Theoretical Advances in Reinforcement Learning: Online Average-Reward and Offline Constrained Settings

by

Kihyuk Hong

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Statistics)
in the University of Michigan
2025

Doctoral Committee:

Professor Ambuj Tewari, Chair
Professor Yaacov Ritov
Professor Stilian Stoev
Professor Lei Ying

Kihyuk Hong

kihyukh@umich.edu

ORCID iD: 0009-0009-3549-1130

ACKNOWLEDGEMENTS

I am deeply grateful to my advisor, Ambuj Tewari, for his guidance and support throughout my PhD. He has been an exceptional mentor whose advice helped me navigate the most difficult parts of this journey. His influence has profoundly and permanently shaped the course of my life for the better.

I also want to thank the faculty and friends in the Department of Statistics for their support, especially during challenging times. Teaching as a graduate student instructor was a highlight of my PhD, and I am glad to have met so many thoughtful and curious students.

Lastly, I thank my family and friends for being part of my life whether from afar or in Ann Arbor. Their presence made my life fruitful, joyful, healthy, flavorful, and melodious.

TABLE OF CONTENTS

CHAPTER

LIST OF TABLES

TABLE

LIST OF APPENDICES

ABSTRACT

This dissertation presents theoretical advancements in reinforcement learning (RL), focusing on two key settings: online RL under the infinite-horizon average-reward criterion and offline constrained RL under partial data coverage.

The first part addresses the online setting, where the objective is to optimize long-run average rewards through interaction with the environment. A family of value-iteration-based algorithms is proposed by approximating the average-reward objective using a carefully tuned discounted surrogate. This part resolves an open problem by establishing a computationally efficient algorithm for linear Markov decision processes under a weak structural assumption. The proposed algorithm employs span-constrained value clipping and a decoupled planning strategy that mitigates statistical inefficiencies arising from the complexity of the function class.

The second part is motivated by safety-critical applications and studies the offline setting, where the agent must learn a policy from a fixed dataset without further interaction. Primal-dual algorithms are developed for both linear MDPs and general function-approximation regimes, based on the linear programming formulation of RL. These methods are oracle-efficient and provably sample-efficient under partial data coverage. Moreover, they extend to the constrained RL setting, where the policy must satisfy additional safety constraints defined by auxiliary reward signals and threshold levels.

Together, the contributions of this dissertation advance the theoretical foundations of reinforcement learning in settings that prioritize safety and long-term performance.

CHAPTER 1

Introduction

Reinforcement learning (RL) is a general framework for solving decision-making problems, and has been applied across a range of domains including robotics, healthcare, education, and industrial engineering. A common feature of these applications is that a decision-making agent encounters varying situations, and its actions influence future situations. The RL framework captures this dynamic by modeling the problem as an interaction between an agent and an environment, where the agent's actions affect the environment's state transitions. The role of the problem designer employing the RL framework is to construct a reward signal that incentivizes a reward-maximizing agent to exhibit the desired behavior. Given such a reward signal, the objective of the agent is to learn a policy that selects actions to maximize cumulative rewards through its interaction with the environment.

Within the reinforcement learning framework, different problem settings arise depending on the nature of the application. One important aspect is the choice of performance measure for the agent. In the finite-horizon formulation, the focus is on the agent's performance over a fixed, finite number of time steps. In contrast, the infinite-horizon formulation considers the agent's performance over an unbounded sequence of interactions. In this setting, there are broadly two approaches to aggregating rewards over time: one based on the discounted sum of rewards and the other based on the long-term average reward.

Another important aspect is the data collection scheme. In online reinforcement learning, data is collected through direct interaction with the environment. In this setting, the agent must balance exploration and exploitation: it needs to explore the environment to learn its dynamics while simultaneously exploiting current knowledge to maximize rewards. In offline reinforcement learning, by contrast, the agent is provided with a fixed, pre-collected dataset and must learn a policy from this data, without further interaction with the environment.

Finally, an additional aspect considered in this thesis is the presence of constraints. In the standard, unconstrained reinforcement learning formulation, the objective of the agent is to maximize cumulative reward. However, in safety-critical applications, the problem designer may wish to enforce constraints on the agent's behavior to ensure safe operation. Constrained reinforcement learning addresses this by formulating the problem as a constrained optimization problem, in which the goal is to find a policy that maximizes the primary reward subject to constraints imposed on auxiliary reward signals.

This thesis examines reinforcement learning problems that span the aforementioned aspects, with a focus on advances in algorithmic design and accompanying theoretical guarantees. The thesis is organized into two parts: the first part addresses online reinforcement learning with an infinite-horizon average-reward performance criterion, while the second part focuses on offline constrained reinforcement learning. The remainder of this chapter presents the necessary preliminaries for the mathematical formulation of these problems.

1.1 Markov Decision Processes

The reinforcement learning problem is typically modeled as Markov Decision Processes (MDPs) [Puterman, 2014], which capture the sequential interaction between an agent and an environment. In an MDP, the actions taken by the agent influence the evolution of the state of the environment, and the agent receives rewards based on the actions it takes. Formally, MDPs are defined by the tuple $(\mathcal{S}, \mathcal{A}, r, P, d_0)$ where $\mathcal{S}$ is the set of all possible states the environment can be in, $\mathcal{A}$ is the set of all actions the agent can take, $r : \mathcal{S} \times \mathcal{A} \to [0, 1]$ is the reward function that determines the reward received by the agent when taking an action in a state, $P : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ is the transition probability kernel that governs the transition of the state of the environment when the agent takes an action in a state. Finally, $d_0 \in \Delta(\mathcal{S})$ is the initial state distribution that governs the distribution of the initial state. Here, we use the notation $\Delta(\mathcal{X})$ to denote the probability distribution over the set $\mathcal{X}$.

Interaction Protocol In this thesis, we restrict our attention to the following interaction protocol between an agent and an environment. The protocol begins with the environment sampling an initial state from d_0. Then, at each time step, the agent selects an action, receives a reward determined by the reward function r, and the environment transitions to the next state sampled from the transition kernel P. Since the interaction proceeds without a fixed terminal time, this setting is referred to as the infinite-horizon setting.

Protocol 1: Infinite-Horizon Setting

```
Environment samples $s_1 \in \mathcal{S}$ from distribution $d_0$.
for $t = 1, 2, \dots$ do
    Agent takes action $a_t \in \mathcal{A}$.
    Agent receives reward $r_t = r(s_t, a_t)$.
    Environment transitions to the next state sampled from $s_{t+1} \sim P(\cdot|s_t, a_t)$.
```

Policy At each time step t, when the agent selects an action $a_t \in \mathcal{A}$, it has access to the full history $(s_1, a_1, \dots, s_{t-1}, a_{t-1}, s_t)$. Thus, in general, the agent may follow a policy that maps the history up to time t to a probability distribution over the action space $\mathcal{A}$. However, all the settings considered in this thesis admit the existence of a *stationary* policy that determines the action distribution solely based on the current state, ignoring the past history. Therefore, we restrict our attention to stationary policies.

A stationary policy is defined as a function $\pi : \mathcal{S} \to \Delta(\mathcal{A})$ that maps each state to a distribution over actions. Given a stationary policy π, the transition kernel P, and the initial state distribution d_0, the distribution over the trajectory $(s_1, a_1, s_2, a_2, \dots)$ is fully specified. We denote by P^π the probability measure induced by the interaction of π, P, and d_0, and by $\mathbb{E}^\pi$ the corresponding expectation.

1.1.1 Linear Markov Decision Processes

In real-world applications, the state space $\mathcal{S}$ may be extremely large or even infinite. In such scenarios, the learner must be able to generalize to states that were not encountered during training. To enable this generalization, it is necessary to impose structural assumptions on the underlying MDP.

A widely used assumption in the reinforcement learning theory literature is the *linear MDP* assumption [Jin et al., 2020]. This assumption introduces structure into the MDP by positing that the dynamics and rewards can be represented through a low-dimensional feature mapping over state-action pairs. The linear MDP framework enables generalization to unseen states by leveraging this shared feature representation.

The additional assumption made under the linear MDP setting is stated as follows.

Assumption 1 (Linear MDP [Jin et al., 2020]). *We assume that the transition and the reward functions can be expressed as a linear function of a known d-dimensional feature map* $\boldsymbol{\varphi} : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$ *such that for any* $(s, a) \in \mathcal{S} \times \mathcal{A}$*, we have*

$$r(s, a) = \langle \boldsymbol{\varphi}(s, a), \boldsymbol{\theta} \rangle, \quad P(s'|s, a) = \langle \boldsymbol{\varphi}(s, a), \boldsymbol{\mu}(s') \rangle$$

where $\boldsymbol{\mu}(\cdot) = (\mu_1(\cdot), \ldots, \mu_d(\cdot))$ is a vector of d unknown measures on $\mathcal{S}$ and $\boldsymbol{\theta} \in \mathbb{R}^d$ is a known parameter for the reward function.

As is commonly done in the literature on linear MDPs [Jin et al., 2020], we further assume, without loss of generality (see Wei et al. [2021] for justification), the following boundedness conditions:

$$\begin{aligned} \|\boldsymbol{\varphi}(s,a)\|_2 &\leq 1 \text{ for all } (s,a) \in \mathcal{S} \times \mathcal{A}, \\ \|\boldsymbol{\theta}\|_2 &\leq \sqrt{d}, \\ \|\boldsymbol{\mu}(\mathcal{S})\|_2 &\leq \sqrt{d}. \end{aligned} \tag{1.1}$$

1.2 Infinite-Horizon Discounted Setting

In the infinite-horizon discounted setting, the agent and the environment interacts under the infinite-horizon protocol (Protocol 1). The performance metric of a stationary policy π in this setting is the expected discounted sum of rewards defined as

$$J_\gamma(\pi) := \mathbb{E}^\pi \left[\sum_{t=1}^{\infty} \gamma^{t-1} r(s_t, a_t) \right]. \tag{1.2}$$

We define related quantities called value functions that condition on the initial state or state-action pair:

$$\begin{aligned} V_\gamma^\pi(s) &= \mathbb{E}^\pi \left[\sum_{t=1}^{\infty} \gamma^{t-1} r(s_t, a_t) | s_1 = s \right] \\ Q_\gamma^\pi(s,a) &= \mathbb{E}^\pi \left[\sum_{t=1}^{\infty} \gamma^{t-1} r(s_t, a_t) | s_1 = s, a_1 = a \right]. \end{aligned}$$

We write the optimal value functions under the discounted setting as

$$V_\gamma^*(s) = \max_\pi V_\gamma^\pi(s), \quad Q_\gamma^*(s,a) = \max_\pi Q_\gamma^\pi(s,a).$$

When clear from the context, we suppress the subscript γ and write V^π, Q^π, V^* and Q^*. The goal of RL in this setting is to find a nearly optimal policy that maximizes $J_\gamma^\pi(s)$ for each s.

1.3 Infinite-Horizon Average-Reward Setting

In the infinite-horizon average-reward setting, the agent and the environment interacts under the infinite-horizon protocol (Protocol 1). The performance metric in this setting is the expected average of rewards, defined as

$$J^{\pi}(s) := \liminf_{T\to\infty} \mathbb{E}^{\pi}\left[\frac{1}{T}\sum_{t=1}^{T} r(s_t, a_t) \mid s_0 = s\right].$$

The goal of RL in this setting is to find a nearly optimal policy that maximizes $J^{\pi}(s)$ for each s.

CHAPTER 2

Infinite-Horizon Average-Reward RL

2.1 Introduction

Among the various RL settings, the infinite-horizon setting is particularly well-suited for applications where optimizing long-term performance is the primary objective. Examples include production system management [Yang et al., 2021, Gosavi, 2004], inventory management [Gijsbrechts et al., 2022, Giannoccaro and Pontrandolfo, 2002] and network routing [Mammeri, 2019], where interactions between the agent and the environment continue indefinitely, and the natural goal is to optimize long-term rewards.

In the infinite-horizon framework, there are two widely-used definitions of long-term rewards. The first is the infinite-horizon discounted setting, where the objective is to maximize the discounted cumulative sum of rewards, with exponentially decaying weight assigned to future rewards. The second is the infinite-horizon average-reward setting, where the objective is to maximize the undiscounted long-term average of rewards, assigning uniform weight to future and present rewards. Learning in the average-reward setting is more challenging because its Bellman operator is not a contraction, and the widely used value iteration algorithm may fail when the transition probability model used for value iteration is not well-behaved. This complicates algorithm design, especially when the underlying transition probability model is unknown and must be estimated.

Seminal work by Auer et al. [2008] introduces a value iteration based algorithm for the infinite-horizon average-reward setting in the tabular case, where the state space and the action space are finite. To address sensitivity of the value iteration algorithm to the transition probability model, they maintain a confidence set that captures the true, well-behaved transition probability model. Their algorithm employs an extended value iteration approach, which optimally selects the transition probability model from the confidence set at each iteration. This extended value iteration method has since been extensively used in the tabular setting [Bartlett and Tewari, 2009, Fruit et al., 2018, Zhang and Ji, 2019]. Beyond the tabular case, the method has also been adapted to the linear mixture MDP setting [Modi et al., 2020, Ayoub et al., 2020], where the transition probability model has a low-dimensional structure [Ayoub et al., 2020, Wu et al., 2022, Chae et al., 2025].

The extended value iteration method is limited to tabular and linear mixture MDPs, as it relies on sample-efficient transition probability estimation, which is infeasible for settings like linear MDPs with large state spaces [Jin et al., 2020]. Due to these limitations, researchers have explored alternative approaches for such settings. For example, Wei et al. [2021] propose a reduction to the finite-horizon episodic setting by dividing the time steps into episodes of a fixed length. This approach achieves a regret bound of $\widetilde{\mathcal{O}}(T^{3/4})$, which is suboptimal, where T denotes the number of time steps. They also introduce a policy-based algorithm that alternates between policy evaluation and policy improvement steps to directly optimize the policy. This approach achieves an order-optimal regret bound of $\widetilde{\mathcal{O}}(\sqrt{T})$, but it requires a strong ergodicity assumption on the transition probability model for sample-efficient policy evaluation. Lastly, they propose another approach that achieves an order-optimal regret bound by directly solving the Bellman optimality equation as a fixed point problem, bypassing the need for value iteration. However, the fixed point problem is computationally intractable.

A recent work by Wei et al. [2020] on infinite-horizon average-reward RL uses a reduction to the discounted setting to leverage value iteration-based algorithms. They propose a Q-learning-based algorithm for the tabular setting that solves the discounted setting problem as a surrogate for the average-reward problem, achieving a regret bound of $\widetilde{\mathcal{O}}(T^{2/3})$. This chapter explores this idea further and presents an computationally efficient algorithm under the infinite-horizon average-reward setting with linear MDPs, that achieves a sharper regret bound of $\widetilde{\mathcal{O}}(\sqrt{T})$.

The chapter is organized as follows. Section 2.2 formalizes the performance measure under the infinite-horizon average-reward setting, and discusses structural assumptions necessary for nonvacuous result. Section 2.3 introduces the key technique of approximating the average-reward setting by the discounted setting, which is used throughout the chapter. Section 2.4 applies the technique to the tabular setting. Section 2.5 applies the technique to the linear MDP setting. Finally, Section 2.6 discusses a modification of the algorithm for computational efficiency.

2.1.1 Key References and Related Work

This chapter is based on the following papers written during my PhD.

1. Reinforcement learning for infinite-horizon average-reward linear MDPs via approxi-

mation by discounted-reward MDPs. AISTATS 2025. [Hong et al., 2025].

This paper proposes a new algorithm for online reinforcement learning under the infinite-horizon average-reward criterion. The central idea is to approximate the average-reward objective using a carefully tuned discounted surrogate, enabling the use of value-iteration-based methods. This paper develops algorithmic techniques and regret analysis for both the tabular setting and the linear MDP setting.

2. A computationally efficient algorithm for infinite-horizon average-reward linear MDPs. ICML 2025. [Hong and Tewari, 2025a].

   This paper improves the computational efficiency of the above approach in the linear MDP setting. It eliminates the dependency of computational complexity on the size of the state space by introducing a novel algorithmic trick and accompanying analysis.

3. Learning infinite-horizon average-reward linear mixture mdps of bounded span. AISTATS 2025. [Chae et al., 2025].

   This paper is a related work that applies a similar discounted-approximation technique to the setting of linear mixture MDPs. While closely related in spirit, this paper is not discussed in detail in the thesis for brevity.

2.2 Preliminaries

Online RL algorithm aims to find an optimal policy by interacting with the environment. Initially, without knowledge about the dynamics of the underlying MDP, the performance of the agent is bound to be bad. The performance of the agent would improve over time as the agent learns about the environment as it explores the MDP. The overall performance of an online RL algorithm employed by the agent under this setting is measured by the regret against the best stationary policy π^* that maximizes $J^\pi(s_1)$. Writing $J^*(s_1) := J^{\pi^*}(s_1)$, we define the T-step regret as

$$R_T := \sum_{t=1}^{T} (J^*(s_1) - r(s_t, a_t)).$$

Since the reward function takes values in the range $[0, 1]$, it follows that $J^*(s_1) \in [0, 1]$, and hence the regret is trivially bounded above by T. The question is whether there exists an algorithm with regret sublinear in T.

As discussed by Bartlett and Tewari [2009], without additional assumptions on the structure of the MDP, the agent may incur linear regret if it enters a state from which it is impossible to reach the states that yield high long-run rewards. This scenario suggests the need for structural assumptions that ensure the agent can eventually recover from visiting suboptimal states. A rather strong assumption that allows sublinear regret is the ergodicity assumption that requires the Markov chain induced by any policy to be ergodic, i.e., irreducible and aperiodic. With this assumption, the agent can always recover from a bad state by running any policy due to irreducibility.

A weaker assumption that still allows sublinear regret is the weakly communicating assumption, which asserts that the state space can be partitioned into a set of transient states and a set of communicating states. The states in the transient set are visited only finitely often with probability 1 under any policy, and they do not affect the long-term performance of any policy. For every pair of states in the communicating set, there exists a policy such that the agent can reach one state to the other in finite number steps, which allows agent to recover from a suboptimal state.

An even weaker assumption, popularized by Wei et al. [2021], is the following.

Assumption 2 (Bellman optimality equation)**.** *There exist $J^* \in \mathbb{R}$ and functions $v^* : \mathcal{S} \to \mathbb{R}$ and $q^* : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ such that for all $(s, a) \in \mathcal{S} \times \mathcal{A}$, we have*

$$J^* + q^*(s,a) = r(s,a) + [Pv^*](s,a)$$
$$v^*(s) = \max_{a \in \mathcal{A}} q^*(s,a).$$

As shown by Wei et al. [2021], under Assumption 2, the policy π^* that deterministically selects an action from $\operatorname{argmax}_a q^*(s,a)$ at each state $s \in \mathcal{S}$ is an optimal policy. Moreover, π^* always gives an optimal average reward $J^{\pi^*}(s_1) = J^*$ for all initial states $s_1 \in \mathcal{S}$. Since the optimal average reward is independent of the initial state, we can simply write the regret as $R_T = \sum_{t=1}^{T}(J^* - r(s_t, a_t))$. Functions $v^*(s)$ and $q^*(s,a)$ are the relative advantage of starting with s and (s,a) respectively. A problem with large $\mathrm{sp}(v^*)$ to be more difficult since starting with a bad state can be more disadvantageous. As is common in the literature [Bartlett and Tewari, 2009, Wei et al., 2020], we assume an upper bound H of the span $\mathrm{sp}(v^*)$ is known to the learner.

2.3 Approximation by Discounted Setting

The key idea of the algorithms presented in this chapter, motivated by Zhang and Xie [2023], is to approximate the infinite-horizon average-reward setting using the infinite-horizon *discounted* setting with a discount factor $\gamma \in [0, 1)$ that is carefully tuned. When γ is close to 1, the optimal policy for the discounted setting becomes nearly optimal for the average-reward setting. This is supported by a classical result from Puterman [2014], which states that the discounted cumulative reward of a stationary policy converges to its average reward as γ approaches 1.

The following lemma formally establishes the connection between the infinite-horizon average-reward setting and the discounted setting.

Lemma 1 (Lemma 2 in Wei et al. [2020])**.** *For any $\gamma \in [0, 1)$, the optimal value function V^*_γ for the infinite-horizon discounted setting with discounting factor γ satisfies*

(i) $sp(V^*) \leq 2sp(v^*)$ *and*

(ii) $|(1-\gamma)V^*(s) - J^*| \leq (1-\gamma)sp(v^*)$ *for all* $s \in \mathcal{S}$.

The lemma above implies that the difference between the optimal average reward J^* and the optimal discounted cumulative reward normalized by the factor $(1-\gamma)$ is small when γ is close to 1. Therefore, the policy that is optimal under the discounted setting can be expected to be nearly optimal for the average-reward setting, provided that the discount factor γ is chosen sufficiently close to 1.

2.4 Tabular Setting

In this section, we introduce an algorithm designed for the tabular setting, where the state space $\mathcal{S}$ and action space $\mathcal{A}$ are both finite, and no specific structure is assumed for the reward function or the transition probabilities. The structure of the algorithm, along with the accompanying analysis, will lay the groundwork for extending these results to the linear MDP setting.

2.4.1 Algorithm

Our algorithm, called *discounted upper confidence bound clipped value iteration* (γ-UCB-CVI), adapts UCBVI [Azar et al., 2017], which was originally designed for the finite-horizon episodic setting, to the infinite-horizon discounted setting. At each time step, the algorithm performs an approximate Bellman backup with an added bonus term $\beta\sqrt{1/N_t(s,a)}$ (Line 10) where $N_t(s,a)$ is the number of times the state-action pair (s,a) is visited. The bonus term is designed to guarantee optimism, ensuring that $Q_t \geq Q^*$ for all $t = 1, \ldots, T$. A key modification from UCBVI is the clipping step (Line 12), which bounds span of the value function estimate V_t by H, where the target span H is an input to the algorithm. Without clipping, the span of the value function V_t can be as large as $\frac{1}{1-\gamma}$, while with clipping, the span can only be as large as H. As we will see in the analysis, this clipping step is crucial to

Algorithm 2: γ-UCB-CVI for Tabular Setting

```
Input: Discounting factor $\gamma \in [0,1)$, span $H$, bonus factor $\beta$.
Initialize: $Q_1(s,a), V_1(s) \leftarrow \frac{1}{1-\gamma}$; $N_0(s,a,s') \leftarrow 0, N_0(s,a) \leftarrow 1$, for all
            $(s,a,s') \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}$.
Receive initial state $s_1$.
for time step $t = 1, \ldots, T$ do
    Take action $a_t = \operatorname{argmax}_a Q_t(s_t, a)$.
    Receive reward $r(s_t, a_t)$.
    Receive next state $s_{t+1}$.
    $N_t(s_t, a_t, s_{t+1}) \leftarrow N_{t-1}(s_t, a_t, s_{t+1}) + 1$
    $N_t(s_t, a_t) \leftarrow N_{t-1}(s_t, a_t) + 1$.
    (Other entries of $N_t$ remain the same as $N_{t-1}$.)
    $\widehat{P}_t(s'|s,a) \leftarrow N_t(s,a,s')/N_t(s,a), \quad \forall (s,a) \in \mathcal{S} \times \mathcal{A}$.
    $Q_{t+1}(s,a) \leftarrow (r(s,a) + \gamma[\widehat{P}_t V_t](s,a) + \beta/\sqrt{N_t(s,a)}) \wedge Q_t(s,a), \quad \forall (s,a) \in \mathcal{S} \times \mathcal{A}$
    $\widetilde{V}_{t+1}(s) \leftarrow (\max_a Q_{t+1}(s,a)) \wedge V_t(s), \quad \forall s \in \mathcal{S}$.
    $V_{t+1}(s) \leftarrow \widetilde{V}_{t+1}(s) \wedge (\min_{s'} \widetilde{V}_{t+1}(s') + H), \quad \forall s \in \mathcal{S}$.
```

achieving a sharp dependence on $\frac{1}{1-\gamma}$ in the regret bound, which enables the $\widetilde{O}(\sqrt{T})$ regret through tuning γ. Running the algorithm with the discounting factor set to $\gamma = 1 - 1/\sqrt{T}$ and the target span set to $H \geq 2 \cdot \mathrm{sp}(v^*)$ guarantees the following regret bound.

Theorem 1. *Under Assumption 2, there exists a constant $c > 0$ such that, for any fixed $\delta \in (0,1)$, if Algorithm 2 is run with $\gamma = 1 - \sqrt{1/T}$, $H \geq 2 \cdot sp(v^*)$, and $\beta = cH\sqrt{S\log(SAT/\delta)}$, then with probability at least $1-\delta$, the total regret is bounded by*

$$R_T \leq \mathcal{O}\left(H\sqrt{S^2AT\log(SAT/\delta)}\right).$$

In the theorem above, the constant c in the definition of β is specified in Lemma 2 in the following subsection. The resulting regret bound matches the best known regret bound for computationally efficient algorithms in this setting. A comprehensive comparison with previous work on infinite-horizon average-reward tabular MDPs is provided in Section 2.4.3.

An interesting direction for future work is to improve the regret bound by a factor of $\sqrt{S}$ through a refined analysis using Bernstein-type concentration inequalities, inspired by the refined analysis of UCBVI in Azar et al. [2017].

2.4.2 Regret Analysis

This section outlines the proof of Theorem 1, with the complete argument deferred to Appendix A.1. The key component of the analysis is the following concentration inequality.

Lemma 2. *Under the setting of Theorem 1, there exists a constant c such that for any fixed $\delta \in (0, 1)$, we have with probability at least $1 - \delta$ that*

$$|[(\widehat{P}_t - P)V_t](s, a)| \leq c \cdot H\sqrt{S \log(SAT/\delta)/N_t(s, a)}$$

for all $(s, a, t) \in \mathcal{S} \times \mathcal{A} \times [T]$.

Without clipping, the span of V_t would be $\frac{1}{1-\gamma}$ rather than H, causing the deviation term $[(\widehat{P}_t - P)V_t](s, a)$ to scale with $\frac{1}{1-\gamma}$ instead of H. Replacing the $\frac{1}{1-\gamma}$ factor with H via clipping is essential for achieving a regret bound of $\widetilde{\mathcal{O}}(\sqrt{T})$ when tuning γ.

In Theorem 1, the bonus factor parameter β is chosen based on the concentration bound established in the preceding lemma. This choice guarantees that the deviation term $[(\widehat{P}_t - P)V_t](s, a)$ is bounded by $\beta/\sqrt{N_t(s, a)}$, which corresponds to the bonus term used by the algorithm. With this result in place, the following optimism result can now be established.

Lemma 3 (Optimism)**.** *Under the setting of Theorem 1, we have with probability at least $1 - \delta$ that*

$$V_t(s) \geq V^*(s), \quad Q_t(s, a) \geq Q^*(s, a)$$

for all $(s, a, t) \in \mathcal{S} \times \mathcal{A} \times [T]$.

The proof uses an induction argument (e.g. Lemma 18 in Azar et al. [2017]) to show $\widetilde{V}_t(s) \geq V^*(s)$. To establish that the clipped value function V_t, no larger than $\widetilde{V}_t$ by design, still satisfies $V_t(s) \geq V^*(s)$, we use $\mathrm{sp}(V^*) \leq 2 \cdot \mathrm{sp}(v^*) \leq H$ (Lemma 1), which guarantees the clipping operation does not clip V_t below V^*.

Now consider the high probability events established in Lemma 2 and Lemma 3 to hold. By the value iteration step (Line 10) of Algorithm 2 and the concentration inequality in Lemma 2, it follows that for all $t = 2, \ldots, T$,

$$r(s_t, a_t) \geq Q_t(s_t, a_t) - \gamma[\widehat{P}_{t-1}V_{t-1}](s_t, a_t) - \beta/\sqrt{N_{t-1}(s_t, a_t)}.$$

Furthermore, by the optimism guarantee of Lemma 3, it holds that

$$V_t(s_t) \leq \widetilde{V}_t(s_t) \leq \max_a Q_t(s_t, a) = Q_t(s_t, a_t),$$

which implies

$$r(s_t, a_t) \geq V_t(s_t) - \gamma[PV_{t-1}](s_t, a_t) - 2\beta/\sqrt{N_{t-1}(s_t, a_t)}.$$

Hence, the regret can be bounded by

$$\begin{aligned} R_T &= \sum_{t=1}^{T}(J^* - r(s_t, a_t)) \\ &\le \sum_{t=2}^{T}(J^* - V_t(s_t) + \gamma[PV_{t-1}](s_t, a_t) + 2\beta/\sqrt{N_{t-1}(s_t, a_t)}) + \mathcal{O}(1), \end{aligned}$$

where the first inequality uses the fact that $J^* \le 1$, which can be decomposed into

$$\begin{aligned} &= \underbrace{\sum_{t=2}^{T}(J^* - (1-\gamma)V_t(s_t))}_{(a)} + \underbrace{\gamma\sum_{t=2}^{T}(V_{t-1}(s_{t+1}) - V_t(s_t))}_{(b)} \\ &\quad + \underbrace{\gamma\sum_{t=2}^{T}(PV_{t-1}(s_t, a_t) - V_{t-1}(s_{t+1}))}_{(c)} + 2\underbrace{\beta\sum_{t=2}^{T}1/\sqrt{N_{t-1}(s_t, a_t)}}_{(d)} + \mathcal{O}(1). \end{aligned}$$

The terms $(a), (b), (c), (d)$ can be bounded separately as follows.

Bounding Term (*a*) Using the optimism result (Lemma 3) that says $V_t(s) \ge V^*(s)$ for all $s \in \mathcal{S}$ and Lemma 1 that bounds $|J^* - (1-\gamma)V^*(s)|$ for all $s \in \mathcal{S}$,

$$\begin{aligned} \sum_{t=2}^{T}(J^* - (1-\gamma)V_t(s_t)) &\le \sum_{t=2}^{T}(J^* - (1-\gamma)V^*(s_t)) \\ &\le T(1-\gamma)\mathrm{sp}(v^*). \end{aligned}$$

Bounding Term (*b*) Note that for any $s \in \mathcal{S}$, the sequence $\{V_t(s)\}_{t=1}^{T}$ is monotonically decreasing due to Line 11-12 in Algorithm 2. Moreover, since $V_t(s) \in [0, \frac{1}{1-\gamma}]$ for all $t = 1, \ldots, T$, the total decrease in $V_t(s)$ from $t = 1$ to T is bounded above by $\frac{1}{1-\gamma}$. Hence,

$$\begin{aligned} \sum_{t=2}^{T}(V_{t-1}(s_{t+1}) - V_t(s_t)) &\le \sum_{t=2}^{T}(V_{t-1}(s_{t+1}) - V_{t+1}(s_{t+1})) + \mathcal{O}\left(\frac{1}{1-\gamma}\right) \\ &\le \sum_{s\in\mathcal{S}}\sum_{t=2}^{T}(V_{t-1}(s) - V_{t+1}(s)) + \mathcal{O}\left(\frac{1}{1-\gamma}\right) \\ &\le \mathcal{O}\left(\frac{S}{1-\gamma}\right). \end{aligned}$$

Bounding Term (*c*) Term (c) is the sum of a martingale difference sequence where each term is bounded by H. Hence, by the Azuma-Hoeffding inequality, with probability at least $1-\delta$, term (c) is bounded by $H\sqrt{2T\log(1/\delta)}$. Without clipping, each term of the martingale difference sequence can only be bounded by $\frac{1}{1-\gamma}$, leading to a bound of $\widetilde{\mathcal{O}}(\frac{1}{1-\gamma}\sqrt{T})$, which is too loose for achieving a regret bound of $\widetilde{\mathcal{O}}(\sqrt{T})$.

Bounding Term (*d*) We can bound the sum of the bonus terms (d) using a standard argument (Azar et al. [2017], Lemma 21 in Appendix A.1) by $\mathcal{O}(\beta\sqrt{SAT}) = \mathcal{O}(H\sqrt{S^2AT\log(SAT/\delta)})$.

Combining the above, and rescaling δ, it follows that with probability at least $1-\delta$,

$$R_T \leq \mathcal{O}\Big(T(1-\gamma)\text{sp}(v^*) + \frac{S}{1-\gamma} + H\sqrt{T\log(1/\delta)} + H\sqrt{S^2AT\log(SAT/\delta)}\Big).$$

Choosing $\gamma = 1 - 1/\sqrt{T}$ gives

$$R_T \leq \mathcal{O}\left(H\sqrt{S^2AT\log(SAT/\delta)}\right),$$

which completes the proof of Theorem 1.

2.4.3 Related Work

Seminal work by Auer et al. [2008] on infinite-horizon average-reward setting in tabular MDPs laid the foundation for the problem. Their model-based algorithm called UCRL2 constructs a confidence set on the transition model and run an extended value iteration that involves choosing the optimistic model in the confidence set each iteration. They achieve a regret bound of $\mathcal{O}(DS\sqrt{AT})$ where D is the diameter of the true MDP. Bartlett and Tewari [2009] improve the regret bound of UCRL2 by restricting the confidence set of the model to only include models such that the span of the induced optimal value function is bounded. Their algorithm, called REGAL, achieves a regret bound that scales with the span of the optimal value function $\text{sp}(v^*)$ instead of the diameter of the MDP. However, REGAL is computationally inefficient. Fruit et al. [2018] propose a model-based algorithm called SCAL, which is a computationally efficient version of REGAL. Zhang and Ji [2019] propose a model-based algorithm called EBF that achieves the minimax optimal regret of $\mathcal{O}(\sqrt{\text{sp}(v^*)SAT})$ by maintaining a tighter model confidence set by making use of the estimate for the optimal bias function. However, their algorithm is computationally inefficient. There is another line of work on model-free algorithms for this setting. Wei et al. [2020] introduce a model-free Q-learning-based algorithm called Optimistic Q-learning. Their algorithm is a

Table 2.1: Comparison of algorithms for infinite-horizon average-reward RL in tabular setting

Algorithm	Regret $\widetilde{\mathcal{O}}(\cdot)$	Assumption
UCRL2 [Auer et al., 2008]	$DS\sqrt{AT}$	Bounded diameter
† REGAL [Bartlett and Tewari, 2009]	$\mathrm{sp}(v^*)\sqrt{SAT}$	Weakly communicating
PSRL [Ouyang et al., 2017]	$\mathrm{sp}(v^*)S\sqrt{AT}$	Weakly communicating
† OSP [Ortner, 2020]	$\sqrt{t_{\mathrm{mix}}SAT}$	Ergodic
SCAL [Fruit et al., 2018]	$\mathrm{sp}(v^*)S\sqrt{AT}$	Weakly communicating
UCRL2B [Fruit et al., 2020]	$S\sqrt{DAT}$	Bounded diameter
† EBF [Zhang and Ji, 2019]	$\sqrt{\mathrm{sp}(v^*)SAT}$	Weakly communicating
Optimistic Q-learning [Wei et al., 2020]	$\mathrm{sp}(v^*)(SA)^{\frac{1}{3}}T^{\frac{2}{3}}$	Weakly communicating
MDP-OOMD [Wei et al., 2020]	$\sqrt{t_{\mathrm{mix}}^3\eta AT}$	Ergodic
UCB-AVG [Zhang and Xie, 2023]	$\mathrm{sp}(v^*)S^5A^2\sqrt{T}$	Weakly communicating
γ-UCB-CVI [Hong et al., 2025]	$\mathrm{sp}(v^*)S\sqrt{AT}$	Bellman optimality equation
Lower bound [Auer et al., 2008]	$\Omega(\sqrt{DSAT})$	

reduction to the discounted setting. Although model-free, their algorithm has a suboptimal regret of $\mathcal{O}(T^{2/3})$. Recently, Zhang and Xie [2023] introduce a Q-learning-based algorithm called UCB-AVG that achieves regret bound of $\mathcal{O}(\sqrt{T})$. Their algorithm, which is also a reduction to the discounted setting, is the first model-free to achieve the order optimal regret bound. Their main idea is to use the optimal bias function estimate to increase statistical efficiency. Agrawal and Agrawal [2024] introduces a model-free Q-learning-based algorithm and provides a unified view of episodic setting and infinite-horizon average-reward setting. However, their algorithm requires additional assumption of the existence of a state with bounded hitting time.

Comparison of algorithms for infinite-horizon average-reward RL in the tabular setting can be see in Table 2.1. The sign † indicates that the corresponding algorithm is computationally inefficient.

2.5 Linear MDP Setting

In this section, the key ideas developed in the previous section are applied to the linear MDP setting (see Section 1.1.1 for the definition). As discussed by Jin et al. [2020], although the transition model P is linear in the d-dimensional feature mapping $\boldsymbol{\varphi}$, the transition kernel P retains infinite degrees of freedom due to the unknown measure $\boldsymbol{\mu}$, making the estimation of P challenging.

