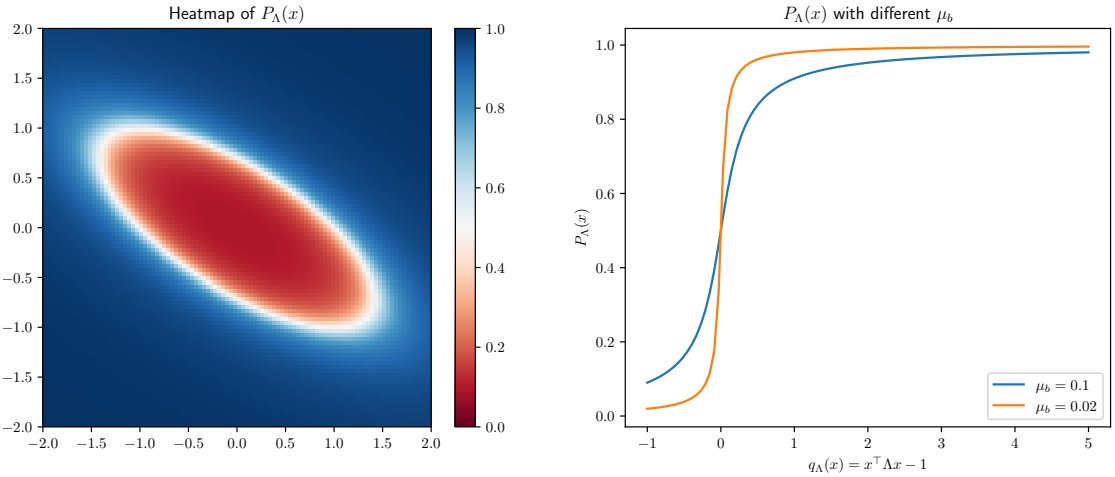


Figure D.1: (left) A heatmap of some P_Λ when problem dimension is $d = 2$, which shows that P_Λ is approximately an 0-1 threshold rule characterized by an ellipsoid. (right) A plot of P_Λ as a function of $q_\Lambda(x) = x^\top \Lambda x - 1$, which shows that the change of P_Λ near the boundary of ellipsoid is sharper when the barrier weight μ is smaller.

- $\frac{\mathrm{d}P_\Lambda}{\mathrm{d}q_\Lambda} = \frac{\mu\sqrt{q_\Lambda^2+4\mu^2}-2\mu^2}{q_\Lambda^2\sqrt{q_\Lambda^2+4\mu^2}}$ *decreases as* q_Λ^2 *increases. Further,* $\frac{\mathrm{d}P_\Lambda}{\mathrm{d}q_\Lambda} \in [0, \frac{1}{8\mu}]$. *Thus,* P_Λ *increases monotonically as* q_Λ *increases and* $P_\Lambda(x) \geq \frac{\mu}{3}$ *for any* $x \in \mathcal{X}$ *and* $\Lambda \succeq \mathbf{0}$.

- $\frac{\mathrm{d}P_\Lambda}{\mathrm{d}q_\Lambda}|_{q_\Lambda=\pm 1} \geq \frac{\mu}{10}$ *and* $\frac{\mathrm{d}P_\Lambda}{\mathrm{d}q_\Lambda} \geq \frac{\mu}{2q_\Lambda^2}$ *when* $q_\Lambda^2 \geq 12\mu^2$.

Proof. For simplicity, we will drop the subscript Λ and just treat P as a function of q. That is, we have

$$P(q) = \frac{1}{2} - \frac{\mu}{q} + \frac{\sqrt{(2\mu - q)^2 + 4\mu q}}{2q}.$$

We prove each bullet point separately.

- Since $P(q)$ also satisfies Eq. (5.6), which in simpler form is $\frac{\mu}{1-P(q)} - \frac{\mu}{P(q)} = q$, we can easily see that $P(q) = \frac{1}{2}$ satisfies this equation when $q = 0$.

- By direction computation, we can get $P(-1) = \frac{1}{2} + \mu - \frac{\sqrt{1+4\mu^2}}{2}$. To show this is greater than $\frac{\mu}{3}$ for any $\mu \in [0, 1]$, consider $\ell(\mu) = P(-1) - \frac{\mu}{3}$. It is easy to check that $\ell(0) = 0$ and $\ell(1) > 0$. Then, since $\ell'(\mu) = \frac{2}{3} - \frac{2\mu}{\sqrt{1+4\mu^4}}$ is initially greater than 0 and

then smaller than 0, we know $\ell(\mu)$ first increases and then decreases on $[0,1]$. Thus, $\ell(\mu) \geq 0$ on $[0,1]$ and thus $P(-1) \geq \frac{\mu}{3}$ for any $\mu \in [0,1]$.

For the second part, define $\tilde{\ell}(\mu) = 2\sqrt{\mu} - P(-1) + \mu\left(\log(1-P(-1)) + \log(P(-1))\right)$. Then, by utilizing the fact that P satisfies Eq. (5.6), we can compute its derivative and get $\frac{\mathrm{d}\tilde{\ell}}{\mathrm{d}\mu} = \frac{1}{\sqrt{\mu}} + \log(1-P(-1)) + \log(P(-1))$. We can check that on the domain $(0,1)$, we have $\frac{\mathrm{d}^2\tilde{\ell}}{\mathrm{d}\mu^2} = -\frac{1}{2\mu^{3/2}} + \frac{1}{\mu} - \frac{2}{\sqrt{1+4\mu^2}} \cdot \frac{2\sqrt{\mu(1+4\mu^2)} - 4\mu^{3/2} - \sqrt{1+4\mu^2}}{2\mu^{3/2}\sqrt{1+4\mu^2}} \leq 0$ on $(0,1)$, which means that $\frac{\mathrm{d}\tilde{\ell}}{\mathrm{d}\mu}$ is monotonically decreasing. To see why the second derivative is smaller than 0, we can compute

$$\left(4\mu^{3/2} + \sqrt{1+4\mu^2}\right) - 4\mu\left(1+4\mu^2\right) = (1-2\mu)^2 + 8\mu^{3/2}\sqrt{1+4\mu^2} \geq 0.$$

Thus, $\frac{\mathrm{d}\tilde{\ell}}{\mathrm{d}\mu}$ is initially greater than 0 and then smaller than 0 on $(0,1)$. It is easy to verify that $\lim_{\mu\to 0}\tilde{\ell} = 0$ and $\tilde{\ell}(1) > 0$. Therefore, we have $\tilde{\ell}(\mu) \geq 0$ for any $\mu \in (0,1)$.

- We can get $\frac{\mathrm{d}P}{\mathrm{d}q} = \frac{\mu\sqrt{q^2+4\mu^2} - 2\mu^2}{q^2\sqrt{q^2+4\mu^2}}$ by direct computation. To show it is decreasing as q^2 increasing, we consider $\tilde{f}(z) = \frac{\mu\sqrt{z+4\mu^2} - 2\mu^2}{z\sqrt{z+4\mu^2}}$ and it is sufficient to show that $\frac{\mathrm{d}\tilde{f}}{\mathrm{d}z} < 0$ for any $z > 0$. Again by direct computation, we have

$$\frac{\mathrm{d}\tilde{f}}{\mathrm{d}z} = \frac{\mu\left(8\mu^3 + 3\mu z - \left(z+4\mu^2\right)^{3/2}\right)}{z^2\left(z+4\mu^2\right)^{3/2}}.$$

By direct computation, We can show that $\left(z+4\mu^2\right)^3 - \left(8\mu^3 + 3\mu z\right)^2 = z^3 + 3z^2\mu^2 > 0$ for any $z > 0$ and $\mu \in [0,1]$. Thus, $\frac{\mathrm{d}\tilde{f}}{\mathrm{d}z} < 0$ and thus $\frac{\mathrm{d}P}{\mathrm{d}q}$ is decreasing as q^2 increases.

It is obvious that $\frac{\mathrm{d}P}{\mathrm{d}q} \geq 0$ for any $q^2 \geq 0$ and $\mu \in [0,1]$ since we always have $\mu\sqrt{q^2+4\mu^2} \geq 2\mu^2$. Thus, the maximum value could potentially happen is when $q^2 \to 0$, which can be evaluated by using L'Hospital's rule. A direct computation gives us $\lim_{q^2\to 0}\frac{\mathrm{d}P}{\mathrm{d}q} = \frac{1}{8\mu}$. Thus, we can conclude that $\frac{\mathrm{d}P}{\mathrm{d}q} \in \left[0, \frac{1}{8\mu}\right]$. Therefore, P increases monotonically as q increases, which implies that $P_\Lambda(x) \geq \frac{\mu}{3}$ for any $x \in \mathcal{X}$ and Λ.

- By direct computation, we have $\frac{\mathrm{d}P_\Lambda}{\mathrm{d}q_\Lambda}|_{q_\Lambda=\pm 1} = \mu\left(1 - \frac{2\mu}{\sqrt{1+4\mu^2}}\right) \geq \mu\left(1 - \frac{2}{\sqrt{5}}\right) \geq \frac{\mu}{10}$

for any $\mu \in [0,1]$. The reason is that we can easily see $\frac{2\mu}{\sqrt{1+4\mu^2}}$ is increasing in μ.

Finally, notice that when $2\mu \le \frac{1}{2}\sqrt{q^2+4\mu^2}$, which is equivalent to $q^2 \ge 12\mu^2$, we have

$$\frac{\mathrm{d}P}{\mathrm{d}q} = \frac{\mu\sqrt{q^2+4\mu^2}-2\mu^2}{q^2\sqrt{q^2+4\mu^2}} \ge \frac{\mu\sqrt{q^2+4\mu^2}-\frac{\mu}{2}\sqrt{q^2+4\mu^2}}{q^2\sqrt{q^2+4\mu^2}} = \frac{\mu}{2q^2}.$$

Thus, the proof is complete. □

D.3.3 An Alternative Approach to OptimizeDesign

Based on the analysis in Section D.3.1, we know that maximizing $\overline{D}(\cdot)$ is equivalent to maximizing $D(\cdot)$. Therefore, as an alternative to Algorithm 9, which maximizes $D(\cdot)$ through stochastic gradient ascent, it is natural to have an algorithm that directly maximizes $\overline{D}(\cdot)$. Here, we will consider subgradient ascent.

Recall that $\overline{D}: \mathbb{S}_+^d \mapsto \mathbb{R}$ is defined as

$$\overline{D}(\Lambda) = \mathbb{E}_{X\sim\nu}\left[P_\Lambda(X) - \mu\left(\log(1-P_\Lambda(X)) + \log(P_\Lambda(X))\right) - P_\Lambda(X)X^\top\Lambda X\right] + \frac{1}{c_\ell^2}\cdot f(\Lambda),$$

where $f(\Lambda)$ is defined in problem (D.3). The subgradient of $\overline{D}(\Lambda)$ is

$$\begin{aligned}\partial\overline{D}(\Lambda) &= \mathbb{E}_{X\sim\nu}\left[\left(1+\frac{\mu}{1-P_\Lambda(x)}-\frac{\mu}{P_\Lambda(X)}-X^\top\Lambda X\right)\nabla P_\Lambda(X) - P_\Lambda(X)XX^\top\right] + \frac{\partial f(\Lambda)}{c_\ell^2} \\ &\qquad\qquad \text{(The first term is differentiable)} \\ &= \frac{\partial f(\Lambda)}{c_\ell^2} - \mathbb{E}_{X\sim\nu}\left[P_\Lambda(X)XX^\top\right]. \qquad \text{(Since } P_\Lambda(X) \text{ solves Eq. (5.6))}\end{aligned}$$

Therefore, to run subgradient ascent, we only need to find an element in $\partial f(\Lambda)$, which can be obtained by solving the following optimization problem as claimed by Lemma D.3.13.

$$\begin{aligned} \min_\Gamma \quad & \langle \Gamma, \Lambda\rangle \\ \text{subject to} \quad & \Gamma \succeq yy^\top, \quad \forall y \in \mathcal{Y}_\ell, \\ & \Gamma \preceq 2\textstyle\sum_{y\in\mathcal{Y}_\ell} yy^\top. \end{aligned} \tag{D.18}$$

As a result, we have Algorithm 16 as an alternative to solve OptimizeDesign. Compared

to Algorithm 9, which needs to maintain $|\mathcal{Y}_\ell|\, d^2$ number of objective variables, Algorithm 16 only has d^2 variables. However, each iteration of Algorithm 16 is computationally more intensive since finding a subgradient needs to solve the problem (D.18).

Algorithm 16 Projected Stochastic Subgradient Ascent to Solve OptimizeDesign

Input: Number of iterations K; number of samples u; barrier weight $\mu_b \in (0,1)$
Initialize $\hat{\Lambda}^{(0)} = \mathbf{0}$
for $k = 0, 1, 2, \ldots, K-1$ **do**
  Sample $x_k \sim \nu$
  Solve problem (D.18) with current $\hat{\Lambda}^{(k)}$ to get $\Gamma^{(k)}$
  Set $g_k = \frac{\Gamma^{(k)}}{c_\ell^2} - P_{\hat{\Lambda}^{(k)}}(x_k) x_k x_k^\top$
  Set $\hat{\Lambda}^{(k+1)} \leftarrow \hat{\Lambda}^{(k)} + \eta_k g_k$, where $\eta_k = \frac{1}{\sqrt{2\sum_{s=1}^k \|g_s\|_2^2}}$
  Update $\hat{\Lambda}^{(k+1)} \leftarrow \Pi_{\mathbb{S}_+^d}(\hat{\Lambda}^{(k+1)})$, a projection to the set of $d \times d$ PSD matrices
end for
Let $\hat{\Lambda} = \frac{1}{K}\sum_{k=1}^K \hat{\Lambda}^{(k)}$
Find $s^* \leftarrow \arg\max_{s\in[0,1]} \overline{D}_E(s \cdot \hat{\Lambda})$, where $\overline{D}_E$ is the empirical version of $\overline{D}$, evaluated using u i.i.d. samples
return $\widetilde{\Lambda} = s^* \cdot \hat{\Lambda}$

A result similar to Theorem D.3.1 can also be obtained for Algorithm 16, which is given in Theorem D.3.12. The bounds are almost identical except that the old lower bound for K depends on $|\mathcal{Y}_\ell|^3$ while the new one depends on $|\mathcal{Y}_\ell|$. Steps identical to the proof of Theorem D.3.1 will be skipped in the proof of Theorem D.3.12.

Theorem D.3.12. *Let $\Lambda^* \in \arg\max_{\Lambda \succeq \mathbf{0}} \overline{D}(\Lambda)$ and take other settings the same as that in Theorem D.3.1.*

Then, Λ^* is unique. Further, for any $\epsilon > 0$ and $\delta > 0$, suppose it holds that

$$\mu \le \min\left\{\sqrt{\frac{3\kappa(\Sigma)\,\|\Lambda^*\|_F\, M^2}{8}\cdot\frac{1+\epsilon}{\epsilon}}, \frac{4}{9}\,\|\Lambda^*\|_F^2\, M^4, \frac{1}{2\sqrt{3}}\right\}$$

$$K \ge \frac{288\kappa(\Sigma)^2\,\|\Lambda^*\|_F^4\, M^4(M^4 + 4\,|\mathcal{Y}_\ell|\, C_\ell^2)\cdot\left(2\,\|\Lambda^*\|_F\, M^2 + 1\right)^4 \log(6/\delta)}{\omega^2\mu^6}\cdot\left(\frac{1+\epsilon}{\epsilon}\right)^2$$

$$u \ge \frac{576\kappa(\Sigma)^2\,\|\Lambda^*\|_F^2\, M^8\cdot\left(2\,\|\Lambda^*\|_F\, M^2 + 1\right)^4 \log(6/\delta)}{\omega^2\mu^6}\cdot\left(\frac{1+\epsilon}{\epsilon}\right)^2.$$

Then, with probability at least $1-\delta$, Algorithm 9 will output $\widetilde{\Lambda}$ that satisfies

- $y^\top \mathbb{E}_{X\sim\nu}\left[P_{\widetilde{\Lambda}}(X)XX^\top\right]^{-1} y \le (1+\epsilon)c_\ell^2, \quad \forall y \in \mathcal{Y}_\ell.$
- $\mathbb{E}_{X\sim\nu}\left[P_{\widetilde{\Lambda}}(X)\right] \le \mathbb{E}_{X\sim\nu}\left[\widetilde{P}(X)\right] + 4\sqrt{\mu}$, *where* $\widetilde{P}$ *is the optimal solution to problem* (D.13).

Proof. **First Bullet Point.** Similar to the proof of Theorem D.3.1, let $\hat{\Lambda}$ be the parameter obtained by Algorithm 16 just before the re-scaling step (line 11). Then, by Theorem 3.13 of Orabona [2019], with probability at least $1-\frac{\delta}{3}$, it holds that

$$\overline{D}(\Lambda^*) - \overline{D}(\hat{\Lambda}) \le \frac{\mathrm{Reg}(K) + 2\sqrt{2K\log(6/\delta)}}{K},$$

where $\mathrm{Reg}(K)$ is the regret of running projected stochastic subgradient ascent for K steps with η_k specified in Algorithm 16. Meanwhile, by Theorem 4.14 of Orabona [2019] also, we have $\mathrm{Reg}(K) = \sqrt{2}B^2\sqrt{\sum_{k=1}^K \|g_k\|_2^2}$, where $B = \|\Lambda^*\|_F$. Since $g_k = \frac{\Gamma^{(k)}}{c_\ell^2} - P_{\hat{\Lambda}^{(k)}}(x_k)x_k x_k^\top$ and $\left\|\Gamma^{(k)}\right\|_F \le 2\left\|\sum_{y\in\mathcal{Y}_\ell} yy^\top\right\|_F$, we can easily get $\|g_k\|_2^2 \le 2M^4 + \frac{8}{c_\ell^2}\sum_{y\in\mathcal{Y}_\ell}\|y\|_2^4 = 2M^4 + 8\left|\mathcal{Y}_\ell\right| C_\ell^2$. Thus, we have

$$\mathrm{Reg}(K) \le 2\|\Lambda^*\|_F^2\sqrt{M^4 + 4\left|\mathcal{Y}_\ell\right| C_\ell^2}\cdot\sqrt{K} := C_{\mathrm{Reg}}\sqrt{K} \tag{D.19}$$

$$\implies \overline{D}(\Lambda^*) - \overline{D}(\hat{\Lambda}) \le \frac{C_{\mathrm{Reg}} + 2\sqrt{2\log(6/\delta)}}{\sqrt{K}}, \tag{D.20}$$

We now consider the effect of using u i.i.d. samples in the re-scaling step. Since re-scaling always increases the function value, we must have $\overline{D}_E(\hat{\Lambda}) \le \overline{D}_E(\widetilde{\Lambda})$.

Then, after **exactly the same** steps of analysis, we can get the following same lower bound for K,

$$K \ge \left(\frac{3\kappa(\Sigma)M^2\left(C_{\mathrm{Reg}} + 2\sqrt{2\log(6/\delta)}\right)}{2G\mu^2}\cdot\frac{1+\epsilon}{\epsilon}\right)^2, \tag{D.21}$$

with a different value of C_{Reg}.

Second Bullet Point. We then prove the upper bound for primal objective value $\mathbb{E}_{X\sim\nu}\left[P_{\widetilde{\Lambda}}(X)\right]$, which explains the reason why an extra re-scaling step is needed. Let $\hat{\mathbf{\Lambda}} = (\hat{\Lambda}_y)_{y\in\mathcal{Y}_\ell}$ be a set of PSD matrices that solves problem (D.3) with parameter $\hat{\Lambda}$ and

$\widetilde{\mathbf{\Lambda}} = s^* \cdot \hat{\mathbf{\Lambda}}$, where $s^* = \arg\max_{s\in[0,1]} \overline{D}_E(s \cdot \hat{\Lambda})$. Since the constraint in problem (D.3) requires $\sum_{y\in\mathcal{Y}_\ell} \hat{\Lambda}_y = \hat{\Lambda}$, we have $\sum_{y\in\mathcal{Y}_\ell} \widetilde{\Lambda}_y = \widetilde{\Lambda}$, which is the output of Algorithm 16.