For sample-efficient learning, the key observation is that $[Pv](s,a)$ is linear in $\boldsymbol{\varphi}(s,a)$ for any function $v : \mathcal{S} \to \mathbb{R}$. That is, there exists a vector $\boldsymbol{w}_v^* \in \mathbb{R}^d$ such that

$$[Pv](s,a) = \langle \boldsymbol{\varphi}(s,a), \boldsymbol{w}_v^* \rangle,$$

since

$$\begin{aligned}[Pv](s,a) :&= \int_{s' \in \mathcal{S}} v(s') P(ds' \mid s,a) \\ &= \int_{s' \in \mathcal{S}} v(s') \langle \boldsymbol{\varphi}(s,a), \boldsymbol{\mu}(ds') \rangle \\ &= \langle \boldsymbol{\varphi}(s,a), \int_{s' \in \mathcal{S}} v(s') \boldsymbol{\mu}(ds') \rangle.\end{aligned}$$

This representation allows the estimation of $[Pv](s,a)$ via linear regression, circumventing the need to estimate the full transition kernel P.

Exploiting the linearity, for a function $V : \mathcal{S} \to \mathbb{R}$, we can estimate $\boldsymbol{w}^*(V)$ given a trajectory data $(s_1, a_1, \ldots, s_{t-1}, a_{t-1}, s_t)$ via linear regression as follows:

$$\widehat{\boldsymbol{w}}_t(V) := \Lambda_t^{-1} \sum_{\tau=1}^{t-1} V(s_{\tau+1}) \cdot \boldsymbol{\varphi}(s_\tau, a_\tau)$$

where $\Lambda_t = \lambda I + \sum_{\tau=1}^{t-1} \boldsymbol{\varphi}(s_t, a_t) \boldsymbol{\varphi}(s_t, a_t)^\top$. With such a regression coefficient, $[PV](s,a)$ can be estimated by

$$[\widehat{P}_t V](s,a) := \langle \boldsymbol{\varphi}(s,a), \widehat{\boldsymbol{w}}_t(V - V(s_1)) \rangle + V(s_1).$$

We estimate $[PV](s,a)$ by estimating $[P(V - V(s_1))](s,a)$ and then adding back $V(s_1)$. This allows bounding the norm of the regression coefficient $\|\widehat{\boldsymbol{w}}_t(V - V(s_1))\|_2$ by a bound that scales with the span of V instead of the magnitude of V, which is required for getting a sharp regret bound.

Naively adapting the algorithm design and analysis for the tabular setting to the linear MDP setting would result in a regret bound that is polynomial in S, the size of the state space, when bounding $\sum_{t=1}^{T} (V_{t-1}(s_{t+1}) - V_t(s_t))$. Also, algorithmically making the state value function monotonically decrease in t by taking minimum with the previous estimate every iteration, as is done in the tabular setting for the telescoping sum argument, would lead to an exponential covering number for the function class of the value function, in either T or S [He et al., 2023]. A major challenge in algorithm design and analysis is sidestepping these issues. The algorithm presented in the next section for the linear MDP setting addresses these issues.

```
Algorithm 3: $\gamma$-LSCVI-UCB
   Input: Discounting factor $\gamma \in (0,1)$, regularization $\lambda > 0$, span $H$, bonus factor $\beta$.
   Initialize: $t \leftarrow 1$, $k \leftarrow 1$, $t_k \leftarrow 1$, $\Lambda_1 \leftarrow \lambda I$, $\bar{\Lambda}_0 \leftarrow \lambda I$, $Q^1_t(\cdot,\cdot) \leftarrow \frac{1}{1-\gamma}$ for $t \in [T]$.
1  Receive state $s_1$.
2  for time step $t = 1, \dots, T$ do
3      Take action $a_t = \mathrm{argmax}_a\, Q^k_t(s_t, a)$. Receive reward $r(s_t, a_t)$. Receive next state $s_{t+1}$.
4      $\bar{\Lambda}_t \leftarrow \bar{\Lambda}_{t-1} + \boldsymbol{\varphi}(s_t, a_t)\boldsymbol{\varphi}(s_t, a_t)^T$.
5      if $2\det(\Lambda_k) < \det(\bar{\Lambda}_t)$ then
6          $k \leftarrow k + 1$, $t_k \leftarrow t + 1$, $\Lambda_k \leftarrow \bar{\Lambda}_t$.
           // Run value iteration to plan for remaining $T - t_k + 1$ time steps in the new episode.
7          $\widetilde{V}^k_{T+1}(\cdot) \leftarrow \frac{1}{1-\gamma}$, $V^k_{T+1}(\cdot) \leftarrow \frac{1}{1-\gamma}$.
8          for $u = T, T-1, \dots, t_k$ do
9              $\boldsymbol{w}^k_{u+1} \leftarrow \Lambda_k^{-1} \sum_{\tau=1}^{t_k-1} \boldsymbol{\varphi}(s_\tau, a_\tau)(V^k_{u+1}(s_{\tau+1}) - V^k_{u+1}(s_1)))$.
10             $Q^k_u(\cdot,\cdot) \leftarrow \Big(r(\cdot,\cdot) + \gamma(\langle \boldsymbol{\varphi}(\cdot,\cdot), \boldsymbol{w}^k_{u+1}\rangle + V^k_{u+1}(s_1) + \beta\|\boldsymbol{\varphi}(\cdot,\cdot)\|_{\Lambda_k^{-1}}\Big) \wedge \frac{1}{1-\gamma}$.
11             $\widetilde{V}^k_u(\cdot) \leftarrow \max_a Q^k_u(\cdot, a)$.
12             $V^k_u(\cdot) \leftarrow \widetilde{V}^k_u(\cdot) \wedge (\min_{s'} \widetilde{V}^k_u(s') + H)$.
```

2.5.1 Algorithm

The algorithm presented in this section, referred to as *discounted least-squares clipped value iteration with upper confidence bound* (γ-LSCVI-UCB), extends the LSVI-UCB algorithm of Jin et al. [2020], originally developed for the finite-horizon episodic setting, to the infinite-horizon discounted setting. The main modifications introduced to adapt to the discounted setting are summarized below.

Clipping the Value Function The value function estimates are clipped to restrict their span (Line 12), following the same strategy employed in the tabular case discussed in the previous section. This clipping operation enforces a bounded span, which is critical in controlling the bias arising from approximation and enables a regret bound with improved dependence on the discount factor, saving a factor of $1/(1-\gamma)$.

In the previous algorithm, γ-UCB-CVI, designed for the tabular setting, the value iteration procedure is interleaved with the decision-making process. Specifically, at each time step t, the agent selects an action greedily with respect to the most recently updated action-value function Q_t. This design structure is common across value iteration-based and Q-learning-based algorithms in both infinite-horizon average-reward tabular MDPs [Zhang and Xie, 2023] and infinite-horizon discounted tabular MDPs [Liu and Su, 2020, He et al., 2021]. Under this interleaved design, the value function sequence $\{V_t\}$ is algorithmically required to be monotonic in order to control the telescoping sum $\sum_t (V_{t-1}(s_{t+1}) - V_t(s_t))$.

However, in the linear MDP setting, enforcing monotonicity of $\{V_t\}$ by defining $V_t := \min(V_{t-1}, V_t)$ would cause the complexity of the function class from which value estimates are drawn to grow with T. In particular, the logarithm of the covering number of the function class would scale with T, leading to vacuous regret bounds.

To avoid this issue, a novel algorithmic structure is adopted that decouples the value iteration updates from the decision-making process. This structural separation allows the algorithm to maintain a fixed function class and leverage concentration tools without incurring complexity penalties tied to time horizon T.

Planning until the End of Horizon Prior to executing any action at time t, the algorithm performs a sequence of $T - t$ value iteration steps to construct a series of action-value functions $Q_T, Q_{T-1}, \ldots, Q_t$ (Lines 7–12). At each decision point t, the agent then selects a greedy action with respect to Q_t. This algorithmic structure resembles that of value iteration-based algorithms for the finite-horizon episodic setting [Azar et al., 2017, Jin et al., 2020], in which Bellman updates are used to generate action-value functions corresponding to each time step of a fixed-length episode, and actions are chosen greedily according to these functions throughout the episode. Under this structure, Q_{t-1} becomes one Bellman update ahead of Q_t. As a result, the key quantity of interest becomes $\sum_{t=1}^{T} V_{t+1}(s_{t+1}) - V_t(s_t)$, which can now be bounded by telescoping sum.

Restarting when Information Doubles If all T action-value functions are generated at the initial time step via approximate value iteration and subsequently used for decision-making over T steps, the collected trajectory data cannot be leveraged for improving future decisions. To address this limitation while retaining the pregeneration scheme, we periodically restart the value iteration process based on the growth of an information measure derived from the observed data. Specifically, we initiate a new sequence of value iterations whenever a chosen information criterion doubles, enabling the algorithm to incorporate newly collected trajectory data into future planning. To implement this mechanism, we adopt the rarely-switching covariance matrix technique introduced by Wang et al. [2021]. This approach triggers a restart when the determinant of the empirical covariance matrix doubles (Line 5).

The algorithm γ-LSCVI-UCB, based on the design discussed above, has the following guarantee.

Theorem 2. *Under Assumptions 2 and 1, running Algorithm 3 with inputs* $\gamma = 1 - \sqrt{\log(T)/T}$*,* $\lambda = 1$*,* $H \geq 2 \cdot sp(v^*)$ *and* $\beta = 2c_\beta \cdot Hd\sqrt{\log(dT/\delta)}$ *guarantees with probability at least* $1 - \delta$*,*

$$R_T \leq \mathcal{O}(H\sqrt{d^3 T \log(dT/\delta)}).$$

The constant c_β is an absolute constant defined in Lemma 4. A refined analysis of the variance of the value estimate [He et al., 2023] may improve our regret by a factor of $\sqrt{d}$, which would be an interesting future work.

2.5.2 Regret Analysis

This section outlines the proof of the regret bound presented in Theorem 2. The first step is to show that the value iteration step in Line 10 with the bonus term $\beta\|\boldsymbol{\phi}(\cdot,\cdot)\|_{\Lambda_k^{-1}}$ with appropriately chosen β ensures the value function estimates V_t and Q_t to be optimistic estimates of V^* and Q^*, respectively. The argument is based on the following concentration inequality for the regression coefficients. See Appendix A.2.1 for a proof.

Lemma 4 (Concentration of regression coefficients). *With probability at least* $1 - \delta$*, there exists an absolute constant* c_β *such that for* $\beta = c_\beta \cdot Hd\sqrt{\log(dT/\delta)}$*, we have*

$$|\langle \boldsymbol{\phi}, \boldsymbol{w}_u^k - \boldsymbol{w}_u^{k*} \rangle| \leq \beta \|\boldsymbol{\phi}\|_{\Lambda_k^{-1}}$$

for all episode indices k *and for all vectors* $\boldsymbol{\phi} \in \mathbb{R}^d$ *where* $\boldsymbol{w}_u^{k*} := \int (V_u^k(s) - V_u^k(s_1)) d\boldsymbol{\mu}(s)$ *is a parameter that satisfies* $\langle \boldsymbol{\varphi}(s,a), \boldsymbol{w}_u^{k*} \rangle = [PV_u^k](s,a) - V_u^k(s_1)$*.*

With the concentration inequality, we can show the following optimism result. See Appendix A.2.2 for an induction-based proof.

Lemma 5 (Optimism). *Under the linear MDP setting, running Algorithm 3 with input* $H \geq 2 \cdot sp(v^*)$ *guarantees with probability at least* $1 - \delta$ *that for all episodes* $k = 1, 2, \ldots,$ $u = t_k, \ldots, T+1$ *and for all* $(s,a) \in \mathcal{S} \times \mathcal{A}$*, we have*

$$V_u^k(s) \geq V^*(s), \quad Q_u^k(s,a) \geq Q^*(s,a).$$

Now, we show the regret bound under the event that the high probability events in the previous two lemmas (Lemma 4, Lemma 5) hold. Let t be a time step in episode k such that both t and $t+1$ are in episode k. By the definition of $Q_u^k(\cdot,\cdot)$ (Line 10), we have for all

$t = t_k, \dots, T+1$ and $(s,a) \in \mathcal{S} \times \mathcal{A}$ that

$$\begin{aligned} r(s,a) &\geq Q_t^k(s,a) - \gamma(\langle \boldsymbol{\varphi}(s,a), \boldsymbol{w}_{t+1}^k \rangle + \min_{s'} V_{t+1}^k(s') - \beta \|\boldsymbol{\varphi}(s,a)\|_{\Lambda_k^{-1}}) \\ &\geq Q_t^k(s,a) - \gamma[PV_{t+1}^k](s,a) - 4\beta \|\boldsymbol{\varphi}(s,a)\|_{\bar{\Lambda}_t^{-1}} \end{aligned}$$

where the second inequality uses the concentration bound for the regression coefficients in Lemma 4. It also uses $\|\boldsymbol{x}\|_{\Lambda_k^{-1}} \leq 2\|\boldsymbol{x}\|_{\Lambda_t^{-1}}$ (Lemma 30). Hence, we can bound the regret in episode k by

$$\begin{aligned} R^k &= \sum_{t=t_k}^{t_{k+1}-1} (J^* - r(s_t,a_t)) \\ &\leq \sum_{t=t_k}^{t_{k+1}-1} (J^* - Q_t^k(s_t,a_t) + \gamma[PV_{t+1}^k](s_t,a_t) + 4\beta\|\boldsymbol{\varphi}(s_t,a_t)\|_{\bar{\Lambda}_t^{-1}}), \end{aligned}$$

which can be decomposed into

$$\begin{aligned} = &\underbrace{\sum_{t=t_k}^{t_{k+1}-1} (J^* - (1-\gamma)V_{t+1}^k(s_{t+1}))}_{(a)} + \gamma \underbrace{\sum_{t=t_k}^{t_{k+1}-1} (V_{t+1}^k(s_{t+1}) - Q_t^k(s_t,a_t))}_{(b)} \\ &+ \gamma \underbrace{\sum_{t=t_k}^{t_{k+1}-1} [PV_{t+1}^k](s_t,a_t) - V_{t+1}^k(s_{t+1}))}_{(c)} + 4\beta \underbrace{\sum_{t=t_k}^{t_{k+1}-1} \|\boldsymbol{\varphi}(s_t,a_t)\|_{\bar{\Lambda}_t^{-1}}}_{(d)} \end{aligned}$$

where the first inequality uses the bound for $r(s_t, a_t)$. With the same argument as in the tabular case, the term (a) summed over all episodes can be bounded by $T(1-\gamma)\mathrm{sp}(v^*)$ using the optimism $V_u^k(s_{t+1}) \geq V^*(s_{t+1})$, and Lemma 1 that bounds $|J^* - (1-\gamma)V^*(s)|$ for all $s \in \mathcal{S}$. Term (d), summed over all episodes, can be bounded by $\mathcal{O}(\beta\sqrt{dT\log T})$ using Cauchy-Schwartz and Lemma 29. Term (c), summed over all episodes, is a sum of a martingale difference sequence, which can be bounded by $\mathcal{O}(\mathrm{sp}(v^*)\sqrt{T\log(1/\delta)})$ since $\mathrm{sp}(V_u^k) \leq 2 \cdot \mathrm{sp}(v^*)$ by the clipping step in Line 12.

Bounding Term (*b*) To bound term (b) note that

$$\begin{aligned} V^k_{t+1}(s_{t+1}) &\leq \widetilde{V}^k_{t+1}(s_{t+1}) \\ &= \max_a Q^k_{t+1}(s_{t+1}, a) \\ &= Q^k_{t+1}(s_{t+1}, a_{t+1}) \end{aligned}$$

as long as the time step $t+1$ is in episode k, since the algorithm chooses a_{t+1} that maximizes $Q^k_{t+1}(s_{t+1}, \cdot)$. Hence,

$$\begin{aligned} \sum_{t=t_k}^{t_{k+1}-1} (V^k_{t+1}(s_{t+1}) - Q^k_t(s_t, a_t)) &\leq \frac{1}{1-\gamma} + \sum_{t=t_k}^{t_{k+1}-2} (Q^k_{t+1}(s_{t+1}, a_{t+1}) - Q^k_t(s_t, a_t)) \\ &\leq \mathcal{O}\left(\frac{1}{1-\gamma}\right) \end{aligned}$$

where the second inequality uses telescoping sum and the fact that $Q^k_t \leq \frac{1}{1-\gamma}$. Since the episode is advanced when the determinant of the covariance matrix doubles, it can be shown that the number of episodes is bounded by $\mathcal{O}(d \log(T))$ (Lemma 31). Combining all the bounds, and using $\beta = \mathcal{O}(\mathrm{sp}(v^*) d \sqrt{\log(dT/\delta)}$, we get

$$R_T \leq \mathcal{O}\Big(T(1-\gamma)\mathrm{sp}(v^*) + \frac{d}{1-\gamma} \log(T) + H\sqrt{T \log(1/\delta)} + H\sqrt{d^3 T \log(dT/\delta)}\Big).$$

Setting $\gamma = 1 - \sqrt{(\log T)/T}$, we get

$$R_T \leq \mathcal{O}\left(H\sqrt{d^3 T \log(dT/\delta)}\right),$$

which concludes the proof of Theorem 2.

2.5.3 Computational Complexity

The algorithm γ-LSCVI-UCB runs in episodes and since a new episode starts only when the determinant of the covariance matrix Λ_t doubles, there can be at most $\mathcal{O}(d \log_2 T)$ episodes (see Lemma 31). In each episode, we run at most T value iterations. In each iteration step u, the algorithm computes $\min_{s'} \widetilde{V}^t_u(s')$ which requires evaluating $\widetilde{V}^t_u(s')$ at all $s' \in \mathcal{S}$, which requires $\mathcal{O}(d^2 SA)$ computations. Also, the algorithm computes $\boldsymbol{w}^k_{u+1}$, which requires $\mathcal{O}(d^2 + Td)$ operations. All other operations runs in $\mathcal{O}(d^2 + A)$ per value iteration. In total, the algorithm runs in $\mathcal{O}((\log_2 T) d^3 SAT^2)$. See Appendix A.2.3 for detailed analysis.

The FOPO algorithm by Wei et al. [2021] that matches the regret bound under the same set of assumptions, has a time complexity of $\mathcal{O}(T^d \log_2 T)$. Although the time complexity of γ-LSCVI-UCB is an improvement over previous work in the sense that the time complexity is polynomial in problem parameters, it has linear dependency on S. The dependency on S arises from taking the minimum of value functions for clipping. The algorithm presented in the next section gets rid of this dependency by employing an efficient clipping.

2.6 Linear MDP Setting: Computational Efficiency

This section presents the algorithm called *discounted Deviation Controlled Least Squares Clipped Value Iteration with Upper Confidence Bound* (γ-DC-LSCVI-UCB, Algorithm 4), which improves computational complexity of the algorithm presented in the previous section.

2.6.1 Computationally Efficient Clipping

The algorithm design for computational efficiency is centered around bounding the term

$$\sum_{t=1}^{T-1} V_{t+1}^t(s_{t+1}) - \widetilde{V}_{t+1}^{t+1}(s_{t+1}),$$

where $\{\widetilde{V}_u^t\}_{u\in[t:T]}$ is the sequence of value functions generated at time step t, and $\{V_u^t\}_{u\in[t:T]}$ is the sequence of clipped value functions generated at time step t. Note that the clipped value function V_{t+1}^t in the summation is generated at time step t, prior to observing the next state s_{t+1}. With unlimited compute power, the γ-LSCVI-UCB algorithm by previous work uses $\min_{s\in\mathcal{S}} \widetilde{V}_{t+1}^t(s)$ as the clipping threshold, which allows bounding V_{t+1}^t evaluated at s_{t+1} by

$$\begin{aligned} V_{t+1}^t(s_{t+1}) &= \textsc{Clip}(\widetilde{V}_{t+1}^t(s_{t+1}); \min_{s\in\mathcal{S}} \widetilde{V}_{t+1}^t(s), \min_{s\in\mathcal{S}} \widetilde{V}_{t+1}^t(s) + H) \\ &\leq \widetilde{V}_{t+1}^t(s_{t+1}) \end{aligned}$$

where the inequality only holds because $\min_{s\in\mathcal{S}} \widetilde{V}_{t+1}^t(s) \leq \widetilde{V}_{t+1}^t(s_{t+1})$. The algorithm γ-LSCVI-UCB also reuses the sequence of value functions most of the time steps, such that $\widetilde{V}_{t+1}^t(s_{t+1}) = \widetilde{V}_{t+1}^{t+1}(s_{t+1})$, allowing the bound $V_{t+1}^t(s_{t+1}) - \widetilde{V}_{t+1}^{t+1}(s_{t+1}) \leq 0$.

For computational efficiency, suppose we use m_t as the clipping threshold instead of $\min_{s\in\mathcal{S}} \widetilde{V}_{t+1}^t(s)$, where m_t is computed using states $s_1, \ldots, s_t$ only. Then, the bound $V_{t+1}^t(s_{t+1}) \leq \widetilde{V}_{t+1}^t(s_{t+1})$ may no longer hold because

$$V_{t+1}^t(s_{t+1}) = \textsc{Clip}(\widetilde{V}_{t+1}^t(s_{t+1}); m_t, m_t + H) \geq m_t$$

and we may have $m_t > \widetilde{V}^t_{t+1}(s_{t+1})$ since we cannot look ahead s_{t+1} when choosing the clipping threshold m_t. We can instead get a bound with an error term:

$$\begin{aligned} V^t_{t+1}(s_{t+1}) &= \textsc{Clip}(\widetilde{V}^t_{t+1}(s_{t+1}); m_t, m_t + H) \\ &\le \widetilde{V}^t_{t+1}(s_{t+1}) + \max\{m_t - \widetilde{V}^t_{t+1}(s_{t+1}), 0\}. \end{aligned}$$

One key idea of handling the sum of the error terms is to choose $m_{t+1} = \widetilde{V}^t_{t+1}(s_{t+1}) \wedge m_t$ (Line 14), leading to

$$V^t_{t+1}(s_{t+1}) \le \widetilde{V}^t_{t+1}(s_{t+1}) + \Delta_t$$

where $\Delta_t = m_t - m_{t+1}$. Then the sum of the errors Δ_t can then be bounded using a telescoping sum.

The clipping threshold $m_{t+1} = \widetilde{V}^t_{t+1}(s_{t+1}) \wedge m_t$ may change every time step. Hence, after advancing to the next time step $t+1$ and computing the new threshold m_{t+1}, the algorithm computes Q^{t+1}_{t+1} afresh, which involves generating a sequence of value functions $V^{t+1}_T, \dots, V^{t+1}_{t+1}$ by running clipped value iteration with the new threshold m_{t+1}. Therefore, unlike previous work that ensures $\widetilde{V}^t_{t+1}(s_{t+1}) = \widetilde{V}^{t+1}_{t+1}(s_{t+1})$ by reusing the sequence of value functions, we need to control the difference between $\widetilde{V}^t_{t+1}(s_{t+1})$ and $\widetilde{V}^{t+1}_{t+1}(s_{t+1})$ to be able to bound

$$V^t_{t+1}(s_{t+1}) \le \widetilde{V}^t_{t+1}(s_{t+1}) + \Delta_t \approx \widetilde{V}^{t+1}_{t+1}(s_{t+1}) + \Delta_t.$$

The next section discusses the algorithm design for ensuring $\widetilde{V}^t_{t+1} \approx \widetilde{V}^{t+1}_{t+1}$.

2.6.2 Deviation-Controlled Value Iteration

Previous discussion suggests we need to bound the difference between sequences of value functions $\{\widetilde{V}^t_u\}_{u\in[T]}$ and $\{\widetilde{V}^{t+1}_u\}_{u\in[T]}$ generated by value iterations using different clipping thresholds m_t and m_{t+1}. We would expect that the difference between sequences of value functions to be bounded by the difference in clipping thresholds $m_t - m_{t+1}$. Surprisingly, a naive adaptation of the previous work γ-LSCVI-UCB, fails to control the difference. To see this, consider the following clipped value iteration procedure that generates a sequence of value functions $\{\widetilde{V}^t_u\}_u$ at time step t using the clipping threshold m_t.

We argue that controlling the difference $\|\widetilde{V}^t_{u+1} - \widetilde{V}^{t+1}_{u+1}\|_\infty \le \Delta$ for $\Delta = m_t - m_{t+1}$ at value iteration index $u+1$ does not necessarily control the difference $\|\widetilde{V}^t_u - \widetilde{V}^{t+1}_u\|_\infty$ at the next value iteration. To see this, suppose $\|\widetilde{V}^t_{u+1} - \widetilde{V}^{t+1}_{u+1}\|_\infty \le \Delta$. Then, by value iteration, we have

$$\|\widetilde{V}^t_u - \widetilde{V}^{t+1}_u\|_\infty \le \|Q^t_u - Q^{t+1}_u\|_\infty \approx \|\widehat{P}_t(V^t_{u+1} - V^{t+1}_{u+1})\|_\infty.$$

```
$V^t_{T+1}(\cdot) \leftarrow \frac{1}{1-\gamma}$.
for $u = T, T-1, \ldots, t$ do
    $Q^t_u(\cdot,\cdot) \leftarrow \Big(r(\cdot,\cdot) + \gamma([\widehat{P}_t V^t_{u+1}](\cdot,\cdot) + \beta\|\boldsymbol{\varphi}(\cdot,\cdot)\|_{\Lambda_t^{-1}})\Big) \wedge \frac{1}{1-\gamma}$.
    $\widetilde{V}^t_u(\cdot) \leftarrow \max_a Q^t_u(\cdot, a)$.
    $V^t_u(\cdot) \leftarrow \textsc{Clip}(\widetilde{V}^t_u(\cdot); m_t, m_t + H)$.
```

It is natural to expect that $\|V^t_{u+1} - V^{t+1}_{u+1}\|_\infty \leq \Delta$ would imply $\|\widehat{P}_t(V^t_{u+1} - V^{t+1}_{u+1})\|_\infty \leq \Delta$. This is true when $[\widehat{P}_t V](s,a)$ is an expectation of $V(\cdot)$ with respect to an empirical probability distribution $\widehat{P}_t(\cdot|\cdot,\cdot)$, which is the case for the tabular setting (see Appendix A.3.3.1 for more discussion). However, in the linear MDP setting, and more generally in general value function approximation setting, $[\widehat{P}_t V](s,a)$ is defined through a regression: $[\widehat{P}_t V](s,a) = \langle \boldsymbol{\varphi}(s,a), \widehat{\boldsymbol{w}}_t(V^t_{u+1} - V^{t+1}_{u+1})\rangle$, which can be arbitrarily larger than Δ as shown in the next lemma.

Lemma 6. *There exist $\boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_n \in \mathbb{R}^d$ with $\|\boldsymbol{\phi}_i\| \leq 1$ for $i = 1, \ldots, n$, and $y_1, \ldots, y_n \in \mathbb{R}$ with $|y_i| \leq \Delta$, $i = 1, \ldots, n$ for any $\Delta > 0$, such that*

$$|\langle \boldsymbol{w}_n, \boldsymbol{\phi}\rangle| \geq \frac{1}{2}\Delta\sqrt{n}$$

for some $\boldsymbol{\phi} \in \mathbb{R}^d$ where $\boldsymbol{w}_n$ is the regression coefficient $\boldsymbol{w}_n = \Lambda_n^{-1}\sum_{i=1}^n y_i \boldsymbol{\phi}_i$ where $\Lambda_n = \sum_{i=1}^n \boldsymbol{\phi}_i \boldsymbol{\phi}_i^\top + \lambda I$.

To address this issue, we propose a novel value iteration procedure that explicitly controls the deviation of a sequence of value functions from its previous sequences. The key idea is to clip the value function $\widetilde{Q}^t_u$ so that its values do not deviate too much from value functions $\widetilde{Q}^{t-1}_u$ and $\widetilde{Q}^{t-2}_u$ from previously generated sequences of value functions (Line 6-8). With this scheme, we can bound the difference between $\widetilde{V}^t_u$ and $\widetilde{V}^{t+1}_u$ as follows.

Lemma 7. *When running γ-DC-LSCVI-UCB (Algorithm 4), we have*

$$|\widetilde{V}^{t+1}_u(s) - \widetilde{V}^t_u(s)| \leq m_{t-1} - m_{t+1}$$

for all $t \in [T]$, $u \in [t:T]$ and for all $s \in \mathcal{S}$.

The lemma above says that the sequence of value functions $\{\widetilde{V}^{t+1}_u\}_{u\in[t+1:T]}$ generated at time step $t+1$ deviates from the chain of value functions $\{\widetilde{V}^t_u\}_{u\in[t:T]}$ by at most $m_{t-1} - m_{t+1}$. This deviation control enables bounding the term $\sum_{t=1}^{T-1} V^t_{t+1}(s_{t+1}) - \widetilde{V}^{t+1}_{t+1}(s_{t+1})$, which we demonstrate in the next section.

Algorithm 4: γ-DC-LSCVI-UCB [Hong and Tewari, 2025a]

```
Input: Discounting factor $\gamma \in [0,1)$, regularization constant $\lambda > 0$, span $H > 0$,
       bonus factor $\beta > 0$.
Initialize: $\Lambda_1 \leftarrow \lambda I$, $m_{-1} \leftarrow \infty$, $m_0 \leftarrow \infty$, $m_1 \leftarrow \frac{1}{1-\gamma}$, $\widetilde{Q}_u^0(\cdot,\cdot) \leftarrow \frac{1}{1-\gamma}$,
            $\widetilde{Q}_u^{-1}(\cdot,\cdot) \leftarrow \frac{1}{1-\gamma}$.
Receive state $s_1$.
for $t = 1, \ldots, T$ do
    $V_{T+1}^t(\cdot) \leftarrow \frac{1}{1-\gamma}$.
    for $u = T, T-1, \ldots, t$ do
        $\widetilde{Q}_u^t(\cdot,\cdot) \leftarrow \Big(r(\cdot,\cdot) + \gamma([\widehat{P}_t V_{u+1}^t](\cdot,\cdot) + \beta\|\boldsymbol{\varphi}(\cdot,\cdot)\|_{\Lambda_t^{-1}})\Big) \wedge \frac{1}{1-\gamma}$.
        $U_u^t(\cdot,\cdot) \leftarrow \widetilde{Q}_u^{t-1}(\cdot,\cdot) \wedge \widetilde{Q}_u^{t-2}(\cdot,\cdot)$.
        $L_u^t(\cdot,\cdot) \leftarrow (\widetilde{Q}_u^{t-1}(\cdot,\cdot) - m_{t-1} + m_t) \vee (\widetilde{Q}_u^{t-2}(\cdot,\cdot) - m_{t-2} + m_t)$.
        $Q_u^t(\cdot,\cdot) \leftarrow \textsc{Clip}(\widetilde{Q}_u^t(\cdot,\cdot); L_u^t(\cdot,\cdot), U_u^t(\cdot,\cdot))$.
        $\widetilde{V}_u^t(\cdot) \leftarrow \max_a Q_u^t(\cdot,a)$.
        $V_u^t(\cdot) \leftarrow \textsc{Clip}(\widetilde{V}_u^t(\cdot); m_t, m_t + H)$.
    Take action $a_t \leftarrow \operatorname{argmax}_{a\in\mathcal{A}} Q_t^t(s_t,a)$.
    Receive reward $r(s_t,a_t)$. Receive next state $s_{t+1}$.
    $\Lambda_{t+1} \leftarrow \Lambda_t + \boldsymbol{\varphi}(s_t,a_t)\boldsymbol{\varphi}(s_t,a_t)^\top$.
    $m_{t+1} \leftarrow \widetilde{V}_{t+1}^t(s_{t+1}) \wedge m_t$.
```

2.6.3 Regret Analysis

In this section, we outline a regret analysis for our algorithm. Central to the regret analysis is the following concentration bound for the estimate $\widehat{P}_t V$.

Lemma 8. *With probability at least $1-\delta$, there exists an absolute constant c_β such that for $\beta = c_\beta \cdot Hd\sqrt{\log(dT/\delta)}$,*

$$|[\widehat{P}_t V_u^t](s,a) - [PV_u^t](s,a)| \leq \beta\|\boldsymbol{\varphi}(s,a)\|_{\Lambda_t^{-1}}$$

for all $t \in [T]$, $u \in [t:T]$ and $(s,a) \in \mathcal{S}\times\mathcal{A}$.

A proof for the lemma above first finds a concentration bound for $\widehat{P}_t V$ for a fixed value function $V : \mathcal{S} \to \mathbb{R}$ using a concentration bound for vector-valued self-normalized processes. Then, an ϵ-net covering argument is used to get a uniform bound on the function class that captures all value functions V_u^t encountered by the algorithm. For this to work, we require the function class to have low covering number. We can show that the log covering number of the function class that captures functions $\widetilde{Q}_u^t$ can be bounded by $\widetilde{\mathcal{O}}(d^2)$, which amounts to covering the $d \times d$ matrices Λ_t. Since Q_u^t is a function of 5 functions in this function class, the log covering number of the function class that captures Q_u^t is bounded by $\widetilde{\mathcal{O}}(d^2)$. With the concentration inequality, and the fact that the algorithm uses $\beta\|\boldsymbol{\varphi}(s,a)\|_{\Lambda_t^{-1}}$ as the bonus term, we get the following results.

Lemma 9 (Optimism)**.** *With probability at least $1-\delta$, for all $t \in [T]$ and $u \in [t:T]$ and $s \in \mathcal{S}$, we have*

$$V_u^t(s) \geq V^*(s),$$

as long as the input argument H is chosen such that $H \geq 2 \cdot sp(v^*)$.

Lemma 10. *With probability at least $1-\delta$, we have for all $t \in [4:T]$ and $u \in [t:T]$ that*

$$Q_u^t(s,a) \leq r(s,a) + \gamma[PV_{u+1}^t](s,a) + 2\beta\|\boldsymbol{\varphi}(s,a)\|_{\Lambda_t^{-1}} + 2(m_{t-3} - m_t)$$

for all $(s,a) \in \mathcal{S} \times \mathcal{A}$.

Using the lemma above, the regret can be bounded by

$$\begin{aligned} R_T &= \textstyle\sum_{t=1}^T (J^* - r(s_t,a_t)) \\ &\leq \textstyle\sum_{t=4}^T (J^* - Q_t^t(s_t,a_t) + \gamma[PV_{t+1}^t](s_t,a_t) + 2\beta\|\boldsymbol{\varphi}(s_t,a_t)\|_{\Lambda_t^{-1}} + 2(m_{t-3} - m_t)) + \mathcal{O}(1) \end{aligned}$$

which can be decomposed into

$$\begin{aligned} = &\underbrace{\textstyle\sum_{t=4}^T (J^* - (1-\gamma)V_{t+1}^t(s_{t+1}))}_{(a)} + \underbrace{\textstyle\sum_{t=4}^T (V_{t+1}^t(s_{t+1}) - \widetilde{V}_t^t(s_t))}_{(b)} \\ &+ \gamma \underbrace{\textstyle\sum_{t=4}^T ([PV_{t+1}^t](s_t,a_t) - V_{t+1}^t(s_{t+1}))}_{(c)} + 2\underbrace{\beta \textstyle\sum_{t=4}^T \|\boldsymbol{\varphi}(s_t,a_t)\|_{\Lambda_t^{-1}}}_{(d)} + \mathcal{O}(\frac{1}{1-\gamma}). \end{aligned}$$

where we use $Q_t^t(s_t,a_t) = \widetilde{V}_t^t(s_t)$ by the choice of a_t by the algorithm. Each term can be bounded as follows.

Bounding (a) By the optimism result (Lemma 9), we have $V_u^t(s) \geq V^*(s)$ for all $t \in [T]$ and $u \in [t:T]$ with high probability. It follows that

$$\begin{aligned} J^* - (1-\gamma)V_{t+1}^t(s_{t+1}) &\leq J^* - (1-\gamma)V^*(s_{t+1}) \\ &\leq (1-\gamma)\mathrm{sp}(v^*) \end{aligned}$$

where the last inequality is by the bound on the error of approximating the average-reward setting by the discounted setting provided in Lemma 1. Hence, the term (a) can be bounded by $T(1-\gamma)\mathrm{sp}(v^*)$.

Bounding (b) Using Lemma 7 that controls the difference between $\widetilde{V}_u^{t+1}$ and $\widetilde{V}_u^t$, we have

$$\begin{aligned}
V_{t+1}^t(s_{t+1}) &= \textsc{Clip}(\widetilde{V}_{t+1}^t(s_{t+1}); m_t, m_t + H) \\
&\leq \textsc{Clip}(\widetilde{V}_{t+1}^t(s_{t+1}); m_{t+1}, m_{t+1} + H) + m_t - m_{t+1} \\
&\leq \widetilde{V}_{t+1}^t(s_{t+1}) + m_t - m_{t+1} \\
&\leq \widetilde{V}_{t+1}^{t+1}(s_{t+1}) + 2m_{t-1} - 2m_{t+1}
\end{aligned}$$

where the second inequality holds because $\widetilde{V}_{t+1}^t(s_{t+1}) \geq m_{t+1}$ by Line 14. Hence, term (b) can be bounded by $\mathcal{O}(\frac{1}{1-\gamma})$ using telescoping sums of $\widetilde{V}_{t+1}^{t+1}(s_{t+1}) - \widetilde{V}_t^t(s_t)$ and $2m_{t-1} - 2m_{t+1}$, and the fact that $V_u^t \leq \frac{1}{1-\gamma}$ and $m_t \leq \frac{1}{1-\gamma}$ for all $t \in [T]$ and $u \in [t : T]$.

Bounding (c) Since V_u^t is $\mathcal{F}_t$-measurable where $\mathcal{F}_t$ is history up to time step t, we have $\mathbb{E}[V_{t+1}^t(s_{t+1})|\mathcal{F}_t] = [PV_{t+1}^t](s_t, a_t)$, making the summation (c) a summation of a martingale difference sequence. Since $\mathrm{sp}(V_{t+1}^t) \leq H$ for all $t \in [T]$, the summation can be bounded by $\widetilde{\mathcal{O}}(\mathrm{sp}(v^*)\sqrt{T})$ using Azuma-Hoeffding inequality.

Bounding (d) The sum of the bonus terms can be bounded by $\widetilde{\mathcal{O}}(\beta\sqrt{dT})$ using a standard analysis from literature on linear MDP.

Combining the bounds, and choosing $H \geq 2 \cdot \mathrm{sp}(v^*)$ and $\beta = \widetilde{\mathcal{O}}(\mathrm{sp}(v^*)d)$ specified in Lemma 8, we get

$$R_T \leq \widetilde{\mathcal{O}}(T(1-\gamma)\mathrm{sp}(v^*) + \tfrac{1}{1-\gamma} + H\sqrt{T} + H\sqrt{d^3T}).$$

Choosing $\gamma = 1 - \sqrt{1/T}$, we get $R_T \leq \widetilde{\mathcal{O}}(H\sqrt{d^3T})$, leading to our main result (see Appendix A.3.4 for a more detailed analysis):

Theorem 3. *Under Assumptions 2 and 1, running Algorithm 4 with inputs $\gamma = 1 - \sqrt{1/T}$, $\lambda = 1$, $H = 2 \cdot sp(v^*)$ and $\beta = 2c_\beta \cdot sp(v^*)d\sqrt{\log(dT/\delta)}$ guarantees with probability at least $1 - \delta$,*

$$R_T \leq \mathcal{O}(sp(v^*)\sqrt{d^3T\log(dT/\delta)\log T}).$$

where c_β is defined in Lemma 8.

The regret bound for our algorithm γ-DC-LSCVI-UCB matches the regret bound of the previous algorithm γ-LSCVI-UCB.

2.6.4 Computational Complexity

Our algorithm γ-LSCVI-UCB+ runs up to T steps of value iteration every time step, resulting in $\mathcal{O}(T^2)$ value iteration steps. This can be seen by the nested loop structure of the algorithm, where the outer loop is indexed by t for the time step and the inner loop is indexed by u for the value iteration step. The computational bottleneck of the algorithm is computing $\widetilde{Q}_u^t(s,a)$ for all $a \in \mathcal{A}$ and all $s \in \{s_1, \dots, s_{t-1}\}$, which involves computing the regression coefficient $\widehat{\boldsymbol{w}}_t(V_{u+1}^t)$. Computing the regression coefficient takes $\mathcal{O}(T + d^2)$ operations.

In total, the computational complexity of our algorithm is $\mathcal{O}(T^3 d^2 A)$, which is polynomial in the problem parameters T, d, A and is independent of the size of the state space. Although our algorithm enjoys a polynomial-time computational complexity, it is super linear in T, just as the the OLSVI.FH algorithm [Wei et al., 2021] and the previous work γ-LSCVI-UCB [Hong et al., 2025]. We leave further improving the computational complexity to be linear in T as future work.

2.6.5 Related Work

Table 2.2 compares our work with previous approaches for infinite-horizon average-reward linear MDPs. All algorithms in this table uses the Bellman optimality equation assumption. FOPO solves the Bellman optimality equation directly as a fixed-point problem, which is computationally intractable, with brute-force solution requiring computational complexity that scales with T^d, where d is the dimension of the feature representation. OLSVI.FH reduces the problem to the finite-horizon episodic setting. This approach is computationally efficient, but has suboptimal regret bound. LOOP generalizes FOPO to the general function approximation setting, but inherits the computational complexity that scales with T^d for solving a fixed-point problem. γ-LSCVI-UCB reduces the average-reward problem to the discounted problem and achieves an order-optimal regret bound. However, its computational complexity scales with the size of the state space S. Our work is the first computationally efficient algorithm to achieve $\widetilde{\mathcal{O}}(\sqrt{T})$ regret without making strong assumptions.

There is another algorithm called MDP-EXP2 [Wei et al., 2021] directly optimizes for the policy by alternating between policy evaluation and policy improvement. This approach is computationally efficient and achieves an order-optimal regret bound, but requires a strong assumption that all policies induce Markov chains that have uniformly bounded mixing time.

Table 2.2: Comparison of algorithms for infinite-horizon average-reward linear MDP

Algorithm	Regret $\widetilde{\mathcal{O}}(\cdot)$	Computation poly$(\cdot)$
FOPO [Wei et al., 2021]	$\mathrm{sp}(v^*)\sqrt{d^3T}$	T^d, A, d
OLSVI.FH [Wei et al., 2021]	$\sqrt{\mathrm{sp}(v^*)}(dT)^{\frac{3}{4}}$	T, A, d
LOOP [He et al., 2024]	$\sqrt{\mathrm{sp}(v^*)^3d^3T}$	T^d, A, d
γ-LSCVI-UCB [Hong et al., 2025]	$\mathrm{sp}(v^*)\sqrt{d^3T}$	T, S, A, d
γ-DC-LSCVI-UCB [Hong and Tewari, 2025a]	$\mathrm{sp}(v^*)\sqrt{d^3T}$	T, A, d
Lower Bound [Wu et al., 2022]	$\Omega(d\sqrt{\mathrm{sp}(v^*)T})$	

2.7 Discussion

2.7.1 Anytime Algorithm

All algorithms γ-UCB-CVI, γ-DC-LSCVI-UCB and γ-LSCVI-UCB presented in this chapter require the knowledge of the time horizon T to tune the discount factor γ in order to achieve a T-step regret bound of $\widetilde{\mathcal{O}}(\sqrt{T})$. This limitation can be addressed using the standard doubling trick, which allows us to obtain a regret bound of $\widetilde{\mathcal{O}}(\sqrt{T})$ for *any* horizon T. The doubling trick is a standard technique in online learning to convert an algorithm with $\mathcal{O}(\sqrt{T})$ regret guarantee for a fixed known T to an *anytime* algorithm that does not take T as an input and guarantee T-step regret of $\mathcal{O}(\sqrt{T})$ for any T. The idea is to run the algorithm in phases, where each phase lasts twice as long as the previous one. At the beginning of each phase, the algorithm is restarted with parameters tuned for that phase length.

2.7.2 Knowledge of Span of Optimal Bias Function

Throughout this chapter, we assume that an upper bound of the span of the optimal bias function $\mathrm{sp}(v^*)$ is known to the learner. In practice, if one has a general sense of the diameter of the MDP, which is the expected number of steps needed to transition between any two states in the worst case, the diameter can serve as an upper bound, since the diameter is guaranteed to be an upper bound of $\mathrm{sp}(v^*)$ when the reward function is bounded by 1.

The problem of relaxing the assumption on the knowledge of $\mathrm{sp}(v^*)$ has been addressed only recently in the tabular setting [Boone and Zhang, 2024]. The proposed approach estimates the difference in values of the optimal bias function between two states by the total reward collected when transitioning from one state to the other. Extending such relaxation techniques to the linear MDP setting remains an open problem and a direction for future research.

2.7.3 Other Related Work

Infinite-Horizon Average-Reward Setting with General Function Approximation He et al. [2024] study infinite-horizon average reward with general function approximation. They propose an algorithm called LOOP which is a modified version of the fitted Q-iteration with optimistic planning and lazy policy updates. Although their algorithm when adapted to the linear MDP set up achieves $\mathcal{O}(\sqrt{\mathrm{sp}(v^*)^3 d^3 T})$, which is comparable to our work, their algorithm is computationally inefficient.

Infinite-Horizon Average-Reward Setting with Linear MDPs There is another work by Ghosh et al. [2023] on the infinite-horizon average-reward setting with linear MDPs. They study a more general constrained MDP setting where the goal is to maximize average reward while minimizing the average cost. They achieve $\widetilde{\mathcal{O}}(\mathrm{sp}(v^*)\sqrt{d^3T})$ regret, same as our work, but they make an additional assumption that the optimal policy is in a smooth softmax policy class. Also, their algorithm requires solving an intractable optimization problem.

Reduction of Average-Reward to Finite-Horizon Episodic Setting There are works that reduce the average-reward setting to the finite-horizon episodic setting. However, in general, this reduction can only give regret bound of $\mathcal{O}(T^{2/3})$. Chen et al. [2022] study the constrained tabular MDP setting and propose an algorithm that uses the finite-horizon reduction. Their algorithm gives regret bound of $\mathcal{O}(T^{2/3})$. Wei et al. [2021] study the linear MDP setting and propose a finite-horizon reduction that uses the LSVI-UCB [Jin et al., 2020]. Their reduction gives regret bound of $\mathcal{O}(T^{2/3})$.

Online RL in Infinite-Horizon Discounted Setting The literature on online RL in the infinite-horizon discounted setting is sparse because there is no natural notion of regret in this setting without additional assumption on the structure of the MDP. The seminal paper by Liu and Su [2020] introduce a notion of regret in the discounted setting and propose a Q-learning-based algorithm for the tabular setting and provides a regret bound. He et al. [2021] propose a model-based algorithm that adapts UCBVI [Azar et al., 2017] to the discounted setting and achieve a nearly minimax optimal regret bound. Ji and Li [2024] propose a model-free algorithm with nearly minimax optimal regret bound.

Approximation by discounted setting The method of approximating the average-reward setting by the discounted setting has been used in various settings. It is used in the problem of finding a nearly optimal policy given access to a simulator in the tabular setting by Jin and Sidford [2021], Wang et al. [2022], Zurek and Chen [2023], Wang et al.

[2023]. It is also used in the online RL setting with tabular MDPs: Wei et al. [2020] propose a Q-learning based algorithm, but has $\widetilde{\mathcal{O}}(T^{2/3})$ regret. Zhang and Xie [2023] improve the regret to $\widetilde{\mathcal{O}}(\sqrt{T})$ by making use of an estimate for the span of optimal bias function. The reduction is also used in the linear mixture MDP setting by Chae et al. [2025].

Span-constraining methods Learning in the infinite-horizon average-reward setting requires an assumption that ensures the agent can recover from a bad state, leading to a bounded span of the optimal value function. For statistical efficiency, previous work makes use of this fact by constraining the span of the value function estimates. Bartlett and Tewari [2009] modify the extended value iteration algorithm by Auer et al. [2008] to constrain the confidence set on the model so that the spans of the models in the set are bounded. Fruit et al. [2018] propose a computationally efficient version of the algorithm proposed by Bartlett and Tewari [2009]. Zhang and Ji [2019] improve the algorithm proposed by Bartlett and Tewari [2009] by constructing tighter confidence sets using a method for directly estimating the bias function. Zhang and Xie [2023] study a Q-learning-based algorithm that projects the value function to a function class of span-constrained functions. Hong et al. [2025] and Chae et al. [2025] propose a value iteration-based algorithm and clips the value function to constrain its span.

CHAPTER 3

Offline Constrained Reinforcement Learning

3.1 Introduction

The previous chapter discussed the online reinforcement learning framework, where the agent interacts with the environment to collect data for policy improvement. However, in many real-world scenarios, such interaction, particularly when guided by a suboptimal policy, may be costly or unsafe [Kumar et al., 2021, Tang and Wiens, 2021, Levine et al., 2018]. For example, in autonomous driving, allowing an agent to learn from scratch through direct interaction may result in collisions. In such settings, learning from a pre-collected dataset without further interaction becomes a desirable alternative, especially when such data is readily available. In the case of autonomous driving, for instance, the agent may have access to trajectory data collected by human drivers.

In safety-critical applications, it is often necessary to impose constraints on the behavior of the agent. For instance, in autonomous driving, the agent may be required to obey speed limits and avoid collisions, while still aiming to reach the destination efficiently. The constrained reinforcement learning framework [Altman, 1999] provides a natural formulation for incorporating such requirements. It aims to learn a policy that maximizes a primary reward signal while satisfying constraints defined by auxiliary reward signals. These constraints are typically specified by designing suitable safety-related signals and setting thresholds that must be satisfied.

This chapter focuses on the offline constrained reinforcement learning problem, where the objective is to solve a constrained RL problem using a pre-collected dataset. Offline constrained RL inherits the central challenge of offline reinforcement learning [Levine et al., 2020], namely the issue of *distribution shift* [Levine et al., 2020], which arises due to the mismatch between the state-action distribution of the offline dataset and those induced by the candidate policies.

A commonly adopted condition that enables sample-efficient learning of an optimal policy is the uniform data coverage assumption [Antos et al., 2007], which requires the offline dataset to provide sufficient coverage over the state-action distributions induced by all policies. However, this assumption is typically impractical, as it demands the dataset to include states visited by suboptimal policies that would never be encountered by an optimal policy. Such a full data coverage is difficult to obtain in real-world scenarios where data collection is limited or costly.

Recent work has sought to relax this requirement by studying offline RL under a partial data coverage assumption, where the dataset is only assumed to cover the distribution of a single target policy [Jin et al., 2021]. However, existing methods under this relaxed condition suffer from either suboptimal sample complexity or computational inefficiency. This chapter introduces a computationally efficient algorithm for offline constrained RL under the partial data coverage assumption. The algorithm is based on a linear programming formulation.

3.1.1 Key References and Related Work

This chapter is based on the following papers written during my PhD.

1. A primal-dual algorithm for offline constrained reinforcement learning with linear MDPs. ICML 2024. [Hong and Tewari, 2024]

   This paper studies offline constrained reinforcement learning in the linear MDP setting. It introduces a primal-dual algorithm based on the linear programming formulation of RL and proves sample-efficiency under partial data coverage assumptions.

2. Offline constrained reinforcement learning under partial data coverage. Preprint 2025. [Hong and Tewari, 2025b]

   This paper extends this line of work to general function approximation settings. It develops an oracle-efficient primal-dual algorithm and generalizes the approach to handle additional safety constraints via a constrained RL formulation.

3. A primal-dual-critic algorithm for offline constrained reinforcement learning. AISTATS 2024. [Hong et al., 2024]

   This earlier paper also addresses offline constrained RL in the linear MDP setting. However, it requires full data coverage for all policies and uses a different algorithmic structure. Due to these differences and the stronger assumptions required, this paper is not discussed in detail in the thesis.

3.2 Preliminaries

3.2.1 Constrained Reinforcement Learning

We formulate constrained RL with a Constrained Markov decision process (CMDP) [Altman, 1999] defined by a tuple $\mathcal{M} = \big(\mathcal{S}, \mathcal{A}, P, r_0, \{r_i\}_{i=1}^{I}, \gamma, d_0\big)$. Unlike an ordinary MDP, the CMDP allows for multiple reward signals: a primary reward r_0, and there are I additional rewards $r_1, \dots, r_I$. For each reward signal r_i, $i = 0, 1, \dots I$, we define the discounted return of a policy π as $J_i(\pi) := \mathbb{E}^{\pi}[\sum_{t=0}^{\infty} \gamma^t r_i(s_t, a_t)]$. Similarly, the state value function and state-action value function of π with respect to r_i are defined as

$$V_i^{\pi}(s) := \mathbb{E}^{\pi}\left[\textstyle\sum_{t=0}^{\infty} \gamma^t r_i(s_t, a_t) | s_0 = s\right], \quad Q_i^{\pi}(s, a) := \mathbb{E}^{\pi}\left[\textstyle\sum_{t=0}^{\infty} \gamma^t r_i(s_t, a_t) | s_0 = s, a_0 = a\right].$$

The goal of CMDP is to find a stationary policy π that solves the following optimization problem:

$$\begin{aligned} \max_{\pi} \quad & J_0(\pi) \\ \text{subject to} \quad & J_i(\pi) \geq \tau_i/(1-\gamma), \quad i = 1, \dots I \end{aligned} \tag{3.1}$$

where $\tau_i \in [0, 1]$, $i = 1, \dots, I$ are thresholds specified by the designer.

We assume the following Slater's condition, a commonly made assumption in constrained RL [Le et al., 2019, Chen et al., 2021, Bai et al., 2023, Ding et al., 2020] for ensuring strong duality of the optimization problem.

Assumption 3 (Slater's condition). *There exist a constant $\varphi > 0$ and a policy π such that $J_i(\pi) \geq (\tau_i + \varphi)/(1-\gamma)$ for all $i = 1, \dots, I$. Assume φ is known.*

As discussed in Hong et al. [2024], Slater's condition is a mild assumption since given the knowledge of the feasibility of the problem, we can guarantee that Slater's condition is met by slightly loosening the cost threshold.

3.2.2 Linear Programming Formulation of RL

Recall that reinforcement learning aims to find an optimal policy of the following optimization problem

$$\max_{\pi \in \Pi} J_{\gamma}(\pi)$$

where Π denotes the set of stationary policies and $J_{\gamma}(\pi)$ represents the discounted value of policy π, as defined in Equation (1.2). The linear programming (LP) formulation of this problem, introduced by Manne [1960], reformulates the optimization over policies into a linear program whose decision variable is the *occupancy measure*. The occupancy measure associated with a policy π, denoted by $\mu^\pi \in \Delta(\mathcal{S} \times \mathcal{A})$, is defined as

$$\mu^\pi(s,a) := (1-\gamma) \sum_{t=1}^{\infty} \gamma^{t-1} P^\pi(s_t = s, a_t = a).$$

Intuitively, $\mu^\pi(s,a)$ represents the expected discounted visitation frequency of the state-action pair (s,a) under policy π, normalized so that μ^π forms a probability distribution over the space $\mathcal{S} \times \mathcal{A}$.

Leveraging the relation $(1-\gamma)J_\gamma(\pi) = \langle \boldsymbol{\mu}^\pi, \boldsymbol{r} \rangle$, the LP optimizes $\langle \boldsymbol{\mu}, \boldsymbol{r} \rangle$ over admissible occupancy measures $\boldsymbol{\mu}$:

$$\begin{aligned} \max_{\boldsymbol{\mu} \geq 0} \quad & \langle \boldsymbol{\mu}, \boldsymbol{r} \rangle \\ \text{subject to} \quad & \boldsymbol{E}^\top \boldsymbol{\mu} = (1-\gamma)\boldsymbol{d}_0 + \gamma \boldsymbol{P}^\top \boldsymbol{\mu}. \end{aligned} \tag{3.2}$$

where the optimization variable is $\boldsymbol{\mu} \in \mathbb{R}^{|\mathcal{S} \times \mathcal{A}|}$. The bold-faced variables $\boldsymbol{r} \in \mathbb{R}^{|\mathcal{S} \times \mathcal{A}|}$, $\boldsymbol{d}_0 \in \mathbb{R}^{|\mathcal{S}|}$ and $\boldsymbol{P} \in \mathbb{R}^{|\mathcal{S} \times \mathcal{A}| \times |\mathcal{S}|}$ denote the vector or matrix representations of the functions r, d_0 and P, respectively. The linear operator $\boldsymbol{E}$ is such that $[\boldsymbol{E}^\top \mu](s) = \sum_{a'} \mu(s,a')$ and $[\boldsymbol{E}V](s,a) = V(s)$ for all s, a. The constraint, known as the Bellman flow constraint, ensures $\boldsymbol{\mu}$ is an admissible occupancy measure induced by some policy.

The linear programming formulation can be adapted to the constrained RL setting as follows:

$$\begin{aligned} \max_{\boldsymbol{\mu} \geq 0} \quad & \langle \boldsymbol{\mu}, \boldsymbol{r}_0 \rangle \\ \text{subject to} \quad & \boldsymbol{E}^\top \boldsymbol{\mu} = (1-\gamma)\boldsymbol{d}_0 + \gamma \boldsymbol{P}^\top \boldsymbol{\mu} \\ & \langle \boldsymbol{\mu}, \boldsymbol{r}_i \rangle \geq \tau_i, \quad i = 1, \dots I. \end{aligned} \tag{3.3}$$

Policy Extraction Consider a procedure for extracting policy from an occupancy measure μ, not necessarily admissible:

$$[\pi(\mu)](a|s) := \frac{\mu(s,a)}{\sum_{a'} \mu(s,a')} \;\; \text{if} \; \sum_{a'} \mu(s,a') > 0, \quad \frac{1}{|\mathcal{A}|} \; \text{otherwise}. \tag{3.4}$$

It is known that given an admissible occupancy measure μ, the extracted policy $\pi(\mu)$ induces the occupancy measure μ, i.e., $\mu = \mu^{\pi(\mu)}$. With this fact, we can find an optimal occupancy measure μ^* by solving the linear program to find optimal μ^*, and extracting policy $\pi^* = \pi(\mu^*)$ from it.

3.2.3 Offline Dataset

The offline reinforcement learning (RL) setting assumes access to a dataset

$$\mathcal{D} = \{(s_j, a_j, s'_j)\}_{j=1}^n.$$

The pairs (s_j, a_j) for $j = 1, \ldots, n$ are assumed to be independent and identically distributed samples drawn from a data distribution $\mu_D \in \Delta(\mathcal{S} \times \mathcal{A})$, and each s'_j is sampled from the transition kernel $P(\cdot \mid s_j, a_j)$. This i.i.d. assumption on the offline dataset is standard in the offline RL literature [Xie et al., 2021, Zhan et al., 2022, Chen and Jiang, 2022, Zhu et al., 2023] and is used to facilitate the derivation of concentration bounds.

A central challenge in sample-efficient offline RL is the issue of *distribution shift*, which refers to the mismatch between the state-action distribution of the offline dataset and that induced by the target (optimal) policy. To enable reliable learning from offline data, it is necessary to impose an assumption on data coverage that ensures the distribution induced by the target policy is adequately represented in the dataset.

A commonly used assumption in this context is the *concentrability assumption*, which places an upper bound on the ratio between the occupancy measure of the target policy and that of the behavior policy. The normalized occupancy measure of a policy π is defined as

$$\mu^\pi(s,a) = (1-\gamma)\mathbb{E}^\pi\left[\sum_{t=0}^{\infty} \gamma^t \mathbb{I}\{s_t = s, a_t = a\}\right].$$

Intuitively, $\mu^\pi(s,a)$ represents the normalized visitation frequency of the state-action pair (s,a) under policy π. The normalization ensures that μ^π is a probability distribution over $\mathcal{S} \times \mathcal{A}$, satisfying $\sum_{s,a} \mu^\pi(s,a) = 1$. We use the notation $\boldsymbol{\mu}^\pi \in \mathbb{R}^{|\mathcal{S}\times\mathcal{A}|}$ to denote the vector representation of the function $\mu^\pi(s,a)$.

The concentrability assumption is formally stated below.

Assumption 4 (Concentrability)**.** *Let π^* denote an optimal policy, and let $\mu^* = \mu^{\pi^*}$. Then, for all $s \in \mathcal{S}$ and $a \in \mathcal{A}$ with $\mu_D(s,a) > 0$,*

$$\frac{\mu^*(s,a)}{\mu_D(s,a)} \leq C^*,$$

where C^* is a known constant.

This assumption is widely adopted in the offline RL literature [Munos, 2003, 2005]. It requires that the ratio $\mu^*(s,a)/\mu_D(s,a)$ is bounded for all state-action pairs (s,a) where $\mu_D(s,a)$ is strictly positive.

3.3 Linear MDP Setting

This section discusses the algorithm design and theoretical analysis for the linear MDP setting (see Section 1.1.1 for the definition). The linear structure induces a low-dimensional factorization of key quantities, which allows learning statistically and computationally efficient.

In this section, a technical assumption is additionally imposed, following Wagenmaker et al. [2022], to control the complexity of the measure embedding $\boldsymbol{\psi}$. Specifically, it is assumed that there exists a constant D_ψ such that

$$\left\| |\boldsymbol{\psi}| (\mathcal{S}) \right\|_2 \leq D_\psi \sqrt{d},$$

where

$$|\boldsymbol{\psi}| (\mathcal{S}) := \sum_{s \in \mathcal{S}} \left(|\psi_1(s)|, \ldots, |\psi_d(s)| \right).$$

This condition holds, for example, when each ψ_i is a probability measure on $\mathcal{S}$. In that case, the ℓ_1 norm of each ψ_i is bounded by 1, and thus the assumption holds with $D_\psi = 1$.

Let $\boldsymbol{P} \in \mathbb{R}^{|\mathcal{S} \times \mathcal{A}| \times |\mathcal{S}|}$ denote the matrix representation of the transition kernel P, where the entry at row (s,a) and column s' is defined as

$$(\boldsymbol{P})_{(s,a),s'} = P(s' \mid s, a), \quad \text{for all } (s,a,s') \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}.$$

The linear structure of the MDP implies that the transition matrix admits a factorized form

$$\boldsymbol{P} = \boldsymbol{\Phi}\boldsymbol{\Psi},$$

where $\boldsymbol{\Phi} \in \mathbb{R}^{|\mathcal{S} \times \mathcal{A}| \times d}$ is the known matrix whose rows are the feature vectors $(\boldsymbol{\varphi}(s,a))_{(s,a) \in \mathcal{S} \times \mathcal{A}}$, and $\boldsymbol{\Psi} \in \mathbb{R}^{d \times |\mathcal{S}|}$ is the unknown matrix whose rows correspond to the measures $(\psi_i(s'))_{s' \in \mathcal{S}}$. This factorization enables the use of linear regression techniques in estimating the transition dynamics and propagating value functions.

We use $\boldsymbol{r} = \boldsymbol{\Phi}\boldsymbol{\theta}$ and $\boldsymbol{P} = \boldsymbol{\Phi}\boldsymbol{\Psi}$, which hold by the linear MDP assumption, to rewrite the linear program (3.2) as

$$\begin{aligned} \max_{\boldsymbol{\mu} \geq 0} \quad & \langle \boldsymbol{\theta}, \boldsymbol{\Phi}^T \boldsymbol{\mu} \rangle \\ \text{subject to} \quad & \boldsymbol{E}^T \boldsymbol{\mu} = (1-\gamma)\boldsymbol{\nu}_0 + \gamma \boldsymbol{\Psi}^T \boldsymbol{\Phi}^T \boldsymbol{\mu} \end{aligned}$$

Note that the optimization variable $\boldsymbol{\mu} \in \mathbb{R}^{|\mathcal{S} \times \mathcal{A}|}$ is high-dimensional and depends on the cardinality of $\mathcal{S}$. Following Gabbianelli et al. [2024a], for the purpose of computational and statistical efficiency, a low-dimensional optimization variable is introduced:

$$\boldsymbol{\lambda} = \boldsymbol{\Phi}^T \boldsymbol{\mu} \in \mathbb{R}^d,$$

which has the interpretation of the average occupancy measure in the feature space.

With this reparametrization, the primal linear program becomes

$$\begin{aligned} \max_{\boldsymbol{\mu} \geq \mathbf{0},\ \boldsymbol{\lambda}} \quad & \langle \boldsymbol{\theta}, \boldsymbol{\lambda} \rangle \\ \text{subject to} \quad & \boldsymbol{E}^T \boldsymbol{\mu} = (1-\gamma)\boldsymbol{\nu}_0 + \gamma \boldsymbol{\Psi}^T \boldsymbol{\lambda} \\ & \boldsymbol{\lambda} = \boldsymbol{\Phi}^T \boldsymbol{\mu}, \end{aligned}$$

where $\boldsymbol{E}$ is the state-action to state marginalization matrix and $\boldsymbol{\nu}_0 \in \mathbb{R}^{|\mathcal{S}|}$ is the initial state distribution.

The corresponding dual program is given by

$$\begin{aligned} \min_{\boldsymbol{v},\ \boldsymbol{\zeta}} \quad & (1-\gamma)\langle \boldsymbol{\nu}_0, \boldsymbol{v} \rangle \\ \text{subject to} \quad & \boldsymbol{\zeta} = \boldsymbol{\theta} + \gamma \boldsymbol{\Psi} \boldsymbol{v} \\ & \boldsymbol{E}\boldsymbol{v} \geq \boldsymbol{\Phi}\boldsymbol{\zeta}, \end{aligned}$$

where the dual variable $\boldsymbol{v} \in \mathbb{R}^{|\mathcal{S}|}$ corresponds to the state value function and $\boldsymbol{\zeta} \in \mathbb{R}^d$ is such that $\boldsymbol{\Phi}\boldsymbol{\zeta} \in \mathbb{R}^{|\mathcal{S} \times \mathcal{A}|}$ represents the state-action value function.

The Lagrangian associated with the primal and dual problems is

$$\begin{aligned} L(\boldsymbol{\lambda}, \boldsymbol{\mu};\ \boldsymbol{v}, \boldsymbol{\zeta}) &= (1-\gamma)\langle \boldsymbol{\nu}_0, \boldsymbol{v} \rangle + \langle \boldsymbol{\lambda},\ \boldsymbol{\theta} + \gamma \boldsymbol{\Psi} \boldsymbol{v} - \boldsymbol{\zeta} \rangle + \langle \boldsymbol{\mu},\ \boldsymbol{\Phi}\boldsymbol{\zeta} - \boldsymbol{E}\boldsymbol{v} \rangle \\ &= \langle \boldsymbol{\lambda}, \boldsymbol{\theta} \rangle + \langle \boldsymbol{v},\ (1-\gamma)\boldsymbol{\nu}_0 + \gamma \boldsymbol{\Psi}^T \boldsymbol{\lambda} - \boldsymbol{E}^T \boldsymbol{\mu} \rangle + \langle \boldsymbol{\zeta},\ \boldsymbol{\Phi}^T \boldsymbol{\mu} - \boldsymbol{\lambda} \rangle. \end{aligned}$$

Note that the optimization variables $\boldsymbol{\lambda}, \boldsymbol{\zeta} \in \mathbb{R}^d$ are low-dimensional, whereas $\boldsymbol{\mu} \in \mathbb{R}^{|\mathcal{S} \times \mathcal{A}|}$ and $\boldsymbol{v} \in \mathbb{R}^{|\mathcal{S}|}$ are not. To enable the use of a primal-dual algorithm that operates only on low-dimensional variables, a parameterization of $\boldsymbol{\mu}$ and $\boldsymbol{v}$ via a policy variable π is introduced, following Gabbianelli et al. [2024a]:

$$\begin{aligned} \mu_{\boldsymbol{\lambda},\pi}(s,a) &= \pi(a \mid s) \left[(1-\gamma)\nu_0(s) + \gamma \langle \psi(s), \boldsymbol{\lambda} \rangle \right], & (3.5) \\ v_{\boldsymbol{\zeta},\pi}(s) &= \sum_{a \in \mathcal{A}} \pi(a \mid s) \langle \boldsymbol{\zeta}, \boldsymbol{\varphi}(s,a) \rangle. & (3.6) \end{aligned}$$

The parameterization in (3.5) ensures that the Bellman flow constraint $\boldsymbol{E}^T \boldsymbol{\mu}_{\boldsymbol{\lambda},\pi} = (1-\gamma)\boldsymbol{\nu}_0 + \gamma \boldsymbol{\Psi}^T \boldsymbol{\lambda}$ is satisfied by construction. Similarly, the choice of $v_{\boldsymbol{\zeta},\pi}$ in (3.6) ensures that the inner product $\langle \boldsymbol{\mu}, \boldsymbol{\Phi}\boldsymbol{\zeta} - \boldsymbol{E}\boldsymbol{v}_{\boldsymbol{\zeta},\pi}\rangle = 0$ holds.

Using the above parameterization, the Lagrangian function can be expressed entirely in terms of the low-dimensional variables $\boldsymbol{\lambda}$, $\boldsymbol{\zeta}$, and the policy π:

$$\begin{aligned} f(\boldsymbol{\lambda}, \boldsymbol{\zeta}, \pi) &= \langle \boldsymbol{\lambda}, \boldsymbol{\theta}_0\rangle + \langle \boldsymbol{\zeta},\ \boldsymbol{\Phi}^T \boldsymbol{\mu}_{\boldsymbol{\lambda},\pi} - \boldsymbol{\lambda}\rangle && (3.7)\\ &= (1-\gamma)\langle \boldsymbol{\nu}_0,\ v_{\boldsymbol{\zeta},\pi}\rangle + \langle \boldsymbol{\lambda},\ \boldsymbol{\theta}_0 + \gamma \boldsymbol{\Psi} v_{\boldsymbol{\zeta},\pi} - \boldsymbol{\zeta}\rangle. && (3.8) \end{aligned}$$

Although this formulation introduces the policy π as an additional object to track, it remains computationally efficient. In particular, only the distribution $\pi(a \mid s)$ for state-action pairs (s, a) that appear in the dataset needs to be maintained. See Appendix B.3 for further computational details.

Previous work on offline linear MDPs [Gabbianelli et al., 2024a] runs a primal-dual algorithm over the variables $\boldsymbol{\zeta}$ and $\boldsymbol{\beta} = \boldsymbol{\Lambda}^{\dagger}\boldsymbol{\lambda}$, where $\boldsymbol{\Lambda}^{\dagger}$ denotes a pseudoinverse of a suitable scaling matrix. Their approach estimates the gradient of the Lagrangian with respect to these variables and performs a gradient descent step on $\boldsymbol{\zeta}$ for each gradient ascent step on $\boldsymbol{\beta}$. This results in a double-loop algorithmic structure.

Since each gradient descent or ascent step requires a fresh batch of independent samples to control the estimation error, the double-loop structure incurs a sample complexity of $\mathcal{O}(\epsilon^{-4})$ for achieving ϵ-optimality.

In contrast, the approach developed here avoids the double-loop structure and achieves a sample complexity of $\mathcal{O}(\epsilon^{-2})$. This is accomplished by constraining the variable $\boldsymbol{\lambda}$ to lie within a carefully constructed confidence set that enables uniform estimation of the gradient of the Lagrangian across all relevant choices of $\boldsymbol{\lambda}$, $\boldsymbol{\zeta}$, and π.

The details of this construction and the associated analysis are presented in the following subsections.

3.3.1 Regret Analysis

We use $\boldsymbol{Q}^{\pi} \in \mathbb{R}^{|\mathcal{S}\times\mathcal{A}|}$ to denote the matrix representation of the function Q^{π} such that $(\boldsymbol{Q}^{\pi})_{s,a} = Q^{\pi}(s, a)$ and $\boldsymbol{V}^{\pi} \in \mathbb{R}^{|\mathcal{S}|}$ the vector representation of the function V^{π} such that $\boldsymbol{V}^{\pi}_s = V^{\pi}(s)$. With these notations, the well known Bellman equation $Q^{\pi}(s, a) = r(s, a) + \gamma P V^{\pi}(s, a)$ (see e.g. Puterman [2014]) can be written compactly as

$$\boldsymbol{Q}^{\pi} = \boldsymbol{r} + \gamma \boldsymbol{P}\boldsymbol{V}^{\pi} = \boldsymbol{\Phi}(\boldsymbol{\theta} + \gamma \boldsymbol{\Psi}\boldsymbol{V}^{\pi}) = \boldsymbol{\Phi}\boldsymbol{\zeta}^{\pi} \tag{3.9}$$

where $\boldsymbol{r} \in \mathbb{R}^{|\mathcal{S}\times\mathcal{A}|}$ is the matrix representation of the reward function r and we define

$$\boldsymbol{\zeta}^{\pi} := \boldsymbol{\theta} + \gamma \boldsymbol{\Psi} \boldsymbol{V}^{\pi} \in \mathbb{R}^d.$$

This shows that the action value function is linear in the feature vector:

$$Q^{\pi}(s,a) = \langle \boldsymbol{\varphi}(s,a), \boldsymbol{\zeta}^{\pi} \rangle.$$

Due to the boundedness assumptions $\|\boldsymbol{\theta}\|_2 \leq \sqrt{d}$ and $\||\boldsymbol{\psi}|(\mathcal{S})\|_2 \leq D_{\psi}\sqrt{d}$, and the fact that $V^{\pi}(s) \in [0, \frac{1}{1-\gamma}]$, the norm of the parameter $\boldsymbol{\zeta}^{\pi}$ is bounded by

$$\|\boldsymbol{\zeta}^{\pi}\|_2 \leq \sqrt{d} + \frac{\gamma D_{\psi}\sqrt{d}}{1-\gamma} = \mathcal{O}\left(\frac{D_{\psi}\sqrt{d}}{1-\gamma}\right).$$

We define $D_{\zeta} := \sqrt{d} + \frac{\gamma D_{\psi}\sqrt{d}}{1-\gamma}$.