Define $g(s) = D_E(s \cdot \widetilde{\mathbf{\Lambda}})$. By construction, we know that $g(s)$ is maximized at $s = 1$ because $\overline{D}_E(s \cdot \hat{\Lambda}) = D_E(s \cdot \hat{\mathbf{\Lambda}})$ for any $s \geq 0$ as shown in Lemma D.3.7, which means that $s^* = \arg\max_{s\in[0,1]} D_E(s \cdot \hat{\mathbf{\Lambda}})$. Therefore, we have $g'(1) \geq 0$, which in turn gives us

$$g'(1) = \frac{1}{c_\ell^2} \sum_{y\in\mathcal{Y}_\ell} y^\top \widetilde{\Lambda}_y y - \frac{1}{u} \sum_{i=1}^{u} P_{\widetilde{\Lambda}}(x_i) x_i^\top \widetilde{\Lambda} x_i \geq 0.$$

Then, after **exactly the same** steps of analysis, we can get $\mathbb{E}_{X\sim\nu}\left[P_{\widetilde{\Lambda}}(X)\right] \leq \mathbb{E}_{X\sim\nu}\left[\widetilde{P}(X)\right] + 4\sqrt{\mu}$, where $\widetilde{P}$ is the optimal solution of the problem (D.13). □

Technical Lemmas

Lemma D.3.13. *The optimal value of the optimization problem (D.18) with parameter $\Lambda \succeq \mathbf{0}$ is equal to $f(\Lambda)$. Further, let $\Gamma^*(\Lambda)$ be an optimal solution to (D.18). Then, it holds that $\Gamma^*(\Lambda) \in \partial f(\Lambda)$ and $\|\Gamma^*(\Lambda)\| \leq 2\left\|\sum_{y\in\mathcal{Y}_\ell} yy^\top\right\|_F$.*

Proof. Alternatively, we first consider the following optimization problem.

$$\begin{array}{rl} \max_{\Lambda_y,\Sigma} & \sum_{y\in\mathcal{Y}_\ell} y^\top (\Lambda_y - 2\Sigma)\, y \\ \text{subject to} & \Lambda = \sum_{y\in\mathcal{Y}_\ell} \Lambda_y - \Sigma, \\ & \Sigma \succeq \mathbf{0}, \Lambda_y \succeq \mathbf{0}, \quad \forall y \in \mathcal{Y}_\ell. \end{array} \tag{D.22}$$

Since $y^\top \Sigma y \geq 0$ for any $y \in \mathcal{Y}_\ell$ and $\Sigma \succeq \mathbf{0}$, it is clear that problem (D.22) has the same optimal value as problem (D.3). Then, let $\Gamma \in \mathbb{R}^{d\times d}$ be the dual variable for the equality constraint in problem (D.22). We can have its dual problem to be

$$\begin{aligned} & \min_{\Gamma} \max_{\substack{\Lambda_y\succeq\mathbf{0},\forall y\in\mathcal{Y}_\ell,\\ \Sigma\succeq\mathbf{0}}} \sum_{y\in\mathcal{Y}_\ell} \left\langle yy^\top, \Lambda_y - 2\Sigma \right\rangle + \left\langle \Gamma, \Lambda + \Sigma - \sum_{y\in\mathcal{Y}_\ell} \Lambda_y \right\rangle \\ \Longrightarrow & \min_{\Gamma} \max_{\substack{\Lambda_y\succeq\mathbf{0},\forall y\in\mathcal{Y}_\ell,\\ \Sigma\succeq\mathbf{0}}} \langle \Gamma, \Lambda \rangle + \left\langle \Sigma, \Gamma - 2\sum_{y\in\mathcal{Y}_\ell} yy^\top \right\rangle + \sum_{y\in\mathcal{Y}_\ell} \left\langle \Lambda_y, yy^\top - \Gamma \right\rangle. \end{aligned}$$

In order for the above optimization problem to have finite value, we must have $\Gamma \preceq$

$2\sum_{y\in\mathcal{Y}_\ell} yy^\top$ and $\Gamma \succeq yy^\top$ for any $y \in \mathcal{Y}_\ell$. Therefore, we obtain the following dual problem.

$$\begin{aligned} \min_{\Gamma} \quad & \langle \Gamma, \Lambda \rangle \\ \text{subject to} \quad & \Gamma \succeq yy^\top, \quad \forall y \in \mathcal{Y}_\ell, \\ & \Gamma \preceq 2\sum_{y\in\mathcal{Y}_\ell} yy^\top. \end{aligned} \cdot$$

This is exactly the problem (D.18). Then, we can notice the Slater's condition is clearly satisfied by problem (D.18), which means the strong duality holds. Therefore, problem (D.18) has the same optimal value as (D.22), which is the same as (D.3).

Since $f(\Lambda)$ is concave in Λ as shown in Lemma D.3.3, to show that $\Gamma^*(\Lambda) \in \partial f(\Lambda)$, consider arbitrary $\Lambda, \Lambda' \succeq \mathbf{0}$. Then, we have

$$f(\Lambda) + \left\langle \Gamma^*(\Lambda), \Lambda' - \Lambda \right\rangle = \langle \Gamma^*(\Lambda), \Lambda \rangle + \left\langle \Gamma^*(\Lambda), \Lambda' - \Lambda \right\rangle = \left\langle \Gamma^*(\Lambda), \Lambda' \right\rangle \geq f(\Lambda').$$

The first equality holds because the optimal value of problem (D.18) is $f(\Lambda)$ as just shown above. The last inequality holds because $\Gamma^*(\Lambda)$ is a feasible solution to the problem (D.18) with parameter Λ'. Therefore, we have $\Gamma^*(\Lambda) \in \partial f(\Lambda)$.

Finally, since the constraint of problem (D.18) requires $\Gamma^*(\Lambda) \preceq 2\sum_{y\in\mathcal{Y}_\ell} yy^\top$, we can obtain $\|\Gamma^*(\Lambda)\|_F \leq 2\left\|\sum_{y\in\mathcal{Y}_\ell} yy^\top\right\|_F$ as a direct consequence of Lemma D.3.14. □

Lemma D.3.14. *For $A, B \in \mathbb{S}^{d\times d}$, if $A \succeq B \succeq \mathbf{0}$, then $\|A\|_F \geq \|B\|_F$.*

Proof. Let $\lambda_1, \ldots, \lambda_d$ and $\gamma_1, \ldots, \gamma_d$ be eigenvalues of A and B, respectively. Let $v_1, \ldots, v_d$ be a set of orthogonal unit eigenvectors of matrix A. Then, we have

$$\|A\|_F = \sqrt{\operatorname{tr}(AA)} = \sqrt{\operatorname{tr}\left(\left(\sum_{i=1}^d \lambda_i v_i v_i^\top\right)\left(\sum_{i=1}^d \lambda_i v_i v_i^\top\right)\right)} = \sqrt{\sum_{i=1}^d \lambda_i^2}.$$

Similarly, we have $\|B\|_F = \sqrt{\sum_{i=1}^d \gamma_i^2}$. By Corollary 7.7.4 in Horn and Johnson [2012], since $A \succeq B \succeq \mathbf{0}$, we know that $\lambda_i \geq \gamma_i \geq 0$ for each i. Therefore, we have $\|A\|_F \geq \|B\|_F$. □

D.4 Selective Sampling Algorithm for Unknown Distribution ν

D.4.1 Statement and proof of Theorem D.4.1

Consider now the case where we do not know ν exactly, and are returned $(\widehat{P}_\ell, \widehat{\Sigma}_{\widehat{P}_\ell})$ that only approximate their ideals. Algorithm 8 can still be employed to solve this case where ν is unknown, but at the cost of sampling some historical data. Note that compared to the case where ν is know, it assumes the knowledge of an upper bound on $\sup_{x\in\text{support}(\nu)} \|x\|$. It also relies on a multiplicative factor change in the constraint of the optimization problem, in order to account for the possible constraint violation of the output of the subroutine. The last difference is the use of an approximation of the covariance matrix to compute the estimator. The covariance matrix is empirically approximated by injecting additional unlabeled samples (historical data). With that, although we do not know ν but we can approximate the relevant quantities, such as the covariance matrix $\mathbb{E}_{X\sim\nu}[XX^\top]$.

Let us detail the properties of the implementation of $\widehat{P}_\ell, \widehat{\Sigma}_{\widehat{P}_\ell} \leftarrow$OptimizeDesign$(\mathcal{Z}_\ell, 2^{-\ell}, \tau)$ we use at each round ℓ.

First, $\widehat{P}_\ell$ has the properties described in Theorem 5.4.1 (by using Algorithm 9). More explicitly, let $\epsilon_\ell := 2^{-\ell}$, $B < \infty$ such that $\max_{x\in\mathcal{X}} |\langle x, \theta_*\rangle| \le B$, and $\sigma < \infty$ such that $\mathbb{E}[(y_s - \langle\theta_*, x_s\rangle)^2 | x_s] \le \sigma^2$. If

$$\beta_{\delta,\ell} := 4(1+\varepsilon)^2 \left(4\sqrt{B^2+\sigma^2}+1\right)^2 \log(4\ell^2|\mathcal{Z}|^2/\delta)$$

then $\widehat{P}_\ell$ is such that

- $\max_{z,z'\in\mathcal{Z}_\ell} \frac{\|z-z'\|^2_{\mathbb{E}_{X\sim\nu}[\tau\widehat{P}_\ell(X)XX^\top]^{-1}}}{\epsilon_\ell^2}\beta_{\delta,\ell} \le 1+\varepsilon.$
- $\mathbb{E}_{X\sim\nu}\left[\widehat{P}_\ell(X)\right] \le \mathbb{E}_{X\sim\nu}\left[\widetilde{P}_\ell(X)\right] + 4\sqrt{\mu_b}$, where $\widetilde{P}_\ell$ is the optimal solution to problem (D.23).

$$\begin{aligned} \min_P \quad & \mathbb{E}_{X\sim\nu}\left[P(X)\right] \\ \text{subject to} \quad & \max_{z,z'\in\mathcal{Z}_\ell} \frac{\|z-z'\|^2_{\mathbb{E}_{X\sim\nu}[\tau P(X)XX^\top]^{-1}}}{\epsilon_\ell^2}\beta_{\delta,\ell} \le 1, \\ & 0 \le P(x) \le 1-\mu_b, \quad \forall x\in\mathcal{X}. \end{aligned} \tag{D.23}$$

where $\mu_b \geq 0$. The quantity $\mathbb{E}_{X\sim\nu}\left[\widetilde{P}_\ell(X)\right]$ that uses $\mu_b > 0$ is easily related to the value when $\mu_b = 0$ through a simple scaling factor of $\frac{1}{1-\mu_b}$ (see proof below).

$\widehat{\Sigma}_{\widehat{P}_\ell}$ is the empirical covariance matrix of $\Sigma_{\widehat{P}_\ell} := \mathbb{E}_{X\sim\nu}[\widehat{P}_\ell(X)XX^\top]$ using historical data and is such that

$$(1-\gamma)\Sigma_{\widehat{P}_\ell} \preceq \widehat{\Sigma}_{\widehat{P}_\ell} \preceq (1+\gamma)\Sigma_{\widehat{P}_\ell}$$

where $\gamma \geq 0$.

Again, while we think of historical data as independent data collected offline before the start of the game, in practice this historical data could just come from previous rounds (which is not technically correct since its use may introduce some dependencies).

Theorem D.4.1 (Upper bound)**.** *Fix any $\delta \in (0,1)$. Let $\Delta = \min_{z\in\mathcal{Z}\setminus z_*}\langle z_* - z, \theta_*\rangle$ and set*

$$\beta_\delta = 256(1+\varepsilon)^2\left(4\sqrt{B^2+\sigma^2}+1\right)^2\log(4\log_2^2(\tfrac{4}{\Delta})|\mathcal{Z}|^2/\delta).$$

For any $\tau \geq \rho(\nu)\beta_\delta$ there exists a δ-PAC selective sampling algorithm that collects $\mathcal{T}$ historical data before the start of the game, observes $\mathcal{U}$ unlabeled examples, and requests just $\mathcal{L}$ labels that satisfies

- $\mathcal{U} \leq \log_2(\frac{4}{\Delta})\tau$,
- $\mathcal{L} \leq \frac{1}{1-\mu_b}\min_{\lambda\in\triangle_\mathcal{X}}\rho(\lambda)\beta_\delta + \frac{5\tau}{1-\mu_b}\sqrt{\mu_b}$ *subject to* $\tau \geq \|\lambda/\nu\|_\infty\rho(\lambda)\,\beta_\delta$, *and*
- $\mathcal{T} \leq \log_2(\frac{4}{\Delta})(K+u+\kappa_\delta)$

with probability at least $1-\delta$.

Here, the sample complexity for estimating the covariance matrix is bounded by $\kappa_\delta = \lceil 2K_{\psi_2}^2(\sqrt{d\ln 9/c_1} + \sqrt{\frac{\log(2/\delta)}{c_1}})\max\{1, 20\|\theta_*\|_{\mathbb{E}_{X\sim\nu}[XX^\top]}\}\rceil$ (where the sub-gaussian norm $K_{\psi_2} = \max_{s,P}\|\sqrt{P(\widetilde{x}_s)}\Sigma_P^{-1/2}\widetilde{x}_s\|_{\psi_2}$), and the contributions from the optimization problem to compute $\{\widehat{P}_\ell\}_\ell$ are

$$K = \widetilde{O}\left(\frac{|\mathcal{Z}|^6\,\kappa(\Sigma)^2\,\|\Lambda^*\|_2^8\,M^{16}}{\omega^2\mu_b^6}\right)\cdot\left(\frac{1+\epsilon}{\epsilon}\right)^2, \quad u = \widetilde{O}\left(\frac{\kappa(\Sigma)^2\,\|\Lambda^*\|_2^6\,M^{16}}{\omega^2\mu_b^6}\right)\cdot\left(\frac{1+\epsilon}{\epsilon}\right)^2,$$

Naturally, we have a trade-off on the subroutine tolerance μ_b. In order to get a better solution of the optimization over the selection rule P (and thus get a smaller $\sum_{t=(\ell-1)\tau+1}^{\ell\tau} P(x_t)$ term), the subroutine needs more unlabeled samples. However, it suffices to take $\mu_b = \frac{1}{\tau^2}$ to make $\mathcal{U}$, and $\mathcal{L}$ roughly match those of the case when ν was known.

The proof of this theorem is established through several results, which we provide in Section D.4.2.

D.4.2 Lemmas for the correctness

We first state here the correctness of Algorithm 8 in the case where ν is unknown.

Lemma D.4.2. *With probability at least $1-\delta$ we have for all stages $\ell \in \mathbb{N}$, we have that $z_* \in \mathcal{Z}_\ell$ and $\max_{z\in\mathcal{Z}_\ell}\langle z_* - z, \theta_*\rangle \le 4\epsilon_\ell$.*

The proof of the correctness lemma is established though several lemmas. First we provide Lemma D.4.3 guaranteeing concentration of empirical covariance matrices, which is obtained by sampling κ additional measurements. Then we show in Proposition D.4.4 that the RIPS estimator does not suffer from using that empirical covariance matrix.

Lemma D.4.3. *For any $P : \mathcal{X} \to [0,1]$, let $\Sigma_P = \mathbb{E}_{X\sim\nu}[P(X)XX^\top]$, $\widehat{\Sigma}_P = \frac{1}{\kappa}\sum_{s=1}^{\kappa} P(\widetilde{x}_s)\widetilde{x}_s\widetilde{x}_s^\top$. Define $K_{\psi_2} = \max_s \|\sqrt{P(\widetilde{x}_i)}\Sigma_P^{-1/2}\widetilde{x}_s\|_{\psi_2}$. With probability at least $1 - 2\exp(-c_1 t^2/K_{\psi_2}^4)$ holds*

$$(1-c)x^\top\Sigma_P x \le x^\top\widehat{\Sigma}_P x \le (1+c)x^\top\Sigma_P x$$

where $c = \max\left\{\frac{C\sqrt{d}+t}{\sqrt{\kappa}}, \left(\frac{C\sqrt{d}+t}{\sqrt{\kappa}}\right)^2\right\}$, $C = K_{\psi_2}^2\sqrt{\ln 9/c_1}$ and c_1 is an absolute constant.

Consequently for $\kappa \ge c_\delta := K_{\psi_2}^2(\sqrt{d\ln 9/c_1} + \sqrt{\frac{\log(2/\delta)}{c_1}})$, holds with probability at least $1-\delta$

$$\left(1 - \frac{c_\delta}{\sqrt{\kappa}}\right)x^\top\Sigma_P x \le x^\top\widehat{\Sigma}_P x \le \left(1 + \frac{c_\delta}{\sqrt{\kappa}}\right)x^\top\Sigma_P x.$$

Proof. Let $A \in \mathbb{R}^{\kappa\times d}$ whose rows A_i are independent sub-gaussian isotropic random vectors in R^d and define $K_{\psi_2} = \max_i \|A_i\|_{\psi_2}$. We can apply Theorem 5.39 of Vershynin [2011] on

A to have that with probability at least $1 - 2\exp(-c_1 t^2/K_{\psi_2}^4)$ holds

$$1 - \frac{C\sqrt{d}+t}{\sqrt{\kappa}} \leq \sigma_{\min}(A) \leq \sigma_{\max}(A) \leq 1 + \frac{C\sqrt{d}+t}{\sqrt{\kappa}},$$

where $C = K_{\psi_2}^2\sqrt{\ln 9/c_1}$ and c_1 is an absolute constant.

With Lemma 5.36 of Vershynin [2011], this implies that with probability at least $1 - 2\exp(-c_0 t^2)$ holds

$$\|A^\top A - I\| \leq \max\left\{\frac{C\sqrt{d}+t}{\sqrt{\kappa}}, \left(\frac{C\sqrt{d}+t}{\sqrt{\kappa}}\right)^2\right\} =: c \tag{D.24}$$

Recall $\Sigma_P = \mathbb{E}_{X\sim\nu}[P(X)XX^\top]$, so $Y = \sqrt{P(X)}\Sigma_P^{-1/2}X$ satisfies $\mathbb{E}[YY^\top] = \mathbb{E}[\Sigma_P^{-1/2}P(X)XX^\top\Sigma_P^{-1/2}] = \Sigma_P^{-1/2}\Sigma_P\Sigma_P^{-1/2} = I$. So we can apply (D.24) to get $\|\Sigma_P^{-1/2}\widehat{\Sigma}_P\Sigma_P^{-1/2} - I\| \leq c$. Thus for any $y \in \mathbb{R}^d$,

$$1 - c \leq \frac{y^\top}{\|y\|}\Sigma_P^{-1/2}\widehat{\Sigma}_P\Sigma_P^{-1/2}\frac{y}{\|y\|} \leq 1 + c$$

so setting $y = \Sigma_P^{1/2}x$

$$(1-c)x^\top\Sigma_P x \leq x^\top\widehat{\Sigma}_P x \leq (1+c)x^\top\Sigma_P x.$$

Also, the sub-gaussian bound becomes $K_{\psi_2} = \max_i \|\sqrt{P(\widetilde{x}_i)}\Sigma_P^{-1/2}\widetilde{x}_i\|_{\psi_2}$. □

Proposition D.4.4 (RIPS guarantees on empirical covariance matrix)**.** *Let $x_1, \ldots, x_n$ and $\widetilde{x}_1, \ldots, \widetilde{x}_\kappa$ be drawn IID from a distribution ν. For $s = 1, \ldots, n$, assume that $|\langle\theta, x_s\rangle| \leq B$ and $\mathbb{E}[|\langle\theta, x_s\rangle - y_s|^2] \leq \sigma_{noise}^2$. For $s = 1, \ldots, \kappa$, assume that $\mathbb{E}[|\langle\theta, x_s\rangle - y_s|^2] \leq \sigma_{noise}^2$. Let $P \in [0,1]$ be arbitrary and let $Q_s(x_s) \sim$ Bernoulli(P) independently for all $s \in [n]$. Let $\Sigma_P = \mathbb{E}_{X\sim\nu}[P(X)XX^\top]$ and $\widehat{\Sigma}_P = \frac{1}{\kappa}\sum_{s=1}^{\kappa} P(\widetilde{x}_s)\widetilde{x}_s\widetilde{x}_s^\top$. Assume that Σ_P is invertible and that there exists $\gamma \geq 0$ such that $(1-\gamma)\Sigma_P \preceq \widehat{\Sigma}_P \preceq (1+\gamma)\Sigma_P$. For a given finite set $\mathcal{V} \subset \mathbb{R}^d$ define*

$$w_v = \text{Catoni}(\{\langle v, \widehat{\Sigma}_P^{-1}Q_s(x_s)x_s y_s\rangle\}_{s=1}^n),$$

If $\widehat{\theta} = \arg\min_\theta \max_v \frac{|w_v - \langle \theta, v\rangle|}{\|v\|_{\widehat{\Sigma}_P^{-1}}}$ *and* $n \geq 4\log(2|\mathcal{V}|/\delta)$*, then with probability at least* $1-\delta$*, it holds that*

$$|\langle v, \widehat{\theta} - \theta\rangle| \leq 4\left(\sqrt{\frac{B^2+\sigma^2}{(1-\gamma)^2}} + \sqrt{n\gamma}\|\theta_*\|_{\mathbb{E}_{X\sim\nu}[XX^\top]}\right)\|v\|_{\mathbb{E}_{X\sim\nu}[nP(X)XX^\top]^{-1}}\sqrt{\log(2\,|\mathcal{V}|\,/\delta)}$$

We first state an intermediate matrix lemma before the proof of Proposition D.4.4.

Lemma D.4.5. *Assume that* Σ_P *is invertible and that there exists* $\gamma \in [0, 1/2]$ *such that* $(1-\gamma)\Sigma_P \preceq \widehat{\Sigma}_P \preceq (1+\gamma)\Sigma_P$*. Then for any* $v \in \mathcal{V}$

$$\|v\|^2_{\widehat{\Sigma}_P^{-1}\Sigma_P\widehat{\Sigma}_P^{-1}} \leq \frac{1}{(1-\gamma)^2}\|v\|^2_{\Sigma_P^{-1}}.$$

and

$$\|v\|_{(I-\Sigma_P^{1/2}\widehat{\Sigma}_P^{-1}\Sigma_P^{1/2})^2} \leq \sqrt{1 - \frac{2}{1+\gamma} + \frac{1}{(1-\gamma)^2}}\|v\|_2 \leq \sqrt{10\gamma}\|v\|_2.$$

Proof. We know that taking the inverse of two ordered positive definite matrices will flip the order, so here

$$\frac{1}{(1+\gamma)}\Sigma_P^{-1} \preceq \widehat{\Sigma}_P^{-1} \preceq \frac{1}{(1-\gamma)}\Sigma_P^{-1}.$$

$(1-\gamma)\Sigma_P \preceq \widehat{\Sigma}_P$ implies that for all $u \in \mathbb{R}^d$ holds $u^\top\Sigma_P u \leq 1/(1-\gamma)u^\top\widehat{\Sigma}_P u$. So taking $u = \widehat{\Sigma}_P^{-1}v$, we get $v^\top\widehat{\Sigma}_P^{-1}\Sigma_P\widehat{\Sigma}_P^{-1}v \leq 1/(1-\gamma)v^\top\widehat{\Sigma}_P^{-1}v$. Conclusion

$$v^\top\widehat{\Sigma}_P^{-1}\Sigma_P\widehat{\Sigma}_P^{-1}v = \frac{1}{1-\gamma}v^\top\widehat{\Sigma}_P^{-1}v \leq \frac{1}{(1-\gamma)^2}v^\top\Sigma_P^{-1}v$$

hence the first result of Lemma D.4.5.