For a given policy π, recall that $\boldsymbol{\zeta}^{\pi} \in \mathbb{R}^d$ is the parameter such that the action-value function satisfies $\boldsymbol{Q}^{\pi} = \boldsymbol{\Phi}\boldsymbol{\zeta}^{\pi}$. It can be shown that for any $\boldsymbol{\lambda} \in \mathbb{R}^d$,

$$f(\boldsymbol{\zeta}^{\pi}, \boldsymbol{\lambda}, \pi) = J(\pi).$$

Furthermore, defining $\boldsymbol{\lambda}^{\pi} = \boldsymbol{\Phi}^T \boldsymbol{\mu}^{\pi}$, which represents the average occupancy in the feature space under policy π, it can also be shown that for any $\boldsymbol{\zeta} \in \mathbb{R}^d$,

$$f(\boldsymbol{\zeta}, \boldsymbol{\lambda}^{\pi}, \pi) = J(\pi).$$

Proofs of these identities are provided in Appendix B.5.2.

As a result, for any sequences $\{\pi_t\}$, $\{\boldsymbol{\zeta}_t\} \subset \mathbb{R}^d$, and $\{\boldsymbol{\lambda}_t\} \subset \mathbb{R}^d$, the suboptimality gap can be decomposed as

$$\begin{aligned} J(\pi^*) - J(\pi_t) &= f(\boldsymbol{\zeta}_t, \boldsymbol{\lambda}^*, \pi^*) - f(\boldsymbol{\zeta}^{\pi_t}, \boldsymbol{\lambda}_t, \pi_t) \\ &= (\underbrace{f(\boldsymbol{\zeta}_t, \boldsymbol{\lambda}^*, \pi^*) - f(\boldsymbol{\zeta}_t, \boldsymbol{\lambda}^*, \pi_t)}_{\textsc{Reg}_t^{\pi}}) + (\underbrace{f(\boldsymbol{\zeta}_t, \boldsymbol{\lambda}^*, \pi_t) - f(\boldsymbol{\zeta}_t, \boldsymbol{\lambda}_t, \pi_t)}_{\textsc{Reg}_t^{\lambda}}) \\ &\quad + (\underbrace{f(\boldsymbol{\zeta}_t, \boldsymbol{\lambda}_t, \pi_t) - f(\boldsymbol{\zeta}^{\pi_t}, \boldsymbol{\lambda}_t, \pi_t)}_{\textsc{Reg}_t^{\zeta}}), \end{aligned}$$

where $\boldsymbol{\lambda}^* = \boldsymbol{\lambda}^{\pi^*}$ denotes the feature occupancy vector under the optimal policy.

This decomposition expresses the suboptimality $J(\pi^*) - J(\pi_t)$ in terms of the regrets of the three players: the policy player, the dual variable player, and the primal variable player.

If the cumulative regrets $\sum_{t=1}^T \textsc{Reg}_t^\pi$, $\sum_{t=1}^T \textsc{Reg}_t^\lambda$, and $\sum_{t=1}^T \textsc{Reg}_t^\zeta$ are all sublinear in T, then the average suboptimality satisfies

$$\frac{1}{T}\sum_{t=1}^{T}(J(\pi^*) - J(\pi_t)) = J(\pi^*) - J(\bar{\pi}) = o(1),$$

where $\bar{\pi} = \text{Unif}(\pi_1, \ldots, \pi_T)$ is the mixture policy that selects one of $\pi_1, \ldots, \pi_T$ uniformly at random and executes the selected policy throughout the episode.

The remainder of this section outlines regret analyses for the three players. These analyses motivate the algorithm introduced in Section 3.3.2.

3.3.1.1 Bounding Regret of π-player

Using Equation (3.8), the regret of π-player simplifies to

$$\begin{aligned}\text{Reg}_t^\pi &= f(\boldsymbol{\zeta}_t, \boldsymbol{\lambda}^*, \pi^*) - f(\boldsymbol{\zeta}_t, \boldsymbol{\lambda}^*, \pi_t)\\ &= \langle \boldsymbol{\nu}^*, \boldsymbol{v}_{\boldsymbol{\zeta}_t,\pi^*} - \boldsymbol{v}_{\boldsymbol{\zeta}_t,\pi_t}\rangle\\ &= \langle \boldsymbol{\nu}^*, \textstyle\sum_a (\pi^*(a|\cdot) - \pi_t(a|\cdot))\langle \boldsymbol{\zeta}_t, \boldsymbol{\varphi}(\cdot, a)\rangle\rangle.\end{aligned}$$

where we define $\boldsymbol{\nu}^\pi = (1-\gamma)\boldsymbol{\nu}_0 + \gamma \boldsymbol{\Psi}^T \boldsymbol{\lambda}^\pi$ as the state occupancy measure induced by π and write $\boldsymbol{\nu}^* = \boldsymbol{\nu}^{\pi^*}$. The regret can be bounded if π-player updates its policy using an exponentiation algorithm [Zanette et al., 2021]

$$\pi_{t+1} = \sigma\left(\alpha \sum_{i=1}^{t} \boldsymbol{\Phi}\boldsymbol{\zeta}_i\right)$$

where $\sigma(\boldsymbol{q})$ for $\boldsymbol{q} \in \mathbb{R}^{|\mathcal{S}\times\mathcal{A}|}$ is a softmax policy with

$$\sigma(\boldsymbol{q})(a|s) := \frac{\exp(q(s,a))}{\sum_{a'}\exp(q(s,a'))}.$$

Based on the standard mirror descent analysis by Gabbianelli et al. [2024a] (Appendix B.4.1) we can show that, choosing $\alpha = \mathcal{O}((1-\gamma)\sqrt{\log|\mathcal{A}|/(dT)})$ gives

$$\frac{1}{T}\sum_{t=1}^{T}\textsc{Reg}_t^\pi \le \mathcal{O}\left(\frac{1}{1-\gamma}\sqrt{(d\log|\mathcal{A}|)/T}\right)$$

which vanishes as T increases. Consequently, choosing T to be at least $\Omega(\frac{d\log|\mathcal{A}|}{(1-\gamma)^2\epsilon^2})$ gives $\frac{1}{T}\sum_{t=1}^T \text{Reg}_t^\pi \le \epsilon$. Note that when the exponentiation algorithm is employed, the π-player does not need to know the value of $\boldsymbol{\zeta}_t$ when choosing π_t, allowing the π-player to play before the ζ-player. Another benefit of the exponentiation algorithm is that the policy chosen by the π-player is restricted to the softmax function class $\Pi(D_\pi)$ where $\Pi(\cdot)$ is defined as

$$\Pi(B) := \{\sigma(\boldsymbol{\Phi}\boldsymbol{z}) : \boldsymbol{z} \in \mathbb{B}_d(B)\}. \tag{3.10}$$

and $D_\pi := \alpha T D_\zeta$. The restriction allows statistically efficient estimation of quantities that depend on policies in $\Pi(B)$ via covering argument on $\Pi(B)$, as we will see in later sections.

3.3.1.2 Bounding Regret of ζ-player

Using Equation (3.7), the regret of ζ-player simplifies to

$$\begin{aligned}\textsc{Reg}_t^\zeta &= f(\boldsymbol{\zeta}_t, \boldsymbol{\lambda}_t, \pi_t) - f(\boldsymbol{\zeta}^{\pi_t}, \boldsymbol{\lambda}_t, \pi_t)\\ &= \langle \boldsymbol{\zeta}_t - \boldsymbol{\zeta}^{\pi_t}, \boldsymbol{\Phi}^T \boldsymbol{\mu}_{\boldsymbol{\lambda}_t, \pi_t} - \boldsymbol{\lambda}_t\rangle.\end{aligned}$$

Recall that $\boldsymbol{\mu}_{\boldsymbol{\lambda},\pi} = \pi \circ \boldsymbol{E}[(1-\gamma)\boldsymbol{\nu}_0 + \gamma \boldsymbol{\Psi}^T \boldsymbol{\lambda}]$. The only unknown quantity in the regret is $\boldsymbol{\Psi}^T\boldsymbol{\lambda} \in \mathbb{R}^{|\mathcal{S}|}$. Note that $\boldsymbol{\Psi}^T\boldsymbol{\varphi}(s,a) = (P(s'|s,a))_{s'\in\mathcal{S}}$ is a next-state distribution given current state-action pair (s,a), and $\boldsymbol{e}_{s'_k}$ is an unbiased estimator for $\boldsymbol{\Psi}^T\boldsymbol{\varphi}(s_k,a_k)$. Hence, if $\boldsymbol{\lambda}$ is a linear combination $\sum_{k=1}^n c_k \boldsymbol{\varphi}(s_k,a_k)$, we can construct an unbiased estimator $\sum_{k=1}^n c_k \boldsymbol{e}_{s'_k}$ for $\boldsymbol{\Psi}^T\boldsymbol{\lambda}$. Motivated by this observation, to facilitate the algorithm design for the ζ-player, we will restrict the λ-player to choose a linear combination of feature vectors that appear in the dataset to allow estimating $\boldsymbol{\Psi}^T\boldsymbol{\lambda}$. Specifically, we strict the λ-player to choose $\boldsymbol{\lambda}_t$ from the following set where the bound B will be chosen later.

$$\mathcal{C}_n(B) := \left\{ \frac{1}{n}\sum_{k=1}^n c_k \boldsymbol{\varphi}(s_k,a_k) : c_1,\dots,c_n \in [-B,B] \right\}. \tag{3.11}$$

Given the restriction, we can parameterize the value of $\boldsymbol{\lambda}_t$ by the coefficients $\boldsymbol{c}_t \in [-B,B]^n$ for some bound B, and write $\boldsymbol{\lambda}_t = \boldsymbol{\lambda}(\boldsymbol{c}_t)$ where we define

$$\boldsymbol{\lambda}(\boldsymbol{c}) := \frac{1}{n}\sum_{k=1}^n c_k \boldsymbol{\varphi}(s_k,a_k)$$

Following the previous discussion, we define the estimates for $\boldsymbol{\Psi}^T\boldsymbol{\lambda}(c)$ and $\boldsymbol{\mu}_{\boldsymbol{\lambda}(\boldsymbol{c}),\pi}$ parameterized by $\boldsymbol{c} \in [-B, B]^n$:

$$\widehat{\boldsymbol{\Psi}^T\boldsymbol{\lambda}}(\boldsymbol{c}) := \frac{1}{n}\sum_{k=1}^{n} c_k \boldsymbol{e}_{s'_k}$$
$$\widehat{\boldsymbol{\mu}}_{\boldsymbol{\lambda}(\boldsymbol{c}),\pi} := \pi \circ \boldsymbol{E}[(1-\gamma)\boldsymbol{\nu}_0 + \gamma\widehat{\boldsymbol{\Psi}^T\boldsymbol{\lambda}}(\boldsymbol{c})].$$

These estimates enjoy the following concentration bound, which can be shown using matrix Bernstein inequality. See Appendix B.4.2 for a proof.

Lemma 11. *For a fixed $\boldsymbol{\lambda}(\boldsymbol{c}) = \frac{1}{n}\sum_{k=1}^{n} c_k \boldsymbol{\varphi}(s_k, a_k)$ with $|c_k| \le B$ for $k = 1, \ldots, n$, and a policy π, we have*

$$\|\boldsymbol{\Phi}^T\boldsymbol{\mu}_{\boldsymbol{\lambda}(\boldsymbol{c}),\pi} - \boldsymbol{\Phi}^T\widehat{\boldsymbol{\mu}}_{\boldsymbol{\lambda}(\boldsymbol{c}),\pi}\|_2 \le \mathcal{O}\left(B\sqrt{\frac{\log(d/\delta)}{n}}\right)$$

with probability at least $1-\delta$ conditional on the data of state-action pairs $\{(s_k, a_k)\}_{k=1}^n$.

For estimating the regret $\langle \boldsymbol{\zeta}_t - \boldsymbol{\zeta}^{\pi_t}, \boldsymbol{\Phi}^T\boldsymbol{\mu}_{\boldsymbol{\lambda}(\boldsymbol{c}_t),\pi_t} - \boldsymbol{\lambda}(\boldsymbol{c}_t)\rangle$, we need a uniform concentration bound on the estimates $\boldsymbol{\Phi}^T\widehat{\boldsymbol{\mu}}_{\boldsymbol{\lambda}(\boldsymbol{c}_t),\pi}$ over $\boldsymbol{\lambda}(\boldsymbol{c})$ and π. The restriction on the π-player to choose a policy in the softmax function class defined in (3.10) allows converting the concentration bound for a fixed policy π in Lemma 11 to a uniform concentration bound over all policies in the softmax function class via a covering argument. The conversion is possible due to the fact that the log covering number for the softmax function class is bounded by $\widetilde{\mathcal{O}}(d)$ (see Lemma 41 in Appendix B.1). However, such a conversion to a uniform concentration bound over all $\boldsymbol{\lambda}(\boldsymbol{c})$ for $\boldsymbol{c} \in [-B, B]^n$ is elusive since a naive covering argument on the space of parameters $[-B, B]^n$ will give a log covering number bound of $\mathcal{O}(n)$. To sidestep this issue, we exploit the fact that $\mathcal{C}_n(B)$ can be spanned by a set of spanners $\{\boldsymbol{\varphi}(s_j, a_j)\}_{j\in\mathcal{I}}$ for some index set $\mathcal{I} \subseteq \{1, \ldots, n\}$ of size at most d. This can be seen by the following lemma by Awerbuch and Kleinberg [2008].

Lemma 12 (Barycentric spanner)**.** *Let $\mathcal{K} \subseteq \mathbb{R}^d$ be compact set. Then, there exists a spanner $\{\boldsymbol{\phi}_1, \cdots \boldsymbol{\phi}_d\} \subset \mathcal{K}$ such that any vector $\boldsymbol{x} \in \mathcal{K}$ can be represented as $\boldsymbol{x} = \sum_{i=1}^{d} c_i \boldsymbol{\phi}_i$ where $c_i \in [-C, C]$ for all $i = 1, \ldots, d$. Such a spanner is called a C-approximate barycentric spanner for $\mathcal{K}$. If $\mathcal{K}$ is finite, we can find a C-approximate barycentric spanner in time complexity $\mathcal{O}(nd^2 \log_C d)$.*

Applying this lemma, we can compute a 2-approximate barycentric spanner $\{\boldsymbol{\varphi}(s_j, a_j)\}_{j\in\mathcal{I}}$ for $\{\boldsymbol{\varphi}(s_k, a_k)\}_{k=1}^n$ where $\mathcal{I} \subseteq \{1, \ldots, n\}$ is an index set of size d. Given any $\boldsymbol{c} \in [-B, B]^n$, we can convert it to $\boldsymbol{c}' \in [-2B, 2B]^n$ with c'_k nonzero only if $k \in \mathcal{I}$ such that $\boldsymbol{\lambda}(\boldsymbol{c}) = \boldsymbol{\lambda}(\boldsymbol{c}')$.

This can be seen by

$$\begin{aligned}\boldsymbol{\lambda}(\boldsymbol{c}) &= \frac{1}{n}\sum_{k=1}^{n} c_k \boldsymbol{\varphi}(s_k, a_k) \\ &= \sum_{j\in\mathcal{I}}\left(\frac{1}{n}\sum_{k=1}^{n} b_{kj} c_k\right)\boldsymbol{\varphi}(s_j, a_j) = \boldsymbol{\lambda}(\boldsymbol{c}')\end{aligned}$$

where the coefficients $b_{kj} \in [-2, 2]$ are such that $\boldsymbol{\varphi}(s_k, a_k) = \sum_{j\in\mathcal{I}} b_{kj}\boldsymbol{\varphi}(s_j, a_j)$, which exist by the fact that $\{\boldsymbol{\varphi}(s_j, a_j)\}_{j\in\mathcal{I}}$ is a 2-approximate barycentric spanner of $\{\boldsymbol{\varphi}(s_k, a_k)\}_{k=1}^{n}$. We summarize the definition of the conversion from $\boldsymbol{c}$ to $\boldsymbol{c}'$ that satisfies $\boldsymbol{\lambda}(\boldsymbol{c}) = \boldsymbol{\lambda}(\boldsymbol{c}')$.

Definition 1. *Given a dataset $\{(s_k, a_k, s_k')\}_{k=1}^{n}$, let $\{\boldsymbol{\varphi}(s_j, a_j)\}_{j\in\mathcal{I}}$ be a 2-approximate barycentric spanner for $\{\boldsymbol{\varphi}(s_k, a_k)\}_{k=1}^{n}$ with $|\mathcal{I}| \le d$. We define the conversion of $\boldsymbol{c} \in \mathbb{R}^n$ to $\boldsymbol{c}' \in \mathbb{R}^n$ as*

$$c_j' = \begin{cases} \frac{1}{n}\sum_{k=1}^{n} b_{kj} c_k & \text{if } j \in \mathcal{I} \\ 0 & \text{otherwise.} \end{cases} \tag{3.12}$$

where b_{kj} are the coefficients such that $\boldsymbol{\varphi}(s_k, a_k) = \sum_{j\in\mathcal{I}} b_{kj}\boldsymbol{\varphi}(s_j, a_j)$ with $b_{kj} = 0$ for $j \notin \mathcal{I}$.

Given $\boldsymbol{c}_t \in [-B, B]^n$ such that $\boldsymbol{\lambda}(\boldsymbol{c}_t) \in \mathcal{C}_n(B)$, let $\boldsymbol{c}_t' \in [-2B, 2B]^n$ be the conversion such that $\boldsymbol{\lambda}(\boldsymbol{c}_t') = \boldsymbol{\lambda}(\boldsymbol{c}_t)$ with only the coefficients with indices in $\mathcal{I}$ nonzero. The converted coefficients $\boldsymbol{c}_t' \in \mathbb{R}^n$ live in a low dimensional space $\{\boldsymbol{c}' \in [-2B, 2B]^n : c_j' = 0 \text{ if } j \notin \mathcal{I}\}$ with the log covering number of $\mathcal{O}(d)$. To use a covering argument, let $\boldsymbol{c}_t''$ be the covering center closest to $\boldsymbol{c}_t'$. Then we can decompose the regret of the ζ-player as

$$\begin{aligned}\textsc{Reg}_t^{\zeta} &= \langle \boldsymbol{\zeta}_t - \boldsymbol{\zeta}^{\pi_t}, \boldsymbol{\Phi}^T\boldsymbol{\mu}_{\boldsymbol{\lambda}(\boldsymbol{c}_t),\pi_t} - \boldsymbol{\lambda}(\boldsymbol{c}_t)\rangle \\ &= \langle \boldsymbol{\zeta}_t - \boldsymbol{\zeta}^{\pi_t}, \boldsymbol{\Phi}^T\boldsymbol{\mu}_{\boldsymbol{\lambda}(\boldsymbol{c}_t'),\pi_t} - \boldsymbol{\Phi}^T\boldsymbol{\mu}_{\boldsymbol{\lambda}_t(\boldsymbol{c}_t''),\pi_t}\rangle + \langle \boldsymbol{\zeta}_t - \boldsymbol{\zeta}^{\pi_t}, \boldsymbol{\Phi}^T\boldsymbol{\mu}_{\boldsymbol{\lambda}(\boldsymbol{c}_t''),\pi_t} - \boldsymbol{\Phi}^T\widehat{\boldsymbol{\mu}}_{\boldsymbol{\lambda}(\boldsymbol{c}_t''),\pi_t}\rangle \\ &\quad + \langle \boldsymbol{\zeta}_t - \boldsymbol{\zeta}^{\pi_t}, \boldsymbol{\Phi}^T\widehat{\boldsymbol{\mu}}_{\boldsymbol{\lambda}(\boldsymbol{c}_t''),\pi_t} - \boldsymbol{\Phi}^T\widehat{\boldsymbol{\mu}}_{\boldsymbol{\lambda}(\boldsymbol{c}_t'),\pi_t}\rangle + \langle \boldsymbol{\zeta}_t - \boldsymbol{\zeta}^{\pi_t}, \boldsymbol{\Phi}^T\widehat{\boldsymbol{\mu}}_{\boldsymbol{\lambda}(\boldsymbol{c}_t'),\pi_t} - \boldsymbol{\lambda}(\boldsymbol{c}_t')\rangle.\end{aligned}$$

The first term can be bounded since $\boldsymbol{\lambda}(c_t') \approx \boldsymbol{\lambda}_t(\boldsymbol{c}_t'')$. The second term can be bounded using a union bound of the concentration inequalities on $\boldsymbol{\Phi}^T\widehat{\boldsymbol{\mu}}_{\boldsymbol{\lambda}(\boldsymbol{c}''),\pi}$ over $\boldsymbol{c}''$ in the cover of $[-2B, 2B]^d$. The third term can be bounded since $c_t' \approx c_t''$. The last term, interpreted as a regret of the ζ-player against a dynamic action $\boldsymbol{\zeta}^{\pi_t}$, can be bounded by a greedy ζ-player that minimizes $\langle \cdot, \boldsymbol{\Phi}^T\widehat{\boldsymbol{\mu}}_{\boldsymbol{\lambda}(\boldsymbol{c}_t'),\pi_t} - \boldsymbol{\lambda}(\boldsymbol{c}_t')\rangle$. The greedy strategy requires ζ-player to play after λ-player and π-player. The bounds lead to

$$\frac{1}{T}\sum_{t=1}^{T}\textsc{Reg}_t^{\zeta} \le \mathcal{O}\left(\frac{Bd}{1-\gamma}\sqrt{\frac{\log(Bdn/\delta)}{n}}\right).$$

In summary, we can bound the regret of ζ-player if ζ-player plays $\boldsymbol{\zeta}_t \in \mathbb{B}_d(D_\zeta)$ that minimizes $\langle \cdot, \boldsymbol{\Phi}^T \widehat{\boldsymbol{\mu}}_{\boldsymbol{\lambda}(\boldsymbol{c}'_t),\pi_t} - \boldsymbol{\lambda}(\boldsymbol{c}'_t)\rangle$. The greedy strategy requires ζ-player to play after λ-player and π-player. See Appendix B.4.2 for detailed analysis.

3.3.1.3 Bounding Regret of λ-player

Using Equation (3.8), the regret of λ-player simplifies to

$$\begin{aligned}\textsc{Reg}_t^{\lambda} &= f(\boldsymbol{\zeta}_t, \boldsymbol{\lambda}^*, \pi_t) - f(\boldsymbol{\zeta}_t, \boldsymbol{\lambda}_t, \pi_t) \\ &= \langle \boldsymbol{\lambda}^* - \boldsymbol{\lambda}_t, \underbrace{\boldsymbol{\theta} + \gamma \boldsymbol{\Psi} \boldsymbol{v}_{\zeta_t,\pi_t} - \boldsymbol{\zeta}_t}_{=\boldsymbol{\xi}_t} \rangle\end{aligned}$$

The sum of $\textsc{Reg}_t^{\lambda}$ over $t = 1, \ldots, T$ is the regret of the λ-player against a fixed action λ^* where the reward function at time t is $\langle \cdot, \boldsymbol{\xi}_t \rangle$. From the previous section, we require λ player to play before ζ-player, whose play affects $\boldsymbol{\xi}_t$. Hence, the decision of λ-player at time t must be made before the knowledge of $\boldsymbol{\xi}_t$. Assuming for now that $\boldsymbol{\xi}_t$ is known ($\boldsymbol{\xi}_t$ is in fact unknown and needs to be estimated since $\boldsymbol{\Psi}$ is unknown), the regret of λ-player can be made sublinear in T by employing a no-regret online convex optimization oracle (defined below) on $\mathcal{C}_n(B)$ as long as $\boldsymbol{\lambda}^* \in \mathcal{C}_n(B)$.

Definition 2. *An algorithm is called a no-regret online convex optimization oracle with respect to a convex set $\mathcal{C}$ if, for any sequence of convex functions $h_1, \ldots, h_T : \mathbb{R}^d \to [-1, 1]$ and for any $\boldsymbol{\lambda} \in \mathcal{C}$, the sequence of vectors $\boldsymbol{\lambda}_1, \ldots, \boldsymbol{\lambda}_T \in \mathcal{C}$ produced by the algorithm satisfies*

$$\frac{1}{T} \sum_{t=1}^{T} h_t(\boldsymbol{\lambda}_t) - h_t(\boldsymbol{\lambda}) \le \epsilon_{opt}^{\lambda}(T)$$

for some $\epsilon_{opt}^{\lambda}(T) > 0$ that converges to 0 as $T \to \infty$.

The online gradient descent algorithm [Hazan et al., 2016] is an example of a computationally efficient online convex optimization oracle. Employing a no-regret online convex optimization oracle with convex set $\mathcal{C}_n(B)$ on the sequence of functions $\langle \cdot, \boldsymbol{\xi}_t \rangle$, the λ-player can enjoy a sublinear regret against any fixed $\boldsymbol{\lambda} \in \mathcal{C}_n(B)$. However, $\boldsymbol{\lambda}^*$ may not lie in $\mathcal{C}_n(B)$ for any B. In fact, we can construct an example where $\boldsymbol{\lambda}^*$ is not in the span of $\{\boldsymbol{\varphi}(s_k, a_k)\}_{k=1}^n$ with probability at least 1/2 (Lemma 44). To sidestep this problem, we show $\boldsymbol{\lambda}^*$ can be approximated by a vector $\widehat{\boldsymbol{\lambda}}^*$ in $\mathcal{C}_n(C^*)$:

Lemma 13. *Under the concentrability assumption 4, there exists $\widehat{\boldsymbol{\lambda}}^* \in \mathbb{R}^d$ of the form*

$\widehat{\boldsymbol{\lambda}}^* = \frac{1}{n}\sum_{k=1}^n c_k \boldsymbol{\varphi}(s_k, a_k)$ *with* $c_k \in [0, C^*]$, $k = 1, \ldots, n$ *such that*

$$\|\widehat{\boldsymbol{\lambda}}^* - \boldsymbol{\lambda}^*\|_2 \leq \mathcal{O}\left(C^* \sqrt{\tfrac{\log(d/\delta)}{n}}\right)$$

with probability at least $1 - \delta$.

Note that this lemma is the only place the data coverage assumption is needed for our analysis and this is where we require choosing $B = C^*$. Also, in our algorithm, computing $\widehat{\boldsymbol{\lambda}}^*$ is not needed. Only the existence of such a vector $\widehat{\boldsymbol{\lambda}}^*$ is needed in the analysis. With the lemma above, we can approximate the regret as

$$\begin{aligned}\text{REG}_t^{\lambda} &= \langle \boldsymbol{\lambda}^* - \boldsymbol{\lambda}_t, \boldsymbol{\theta} + \gamma \boldsymbol{\Psi} \boldsymbol{v}_{\boldsymbol{\zeta}_t, \pi_t} - \boldsymbol{\zeta}_t \rangle \\ &\approx \langle \widehat{\boldsymbol{\lambda}}^* - \boldsymbol{\lambda}_t, \boldsymbol{\theta} + \gamma \boldsymbol{\Psi} \boldsymbol{v}_{\boldsymbol{\zeta}_t, \pi_t} - \boldsymbol{\zeta}_t \rangle\end{aligned}$$

and argue that the sum of the above quantity over $t = 1, \ldots, T$ is sublinear since the regret against $\widehat{\boldsymbol{\lambda}}^* \in \mathcal{C}_n(C^*)$ is sublinear when employing a no-regret online convex optimization oracle. Now, we deal with the fact that the term $\boldsymbol{\Psi} \boldsymbol{v}_{\boldsymbol{\zeta}_t, \pi_t}$ is unknown and needs to be estimated. Observing that $\langle \boldsymbol{\varphi}(s, a), \boldsymbol{\Psi} \boldsymbol{v} \rangle = \mathbb{E}_{s' \sim P(\cdot|s,a)}[\boldsymbol{v}(s')]$, we can estimate $\boldsymbol{\Psi} \boldsymbol{v} \in \mathbb{R}^d$ for any $\boldsymbol{v} \in \mathbb{R}^{|\mathcal{S}|}$ by regressing $v(s')$ on $\boldsymbol{\varphi}(s, a)$ using the triplets (s, a, s') in the dataset $\mathcal{D}$. Following the literature on linear bandits [Abbasi-Yadkori et al., 2011], we use the regularized least squares estimate

$$\widehat{\boldsymbol{\Psi} \boldsymbol{v}} := (n \widehat{\boldsymbol{\Lambda}}_n + \boldsymbol{I})^{-1} \textstyle\sum_{k=1}^n v(s'_k) \boldsymbol{\varphi}(s_k, a_k) \tag{3.13}$$

where $\widehat{\boldsymbol{\Lambda}}_n := \frac{1}{n}\sum_{k=1}^n \boldsymbol{\varphi}(s_k, a_k)\boldsymbol{\varphi}(s_k, a_k)^T$ is the empirical Gram matrix. By the well-known result for linear bandits (e.g. Theorem 2 in Abbasi-Yadkori et al. [2011]), we have the following high-probability concentration bound (Lemma 46) for the estimate $\widehat{\boldsymbol{\Psi} \boldsymbol{v}}$ where $v : \mathcal{S} \to [0, D_v]$:

$$\|\boldsymbol{\Psi} \boldsymbol{v} - \widehat{\boldsymbol{\Psi} \boldsymbol{v}}\|_{n \widehat{\boldsymbol{\Lambda}}_n + \boldsymbol{I}} \leq \mathcal{O}\left(D_v \sqrt{d \log(n/\delta)}\right).$$

Since we need concentration bound of $\widehat{\boldsymbol{\Psi} \boldsymbol{v}}_{\boldsymbol{\zeta}_t, \pi_t}$ where $\boldsymbol{v}_{\boldsymbol{\zeta}_t, \pi_t}$ are random, we need a uniform bound over all possible functions $v_{\boldsymbol{\zeta}_t, \pi_t}$. Since the domain of v has cardinality $|\mathcal{S}|$, a naive covering argument on the function space of v will make the bound scale with poly$(|\mathcal{S}|)$. To avoid this, we use a careful covering argument exploiting the fact that π_t is a softmax function parameterized by a d-dimensional vector and $\boldsymbol{\zeta}_t$ are d-dimensional vectors. With covering, we can show the following uniform concentration bound.

Lemma 14. *Consider a function class*

$$\mathcal{V} = \left\{ v_{\boldsymbol{\zeta},\pi} \in (\mathcal{S} \to [0, D_v]) : \boldsymbol{\zeta} \in \mathbb{B}(D_\zeta), \pi \in \Pi(D_\pi) \right\}.$$

With probability at least $1-\delta$*, we have*

$$\|\boldsymbol{\Psi v} - \widehat{\boldsymbol{\Psi v}}\|_{n\widehat{\boldsymbol{\Lambda}}_n + \boldsymbol{I}} \leq \mathcal{O}\left(D_v \sqrt{d \log(D_\zeta D_\pi n/\delta)} \right)$$

uniformly over $\boldsymbol{v} \in \mathcal{V}$ *where* $\widehat{\boldsymbol{\Psi v}}$ *is the least squares estimate defined in* (3.13).

See Lemma 14 in the Appendix for detail. With the uniform concentration bound, we can continue bounding the regret of λ-player as follows.

$$\langle \widehat{\boldsymbol{\lambda}}^* - \boldsymbol{\lambda}_t, \boldsymbol{\theta} + \gamma \boldsymbol{\Psi} \boldsymbol{v}_{\boldsymbol{\zeta}_t,\pi_t} - \boldsymbol{\zeta}_t \rangle = \langle \widehat{\boldsymbol{\lambda}}^* - \boldsymbol{\lambda}_t, \boldsymbol{\theta} + \gamma \widehat{\boldsymbol{\Psi v}}_{\boldsymbol{\zeta}_t,\pi_t} - \boldsymbol{\zeta}_t \rangle + \gamma \langle \widehat{\boldsymbol{\lambda}}^* - \boldsymbol{\lambda}_t, \boldsymbol{\Psi} \boldsymbol{v}_{\boldsymbol{\zeta}_t,\pi_t} - \widehat{\boldsymbol{\Psi v}}_{\boldsymbol{\zeta}_t,\pi_t} \rangle.$$

The sum of the first term across $t = 1, \ldots, T$ can be bounded by employing online convex optimization algorithm. We can bound the second term as follows.

$$\begin{aligned} \langle \widehat{\boldsymbol{\lambda}}^* - \boldsymbol{\lambda}_t, \boldsymbol{\Psi} \boldsymbol{v}_{\boldsymbol{\zeta}_t,\pi_t} - \widehat{\boldsymbol{\Psi v}}_{\boldsymbol{\zeta}_t,\pi_t} \rangle &\leq \|\widehat{\boldsymbol{\lambda}}^* - \boldsymbol{\lambda}_t\|_{(\widehat{\boldsymbol{\Lambda}}_n + \boldsymbol{I}/n)^{-1}} \|\boldsymbol{\Psi} \boldsymbol{v}_{\boldsymbol{\zeta}_t,\pi_t} - \widehat{\boldsymbol{\Psi v}}_{\boldsymbol{\zeta}_t,\pi_t}\|_{\widehat{\boldsymbol{\Lambda}}_n + \boldsymbol{I}/n} \\ &\leq \underbrace{\|\widehat{\boldsymbol{\lambda}}^* - \boldsymbol{\lambda}_t\|_{\widehat{\boldsymbol{\Lambda}}_n^\dagger}}_{(i)} \underbrace{\|\boldsymbol{\Psi} \boldsymbol{v}_{\boldsymbol{\zeta}_t,\pi_t} - \widehat{\boldsymbol{\Psi v}}_{\boldsymbol{\zeta}_t,\pi_t}\|_{\widehat{\boldsymbol{\Lambda}}_n + \boldsymbol{I}/n}}_{(ii)} \end{aligned}$$

where the second inequality follows since $\widehat{\boldsymbol{\lambda}}^* - \boldsymbol{\lambda}_t$ is in the column space of $\widehat{\boldsymbol{\Lambda}}_n$. (ii) can be bounded using Lemma 14. (i) can be bounded by the following technical lemma. See Appendix B.2.1 for a proof.

Lemma 15. *For any* $\boldsymbol{\lambda}(\boldsymbol{c}) = \frac{1}{n}\sum_{k=1}^n c_k \boldsymbol{\varphi}(s_k, a_k)$ *with* $c_k \in [-B, B]$*,* $k = 1, \ldots, n$*, we have*

$$\|\boldsymbol{\lambda}(\boldsymbol{c})\|^2_{\widehat{\boldsymbol{\Lambda}}_n^\dagger} \leq dB^2.$$

Combining all the bounds, we get the following.

$$\frac{1}{T}\sum_{t=1}^{T} \textsc{Reg}_t^\lambda \leq \widetilde{\mathcal{O}}\left(\frac{C^* d^{3/2}}{1-\gamma} \sqrt{\frac{\log(dnT/\delta)}{n}} \right) + \epsilon_{\text{opt}}^\lambda(T)$$

where $\widetilde{\mathcal{O}}$ hides $\log\log|\mathcal{A}|$. See Appendix B.4.3 for details.

3.3.2 Algorithm and Regret Bound

Motivated by the regret decomposition analysis presented in the previous section, a primal-dual algorithm is introduced that operates over T steps. At each step t, the three players—the λ-player, the ζ-player, and the π-player—select actions denoted by $\boldsymbol{\lambda}_t$, $\boldsymbol{\zeta}_t$, and π_t, respectively.

Since the analysis requires the ζ-player to act greedily with respect to the previously chosen strategies of the other players, the algorithm is structured such that the λ-player and the π-player play first, selecting $\boldsymbol{\lambda}_t$ and π_t, respectively, after which the ζ-player responds with a greedy choice of $\boldsymbol{\zeta}_t$.

```
Algorithm 5: Primal-Dual Algorithm for Offline Linear MDPs
Input: Dataset $\mathcal{D} = \{(s_j, a_j, r_j, s'_j)\}_{j=1}^n$, no-regret online convex optimization oracle OCO.
Initialize: $\pi_1$ uniform, $\boldsymbol{c}'_1 \leftarrow \boldsymbol{0}$, $\alpha \leftarrow \sqrt{\log|\mathcal{A}|/T}$.
for $t = 1, \dots, T$ do
    $\boldsymbol{\zeta}_t \leftarrow \mathrm{argmin}_{\boldsymbol{\zeta}\in\mathbb{B}_d(D_\zeta)}\langle \boldsymbol{\zeta}, \boldsymbol{\Phi}^T\widehat{\boldsymbol{\mu}}_{\boldsymbol{\lambda}(\boldsymbol{c}'_t),\pi_t} - \boldsymbol{\lambda}(\boldsymbol{c}'_t)\rangle$.
    $\boldsymbol{\lambda}(\boldsymbol{c}_{t+1}) \leftarrow \mathrm{OCO}(\boldsymbol{\theta} - \boldsymbol{\zeta}_t + \gamma\widehat{\boldsymbol{\Psi}\boldsymbol{v}}_{\boldsymbol{\zeta}_t,\pi_t}; \mathcal{C}_n(C^*))$.
    Convert $\boldsymbol{c}_{t+1}$ to $\boldsymbol{c}'_{t+1}$ using Definition 1.
    $\pi_{t+1} \leftarrow \sigma(\alpha\sum_{i=1}^t \boldsymbol{\Phi}\boldsymbol{\zeta}_i)$.
Return: $\bar{\pi} = \mathrm{Unif}(\pi_1, \dots, \pi_T)$
```

The regret analysis of the three players in the previous section leads to our main result in the following theorem.