For the second one, we get

$$\begin{aligned}\|v\|^2_{\left(I-\Sigma_P^{1/2}\hat{\Sigma}_P^{-1}\Sigma_P^{1/2}\right)^2} &= v^\top\left(I - \Sigma_P^{1/2}\hat{\Sigma}_P^{-1}\Sigma_P^{1/2}\right)^2 v \\ &= \|v\|_2^2 - 2v^\top\Sigma_P^{1/2}\hat{\Sigma}_P^{-1}\Sigma_P^{1/2}v + v^\top\Sigma_P^{1/2}\hat{\Sigma}_P^{-1}\Sigma_P\hat{\Sigma}_P^{-1}\Sigma_P^{1/2}v\end{aligned}$$

$$
\begin{aligned}
&\overset{\text{(i)}}{\leq} \|v\|_2^2 - \frac{2}{1+\gamma}\|v\|_2^2 + \frac{1}{1-\gamma} v^\top \Sigma_P^{1/2} \hat{\Sigma}_P^{-1} \Sigma_P^{1/2} v \\
&\leq \|v\|_2^2 - \frac{2}{1+\gamma}\|v\|_2^2 + \frac{1}{(1-\gamma)^2}\|v\|_2^2 \qquad (\text{Since } \hat{\Sigma}_P \preceq \tfrac{1}{1-\gamma}\Sigma_P) \\
&\leq \left(1 - \frac{2}{1+\gamma} + \frac{1}{(1-\gamma)^2}\right)\|v\|_2^2 \\
&\overset{\text{(ii)}}{\leq} 10\gamma \|v\|_2^2 .
\end{aligned}
$$

The inequality (i) above holds because $\frac{1}{1+\gamma}\Sigma_P^{-1} \preceq \hat{\Sigma}_P^{-1}$ and $(1-\gamma)\Sigma_P \preceq \hat{\Sigma}_P \implies \Sigma_P \preceq \frac{1}{1-\gamma}\hat{\Sigma}_P$. The inequality (ii) above holds because for $\gamma \in \left[0, \frac{1}{2}\right]$, we have

$$
1 - \frac{2}{1+\gamma} + \frac{1}{(1-\gamma)^2} \leq 1 - 2(1-\gamma) + (1+2\gamma)^2 \leq 10\gamma.
$$

Taking square root on both sides gives us the results. □

Proof of Proposition D.4.4. This proof is analogous to the proof of Proposition D.2.4. We first note that

$$
\begin{aligned}
\max_{v \in \mathcal{V}} \frac{|\langle \widehat{\theta}, v\rangle - \langle \theta, v\rangle|}{\|v\|_{\widehat{\Sigma}_P^{-1}}} &= \max_{v \in \mathcal{V}} \frac{|\langle \widehat{\theta}, v\rangle - w_v + w_v - \langle \theta, v\rangle|}{\|v\|_{\widehat{\Sigma}_P^{-1}}} \\
&\leq \max_{v \in \mathcal{V}} \frac{|\langle \widehat{\theta}, v\rangle - w_v|}{\|v\|_{\widehat{\Sigma}_P^{-1}}} + \max_{v \in \mathcal{V}} \frac{|w_v - \langle \theta, v\rangle|}{\|v\|_{\widehat{\Sigma}_P^{-1}}} \\
&= \min_{\theta'} \max_{v \in \mathcal{V}} \frac{|\langle \theta', v\rangle - w_v|}{\|v\|_{\widehat{\Sigma}_P^{-1}}} + \max_{v \in \mathcal{V}} \frac{|w_v - \langle \theta', v\rangle|}{\|v\|_{\widehat{\Sigma}_P^{-1}}} \\
&\leq 2 \max_{v \in \mathcal{V}} \frac{|\langle \theta, v\rangle - w_v|}{\|v\|_{\widehat{\Sigma}_P^{-1}}}
\end{aligned}
$$

So it suffices to show that each $|\langle \theta, v\rangle - w_v|$ is small. We begin by fixing some $v \in \mathcal{V}$ and bounding the variance of $v^\top \widehat{\Sigma}_P^{-1} Q_s(x_s) x_s y_s$ for any $s \leq n$ which is necessary to use the robust estimator. Note that

$$
\begin{aligned}
\mathbb{V}\mathrm{ar}_{x_s \sim \nu, Q_s(x_s) \sim P(x_s)}(v^\top \widehat{\Sigma}_P^{-1} Q_s(x_s) x_s y_s) =& \mathbb{E}_{x_s \sim \nu, Q_s(x_s) \sim P(x_s)}[(v^\top \widehat{\Sigma}_P^{-1} Q_s(x_s) x_s y_s)^2] \\
&- \mathbb{E}_{x_s \sim \nu, Q_s(x_s) \sim P(x_s)}[v^\top \widehat{\Sigma}_P^{-1} Q_s(x_s) x_s y_s]^2
\end{aligned}
$$

which means we can drop the second term to bound the variance by

$$
\begin{aligned}
&\mathbb{E}_{x_s\sim\nu, Q_s(x_s)\sim P(x_s)}[\Big((v^\top \widehat{\Sigma}_P^{-1} Q_s(x_s) x_s y_s\Big)^2] \\
&= \mathbb{E}_{x_s\sim\nu, Q_s(x_s)\sim P(x_s)}[\Big(v^\top \widehat{\Sigma}_P^{-1} Q_s(x_s) x_s (x_s^\top \theta + \xi_s)\Big)^2] \\
&= \mathbb{E}_{x_s\sim\nu}\left[\mathbb{E}_{Q_s(x_s)\sim P(s_s)}[\Big(v^\top \widehat{\Sigma}_P^{-1} Q_s(x_s) x_s (x_s^\top \theta)\Big)^2] + \mathbb{E}_{Q_s(x_s)\sim P(s_s)}[\Big(v^\top \widehat{\Sigma}_P^{-1} Q_s(x_s) x_s\Big)^2 \xi_t^2]\right] \\
&\leq \mathbb{E}_{x_s\sim\nu}\left[B^2 \mathbb{E}_{Q_s(x_s)\sim P(s_s)}[\Big(v^\top \widehat{\Sigma}_P^{-1} Q_s(x_s) x_s\Big)^2] + \sigma^2 \mathbb{E}_{Q_s(x_s)\sim P(s_s)}[\Big(v^\top \widehat{\Sigma}_P^{-1} Q_s(x_s) x_s\Big)^2]\right] \\
&= \mathbb{E}_{x_s\sim\nu}\left[(B^2+\sigma^2)\mathbb{E}_{Q_s(x_s)\sim P(s_s)}[v^\top \widehat{\Sigma}_P^{-1} Q_s(x_s) x_s x_s^\top Q_s(x_s) \widehat{\Sigma}_P^{-1} v]\right] \\
&= \mathbb{E}_{x_s\sim\nu}\left[(B^2+\sigma^2)\mathbb{E}_{Q_s(x_s)\sim P(s_s)}[v^\top \widehat{\Sigma}_P^{-1} Q_s(x_s) x_s x_s^\top \widehat{\Sigma}_P^{-1} v]\right] \\
&\leq \mathbb{E}_{x_s\sim\nu}\left[(B^2+\sigma^2) v^\top \widehat{\Sigma}_P^{-1} P(x_s) x_s x_s^\top \widehat{\Sigma}_P^{-1} v\right],
\end{aligned}
$$

where we used that $Q_s^2(x_s) = Q_s(x_s)$. Thus, we have with Lemma D.4.5

$$
\begin{aligned}
\mathbb{V}\mathrm{ar}(v^\top \widehat{\Sigma}_P^{-1} Q_s(x_s) x_s y_s) &\leq (B^2+\sigma^2) v^\top \widehat{\Sigma}_P^{-1} \mathbb{E}_{x_s\sim\nu}[P(x_s) x_s x_s^\top] \widehat{\Sigma}_P^{-1} v \\
&= (B^2+\sigma^2)\|v\|^2_{\widehat{\Sigma}_P^{-1}\Sigma_P\widehat{\Sigma}_P^{-1}} \\
&\leq \frac{B^2+\sigma^2}{(1-\gamma)^2}\|v\|^2_{\Sigma_P^{-1}}.
\end{aligned}
$$

We have

$$
\begin{aligned}
|\langle \theta_*, v\rangle - w_v| &= |\langle \theta_*, v\rangle - \mathbb{E}[v^\top \widehat{\Sigma}_P^{-1} P(x_1) x_1 y_1] + \mathbb{E}[v^\top \widehat{\Sigma}_P^{-1} P(x_1) x_1 y_1] - w_v| \\
&\leq |\langle \theta_*, v\rangle - \mathbb{E}[v^\top \widehat{\Sigma}_P^{-1} P(x_1) x_1 y_1]| \\
&\quad + |\mathrm{Catoni}(\{\langle v, \widehat{\Sigma}_P^{-1} Q_s(x_s) x_s y_s\rangle\}_{s=1}^n) - \mathbb{E}_{X\sim\nu}[v^\top \widehat{\Sigma}_P^{-1} P(X) X Y]|.
\end{aligned}
$$

We now recall that we can write $y_t = x_t^\top \theta_* + \xi_t$ where ξ_t is a mean-zero, independent random variable with variance at most σ^2. Thus, using Cauchy-Schwarz and applying Lemma D.4.5, we get

$$
\begin{aligned}
|\langle \theta_*, v\rangle - \mathbb{E}[v^\top \widehat{\Sigma}_P^{-1} P(x_1) x_1 y_1]| &= |v^\top \theta_* - v^\top \widehat{\Sigma}_P^{-1} \Sigma_P \theta_*| \\
&= |v^\top (I - \widehat{\Sigma}_P^{-1}\Sigma_P)\theta_*|
\end{aligned}
$$

$$
\begin{aligned}
&= |v^\top \Sigma_P^{-1/2}(I - \Sigma_P^{1/2}\widehat{\Sigma}_P^{-1}\Sigma_P^{1/2})\Sigma_P^{1/2}\theta_*| \\
&\leq \|\Sigma_P^{-1/2} v\| \, \|\Sigma_P^{1/2}\theta_*\|_{(I-\Sigma_P^{1/2}\widehat{\Sigma}_P^{-1}\Sigma_P^{1/2})^2} \\
&\leq \sqrt{10\gamma}\|\Sigma_P^{-1/2} v\| \, \|\Sigma_P^{1/2}\theta_*\| \\
&= \sqrt{10\gamma}\|v\|_{\Sigma_P^{-1}}\|\theta_*\|_{\Sigma_P}.
\end{aligned}
$$

By using the property of Catoni estimator stated in Definition D.2.3, we have

$$
\begin{aligned}
&|\langle\theta_*, v\rangle - w_v| \\
\leq &|\mathrm{Catoni}(\{\langle v, \mathbb{E}_{X\sim\nu}[P(X)XX^\top]^{-1}Q_s(x_s)x_s y_s\rangle\}_{s=1}^n) - \mathbb{E}[\langle v, \mathbb{E}_{X\sim\nu}[P(X)XX^\top]^{-1}Q_s(x_s)x_s y_s\rangle]| \\
&\quad + \sqrt{10\gamma}\|\theta_*\|_{\mathbb{E}_{X\sim\nu}[XX^\top]}\|v\|_{(\mathbb{E}_{X\sim\nu}[P(X)XX^\top]^{-1}} \\
\leq &\sqrt{2}\sqrt{(\mathbb{V}\mathrm{ar}(\langle v, \mathbb{E}_{X\sim\nu}[P(X)XX^\top]^{-1}Q_s(x_s)x_s y_s\rangle))\frac{\log(\frac{2}{\delta})}{n/2}} \\
&\quad + \sqrt{10\gamma}\|\theta_*\|_{\mathbb{E}_{X\sim\nu}[XX^\top]}\|v\|_{(\mathbb{E}_{X\sim\nu}[P(X)XX^\top]^{-1}} \\
&\qquad\qquad\qquad\qquad\qquad\qquad (\text{with probability at least } 1-\delta \text{ if } n \geq 4\log(2/\delta)) \\
\leq &\left(\sqrt{4}\sqrt{\frac{B^2+\sigma^2}{(1-\gamma)^2}} + \sqrt{10n\gamma}\|\theta_*\|_{\mathbb{E}_{X\sim\nu}[XX^\top]}\right)\|v\|_{(\mathbb{E}_{X\sim\nu}[P(X)XX^\top]^{-1}}\sqrt{\frac{\log(\frac{2}{\delta})}{n}} \\
= &\left(\sqrt{4}\sqrt{\frac{B^2+\sigma^2}{(1-\gamma)^2}} + \sqrt{10n\gamma}\|\theta_*\|_{\mathbb{E}_{X\sim\nu}[XX^\top]}\right)\|v\|_{\mathbb{E}_{X\sim\nu}[nP(X)XX^\top]^{-1}}\sqrt{\log(2/\delta)}.
\end{aligned}
$$

Finally, the proof is complete by taking union bounding over all $v \in \mathcal{V}$. □

Proof of Lemma D.4.2. Most of this proof is exactly the one of Section D.2.1 and Section D.2.1 so we only state the concentration bound. For any $\mathcal{V} \subseteq \mathcal{Z}$ and $z, z' \in \mathcal{V}$ define

$$
\mathcal{E}_{z,z',\ell}(\mathcal{V}) = \{|\langle z - z', \widehat{\theta}_\ell(\mathcal{V}) - \theta_*\rangle| \leq \epsilon_\ell\}
$$

where $\widehat{\theta}_\ell(\mathcal{V})$ is the estimator that would be constructed by the algorithm at stage ℓ with $\mathcal{Z}_\ell = \mathcal{V}$. Naturally we want to apply Proposition D.4.4 with τ labeled samples to obtain that $\mathcal{E}_{z,z',\ell}(\mathcal{V})$ holds with probability at least $1 - \frac{\delta}{2\ell^2|\mathcal{Z}|^2}$. Note that as Lemma D.3.11 gives

$P(x) \geq \mu/3$ so

$$\Sigma_P = \mathbb{E}_{X\sim\nu}[P(X)XX^\top] \geq \frac{\mu}{3}\mathbb{E}_{X\sim\nu}[XX^\top]$$

Σ_P is invertible.

Defining $\delta_0 := \frac{\delta}{4\ell^2|\mathcal{Z}|^2}$ and setting $\kappa \geq 2c_{\delta_0}\max\{1, 20\|\theta_*\|^2_{\mathbb{E}_{X\sim\nu}[XX^\top]}\}$ where we recall that was defined $c_\delta = K^2_{\psi_2}(\sqrt{d\ln 9/c_1} + \sqrt{\frac{\log(2/\delta)}{c_1}})$, Lemma D.4.3 leads to

$$\frac{c_{\delta_0}}{\kappa} \leq \frac{1}{2}\min\left\{1, \frac{1}{20\|\theta_*\|^2_{\mathbb{E}_{X\sim\nu}[XX^\top]}}\right\}$$

so that we can set $\gamma = c_{\delta_0}/(\tau\kappa)$ in the bound of Proposition D.4.4 to get

$$\sqrt{10\tau\gamma}\|\theta_*\|_{\mathbb{E}_{X\sim\nu}[XX^\top]} \leq \frac{1}{2}$$

and

$$\sqrt{\frac{B^2+\sigma^2}{(1-\gamma)^2}} \leq 2\sqrt{B^2+\sigma^2}$$

So for $\delta_0 = \frac{\delta}{4\ell^2|\mathcal{Z}|^2}$ the event $\widetilde{\mathcal{E}}_{\text{cov}}$ defined as

$$\widetilde{\mathcal{E}}_{\text{cov}} := \left\{\left(1 - \frac{c_{\delta_0}}{\sqrt{\kappa}}\right)x^\top\Sigma_P x \leq x^\top\widehat{\Sigma}_P x \leq \left(1 + \frac{c_{\delta_0}}{\sqrt{\kappa}}\right)x^\top\Sigma_P x\right\}.$$

happen with probability at least $1-\delta_0$.

Now, let us for now condition on $\widetilde{\mathcal{E}}_{\text{cov}}$. For fixed $\mathcal{V} \subset \mathcal{Z}$ and $\ell \in \mathbb{N}$ we apply Proposition D.4.4, instantiating the arbitrary P to $\widehat{P}_\ell$ (obtained with OptimizeDesign, recall Section D.4.1) so that with probability at least $1 - \frac{\delta}{4\ell^2|\mathcal{Z}|^2}$ we have that for any $z, z' \in \mathcal{V}$ holds that the event $\widetilde{\mathcal{E}}_{\text{RIPS},z,z'}$ defined as

$$\widetilde{\mathcal{E}}_{\text{RIPS},z,z'} := \Bigg\{|\langle z - z', \widehat{\theta}_\ell(\mathcal{V}) - \theta_*\rangle| \\ \leq 2\|z - z'\|_{\mathbb{E}_{X\sim\nu}[\tau\widehat{P}_\ell(X)XX^\top]^{-1}}\left(4\sqrt{B^2+\sigma^2}+1\right)\sqrt{\log(4\ell^2|\mathcal{Z}|^2/\delta)}\Bigg\}$$

happen with probability at least $1-\delta_0$.

So with probability at least $1-\mathbb{P}(\widetilde{\mathcal{E}}^c_{\text{RIPS},z,z'}) - \mathbb{P}(\widetilde{\mathcal{E}}^c_{\text{cov}}) \geq 1 - \frac{\delta}{4\ell^2|\mathcal{Z}|^2} - \frac{\delta}{4\ell^2|\mathcal{Z}|^2} = 1 - \frac{\delta}{2\ell^2|\mathcal{Z}|^2}$,

both events hold and we have that for any $z, z' \in \mathcal{V}$ holds

$$\begin{aligned}|\langle z - z', \widehat{\theta}_\ell(\mathcal{V}) - \theta_* \rangle| &\leq 2\|z - z'\|_{\mathbb{E}_{X\sim\nu}[\tau \widehat{P}_\ell(X) X X^\top]^{-1}} \left(4\sqrt{B^2 + \sigma^2} + 1\right) \sqrt{\log(4\ell^2|\mathcal{Z}|^2/\delta)} \\ &\leq 2(1+\varepsilon) \left(4\sqrt{B^2 + \sigma^2} + 1\right) \|z - z'\|_{\mathbb{E}_{X\sim\nu}[\tau \widehat{P}_\ell(X) X X^\top]^{-1}} \sqrt{\log(4\ell^2|\mathcal{Z}|^2/\delta)} \\ &\leq \epsilon_\ell.\end{aligned}$$

where we used the property of $\widehat{P}_\ell$ as detailed in Section D.4.1 to conclude. □

Proof of Theorem D.4.1. The total number of labels requested after L rounds is equal to $\sum_{\ell=1}^{L} \sum_{t=(\ell-1)\tau+1}^{\ell\tau} \widehat{P}_\ell(x_t)$. Again by Freedman's inequality we have that

$$\sum_{\ell=1}^{L} \sum_{t=(\ell-1)\tau+1}^{\ell\tau} \widehat{P}_\ell(x_t) \leq 2 \sum_{\ell=1}^{L} \tau \mathbb{E}_{X\sim\nu}[\widehat{P}_\ell(X)|\mathcal{Z}_\ell] + \log(1/\delta)$$

From Theorem 5.4.1, it holds for any ℓ that $\mathbb{E}_{X\sim\nu}[\widehat{P}_\ell(X)] \leq \mathbb{E}_{X\sim\nu}[\widetilde{P}_\ell(X)] + 4\sqrt{\mu}$ where $\widetilde{P}_\ell$ is the optimal solution to problem (D.13). So now, for some $\widetilde{\tau}$, we want to relate $\mathbb{E}_{X\sim\nu}[\widetilde{\tau}\widetilde{P}_\ell(X)]$ to $\mathbb{E}_{X\sim\nu}[\tau P_\ell(X)]$ where P_ℓ is the solution of problem (5.4). To do so, we rewrite problem (5.4) and problem (D.13) as

$$\begin{aligned} \min_P \quad & \mathbb{E}_{X\sim\nu}\left[\tau P(X)\right] \\ \text{subject to} \quad & y^\top \mathbb{E}_{X\sim\nu}\left[\tau P(X) X X^\top\right]^{-1} y \leq c_\ell^2, \quad \forall y \in \mathcal{Y}_\ell, \\ & 0 \leq \tau P(x) \leq \tau, \quad \forall x \in \mathcal{X}. \end{aligned} \tag{D.25}$$

and

$$\begin{aligned} \min_P \quad & \mathbb{E}_{X\sim\nu}\left[\widetilde{\tau} P(X)\right] \\ \text{subject to} \quad & y^\top \mathbb{E}_{X\sim\nu}[\widetilde{\tau} P(X) X X^\top]^{-1} y \leq c_\ell^2, \quad \forall y \in \mathcal{Y}_\ell, \\ & 0 \leq \widetilde{\tau} P(x) \leq \widetilde{\tau}(1 - \mu_b), \quad \forall x \in \mathcal{X}. \end{aligned} \tag{D.26}$$

where problem (D.25) is equivalent to problem (5.4) and problem (D.26) is equivalent to

problem (D.13). Thus taking $\widetilde{\tau} = \frac{\tau}{1-\mu_b}$, problem (D.26) becomes

$$\begin{aligned} \min_P \quad & \mathbb{E}_{X\sim\nu}\left[\tfrac{\tau}{1-\mu_b}P(X)\right] \\ \text{subject to} \quad & y^\top \mathbb{E}_{X\sim\nu}[\tfrac{\tau}{1-\mu_b}P(X)XX^\top]^{-1}y \le c_\ell^2, \quad \forall y \in \mathcal{Y}_\ell, \\ & 0 \le \tfrac{\tau}{1-\mu_b}P(x) \le \tau, \quad \forall x \in \mathcal{X}. \end{aligned}$$

which, using $Q = \frac{P}{1-\mu_b}$ is equivalent to

$$\begin{aligned} \min_Q \quad & \mathbb{E}_{X\sim\nu}\left[\tau Q(X)\right] \\ \text{subject to} \quad & y^\top \mathbb{E}_{X\sim\nu}[\tau Q(X)XX^\top]^{-1}y \le c_\ell^2, \quad \forall y \in \mathcal{Y}_\ell, \\ & 0 \le \tau Q(x) \le \tau, \quad \forall x \in \mathcal{X}. \end{aligned} \tag{D.27}$$

And we can now see that (D.27) and (D.25) are the same optimization problem. And Q_ℓ^* the solution of (D.27) is equal to $\frac{\widetilde{P}_\ell}{1-\mu_b}$. Thus the result $\mathbb{E}_{X\sim\nu}\left[\tilde{\tau}\widetilde{P}_\ell(X)\right] = \mathbb{E}_{X\sim\nu}\left[\tau P_\ell(X)\right]$.

Remains to bound $\sum_{\ell=1}^L \tau\mathbb{E}_{X\sim\nu}[P_\ell(X)]$ where

$$\begin{aligned} &\sum_{\ell=1}^L \tau\mathbb{E}_{X\sim\nu}[P_\ell(X)|\mathcal{Z}_\ell] \\ &= \sum_{\ell=1}^L \left[\min_{P:\mathcal{X}\to[0,1]} \tau\mathbb{E}_{X\sim\nu}[P(X)] \quad \text{subject to} \quad \max_{z,z'\in\mathcal{Z}_\ell} \frac{\|z-z'\|^2_{\mathbb{E}_{X\sim\nu}[\tau P(X)XX^\top]^{-1}}}{\epsilon_\ell^2}\beta_{\delta,\ell} \le 1\right], \end{aligned}$$

where $\beta_{\delta,\ell}$ is defined in Section D.4.1 as

$$\beta_{\delta,\ell} := 4(1+\varepsilon)^2\left(4\sqrt{B^2+\sigma^2}+1\right)^2\log(4\ell^2|\mathcal{Z}|^2/\delta).$$

As in the case where the distribution ν is known (Section D.2.1), we use Lemma D.2.5 to bound $\max_{z,z'\in\mathcal{Z}_\ell} \frac{\|z-z'\|^2_{\mathbb{E}_{X\sim\nu}[\tau P(X)XX^\top]^{-1}}}{\epsilon_\ell^2}\beta_{\delta,\ell}$ by $\max_{z\in\mathcal{Z}\setminus z_*} \frac{\|z-z_*\|^2_{\mathbb{E}_{X\sim\nu}[\tau P(X)XX^\top]^{-1}}}{\langle z-z_*,\theta_*\rangle^2}64\beta_{\delta,L}$. Last, the reparameterization of Proposition D.2.6 also applies here.

In the unlabeled sample complexity, we get an additional $L\kappa = L\lceil 2K_{\psi_2}^2(\sqrt{d\ln 9/c_1} + \sqrt{\frac{\log(2/\delta)}{c_1}})\max\{1, 20\|\theta_*\|_{\mathbb{E}_{X\sim\nu}[XX^\top]}\}\rceil$ term from the estimation of the covariance matrix.

Last, we get an additional $L(K+u)$, where K and u are such that

$$K \geq \widetilde{O}\left(\frac{|\mathcal{Z}|^3 \kappa(\Sigma)^2 \|\Lambda^*\|_2^8 M^{16}}{\beta^2 \mu_b^6}\right) \cdot \left(\frac{1+\epsilon}{\epsilon}\right)^2, \quad u \geq \widetilde{O}\left(\frac{\kappa(\Sigma)^2 \|\Lambda^*\|_2^6 M^{16}}{\beta^2 \mu_b^6}\right) \cdot \left(\frac{1+\epsilon}{\epsilon}\right)^2,$$

from the sample complexity of the subroutine. □

D.5 Classification

In this section we adopt the implementation described in Section D.2.1. As described in the text, given a distribution $\pi \in \Delta_{\mathcal{X}}$, and a class of hypothesis $\mathcal{H}$, we can reduce classification to linear bandits by setting $\theta^* = [\theta_x^*]_{x \in \Delta_{\mathcal{X}}}$ where $\theta_x^* = 2\eta(x) - 1$, and $\mathcal{Z} := \{z^{(h)}\}_{h \in \mathcal{H}} \subset [0,1]^{|\mathcal{X}|}$ where $z_x^{(h)} = \pi(x)\mathbf{1}\{h(x) = 1\}$. With the quantities computed in Section 5.3, we now prove Theorem 5.3.1.

Proof of Theorem 5.3.1. We consider a slightly modified version of Algorithm 8 where we stop at round L where $L_\epsilon = \lceil \log_2(4/\epsilon) \rceil$ and return $\arg\max_{z^{(h)} \in \mathcal{Z}_\ell} \langle z^{(h)}, \widehat{\theta}_\ell \rangle$. By an identical analysis to that in the proof of Theorem 2, we are guaranteed that $h \in \mathcal{S}_\ell$, i.e. $R_\nu(h) - R_\nu(z^*) = \langle z^* - z, \theta_* \rangle \leq 4\epsilon_\ell$. In addition the analysis of the sample complexity given there immediately gives the first part of the theorem.

It remains to bound the sample complexity in terms of the disagreement coefficient. The total sample complexity is given by,

$$\sum_{\ell=1}^{L} \left[\min_{P:\mathcal{X}\to[0,1]} \tau \mathbb{E}_{X\sim\nu}[P(X)] \quad \text{subject to} \quad \max_{z \in \mathcal{S}_\ell} \frac{\|z - z_*\|^2_{\mathbb{E}_{X\sim\nu}[\tau P(X) X X^\top]^{-1}}}{\epsilon_\ell^2} \beta_\delta \leq 1 \right]$$

where we recall $\beta_\delta = 2048 \log(2L^2|\mathcal{H}|/\delta)$ since we can take $B = 1$ and $\sigma = 1$.

We recall the proof of Theorem 2. From the proof, we see that with probability greater than $1 - \delta$, our sample complexity is obtained by summing up to round L

$$\sum_{\ell=1}^{L} \left[\min_{P:\mathcal{X}\to[0,1]} \tau \mathbb{E}_{X\sim\nu}[P(X)] \quad \text{subject to} \quad \max_{z \in \mathcal{S}_\ell} \frac{\|z - z_*\|^2_{\mathbb{E}_{X\sim\nu}[\tau P(X) X X^\top]^{-1}}}{\epsilon_\ell^2} \beta_\delta \leq 1 \right]$$

By proposition 2 this is equivalent to

$$\sum_{\ell=1}^{L}\left[\min_{\lambda\in\Delta_X}\rho_\ell(\lambda)\beta_\delta \quad \text{subject to} \quad \left\|\frac{\lambda}{\nu}\right\|_\infty \rho_\ell(\lambda)\beta_\delta \le \tau\right], \text{ where } \rho_\ell(\lambda) := \max_{z\in\mathcal{S}_\ell}\frac{\|z-z_*\|^2_{\mathbb{E}_{X\sim\lambda}[XX^\top]^{-1}}}{\epsilon_\ell^2}.$$

Define

$$A_\ell = \{x\in\mathcal{X} : \exists h, h(x)\neq h^*(x), R_\nu(h)-R_\nu(h^*)\le 4\epsilon_\ell\}, \ell\le L$$

and let $\lambda_\ell = \dfrac{\mathbf{1}\{x\in A_\ell\}\nu(x)}{\mathbb{E}[\mathbf{1}\{x\in A_\ell\}]}$, so $\left\|\dfrac{\lambda}{\nu}\right\|_\infty = \dfrac{1}{\mathbb{E}[\mathbf{1}\{x\in A_i\}]}$.

We first argue that λ_ℓ is feasible for the previous program. Note,

$$\begin{aligned}
\rho_\ell(\lambda_\ell) &= \max_{h:R_\nu(h)-R_\nu(h^*)\le 4\epsilon_\ell}\frac{\mathbb{E}_{X\sim\nu}[\frac{\mathbf{1}\{h(x)\neq h^*(x)}{\lambda_\ell(x)/\nu(x)}\}]}{\epsilon_\ell^2}\\
&\overset{\text{(i)}}{=} \mathbb{E}[\mathbf{1}\{x\in A_\ell\}]\max_{h:R_\nu(h)-R_\nu(h^*)\le 4\epsilon_\ell}\frac{\mathbb{E}_{X\sim\nu}[\mathbf{1}\{h(x)\neq h^*(x)\}]}{\epsilon_\ell^2}\\
&\le \mathbb{E}[\mathbf{1}\{x\in A_\ell\}]\max_{h:R_\nu(h)-R_\nu(h^*)\le 4\epsilon_\ell}\frac{16\mathbb{E}_{X\sim\nu}[\mathbf{1}\{h(x)\neq h^*(x)\}]}{\max\{\epsilon_\ell^2,(R_\nu(h)-R_\nu(h^*))^2\}}\\
&\le \mathbb{E}[\mathbf{1}\{x\in A_\ell\}]\max_{h:R_\nu(h)-R_\nu(h^*)\le 4\epsilon_\ell}\frac{16\mathbb{E}_{X\sim\nu}[\mathbf{1}\{h(x)\neq h^*(x)\}]}{\max\{(4\epsilon_\ell)^2,(R_\nu(h)-R_\nu(h^*))^2\}}\\
&\overset{\text{(ii)}}{\le} \mathbb{E}[\mathbf{1}\{x\in A_\ell\}]\max_{h:R_\nu(h)-R_\nu(h^*)\le 4\epsilon_\ell}\frac{16\mathbb{E}_{X\sim\nu}[\mathbf{1}\{h(x)\neq h^*(x)\}]}{\max\{\epsilon^2,(R_\nu(h)-R_\nu(h^*))^2\}}\\
&\le \mathbb{E}[\mathbf{1}\{x\in A_\ell\}]\max_{h\in H}\frac{16\mathbb{E}_{X\sim\nu}[\mathbf{1}\{h(x)\neq h^*(x)\}]}{\max\{\epsilon^2,(R_\nu(h)-R_\nu(h^*))^2\}}\\
&\le 16\mathbb{E}[\mathbf{1}\{x\in A_\ell\}]\rho(\nu,\epsilon)
\end{aligned}$$

where the equality (i) holds because the following is true when we only consider h such that $R_\nu(h)-R_\nu(h^*)\le 4\epsilon_\ell$

$$\frac{\mathbf{1}\{h(x)\neq h^*(x)\}}{\mathbf{1}\{x:\exists h, h(x)\neq h^*(x),(R_\nu(h)-R_\nu(h^*))\le 4\epsilon_\ell\}} = \mathbf{1}\{h(x)\neq h^*(x)\}.$$

The inequality (ii) above is true because $4\epsilon_\ell \ge \epsilon$. Thus we see that $\rho_\ell(\lambda_\ell)\|\lambda/\nu\|_\infty\beta_\delta \le 16\rho(\nu,\epsilon)\beta_\delta \le \tau$. It remains to argue about the disagreement coefficient. Firstly note that

for any h such that $R_\nu(h) - R_\nu(h^*) \le 4\epsilon_\ell$.

$$d_\nu(h, h^*) = \mathbb{E}_{X\sim\nu}[\mathbf{1}\{h(X) \neq h^*(X)\}] \le \mathbb{E}_{X\sim\nu}[\mathbf{1}\{h(X) \neq Y\}] + \mathbb{E}_{X\sim\nu}[\mathbf{1}\{h^*(X) \neq Y\}] \tag{D.28}$$

$$\le R_\nu(h) + R_\nu(h^*) \tag{D.29}$$

$$\le 2R_\nu(h^*) + 4\epsilon_\ell \tag{D.30}$$

Using this we see that,

$$\begin{aligned}
&\min_{\lambda\in\Delta} \rho_\ell(\lambda) \\
&\text{subject to } \rho_\ell(\lambda)\|\lambda/\nu\|_\infty\beta_\delta \le \tau \\
&\le \rho_\ell(\lambda_\ell)\beta_\delta && \text{(since } \lambda_\ell \text{ is feasible.)} \\
&\le \mathbb{E}[\mathbf{1}\{x \in A_\ell\}] \max_{h:R_\nu(h)-R_\nu(h^*)\le 4\epsilon_\ell} \frac{\mathbb{E}_{X\sim\nu}[\mathbf{1}\{h(x) \neq h^*(x)\}]}{\epsilon_\ell^2}\beta_\delta \\
& && \text{(imitating the above computation)} \\
&\le \frac{(2R(h^*) + 4\epsilon_\ell)\mathbb{E}_{X\sim\nu}[\mathbf{1}\{\exists h : h(X) \neq h^*(X), d_\nu(h, h^*) \le 2R(h^*) + 4\epsilon_\ell\}]}{\epsilon_\ell^2}\beta_\delta \\
& && \text{(Equation (D.28))} \\
&\le \beta_\delta \begin{cases} \frac{9R(h^*)^2}{\epsilon_\ell^2} \frac{\mathbb{E}_{X\sim\nu}[\mathbf{1}\{\exists h:h(X)\neq h^*(X),d_\nu(h,h^*)\le 2R(h^*)+4\epsilon_\ell\}]}{2R(h^*)+4\epsilon_\ell} & 4\epsilon_\ell \le R(h^*) \\ \frac{144\mathbb{E}_{X\sim\nu}[\mathbf{1}\{\exists h:h(X)\neq h^*(X),d_\nu(h,h^*)\le 2R(h^*)+4\epsilon_\ell\}]}{2R(h^*)+4\epsilon_\ell} & 4\epsilon_\ell > R(h^*) \end{cases} \\
&\le \left(\frac{9R(h^*)^2}{\epsilon_\ell^2} + 144\right) \frac{\mathbb{E}_{X\sim\nu}[\mathbf{1}\{\exists h : h(X) \neq h^*(X), d_\nu(h, h^*) \le 2R(h^*) + 4\epsilon_\ell\}]}{2R(h^*) + 4\epsilon_\ell}\beta_\delta
\end{aligned}$$

Thus,

$$\begin{aligned}
&\sum_{\ell=1}^{L} \left[\min_{\lambda\in\Delta_X} \rho_\ell(\lambda)\beta_\delta \quad \text{subject to} \quad \left\|\frac{\lambda}{\nu}\right\|_\infty \rho_\ell(\lambda)\beta_\delta \le \tau\right] \\
&\le \sum_{\ell=1}^{L} \rho_\ell(\lambda_\ell)\beta_\delta \\
&\le \sum_{\ell=1}^{L} \left(\frac{9R(h^*)^2}{\epsilon_\ell^2} + 144\right) \frac{\mathbb{E}_{X\sim\nu}[\mathbf{1}\{\exists h : h(X) \neq h^*(X), d_\nu(h, h^*) \le 2R(h^*) + 4\epsilon_\ell\}]}{2R(h^*) + 4\epsilon_\ell}\beta_\delta
\end{aligned}$$

$$\leq \log_2\left(\frac{4}{\epsilon}\right)\sup_{\ell\leq L}\left(\frac{9R(h^*)^2}{\epsilon_\ell^2}+144\right)\frac{\mathbb{E}_{X\sim\nu}[\mathbf{1}\{\exists h: h(X)\neq h^*(X), d_\nu(h,h^*)\leq 2R(h^*)+\epsilon_\ell\}]}{2R(h^*)+4\epsilon_\ell}\beta_\delta$$

$$\leq \log_2\left(\frac{4}{\epsilon}\right)\left(\frac{36R(h^*)^2}{\epsilon^2}+144\right)\sup_{\ell\leq L}\frac{\mathbb{E}_{X\sim\nu}[\mathbf{1}\{\exists h: h(X)\neq h^*(X), d_\nu(h,h^*)\leq 2R(h^*)+4\epsilon_\ell\}]}{2R(h^*)+4\epsilon_\ell}\beta_\delta$$

$$\leq 36\log_2\left(\frac{4}{\epsilon}\right)\left(\frac{R(h^*)^2}{\epsilon^2}+4\right)\sup_{\xi\geq\epsilon}\theta^*(2R(h^*)+\xi,\nu)\beta_\delta$$

from which the result follows.

□

Appendix E

OMITTED PROOFS AND EXPERIMENT DETAILS IN CHAPTER 6

E.1 Convergence Analysis of DAPO

The analysis starts by proving an approximate version of the Pythagorean theorem, which controls the error in three-point identity by the corresponding Bregman divergence and will serve as the key tool of our analysis.

E.1.1 Approximate Pythagorean Theorem

We begin with a general upper bound and then, we will derive its extensions under specific choices of mirror maps.

Lemma E.1.1 (Approximate Pythagorean Theorem)**.** *Let $\Phi : \mathcal{C} \mapsto \mathbb{R}$ be a proper closed convex mirror map, $\mathcal{D} \subseteq \mathcal{C}$ be a closed convex set and $v, c \in \mathcal{C}$ be two points. Suppose $u^\star = \arg\min_{u \in \mathcal{D}} D_\Phi(u, v)$. Then, for any $u \in \mathcal{D}$, we have*

$$D_\Phi(u, u^\star) + D_\Phi(u^\star, c) - D_\Phi(u, c) \leq \langle \nabla\Phi(v) - \nabla\Phi(c), u^\star - u \rangle .$$

Proof. Using the definition of Bregman divergence in Eq. (6.8), we have

$$\begin{aligned}
& D_\Phi(u, u^\star) + D_\Phi(u^\star, c) - D_\Phi(u, c) && \text{(E.1)} \\
&= \langle \nabla\Phi(u^\star), u - u^\star \rangle - \langle \nabla\Phi(c), u^\star - c \rangle + \langle \nabla\Phi(c), u - c \rangle \\
&= \langle \nabla\Phi(u^\star) - \nabla\Phi(c), u^\star - u \rangle && \text{(E.2)} \\
&= \underbrace{\langle \nabla\Phi(u^\star) - \nabla\Phi(v), u^\star - u \rangle}_{\leq 0 \text{ by Lemma 4.1 in Bubeck et al. [2012]}} + \langle \nabla\Phi(v) - \nabla\Phi(c), u^\star - u \rangle \\
&\leq \langle \nabla\Phi(v) - \nabla\Phi(c), u^\star - u \rangle .
\end{aligned}$$

□

Extension under Squared L_2-Norm

Lemma E.1.2. *Under the condition of Lemma E.1.1, if we take Φ to be the squared L_2-norm (see Example 6.3.4) and $\mathcal{D} = \Delta(\mathcal{A})$, then for any $u \in \mathcal{D}$, we have*

$$D_\Phi(u, u^\star) + D_\Phi(u^\star, c) - D_\Phi(u, c) \le \sqrt{2D_\Phi(v, c)}.$$

Proof. By Lemma E.1.1, we only need to bound $\langle \nabla\Phi(v) - \nabla\Phi(c), u^\star - u\rangle$. Then, since $\nabla\Phi(x) = x$, we have

$$\langle \nabla\Phi(v) - \nabla\Phi(c), u^\star - u\rangle \le \|v - c\|_2 \|u^\star - u\|_2 \le \sqrt{2D_\Phi(v, c)},$$

where $\|u^\star - u\|_2 \le 1$ since $u^\star, u \in \Delta(\mathcal{A})$. □

Extension under Negative Entropy

Lemma E.1.3. *Under the condition of Lemma E.1.1, if we take Φ to be the negative entropy restricted on $\Delta(\mathcal{A})$ (see Example 6.3.6) and assume $\mathcal{C} = \mathcal{D} = \Delta(\mathcal{A})$, then for any $u \in \mathcal{D}$, we have*

$$D_\Phi(u, u^\star) + D_\Phi(u^\star, c) - D_\Phi(u, c) \le \left(1 + \left\|\frac{u}{v}\right\|_\infty\right)\left(D_\Phi(v, c) + \sqrt{2D_\Phi(v, c)}\right).$$

Proof. By Lemma E.1.1, we only need to bound $\langle \nabla\Phi(v) - \nabla\Phi(c), u^\star - u\rangle$. Then, since $\mathcal{C} = \mathcal{D} = \Delta(\mathcal{A})$, we have $v = u^\star$. Therefore, we have

$$\begin{aligned}
\langle \nabla\Phi(v) - \nabla\Phi(c), u^\star - u\rangle &= \langle \nabla\Phi(v) - \nabla\Phi(c), v - u\rangle \\
&= \left\langle \log\frac{v}{c}, v - u\right\rangle \qquad \text{(Since } \Phi \text{ is the negative Shannon entropy)} \\
&= D_{\mathrm{KL}}(v\|c) - \left\langle \log\frac{v}{c}, u\right\rangle \\
&\le D_{\mathrm{KL}}(v\|c) + \left\|\frac{u}{v}\right\|_\infty \left\langle \left|\log\frac{v}{c}\right|, v\right\rangle \\
&\le D_{\mathrm{KL}}(v\|c) + \left\|\frac{u}{v}\right\|_\infty \left(D_{\mathrm{KL}}(v\|c) + \sqrt{2D_{\mathrm{KL}}(v\|c)}\right). \\
&\qquad \text{(By Lemma E.1.4)}
\end{aligned}$$

$$\leq \left(1 + \left\|\frac{u}{v}\right\|_\infty\right)\left(D_\Phi(v,c) + \sqrt{2D_\Phi(v,c)}\right)$$

□

Lemma E.1.4. *For any distributions $p, q \in \Delta(\mathcal{A})$ (or $\mathscr{P}(\mathcal{A})$ in continuous case) such that p is absolutely continuous with respect to q, we have $\left\langle \left|\log \frac{p}{q}\right|, p\right\rangle \leq D_{\mathrm{KL}}(p\|q) + \sqrt{2D_{\mathrm{KL}}(p\|q)}$.*

Proof. Without loss of generality, assume $\mathrm{supp}(p) = \mathcal{A}$. Now, we define $\mathcal{A}^+ = \{a \in \mathcal{A} \mid p_a \geq q_a\}$ and $\mathcal{A}^- = \{a \in \mathcal{A} \mid p_a < q_a\}$. Then, when p, q are discrete distributions, we have

$$\begin{aligned}
\left\langle \left|\log \frac{p}{q}\right|, p\right\rangle =& \sum_{a\in\mathcal{A}^+} p_a \log \frac{p_a}{q_a} + \sum_{a\in\mathcal{A}^-} p_a \log \frac{q_a}{p_a} \\
=& \sum_{a\in\mathcal{A}^+} p_a \log \frac{p_a}{q_a} - \sum_{a\in\mathcal{A}^-} p_a \log \frac{q_a}{p_a} + \sum_{a\in\mathcal{A}^-} p_a \log \frac{q_a}{p_a} + \sum_{a\in\mathcal{A}^-} p_a \log \frac{q_a}{p_a} \\
=& D_{\mathrm{KL}}(p\|q) + 2\sum_{a\in\mathcal{A}^-} p_a \log \frac{q_a}{p_a} \\
\leq& D_{\mathrm{KL}}(p\|q) + 2\sum_{a\in\mathcal{A}^-} p_a \left(\frac{q_a}{p_a} - 1\right) && \text{(Since } \log x \leq x - 1 \text{ for any } x > 0.\text{)} \\
=& D_{\mathrm{KL}}(p\|q) + 2\sum_{a\in\mathcal{A}^-} (q_a - p_a) \\
=& D_{\mathrm{KL}}(p\|q) + 2\|q - p\|_{\mathrm{TV}} && \text{(By definition of total variation distance.)} \\
\leq& D_{\mathrm{KL}}(p\|q) + \sqrt{2D_{\mathrm{KL}}(p\|q)}. && \text{(By Pinsker's inequality.)}
\end{aligned}$$

When p, q are continuous distributions, we only need to replace all summations to integrations and the proof is complete. □

E.1.2 Sublinear Convergence of DAPO

We first recall the Assumption (A1), (A1'), (A2) and (A3) listed in Section 6.5. Here, we prove Theorem 6.5.1 and Theorem 6.5.3 in a slightly more general version in which the training data distributions can be different from $d_\rho^{(k)}$ as long as they satisfy the assumptions. This slight extension makes our result applicable to the offline training setting where $\nu^{(k)} \in \Delta(\mathcal{S})$

is the reply buffer distribution at k-th iteration. Taking $\nu^{(k)} = d_\rho^{(k)}$ recovers the online training setting.