Theorem 4. *Under Assumptions 1 and 4, as long as T is at least $\Omega(\frac{d\log|\mathcal{A}|}{(1-\gamma)^2\epsilon^2})$, the policy $\bar{\pi}$ produced by Algorithm 5 satisfies $J(\bar{\pi}) \geq J(\pi^*) - \epsilon$ with probability at least $1-\delta$ as long as*

$$n \geq \Omega\left(\frac{(C^*)^2 d^3 \log(dn(\log|\mathcal{A}|)/(\delta\epsilon(1-\gamma)))}{(1-\gamma)^2\epsilon^4}\right).$$

Our work is an improvement over the work by Gabbianelli et al. [2024a] who give $\widetilde{\mathcal{O}}(\frac{(C^*)^2 d^2 \log|\mathcal{A}|}{(1-\gamma)^4\epsilon^2})$ sample complexity.

3.3.3 Extension to Constrained RL Setting

This section extends the algorithm and accompanying analysis developed in the previous section for offline unconstrained reinforcement learning to the offline constrained reinforcement learning setting. Recall that the constrained reinforcement learning problem is formulated as the optimization problem in (3.1).

Algorithm 6: Primal-Dual Algorithm for Offline Linear CMDPs

```
Input: Dataset $\mathcal{D} = \{(s_j, a_j, r_j, s'_j)\}_{j=1}^n$, $D_w$, $\boldsymbol{\tau}$
Initialize: $\pi_1$ uniform, $\boldsymbol{c}'_1 \leftarrow \mathbf{0}$, $\alpha \leftarrow \sqrt{\log|\mathcal{A}|/T}$.
for $t = 1, \dots, T$ do
    $\boldsymbol{\zeta}_t \leftarrow \mathrm{argmin}_{\boldsymbol{\zeta} \in \mathbb{B}_d(D_\zeta)} \langle \boldsymbol{\zeta}, \boldsymbol{\Phi}^T \widehat{\boldsymbol{\mu}}_{\boldsymbol{\lambda}(\boldsymbol{c}'_t), \pi_t} - \boldsymbol{\lambda}(\boldsymbol{c}'_t) \rangle$.
    $\boldsymbol{w}_t \leftarrow \mathrm{argmin}_{\boldsymbol{w} \in D_w \boldsymbol{\Delta}^I} \langle \boldsymbol{w}, \boldsymbol{\tau} - \boldsymbol{\Theta}^T \boldsymbol{\lambda}_t \rangle$.
    $\boldsymbol{\lambda}(\boldsymbol{c}_{t+1}) \leftarrow \mathrm{OCO}(\boldsymbol{\theta}_0 - \boldsymbol{\zeta}_t + \boldsymbol{\Theta} \boldsymbol{w}_t + \gamma \widehat{\boldsymbol{\Psi} \boldsymbol{v}}_{\boldsymbol{\zeta}_t, \pi_t}; \mathcal{C}_n(C^*))$.
    Convert $\boldsymbol{c}_{t+1}$ to $\boldsymbol{c}'_{t+1}$ using Definition 1.
    $\pi_{t+1} \leftarrow \sigma(\alpha \sum_{i=1}^t \boldsymbol{\Phi} \boldsymbol{\zeta}_i)$.
Return: $\bar{\pi} = \mathrm{Unif}(\pi_1, \dots, \pi_T)$
```

Analogously to the linear MDP setting discussed previously, to enable sample efficient learning over arbitrarily large state spaces, the following linear structure is assumed on the CMDP.

Assumption 5 (Linear CMDP). *We assume that the transition and the reward functions can be expressed as a linear function of a known feature map $\boldsymbol{\varphi} : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$ such that*

$$r_i(s, a) = \langle \boldsymbol{\varphi}(s, a), \boldsymbol{\theta}_i \rangle, \quad P(s'|s, a) = \langle \boldsymbol{\varphi}(s, a), \boldsymbol{\psi}(s') \rangle$$

for all $(s, a, s') \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}$ and $i = 1, \dots, I$, where $\boldsymbol{\theta}_i \in \mathbb{R}^d$ are known parameters and $\boldsymbol{\psi} = (\psi_1, \dots, \psi_d)$ is a vector of d unknown (signed) measures on $\mathcal{S}$.

Also, similarly to the linear MDP setting, we require the data coverage assumption (Assumption 4) that says

$$\frac{\mu^{\pi^*}(s, a)}{\mu_D(s, a)} \leq C^*$$

for a known upper bound C^* for all $(s, a) \in \mathcal{S} \times \mathcal{A}$ where π^* is optimal for the optimization problem (3.1).

The structure of our algorithm for the constrained RL setting is similar to that for the unconstrained RL setting. The difference is that we add the w-player that adjusts the weights on the rewards $r_1, \dots, r_I$. Closely following the analysis for the unconstrained setting, we can show the following sample complexity for the constrained setting.

Theorem 5. *Under Assumptions 4,3 and 5, the policy $\bar{\pi}$ produced by Algorithm 5 with threshold $\boldsymbol{\tau}$ and $D_w = 1 + \frac{1}{\phi}$ and T at least $\Omega(\frac{d \log |\mathcal{A}|}{(1-\gamma)^2 \epsilon^2})$ and large enough such that $\epsilon^\lambda_{opt}(T) \leq \epsilon$ satisfies $J_0(\bar{\pi}) \geq J_0(\pi^*) - \epsilon$ and $J_i(\bar{\pi}) \geq \tau_i - \epsilon$ with probability at least $1 - \delta$ as long as the sample size is*

$$n \geq \Omega\left(\frac{(C^*)^2 d^3 \log(dn(\log|\mathcal{A}|)/(\delta \phi \epsilon (1-\gamma)))}{(1-\gamma)^2 \phi^2 \epsilon^2} \right).$$

See Appendix B.5 for details.

3.3.4 Discussion

The preceding discussion relies on the concentrability assumption (Assumption 4), which requires that $\mu^*(s,a)/\mu(s,a) \leq C^*$ for all $s \in \mathcal{S}$ and $a \in \mathcal{A}$. With a careful analysis, the same result can be shown under a weaker assumption based on the notion of feature coverage introduced in Gabbianelli et al. [2024a]:

Assumption 6 (Feature coverage)**.** *For an optimal policy π^*, we have*

$$(\boldsymbol{\lambda}^*)^T(\boldsymbol{\Lambda}^\dagger)^2\boldsymbol{\lambda}^* \leq C^* \quad \textit{and} \quad \boldsymbol{\lambda}^* \in \textit{Col}(\boldsymbol{\Lambda})$$

where $\boldsymbol{\lambda}^* := \mathbb{E}_{\mu^*}[\boldsymbol{\varphi}(s,a)]$ and $\textit{Col}(\cdot)$ is the column space.

The feature coverage assumption requires $\boldsymbol{\lambda}^*$, the expected occupancy of the target policy in the feature space, to be covered by covariance matrix induced by the data distribution μ_B. Under the feature coverage assumption, $\boldsymbol{\lambda}^*$ can be approximated by a linear combination of $\boldsymbol{\varphi}(s_k, a_k)$, $k = 1, \ldots, n$ (Lemma 45). This result is analogous to Lemma 13 that uses concentrability assumption instead. It follows that the result in Theorem 4 with the concentrability assumption (Assumption 4) replaced by the feature coverage assumption (Assumption 6). A limitation of our work is that we use a stronger notion of feature coverage compared to the one used by Gabbianelli et al. [2024a], who assume $(\boldsymbol{\lambda}^*)^T(\boldsymbol{\Lambda}^\dagger)\boldsymbol{\lambda}^*$ is bounded. However, they require the knowledge of $\boldsymbol{\Lambda}$ and their sample complexity is $\widetilde{\mathcal{O}}(\epsilon^{-4})$.

Very recently, Neu and Okolo [2025b] improve the primal-dual algorithm presented in this thesis to work with the weaker notion of feature coverage.

3.3.5 Related Work

In Table 3.1, we compare our work to previous works. The column N shows how the sample complexity bound scales with the error tolerance ϵ. The first five algorithms are for offline RL with general function approximation. The algorithms can be reduced to the linear function approximation setting by taking a value function class consisting of linear functions. The computational efficiency of algorithms for the general function approximation setting is judged based on the efficiency when applied to linear function class. As the table shows, our algorithm is the first computationally efficient algorithm with sample complexity $\mathcal{O}(\epsilon^{-2})$ for finding ϵ-optimal policy under partial data coverage assumption. Moreover, our algorithm supports constraints on additional reward signals.

Table 3.1: Comparison of algorithms for offline RL with linear MDPs

Algorithm	Partial coverage	Comp' efficient	Support constraints	N
FQI [Munos and Szepesvári, 2008]	No	Yes	No	ϵ^{-2}
CBPL [Le et al., 2019]	No	Yes	Yes	ϵ^{-2}
Minimax [Xie et al., 2021]	Yes	No	No	ϵ^{-2}
CPPO [Uehara and Sun, 2022]	Yes	No	No	ϵ^{-2}
Minimax [Zanette, 2023]	No	No	No	ϵ^{-2}
Primal-Dual [Gabbianelli et al., 2024a]	Yes	Yes	No	ϵ^{-4}
Primal-Dual [Neu and Okolo, 2025b]	Yes	Yes	Yes	ϵ^{-2}
Primal-Dual [Hong and Tewari, 2024]	Yes	Yes	Yes	ϵ^{-2}

Offline RL with General Function Approximation Offline RL with general function approximation is widely studied in the discounted infinite-horizon setting. When casting the linear function approximation setting to the general function approximation setting, we get the realizability and Bellman completeness for free when using linear function class since the value function under linear function approximation is linear. In Table 3.1, we only compared works on general function approximation that assumes realizability, Bellman completeness and data coverage. There are other works that relax Bellman completeness assumption at the cost of introducing another assumption. For example, Xie and Jiang [2020], Zhan et al. [2022], Zhu et al. [2023] relax Bellman completeness assumption and introduce marginalized importance weight assumption.

Offline RL with Episodic Setting Offline RL with linear function approximation has been studied in the finite-horizon episodic setting. Zanette et al. [2021] propose a computationally efficient actor-critic algorithm with pessimism to achieve $\mathcal{O}(\epsilon^{-2})$ sample complexity under partial data coverage. Jin et al. [2021] propose a computationally efficient value iteration based algorithm with pessimism to achieve $\mathcal{O}(\epsilon^{-2})$ sample complexity under partial data coverage. However, they require the knowledge of the covariance matrix induced by the state-action data distribution. Although their results are computationally efficient and work under partial data coverage, they do not apply to the infinite-horizon discounted setting. Wu et al. [2021] study offline constrained RL with a more general way of specifying constraints. Their focus is on episodic setting with linear mixture MDP.

3.4 General Function Approximation Setting

In this section, we study a more general function approximation setting. Introducing Lagrangian multipliers $\boldsymbol{V} \in \mathbb{R}^{|\mathcal{S}|}$ for the Bellman flow constraints in the linear program (3.2), the Lagrangian function is

$$L(\boldsymbol{\mu}, \boldsymbol{V}) := \langle \boldsymbol{\mu}, \boldsymbol{r} \rangle + \langle \boldsymbol{V}, (1-\gamma)\boldsymbol{d}_0 + \gamma \boldsymbol{P}^\top \boldsymbol{\mu} - \boldsymbol{E}^\top \boldsymbol{\mu} \rangle$$

By the saddle point theorem for convex optimization (e.g. [Boyd and Vandenberghe, 2004]), a saddle point $(\boldsymbol{\mu}^*, \boldsymbol{V}^*)$ of L over $\mathbb{R}_+^{|\mathcal{S}\times\mathcal{A}|} \times \mathbb{R}^{|\mathcal{S}|}$ is optimal. Also, as shown by Zhan et al. [2022], such a saddle point $(\boldsymbol{\mu}^*, \boldsymbol{V}^*)$ is a saddle point of L when the domains are restricted to $\mathcal{U} \times \mathcal{V}$ provided that $\boldsymbol{\mu}^* \in \mathcal{U}$ and $\boldsymbol{V}^* \in \mathcal{V}$ (see Lemma 56 in Appendix). However, they show that a saddle point of L over $\mathcal{U} \times \mathcal{V}$ is not necessarily optimal as shown in the following proposition.

Proposition 1. *There exists a MDP $(\mathcal{S}, \mathcal{A}, r, \gamma, d_0)$ and function classes $\mathcal{U} \subseteq \mathbb{R}_+^{\mathcal{S}\times\mathcal{A}}$ and $\mathcal{V} \subseteq \mathbb{R}^{\mathcal{S}}$ with $\mu^* \in \mathcal{U}$ and $V^* \in \mathcal{V}$ such that there exists a saddle point $(\widehat{\mu}, \widehat{V})$ of the Lagrangian function L associated with the linear programming formulation of the MDP over $\mathcal{U} \times \mathcal{V}$, such that the extracted policy $\pi(\widehat{\mu})$ is not optimal.*

This proposition suggests that finding a saddle point of L over $\mathcal{U} \times \mathcal{V}$ may not give an optimal solution even when $\mathcal{U} \times \mathcal{V}$ contains an optimal solution. As a workaround of this problem, Zhan et al. [2022] add a regularization term $-\alpha \mathbb{E}_\mu[f(\mu/\mu_D)]$ to enforce strong concavity L in terms of μ. However, this approach leads to sample complexity for finding ε near optimal policy scales with $1/\varepsilon^4$. In our paper, we propose a simple workaround that does not affect the sample complexity. Specifically, we add an assumption that the function class $\mathcal{V}$ is rich enough:

Assumption 7 (All-Policy State-Value Realizability)**.** *We have $V^\pi \in \mathcal{V}$ for all stationary policy π.*

With this stronger assumption on $\mathcal{V}$, we can show that a saddle point of L over $\mathcal{U} \times \mathcal{V}$ is optimal, as formally stated in the following:

Lemma 16. *Consider function classes $\mathcal{U} \subseteq \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$ and $\mathcal{V} \subseteq \mathbb{R}^{\mathcal{S}}$ such that $\mu^* \in \mathcal{U}$ and $V^* \in \mathcal{V}$. Suppose $V^\pi \in \mathcal{V}$ for all policy π. Then, for any saddle point $(\widehat{\mu}, \widehat{V})$ of $L(\cdot,\cdot)$ over $\mathcal{U} \times \mathcal{V}$, the extracted policy $\pi(\widehat{\mu})$ is optimal.*

With the result in the proposition above, we consider an algorithm that finds a saddle point $(\widehat{\mu}, \widehat{V})$ that solves $\max_{\mu\in\mathcal{U}} \min_{V\in\mathcal{V}} L(\mu, V)$, then returns the policy $\widehat{\pi} = \pi(\widehat{\mu})$ extracted from $\widehat{\mu}$. Since computing the Lagrangian function requires the knowledge of the transition probability kernel P, we need to estimate the Lagrangian function. To facilitate the estimation, we change the measure from μ to μ_D by multiplying the importance weight $w = \mu/\mu_D$ and express the Lagrangian in terms of w instead of μ as follows.

$$L(\boldsymbol{w}, \boldsymbol{V}) := (1-\gamma)V(s_0) + \mathbb{E}_{\mu_D}[(w(s,a)(r(s,a) + \gamma[PV](s,a) - V(s))].$$

Analogous to Lemma 16, we can show that if a function class $\mathcal{W}$ for the MIW contains the optimal $w^* = \mu^*/\mu_D$ and a function class $\mathcal{V}$ contains $V^\pi \in \mathcal{V}$ for all stationary policy π, then a saddle point of $L(w, V)$ over $\mathcal{W} \times \mathcal{V}$ is optimal. In addition to the realizability assumption on $\mathcal{V}$ (Assumption 7), we make the following realizability assumption on $\mathcal{W}$.

Assumption 8. *For an optimal policy π^*, we have $w^* = \mu^{\pi^*}/\mu_D \in \mathcal{W}$.*

With the realizability assumptions, we consider an algorithm that finds a saddle point of the estimated Lagrangian:

$$\max_{w \in \mathcal{W}} \min_{V \in \mathcal{V}} \; \widehat{L}(w, V) := (1-\gamma)V(s_0) + \frac{1}{n}\sum_{i=1}^{n} w(s_i, a_i)(r(s_i, a_i) + \gamma V(s_i') - V(s_i)) \tag{3.14}$$

where $\widehat{L}(w, V)$ is an unbiased estimate of $L(w, V)$. After finding a saddle point $(\widehat{w}, \widehat{V})$, we need to extract policy from $\widehat{w}$. In general, the policy extraction from an importance weight $w \in \mathcal{W}$ can be done by the policy extraction from the corresponding occupancy measure $\mu = w \cdot \mu_D$ and using (3.4):

$$[\pi(w)](a|s) = \frac{w(s,a)\mu_D(s,a)}{\sum_{a'} w(s,a')\mu_D(s,a')} \;\text{ if } \sum_{a'} w(s,a')\mu_D(s,a') > 0, \quad \frac{1}{|\mathcal{A}|} \text{ otherwise.} \tag{3.15}$$

The policy $\pi(\widehat{w})$ extracted from a solution $(\widehat{w}, \widehat{V})$ of the saddle point problem (3.14) gives the following guarantee.

Theorem 6. *Under Assumptions 4, 7 and 8, if $(\widehat{w}, \widehat{V})$ is a saddle point that solves (3.14), then the extracted policy $\widehat{\pi} = \pi(\widehat{w})$ satisfies*

$$J(\widehat{\pi}) \geq J(\pi^*) - \varepsilon_n$$

with probability at least $1-\delta$ where $\varepsilon_n = \mathcal{O}(\frac{C^*}{1-\gamma}\sqrt{\log(|\mathcal{W}||\mathcal{V}|/\delta)/n})$.

Note that the theorem statement assumes the function classes $\mathcal{W}$ and $\mathcal{V}$ are finite. We can generalize this result to infinite function classes using a covering argument, which replaces the cardinalities of the function classes with their covering numbers. One drawback of this approach, common in approaches for offline RL that uses linear programming formulation, is that the policy extraction step requires the knowledge of μ_D. However, it may be impractical to have such a knowledge. As a workaround, Zhan et al. [2022] propose a behavior cloning to learn the policy $\pi(\mu_D)$. However, such an approach requires access to an additional function class for the behavior policy, which may be impractical. In the next section, we propose a Lagrangian decomposition and reparameterizing trick as a workaround.

3.4.1 Lagrangian Decomposition

To handle the case where the data generating distribution μ_D is unknown, we use the Lagrangian decomposition that adds a dummy occupancy measure variable μ to the original linear program (3.2):

$$\begin{aligned} \max_{\boldsymbol{\mu}\geq \mathbf{0}, \boldsymbol{\nu}\geq \mathbf{0}} \quad & \langle \boldsymbol{\mu}, \boldsymbol{r} \rangle \\ \text{subject to} \quad & \boldsymbol{E}^\top \boldsymbol{\nu} = (1-\gamma)\boldsymbol{d}_0 + \gamma \boldsymbol{P}^\top \boldsymbol{\mu} \\ & \boldsymbol{\mu} = \boldsymbol{\nu}. \end{aligned}$$

Introducing the Lagrangian multiplier $\boldsymbol{Q} \in \mathbb{R}^{|\mathcal{S}\times\mathcal{A}|}$ for the constraint $\boldsymbol{\mu} = \boldsymbol{\nu}$, the Lagrangian function becomes

$$\begin{aligned} L(\boldsymbol{\mu}, \boldsymbol{\nu}; \boldsymbol{Q}, \boldsymbol{V}) &= (1-\gamma)\langle \boldsymbol{d}_0, \boldsymbol{V} \rangle + \langle \boldsymbol{\mu}, \boldsymbol{r} + \gamma \boldsymbol{P}\boldsymbol{V} - \boldsymbol{Q} \rangle + \langle \boldsymbol{\nu}, \boldsymbol{Q} - \boldsymbol{E}\boldsymbol{V} \rangle \\ &= \langle \boldsymbol{r}, \boldsymbol{\mu} \rangle + \langle \boldsymbol{V}, (1-\gamma)\boldsymbol{d}_0 + \gamma \boldsymbol{P}^\top \boldsymbol{\mu} - \boldsymbol{E}^\top \boldsymbol{\nu} \rangle + \langle \boldsymbol{Q}, \boldsymbol{\nu} - \boldsymbol{\mu} \rangle. \end{aligned}$$

The Lagrangian decomposition [Shepardson and Marsten, 1980, Guignard and Kim, 1987] is a classical method for decomposing an optimization into subproblems by introducing copies of optimization variables. To our knowledge the method is first used in the context of MDP by Mehta and Meyn [2009]. Since then, the method has been used in RL in various settings [Neu and Okolo, 2023, Gabbianelli et al., 2024b, Hong and Tewari, 2024, Neu and Okolo, 2025a].

It appears that we have made the problem more complex for nothing, but with a reparameterization trick, we can expose a policy variable π, so that the policy extraction step (3.15) can be bypassed, eliminating the need to know the data generating distribution μ_D. The reparameterization trick is first introduced by Gabbianelli et al. [2024b] for offline RL with linear MDPs, but without a clear justification. In the remainder of this section we discuss the reparameterization trick, and argue that naive adaptation of the trick to the general function approximation setting may fail.

We motivate the trick by observing that an optimal μ^* and ν^* satisfy

$$\nu^*(s,a) = \pi^*(a|s)\nu^*(s) = \pi^*(a|s)((1-\gamma)d_0(s) + \gamma[P^\top \mu^*](s))$$

where π^* is the policy extracted from ν^*. Note that we expressed ν^* in terms of π^* and μ^*. For the purpose of finding an optimal μ^* and ν^*, we can restrict the search space for (μ^*, ν^*) from $\mathbb{R}_+^{|\mathcal{S}\times\mathcal{A}|} \times \mathbb{R}_+^{|\mathcal{S}\times\mathcal{A}|}$ to

$$\mathcal{F}(\mathcal{U}, \Pi) := \{(\mu, \nu_{\pi,\mu}) : \mu \in \mathcal{U}, \pi \in \Pi\}$$

where Π is a policy class that contains an optimal policy π^* and $\mathcal{U}$ is a function class that contains μ^*, and

$$\nu_{\pi,\mu}(s,a) := \pi(a|s)((1-\gamma)d_0(s) + \gamma[P^\top \mu](s)).$$

Necessarily, any $(\mu, \nu) \in \mathcal{F}(\mathcal{U}, \Pi)$ satisfies the Bellman flow constraint, and it follows that the Lagrangian function is not dependent on the corresponding Lagrangian multiplier V. With a slight abusive notation, we write the Lagrangian function as $L(\mu, \pi; Q) = L(\mu, \nu_{\pi,\mu}; Q)$, suppressing the dependency on V, when $(\boldsymbol{\mu}, \boldsymbol{\nu})$ is restricted to $\mathcal{F}(\mathcal{U}, \Pi)$. With this restriction, a natural algorithm is to find a saddle point of L over $\mathcal{F}(\mathcal{U}, \Pi) \times \mathcal{Q}$, then extract a policy. However, the following shows that such an algorithm may fail.

Proposition 2. *There exists a MDP $(\mathcal{S}, \mathcal{A}, r, \gamma, d_0)$ and function classes $\mathcal{U} \subseteq \mathbb{R}_+^{\mathcal{S}\times\mathcal{A}}$, $\mathcal{Q} \subseteq \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$ and policy class Π, with $\mu^* \in \mathcal{U}$, $Q^* \in \mathcal{Q}$ and $\pi^* \in \Pi$ such that there exists a saddle point $(\widehat{\mu}, \nu_{\widehat{\mu},\pi}; \widehat{Q})$ of the Lagrangian function L associated with the linear programming formulation of the MDP over $\mathcal{F}(\mathcal{U}, \Pi) \times \mathcal{Q}$, where the policy $\widehat{\pi}$ is not optimal.*

To deal with the problem of spurious saddle points, we make the following assumption.

Assumption 9 (All-Policy State-Action Value Function Realizability)**.** *We have $Q^\pi \in \mathcal{Q}$ for all stationary policy π.*

Indeed, with the assumption above, we get the following guarantee.

Lemma 17. *Consider function classes $\mathcal{U} \subseteq \mathbb{R}^{|\mathcal{S}\times\mathcal{A}|}$ and $\mathcal{Q} \subseteq \mathbb{R}^{|\mathcal{S}\times\mathcal{A}|}$, such that $\mu^* \in \mathcal{U}$ and $Q^\pi \in \mathcal{Q}$ for all policy $\pi \in \Pi$, where Π is a function class that contains an optimal policy π^*. If $(\widehat{\mu}, \nu_{\widehat{\mu},\widehat{\pi}}; \widehat{Q})$ is a saddle point of L over $\mathcal{F}(\mathcal{U}, \Pi) \times \mathcal{Q}$, then the policy $\widehat{\pi}$ is optimal.*

As done in the previous section, we can change variable $w = \mu/\mu_D$ for estimating the

Lagrangian as

$$\begin{aligned} L(\mu,\pi;Q) &= \langle \boldsymbol{\mu}, \boldsymbol{r}\rangle + \langle \boldsymbol{Q}, \boldsymbol{\nu}_{\mu,\pi} - \boldsymbol{\mu}\rangle = (1-\gamma)\langle Q(\cdot,\pi), \boldsymbol{d}_0\rangle + \langle \boldsymbol{\mu}, \boldsymbol{r} + \gamma \boldsymbol{P} Q(\cdot,\pi) - \boldsymbol{Q}\rangle \\ &\approx (1-\gamma)Q(s_0,\pi) + \tfrac{1}{n}\textstyle\sum_{i=1}^n w(s_i,a_i)(r(s_i,a_i) + \gamma Q(s_i',\pi) - Q(s_i,a_i)) \\ &=: \widehat{L}(w,\pi;Q) \end{aligned}$$

where $\pi \in \Pi$ is the policy that parameterizes $\nu = \nu_{\mu,\pi}$ and $Q(s,\pi) = \sum_a \pi(a|s)Q(s,a)$. With the estimate, we consider an algorithm that finds a saddle point $(\widehat{w}, \widehat{\pi}; \widehat{Q})$ of the estimate

$$\max_{w\in\mathcal{W},\pi\in\Pi} \min_{Q\in\mathcal{Q}} \widehat{L}(w,\pi;Q), \tag{3.16}$$

then returns $\widehat{\pi}$. This algorithm has the following guarantee.

Theorem 7. *Let Π be a function class for policies such that $\pi^* \in \Pi$ where π^* is an optimal policy. Under Assumptions 4, 8 and 9, if $(\widehat{w}, \widehat{\pi}; \widehat{Q})$ is a saddle point that solves* (3.16)*, then we have*

$$J(\widehat{\pi}) \geq J(\pi^*) - \varepsilon_n$$

with probability at least $1-\delta$ where $\varepsilon_n = \mathcal{O}(\frac{C^*}{1-\gamma}\sqrt{\log(|\mathcal{W}||\mathcal{Q}||\Pi|/\delta)/n})$.

We have not yet addressed the computational challenges of finding the saddle point. In the next section, we introduce an oracle-efficient primal-dual algorithm that approximately solves the saddle point problem. As discussed there, this algorithm does not require access to a policy class Π.

3.4.2 Oracle-Efficient Primal-Dual Algorithm

We propose an oracle-efficient primal-dual algorithm that solves the saddle point problem (3.16). Inspired by Xie et al. [2021], Cheng et al. [2022], we view the problem of finding a saddle point as a Stackelberg game between the primal variables w and π, and the dual variable Q, with the w-player and the π-player as the leaders and the Q-player as the follower:

$$\begin{aligned} (w^*,\pi^*) &\in \operatorname*{argmax}_{w\in\mathcal{W},\pi\in\Pi} \widehat{L}(w,\pi;Q_{w,\pi}) \\ \text{s.t. } \; Q_{w,\pi} &\in \operatorname*{argmin}_{Q\in\mathcal{Q}} \widehat{L}(w,\pi;Q). \end{aligned}$$

Our primal-dual algorithm aims to solve this game iteratively by optimizing for the leader variables and the follower variable, alternately. Specifically, the w-player employs a no-regret algorithm using a no-regret oracle that enjoys the following guarantee.

Definition 3 (No-Regret Oracle)**.** *Given a set $\mathcal{X} \subseteq \mathbb{R}^{|\mathcal{S}\times\mathcal{A}|}$, we say $\mathcal{O}$ is a no-regret oracle with respect to $\mathcal{X}$ if for any sequence of linear functions $h_t : \mathcal{X} \to \mathbb{R}$ with $|h_t(x)| \leq 1$ for all $x \in \mathcal{X}$, $t = 1, \ldots, T$, the oracle sequentially outputs $x_t \in \mathcal{X}$ for each t given the sequence $h_1, \ldots, h_{t-1}$ as an input such that*

$$\sum_{t=1}^{T} h_t(x^*) - h_t(x_t) \leq o(T)$$

where $x^* = \operatorname{argmax}_{x\in\mathcal{F}} \sum_{t=1}^{T} h_t(x)$.

The w-player employs a no-regret oracle with respect to the function class $\mathcal{W}$ to generate a sequence $w_1, \ldots, w_T \in \mathcal{W}$, based on the sequence of functions $\{\widehat{L}(\cdot, \pi_t; Q_t)\}_{t=1}^{T}$ (Line 2). The objective is to maximize $\sum_{t=1}^{T} \widehat{L}(w_t, \pi_t; Q_t)$, and the no-regret oracle ensures that the sequence competes effectively with the optimal fixed w^* in hindsight. An example of such an oracle is the online gradient ascent algorithm, which takes a gradient step followed by projection onto $\mathcal{W}$. See Appendix C.3 for details.

The π-player optimizes the policy using mirror descent updates (Line 3). Intuitively, the mirror descent updates increase the weight on the action with high action value. A key advantage is that they eliminate the need for an explicit policy class Π, as the policy class is implicitly defined by the value function class $\mathcal{Q}$. Specifically, with an additional assumption that $\mathcal{Q}$ is convex, the policies encountered during mirror descent are softmax policies of the form:

$$\Pi(\mathcal{Q}, C) := \left\{ \pi : \pi(\cdot|s) = \frac{\exp(cQ(s,\cdot))}{\sum_{a'} \exp(cQ(s,a'))}, Q \in \mathcal{Q}, c \in [0, C] \right\}.$$

The Q-player chooses a function $\mathcal{Q}$ that minimizes $\widehat{L}(w, \pi; \cdot)$ (Line 4). Due to the linearity of $\widehat{L}(w_t, \pi_t; \cdot)$, the minimizer can be computed given a linear optimization oracle that solves the problem of the form $\operatorname{argmin}_{Q\in\mathcal{Q}}\langle c, Q\rangle$. The algorithm runs for T iterations and returns a policy sampled uniformly at random from $\pi_1, \ldots, \pi_T$.

Algorithm 7: PDORL

```
Input: Dataset $\mathcal{D} = \{(s_i, a_i, s'_i)\}_{i=1}^{n}$, function classes $\mathcal{W}$ and $\mathcal{Q}$, no-regret oracle $\mathcal{O}$,
       learning rate $\alpha > 0$, number of iterations $T$.
1 for $t = 1, \ldots, T$ do
2     $w_t \leftarrow \mathcal{O}(\widehat{L}(\cdot, \pi_{t-1}; Q_{t-1}); \mathcal{W})$.
3     $\pi_t(\cdot|s) \propto \pi_{t-1}(\cdot|s) \exp(\alpha Q_{t-1}(\cdot, s))$, for $s \in \{s'_1, \ldots, s'_n\}$.
4     $Q_t \leftarrow \operatorname{argmin}_{Q\in\mathcal{Q}} \widehat{L}(w_t, \pi_t; Q)$.
Return: Uniform($\pi_1, \ldots, \pi_T$)
```

The policy returned by the algorithm has the following guarantee.

Theorem 8. *Under Assumptions 4, 8 and 9, running Algorithm 7 with a no-regret oracle* $\mathcal{O}$ *with sublinear regret* Reg_T, $\alpha = (1-\gamma)\sqrt{\log|\mathcal{A}|}/\sqrt{T}$ *and* T *large enough such that* $Reg_T/T \leq 1/\sqrt{n}$ *and* $T \geq n$*, we have with probability at least* $1-\delta$ *that*

$$J(\widehat{\pi}) \geq J(\pi^*) - \varepsilon_n$$

where $\varepsilon_n = \mathcal{O}\left(\frac{C^*}{1-\gamma}\sqrt{\log(\mathcal{N}_{1/\sqrt{n}}(\mathcal{W}, \|\cdot\|_\infty)\mathcal{N}_{1/(8\sqrt{n}T)}(\mathcal{Q}, \|\cdot\|_\infty)/\delta)/n}\right)$.

The proof of the theorem relies on the fact that the suboptimality of the returned policy can be decomposed into the regrets of the three players, and that each regret can be bounded using the property of the strategy employed by each player. See Appendix C.1.7 for detail.

3.4.3 Extension to Constrained RL

The corresponding linear program, with Lagrangian decomposition, is

$$\begin{aligned} \max_{\boldsymbol{\mu}\geq 0, \boldsymbol{\nu}\geq 0} \quad & \langle \boldsymbol{\mu}, \boldsymbol{r}_0 \rangle \\ \text{subject to} \quad & \boldsymbol{E}^\top \boldsymbol{\nu} = (1-\gamma)\boldsymbol{d}_0 + \gamma \boldsymbol{P}^\top \boldsymbol{\mu} \\ & \boldsymbol{\mu} = \boldsymbol{\nu} \\ & \langle \boldsymbol{\mu}, \boldsymbol{r}_i \rangle \geq \tau_i, \quad i = 1, \dots I. \end{aligned}$$

The corresponding Lagrangian function is

$$L(\boldsymbol{\mu}, \boldsymbol{\nu}; \boldsymbol{Q}, \boldsymbol{V}, \boldsymbol{\lambda}) = \langle \boldsymbol{\mu}, \boldsymbol{r}_0 \rangle + \langle \boldsymbol{V}, (1-\gamma)\boldsymbol{d}_0 + \gamma \boldsymbol{P}^\top \boldsymbol{\mu} - \boldsymbol{E}^\top \boldsymbol{\nu} \rangle + \langle \boldsymbol{Q}, \boldsymbol{\nu} - \boldsymbol{\mu} \rangle + \langle \boldsymbol{\lambda}, \boldsymbol{R}^\top \boldsymbol{\mu} - \boldsymbol{\tau} \rangle$$

where $\boldsymbol{R} \in \mathbb{R}^{|\mathcal{S}\times\mathcal{A}|\times I}$ is the matrix containing the reward functions $r_1, \dots, r_I$. As done for the unconstrained RL setting, we restrict (μ, ν) to be in $\mathcal{F}(\mathcal{U}, \Pi)$, so that $\nu(s,a) = \pi(a|s)((1-\gamma)d_0(s) + \gamma[P^\top \mu](s))$ for some $\pi \in \Pi$, and reparameterize μ as an importance weight $w = \mu/\mu_D$. Then, we can rewrite the Lagrangian and estimate it as:

$$\begin{aligned} & L(\boldsymbol{\mu}, \pi; \boldsymbol{Q}, \boldsymbol{\lambda}) \\ & = (1-\gamma)\langle Q(\cdot, \pi), \boldsymbol{d}_0 \rangle + \langle \boldsymbol{\mu}, \boldsymbol{r}_0 + \boldsymbol{R}\lambda + \gamma \boldsymbol{P} Q(\cdot, \pi) - \boldsymbol{Q} \rangle - \langle \boldsymbol{\lambda}, \boldsymbol{\tau} \rangle \\ & \approx (1-\gamma)Q(s_0, \pi) + \frac{1}{n}\sum_{j=1}^n w(s_j, a_j)((r_0 + \sum_{i=1}^I \lambda_i r_i)(s_j, a_j) + \gamma Q(s'_j, \pi) - Q(s_j, a_j)) - \langle \lambda, \tau \rangle \\ & =: \widehat{L}(w, \pi; Q, \lambda). \end{aligned}$$

To derive a uniform concentration bound for the Lagrangian estimate across all λ, we require a bound on λ. To achieve this, we assume the Slater's condition (Assumption 3)

Algorithm 8: PDOCRL

```
Input: Dataset $\mathcal{D}$, function classes $\mathcal{W}$ and $\mathcal{Q}$, learning rate $\alpha > 0$, dual bound
       $B > 0$, no-regret oracle $\mathcal{O}$, number of iterations $T$.
for $t = 1, \ldots, T$ do
    $w_t \leftarrow \mathcal{O}(\widehat{L}(\cdot, \pi_{t-1}; Q_{t-1}, \lambda_{t-1}); \mathcal{W})$.
    $\pi_t(\cdot|\cdot) \leftarrow \pi_{t-1}(\cdot|\cdot) \exp(\alpha Q_{t-1}(\cdot, \cdot))$.
    $Q_t \leftarrow \mathrm{argmin}_{Q \in \mathcal{Q}} \widehat{L}(w_t, \pi_t; Q, \lambda_{t-1})$.
    $\lambda_t \leftarrow \mathrm{argmin}_{\lambda \in B\Delta^I} \widehat{L}(w_t, \pi_t; Q_t, \lambda)$.
Return: Uniform$(\pi_1, \ldots, \pi_T)$
```

that ensures the existence of a feasible policy that satisfies the constraints with a positive margin φ. Then, it can be shown that the optimal dual variable is bounded as stated in the following lemma.