(A1) With initial distribution $\rho \in \Delta(\mathcal{S})$ and reply buffer distribution $\nu^{(k)} \in \Delta(\mathcal{S})$, there exist constants $\epsilon_{\text{critic}}, \epsilon_{\text{actor}} > 0$ such that for any k, it holds

$$\mathbb{E}_{s\sim\nu^{(k)}}\left[\left\|\widehat{Q}_s^{(k)} - Q_s^{(k)}\right\|_\infty\right] \le \epsilon_{\text{critic}},$$
$$\mathbb{E}_{s\sim\nu^{(k)}}\left[\frac{1}{2}\left\|f_s^{(k+1)} - \left(\pi_s^{(k)} + \eta_k \widehat{Q}_s^{(k)}\right)\right\|_2^2\right] \le \eta_k^2 \epsilon_{\text{actor}},$$

(A1') Under the same setting as (A1), we instead have

$$\mathbb{E}_{s\sim\nu^{(k)}}\left[D_{\text{KL}}\left(\pi_s^{(k+1)} \middle\| \pi_s^{(k)} \exp\left(-\eta_k \widehat{Q}_s^{(k)}\right)/Z_s^{(k)}\right)\right] \le \eta_k \epsilon_{\text{actor}},$$

(A2) With initial distribution ρ and replay buffer distribution $\nu^{(k)} \in \Delta(\mathcal{S})$, there exists constant $\vartheta_\rho \ge 1$ such that for any k, it holds

$$\max\left\{\left\|\frac{d_\rho^\star}{d_\rho^{(k+1)}}\right\|_\infty, \left\|\frac{d_\rho^{(k+1)}}{\nu^{(k)}}\right\|_\infty, \left\|\frac{d_{d_\rho^\star}^{(k+1)}}{\nu^{(k)}}\right\|_\infty, \left\|\frac{d_{d_\rho^\star}^{(k+1)}}{d_\rho^\star}\right\|_\infty\right\} \le \vartheta_\rho.$$

(A3) There exists constant $C_\rho > 0$ such that for any k, it holds

$$\max_{s\in\text{supp}(\nu^{(k)})}\left\{\left\|\frac{\pi_s^\star}{\pi_s^{(k+1)}}\right\|_\infty, \left\|\frac{\pi_s^{(k)}}{\pi_s^{(k+1)}}\right\|_\infty\right\} \le C_\rho.$$

Here, notice that conditions in Assumption (A1) and (A1') can be unifiedly written as

$$\mathbb{E}_{s\sim\nu^{(k)}}\left[D_{\Phi^*}\left(\nabla\Phi(\pi_s^{(k)}) - \eta_k \widehat{Q}_s^{(k)}, f_s^{(k+1)}\right)\right] \le \eta_k^{\omega^\Phi} \epsilon_{\text{actor}},$$

where we define ω^{Φ} as

$$\omega^{\Phi} = \begin{cases} 2, & \text{if } \Phi \text{ is the squared } L_2\text{-norm}, \\ 1, & \text{if } \Phi \text{ is the negative entropy on } \Delta(\mathcal{A}). \end{cases}$$

To present the proof in an unified way, we define $C_{\rho,s} = \max\left\{ \left\| \frac{\pi_s^\star}{\pi_s^{(k+1)}} \right\|_\infty, \left\| \frac{\pi_s^{(k)}}{\pi_s^{(k+1)}} \right\|_\infty \right\}$ for some state $s \in \mathcal{S}$ and define $\psi_s^{\Phi} : \mathbb{R}_+ \mapsto \mathbb{R}_+$ as

$$\psi_s^{\Phi}(x) = \begin{cases} \sqrt{2x}, & \text{if } \Phi \text{ is the squared } L_2\text{-norm}, \\ (1 + C_{\rho,s})\left(x + \sqrt{2x}\right), & \text{if } \Phi \text{ is the negative entropy on } \Delta(\mathcal{A}). \end{cases} \tag{E.3}$$

Then, applying Lemma E.1.2 and E.1.3 to Algorithm 10 gives us the following key lemma.

Lemma E.1.5. *Consider running Algorithm 10. Then, for policy $\pi = \pi^{(k)}$ or $\pi = \pi^\star$, for any $s \in \mathcal{S}$, if Φ is either squared L_2-norm or negative entropy on $\Delta(\mathcal{A})$, we have*

$$\begin{aligned} &\eta_k \left\langle \widehat{Q}_s^{(k)}, \pi_s^{(k+1)} - \pi_s \right\rangle + D_\Phi\left(\pi_s, \pi_s^{(k+1)}\right) + D_\Phi\left(\pi_s^{(k+1)}, \pi_s^{(k)}\right) - D_\Phi\left(\pi_s, \pi_s^{(k)}\right) \\ \leq &\psi_s^{\Phi}\left(D_{\Phi^*}\left(\nabla\Phi(\pi_s^{(k)}) - \eta_k \widehat{Q}_s^{(k)}, f_s^{(k+1)}\right)\right). \end{aligned}$$

Proof. Fix some $s \in \mathcal{S}$. Since line 5 of Algorithm 10 states that

$$\pi_s^{(k+1)} \in \operatorname*{arg\,min}_{\pi_s' \in \Delta(\mathcal{A})} D_\Phi\left(\pi_s', \nabla\Phi^*(f_s^{(k+1)})\right),$$

Then, We can apply Lemma E.1.2 or E.1.3 with $\mathcal{D} = \Delta(\mathcal{A})$, $u = \pi_s$, $u^\star = \pi_s^{(k+1)}$, $v = \nabla\Phi^*(f_s^{(k+1)})$ and $c = \nabla\Phi^*\left(\nabla\Phi(\pi_s^{(k)}) - \eta_k \widehat{Q}_s^{(k)}\right)$, which gives us

$$\begin{aligned} &D_\Phi\left(\pi_s^{(k+1)}, \nabla\Phi^*\left(\nabla\Phi(\pi_s^{(k)}) - \eta_k \widehat{Q}_s^{(k)}\right)\right) - D_\Phi\left(\pi_s, \nabla\Phi^*\left(\nabla\Phi(\pi_s^{(k)}) - \eta_k \widehat{Q}_s^{(k)}\right)\right) \\ &+ D_\Phi\left(\pi_s, \pi_s^{(k+1)}\right) \leq \psi_s^{\Phi}\left(D_\Phi\left(\nabla\Phi^*(f_s^{(k+1)}), \nabla\Phi^*\left(\nabla\Phi(\pi_s^{(k)}) - \eta_k \widehat{Q}_s^{(k)}\right)\right)\right). \end{aligned} \tag{E.4}$$

By using the duality of Bregman divergence in Eq. (6.21), we have

$$\psi_s^{\Phi}\left(D_{\Phi}\left(\nabla\Phi^*(f_s^{(k+1)}),\nabla\Phi^*\left(\nabla\Phi(\pi_s^{(k)})-\eta_k\widehat{Q}_s^{(k)}\right)\right)\right)=\psi_s^{\Phi}\left(D_{\Phi^*}\left(\nabla\Phi(\pi_s^{(k)})-\eta_k\widehat{Q}_s^{(k)},f_s^{(k+1)}\right)\right).$$

Meanwhile, by using the identity in Eq. (E.2), for the left-hand side of the Eq. (E.4), we have

$$\begin{aligned}
\text{LHS}=&\left\langle\nabla\Phi(\pi_s^{(k+1)})-\nabla\Phi\left(\nabla\Phi^*\left(\nabla\Phi(\pi_s^{(k)})-\eta_k\widehat{Q}_s^{(k)}\right)\right),\pi_s^{(k+1)}-\pi_s\right\rangle\\
=&\left\langle\nabla\Phi(\pi_s^{(k+1)})-\left(\nabla\Phi(\pi_s^{(k)})-\eta_k\widehat{Q}_s^{(k)}\right),\pi_s^{(k+1)}-\pi_s\right\rangle && \text{(By Lemma 6.3.2.)}\\
=&\eta_k\left\langle\widehat{Q}_s^{(k)},\pi_s^{(k+1)}-\pi_s\right\rangle+\left\langle\nabla\Phi(\pi_s^{(k+1)})-\nabla\Phi(\pi_s^{(k)}),\pi_s^{(k+1)}-\pi_s\right\rangle\\
=&\eta_k\left\langle\widehat{Q}_s^{(k)},\pi_s^{(k+1)}-\pi_s\right\rangle+D_{\Phi}\left(\pi_s,\pi_s^{(k+1)}\right)+D_{\Phi}\left(\pi_s^{(k+1)},\pi_s^{(k)}\right)-D_{\Phi}\left(\pi_s,\pi_s^{(k)}\right)\\
&\text{(By using the identity in Eq. (E.2) again on the second term above.)}
\end{aligned}$$

The proof is then complete by plugging this inequality back. □

Now, we can summarize both Theorem 6.5.1 and 6.5.3 into the following theorem and present its proof.

Theorem E.1.6. *Consider Algorithm 10 with initial policy $\pi^{(0)}$, initial distribution $\rho\in\Delta(\mathcal{S})$, arbitrary comparator policy $\pi^\star$ and Φ being the negative entropy restricted on $\Delta(\mathcal{A})$. Suppose Assumptions (A1), (A1'), (A2) and (A3) hold and we have constant step size $\eta_k=\eta$ for all $k\geq 0$. Then, for any comparator policy $\pi^\star$, it holds that*

$$\frac{1}{K}\sum_{k=0}^{K-1}\left(V_\rho^{(k)}-V_\rho^\star\right)\leq\frac{1}{K}\left(\frac{D_0^\star}{(1-\gamma)\eta}+\frac{V_{d_\rho^\star}^{(0)}}{1-\gamma}\right)+\frac{\vartheta_\rho\psi^\Phi(\epsilon_{\text{actor}})+(2-\gamma)\vartheta_\rho\epsilon_{\text{critic}}}{(1-\gamma)^2},$$

where $D_0^\star=\mathbb{E}_{s\sim d_\rho^\star}\left[D_\Phi(\pi_s^\star,\pi_s^{(0)})\right]$ and $\psi^\Phi:\mathbb{R}_+\mapsto\mathbb{R}_+$ is defined as

$$\psi^\Phi(x)=\begin{cases}\sqrt{2x}, & \text{if } \Phi \text{ is the } L_2\text{-norm square},\\ (1+C_\rho)\left(x+\sqrt{2x}\right), & \text{if } \Phi \text{ is the negative Shannon entropy.}\end{cases}\tag{E.5}$$

Proof. Fix some $s \in \mathcal{S}$ and $k < K$. First, by Lemma E.1.5 with $\pi_s = \pi_s^{(k)}$, we have

$$\eta_k \left\langle \widehat{Q}_s^{(k)}, \pi_s^{(k+1)} - \pi_s^{(k)} \right\rangle + D_\Phi \left(\pi_s^{(k)}, \pi_s^{(k+1)} \right) \leq \psi_s^\Phi \left(D_{\Phi^*} \left(\nabla \Phi(\pi_s^{(k)}) - \eta_k \widehat{Q}_s^{(k)}, f_s^{(k+1)} \right) \right),$$

where we dropped the term $D_\Phi \left(\pi_s^{(k+1)}, \pi_s^{(k)} \right)$ since it is always non-negative.

Then, since $D_\Phi \left(\pi_s^{(k)}, \pi_s^{(k+1)} \right) \geq 0$ as a Bregman divergence, we have

$$\Delta_s^{(k)} \stackrel{\text{def}}{=} \eta_k \left\langle \widehat{Q}_s^{(k)}, \pi_s^{(k+1)} - \pi_s^{(k)} \right\rangle - \psi_s^\Phi \left(D_{\Phi^*} \left(\nabla \Phi(\pi_s^{(k)}) - \eta_k \widehat{Q}_s^{(k)}, f_s^{(k+1)} \right) \right) \leq 0. \tag{E.6}$$

Then, by using Lemma E.1.5 with $\pi_s = \pi_s^\star$, the comparator policy, and similarly dropping $D_\Phi \left(\pi_s^{(k+1)}, \pi_s^{(k)} \right)$, we have

$$\begin{aligned} &\eta_k \left\langle \widehat{Q}_s^{(k)}, \pi_s^{(k+1)} - \pi_s^\star \right\rangle + D_\Phi \left(\pi_s^\star, \pi_s^{(k+1)} \right) - D_\Phi \left(\pi_s^\star, \pi_s^{(k)} \right) \\ \leq &\psi_s^\Phi \left(D_{\Phi^*} \left(\nabla \Phi(\pi_s^{(k)}) - \eta_k \widehat{Q}_s^{(k)}, f_s^{(k+1)} \right) \right). \end{aligned}$$

By adding and subtracting $\eta_k \left\langle \widehat{Q}_s^{(k)}, \pi_s^{(k)} \right\rangle$ together with some rearrangement, we have

$$\begin{aligned} &\eta_k \left\langle \widehat{Q}_s^{(k)}, \pi_s^{(k+1)} - \pi_s^{(k)} \right\rangle - \psi_s^\Phi \left(D_{\Phi^*} \left(\nabla \Phi(\pi_s^{(k)}) - \eta_k \widehat{Q}_s^{(k)}, f_s^{(k+1)} \right) \right) + \eta_k \left\langle \widehat{Q}_s^{(k)}, \pi_s^{(k)} - \pi_s^\star \right\rangle \\ \leq &D_\Phi \left(\pi_s^\star, \pi_s^{(k)} \right) - D_\Phi \left(\pi_s^\star, \pi_s^{(k+1)} \right). \end{aligned}$$

Taking expectation on both sides with respect to distribution $d_\rho^\star$, we have

$$\begin{aligned} &\mathbb{E}_{s \sim d_\rho^\star} \left[\eta_k \left\langle \widehat{Q}_s^{(k)}, \pi_s^{(k+1)} - \pi_s^{(k)} \right\rangle - \psi_s^\Phi \left(D_{\Phi^*} \left(\nabla \Phi(\pi_s^{(k)}) - \eta_k \widehat{Q}_s^{(k)}, f_s^{(k+1)} \right) \right) \right] \\ &+ \eta_k \mathbb{E}_{s \sim d_\rho^\star} \left[\left\langle \widehat{Q}_s^{(k)}, \pi_s^{(k)} - \pi_s^\star \right\rangle \right] \leq D_k^\star - D_{k+1}^\star, \end{aligned} \tag{E.7}$$

where $D_k^\star = \mathbb{E}_{s \sim d_\rho^\star} \left[D_\Phi \left(\pi_s^\star, \pi_s^{(k)} \right) \right]$.

For the first expectation above, we have

$$\begin{aligned} &\mathbb{E}_{s \sim d_\rho^\star} \left[\Delta_s^{(k)} \right] \\ \stackrel{\text{(i)}}{\geq} &\frac{1}{1-\gamma} \mathbb{E}_{s \sim d_{d_\rho^\star}^{(k+1)}} \left[\Delta_s^{(k)} \right] \end{aligned}$$

$$
\begin{aligned}
\geq & \frac{\eta}{1-\gamma}\mathbb{E}_{s\sim d_{d_\rho^\star}^{(k+1)}}\left[\left\langle Q_s^{(k)}, \pi_s^{(k+1)}-\pi_s^{(k)}\right\rangle\right] - \frac{\eta\vartheta_\rho}{1-\gamma}\mathbb{E}_{s\sim\nu^{(k)}}\left[\left|\left\langle \widehat{Q}_s^{(k)}-Q_s^{(k)}, \pi_s^{(k+1)}-\pi_s^{(k)}\right\rangle\right|\right] \\
& - \frac{\vartheta_\rho}{1-\gamma}\mathbb{E}_{s\sim\nu^{(k)}}\left[\psi_s^\Phi\left(D_{\Phi^*}\left(\nabla\Phi(\pi_s^{(k)})-\eta_k\widehat{Q}_s^{(k)}, f_s^{(k+1)}\right)\right)\right] \quad \text{(By Assumption (A2))} \\
\geq & \frac{\eta}{1-\gamma}\mathbb{E}_{s\sim d_{d_\rho^\star}^{(k+1)}}\left[\left\langle Q_s^{(k)}, \pi_s^{(k+1)}-\pi_s^{(k)}\right\rangle\right] - \frac{\eta\vartheta_\rho}{1-\gamma}\mathbb{E}_{s\sim\nu^{(k)}}\left[\left|\left\langle \widehat{Q}_s^{(k)}-Q_s^{(k)}, \pi_s^{(k+1)}-\pi_s^{(k)}\right\rangle\right|\right] \\
& - \frac{\vartheta_\rho}{1-\gamma}\psi^\Phi\left(\mathbb{E}_{s\sim\nu^{(k)}}\left[D_{\Phi^*}\left(\nabla\Phi(\pi_s^{(k)})-\eta_k\widehat{Q}_s^{(k)}, f_s^{(k+1)}\right)\right]\right) \\
& \qquad \text{(By Assumption (A3), Jensen's inequality and concavity of } \psi_s^\Phi) \\
\overset{\text{(ii)}}{\geq} & \frac{\eta}{1-\gamma}\mathbb{E}_{s\sim d_{d_\rho^\star}^{(k+1)}}\left[\left\langle Q_s^{(k)}, \pi_s^{(k+1)}-\pi_s^{(k)}\right\rangle\right] - \frac{\eta\vartheta_\rho\epsilon_{\text{critic}}}{1-\gamma} - \frac{\vartheta_\rho}{1-\gamma}\psi^\Phi(\eta^{\omega^\Phi}\epsilon_{\text{actor}}) \\
& \qquad \text{(By Assumption (A1), (A1') and monotonicity of } \psi^\Phi) \\
= & \eta\left(V_{d_\rho^\star}^{(k+1)} - V_{d_\rho^\star}^{(k)}\right) - \frac{\eta\vartheta_\rho\epsilon_{\text{critic}}}{1-\gamma} - \frac{\vartheta_\rho\psi^\Phi(\eta^{\omega^\Phi}\epsilon_{\text{actor}})}{1-\gamma} \qquad \text{(By Lemma E.1.9.)}
\end{aligned}
$$

Here, the above inequality (i) holds because Eq. (E.6) holds and we have $d_{d_\rho^\star,s}^{(k+1)} \geq (1-\gamma)d_{\rho,s}^\star$ for any $s \in \mathcal{S}$ as introduced in Section 6.3. The above inequality (ii) holds because by Hölder's inequality, we have

$$
\left\langle \widehat{Q}_s^{(k)}-Q_s^{(k)}, \pi_s^{(k+1)}-\pi_s^{(k)}\right\rangle \leq \left\|\widehat{Q}_s^{(k)}-Q_s^{(k)}\right\|_\infty \left\|\pi_s^{(k+1)}-\pi_s^{(k)}\right\|_1 \leq \left\|\widehat{Q}_s^{(k)}-Q_s^{(k)}\right\|_\infty.
$$

Then, for the second expectation in Eq. (E.7), we can similarly apply Lemma E.1.9 and obtain

$$
\begin{aligned}
\eta\mathbb{E}_{s\sim d_\rho^\star}\left[\left\langle \widehat{Q}_s^{(k)}, \pi_s^{(k)}-\pi_s^\star\right\rangle\right] = & \eta\mathbb{E}_{s\sim d_\rho^\star}\left[\left\langle Q_s^{(k)}, \pi_s^{(k)}-\pi_s^\star\right\rangle\right] + \eta\mathbb{E}_{s\sim d_\rho^\star}\left[\left\langle \widehat{Q}_s^{(k)}-Q_s^{(k)}, \pi_s^{(k)}-\pi_s^\star\right\rangle\right] \\
\geq & (1-\gamma)\eta\left(V_\rho^{(k)}-V_\rho^\star\right) - \eta\vartheta_\rho\epsilon_{\text{critic}}.
\end{aligned}
$$

By plugging these bounds back into Eq. (E.7), we then have

$$
(1-\gamma)\left(V_\rho^{(k)}-V_\rho^\star\right) \leq \frac{D_k^\star}{\eta} - \frac{D_{k+1}^\star}{\eta} + V_{d_\rho^\star}^{(k)} - V_{d_\rho^\star}^{(k+1)} + \frac{(2-\gamma)\vartheta_\rho\epsilon_{\text{critic}}}{1-\gamma} + \frac{\vartheta_\rho\psi^\Phi(\eta^{\omega^\Phi}\epsilon_{\text{actor}})}{(1-\gamma)\eta}
$$

Finally, by noticing that $\frac{\psi^\Phi(\eta^{\omega^\Phi}\epsilon_{\text{actor}})}{\eta} \leq \psi^\Phi(\epsilon_{\text{actor}})$ for either choice of Φ and taking sum

from $k = 0$ to $K - 1$, we can get

$$\frac{1}{K}\sum_{k=0}^{K-1}\left(V_\rho^{(k)} - V_\rho^\star\right) \le \frac{1}{K}\left(\frac{D_0^\star}{(1-\gamma)\eta} + \frac{V_{d_\rho^\star}^{(0)}}{1-\gamma}\right) + \frac{\vartheta_\rho \psi^\Phi(\epsilon_{\text{actor}}) + (2-\gamma)\vartheta_\rho \epsilon_{\text{critic}}}{(1-\gamma)^2}.$$

□

As a result, we can see that Theorem 6.5.1 and 6.5.3 are immediate consequences of Theorem E.1.6 by applying different choices of ψ^Φ for specific mirror map Φ.

E.1.3 Linear Convergence of DAPO

Similarly, we can summarize both Theorem 6.5.2 and 6.5.4 into the following theorem.

Theorem E.1.7. *Consider Algorithm 10 with initial policy $\pi^{(0)}$, initial distribution $\rho \in \Delta(\mathcal{S})$, arbitrary comparator policy $\pi^\star$ and Φ being either the squared L_2-norm or negative entropy on $\Delta(\mathcal{A})$. Let Assumption (A1), (A1'), (A2), (A3) hold and suppose the learning rates satisfy $\eta_0 \ge 1$ and $\eta_{k+1} \ge \frac{\vartheta_\rho}{\vartheta_\rho - 1}\eta_k$ for any $k \in [K]$. Then, it holds that*

$$V_\rho^{(K)} - V_\rho^\star \le \left(1 - \frac{1}{\vartheta_\rho}\right)^K \left(V_\rho^{(0)} - V_\rho^\star + \frac{D_0^\star/(\vartheta_\rho - 1)}{(1-\gamma)\eta_0}\right) + \frac{\vartheta_\rho^2 \psi^\Phi(\epsilon_{\text{actor}}) + 2\vartheta_\rho^2 \epsilon_{\text{critic}}}{1-\gamma},$$

where $D_0^\star = \mathbb{E}_{s\sim d_\rho^\star}\left[D_\Phi(\pi_s^\star, \pi_s^{(0)})\right]$ and $\psi^\Phi : \mathbb{R}_+ \mapsto \mathbb{R}_+$ is defined in Eq. (E.5).

Proof. Step 1. We can start from Eq. (E.7) in the proof of Theorem E.1.6. For the first expectation in Eq. (E.7), we have

$$\begin{aligned}
&\mathbb{E}_{s\sim d_\rho^\star}\left[\Delta_s^{(k)}\right] \\
\ge &\left\|\frac{d_\rho^\star}{d_\rho^{(k+1)}}\right\|_\infty^{-1} \mathbb{E}_{s\sim d_\rho^{(k+1)}}\left[\Delta_s^{(k)}\right] && \text{(By Eq. (E.6).)} \\
\ge &\vartheta_\rho \mathbb{E}_{s\sim d_\rho^{(k+1)}}\left[\Delta_s^{(k)}\right] && \text{(By Assumption (A2).)} \\
= &\eta_k \vartheta_\rho \mathbb{E}_{s\sim d_\rho^{(k+1)}}\left[\left\langle \widehat{Q}_s^{(k)}, \pi_s^{(k+1)} - \pi_s^{(k)}\right\rangle\right] - \vartheta_\rho^2 \mathbb{E}_{s\sim\nu^{(k)}}\left[\psi_s^\Phi\left(D_{\Phi^*}\left(\nabla\Phi(\pi_s^{(k)}) - \eta_k \widehat{Q}_s^{(k)}, f_s^{(k+1)}\right)\right)\right] \\
& && \text{(By Assumption (A2).)}
\end{aligned}$$

$$= (1-\gamma)\,\eta_k\vartheta_\rho\left(V_\rho^{(k+1)} - V_\rho^{(k)}\right) - \eta_k\vartheta_\rho^2\epsilon_{\text{critic}} - \vartheta_\rho^2\psi^\Phi\left(\eta_k^{\omega^\Phi}\epsilon_{\text{actor}}\right).$$

(By Lemma E.1.9 and deriviation similar to Theorem E.1.6.)

Similarly, for the second expectation in Eq. (E.7), we have

$$\begin{aligned}\eta_k\mathbb{E}_{s\sim d_\rho^\star}\left[\left\langle \widehat{Q}_s^{(k)}, \pi_s^{(k)} - \pi_s^\star\right\rangle\right] =& \eta_k\mathbb{E}_{s\sim d_\rho^\star}\left[\left\langle Q_s^{(k)}, \pi_s^{(k)} - \pi_s^\star\right\rangle\right] + \eta_k\mathbb{E}_{s\sim d_\rho^\star}\left[\left\langle \widehat{Q}_s^{(k)} - Q_s^{(k)}, \pi_s^{(k)} - \pi_s^\star\right\rangle\right] \\ \geq& (1-\gamma)\eta_k\left(V_\rho^{(k)} - V_\rho^\star\right) - \eta_k\vartheta_\rho\epsilon_{\text{critic}}.\end{aligned}$$

By plugging the results above back into Eq. (E.7) and defining $\delta_k \stackrel{\text{def}}{=} V_\rho^{(k)} - V_\rho^\star$, we have

$$\vartheta_\rho\left(\delta_{k+1} - \delta_k\right) + \delta_k \leq \frac{1}{(1-\gamma)\eta_k}D_k^\star - \frac{1}{(1-\gamma)\eta_k}D_{k+1}^\star + \frac{\vartheta_\rho^2\psi^\Phi\left(\eta_k^{\omega^\Phi}\epsilon_{\text{actor}}\right)}{(1-\gamma)\eta_k} + \frac{2\vartheta_\rho^2\epsilon_{\text{critic}}}{1-\gamma}. \quad \text{(E.8)}$$

Step 2. Now, dividing both sides of Eq. (E.8) by ϑ_ρ together with some rearrangement, we can have

$$\delta_{k+1} + \frac{D_{k+1}^\star}{(1-\gamma)\eta_k\vartheta_\rho} \leq \left(1 - \frac{1}{\vartheta_\rho}\right)\left(\delta_k + \frac{D_k^\star}{(1-\gamma)\eta_k\left(\vartheta_\rho - 1\right)}\right) + \frac{\vartheta_\rho\psi^\Phi\left(\eta_k^{\omega^\Phi}\epsilon_{\text{actor}}\right)}{(1-\gamma)\eta_k} + \frac{2\vartheta_\rho^2\epsilon_{\text{critic}}}{(1-\gamma)\vartheta_\rho}.$$

Since the learning rates satisfy $\eta_{k+1}(\vartheta_\rho - 1) \geq \eta_k\vartheta_\rho$, we have

$$\begin{aligned}&\delta_{k+1} + \frac{D_{k+1}^\star}{(1-\gamma)\eta_{k+1}(\vartheta_\rho - 1)} \\ \leq& \left(1 - \frac{1}{\vartheta_\rho}\right)\left(\delta_k + \frac{D_k^\star}{(1-\gamma)\eta_k\left(\vartheta_\rho - 1\right)}\right) + \frac{\vartheta_\rho\psi^\Phi\left(\eta_k^{\omega^\Phi}\epsilon_{\text{actor}}\right)}{(1-\gamma)\eta_k} + \frac{2\vartheta_\rho^2\epsilon_{\text{critic}}}{(1-\gamma)\vartheta_\rho} \\ \leq& \left(1 - \frac{1}{\vartheta_\rho}\right)\left(\delta_k + \frac{D_k^\star}{(1-\gamma)\eta_k\left(\vartheta_\rho - 1\right)}\right) + \frac{\vartheta_\rho\psi^\Phi\left(\epsilon_{\text{actor}}\right)}{(1-\gamma)} + \frac{2\vartheta_\rho^2\epsilon_{\text{critic}}}{(1-\gamma)\vartheta_\rho},\end{aligned}$$

where the second inequality above holds because we can easily verify that $\frac{\psi^\Phi(\eta_k^{\omega^\Phi}\epsilon_{\text{actor}})}{\eta_k} \leq$

$\psi^{\Phi}(\epsilon_{\text{actor}})$ for either choice of Φ. Then, applying the above relation recursively, we have

$$\begin{aligned}\delta_K+\frac{D_K^\star}{(1-\gamma)\eta_K(\vartheta_\rho-1)} \leq \left(1-\frac{1}{\vartheta_\rho}\right)^K \left(\delta_0+\frac{D_0^\star}{(1-\gamma)\eta_0(\vartheta_\rho-1)}\right) \\ + \frac{\vartheta_\rho \psi^{\Phi}\left(\epsilon_{\text{actor}}\right)}{1-\gamma} \sum_{k=0}^{K-1}\left(1-\frac{1}{\vartheta_\rho}\right)^k + \frac{2\vartheta_\rho^2 \epsilon_{\text{critic}}}{(1-\gamma)\vartheta_\rho} \sum_{k=0}^{K-1}\left(1-\frac{1}{\vartheta_\rho}\right)^k.\end{aligned} \tag{E.9}$$

We can notice that

$$\sum_{k=0}^{K-1}\left(1-\frac{1}{\vartheta_\rho}\right)^k \leq \frac{1}{1-\left(1-\frac{1}{\vartheta_\rho}\right)} = \vartheta_\rho.$$

Therefore, dropping the term with $D_K^\star$ in Eq. (E.9), we can finally have

$$V_\rho^{(K)} - V_\rho^\star \leq \left(1-\frac{1}{\vartheta_\rho}\right)^K \left(V_\rho^{(0)} - V_\rho^\star + \frac{D_0^\star/(\vartheta_\rho-1)}{(1-\gamma)\eta_0}\right) + \frac{\vartheta_\rho^2 \psi^{\Phi}(\epsilon_{\text{actor}}) + 2\vartheta_\rho^2 \epsilon_{\text{critic}}}{1-\gamma}.$$

□

Then, Theorem 6.5.2 and 6.5.4 are immediate consequences of Theorem E.1.7 under different Φ.

E.1.4 Technical Lemmas

Lemma E.1.8. *If we take $\rho = \mathsf{Unif}(\mathcal{S})$, then for any k and any distribution $\mu \in \Delta(\mathcal{S})$, we have $\left\|\frac{\mu}{d_\rho^{(k)}}\right\|_\infty \leq \frac{|\mathcal{S}|}{1-\gamma}$.*

Proof. By definition of state-visitation distribution in Eq. (6.3), we can immediately get $d_{\rho,s}^{(k)} \geq (1-\gamma)\rho_s$ for any $s \in \mathcal{S}$ by truncating all terms with $t \geq 1$. Since $\rho = \mathsf{Unif}(\mathcal{S})$, we have $\rho_s = \frac{1}{|\mathcal{S}|}$ for any $s \in \mathcal{S}$. Thus, we have

$$\left\|\frac{\mu}{d_\rho^{(k)}}\right\|_\infty = \max_{s\in\mathcal{S}} \frac{\mu_s}{d_{\rho,s}^{(k)}} \leq \frac{1}{(1-\gamma)\rho_s} \leq \frac{|\mathcal{S}|}{1-\gamma}.$$

□

Lemma E.1.9 (Performance Difference Lemma)**.** *For any two policies $\pi, \tilde{\pi} : \mathcal{S} \mapsto \Delta(\mathcal{A})$ and*

initial distribution $\rho \in \Delta(\mathcal{S})$, it holds that

$$V_\rho^\pi - V_\rho^{\tilde{\pi}} = \frac{1}{1-\gamma}\mathbb{E}_{s\sim d_\rho^\pi}\left[\left\langle Q_s^{\tilde{\pi}}, \pi_s - \tilde{\pi}_s \right\rangle\right] = \frac{1}{1-\gamma}\mathbb{E}_{s\sim d_\rho^{\tilde{\pi}}}\left[\langle Q_s^\pi, \pi_s - \tilde{\pi}_s\rangle\right].$$

Proof. See Lemma 1 in Xiao [2022]. □

E.2 Convergence Analysis of SAC

In this section, we prove a sublinear convergence rate for SAC under general function approximation by using our framework. It essentially adopts our proof techniques in Theorem E.1.7 to an entropy-regularized objective.

We start by presenting a modified version of the performance difference lemma under entropy-regularized reinforcement learning.

Lemma 6.5.5 (Modified Performance Difference Lemma)**.** *For any two policies $\pi, \tilde{\pi} : \mathcal{S} \mapsto \Delta(\mathcal{A})$, initial distribution $\rho \in \Delta(\mathcal{S})$ and regularization strength $\tau > 0$, it holds that*

$$\begin{aligned} V_{\tau,\rho}^\pi - V_{\tau,\rho}^{\tilde{\pi}} &= \frac{1}{1-\gamma}\mathbb{E}_{s\sim d_\rho^\pi}\left[\left\langle Q_{\tau,s}^{\tilde{\pi}}, \pi_s - \tilde{\pi}_s \right\rangle + \tau D_{\mathrm{KL}}(\pi_s \| \tilde{\pi}_s)\right] \\ &= \frac{1}{1-\gamma}\mathbb{E}_{s\sim d_\rho^{\tilde{\pi}}}\left[\langle Q_{\tau,s}^\pi, \pi_s - \tilde{\pi}_s\rangle - \tau D_{\mathrm{KL}}(\tilde{\pi}_s \| \pi_s)\right]. \end{aligned}$$

Proof. By definition of the value function, we have

$$\begin{aligned} V_{\tau,\rho}^\pi - V_{\tau,\rho}^{\tilde{\pi}} =& \mathbb{E}_{a_t\sim\pi_{s_t}}\left[\sum_{t=0}^\infty \gamma^t \left(c(s_t,a_t) + \tau\log\pi(a_t \mid s_t)\right) \,\middle|\, s_0 \sim \rho\right] - V_{\tau,\rho}^{\tilde{\pi}} \\ =& \mathbb{E}_{a_t\sim\pi_{s_t}}\left[\sum_{t=0}^\infty \gamma^t \left(c(s_t,a_t) + \tau\log\pi(a_t \mid s_t) + \gamma V_\tau^{\tilde{\pi}}(s_{t+1}) - V_\tau^{\tilde{\pi}}(s_t)\right) \,\middle|\, s_0 \sim \rho\right] \\ =& \mathbb{E}_{a_t\sim\pi_{s_t}}\left[\sum_{t=0}^\infty \gamma^t \left(Q_\tau^{\tilde{\pi}}(s_t,a_t) - V_\tau^{\tilde{\pi}}(s_t) + \tau\log\frac{\pi(a_t \mid s_t)}{\tilde{\pi}(a_t \mid s_t)}\right) \,\middle|\, s_0 \sim \rho\right] \\ =& \mathbb{E}_{a_t\sim\pi_{s_t}}\left[\sum_{t=0}^\infty \gamma^t \left(A_\tau^{\tilde{\pi}}(s_t,a_t) + \tau\log\frac{\pi(a_t \mid s_t)}{\tilde{\pi}(a_t \mid s_t)}\right) \,\middle|\, s_0 \sim \rho\right] \\ & \qquad \text{(We define } A_\tau^\pi(s,a) = Q_\tau^\pi(s,a) - V_\tau^\pi(s).\text{)} \\ =& \frac{1}{1-\gamma}\mathbb{E}_{s\sim d_\rho^\pi}\left[\left\langle Q_{\tau,s}^{\tilde{\pi}}, \pi_s - \tilde{\pi}_s \right\rangle + \tau D_{\mathrm{KL}}(\pi_s \| \tilde{\pi}_s)\right]. \end{aligned}$$

Then, by similarly expanding the term $V_{\tau,\rho}^{\tilde{\pi}}$, we can get

$$V_{\tau,\rho}^{\pi} - V_{\tau,\rho}^{\tilde{\pi}} = \frac{1}{1-\gamma}\mathbb{E}_{s\sim d_\rho^{\tilde{\pi}}}\left[\left\langle Q_{\tau,s}^{\pi}, \pi_s - \tilde{\pi}_s\right\rangle - \tau D_{\mathrm{KL}}(\tilde{\pi}_s \| \pi_s)\right].$$

Thus, the proof is complete. □

Now, we start to prove Theorem 6.5.6. Here, for simplicity, we keep using the function ψ_s and ψ defined in Eq. (E.3) and (E.5). However, we ignore the superscript Φ since in this concrete example, we only take Φ to be the negative entropy.

Theorem 6.5.6 (Sublinear Convergence of SAC)**.** *Consider running Algorithm 10 for entropy-regularized reinforcement learning with initial policy $\pi^{(0)}$, regularization strength τ, initial distribution $\rho \in \Delta(\mathcal{S})$ and Φ being the negative entropy restricted on $\Delta(\mathcal{A})$. Let $\pi^\star$ be an arbitrary comparator policy. Suppose Assumptions (A1'), (A2) and (A3) hold and the step sizes satisfy $\eta_k = \eta \le \frac{1}{\tau\vartheta_\rho}$ for any k. Then, we have*

$$\frac{1}{K}\sum_{k=0}^{K-1}\left(V_{\tau,\rho}^{(k)} - V_{\tau,\rho}^{\star}\right) \le \frac{1}{K}\left(\frac{D_0^\star}{(1-\gamma)\eta} + \frac{V_{\tau,d_\rho^\star}^{(0)}}{1-\gamma}\right) + \frac{\vartheta_\rho\psi(\epsilon_{\mathrm{actor}}) + (2-\gamma)\vartheta_\rho\epsilon_{\mathrm{critic}}}{(1-\gamma)^2}.$$

where $D_0^\star = \mathbb{E}_{s\sim d_\rho^\star}\left[D_{\mathrm{KL}}\left(\pi_s^\star \,\middle\|\, \pi_s^{(0)}\right)\right]$ and $\psi(x) = (1+C_\rho)\left(x + \sqrt{2x}\right)$ for $x \ge 0$.

Proof. First, it is straightforward to check that Lemma E.1.5 still holds under entropy-regularized reinforcement learning. Then, fix some $s \in \mathcal{S}$ and $k < K$, similar to the proof of Theorem E.1.6, just like Eq. (E.6), we also have

$$\Delta_{\tau,s}^{(k)} \stackrel{\mathrm{def}}{=} \eta\left\langle \widehat{Q}_{\tau,s}^{(k)}, \pi_s^{(k+1)} - \pi_s^{(k)}\right\rangle - \psi_s\left(D_{\mathrm{KL}}\left(\pi_s^{(k+1)} \,\middle\|\, \pi_s^{(k)}\exp\left(-\eta\widehat{Q}_{\tau,s}^{(k)}\right)/Z_s^{(k)}\right)\right) \le 0. \tag{E.10}$$

Then, by using Lemma E.1.5 with $\pi_s = \pi_s^\star$, the comparator policy, we have

$$\begin{aligned}&\eta\left\langle \widehat{Q}_{\tau,s}^{(k)}, \pi_s^{(k+1)} - \pi_s^\star\right\rangle + D_{\mathrm{KL}}\left(\pi_s^\star \,\middle\|\, \pi_s^{(k+1)}\right) + D_{\mathrm{KL}}\left(\pi_s^{(k+1)} \,\middle\|\, \pi_s^{(k)}\right) - D_{\mathrm{KL}}\left(\pi_s^\star \,\middle\|\, \pi_s^{(k)}\right)\\ \le&\psi_s\left(D_{\mathrm{KL}}\left(\pi_s^{(k+1)} \,\middle\|\, \pi_s^{(k)}\exp\left(-\eta\widehat{Q}_{\tau,s}^{(k)}\right)/Z_s^{(k)}\right)\right).\end{aligned}$$

Notice here the key difference to the proof of Theorem E.1.7 is that we do **not** drop the term

$D_{\mathrm{KL}}\left(\pi_s^{(k+1)} \,\middle\|\, \pi_s^{(k)}\right)$.

By some algebraic rearrangement and taking expectation with respect to distribution $d_\rho^\star$, we then get

$$\begin{aligned}&\mathbb{E}_{s\sim d_\rho^\star}\left[\eta\left\langle \widehat{Q}_{\tau,s}^{(k)}, \pi_s^{(k+1)}-\pi_s^{(k)}\right\rangle - \psi_s\left(D_{\mathrm{KL}}\left(\pi_s^{(k+1)} \,\middle\|\, \pi_s^{(k)}\exp\left(-\eta \widehat{Q}_{\tau,s}^{(k)}\right)/Z_s^{(k)}\right)\right)\right]\\ &\quad+\eta\mathbb{E}_{s\sim d_\rho^\star}\left[\left\langle \widehat{Q}_{\tau,s}^{(k)}, \pi_s^{(k)}-\pi_s^\star\right\rangle\right] + \mathbb{E}_{s\sim d_\rho^\star}\left[D_{\mathrm{KL}}\left(\pi_s^{(k+1)} \,\middle\|\, \pi_s^{(k)}\right)\right] \le D_k^\star - D_{k+1}^\star,\end{aligned} \tag{E.11}$$

where $D_k^\star = \mathbb{E}_{s\sim d_\rho^\star}\left[D_{\mathrm{KL}}\left(\pi_s^\star \,\middle\|\, \pi_s^{(k)}\right)\right]$.