Lemma 18. *Consider a constrained optimization problem* (3.1) *with thresholds $\boldsymbol{\tau} = (\tau_1, \ldots, \tau_I)$. Suppose the problem satisfies Assumption 3 with margin $\varphi > 0$. Then, the optimal dual variable $\boldsymbol{\lambda}^\star$ of the problem satisfies $\|\boldsymbol{\lambda}^\star\|_1 \leq \frac{1}{\varphi}$.*

The lemma motivates an algorithm that solves the following saddle point problem:

$$\max_{w \in \mathcal{W}, \pi \in \Pi} \min_{Q \in \mathcal{Q}, \lambda \in \frac{1}{\varphi}\Delta^I} \widehat{L}(w, \pi; Q, \lambda) \tag{3.17}$$

where $\Delta^I = \{\lambda \in \mathbb{R}^I_+ : \sum_{i=1}^I \lambda_i \leq 1\}$. We can show that a solution $(\widehat{w}, \widehat{\pi}; \widehat{Q}, \widehat{\lambda})$ to the saddle point problem has the following guarantee.

Theorem 9. *Let Π be a function class for policies such that $\pi^* \in \Pi$ where π^* is an optimal policy. Under Assumptions 4, 8, 9 and 3, if $(\widehat{w}, \widehat{\pi}; \widehat{Q}, \widehat{\lambda})$ is a saddle point that solves* (3.17), *then $\widehat{\pi}$ satisfies*

$$J_0(\widehat{\pi}) \geq J_0(\pi^*) - \varepsilon_n, \quad J_i(\widehat{\pi}) \geq \tau_i - \varepsilon_n, \quad i = 1, \ldots, I.$$

where $\varepsilon_n = \mathcal{O}\left((\frac{1}{\varphi} + \frac{1}{1-\gamma})C^* \sqrt{\log(\mathcal{N}_{1/\sqrt{n}}(\mathcal{W}, \|\cdot\|_\infty)\mathcal{N}_{1/(8\sqrt{n}T)}(\mathcal{Q}, \|\cdot\|_\infty)/\delta)/n}\right)$.

For computational efficiency, we present a primal-dual algorithm. The algorithm design is similar to PDORL (Algorithm 7) for the unconstrained problem. The main difference is the introduction of the λ-player that greedily minimizes $\widehat{L}(w_t, \pi_t; Q_t, \lambda)$ at each time step t. We present below our algorithm called Primal-Dual algorithm for Offline Constrained RL (PDOCRL).

The policy returned by the algorithm has the following guarantee.

Table 3.2: Comparison of algorithms for offline RL with value function approximation

Algorithm	Partial coverage	Comp' efficient	Requires μ_D	N
FQI [Munos and Szepesvári, 2008]	No	Yes	No	ϵ^{-2}
Minimax [Xie et al., 2021]	Yes	No	No	ϵ^{-5}
Minimax [Zanette, 2023]	Yes	No	No	ϵ^{-2}
PRO-RL [Zhan et al., 2022]	Yes	No	Yes	ϵ^{-6}
A-Crab [Zhu et al., 2023]	Yes	No	No	ϵ^{-2}
ATAC [Cheng et al., 2022]	Yes	No	No	ϵ^{-3}
CORAL [Rashidinejad et al., 2022]	Yes	No	Yes	ϵ^{-2}
PDORL [Hong and Tewari, 2025b]	Yes	Yes	No	ϵ^{-2}

Theorem 10. *Under Assumptions 4, 8, 3 and $Q_i^\pi \in \mathcal{Q}$ for all $i = 0, 1, \ldots, I$ and for all stationary policy π, running Algorithm 8 with a no-regret oracle $\mathcal{O}$ with sublinear regret Reg_T, learning rate $\alpha = (1-\gamma)\sqrt{\log|\mathcal{A}|}/\sqrt{T}$, dual bound $B = 1 + \frac{1}{\varphi}$ and T large enough such that $Reg_T/T \leq 1/\sqrt{n}$ and $T \geq n$, we have with probability at least $1-\delta$ that*

$$J_0(\widehat{\pi}) \geq J_0(\pi^*) - \varepsilon_n, \quad J_i(\widehat{\pi}) \geq \tau_i - \varepsilon_n, \quad i = 1, \ldots, I.$$

where $\varepsilon_n = \mathcal{O}\left((\frac{1}{\varphi} + \frac{1}{1-\gamma})C^* \sqrt{\log(\mathcal{N}_{1/\sqrt{n}}(\mathcal{W}, \|\cdot\|_\infty)\mathcal{N}_{1/(8\sqrt{n}T)}(\mathcal{Q}, \|\cdot\|_\infty)/\delta)/n}\right)$.

3.4.4 Related Work

In this section, we compare our algorithm, PDORL (Algorithm 7), with prior work on offline reinforcement learning with function approximation, where the performance objective is the infinite-horizon discounted sum of rewards. See Table 3.2 for the comparison.

Note that our algorithm is the only oracle-efficient algorithm that achieves sample complexity of ϵ^{-2} under partial data coverage that does not require the knowledge of the data generating distribution μ_D.

Hong and Tewari [2024] propose a primal-dual algorithm where the primal variable is the policy and the dual variable is the Lagrangian multiplier associated with the constraint. The update rule for the dual variable requires checking whether the current policy violates the constraint, which demands data coverage for every policy encountered. Le et al. [2019] propose a primal-dual algorithm, using fitted Q-iteration and fitted Q-evaluation algorithms as subroutines. These likewise assume full data coverage and require a function class for the value function that is closed under the Bellman operation. Zhang et al. [2024] propose an oracle-efficient, linear programming based algorithm for offline constrained RL that achieves

Table 3.3: Comparison of algorithms for offline constrained RL

Algorithm	Partial coverage	Oracle efficient	Requires μ_D	N
PDCA [Hong and Tewari, 2024]	No	No	No	$\frac{(C^*)^2}{(1-\gamma)^2\epsilon^2}$
MBCL [Le et al., 2019]	No	Yes	No	$\frac{(C^*)^2}{(1-\gamma)^8\epsilon^2}$
POCC [Zhang et al., 2024]	Yes	Yes	Yes	$\frac{(C^*)^2}{(1-\gamma)^2\epsilon^2}$
PDOCRL [Hong and Tewari, 2025b]	Yes	Yes	No	$\frac{(C^*)^2}{(1-\gamma)^2\epsilon^2}$
Lower bound [Rashidinejad et al., 2021]				$\frac{C^*}{(1-\gamma)\epsilon^2}$

$\mathcal{O}(\epsilon^{-2})$ sample complexity with only partial data coverage. However, their approach rely on an auxiliary function class $\mathcal{X}$ that maps the Bellman flow error to its ℓ_1 norm. See Table 3.3 for a detailed comparison. Entries highlighted in red indicate suboptimality relative to our work. The last column is sample complexity for finding ϵ-optimal policy. C^* is the coefficient that measures the data coverage of an optimal policy.

3.5 Discussion

Throughout this chapter, we assumed that an upper bound C^* of the concentrability coefficient of an optimal policy is known to the learner. In practice, it is unreasonable to have such a knowledge.

By modifying the analysis, it can be shown that the algorithm returns a policy that competes with the best policy among those covered by the dataset. Therefore, even if the value of C^* is misspecified in the sense that it does not upper bound the concentrability coefficient of the optimal policy, the returned policy remains meaningful as long as C^* upper bounds the concentrability coefficient of a sufficiently good policy.

When the well-specified upper bound C^* is unknown, the learner faces a tradeoff. Selecting a large value of C^* allows the returned policy to compete with a broader class of policies, but at the expense of a looser suboptimality guarantee relative to the best policy in this larger set. On the other hand, selecting a smaller value of C^* yields a tighter suboptimality bound, but limits the comparison to a smaller set of policies. An interesting direction for future work is to develop principled methods for selecting C^* that effectively balance this tradeoff.

CHAPTER 4

Conclusion

This thesis advances the theoretical foundations of reinforcement learning (RL) by addressing two important challenges: online RL with the infinite-horizon average-reward criterion, and offline constrained RL under partial data coverage. In both settings, the emphasis lies in designing algorithms that are both statistically and computationally efficient, and in establishing finite-time performance guarantees.

In the first part (Chapter 2), we focused on the *online average-reward setting*, which is particularly relevant in applications where long-term performance is important. We introduced a family of algorithms that reduce the average-reward problem to the discounted setting using a carefully tuned discount factor. This reduction enabled us to leverage value-iteration-based methods while maintaining tight control over approximation errors. In the *tabular case*, we proposed a clipped value iteration algorithm (γ-`UCB-CVI`) that achieves sublinear regret by combining optimism with a span-constrained value function update. In the *linear MDP setting*, we introduced a novel planning architecture (γ-`LSCVI-UCB`) that decouples value iteration from policy execution and restarts planning episodes based on information growth. We also proposed a computationally efficient variant (γ-`DC-LSCVI-UCB`) that avoids explicit state enumeration by introducing an efficient clipping scheme and a deviation-controlled planning strategy.

In the second part (Chapter 3), we turned our attention to *offline constrained reinforcement learning*, where the agent must optimize a policy under safety constraints using only a fixed dataset. This setting is motivated by safety-critical domains where online exploration is infeasible. We developed a primal-dual algorithmic framework grounded in the linear programming formulation of constrained RL. In the *linear MDP case*, we showed that under partial feature coverage, it is possible to learn near-optimal policies while satisfying constraints. In the *general function approximation setting*, we extended our methods using a Lagrangian decomposition and proposed an oracle-efficient primal-dual algorithm that adapts to both unconstrained and constrained objectives. Importantly, our analysis establishes finite-sample guarantees without requiring the knowledge of the underlying data distribution of the offline dataset.

Future directions include designing *memory-efficient algorithms* for both settings. In the online setting, our algorithms require maintaining full value functions across value iterations, which may be costly when run for many time steps. For the offline constrained setting, our current algorithms provide guarantees on the *average policy* across primal-dual iterates, requiring storage of all intermediate policies. Developing a primal-dual algorithm with *last-iterate convergence*, guaranteeing near-optimality and constraint satisfaction at the final policy, would not only reduce storage requirements but also simplify policy deployment in practical systems.

Another important direction for future work is to relax assumptions such as the knowledge of $\mathrm{sp}(v^*)$ and C^*. While recent work in the tabular setting explores estimating $\mathrm{sp}(v^*)$ from data, extending such ideas to the linear or general function approximation settings remains open. Similarly, treating C^* as a tunable hyperparameter raises the question of how to balance the tradeoff between statistical guarantees and policy class coverage. Developing adaptive or data-driven methods for selecting these quantities would improve the practicality and robustness of algorithms in both online and offline settings.

APPENDIX A

Analysis of Infinite-Horizon Average-Reward RL

A.1 Tabular Setting

Central to the analysis of the concentration bound for the approximate Bellman backup is the following concentration bound for scalar-valued self-normalized processes.

Lemma 19 (Concentration of Scalar-Valued Self-Normalized Processes [Abbasi-Yadkori et al., 2012])**.** *Let $\{\varepsilon_t\}_{t=1}^{\infty}$ be a real-valued stochastic process with corresponding filtration $\{\mathcal{F}_t\}_{t=0}^{\infty}$. Let $\varepsilon_t|\mathcal{F}_{t-1}$ be zero-mean and σ-subgaussian. Let $\{Z_t\}_{t=0}^{\infty}$ be an $\mathbb{R}$-valued stochastic process where $Z_t \in \mathcal{F}_{t-1}$. Assume $W > 0$ is deterministic. Then for any $\delta > 0$, with probability at least $1-\delta$, we have for all $t \geq 0$ that*

$$\frac{(\sum_{s=1}^{t} Z_s\varepsilon_s)^2}{W + \sum_{s=1}^{t} Z_s^2} \leq 2\sigma^2 \log\left(\frac{\sqrt{W + \sum_{s=1}^{t} Z_s^2}}{\delta\sqrt{W}}\right).$$

A.1.1 Proof of Lemma 2

To show a bound for $|(\widehat{P}_t - P)V_t(s,a)|$ uniformly on $t \in [T]$, we use a covering argument on the function class that captures V_t. Note that the value functions V_t defined in the algorithm always lie in the following function class.

$$\mathcal{V}_{\text{tabular}} = \{v \in \mathbb{R}^{\mathcal{S}} : v(s) \in [0, \tfrac{1}{1-\gamma}] \text{ for all } s \in \mathcal{S}\}.$$

We first bound the error for a fixed value function in $\mathcal{V}_{\text{tabular}}$. Afterward, we will use a covering argument to get a uniform bound over $\mathcal{V}_{\text{tabular}}$.

Lemma 20. *Fix any $V \in \mathcal{V}_{tabular}$. There exists some constant C such that for any $\delta \in (0,1)$, with probability at least $1-\delta$, we have:*

$$|[(\widehat{P}_t - P)V](s,a)| \le C sp(v^*)\sqrt{\frac{\log(SAT/\delta)}{N_t(s,a)}}$$

for all $(s,a) \in \mathcal{S} \times \mathcal{A}$ and $t = 1, \dots, T$.

Proof. Fix any $(s,a) \in \mathcal{S} \times \mathcal{A}$. By definition, we have:

$$[(\widehat{P}_t - P)V](s,a) = \frac{1}{N_t(s,a)} \sum_{\tau=1}^{t} \mathbb{I}\{s_\tau = s, a_\tau = a\}[V(s_{\tau+1}) - [PV](s,a)].$$

Let $\varepsilon_t = V(s_{t+1}) - [PV](s_t, a_t)$, $Z_t = \mathbb{I}\{s_t = s, a_t = a\}$, and $W = 1$. Since the range of ε_t is bounded by $2 \cdot \mathrm{sp}(v^*)$, it is $\mathrm{sp}(v^*)$-subgaussian. By Lemma 19, we know for some constant C, with probability at least $1-\delta$, for all $t = 1, \dots, T$, we have

$$\begin{aligned}|[(\widehat{P}_t - P)V](s,a)| &= \frac{|\sum_{s=1}^t Z_s \varepsilon_s|}{1 + \sum_{s=1}^t Z_s^2} \\ &\le C \cdot \mathrm{sp}(v^*)\sqrt{\frac{\log(\sqrt{N_t(s,a)}/\delta)}{N_t(s,a)}} \\ &\le C \cdot \mathrm{sp}(v^*)\sqrt{\frac{\log(T/\delta)}{N_t(s,a)}}.\end{aligned}$$

Applying a union bound for all $(s,a) \in \mathcal{S} \times \mathcal{A}$ gives us the desired inequality.

□

We use $\mathcal{N}_\epsilon$ to denote the ϵ-covering number of $\mathcal{V}_{\text{tabular}}$ with respect to the distance $\text{dist}(V, V') = \|V - V'\|_\infty$. Using a grid of size ϵ, since functions in $\mathcal{V}_{\text{tabular}}$ has the range $[0, \frac{1}{1-\gamma}]$, it can be seen that $\log \mathcal{N}_\epsilon \le S \log \frac{1}{\epsilon(1-\gamma)}$. Now, we prove a uniform concentration bound using a covering argument on $\mathcal{V}_{\text{tabular}}$.

Proof of Lemma 2. Note that $V_t \in \mathcal{V}_{\text{tabular}}$ for all t. Consider an ϵ-cover of $\mathcal{V}_{\text{tabular}}$. For any $V_t \in \mathcal{V}_{\text{tabular}}$, there exists $\widetilde{V}_t$ in the ϵ-cover such that $\sup_s |V_t(s) - \widetilde{V}_t(s)| \le \epsilon$. Thus, we have

$$|[(\widehat{P}_t - P)V_t](s,a)| \le |[(\widehat{P}_t - P)\widetilde{V}_t](s,a)| + |[(\widehat{P}_t - P)(V_t - \widetilde{V}_t)](s,a)| \le |[(\widehat{P}_t - P)\widetilde{V}](s,a)| + 2\epsilon.$$

We then apply Lemma 20 and a union bound to obtain:

$$
\begin{aligned}
|[(\widehat{P}_t - P)V_t](s,a)| &\leq C \cdot \mathrm{sp}(v^*)\sqrt{\frac{\log(SAT\mathcal{N}_\epsilon/\delta)}{N_t(s,a)}} + 2\epsilon \\
&\leq C \cdot \mathrm{sp}(v^*)\sqrt{\frac{\log(SAT/\delta) + S\log(1/(\epsilon(1-\gamma)))}{N_t(s,a)}} + 2\epsilon.
\end{aligned}
$$

Picking $\epsilon = \frac{1}{\sqrt{T}}$ concludes the proof. □

A.1.2 Proof of Lemma 3

Proof of Lemma 3. We prove by induction on $t \geq 1$. The base case $t = 1$ is trivial since Algorithm 2 initializes $V_1(\cdot) = \frac{1}{1-\gamma}$, $Q_1(\cdot,\cdot) = \frac{1}{1-\gamma}$. Now, suppose $V_1, \ldots, V_t \geq V^*$ and $Q_1, \ldots, Q_t \geq Q^*$.

We first show that $Q_{t+1}(s,a) \geq Q^*(s,a)$. By the Bellman optimality equation for the discounted setting, we have for all $(s,a) \in \mathcal{S} \times \mathcal{A}$ that

$$
Q^*(s,a) = r(s,a) + \gamma[PV^*](s,a).
$$

Fix any pair $(s,a) \in \mathcal{S} \times \mathcal{A}$. By the definition of Q_{t+1} in Line 10 of Algorithm 2, we have

$$
\begin{aligned}
Q_{t+1}(s,a) &= (r(s,a) + \gamma[\widehat{P}_t V_t](s,a) + \beta/\sqrt{N_t(s,a)}) \wedge Q_t(s,a) \\
&\geq (r(s,a) + \gamma[PV_t](s,a)) \wedge Q_t(s,a) \\
&\geq (r(s,a) + \gamma[PV^*](s,a)) \wedge Q^*(s,a) \\
&= Q^*(s,a)
\end{aligned}
$$

where the first inequality is by the concentration inequality in Lemma 2 and our choice of β in Theorem 1, and the second inequality is by the induction hypotheses $V_t \geq V^*$ and $Q_t \geq Q^*$. The last equality is by the Bellman optimality equation.

Now, we show $V_{t+1}(s) \geq V^*(s)$. By the definition of $\widetilde{V}_{t+1}$ in Line 11 of Algorithm 2, we have

$$
\begin{aligned}
\widetilde{V}_{t+1}(s) &= (\max_a Q_{t+1}(s,a)) \wedge V_t(s) \\
&\geq (\max_a Q^*(s,a)) \wedge V^*(s) \\
&= V^*(s)
\end{aligned}
$$

where the inequality is by the optimism of Q_{t+1} we just proved, and the induction hypothesis $V_t \geq V^*$.

Finally, by the definition of V_{t+1} in Line 12 of Algorithm 2, we have

$$V_{t+1}(s) = \widetilde{V}_{t+1}(s) \wedge (\min_{s'} \widetilde{V}_{t+1}(s') + \mathrm{sp}(v^*)) \geq \widetilde{V}_{t+1}(s) \geq V^*(s),$$

which completes the proof by induction. □

A.1.3 Omitted Proofs

Lemma 21 (Sum of Bonus Terms). *Consider running Algorithm 2. The sum of the bonus terms can be bounded by*

$$\sum_{t=1}^{T} \sqrt{1/N_{t-1}(s_t, a_t)} \leq 2\sqrt{SAT}.$$

Proof. Recall that $N_t(s,a) = 1 + \sum_{s=1}^{t} \mathbb{I}\{s_t = s, a_t = a\}$. For convenience, write $n_t(s,a) = \sum_{s=1}^{t} \mathbb{I}\{s_t = s, a_t = a\}$ such that $N_t(s,a) = 1 + n_t(s,a)$. Then,

$$\begin{aligned} \sum_{t=1}^{T} \sqrt{1/N_{t-1}(s_t, a_t)} &\leq \sum_{s \in \mathcal{S}} \sum_{a \in \mathcal{A}} \sum_{n=1}^{n_T(s,a)} \sqrt{1/n} \\ &\leq 2 \sum_{s \in \mathcal{S}} \sum_{a \in \mathcal{A}} \sqrt{n_T(s,a)} \\ &\leq 2\sqrt{SAT} \end{aligned}$$

where the second inequality uses the identity $\sum_{n=1}^{N} 1/\sqrt{n} \leq 2\sqrt{N}$, and the last inequality is by Cauchy-Schwarz and the fact that $\sum_s \sum_a n_T(s,a) = T$. □

A.2 Linear MDP Setting

A.2.1 Concentration Bound for Regression Coefficients

Central to the analysis of the concentration bound for the approximate Bellman backup is the following concentration bound for vector-valued self-normalized processes.

Lemma 22 (Concentration of vector-valued self-normalized processes [Abbasi-Yadkori et al., 2011]). *Let $\{\varepsilon_t\}_{t=1}^{\infty}$ be a real-valued stochastic process with corresponding filtration $\{\mathcal{F}_t\}_{t=0}^{\infty}$. Let $\varepsilon_t | \mathcal{F}_{t-1}$ be zero-mean and σ-subgaussian. Let $\{\phi_t\}_{t=0}^{\infty}$ be an $\mathbb{R}^d$-valued stochastic process where $\phi_t \in \mathcal{F}_{t-1}$. Assume Λ_0 is a $d \times d$ positive definite matrix, and let $\Lambda_t = \Lambda_0 + \sum_{s=1}^{t} \phi_s \phi_s^T$.*

Then for any $\delta > 0$, with probability at least $1 - \delta$, we have for all $t \geq 0$ that

$$\left\| \sum_{s=1}^{t} \phi_s \varepsilon_s \right\|_{\Lambda_t^{-1}}^2 \leq 2\sigma^2 \log \left(\frac{det(\Lambda_t)^{1/2} det(\Lambda_0)^{-1/2}}{\delta} \right).$$

Lemma 23. *Let $V : \mathcal{S} \to [-B, B]$ be a bounded function. Then, $\boldsymbol{w}^*(V) = \int_{\mathcal{S}} V(s') d\boldsymbol{\mu}(s')$ which satisfies $[PV](s, a) = \langle \varphi(s, a), \boldsymbol{w}^*(V) \rangle$ for all $(s, a) \in \mathcal{S} \times \mathcal{A}$, satisfies*

$$\|\boldsymbol{w}^*(V)\|_2 \leq B\sqrt{d}.$$

Proof.

$$\|\boldsymbol{w}^*(V)\|_2 = \left\| \int_{\mathcal{S}} V(s') d\boldsymbol{\mu}(s') \right\|_2 \leq B \left\| \int_{\mathcal{S}} d\boldsymbol{\mu}(s') \right\|_2 \leq B\sqrt{d}$$

where the first inequality holds since $\boldsymbol{\mu}$ is a vector of positive measures and $V(s') \geq 0$. The last inequality is by the boundedness assumption (1.1) on $\boldsymbol{\mu}(\mathcal{S})$. □

Lemma 24. *Let $\boldsymbol{w}$ be a ridge regression coefficient obtained by regressing $y \in [0, B]$ on $\boldsymbol{x} \in \mathbb{R}^d$ using the dataset $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ so that $\boldsymbol{w} = \Lambda^{-1} \sum_{i=1}^n \boldsymbol{x}_i y_i$ where $\Lambda = \sum_{i=1}^n \boldsymbol{x}\boldsymbol{x}^T + \lambda I$. Then,*

$$\|\boldsymbol{w}\|_2 \leq B\sqrt{dn/\lambda}.$$

Proof. For any unit vector $\boldsymbol{u} \in \mathbb{R}^d$ with $\|\boldsymbol{u}\|_2 = 1$, we have

$$\begin{aligned}
|\boldsymbol{u}^T \boldsymbol{w}| &= \left| \boldsymbol{u}^T \Lambda^{-1} \sum_{i=1}^n \boldsymbol{x}_i y_i \right| \\
&\leq B \sum_{i=1}^n |\boldsymbol{u}^T \Lambda^{-1} \boldsymbol{x}_i| \\
&\leq B \sum_{i=1}^n \sqrt{\boldsymbol{u}^T \Lambda^{-1} \boldsymbol{u}} \sqrt{\boldsymbol{x}_i^T \Lambda^{-1} \boldsymbol{x}_i} \\
&\leq \frac{B}{\sqrt{\lambda}} \sum_{i=1}^n \sqrt{\boldsymbol{x}_i^T \Lambda^{-1} \boldsymbol{x}_i} \\
&\leq \frac{B}{\sqrt{\lambda}} \sqrt{n} \sqrt{\sum_{i=1}^n \boldsymbol{x}_i^T \Lambda^{-1} \boldsymbol{x}_i} \\
&\leq B\sqrt{dn/\lambda}
\end{aligned}$$

where the second inequality and the fourth inequality are by Cauchy-Schwartz, the third inequality is by $\Lambda \succeq \lambda I$, and the last inequality is by Lemma 28.

The desired result follows from the fact that $\|\boldsymbol{w}\|_2 = \max_{\boldsymbol{u}:\|\boldsymbol{u}\|_2=1} |\boldsymbol{u}^T\boldsymbol{w}|$. □

The following self-normalized process bound is an adaptation of Lemma D.4 in Jin et al. [2020]. Their proof defines the bound B to be a value that satisfies $\|V\|_\infty \leq B$. Upon observing their proof, it is easy to see that we can strengthen their result to require only $\mathrm{sp}(V) \leq B$. The following lemma is the strengthened version.

Lemma 25 (Adaptation of Lemma D.4 in Jin et al. [2020])**.** *Let $\{x_t\}_{t=1}^\infty$ be a stochastic process on state space $\mathcal{S}$ with corresponding filtration $\{\mathcal{F}_t\}_{t=0}^\infty$. Let $\{\phi_t\}_{t=0}^\infty$ be a $\mathbb{R}^d$-valued stochastic process where $\phi_t \in \mathcal{F}_{t-1}$, and $\|\phi_t\|_2 \leq 1$. Let $\Lambda_n = \lambda I + \sum_{t=1}^n \phi_t\phi_t^T$. Then for any $\delta > 0$ and any given function class $\mathcal{V}$, with probability at least $1-\delta$, for all $n \geq 0$, and any $V \in \mathcal{V}$ satisfying $sp(V) \leq H$, we have*

$$\left\|\sum_{t=1}^n \phi_t(V(x_t) - \mathbb{E}[V(x_t)|\mathcal{F}_{t-1}])\right\|_{\Lambda_n^{-1}}^2 \leq 4H^2\left[\frac{d}{2}\log\left(\frac{n+\lambda}{\lambda}\right) + \log\frac{\mathcal{N}_\varepsilon}{\delta}\right] + \frac{8n^2\varepsilon^2}{\lambda}$$

where $\mathcal{N}_\varepsilon$ is the ε-covering number of $\mathcal{V}$ with respect to the distance $dist(V,V') = \sup_x |V(x) - V'(x)|$.

Lemma 26 (Adaptation of Lemma D.6 in Jin et al. [2020])**.** *Let $\mathcal{V}_{linear}$ be a class of functions mapping from $\mathcal{S}$ to $\mathbb{R}$ with the following parametric form*

$$V(\cdot) = (\max_a \boldsymbol{w}^T\boldsymbol{\varphi}(\cdot,a) + v + \beta\sqrt{\boldsymbol{\varphi}(\cdot,a)^T\Lambda^{-1}\boldsymbol{\varphi}(\cdot,a)}) \wedge M \tag{A.1}$$

where the parameters $(\boldsymbol{w},\beta,v,\Lambda,M)$ satisfy $\|\boldsymbol{w}\| \leq L$, $\beta \in [0,B]$, $v \in [0,D]$, $M \geq 0$ and the minimum eigenvalue satisfies $\lambda_{min}(\Lambda) \geq \lambda$. Assume $\|\boldsymbol{\varphi}(s,a)\| \leq 1$ for all (s,a) pairs, and let $\mathcal{N}_\varepsilon$ be the ε-covering number of $\mathcal{V}$ with respect to the distance $dist(V,V') = \sup_x |V(x) - V'(x)|$. Then

$$\log\mathcal{N}_\varepsilon \leq d\log(1+8L/\varepsilon) + \log(1+4D/\varepsilon) + d^2\log[1+8d^{1/2}B^2/(\lambda\varepsilon^2)].$$

For the next lemma, we define value functions $V_u^{(t)}$ to be the functions obtained by the following value iteration (analogous to Line 7-12 in Algorithm 3):

With this definition, we show a high-probability bound on $\|\sum_{\tau=1}^t \boldsymbol{\varphi}(s_\tau,a_\tau)[V_u^{(t)}(s_{\tau+1}) - PV_u^{(t)}(s_\tau,a_\tau)]\|_{\Lambda_t^{-1}}$ uniformly on $u \in [T]$ and $t \in [T]$. Since the tuple $(t_k - 1, V_u^t, \Lambda_{t_k})$ encountered in Algorithm 3 is the same as the pair $(t, V_u^{(t)}, \Lambda_t)$ for some $t \in [T]$, the uniform bound implies bound on $\|\sum_{\tau=1}^{t_k-1} \boldsymbol{\varphi}(s_\tau,a_\tau)[V_u^k(s_{\tau+1}) - PV_u^k(s_\tau,a_\tau)]\|_{\Lambda_{t_k}^{-1}}$ for all episode k.

Lemma 27 (Adaptation of Lemma B.3 in Jin et al. [2020])**.** *Under the linear MDP setting in Theorem 2 for the γ-LSCVI-UCB algorithm with clipping oracle (Algorithm 3), let c_β be*

```
$V^{(t)}_{T+1}(\cdot) \leftarrow \frac{1}{1-\gamma}$.
for $u = T, T-1, \ldots, 1$ do
    $\boldsymbol{w}^{(t)}_{u+1} \leftarrow \Lambda_t^{-1} \sum_{\tau=1}^{t-1} \boldsymbol{\varphi}(s_\tau, a_\tau)(V^{(t)}_{u+1}(s_{\tau+1}) - \min_{s'} \widetilde{V}^{(t)}_{u+1}(s'))$.
    $\widetilde{Q}^{(t)}_u(\cdot,\cdot) \leftarrow \Big(r(\cdot,\cdot) + \gamma(\langle \boldsymbol{\varphi}(\cdot,\cdot), \boldsymbol{w}^{(t)}_{u+1}\rangle + \min_{s'} \widetilde{V}^{(t)}_{u+1}(s') + \beta \|\boldsymbol{\varphi}(\cdot,\cdot)\|_{\Lambda_t^{-1}})\Big) \wedge \frac{1}{1-\gamma}$.
    $\widetilde{V}^{(t)}_u(\cdot) \leftarrow \max_a \widetilde{Q}^{(t)}_u(\cdot, a)$.
    $V^{(t)}_u(\cdot) \leftarrow \widetilde{V}^{(t)}_u(\cdot) \wedge (\min_{s'} \widetilde{V}^{(t)}_u(s') + H)$.
```

the constant in the definition of $\beta = c_\beta H d \sqrt{\log(dT/\delta)}$. There exists an absolute constant C that is independent of c_β such that for any fixed $\delta \in (0,1)$, the event $\mathcal{E}$ defined by

$$\forall u \in [T],\ t \in [T]: \quad \left\| \sum_{\tau=1}^{t} \boldsymbol{\varphi}(s_\tau, a_\tau)[V^{(t)}_u(s_{\tau+1}) - [PV^{(t)}_u](s_\tau, a_\tau)] \right\|_{\Lambda_t^{-1}} \le C \cdot Hd\sqrt{\log((c_\beta + 1)dT/\delta)}$$

satisfies $P(\mathcal{E}) \ge 1 - \delta$.

Proof. For all $t = 1, \ldots, T$, by Lemma 24, we have $\|\boldsymbol{w}_t\|_2 \le H\sqrt{dt/\lambda}$. Hence, by combining Lemma 26 and Lemma 25, for any $\varepsilon > 0$ and any fixed pair $(u,t) \in [T] \times [T]$, we have with probability at least $1 - \delta/T^2$ that

$$\begin{aligned}
&\left\| \sum_{\tau=1}^{t} \boldsymbol{\varphi}(s_\tau, a_\tau)[V^{(t)}_u(s_{\tau+1}) - [PV^{(t)}_u](s_\tau, a_\tau)] \right\|^2_{\Lambda_t^{-1}} \\
&\qquad \le 4H^2 \left[\frac{2}{d} \log\left(\frac{t+\lambda}{\lambda}\right) + d\log\left(1 + \frac{4H\sqrt{dt}}{\varepsilon\sqrt{\lambda}}\right) + d^2 \log\left(1 + \frac{8d^{1/2}\beta^2}{\varepsilon^2\lambda}\right) \right. \\
&\qquad\qquad \left. + \log\left(\frac{T^2}{\delta}\right) \right] + \frac{8t^2\varepsilon^2}{\lambda}.
\end{aligned}$$

Using a union bound over $(u,t) \in [T] \times [T]$ and choosing $\varepsilon = Hd/t$ and $\lambda = 1$, there exists an absolute constant $C > 0$ independent of c_β such that, with probability at least $1 - \delta$,

$$\left\| \sum_{\tau=1}^{t} \varphi(s_\tau, a_\tau)[V^{(t)}_u(s_{\tau+1}) - [PV^{(t)}_u](s_\tau, a_\tau)] \right\|^2_{\Lambda_t^{-1}} \le C^2 \cdot d^2 H^2 \log((c_\beta + 1)dT/\delta),$$

which concludes the proof. □

Proof of Lemma 4. We prove under the event $\mathcal{E}$ defined in Lemma 27. For convenience, we introduce the notation $\bar{V}_u^k(s) = V_u^k(s) - \min_{s'} V_u^k(s')$. With this notation, we can write

$$\boldsymbol{w}_u^k = \Lambda_{t_k}^{-1} \sum_{\tau=1}^{t_k-1} \boldsymbol{\varphi}(s_\tau, a_\tau) \bar{V}_u^k(s_{\tau+1}).$$

We can decompose $\langle \boldsymbol{\phi}, \boldsymbol{w}_u^k \rangle$ as

$$\begin{aligned} \langle \boldsymbol{\phi}, \boldsymbol{w}_u^k \rangle = & \underbrace{\langle \boldsymbol{\phi}, \Lambda_{t_k}^{-1} \sum_{\tau=1}^{t_k-1} \boldsymbol{\varphi}(s_\tau, a_\tau)[P\bar{V}_u^k](s_\tau, a_\tau) \rangle}_{(a)} \\ & + \underbrace{\langle \boldsymbol{\phi}, \Lambda_{t_k}^{-1} \sum_{\tau=1}^{t_k-1} \boldsymbol{\varphi}(s_\tau, a_\tau)(\bar{V}_u^k(s_{\tau+1}) - P\bar{V}_u^k(s_\tau, a_\tau))}_{(b)}. \end{aligned}$$

Since $\boldsymbol{w}_u^{k*} = \int \bar{V}_u^k(s) d\boldsymbol{\mu}(s) = \boldsymbol{w}^*(\bar{V}_u^k)$ and $\bar{V}_u^k(s) \in [0, H]$ for all $s \in \mathcal{S}$, it follows by Lemma 23 that $\|\boldsymbol{w}_u^{k*}\|_2 \leq H\sqrt{d}$. Hence, the first term (a) in the display above can be bounded as

$$\begin{aligned} \langle \boldsymbol{\phi}, \Lambda_k^{-1} \sum_{\tau=1}^{t_k-1} \boldsymbol{\varphi}(s_\tau, a_\tau)[P\bar{V}_u^k](s_\tau, a_\tau) \rangle &= \langle \boldsymbol{\phi}, \Lambda_k^{-1} \sum_{\tau=1}^{t_k-1} \boldsymbol{\varphi}(s_\tau, a_\tau) \boldsymbol{\varphi}(s_\tau, a_\tau)^T \boldsymbol{w}_u^{k*} \rangle \\ &= \langle \boldsymbol{\phi}, \boldsymbol{w}_u^{k*} \rangle - \lambda \langle \boldsymbol{\phi}, \Lambda_k^{-1} \boldsymbol{w}_u^{k*} \rangle \\ &\leq \langle \boldsymbol{\phi}, \boldsymbol{w}_u^{k*} \rangle + \lambda \|\boldsymbol{\phi}\|_{\Lambda_k^{-1}} \|\boldsymbol{w}_u^{k*}\|_{\Lambda_k^{-1}} \\ &\leq \langle \boldsymbol{\phi}, \boldsymbol{w}_u^{k*} \rangle + H\sqrt{\lambda d} \|\boldsymbol{\phi}\|_{\Lambda_k^{-1}} \end{aligned}$$

where the first inequality is by Cauchy-Schwartz and the second inequality is by Lemma 23. Under the event $\mathcal{E}$ defined in Lemma 27, the second term (b) can be bounded by

$$\begin{aligned} & \langle \boldsymbol{\phi}, \Lambda_k^{-1} \sum_{\tau=1}^{t_k-1} \boldsymbol{\varphi}(s_\tau, a_\tau)(\bar{V}_u^k(s_{\tau+1}) - [P\bar{V}_u^k](s_\tau, a_\tau)) \\ & \qquad \leq \|\boldsymbol{\phi}\|_{\Lambda_k^{-1}} \left\| \sum_{\tau=1}^{t_k-1} \boldsymbol{\varphi}(s_\tau, a_\tau)(V_u^k(s_{\tau+1}) - [PV_u^k](s_\tau, a_\tau)) \right\|_{\Lambda_k^{-1}} \\ & \qquad \leq C \cdot Hd\sqrt{\log((c_\beta + 1)dT/\delta)} \cdot \|\boldsymbol{\phi}\|_{\Lambda_k^{-1}}. \end{aligned}$$

Combining the two bounds and rearranging, we get

$$\langle \boldsymbol{\phi}, \boldsymbol{w}_u^k - \boldsymbol{w}_u^{k*} \rangle \leq C \cdot Hd\sqrt{(\log(c_\beta + 1)dT/\delta)} \cdot \|\boldsymbol{\phi}\|_{\Lambda_k^{-1}}$$

for some absolute constant C independent of c_β. Lower bound of $\langle \boldsymbol{\phi}, \boldsymbol{w}_u^k - \boldsymbol{w}_u^{k*} \rangle$ can be shown similarly, establishing

$$|\langle \boldsymbol{\phi}, \boldsymbol{w}_u^k - \boldsymbol{w}_u^{k*} \rangle| \leq C \cdot Hd\sqrt{\log((c_\beta + 1)dT/\delta)} \cdot \|\boldsymbol{\phi}\|_{\Lambda_k^{-1}}.$$

It remains to show that there exists a choice of absolute constant c_β such that

$$C\sqrt{\log(c_\beta + 1) + \log(dT/\delta)} \leq c_\beta \sqrt{\log(dT/\delta)}.$$

Noting that $\log(dT/\delta) \geq \log 2$, this can be done by choosing an absolute constant c_β that satisfies $C\sqrt{\log 2 + \log(c_\beta + 1)} \leq c_\beta \sqrt{\log 2}$. □

A.2.2 Optimism

Proof of Lemma 5. We prove under the event $\mathcal{E}$ defined in Lemma 27. Fix any episode index $k > 1$. We prove by induction on $u = T+1, T, \ldots, 1$. The base case $u = T+1$ is trivial since $V_{T+1}^k(s) = \frac{1}{1-\gamma} \geq V^*(s)$ for all $s \in \mathcal{S}$ and $Q_{T+1}^k(s,a) = \frac{1}{1-\gamma} \geq Q^*(s,a)$ for all $(s,a) \in \mathcal{S} \times \mathcal{A}$.

Now, suppose the optimism results $V_{u+1}^k(s) \geq V^*(s)$ and $Q_{u+1}^k(s,a) \geq Q^*(s,a)$ for all $(s,a) \in \mathcal{S} \times \mathcal{A}$ hold for some $u \in [T]$. For convenience, we use the notation $\bar{V}_u^k(s) = V_u^k(s) - \min_{s'} V_u^k(s')$. Using the concentration bounds of regression coefficients $\boldsymbol{w}_u^k$ provided in Lemma 4, which holds under the event $\mathcal{E}$, we can lower bound $Q_u^k(s,a)$ as follows.

$$\begin{aligned}
Q_u^k(s,a) &= \Big(r(s,a) + \gamma(\langle \boldsymbol{\varphi}(s,a), \boldsymbol{w}_{u+1}^k \rangle + \min_{s'} V_{u+1}^k(s') + \beta \|\boldsymbol{\varphi}(s,a)\|_{\Lambda_k^{-1}}\Big) \wedge \frac{1}{1-\gamma} \\
&\geq \Big(r(s,a) + \gamma(\langle \boldsymbol{\varphi}(s,a), \boldsymbol{w}_{u+1}^k{}^* \rangle + \min_{s'} V_{u+1}^k(s'))\Big) \wedge \frac{1}{1-\gamma} \\
&= (r(s,a) + \gamma P V_{u+1}^k(s,a)) \wedge \frac{1}{1-\gamma} \\
&\geq (r(s,a) + \gamma P V^*(s,a)) \wedge \frac{1}{1-\gamma} \\
&= Q^*(s,a)
\end{aligned}$$

where $\boldsymbol{w}_{u+1}^k{}^*$ is a parameter that satisfies $\langle \boldsymbol{\varphi}(s,a), \boldsymbol{w}_{u+1}^k{}^* \rangle = [P\bar{V}_{u+1}^k](s,a)$. The second inequality is by the induction hypothesis $V_{u+1}^k \geq V^*$ and the last equality is by the Bellman optimality equation for the discounted setting and the fact that $Q^* \leq \frac{1}{1-\gamma}$.

We established $Q_u^k(s,a) \geq Q^*(s,a)$ for all $(s,a) \in \mathcal{S} \times \mathcal{A}$. It remains to show that $V_u^k(s) \geq V^*(s)$ for all $s \in \mathcal{S}$. Recall that the algorithm defines $\widetilde{V}_u^k(\cdot) = \max_a Q_u^k(\cdot, a)$.

Hence, for all $s \in \mathcal{S}$, we have

$$\begin{aligned}\widetilde{V}_u^k(s) - V^*(s) &= \max_a Q_u^k(s,a) - V^*(s) \\ &\geq Q_u^k(s, a_s^*) - Q^*(s, a_s^*) \\ &\geq 0\end{aligned}$$

where we use the notation $a_s^* = \operatorname{argmax}_a Q^*(s,a)$ so that $V^*(s) = Q^*(s, a_s^*)$, establishing $\widetilde{V}_u^k(s) \geq V^*(s)$ for all $s \in \mathcal{S}$. Hence, for all $s \in \mathcal{S}$, we have

$$\begin{aligned}V_u^k(s) &= \widetilde{V}_u^k(s) \wedge (\min_{s'} \widetilde{V}_u^k(s') + 2 \cdot \mathrm{sp}(v^*)) \\ &\geq V^*(s) \wedge (\min_{s'} V^*(s') + 2 \cdot \mathrm{sp}(v^*)) \\ &= V^*(s)\end{aligned}$$

where the last equality is due to $\mathrm{sp}(V^*) \leq 2 \cdot \mathrm{sp}(v^*)$ by Lemma 1. By induction, the proof for the optimism results $V_u^k(s) \geq V^*(s)$ and $Q_u^k(s,a) \geq Q^*(s,a)$ for $u = T+1, T, \ldots, 1$ is complete. □

A.2.3 Computational Complexity

The algorithm γ-LSCVI-UCB runs in episodes and the number of episodes is bounded by $\mathcal{O}(d \log T)$. In each episode, value iteration is run for at most T iterations. In each iteration u in episode k, one evaluation of $\min_{s'} \widetilde{V}_u^k(s')$, t_k evaluations of $V_u^k(\cdot)$ of $V_u^k(\cdot)$ and a multiplication of $d \times d$ matrix (Λ_k^{-1}) and a d-dimensional vector is required.

One evaluation of $\min_{s'} \widetilde{V}_u^k(s')$ involves S evaluations of $\widetilde{V}_u^k(\cdot)$. One evaluation of $\widetilde{V}_u^k(\cdot)$ involves A evaluations of $Q_u^k(\cdot, \cdot)$. One evaluation of $Q_u^k(\cdot, \cdot)$ requires $\mathcal{O}(d^2)$ operations. In total, one evaluation of $\min_{s'} \widetilde{V}_u^k(s')$ requires $\mathcal{O}(d^2 SA)$ operations.

Now, computing $\boldsymbol{w}_u^k$ requires evaluating $V_{u+1}^k(\cdot)$ for at most T states, which requires $\mathcal{O}(d^2 AT)$ operations; adding at most T d-dimensional vectors, which requires Td operations; and multiplying by $d \times d$ matrix, which requires d^2 operations.

In total, computing $\boldsymbol{w}_u^k$ requires $\mathcal{O}(d^2 A(S+T))$ operations. Hence, running at most T value iterations in each episode requires $\mathcal{O}(d^2 A(S+T)T)$ operations, and since there are at most $\mathcal{O}(d \log T)$ episodes, total operations for the algorithm is $\widetilde{\mathcal{O}}(d^3 A(S+T)T)$, which is polynomial in d, S, A, T.

A.2.4 Other Technical Lemmas

Lemma 28 (Lemma D.1 in Jin et al. [2020])**.** *Let $\Lambda_t = \sum_{i=1}^t \boldsymbol{\phi}_i \boldsymbol{\phi}_i^T + \lambda I$ where $\boldsymbol{\phi}_i \in \mathbb{R}^d$ and $\lambda > 0$. Then,*

$$\sum_{i=1}^{t} \boldsymbol{\phi}_i^T \Lambda_t^{-1} \boldsymbol{\phi}_i \leq d.$$

Lemma 29 (Lemma 11 in Abbasi-Yadkori et al. [2011])**.** *Let $\{\boldsymbol{\phi}_t\}_{t \geq 1}$ be a bounded sequence in $\mathbb{R}^d$ with $\|\boldsymbol{\phi}_t\|_2 \leq 1$ for all $t \geq 1$. Let $\Lambda_0 = I$ and $\Lambda_t = \sum_{i=1}^t \boldsymbol{\phi}_i \boldsymbol{\phi}_i^T + I$ for $t \geq 1$. Then,*

$$\sum_{i=1}^{t} \boldsymbol{\phi}_i^T \Lambda_{i-1}^{-1} \boldsymbol{\phi}_i \leq 2 \log \det(\Lambda_t) \leq 2d \log(1+t).$$

Lemma 30 (Lemma 12 in Abbasi-Yadkori et al. [2011])**.** *Suppose $A, B \in \mathbb{R}^{d \times d}$ are two positive definite matrices satisfying $A \succeq B$. Then, for any $\boldsymbol{x} \in \mathbb{R}^d$, we have*

$$\|\boldsymbol{x}\|_A \leq \|\boldsymbol{x}\|_B \sqrt{\frac{\det(A)}{\det(B)}}.$$

Lemma 31 (Bound on number of episodes)**.** *The number of episodes K in Algorithm 3 is bounded by*

$$K \leq d \log_2 \left(1 + \frac{T}{\lambda d}\right).$$

Proof. Let $\{\Lambda_k\}_{k=1}^K$ and $\{\bar{\Lambda}_t\}_{t=0}^T$ be as defined in Algorithm 3. Note that

$$\operatorname{tr}(\bar{\Lambda}_T) = \operatorname{tr}(\lambda I_d) + \sum_{t=1}^{T} \operatorname{tr}(\varphi(s_t, a_t)\varphi(s_t, a_t)^T) = \lambda d + \sum_{t=1}^{T} \|\varphi(s_t, a_t)\|_2^2 \leq \lambda d + T.$$

By the AM–GM inequality, we have

$$\det(\bar{\Lambda}_T) \leq \left(\frac{\operatorname{tr}(\bar{\Lambda}_T)}{d}\right)^d \leq \left(\frac{\lambda d + T}{d}\right)^d.$$

Since we update Λ_k only when $\det(\bar{\Lambda}_t)$ doubles, $\det(\bar{\Lambda}_T) \geq \det(\Lambda_K) \geq \det(\Lambda_1) \cdot 2^K = \lambda^d \cdot 2^K$. Thus, we obtain

$$K \leq d \log_2 \left(1 + \frac{T}{\lambda d}\right)$$

as desired.

□

A.3 Linear MDP Setting: Computational Efficiency

A.3.1 Concentration Inequalities

Lemma 32 (Adaptation of Lemma B.3 in Jin et al. [2020])**.** *Under the linear MDP setting in Theorem 2 for the γ-LSCVI-UCB algorithm with clipping oracle (Algorithm 3), let c_β be the constant in the definition of $\beta = c_\beta H d\sqrt{\log(dT/\delta)}$. There exists an absolute constant C that is independent of c_β such that for any fixed $\delta \in (0,1)$, the event $\mathcal{E}$ defined by*

$$\forall u \in [T],\, t \in [T]: \\ \left\|\sum_{\tau=1}^{t-1} \boldsymbol{\varphi}(s_\tau, a_\tau)[V_u^t(s_{\tau+1}) - [PV_u^t](s_\tau, a_\tau)]\right\|_{\Lambda_t^{-1}} \leq C \cdot Hd\sqrt{\log((c_\beta + 1)dT/\delta)}$$

satisfies $P(\mathcal{E}) \geq 1 - \delta$.

Proof. By Lemma 24, we have $\|\boldsymbol{w}_t\|_2 \leq H\sqrt{dt/\lambda}$ for all $t = 1, \ldots, T$. Hence, by combining Lemma 38 and Lemma 25, for any $\varepsilon > 0$ and any fixed pair $(u,t) \in [T] \times [T]$, we have with probability at least $1 - \delta/T^2$ that

$$\begin{aligned}&\left\|\sum_{\tau=1}^{t-1} \boldsymbol{\varphi}(s_\tau, a_\tau)[V_u^t(s_{\tau+1}) - [PV_u^t](s_\tau, a_\tau)]\right\|_{\Lambda_t^{-1}}^2 \\ &\qquad \leq 4H^2\left[\frac{2}{d}\log\left(\frac{t+\lambda}{\lambda}\right) + d\log\left(1 + \frac{4H\sqrt{dt}}{\varepsilon\sqrt{\lambda}}\right) + d^2\log\left(1 + \frac{8d^{1/2}\beta^2}{\varepsilon^2\lambda}\right) + \log\left(\frac{T^2}{\delta}\right)\right] \\ &\qquad\quad + \frac{8t^2\varepsilon^2}{\lambda}.\end{aligned}$$

Using a union bound over $(u,t) \in [T] \times [T]$ and choosing $\varepsilon = Hd/t$ and $\lambda = 1$, there exists an absolute constant $C > 0$ independent of c_β such that, with probability at least $1 - \delta$,

$$\left\|\sum_{\tau=1}^{t-1} \varphi(s_\tau, a_\tau)[V_u^t(s_{\tau+1}) - [PV_u^t](s_\tau, a_\tau)]\right\|_{\Lambda_t^{-1}}^2 \leq C^2 \cdot d^2H^2\log((c_\beta + 1)dT/\delta),$$

which concludes the proof. □

A.3.2 Proof of Lemma 4

Proof of Lemma 4. We prove under the event $\mathcal{E}$ defined in Lemma 32. Recall the definition

$$[\widehat{P}_t V_u^t](s,a) = \langle \boldsymbol{\varphi}(s,a), \widehat{\boldsymbol{w}}_t(V_u^t - V_u^t(s_1))\rangle + V_u^t(s_1)$$

where $\widehat{\boldsymbol{w}}_t(V_u^t - V_u^t(s_1)) = \Lambda_t^{-1} \sum_{\tau=1}^{t-1}(V_u^t(s_{\tau+1}) - V_u^t(s_1)) \cdot \boldsymbol{\varphi}(s_\tau, a_\tau)$. For convenience, we introduce the notation $\bar{V}_u^k(s) = V_u^k(s) - V_u^k(s_1)$ and $\boldsymbol{w}_u^t = \widehat{\boldsymbol{w}}_t(\bar{V}_u^t)$. With these notations, we have

$$[\widehat{P}_t V_u^t](s,a) = \langle \boldsymbol{\varphi}(s,a), \boldsymbol{w}_u^t \rangle + V_u^t(s_1), \quad \boldsymbol{w}_u^t = \Lambda_t^{-1} \sum_{\tau=1}^{t-1} \boldsymbol{\varphi}(s_\tau, a_\tau) \bar{V}_u^k(s_{\tau+1}).$$

We can decompose $\langle \boldsymbol{\varphi}(s,a), \boldsymbol{w}_u^t \rangle$ as

$$\langle \boldsymbol{\varphi}(s,a), \boldsymbol{w}_u^t \rangle = \underbrace{\langle \boldsymbol{\varphi}(s,a), \Lambda_t^{-1} \sum_{\tau=1}^{t-1} \boldsymbol{\varphi}(s_\tau, a_\tau) [P\bar{V}_u^t](s_\tau, a_\tau) \rangle}_{(a)}$$
$$+ \underbrace{\langle \boldsymbol{\varphi}(s,a), \Lambda_t^{-1} \sum_{\tau=1}^{t-1} \boldsymbol{\varphi}(s_\tau, a_\tau) (\bar{V}_u^t(s_{\tau+1}) - [P\bar{V}_u^t](s_\tau, a_\tau)) }_{(b)}.$$

Since $\bar{V}_u^t(s) \in [-H, H]$ for all $s \in \mathcal{S}$, it follows by Lemma 23 that $\|\boldsymbol{w}^*(\bar{V}_u^t)\|_2 \le H\sqrt{d}$. Hence, the first term (a) in the display above can be bounded as

$$\begin{aligned}
\langle \boldsymbol{\varphi}(s,a), \Lambda_t^{-1} \sum_{\tau=1}^{t-1} \boldsymbol{\varphi}(s_\tau, a_\tau) [P\bar{V}_u^t](s_\tau, a_\tau) \rangle &= \langle \boldsymbol{\varphi}(s,a), \Lambda_t^{-1} \sum_{\tau=1}^{t-1} \boldsymbol{\varphi}(s_\tau, a_\tau) \boldsymbol{\varphi}(s_\tau, a_\tau)^T \boldsymbol{w}^*(\bar{V}_u^t) \rangle \\
&= \langle \boldsymbol{\varphi}(s,a), \boldsymbol{w}^*(\bar{V}_u^t) \rangle - \lambda \langle \boldsymbol{\varphi}(s,a), \Lambda_t^{-1} \boldsymbol{w}^*(\bar{V}_u^t) \rangle \\
&\le \langle \boldsymbol{\varphi}(s,a), \boldsymbol{w}^*(\bar{V}_u^t) \rangle + \lambda \|\boldsymbol{\varphi}(s,a)\|_{\Lambda_t^{-1}} \|\boldsymbol{w}^*(\bar{V}_u^t)\|_{\Lambda_t^{-1}} \\
&\le \langle \boldsymbol{\varphi}(s,a), \boldsymbol{w}^*(\bar{V}_u^t) \rangle + H\sqrt{\lambda d} \|\boldsymbol{\varphi}(s,a)\|_{\Lambda_t^{-1}}
\end{aligned}$$

where the first inequality is by Cauchy-Schwartz and the second inequality is by Lemma 23. Under the event $\mathcal{E}$ defined in Lemma 32, the second term (b) can be bounded by

$$\begin{aligned}
&\langle \boldsymbol{\varphi}(s,a), \Lambda_t^{-1} \sum_{\tau=1}^{t-1} \boldsymbol{\varphi}(s_\tau, a_\tau) (\bar{V}_u^t(s_{\tau+1}) - [P\bar{V}_u^t](s_\tau, a_\tau)) \\
&\qquad \le \|\boldsymbol{\varphi}(s,a)\|_{\Lambda_t^{-1}} \left\| \sum_{\tau=1}^{t-1} \boldsymbol{\varphi}(s_\tau, a_\tau) (V_u^t(s_{\tau+1}) - [PV_u^t](s_\tau, a_\tau)) \right\|_{\Lambda_t^{-1}} \\
&\qquad \le C \cdot Hd\sqrt{\log((c_\beta + 1)dT/\delta)} \cdot \|\boldsymbol{\varphi}(s,a)\|_{\Lambda_t^{-1}}.
\end{aligned}$$

Combining the two bounds and rearranging, we get

$$\langle \boldsymbol{\phi}, \boldsymbol{w}_u^t - \boldsymbol{w}^*(\bar{V}_u^t) \rangle \le C \cdot Hd\sqrt{(\log(c_\beta + 1)dT/\delta)} \cdot \|\boldsymbol{\phi}\|_{\Lambda_t^{-1}}$$

for some absolute constant C independent of c_β. Lower bound of $\langle \boldsymbol{\phi}, \boldsymbol{w}_u^t - \boldsymbol{w}^*(\bar{V}_u^t)\rangle$ can be shown similarly, establishing

$$|\langle \boldsymbol{\phi}, \boldsymbol{w}_u^t - \boldsymbol{w}^*(\bar{V}_u^t)\rangle| \leq C \cdot Hd\sqrt{\log((c_\beta + 1)dT/\delta)} \cdot \|\boldsymbol{\phi}\|_{\Lambda_t^{-1}}.$$

Hence,

$$\begin{aligned}|[\widehat{P}_t V_u^t](s,a) - [PV_u^t](s,a)| &= |\langle \boldsymbol{\varphi}(s,a), \widehat{\boldsymbol{w}}_t(V_u^t - V_u^t(s_1))\rangle + V_u^t(s_1) - \langle \boldsymbol{\varphi}(s,a), \boldsymbol{w}^*(V_u^t)\rangle \\ &= |\langle \boldsymbol{\varphi}(s,a), \boldsymbol{w}_u^t - \boldsymbol{w}^*(\bar{V}_u^t)\rangle| \\ &\leq C \cdot Hd\sqrt{\log((c_\beta + 1)dT/\delta)} \cdot \|\boldsymbol{\phi}\|_{\Lambda_t^{-1}}\end{aligned}$$

where the last equality uses the fact that $\boldsymbol{w}^*(V) = \int_{\mathcal{S}} V(s')\boldsymbol{\mu}(s')$ is linear. It remains to show that there exists a choice of absolute constant c_β such that

$$C\sqrt{\log(c_\beta + 1) + \log(dT/\delta)} \leq c_\beta\sqrt{\log(dT/\delta)}.$$

Noting that $\log(dT/\delta) \geq \log 2$, this can be done by choosing an absolute constant c_β that satisfies $C\sqrt{\log 2 + \log(c_\beta + 1)} \leq c_\beta\sqrt{\log 2}$. □

Lemma 33. *The clipping operation* $\textsc{Clip}(x; L, U)$ *has the following properties:*

(i) $\textsc{Clip}(x; L, U) = \textsc{Clip}(x - c; L - c, U - c) + c$.

(ii) $\textsc{Clip}(x; L, U) \leq \textsc{Clip}(y; L, U)$ *if* $x \leq y$.

(iii) $\textsc{Clip}(x; L, U) \leq x$ *if and only if* $x \geq L$.

(iv) $\textsc{Clip}(x; L, U) \geq \textsc{Clip}(x; L', U')$ *if* $L \geq L'$ *and* $U \geq U'$.

Proof. The proofs are straight from the definition. □

A.3.3 Deviation-Controlled Value Iteration

A.3.3.1 Positive Result for Tabular MDPs

In this section, we show that the scheme used in the algorithm γ-LSCVI-UCB+ for controlling the deviation between chains of value functions with different clipping thresholds is not necessary in the tabular setting.

To reuse the notations developed for the linear setting, we treat the tabular setting with the size of the state space S and the size of the action space A as the SA-dimensional linear MDP setting where each pair $(s,a) \in \mathcal{S} \times \mathcal{A}$ is mapped to a one-hot encoded vector $\boldsymbol{\varphi}(s,a) = \boldsymbol{e}_{(s,a)} \in \mathbb{R}^{SA}$ where the entry associated to (s,a) is equal to 1 and all other entries 0. We show that under the tabular setting, Algorithm 9 that removes the step for clipping Q_u^t from γ-LSCVI-UCB+ successfully control the deviation of a chain of value functions from its previous chain. Note that the algorithm uses the doubling-trick that updates the covariance matrix used for regression only when its determinant doubles. The trick is used to facilitate the analysis of the difference $Q_u^t(s,a) - Q_u^{t+1}(s,a)$ shown in the proof of the lemma below.

We use $\lambda = 0$ and treat Λ_t^{-1} as the pseudoinverse of Λ, and set $\|\boldsymbol{\varphi}(s,a)\|_{\Lambda_t^{-1}} = \frac{1}{1-\gamma}$ when $\|\boldsymbol{\varphi}(s,a)\|_{\Lambda_t^{-1}} = 0$, that is, when the direction $\boldsymbol{\varphi}(s,a)$ is never explored. Then, as shown in the following lemma, the deviation between chains of value iterations is controlled even without the extra scheme used for the linear MDP setting.

Lemma 34. *When running γ-LSCVI-UCB+ algorithm without deviation control under the tabular setting, for all $t \in [T]$, $u \in [t:T]$, we have*

$$|\widetilde{V}_u^{t+1}(s) - \widetilde{V}_u^t(s)| \le m_t - m_{t+1}$$
$$|V_u^{t+1}(s) - V_u^t(s)| \le m_t - m_{t+1}$$

for all $s \in \mathcal{S}$.

Proof. We introduce the notation $N_t(s,a) = \sum_{\tau=1}^{t-1} \mathbb{I}\{s_\tau = s, a_\tau = a\}$ and $N_t(s,a,s') = \sum_{\tau=1}^{t-1} \mathbb{I}\{s_\tau = s, a_\tau = a, s_{\tau+1} = s'\}$, which is the visitation counts up to (excluding) time step t of the state-action pair (s,a) and state-action-state triplet (s,a,s'), respectively. Note that in the tabular setting, we have $[\widehat{P}_t V](s,a) = \sum_{s':N_t(s,a,s')>0} (N_t(s,a,s')/N_t(s,a)) V(s')$, which is the expectation of V with respect to the empirical transition probability kernel $\widehat{P}_t$: $\widehat{P}_t(s'|s,a) = N_t(s,a,s')/N_t(s,a)$. Hence, $\widehat{P}_t$ is linear such that $[\widehat{P}_t V_1](s,a) - [\widehat{P}_t V_2](s,a) = [\widehat{P}_t(V_1 - V_2)](s,a)$, and it satisfies $[\widehat{P}_t \Delta](s,a) \le \|\Delta\|_\infty$ for any function $\Delta : \mathcal{S} \to \mathbb{R}$. We exploit these facts to prove the lemma.

We show by induction on $u = T+1, \ldots, 1$. Fix t such that both t and $t+1$ are in the same episode k. For the base case $u = T+1$, we have $V_{T+1}^{t+1}(s) = V_{T+1}^t(s) = \frac{1}{1-\gamma}$ for all $s \in \mathcal{S}$, and trivially, we have $|V_{T+1}^{t+1}(s) - V_{T+1}^t(s)| \le m_t - m_{t+1}$. Now, suppose $|V_{u+1}^{t+1}(s) - V_{u+1}^t(s)| \le m_t - m_{t+1}$ for all $s \in \mathcal{S}$ for some $u \in [T]$. Then,

$$|Q_u^t(s,a) - Q_u^{t+1}(s,a)| \le \gamma([\widehat{P}_{t_k} V_{u+1}^t](s,a) - [\widehat{P}_{t_k} V_{u+1}^{t+1}](s,a)) \le m_t - m_{t+1}$$

where the first inequality is by the fact that $(\cdot \wedge \frac{1}{1-\gamma})$ is a contraction and the second inequality is by the previous discussion on $\widehat{P}_{t_k}$ being a expectation with respect to a proper probability kernel in the tabular setting. Since $\max_a Q(\cdot, a)$ is a contraction, it follows that

$|\widetilde{V}_u^t(s) - \widetilde{V}_u^{t+1}(s)| \leq m_t - m_{t+1}$. Hence, using the fact that $m_t \geq m_{t+1}$, we have

$$\begin{aligned}
&V_u^t(s) - V_u^{t+1}(s) \\
&\quad = \textsc{Clip}(\widetilde{V}_u^t(s); m_t, m_t + H) - \textsc{Clip}(\widetilde{V}_u^{t+1}(s); m_{t+1}, m_{t+1} + H) \\
&\quad \leq \textsc{Clip}(\widetilde{V}_u^{t+1}(s) + m_t - m_{t+1}; m_t, m_t + H) - \textsc{Clip}(\widetilde{V}_u^{t+1}(s); m_{t+1}, m_{t+1} + H) \\
&\quad = \textsc{Clip}(\widetilde{V}_u^{t+1}(s); m_{t+1}, m_{t+1} + H) + m_t - m_{t+1} - \textsc{Clip}(\widetilde{V}_u^{t+1}(s); m_{t+1}, m_{t+1} + H) \\
&\quad = m_t - m_{t+1}
\end{aligned}$$

where the second equality uses the property (i) of the clipping operation. Similarly, we have

$$\begin{aligned}
V_u^t(s) - V_u^{t+1}(s) &= \textsc{Clip}(\widetilde{V}_u^t(s); m_t, m_t + H) - \textsc{Clip}(\widetilde{V}_u^{t+1}(s); m_{t+1}, m_{t+1} + H) \\
&\geq \textsc{Clip}(\widetilde{V}_u^t(s); m_t, m_t + H) - \textsc{Clip}(\widetilde{V}_u^{t+1}(s); m_t, m_t + H) \\
&\geq \textsc{Clip}(\widetilde{V}_u^t(s); m_t, m_t + H) - \textsc{Clip}(\widetilde{V}_u^t(s) - m_t + m_{t+1}; m_t, m_t + H) \\
&= \textsc{Clip}(\widetilde{V}_u^t(s); m_t, m_t + H) - \textsc{Clip}(\widetilde{V}_u^t(s); m_{t+1}, m_{t+1} + H) - m_t + m_{t+1} \\
&\geq -m_t + m_{t+1}
\end{aligned}$$

where the second equality uses the property (i) of the clipping operation. The two inequalities establish $|V_u^t(s) - V_u^{t+1}(s)| \leq m_t - m_{t+1}$ as desired. By induction, the proof is complete. □

A.3.3.2 Negative Result for Linear MDPs

Proof of Lemma 6. For convenience, let $n = 2m$. If n is odd, we can take $\boldsymbol{\phi}_n = \mathbf{0}$ and similar argument holds. Take $\boldsymbol{\phi}_1, \cdots \boldsymbol{\phi}_m = (\eta, 1/2, 0, \cdots, 0)$ and $\boldsymbol{\phi}_{m+1}, \cdots, \boldsymbol{\phi}_{2m} = (\eta, -1/2, 0, \cdots, 0)$ where $\eta > 0$ is to be chosen later. Take $y_1 = \cdots = y_{2m} = \Delta$ and $\lambda = 1$. Then, $\Lambda_n = \operatorname{diag}(\eta^2 n, n/4, 0, \ldots, 0) + I$ and $\sum_{i=1}^n y_i \boldsymbol{\phi}_i = (\eta \Delta n, 0, \ldots, 0)$. Hence, $\boldsymbol{w}_n = (\frac{\eta \Delta n}{\eta^2 n + 1}, 0, \ldots, 0)$. It follows that, choosing $\boldsymbol{\phi} = (1, 0, \ldots, 0)$, we get

$$|\langle \boldsymbol{w}_n, \boldsymbol{\phi} \rangle| = \frac{\eta \Delta n}{\eta^2 n + 1}.$$

Choosing $\eta = 1/\sqrt{n}$, we get $|\langle \boldsymbol{w}_n, \boldsymbol{\phi} \rangle| = \frac{1}{2}\Delta\sqrt{n}$, which completes the proof. □

Algorithm 9: γ-LSCVI-UCB+ without Deviation Control

```
Input: Discounting factor $\gamma \in [0, 1)$, regularization constant $\lambda > 0$, span $H > 0$,
       bonus factor $\beta > 0$.
Initialize: $k \leftarrow 1$, $t_k \leftarrow 1$, $\Lambda_1 \leftarrow \lambda I$, $m_1 \leftarrow \frac{1}{1-\gamma}$.
Receive state $s_1$.
for $t = 1, \dots, T$ do
    $V^t_{T+1}(\cdot) \leftarrow \frac{1}{1-\gamma}$.
    for $u = T, T-1, \dots, t$ do
        $Q^t_u(\cdot,\cdot) \leftarrow \Big( r(\cdot,\cdot) + \gamma([\widehat{P}_{t_k} V^t_{u+1}](\cdot,\cdot) + \beta \|\boldsymbol{\varphi}(\cdot,\cdot)\|_{\Lambda^{-1}_{t_k}}) \Big) \wedge \frac{1}{1-\gamma}$.
        $\widetilde{V}^t_u(\cdot) \leftarrow \max_a Q^t_u(\cdot, a)$.
        $V^t_u(\cdot) \leftarrow \textsc{Clip}(\widetilde{V}^t_u(\cdot); m_t, m_t + H)$.
    Take action $a_t \leftarrow \operatorname{argmax}_{a \in \mathcal{A}} Q^t_t(s_t, a)$. Receive reward $r(s_t, a_t)$. Receive next
     state $s_{t+1}$.
    $\Lambda_{t+1} \leftarrow \Lambda_t + \boldsymbol{\varphi}(s_t, a_t)\boldsymbol{\varphi}(s_t, a_t)^\top$.
    $m_{t+1} \leftarrow \widetilde{V}^t_{t+1}(s_{t+1}) \wedge m_t$.
    if $2\det(\Lambda_{t_k}) < \det(\Lambda_{t+1})$ then
        $k \leftarrow k + 1$, $t_k \leftarrow t + 1$.
```

A.3.3.3 Deviation-Controlled Value Iteration for Linear MDPs

Lemma 35. *For all $t \in [T]$, $u \in [t : T]$, we have*

$$|\widetilde{V}^{t+1}_u(s) - \widetilde{V}^t_u(s)| \le m_{t-1} - m_{t+1}$$
$$|V^{t+1}_u(s) - V^t_u(s)| \le m_{t-1} - m_{t+1}$$

for all $s \in \mathcal{S}$.

Proof. We first show that $\widetilde{V}^{t+1}_u(s) - \widetilde{V}^t_u(s) \ge -m_{t-1} + m_{t+1}$ and $V^{t+1}_u(s) - V^t_u(s) \ge -m_{t-1} + m_{t+1}$. By definitions of Q^{t+1}_u and Q^t_u, we have

$$\begin{aligned}
Q^{t+1}_u(s,a) &= \textsc{Clip}(\widetilde{Q}^{t+1}_u(s,a); L^{t+1}_u(s,a), U^{t+1}_u(s,a)) \\
&\ge L^{t+1}_u(s,a) \\
&= (\widetilde{Q}^t_u(s,a) - m_t + m_{t+1}) \vee (\widetilde{Q}^{t-1}_u(s,a) - m_{t-1} + m_{t+1}) \\
&\ge \widetilde{Q}^{t-1}_u(s,a) - m_{t-1} + m_{t+1},
\end{aligned}$$

and

$$\begin{aligned}Q_u^t(s,a) &= \textsc{Clip}(\widetilde{Q}_u^t(s,a); L_u^t(s,a), U_u^t(s,a)) \\ &\le U_u^t(s,a) \\ &= \widetilde{Q}_u^{t-1}(s,a) \wedge \widetilde{Q}_u^{t-2}(s,a) \\ &\le \widetilde{Q}_u^{t-1}(s,a).\end{aligned}$$

Chaining the two inequalities, we get $Q_u^{t+1}(s,a) \ge Q_u^t(s,a) - m_{t-1} + m_{t+1}$. It follows that

$$\begin{aligned}\widetilde{V}_u^{t+1}(s) &= \max_a Q_u^{t+1}(s,a) \\ &\ge \max_a Q_u^t(s,a) - m_{t-1} + m_{t+1} \\ &= \widetilde{V}_u^t(s) - m_{t-1} + m_{t+1},\end{aligned}$$

which shows the first claim. Hence,

$$\begin{aligned}V_u^{t+1}(s) &= \textsc{Clip}(\widetilde{V}_u^{t+1}(s); m_{t+1}, m_{t+1} + H) \\ &\ge \textsc{Clip}(\widetilde{V}_u^t(s) - m_{t-1} + m_{t+1}; m_{t+1}, m_{t+1} + H) \\ &= \textsc{Clip}(\widetilde{V}_u^t(s); m_{t-1}, m_{t-1} + H) - m_{t-1} + m_{t+1} \\ &\ge \textsc{Clip}(\widetilde{V}_u^t(s); m_t, m_t + H) - m_{t-1} + m_{t+1} \\ &= V_u^t(s) - m_{t-1} + m_{t+1},\end{aligned}$$

where the second equality is by Property (i) of the clipping operation and the second inequality is by Property (iv) of the clipping operation and the fact that $m_{t-1} \ge m_t$. This shows the second claim.