For the first expectation above, we have

$$\begin{aligned}&\mathbb{E}_{s\sim d_\rho^\star}\left[\Delta_{\tau,s}^{(k)}\right]\\ \ge&\frac{1}{1-\gamma}\mathbb{E}_{s\sim d_{d_\rho^\star}^{(k+1)}}\left[\Delta_{\tau,s}^{(k)}\right]\\ =&\frac{\eta}{1-\gamma}\mathbb{E}_{s\sim d_{d_\rho^\star}^{(k+1)}}\left[\left\langle Q_{\tau,s}^{(k)}, \pi_s^{(k+1)}-\pi_s^{(k)}\right\rangle\right] + \frac{\eta}{1-\gamma}\mathbb{E}_{s\sim d_{d_\rho^\star}^{(k+1)}}\left[\left\langle \widehat{Q}_{\tau,s}^{(k)} - Q_{\tau,s}^{(k)}, \pi_s^{(k+1)}-\pi_s^{(k)}\right\rangle\right]\\ &\quad-\frac{\vartheta_\rho}{1-\gamma}\mathbb{E}_{s\sim\nu^{(k)}}\left[\psi_s\left(D_{\mathrm{KL}}\left(\pi_s^{(k+1)} \,\middle\|\, \pi_s^{(k)}\exp\left(-\eta\widehat{Q}_{\tau,s}^{(k)}\right)/Z_s^{(k)}\right)\right)\right]\\ &\hfill\text{(By Assumption (A2))}\\ \ge&\frac{\eta}{1-\gamma}\mathbb{E}_{s\sim d_{d_\rho^\star}^{(k+1)}}\left[\left\langle Q_{\tau,s}^{(k)}, \pi_s^{(k+1)}-\pi_s^{(k)}\right\rangle\right] - \frac{\eta\vartheta_\rho\epsilon_{\text{critic}}}{1-\gamma} - \frac{\vartheta_\rho}{1-\gamma}\psi(\eta\epsilon_{\text{actor}})\\ &\hfill\text{(By Assumption (A1'), (A3) and concavity of } \psi_s.)\\ =&\eta\left(V_{\tau,d_\rho^\star}^{(k+1)} - V_{\tau,d_\rho^\star}^{(k)}\right) - \eta\tau\mathbb{E}_{s\sim d_{d_\rho^\star}^{(k+1)}}\left[D_{\mathrm{KL}}\left(\pi_s^{(k+1)} \,\middle\|\, \pi_s^{(k)}\right)\right] - \frac{\eta\vartheta_\rho\epsilon_{\text{critic}}}{1-\gamma} - \frac{\vartheta_\rho\psi(\eta\epsilon_{\text{actor}})}{1-\gamma}\\ &\hfill\text{(By Lemma 6.5.5, the modified performance difference lemma.)}\\ \ge&\eta\left(V_{\tau,d_\rho^\star}^{(k+1)} - V_{\tau,d_\rho^\star}^{(k)}\right) - \eta\tau\vartheta_\rho\mathbb{E}_{s\sim d_\rho^\star}\left[D_{\mathrm{KL}}\left(\pi_s^{(k+1)} \,\middle\|\, \pi_s^{(k)}\right)\right] - \frac{\eta\vartheta_\rho\epsilon_{\text{critic}}}{1-\gamma} - \frac{\vartheta_\rho\psi(\eta\epsilon_{\text{actor}})}{1-\gamma}\end{aligned}$$

Then, for the second expectation in Eq. (E.11), we can similarly apply Lemma 6.5.5 and obtain

$$\begin{aligned}\eta\mathbb{E}_{s\sim d_\rho^\star}\left[\left\langle \widehat{Q}_{\tau,s}^{(k)}, \pi_s^{(k)}-\pi_s^\star\right\rangle\right] =&\eta\mathbb{E}_{s\sim d_\rho^\star}\left[\left\langle Q_{\tau,s}^{(k)}, \pi_s^{(k)}-\pi_s^\star\right\rangle\right] + \eta\mathbb{E}_{s\sim d_\rho^\star}\left[\left\langle \widehat{Q}_{\tau,s}^{(k)} - Q_{\tau,s}^{(k)}, \pi_s^{(k)}-\pi_s^\star\right\rangle\right]\\ \ge&(1-\gamma)\eta\left(V_{\tau,\rho}^{(k)} - V_{\tau,\rho}^\star\right) + (1-\gamma)\eta\tau D_k^\star - \eta\vartheta_\rho\epsilon_{\text{critic}}.\end{aligned}$$

By plugging these bounds back into Eq. (E.11), we then have

$$\begin{aligned}(1-\gamma)\left(V_{\tau,\rho}^{(k)}-V_{\tau,\rho}^{\star}\right)\leq&\frac{D_k^\star}{\eta}-\frac{D_{k+1}^\star}{\eta}+V_{\tau,d_\rho^\star}^{(k)}-V_{\tau,d_\rho^\star}^{(k+1)}+\frac{(2-\gamma)\vartheta_\rho\epsilon_{\text{critic}}}{1-\gamma}+\frac{\vartheta_\rho\psi(\eta\epsilon_{\text{actor}})}{(1-\gamma)\eta}\\&+\left(\tau\vartheta_\rho-\frac{1}{\eta}\right)\mathbb{E}_{s\sim d_\rho^\star}\left[D_{\text{KL}}\left(\pi_s^{(k+1)}\,\middle\|\,\pi_s^{(k)}\right)\right]-(1-\gamma)\tau D_k^\star\\\leq&\frac{D_k^\star}{\eta}-\frac{D_{k+1}^\star}{\eta}+V_{\tau,d_\rho^\star}^{(k)}-V_{\tau,d_\rho^\star}^{(k+1)}+\frac{(2-\gamma)\vartheta_\rho\epsilon_{\text{critic}}}{1-\gamma}+\frac{\vartheta_\rho\psi(\eta\epsilon_{\text{actor}})}{(1-\gamma)\eta}\\&\text{(By taking }\eta\leq\tfrac{1}{\tau\vartheta_\rho}\text{ and noticing KL divergence is non-negative.)}\end{aligned}$$

Finally, by noticing that $\frac{\psi(\eta\epsilon_{\text{actor}})}{\eta}\leq\psi(\epsilon_{\text{actor}})$ and taking sum from $k=0$ to $K-1$, we can get

$$\frac{1}{K}\sum_{k=0}^{K-1}\left(V_{\tau,\rho}^{(k)}-V_{\tau,\rho}^{\star}\right)\leq\frac{1}{K}\left(\frac{D_0^\star}{(1-\gamma)\eta}+\frac{V_{\tau,d_\rho^\star}^{(0)}}{1-\gamma}\right)+\frac{\vartheta_\rho\psi(\epsilon_{\text{actor}})+(2-\gamma)\vartheta_\rho\epsilon_{\text{critic}}}{(1-\gamma)^2}.$$

□

E.3 Technical Lemmas for Continuous-Space MDPs and DAPO

Lemma E.3.1. *For any two policies $\pi,\tilde{\pi}:\mathcal{S}\mapsto\mathscr{P}(\mathcal{A})$ and initial distribution $\rho\in\mathscr{P}(\mathcal{S})$, it holds that*

$$V_\rho^\pi-V_\rho^{\tilde{\pi}}=\frac{1}{1-\gamma}\mathbb{E}_{s\sim d_\rho^\pi}\left[\left\langle Q_s^{\tilde{\pi}},\pi_s-\tilde{\pi}_s\right\rangle\right]=\frac{1}{1-\gamma}\mathbb{E}_{s\sim d_\rho^{\tilde{\pi}}}\left[\langle Q_s^\pi,\pi_s-\tilde{\pi}_s\rangle\right].$$

Proof. We have

$$\begin{aligned}V_\rho^\pi-V_\rho^{\tilde{\pi}}\overset{\text{(i)}}{=}&\mathbb{E}_{s_0\sim\rho}\left[\sum_{t=0}^\infty\gamma^t c(s_t,a_t)\right]-V_\rho^{\tilde{\pi}}\\=&\mathbb{E}_{s_0\sim\rho}\left[\sum_{t=0}^\infty\gamma^t\left(c(s_t,a_t)+V_{s_{t+1}}^{\tilde{\pi}}-V_{s_t}^{\tilde{\pi}}\right)\right]\\=&\mathbb{E}_{s_0\sim\rho}\left[\sum_{t=0}^\infty\gamma^t\left(Q_{s_t,a_t}^{\tilde{\pi}}-V_{s_t}^{\tilde{\pi}}\right)\right]\\=&\mathbb{E}_{s_0\sim\rho}\left[\sum_{t=0}^\infty\gamma^t\left\langle Q_{s_t}^{\tilde{\pi}},\pi_{s_t}-\tilde{\pi}_{s_t}\right\rangle\right]\end{aligned}$$

$$
\begin{aligned}
&= \int \left(\sum_{t=0}^{\infty} \gamma^t \left\langle Q_{s_t}^{\tilde{\pi}}, \pi_{s_t} - \tilde{\pi}_{s_t} \right\rangle \right) p_{s_0 \sim \rho}^{\pi}(\tau) \, \mathrm{d}\tau \\
&\qquad\qquad (\tau = (s_0, s_1, \dots) \text{ represents the whole states trajectory.}) \\
&\overset{\text{(ii)}}{=} \sum_{t=0}^{\infty} \int_{\mathcal{S}} \gamma^t \langle Q_s^{\pi}, \pi_s - \tilde{\pi}_s \rangle \, p_t^{\pi}(s \mid \rho) \mathrm{d}s \qquad \text{(By dominated convergence theorem.)} \\
&= \int_{\mathcal{S}} \langle Q_s^{\pi}, \pi_s - \tilde{\pi}_s \rangle \sum_{t=0}^{\infty} \gamma^t p_t^{\pi}(s \mid \rho) \mathrm{d}s \\
&= \frac{1}{1-\gamma} \mathbb{E}_{s \sim d_\rho^{\pi}} \left[\left\langle Q_s^{\tilde{\pi}}, \pi_s - \tilde{\pi}_s \right\rangle \right].
\end{aligned}
$$

The above equality (ii) holds because or each term $\left\langle Q_{s_t}^{\tilde{\pi}}, \pi_{s_t} - \tilde{\pi}_{s_t} \right\rangle$, the densities for $s_{t'}$ with $t' \neq t$ are marginalized. The second form can be obtained by similarly expand $V_\rho^{\tilde{\pi}}$ in equality (i) above and thus the proof is complete. □

Lemma E.3.2. *Under Assumption (A4), for any initial distribution $\rho \in \mathscr{P}(\mathcal{S})$ and policy π, it holds that $\left\| d_\rho^{\pi} \right\|_\infty \leq \alpha$.*

Proof. By Markovian property, for any $s \in \mathcal{S}$ and $t \geq 1$, we have

$$
\begin{aligned}
p_t^{\pi}(s \mid \rho) &= \int_{\mathcal{S}} \cdots \int_{\mathcal{S}} p_1^{\pi}(s_t \mid s_{t-1}) \dots p_1^{\pi}(s_1 \mid s_0) \rho(s_0) \mathrm{d}s_{t-1} \dots \mathrm{d}s_1 \mathrm{d}s_0 \\
&\overset{\text{(i)}}{\leq} \alpha \int_{\mathcal{S}} \cdots \int_{\mathcal{S}} p_1^{\pi}(s_{t-1} \mid s_{t-2}) \dots p_1^{\pi}(s_1 \mid s_0) \rho(s_0) \mathrm{d}s_{t-1} \dots \mathrm{d}s_1 \mathrm{d}s_0 \\
&= \alpha.
\end{aligned}
$$

The abovve inequality (i) holds because by Assumption (A4), we have $\left\| p_t^{\pi}(\cdot \mid s_{t-1}) \right\|_\infty \leq \alpha$ for any $s_{t-1} \in \mathcal{S}$. Then, for any $s \in \mathcal{S}$, we can straightforwardly have $d_{\rho,s}^{\pi} = (1-\gamma) \sum_{t=0}^{\infty} \gamma^t p_t^{\pi}(s \mid \rho) \leq (1-\gamma) \sum_{t=0}^{\infty} \gamma^t \alpha = \alpha$. □

Lemma 6.6.6. *If a proper closed convex functional $\Phi : \mathcal{C} \mapsto \mathbb{R}$ is Fréchet differentiable at $x \in \mathcal{C}$, then $\partial\Phi(x) = \{\nabla\Phi(x)\}$. That is, the subgradient is a singleton.*

Proof. Let $x^* \in \partial\Phi(x)$ and $h \in \mathcal{X}$ be arbitrary. Then, by definition of the subgradient, for any real number $\lambda \in \mathbb{R}$, we have

$$
\lambda \langle h, x^* \rangle + \Phi(x) \leq \Phi(x + \lambda h).
$$

By Lemma 6.6.3, since Φ is Fréchet differentiable, it is also Gâteaux differentiable. That is, if we restrict $\lambda > 0$, we then have

$$\langle h, x^* \rangle \leq \frac{\Phi(x + \lambda h) - \Phi(x)}{\lambda} \to \langle h, \nabla \Phi(x) \rangle \quad \text{as } \lambda \downarrow 0.$$

Similarly, if we restrict $\lambda < 0$, we can have

$$\langle h, x^* \rangle \geq \frac{\Phi(x + \lambda h) - \Phi(x)}{\lambda} \to \langle h, \nabla \Phi(x) \rangle \quad \text{as } \lambda \uparrow 0.$$

Therefore, we must have $\langle h, x^* \rangle = \langle h, \nabla \Phi(x) \rangle$ for any $h \in \mathcal{X}$, which implies $x^* = \nabla \Phi(x)$. Therefore, we have $\partial \Phi(x) = \{\nabla \Phi(x)\}$. □

Lemma 6.6.7. *Let $\Phi : \mathcal{C} \mapsto \mathbb{R}$ be a proper closed convex functional. If Φ is in addition also strictly convex over $\mathcal{C}$, then for any $x_1, x_2 \in \mathcal{C}$, we have $\partial \Phi(x_1) \cap \partial \Phi(x_2) = \emptyset$.*

Proof. We prove the contraposition of the statement. That is, if there exists $x^* \in \partial \Phi(x_1) \cap \partial \Phi(x_2)$, then, for any $x' \in \mathcal{C}$, we have

$$\Phi(x') \geq \Phi(x_1) + \left\langle x' - x_1, x^* \right\rangle \quad \text{and} \quad \Phi(y) \geq \Phi(x_2) + \left\langle x' - x_2, x^* \right\rangle.$$

Plugging $x' = x_2$ and $x' = x_1$ into the above two inequalities, respectively, we can then get

$$\Phi(x_1) - \Phi(x_2) = \langle x_1 - x_2, x^* \rangle.$$

Therefore, for some $\lambda \in (0, 1)$, we have

$$\Phi(\lambda x_1 + (1 - \lambda) x_2) \geq f(x_1) + \lambda \langle x_1 - x_2, x^* \rangle = \lambda \Phi(x_1) + (1 - \lambda) \Phi(x_2).$$

However, since Φ is convex, we also have $\Phi(\lambda x_1 + (1 - \lambda) x_2) \leq \lambda \Phi(x_1) + (1 - \lambda) \Phi(x_2)$. Thus, we must have

$$\Phi(\lambda x_1 + (1 - \lambda) x_2) = \lambda \Phi(x_1) + (1 - \lambda) \Phi(x_2),$$

which implies that Φ cannot be strictly convex and the proof is complete. □

Lemma 6.6.8. *Let $\Phi : \mathcal{C} \mapsto \mathbb{R}$ be a proper closed Legendre-type functional. Then, for any $x \notin \operatorname{int}(\mathcal{C})$, we have $\partial\Phi(x) = \emptyset$.*

Proof. For $x \notin \mathcal{C}$, this is obvious since $\Phi(x) = \infty$. For $x \in \operatorname{bd} \mathcal{C}$ with $\Phi(x) < \infty$, we take some $x_0 \in \operatorname{int}(\mathcal{C})$ and consider

$$\lim_{\lambda \downarrow 0} \frac{\Phi(x + \lambda(x_0 - x)) - \Phi(x)}{\lambda} = \lim_{\lambda \downarrow 0} \frac{\Phi((1-\lambda)x + \lambda x_0) - \Phi(x)}{\lambda} \leq \Phi(x_0) - \Phi(x), \tag{E.12}$$

where the last inequality comes from the convexity of Φ. Since Φ is a Legendre-type functional, the norm of its gradient goes to infinity as the point is approaching the boundary of $\mathcal{C}$. Thus, because of the finite upper bound in Eq. (E.12), there exists some $x_0 \in \operatorname{int}(\mathcal{C})$ such that

$$\lim_{\lambda \downarrow 0} \frac{\Phi(x + \lambda(x_0 - x)) - \Phi(x)}{\lambda} = -\infty.$$

Finally, as given in Brezis and Brézis [2011] (as an exercise in Section Problems), since

$$\lim_{\lambda \downarrow 0} \frac{\Phi(x + \lambda(x_0 - x)) - \Phi(x)}{\lambda} = \sup_{x^* \in \partial\Phi(x)} \langle x^*, x_0 - x \rangle ,$$

we have $\sup_{x^* \in \partial\Phi(x)} \langle x^*, x_0 - x \rangle = -\infty$, which implies $\partial\Phi(x) = \emptyset$.

That is, we have $\partial\Phi(x) = \emptyset$ for any $x \notin \operatorname{int}(\mathcal{C})$ □

E.4 Convergence of DAPO-KL in Continuous-Space MDPs

Theorem 6.6.11 (Linear Convergence for Continuous-space MDPs)**.** *Consider Algorithm 10 with initial policy $\pi^{(0)}$, initial distribution $\rho \in \mathscr{P}(\mathcal{S})$, arbitrary comparator policy $\pi^\star$ and Φ being the negative differential entropy. Suppose Assumptions (A1')-(A5) hold and the step sizes satisfy $\eta_0 > 1$ and $\eta_{k+1} \geq \left(\vartheta_\rho/(\vartheta_\rho - 1)\right)\eta_k$ for all $k \geq 0$. Then, we have*

$$V_\rho^{(K)} - V_\rho^\star \leq \left(1 - \frac{1}{\vartheta_\rho}\right)^K \left(V_\rho^{(0)} - V_\rho^\star + \frac{D_0^\star/(\vartheta_\rho - 1)}{(1-\gamma)\eta_0}\right) + \frac{\vartheta_\rho^2 \psi(\epsilon_{\text{actor}}) + (1+\vartheta_\rho)\epsilon_{\text{critic}}}{1-\gamma},$$

where $D_0^\star = \mathbb{E}_{s \sim d_\rho^\star}\left[D_\Phi(\pi_s^\star, \pi_s^{(0)})\right]$ and $\psi(x) = (1 + C_\rho)\left(x + \sqrt{2x}\right)$ for $x \geq 0$.

Proof. The algebra is the same as the proof of Theorem E.1.7 with $\psi^\Phi(x) = (1+C_\rho)\left(x + \sqrt{2x}\right)$

and additional assumptions ensure that he same algebraic manipulations are still valid. □

Lemma E.4.1. *Under Assumption (A4), if we take $\rho = \mathsf{Unif}(\mathcal{S})$, then for any k and any policy π, it holds that $\left\| \frac{d^\pi_\rho}{d^{(k)}_\rho} \right\|_\infty \leq \frac{\alpha \mathrm{vol}(\mathcal{S})}{1-\gamma}$, where $\mathrm{vol}(\mathcal{S})$ is the volume of the state space $\mathcal{S}$.*

Proof. By definition of state-visitation distribution in Eq. (6.3), we can immediately get $d^{(k)}_{\rho,s} \geq (1-\gamma)\rho_s$ for any $s \in \mathcal{S}$ by truncating all terms with $t \geq 1$. Since $\rho = \mathsf{Unif}(\mathcal{S})$, we have $\rho_s = \frac{1}{\mathrm{vol}(\mathcal{S})}$ for any $s \in \mathcal{S}$. Furthermore, by Lemma E.3.2, we know that $d^\pi_{\rho,s} \leq \alpha$ for any $s \in \mathcal{S}$. Thus, we have

$$\left\| \frac{d^\pi_\rho}{d^{(k)}_\rho} \right\|_\infty = \sup_{s \in \mathcal{S}} \frac{d^\pi_{\rho,s}}{d^{(k)}_{\rho,s}} \leq \frac{\alpha}{(1-\gamma)\rho_s} \leq \frac{\alpha \mathrm{vol}(\mathcal{S})}{1-\gamma}.$$

□

Lemma E.4.2. *Let $\Phi(\pi) = \int_{\mathcal{A}} \pi_a \log \pi_a \mathrm{d}a$ for $\pi \in \mathscr{P}(\mathcal{A})$ be the negative differential entropy. Its conjugate is $\Phi^*(x^*) = \log\left(\int_{\mathcal{A}} \exp(x^*_a)\mathrm{d}a\right)$ for $x^* \in \mathscr{L}_\infty(\mathcal{A})$ such that $\int_{\mathcal{A}} \exp(x^*_a)\mathrm{d}a$ exists.*

Proof. We first define what *regular point* means [Luenberger, 1997]. Let $T : \mathcal{C} \mapsto \mathbb{R}$ be a Fréchet differentiable operator and $\mathcal{C} \subseteq \mathcal{X}$. Then, for $x \in \mathcal{C}$, if the Fréchet derivative $T'(x) \in \mathcal{X}^*$ maps $\mathcal{X}$ onto $\mathbb{R}$, x is called a regular point of T. Here, treating the Fréchet derivative as a map means to have $T'(x)(\tilde{x}) := \langle T'(x), \tilde{x} \rangle$.

Now, by definition, its conjugate function is

$$\Phi^*(x^*) = \sup_{\pi \in \mathscr{P}(\mathcal{A})} \int_{\mathcal{A}} (\pi_a x^*_a - \pi_a \log \pi_a)\, \mathrm{d}a.$$

The right-hand side of the above is an optimization problem with an equality constraint $\int_{\mathcal{A}} \pi_a \mathrm{d}a - 1 = 0$. The Fréchet derivative of this operator ($T(\pi) = \int_{\mathcal{A}} \pi_a \mathrm{d}a - 1$) is $T'(\pi) = \pi \in \mathscr{L}_\infty(\mathcal{A})$. Since for any $r \in \mathbb{R}$, we can take a constant function $x_a = r$ for any $a \in \mathcal{A}$ such that $T'(\pi)(x) = \langle \pi, x \rangle = r$, we know that π is a regular point of T. Therefore, as stated in Luenberger [1997, Section 9.3], this optimization problem exists a multiplier β such that

$$x^*_a - \log \pi_a - 1 + \beta = 0, \quad \forall a \in \mathcal{A} \implies \pi_a \propto \exp(x^*_a).$$

Plugging it back, we can have

$$\Phi^*(x^*) = \log\left(\int_{\mathcal{A}} \exp(x_a^*)\mathrm{d}a\right).$$

□

E.5 Implementation Details

In this section, we will provide details about specific implementations of DAPO-KL, AMPO and MAMPO and their hyperparameters choices. These implementations are based on modifying the actor loss in SAC while keeping other parts unchanged. Therefore, we will first present the pseudocode of SAC and then give modified actor losses for DAPO-KL, AMPO and MAMPO.

E.5.1 SAC

The pseudocode of SAC is given in Algorithm 17, where $J(\tau, \theta)$ and $J_q(\phi_i, \mathcal{B}, \phi_{\text{targ},1}, \phi_{\text{targ},2})$ in line 9 and 10 represent the loss functions to update regularization parameter τ and q-value networks, respectively. More details of these two loss functions can be found in Haarnoja et al. [2018b].

E.5.2 DAPO-KL

To implement DAPO-KL, we will basically replace the update rule in 13 of Algorithm 17 by DAPO-KL's update rule. To do this, we first need to rewrite DAPO-KL's update rule in Eq. (6.27) in terms of $q_\tau^\pi = Q_\tau^\pi(s,a) + \tau \log \pi(a \mid s)$. In particular, we have

$$\begin{aligned}
\theta^{(k+1)} &\in \arg\min_\theta \mathbb{E}_{s\sim d_\rho^{(k)}}\left[D_{\mathrm{KL}}\left(\pi_s^\theta \;\middle\|\; \pi_s^{(k)} \exp\left(\eta_k Q_{\tau,s}^{(k)}\right)/Z_s^{(k)}\right)\right] \\
&= \arg\min_\theta \mathbb{E}_{s\sim d_\rho^{(k)}}\left[D_{\mathrm{KL}}\left(\pi_s^\theta \;\middle\|\; \pi_s^{(k)} \exp\left(\eta_k q_{\tau,s}^{(k)} - \eta_k \tau \log \pi_s^{(k)}\right)/Z_s^{(k)}\right)\right] \\
&= \arg\min_\theta \mathbb{E}_{s\sim d_\rho^{(k)}, a\sim\pi_s^\theta}\left[\tau \log \pi^\theta(a \mid s) - (1-\beta)\tau \log \pi^{(k)}(a \mid s) - \beta q_\tau^{(k)}(s,a)\right], \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{(By ignoring normalization constants.)}
\end{aligned}$$

where $\beta \overset{\text{def}}{=} \eta_k \tau < 1$. We can see that the update rule exactly becomes the SAC's update rule when $\beta = 1$, which means to have $\eta_k = \frac{1}{\tau}$, consistent with our derivation in Section 6.4.3.

Thus, to implement DAPO-KL, we replace the gradient in line 13 of Algorithm 17 by

$$\lambda_\pi \nabla_\theta \mathop{\mathbb{E}}_{\substack{s\sim\mathcal{B}\\ a\sim\pi_s^\theta}} \left[\tau^{(k+1)} \log \pi^\theta(a \mid s) - (1-\beta)\tau^{(k+1)} \log \pi^{\theta^{(k)}}(a \mid s) - \beta \min_{i=1,2} q^{\phi_i^{(k+1)}}(s,a)\right]\Big|_{\theta=\theta_{j-1}^{(k+1)}},$$

where β is a user-specified hyperparameter. Note that we use the standard reparameterization trick to compute the above gradient [Kingma and Welling, 2013].

E.5.3 AMPO

To implement AMPO in Alfano et al. [2024], we will need to replace the update rule in line 13 of Algorithm 17 by AMPO's loss in Eq. (6.19) in a more concrete form. That is, we have

$$\begin{aligned}\theta^{(k+1)} &\in \arg\min_\theta \mathbb{E}_{s\sim d_\rho^{(k)}} \left[\left\| f_s^\theta - \left(\log \pi_s^{(k)} + \eta_k Q_s^{(k)}\right)\right\|_2^2\right] \\ &= \arg\min_\theta \mathbb{E}_{s\sim d_\rho^{(k)}} \left[\left\| f_s^\theta - (1-\eta_k\tau)\log \pi_s^{(k)} - \eta_k q_{\tau,s}^{(k)}\right\|_2^2\right].\end{aligned} \tag{E.13}$$

Here, f^θ should be the exponent of a Gaussian distribution. Therefore, if $g^\theta : \mathcal{S} \mapsto \mathbb{R}^{2n_\mathcal{A}}$ is the policy network, where $n_\mathcal{A}$ is the dimension of the action space. Then, we have

$$f^\theta(s,a) = -\sum_{i=1}^{n_\mathcal{A}} \frac{\left(g^\theta(s)_i\right)^2 - 2a_i g^\theta(s)_i}{2\left(g^\theta(s)_{n_\mathcal{A}+i}\right)^2}. \tag{E.14}$$

Therefore, to implement AMPO, we replace the gradient in line 13 of Algorithm 17 by

$$\lambda_\pi \nabla_\theta \mathop{\mathbb{E}}_{\substack{s\sim\mathcal{B}\\ a\sim\mathsf{Unif}(\mathcal{A})}} \left[\left(f^\theta(s,a) - (1-\eta\tau^{(k+1)})\log \pi^{\theta^{(k)}}(a \mid s) - \eta \min_{i=1,2} q^{\phi_i^{(k)}}(s,a)\right)^2\right]\Big|_{\theta=\theta_{j-1}^{(k+1)}},$$

where η is the mirror descent learning rate.

E.5.4 MAMPO

As discussed in Section 6.7, MAMPO tries to optimize

$$\theta^{(k+1)} \in \arg\min_{\theta} \mathbb{E}_{s\sim d_\rho^{(k)}} \left[\left\| f_s^\theta - \left(\pi_s^{(k)} + \eta_k Q_s^{(k)} \right) \right\|_2^2 \right],$$

where $f^\theta(s,a)$ is defined in Eq. (E.14). Therefore, to implement MAMPO, we replace the gradient in line 13 of Algorithm 17 by

$$\lambda_\pi \nabla_\theta \mathop{\mathbb{E}}_{\substack{s\sim\mathcal{B} \\ a\sim\mathsf{Unif}(\mathcal{A})}} \left[\left(f^\theta(s,a) - \pi^{\theta^{(k)}}(a \mid s) - \eta \min_{i=1,2} q^{\phi_i^{(k)}}(s,a) + \eta\tau^{(k+1)} \log \pi^{\theta^{(k)}}(a \mid s) \right)^2 \right] \Bigg|_{\theta=\theta_{j-1}^{(k+1)}}.$$

E.5.5 Hyperparameter Settings

We use the implementation of SAC from the *Stable Baseline 3* under the MIT license [Raffin et al., 2021]. Then, we implement DAPO-KL, AMPO-KL, and MAMPO as modifications of its SAC's implementation. All model trainings were completed on 8 NVIDIA V100 GPUs.

We use SAC's default hyperparameters on all environments for both SAC and DAPO-KL, while AMPO-KL and MAMPO contain some tuning. Full hyperparameter details are provided in Table E.1. Particularly for hyperparameter β in DAPO-KL, we take different values for different tasks as shown in Table E.2.

Table E.1: Hyperparameters of all algorithms

Hyperparameter	SAC	**DAPO-KL**	AMPO	MAMPO
Adam learning rate	3×10^{-4}	3×10^{-4}	2×10^{-5}	3×10^{-4}
MD learning rate (η)	NA	NA	1.0	1.0
Entropy regularization (τ)	auto*	auto*	0	0
Number of hidden layers	2			
Hidden layer size	256			
Batch size	256			
Discount factor (γ)	0.99			
Target mixture weight (ω)	0.005			
Replay buffer size	1×10^6			

* Being "auto" in entropy regularization means to use the update rule at line 9 of Algorithm 17 to automatically adjust τ.

Table E.2: Values of hyperparameter β in DAPO-KL for different MuJoCo tasks.

Environments	HalfCheetah-v4	Hopper-v4	Walker2d-v4	Ant-v4
β	0.7	0.6	0.4	0.7

E.6 Additional Experiment Results on AMPO

In this section, we first introduce the original version of AMPO proposed in Alfano et al. [2024], which is slightly different from what we have in Eq. (E.13). Then, we present a partial record of our efforts in tuning AMPO, which shows the difficulty of using this algorithm in practical scenario. Nevertheless, we retain the possibility that our implementation of AMPO may not be the optimal.

E.6.1 Variants of AMPO

The original version of AMPO proposed in Alfano et al. [2024] is given as

$$\theta^{(k+1)} \in \arg\min_{\theta} \mathbb{E}_{s\sim d_\rho^{(k)}} \left[\left\| f_s^\theta - \left(f_s^{(k)} + \eta_k Q_s^{(k)} \right) \right\|_2^2 \right]. \tag{E.15}$$

While seemingly different from Eq. (E.13), these two are essentially the same from a theoretical perspective. To see this, as shown in Corollary 6.3.3, when Φ is the negative entropy restricted on $\Delta(\mathcal{A})$, we can freely take $\nabla\Phi(\pi)$ to be any vector in $\partial\Phi(\pi)$ while the Bregman divergence D_Φ is still well-defined. In particular, we have $\partial\Phi(\pi) = \{\log\pi + c\mathbf{1} \mid c \in \mathbb{R}\}$ with $\mathbf{1} = \begin{bmatrix} 1 & \cdots & 1 \end{bmatrix}^\top$. As a result, since the difference between f_s^θ and $\log\pi_s^{(k)}$ is only an action-independent normalization constant, Eq. (E.13) and Eq. (E.15) are theoretically equivalent.*

Nevertheless, Eq. (E.13) and Eq. (E.15) may still be empirically different since the constant difference can still affect the L_2-loss minimization. Therefore, we consider and

*While Alfano et al. [2024] claims to obtain Eq. (E.15) by taking Φ to be the negative entropy on $\mathbb{R}_+^{|\mathcal{A}|}$, this is not an appropriate argument because such a choice of Φ will enforce $\nabla\Phi(\pi) = \log\pi + \mathbf{1}$, excluding the freedom of choosing action-independent constant.

empirically compare these two different theoretically equivalent variants of AMPO-KL.[†]

E.6.2 Comparison between MAMPO and AMPO-KL

Here, we provide a comparison between MAMPO and the two variants of AMPO-KL in Fig. E.1, where both variants use the same set of hyperparameters as given in Table E.1. From the plots, we can see that both variants of AMPO-KL almost cannot learn anything non-trivial in all tasks.

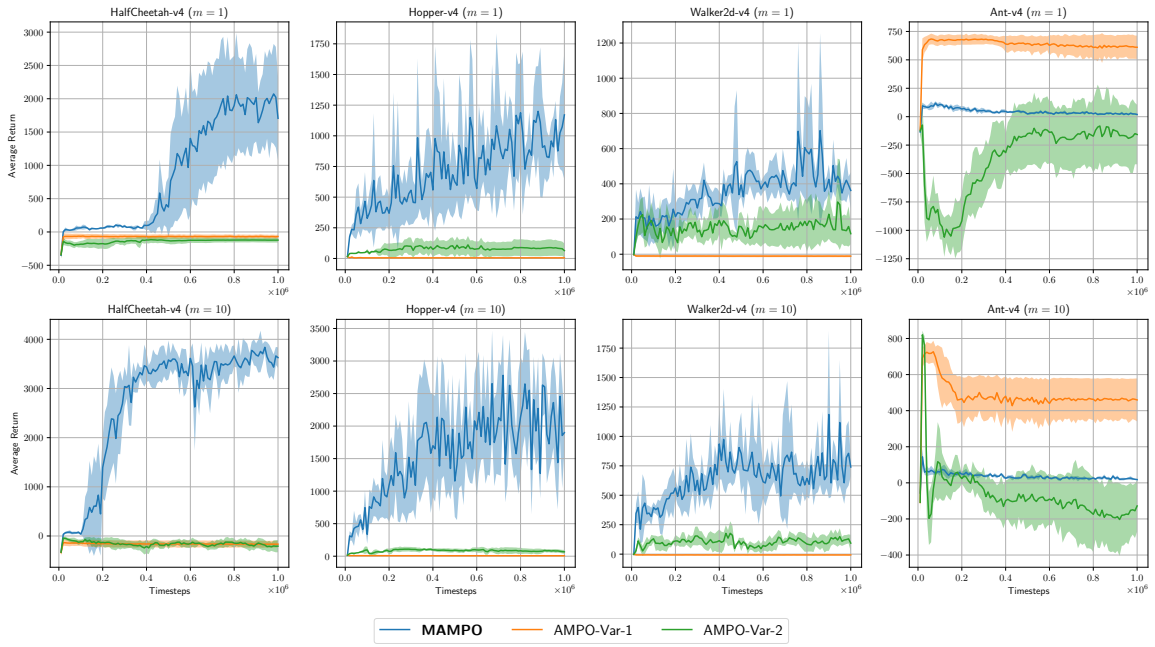


Figure E.1: Comparison under $m = 1$ and $m = 10$ gradient steps per iteration between MAMPO and variants of AMPO-KL. Here, "AMPO-Var-1" refers to Eq. (E.13) and "AMPO-Var-2" refers to Eq. (E.15). Each curve is averaged over 5 different random seeds and the shaded area represents the 95% confidence interval.

E.6.3 AMPO-KL under Different Hyperparameters

Finally, we also provide a performance comparison of variants of AMPO-KL under different hyperparameter settings, where we only show the final-step performance under each setting,

[†]We use the variant in Eq. (E.13) in all previous experiments.

given in Table E.3, E.4, E.5 and E.6, in which each data point is averaged over 3 different random seeds and $\pm$ represents the 95% confidence interval. Nevertheless, we can easily see that AMPO-KL still cannot learn anything non-trivial under all of these settings.

Table E.3: Final-step performance of AMPO-Var-1 (Eq. (E.13)) in HalfCheetah-v4 with entropy regularization ($\tau = 1.0$).

	$\text{lr} = 5 \times 10^{-6}$	$\text{lr} = 1 \times 10^{-5}$	$\text{lr} = 5 \times 10^{-5}$	$\text{lr} = 1 \times 10^{-4}$
$\eta_k = 0.1$	-94.08 ± 37.82	-77.18 ± 16.73	-3.33 ± 0.17	-178.56 ± 41.57
$\eta_k = 1$	-79.8 ± 47.96	-61.83 ± 20.85	-4.19 ± 0.49	-14.93 ± 11.05
$\eta_k = 10$	-27.83 ± 28.96	-201.02 ± 71.98	-8.37 ± 0.33	-7.58 ± 0.73
$\eta_k = 100$	-220.2 ± 124.88	-210.46 ± 168.3	-8.12 ± 0.35	-7.36 ± 0.75

Table E.4: Final-step performance of AMPO-Var-1 (Eq. (E.13)) in HalfCheetah-v4 without entropy regularization ($\tau = 0$).

	$\text{lr} = 5 \times 10^{-6}$	$\text{lr} = 1 \times 10^{-5}$	$\text{lr} = 5 \times 10^{-5}$	$\text{lr} = 1 \times 10^{-4}$
$\eta_k = 0.1$	-131.27 ± 62.23	-129.24 ± 51.24	-123.17 ± 63.02	-109.34 ± 84.59
$\eta_k = 1$	-93.94 ± 46.46	-95.82 ± 49.19	-83.12 ± 59.42	-97.82 ± 47.88
$\eta_k = 10$	-50.57 ± 45.8	-98.75 ± 24.9	-81.51 ± 29.18	-57.17 ± 39.36
$\eta_k = 100$	-301.4 ± 128.66	-295.53 ± 63.03	-196.38 ± 136.29	-255.11 ± 143.2

Table E.5: Final-step performance of AMPO-Var-2 (Eq. (E.15)) in HalfCheetah-v4 with entropy regularization ($\tau = 1.0$).

	$\mathrm{lr} = 5 \times 10^{-6}$	$\mathrm{lr} = 1 \times 10^{-5}$	$\mathrm{lr} = 5 \times 10^{-5}$	$\mathrm{lr} = 1 \times 10^{-4}$
$\eta_k = 0.1$	-3.56 ± 0.29	-3.48 ± 0.52	-3.39 ± 0.22	-3.59 ± 0.22
$\eta_k = 1$	-4.34 ± 0.47	-4.24 ± 0.32	-4.29 ± 0.26	-4.22 ± 0.34
$\eta_k = 10$	-8.38 ± 0.23	-8.33 ± 0.12	-8.4 ± 0.29	-8.08 ± 0.57
$\eta_k = 100$	41.59 ± 79.77	-8.38 ± 0.1	-8.33 ± 0.56	-8.36 ± 0.52

Table E.6: Final-step performance of AMPO-Var-2 (Eq. (E.15)) in HalfCheetah-v4 without entropy regularization ($\tau = 0$).

	$\mathrm{lr} = 5 \times 10^{-6}$	$\mathrm{lr} = 1 \times 10^{-5}$	$\mathrm{lr} = 5 \times 10^{-5}$	$\mathrm{lr} = 1 \times 10^{-4}$
$\eta_k = 0.1$	-120.73 ± 34.97	-178.62 ± 34.49	-174.38 ± 67.63	-144.4 ± 54.12
$\eta_k = 1$	-87.45 ± 5.28	-121.66 ± 56.37	-129.73 ± 43.94	-157.94 ± 49.55
$\eta_k = 10$	-534.72 ± 205.25	-237.98 ± 127.79	-176.42 ± 235.9	-199.77 ± 12.45
$\eta_k = 100$	-341.7 ± 94.17	139.05 ± 118.93	-271.07 ± 149.01	-411.67 ± 56.92

Algorithm 17 Soft Actor-Critic (SAC) [Haarnoja et al., 2018b]

Input: Initial policy network parameter θ; initial q-value network parameters ϕ_1, ϕ_2; replay buffer $\mathcal{D}$; learning rates λ_q, λ_π, λ; target mixture weight $\omega \in (0,1)$; initial regularization power $\tau > 0$; number of gradient steps per iteration m
Set $\phi_{\text{targ},1}^{(0)} \leftarrow \phi_1$ and $\phi_{\text{targ},2}^{(0)} \leftarrow \phi_2$ and initialize $k \leftarrow 0$
while *not done* **do**
  Observe state s and take action $a \sim \pi_\theta(\cdot \mid s)$
  Observe next state s', reward r, done signal d and add (s, a, s', r, d) to buffer $\mathcal{D}$
  // $\{d = 1$ if s' is a terminal state; otherwise, $d = 0\}$
  If s' is a terminal state, reset the environment
  if *it's time to update* **then**
    Randomly sample a batch of transitions $\mathcal{B} = \{(s, a, r, s', d)\} \subseteq \mathcal{D}$
    Update regularization parameter by

$$\tau^{(k+1)} \leftarrow \tau^{(k)} - \lambda \nabla_\tau J(\tau, \theta^{(k)})$$

    Update q-value networks by

$$\phi_i^{(k+1)} \leftarrow \phi_i^{(k)} - \lambda_q \nabla_{\phi_i} J_q(\phi_i, \mathcal{B}, \phi_{\text{targ},1}^{(k)}, \phi_{\text{targ},2}^{(k)}, \theta^{(k)}, \tau^{(k+1)}), \quad i = 1, 2$$

    Set $\theta_0^{(k+1)} \leftarrow \theta^{(k)}$
    for $j = 1, \ldots, m$ **do**
      Update policy network by

$$\theta_j^{(k+1)} \leftarrow \theta_{j-1}^{(k+1)} - \lambda_\pi \nabla_\theta \mathop{\mathbb{E}}_{\substack{s \sim \mathcal{B} \\ a \sim \pi_s^\theta}} \left[\tau^{(k+1)} \log \pi^\theta(a \mid s) - \min_{i=1,2} q^{\phi_i^{(k+1)}}(s, a) \right] \Big|_{\theta = \theta_{j-1}^{(k+1)}}$$

    end for
    Set $\theta^{(k+1)} \leftarrow \theta_m^{(k+1)}$
    Update target networks with

$$\phi_{\text{targ},i}^{(k+1)} \leftarrow \omega \phi_{\text{targ},i}^{(k)} + (1 - \omega) \phi_i^{(k+1)}, \quad i = 1, 2$$

    Update $k \leftarrow k + 1$
  end if
end while