Now, we show that $\widetilde{V}_u^{t+1}(s) - \widetilde{V}_u^t(s) \le m_{t-1} - m_{t+1}$ and $V_u^{t+1}(s) - V_u^t(s) \le m_{t-1} - m_{t+1}$. By definitions of Q_u^{t+1} and Q_u^t, we have

$$\begin{aligned}Q_u^{t+1}(s,a) &= \textsc{Clip}(\widetilde{Q}_u^{t+1}(s,a); L_u^{t+1}(s,a), U_u^{t+1}(s,a)) \\ &\le U_u^{t+1}(s,a) \\ &= \widetilde{Q}_u^t(s,a) \wedge \widetilde{Q}_u^{t-1}(s,a) \\ &\le \widetilde{Q}_u^{t-1}(s,a),\end{aligned}$$

and

$$\begin{aligned}
Q_u^t(s,a) &= \textsc{Clip}(\widetilde{Q}_u^t(s,a); L_u^t(s,a), U_u^t(s,a)) \\
&\geq L_u^t(s,a) \\
&= (\widetilde{Q}_u^{t-1}(s,a) - m_t + m_{t+1}) \vee (\widetilde{Q}_u^{t-2}(s,a) - m_{t-1} + m_{t+1}) \\
&\geq \widetilde{Q}_u^{t-1}(s,a) - m_t + m_{t+1} \\
&\geq \widetilde{Q}_u^{t-1}(s,a) - m_{t-1} + m_{t+1}.
\end{aligned}$$

Chaining the two inequalities, we get $Q_u^{t+1}(s,a) \leq Q_u^t(s,a) + m_{t-1} - m_{t+1}$, and it follows that

$$\begin{aligned}
\widetilde{V}_u^{t+1} &= \max_a Q_u^{t+1}(s,a) \\
&\leq \max_a Q_u^t(s,a) + m_{t-1} - m_{t+1} \\
&= \widetilde{V}_u^t(s) + m_{t-1} - m_{t+1},
\end{aligned}$$

which shows the first claim. Hence,

$$\begin{aligned}
V_u^{t+1}(s) &= \textsc{Clip}(\widetilde{V}_u^{t+1}(s); m_{t+1}, m_{t+1} + H) \\
&\leq \textsc{Clip}(\widetilde{V}_u^t(s) + m_t - m_{t+1}; m_{t+1}, m_{t+1} + H) \\
&\leq \textsc{Clip}(\widetilde{V}_u^t(s) + m_t - m_{t+1}; m_t, m_t + H) \\
&= \textsc{Clip}(\widetilde{V}_u^t(s); m_{t+1}, m_{t+1} + H) + m_t - m_{t+1} \\
&\leq \textsc{Clip}(\widetilde{V}_u^t(s); m_t, m_t + H) + m_t - m_{t+1} \\
&= V_u^t(s) + m_t - m_{t+1} \\
&\leq V_u^t(s) + m_{t-1} - m_{t+1}.
\end{aligned}$$

□

A.3.4 Regret Analysis

We first prove the optimism result that says the value function estimates are optimistic estimates of the true value function.

A.3.4.1 Proof of Lemma 9

Proof of Lemma 9. We prove under the event $\mathcal{E}$ defined in Lemma 32, which holds with probability at least $1 - \delta$. We prove by induction on t and u.

Suppose $V_u^\tau(s) \geq V^*(s)$, $\widetilde{V}_u^\tau(s) \geq V^*(s)$ and $\widetilde{Q}_u^\tau(s,a) \geq Q^*(s,a)$ hold for all $\tau = 1, \ldots, t-1$ and $u \in [\tau : T]$ and $(s,a) \in \mathcal{S} \times \mathcal{A}$. If we show that $V_u^t(s) \geq V^*(s)$, $\widetilde{V}_u^t(s)$ and $\widetilde{Q}_u^t(s,a) \geq Q^*(s,a)$ for all $u \in [t : T]$ and $(s,a) \in \mathcal{S} \times \mathcal{A}$, the proof is complete by induction on t. We show this by induction on $u = T+1, T, \ldots, t$.

The base case $u = T+1$ holds since $V_{T+1}^t(s) = \frac{1}{1-\gamma} \geq V^*(s)$ for all $s \in \mathcal{S}$. Now, suppose $V_{u+1}^t(s) \geq V^*(s)$ for all $s \in \mathcal{S}$ for some $u \in [t+1 : T]$. Then,

$$\begin{aligned}
\widetilde{Q}_u^t(s,a) &= (r(s,a) + \gamma([\widehat{P}_t V_{u+1}^t](s,a) + \beta \|\boldsymbol{\varphi}(s,a)\|_{\Lambda_t^{-1}}) \wedge \frac{1}{1-\gamma} \\
&\geq (r(s,a) + \gamma[P V_{u+1}^t](s,a)) \wedge \frac{1}{1-\gamma} \\
&\geq (r(s,a) + \gamma[P V^*](s,a)) \wedge \frac{1}{1-\gamma} \\
&= Q^*(s,a) \wedge \frac{1}{1-\gamma} \\
&= Q^*(s,a)
\end{aligned}$$

where the first inequality is by the event $\mathcal{E}$, the second inequality by the induction hypothesis. The second equality is by the Bellman optimality equation. This shows $\widetilde{Q}_u^t(s,a) \geq Q^*(s,a)$ for all $(s,a) \in \mathcal{S} \times \mathcal{A}$ as desired. Additionally,

$$\begin{aligned}
Q_u^t(s,a) &= \textsc{Clip}(\widetilde{Q}_u^t(s,a); L_u^t(s,a), U_u^t(s,a)) \\
&\geq \textsc{Clip}(Q^*(s,a); L_u^t(s,a), U_u^t(s,a)) \\
&\geq Q^*(s,a) \wedge U_u^t(s,a) \\
&= Q^*(s,a) \wedge (\widetilde{Q}_u^{t-1}(s,a) \wedge \widetilde{Q}_u^{t-2}(s,a)) \\
&\geq Q^*(s,a)
\end{aligned}$$

where the second inequality is by the clipping property (ii), and the last inequality holds by induction hypothesis. It follows that

$$\widetilde{V}_u^t(s) = \max_a Q_u^t(s,a) \geq \max_a Q^*(s,a) = V^*(s).$$

Note that by induction hypothesis, $\widetilde{V}_u^\tau(s) \geq V^*(s)$ for all $\tau \in [t-1]$, $u \in [\tau : T]$ and $s \in \mathcal{S}$. Hence, $m_t = \min\{\widetilde{V}_t^{t-1}(s_t), \widetilde{V}_{t-1}^{t-2}(s_{t-1}), \ldots, \widetilde{V}_2^1(s_2) \geq$

$\min\{V^*(s_t), V^*(s_{t-1}), \ldots, V^*(s_2), \frac{1}{1-\gamma}\} \geq \min_{s\in\mathcal{S}} V^*(s)$. It follows that

$$\begin{aligned}
V_u^t(s) &= \textsc{Clip}(\widetilde{V}_u^t(s); m_t, m_t + H) \\
&\geq \textsc{Clip}(V^*(s); m_t, m_t + H) \\
&\geq \textsc{Clip}(V^*(s); \min_{s'\in\mathcal{S}} V^*(s'), \min_{s'\in\mathcal{S}} V^*(s') + H) \\
&\geq V^*(s)
\end{aligned}$$

where the last inequality uses the fact that $H \geq 2 \cdot \mathrm{sp}(v^*)$ is chosen such that $\mathrm{sp}(V^*) \leq H$. We have shown that if $V_{u+1}^t(s) \geq V^*(s)$ holds for all $s \in \mathcal{S}$, then $V_u^t(s) \geq V^*(s)$, $\widetilde{V}_u^t(s)$ and $\widetilde{Q}_u^t(s,a)$ hold for all $(s,a) \in \mathcal{S} \times \mathcal{A}$. By induction on $u = T, \ldots, 1$, it follows that $V_u^t(s) \geq V^*(s)$, $\widetilde{V}_u^t(s) \geq V^*(s)$ and $\widetilde{Q}_u^t(s,a) \geq Q^*(s,a)$ hold for all $(s,a) \in \mathcal{S} \times \mathcal{A}$. The proof is complete by induction on t. □

Now, we show an upper bound of the action value function estimate, which is a direct consequence of the concentration inequality in Lemma 32.

A.3.4.2 Proof of Lemma 10

Proof of Lemma 10. We prove under the event $\mathcal{E}$ defined in Lemma 32, which holds with probability at least $1 - \delta$. Fix any $t \in [T]$ and $u \in [t : T]$. By event $\mathcal{E}$, we have

$$\begin{aligned}
\widetilde{Q}_u^t(s,a) &= \Big(r(s,a) + \gamma([\widehat{P}_t V_{u+1}^t](s,a) + \beta \|\boldsymbol{\varphi}(\cdot,\cdot)\|_{\Lambda_t^{-1}}\Big) \wedge \frac{1}{1-\gamma} \\
&\leq r(s,a) + \gamma[PV_{u+1}^t](s,a) + 2\beta\|\boldsymbol{\varphi}(s,a)\|_{\Lambda_t^{-1}}
\end{aligned}$$

for all $t \in [T]$. Hence, by Lemma 35, we have for $t \geq 4$ that

$$\begin{aligned}
\widetilde{Q}_u^{t-2}(s,a) &\leq r(s,a) + \gamma[PV_{u+1}^{t-2}](s,a) + 2\beta\|\boldsymbol{\varphi}(s,a)\|_{\Lambda_t^{-1}} \\
&\leq r(s,a) + \gamma[P(V_{u+1}^t)](s,a) + 2\beta\|\boldsymbol{\varphi}(s,a)\|_{\Lambda_t^{-1}} + m_{t-3} - m_{t-1} + m_{t-2} - m_t \\
&\leq r(s,a) + \gamma[PV_{u+1}^t](s,a) + 2\beta\|\boldsymbol{\varphi}(s,a)\|_{\Lambda_t^{-1}} + 2(m_{t-3} - m_t).
\end{aligned}$$

Therefore, for $t \geq 4$, we have

$$\begin{aligned}
Q_u^t(s,a) &= \textsc{Clip}(\widetilde{Q}_u^t(s,a); L_u^t(s,a), U_u^t(s,a)) \\
&\leq U_u^t(s,a) \\
&= \widetilde{Q}_u^{t-1}(s,a) \wedge \widetilde{Q}_u^{t-2}(s,a) \\
&\leq r(s,a) + \gamma[PV_{u+1}^t](s,a) + 2\beta\|\boldsymbol{\varphi}(s,a)\|_{\Lambda_t^{-1}} + 2(m_{t-3} - m_t)
\end{aligned}$$

□

A.3.4.3 Proof of Main Theorem

Now, we are ready to prove the main theorem.

Proof of Theorem 2. We prove under the event $\mathcal{E}$ defined in Lemma 32, which occurs with probability at least $1-\delta$. By Lemma 10, we have for $t \geq 4$,

$$Q_u^t(s,a) \leq r(s,a) + \gamma[PV_{u+1}^t](s,a) + 2\beta\|\boldsymbol{\varphi}(s,a)\|_{\Lambda_t^{-1}} + 2(m_{t-3} - m_t).$$

Plugging in $u \leftarrow t$, $s \leftarrow s_t$, $a \leftarrow a_t$, we get

$$\begin{aligned}
R_T &= \sum_{t=1}^{T}(J^* - r(s_t,a_t)) \\
&\leq \sum_{t=4}^{T}(J^* - Q_t^t(s_t,a_t) + \gamma[PV_{t+1}^t](s_t,a_t) + 2\beta\|\boldsymbol{\varphi}(s_t,a_t)\|_{\Lambda_t^{-1}} + 2(m_{t-3} - m_t)) + \mathcal{O}(1) \\
&= \underbrace{\sum_{t=4}^{T}(J^* - (1-\gamma)V_{t+1}^t(s_{t+1}))}_{(a)} + \underbrace{\sum_{t=4}^{T}(V_{t+1}^t(s_{t+1}) - Q_t^t(s_t,a_t))}_{(b)} \\
&\qquad\qquad + \gamma\underbrace{\sum_{t=4}^{T}([PV_{t+1}^t](s_t,a_t) - V_{t+1}^t(s_{t+1}))}_{(c)} + 2\,\beta\underbrace{\sum_{t=4}^{T}\|\boldsymbol{\varphi}(s_t,a_t)\|_{\Lambda_t^{-1}}}_{(d)} + \mathcal{O}(\frac{1}{1-\gamma}).
\end{aligned}$$

Bounding (a) By the optimism result (Lemma 9), we have $V_u^t(s) \geq V^*(s)$ for all $t \in [T]$ and $u \in [t:T]$ with high probability. It follows that

$$\begin{aligned}
J^* - (1-\gamma)V_{t+1}^t(s_{t+1}) &\leq J^* - (1-\gamma)V^*(s_{t+1}) \\
&\leq (1-\gamma)\mathrm{sp}(v^*)
\end{aligned}$$

where the last inequality is by the bound on the error of approximating the average-reward setting by the discounted setting provided in Lemma 1. Hence, the term (a) can be bounded by $T(1-\gamma)\mathrm{sp}(v^*)$.

Bounding (b) Using Lemma 7 that controls the difference between $\widetilde{V}_u^{t+1}$ and $\widetilde{V}_u^t$, we have

$$\begin{aligned}
V_{t+1}^t(s_{t+1}) &= \textsc{Clip}(\widetilde{V}_{t+1}^t(s_{t+1}); m_t, m_t + H) \\
&= \textsc{Clip}(\widetilde{V}_{t+1}^t(s_{t+1}) - m_t + m_{t+1}; m_{t+1}, m_{t+1} + H) + m_t - m_{t+1} \\
&\le \textsc{Clip}(\widetilde{V}_{t+1}^t(s_{t+1}); m_{t+1}, m_{t+1} + H) + m_t - m_{t+1} \\
&\le \widetilde{V}_{t+1}^t(s_{t+1}) + m_t - m_{t+1} \\
&\le \widetilde{V}_{t+1}^{t+1}(s_{t+1}) + 2m_{t-1} - 2m_{t+1}
\end{aligned}$$

where the second inequality holds because $\widetilde{V}_{t+1}^t(s_{t+1}) \ge m_{t+1}$ by Line 14. Hence, Term (b) can be bounded by $\mathcal{O}(\frac{1}{1-\gamma})$ using telescoping sums of $\widetilde{V}_{t+1}^{t+1}(s_{t+1}) - \widetilde{V}_t^t(s_t)$ and $2m_{t-1} - 2m_{t+1}$, and the fact that $V_u^t \le \frac{1}{1-\gamma}$ and $m_t \le \frac{1}{1-\gamma}$ for all $t \in [T]$ and $u \in [t : T]$.

Bounding (c) Since V_u^t is $\mathcal{F}_t$-measurable where $\mathcal{F}_t$ is history up to time step t, we have $\mathbb{E}[V_{t+1}^t(s_{t+1})|\mathcal{F}_t] = [PV_{t+1}^t](s_t, a_t)$, making the summation (c) a summation of a martingale difference sequence. Since $\mathrm{sp}(V_{t+1}^t) \le H$ for all $t \in [T]$, the summation can be bounded by $\mathcal{O}(\mathrm{sp}(v^*)\sqrt{T\log(1/\delta)})$ using Azuma-Hoeffding inequality.

Bounding (d) The sum of the bonus terms can be bounded by

$$\begin{aligned}
\beta \sum_{t=1}^T \|\boldsymbol{\varphi}(s_t, a_t)\|_{\Lambda_t^{-1}} &\le \beta\sqrt{T}\left(\sum_{t=1}^T \|\boldsymbol{\varphi}(s_t, a_t)\|_{\Lambda_t^{-1}}^2\right)^{1/2} \\
&\le \mathcal{O}(\beta\sqrt{dT\log T})
\end{aligned}$$

where the first inequality is by Cauchy-Schwartz and the last inequality is by Lemma 29.

Combining the four bounds, and choosing $H = 2 \cdot \mathrm{sp}(v^*)$ and choosing $\beta = \mathcal{O}(\mathrm{sp}(v^*)d\sqrt{\log(dT/\delta)})$ specified in Lemma 4, we get

$$R_T \le \mathcal{O}(T(1-\gamma)\mathrm{sp}(v^*) + \tfrac{1}{1-\gamma} + \mathrm{sp}(v^*)\sqrt{T\log(1/\delta)} + \mathrm{sp}(v^*)\sqrt{d^3T\log(dT/\delta)\log T}).$$

Choosing γ such that $\frac{1}{1-\gamma} = \sqrt{T}$, we get

$$R_T \le \mathcal{O}(\mathrm{sp}(v^*)\sqrt{d^3T\log(dT/\delta)\log T}).$$

□

A.3.5 Covering Numbers

In this section, we provide results on covering numbers of function classes used in this paper. We use the notation $\mathcal{N}_\epsilon(\mathcal{F}, \|\cdot\|)$ to denote the ε-covering number of the function class $\mathcal{F}$ with respect to the distance measure induced by the norm $\|\cdot\|$.

We first present a classical result that bounds the covering number of Euclidean ball.

Lemma 36. *For any $\varepsilon > 0$, the d-dimensional Euclidean ball $\mathbb{B}_d(R)$ with radius $R > 0$ has log-covering number upper bounded by*

$$\log \mathcal{N}_\varepsilon(\mathbb{B}_d(R), \|\cdot\|_2) \leq d \log(1 + 2R/\varepsilon).$$

Using this classical result, we bound the covering number of the function class that captures the functions $\widetilde{Q}_u^t(\cdot,\cdot)$ encountered by our algorithm.

Lemma 37 (Adaptation of Lemma D.6 in Jin et al. [2020])**.** *Let $\mathcal{Q}$ be a class of functions mapping from $\mathcal{S} \times \mathcal{A}$ to $\mathbb{R}$ with the following parametric form*

$$Q(\cdot,\cdot) = (\boldsymbol{w}^T \boldsymbol{\varphi}(\cdot,\cdot) + v + \beta\sqrt{\boldsymbol{\varphi}(\cdot,\cdot)^T \Lambda^{-1} \boldsymbol{\varphi}(\cdot,\cdot)}) \wedge M \tag{A.2}$$

where the parameters $(\boldsymbol{w}, \beta, v, \Lambda)$ satisfy $\|\boldsymbol{w}\| \leq L$, $\beta \in [0, B]$ and $v \in [0, D]$, and Λ is a positive definite matrix with minimum eigenvalue satisfying $\lambda_{min}(\Lambda) \geq \lambda > 0$. The constant $M > 0$ is fixed. Assume $\|\boldsymbol{\varphi}(s,a)\| \leq 1$ for all (s,a) pairs. Then

$$\log \mathcal{N}_\varepsilon(\mathcal{Q}, \|\cdot\|_\infty) \leq d \log(1 + 8L/\varepsilon) + \log(1 + 8D/\varepsilon) + d^2 \log[1 + 8d^{1/2}B^2/(\lambda\varepsilon^2)].$$

Proof. Introducing $\boldsymbol{A} = \beta^2 \Lambda^{-1}$, we can reparameterize as

$$Q(\cdot,\cdot) = (\boldsymbol{w}^T \boldsymbol{\varphi}(\cdot,\cdot) + v + \sqrt{\boldsymbol{\varphi}(\cdot,\cdot)^T \boldsymbol{A} \boldsymbol{\varphi}(\cdot,\cdot)}) \wedge M$$

where the parameters $(\boldsymbol{w}, v, \boldsymbol{A})$ satisfy $\|\boldsymbol{w}\|_2 \leq L$, $\|\boldsymbol{A}\| \leq B^2 \lambda^{-1}$, $v \in [0, D]$. For any pair of functions $Q_1, Q_2 \in \mathcal{Q}$ with parameterization $(\boldsymbol{w}_1, v_1, \boldsymbol{A}_1)$ and $(\boldsymbol{w}_2, v_2, \boldsymbol{A}_2)$, respectively, using the fact that $\cdot \wedge M$ is a contraction, we get

$$\begin{aligned}
\|Q_1 - Q_2\|_\infty &\le \sup_{s,a} |(\boldsymbol{w}_1^\top \boldsymbol{\varphi}(s,a) + v_1 + \sqrt{\boldsymbol{\varphi}(s,a)^\top \boldsymbol{A}_1 \boldsymbol{\varphi}(s,a)}) \qquad (A.3)\\
&\qquad - (\boldsymbol{w}_2^\top \boldsymbol{\varphi}(s,a) + v_2 + \sqrt{\boldsymbol{\varphi}(s,a)^\top \boldsymbol{A}_2 \boldsymbol{\varphi}(s,a)})|\\
&\le \sup_{\boldsymbol{\phi}:\|\boldsymbol{\phi}\|_2 \le 1} |(\boldsymbol{w}_1^\top \boldsymbol{\phi} + v_1 + \sqrt{\boldsymbol{\phi}^\top \boldsymbol{A}_1 \boldsymbol{\phi}}) - (\boldsymbol{w}_2^\top \boldsymbol{\phi} + v_2 + \sqrt{\boldsymbol{\phi}^\top \boldsymbol{A}_2 \boldsymbol{\phi}})|\\
&\le \sup_{\boldsymbol{\phi}:\|\boldsymbol{\phi}\|_2 \le 1} |(\boldsymbol{w}_1 - \boldsymbol{w}_2)^\top \boldsymbol{\phi}| + |v_1 - v_2| + \sup_{\boldsymbol{\phi}:\|\boldsymbol{\phi}\|_2 \le 1} \sqrt{|\boldsymbol{\phi}^\top (\boldsymbol{A}_1 - \boldsymbol{A}_2) \boldsymbol{\phi}|}\\
&= \|\boldsymbol{w}_1 - \boldsymbol{w}_2\|_2 + |v_1 - v_2| + \sqrt{\|\boldsymbol{A}_1 - \boldsymbol{A}_2\|_2}\\
&\le \|\boldsymbol{w}_1 - \boldsymbol{w}_2\|_2 + |v_1 - v_2| + \sqrt{\|\boldsymbol{A}_1 - \boldsymbol{A}_2\|_F} \qquad (A.4)
\end{aligned}$$

where the third inequality uses the fact that $|\sqrt{x} - \sqrt{y}| \le \sqrt{|x-y|}$ holds for any $x, y \ge 0$ and $\|\cdot\|_F$ denotes the Frobenius norm.

Let $\mathcal{C}_{\boldsymbol{w}}$ be an $\varepsilon/4$-cover of $\{\boldsymbol{w} \in \mathbb{R}^d : \|\boldsymbol{w}\| \le L\}$ with respect to the L2-norm, $\mathcal{C}_{\boldsymbol{A}}$ an $\varepsilon^2/4$-cover of $\{\boldsymbol{A} \in \mathbb{R}^{d\times d} : \|\boldsymbol{A}\|_F \le d^{1/2} B^2 \lambda^{-1}\}$ with respect to the Frobenius norm, and $\mathcal{C}_v$ an $\varepsilon/2$-cover of the interval $[0, D]$. Then, treating the matrix $\boldsymbol{A} \in \mathbb{R}^{d\times d}$ as a long vector of dimension $d \times d$, and applying Lemma 36, we know that we can find such covers with

$$\log|\mathcal{C}_{\boldsymbol{w}}| \le d\log(1 + 8L/\varepsilon), \quad \log|\mathcal{C}_{\boldsymbol{A}}| \le d^2 \log(1 + 8d^{1/2}B^2/(\lambda\varepsilon^2)), \quad \log|\mathcal{C}_v| \le \log(1 + 8D/\varepsilon).$$

Hence, the set of functions

$$\mathcal{C}_Q = \{Q \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}} : Q(\cdot,\cdot) = \boldsymbol{w}^T \boldsymbol{\varphi}(\cdot,\cdot) + v + \sqrt{\boldsymbol{\varphi}(\cdot,\cdot)^T \boldsymbol{A} \boldsymbol{\varphi}(\cdot,\cdot)}, \boldsymbol{w} \in \mathcal{C}_{\boldsymbol{w}}, \boldsymbol{A} \in \mathcal{C}_{\boldsymbol{A}}, v \in \mathcal{C}_v\}$$

has cardinality bounded by $\log|\mathcal{C}_Q| \le d\log(1 + 8L/\varepsilon) + d^2\log(1 + 8d^{1/2}B^2/(\lambda\varepsilon^2)) + \log(1 + 8D/\varepsilon)$. We can show that $\mathcal{C}_Q$ defined above is an ε-cover for $\mathcal{Q}$ as follows. Fix any $Q \in \mathcal{Q}$ parameterized by $(\boldsymbol{w}, v, \boldsymbol{A})$ and consider $\widetilde{Q} \in \mathcal{Q}$ parameterized by $(\widetilde{\boldsymbol{w}}, \widetilde{v}, \widetilde{\boldsymbol{A}})$ where $\widetilde{\boldsymbol{w}} \in \mathcal{C}_{\boldsymbol{w}}$ with $\|\boldsymbol{w} - \widetilde{\boldsymbol{w}}\|_2 \le \varepsilon/4$, $\widetilde{v} \in \mathcal{C}_v$ with $|v - \widetilde{v}| \le \varepsilon/4$ and $\widetilde{\boldsymbol{A}} \in \mathcal{C}_{\boldsymbol{A}}$ with $\|\boldsymbol{A} - \widetilde{\boldsymbol{A}}\|_F \le \varepsilon^2/4$. Then, by the bound (A.4), we have $\|Q - \widetilde{Q}\|_\infty \le \varepsilon$ as desired. This concludes the proof. □

Lemma 38. *Let $\mathcal{V}$ be a class of functions mapping from $\mathcal{S}$ to $\mathbb{R}$ defined as*

$$\mathcal{V} = \{\max_a Q(\cdot, a) : Q(\cdot,\cdot) = \textsc{Clip}(Q_1(\cdot,\cdot); Q_2(\cdot,\cdot) \vee Q_3(\cdot,\cdot), Q_4(\cdot,\cdot) \wedge Q_5(\cdot,\cdot), \;\; Q_1, \ldots, Q_5 \in \mathcal{Q}\}$$

where the function class $\mathcal{Q}$ is defined in Lemma 37. Then,

$$\log \mathcal{N}_\epsilon(\mathcal{V}, \|\cdot\|_\infty) \le 5d\log(1 + 8L/\varepsilon) + 5\log(1 + 8D/\varepsilon) + 5d^2\log[1 + 8d^{1/2}B^2/(\lambda\varepsilon^2)].$$

Proof. Let $\mathcal{W}$ be a class of functions mapping from $\mathcal{S} \times \mathcal{A} \to \mathbb{R}$ of the form

$$Q(\cdot,\cdot) = \textsc{Clip}(Q_1(\cdot,\cdot); Q_2(\cdot,\cdot) \vee Q_3(\cdot,\cdot), Q_4(\cdot,\cdot) \wedge Q_5(\cdot,\cdot))$$

where $Q_1, \ldots, Q_5 \in \mathcal{Q}$. Let $\mathcal{C}_0$ be an ϵ-cover of the function class $\mathcal{Q}$ with size $\log|\mathcal{C}_0| \leq d\log(1+8L/\varepsilon)+\log(1+4D/\varepsilon)+d^2\log[1+8d^{1/2}B^2/(\lambda\varepsilon^2)]$. Such a cover exists by Lemma 37. Let $\mathcal{C}$ be defined as

$$\mathcal{C} = \{Q \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}} : Q(\cdot,\cdot) = \textsc{Clip}(Q_1(\cdot,\cdot); Q_2(\cdot,\cdot)\vee Q_3(\cdot,\cdot), Q_4(\cdot,\cdot)\wedge Q_5(\cdot,\cdot),\ \ Q_1,\ldots,Q_5 \in \mathcal{C}_0\}.$$

Then, we have $\log|\mathcal{C}| \leq 5\log|\mathcal{C}_0|$, and we can show that $\mathcal{C}$ is an ε-cover of $\mathcal{W}$ as follows. Consider a function $W \in \mathcal{W}$, with $W(\cdot,\cdot) = \textsc{Clip}(Q_1(\cdot,\cdot); Q_2(\cdot,\cdot) \vee Q_3(\cdot,\cdot), Q_4(\cdot,\cdot) \wedge Q_5(\cdot,\cdot))$ where $Q_1, \ldots, Q_5 \in \mathcal{Q}$. Let $\widetilde{Q}_i \in \mathcal{C}_0$ be the approximation of Q_i for $i = 1, \ldots, 5$ such that $\|\widetilde{Q}_i - Q_i\|_\infty \leq \varepsilon$. Such a $\widetilde{Q}_i$ exists since $\mathcal{C}_0$ is an ε-cover of $\mathcal{Q}$. Let $\widetilde{Q}(\cdot,\cdot) = \textsc{Clip}(\widetilde{Q}_1(\cdot,\cdot); \widetilde{Q}_2(\cdot,\cdot) \vee \widetilde{Q}_3(\cdot,\cdot), \widetilde{Q}_4(\cdot,\cdot) \wedge Q_5(\cdot,\cdot))$. Then, $\widetilde{Q} \in \mathcal{C}$ and

$$\begin{aligned}
Q(\cdot,\cdot) &= \textsc{Clip}(Q_1(\cdot,\cdot); Q_2(\cdot,\cdot) \vee Q_3(\cdot,\cdot), Q_4(\cdot,\cdot) \wedge Q_5(\cdot,\cdot)) \\
&\leq \textsc{Clip}(\widetilde{Q}_1(\cdot,\cdot) + \varepsilon; (\widetilde{Q}_2(\cdot,\cdot) + \varepsilon) \vee (\widetilde{Q}_3(\cdot,\cdot) + \varepsilon), (\widetilde{Q}_4(\cdot,\cdot) + \varepsilon) \wedge (\widetilde{Q}_5(\cdot,\cdot) + \varepsilon)) \\
&= \textsc{Cilp}(\widetilde{Q}_1(\cdot,\cdot); \widetilde{Q}_2(\cdot,\cdot) \vee \widetilde{Q}_3(\cdot,\cdot), \widetilde{Q}_4(\cdot,\cdot) \wedge \widetilde{Q}_5(\cdot,\cdot)) + \varepsilon \\
&= \widetilde{Q}(\cdot,\cdot) + \varepsilon.
\end{aligned}$$

Similarly, we have

$$\begin{aligned}
Q(\cdot,\cdot) &= \textsc{Clip}(Q_1(\cdot,\cdot); Q_2(\cdot,\cdot) \vee Q_3(\cdot,\cdot), Q_4(\cdot,\cdot) \wedge Q_5(\cdot,\cdot)) \\
&\geq \textsc{Clip}(\widetilde{Q}_1(\cdot,\cdot) - \varepsilon; (\widetilde{Q}_2(\cdot,\cdot) - \varepsilon) \vee (\widetilde{Q}_3(\cdot,\cdot) - \varepsilon), (\widetilde{Q}_4(\cdot,\cdot) - \varepsilon) \wedge (\widetilde{Q}_5(\cdot,\cdot) - \varepsilon)) \\
&= \textsc{Cilp}(\widetilde{Q}_1(\cdot,\cdot); \widetilde{Q}_2(\cdot,\cdot) \vee \widetilde{Q}_3(\cdot,\cdot), \widetilde{Q}_4(\cdot,\cdot) \wedge \widetilde{Q}_5(\cdot,\cdot)) - \varepsilon \\
&= \widetilde{Q}(\cdot,\cdot) - \varepsilon,
\end{aligned}$$

which shows $\|Q - \widetilde{Q}\|_\infty \leq \varepsilon$, and that $\mathcal{C}$ is an ε-cover of $\mathcal{W}$. Since $\max_a$ is a contraction map, it follows that $\mathcal{V} = \{\max_a Q(\cdot, a) : Q \in \mathcal{W}\}$ is covered by $\widetilde{\mathcal{V}} = \{\max_a Q(\cdot, a) : Q \in \mathcal{C}\}$. The proof is complete by observing that $\log|\widetilde{\mathcal{V}}| \leq \log|\mathcal{C}| \leq 5\log|\mathcal{C}_0|$, and that there exists ε-cover $\mathcal{C}_0$ for $\mathcal{Q}$ with $\log|\mathcal{C}_0| \leq d\log(1 + 8L/\varepsilon) + \log(1 + 8D/\varepsilon) + d^2\log[1 + 8d^{1/2}B^2/(\lambda\varepsilon^2)]$ by Lemma 37. □

APPENDIX B

Analysis of Offline Constrained RL with Linear MDPs

B.1 Covering Numbers

Lemma 39 (Covering balls. e.g. Wainwright [2019])**.** *For any $\epsilon \in (0,1)$, we have*

$$\log \mathcal{N}(\mathbb{B}_d(r), \|\cdot\|_\infty, \epsilon) \le d \log\left(1 + \frac{2r}{\epsilon}\right).$$

Lemma 40 (Lemma 7 in Zanette et al. [2021])**.** *Consider a feature mapping $\boldsymbol{\varphi} : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$ such that $\|\boldsymbol{\varphi}(s,a)\|_2 \le 1$ for all $(s,a) \in \mathcal{S} \times \mathcal{A}$. Then for all $s \in \mathcal{S}$, we have*

$$\sum_{a\in\mathcal{A}} |\pi_{\boldsymbol{\theta}'}(a|s) - \pi_{\boldsymbol{\theta}}(a|s)| \le 8\|\boldsymbol{\theta} - \boldsymbol{\theta}'\|_2$$

for any pair $\boldsymbol{\theta}, \boldsymbol{\theta}' \in \mathbb{R}^d$ such that $\|\boldsymbol{\theta} - \boldsymbol{\theta}'\|_2 \le \frac{1}{2}$.

Lemma 41 (Covering softmax function class. Lemma 6 in Zanette et al. [2021])**.** *For any $\epsilon \in (0,1)$, we have*

$$\log \mathcal{N}(\Pi(B), \|\cdot\|_{\infty,1}, \epsilon) \le d \log\left(1 + \frac{16B}{\epsilon}\right)$$

where the norm $\|\cdot\|_{\infty,1}$ is defined by

$$\|\pi - \pi'\|_{\infty,1} := \sup_{s\in\mathcal{S}} \sum_{a\in\mathcal{A}} |\pi(a|s) - \pi'(a|s)|.$$

Lemma 42 (Covering number bound for the space of v)**.** *Consider the function class*

$$\mathcal{V} = \{v_{\boldsymbol{\zeta},\pi} : \boldsymbol{\zeta} \in \mathbb{B}(D_\zeta), \pi \in \Pi(D_\pi)\}$$

where $v_{\boldsymbol{\zeta},\pi} : \mathcal{S} \to \mathbb{R}$ is defined by $v_{\boldsymbol{\zeta},\pi}(s) = \sum_a \pi(a|s)\langle \boldsymbol{\zeta}, \varphi(s,a)\rangle$. Then,

$$\mathcal{N}(\mathcal{V}, \|\cdot\|_\infty, \epsilon) \leq \mathcal{N}(\mathbb{B}(D_\zeta), \|\cdot\|_2, \epsilon/2) \times \mathcal{N}(\Pi(D_\pi), \|\cdot\|_{\infty,1}, \epsilon/(2D_\zeta)).$$

and it follows that

$$\log \mathcal{N}(\mathcal{V}, \|\cdot\|_\infty, \epsilon) \leq \mathcal{O}(d \log(D_\zeta D_\pi/\epsilon)).$$

Proof. Consider $\mathcal{C}_v = \{v_{\boldsymbol{\zeta},\pi} \in (\mathcal{S} \to [0, \frac{1}{1-\gamma}]) : \boldsymbol{\zeta} \in \mathcal{C}_\zeta, \pi \in \mathcal{C}_\pi\}$ where $\mathcal{C}_\zeta$ is an $\epsilon/2$-cover of $\mathbb{B}(D_\zeta)$ with respect to $\|\cdot\|_2$ and $\mathcal{C}_\pi$ is an $\epsilon/(2D_\zeta)$-cover of $\Pi(D_\pi)$ with respect to $\|\cdot\|_{\infty,1}$. Such covers with $|\mathcal{C}_\zeta| \leq (1 + 4D_\zeta/\epsilon)^d$ and $|\mathcal{C}_\pi| \leq (1 + 32D_\zeta D_\pi/\epsilon)^d$ exist by previous lemmas. Consider any $v_{\boldsymbol{\zeta},\pi} \in \mathcal{V}$. Then, there exists $\boldsymbol{\zeta}' \in \mathcal{C}_\zeta$ and $\pi' \in \mathcal{C}_\pi$ with $\|\boldsymbol{\zeta} - \boldsymbol{\zeta}'\|_2 \leq \epsilon/2$ and $\|\pi - \pi'\|_{\infty,1} = \sup_{s\in\mathcal{S}} \sum_{a\in\mathcal{A}} |\pi(a|s) - \pi'(a|s)| \leq \epsilon/(2D_\zeta)$. Then for any $s \in \mathcal{S}$, $v_{\boldsymbol{\zeta}',\pi'} \in \mathcal{C}_v$ satisfies

$$\begin{aligned}
|v_{\boldsymbol{\zeta},\pi}(s) - v_{\boldsymbol{\zeta}',\pi'}(s)| &= |\sum_a \pi(a|s)\langle \boldsymbol{\zeta}, \varphi(s,a)\rangle - \sum_a \pi'(a|s)\langle \boldsymbol{\zeta}', \varphi(s,a)\rangle| \\
&= |\sum_a (\pi(a|s) - \pi'(a|s))\langle \boldsymbol{\zeta}, \varphi(s,a)\rangle + \pi'(a|s)\langle \boldsymbol{\zeta} - \boldsymbol{\zeta}', \varphi(s,a)\rangle| \\
&\leq D_\zeta \sum_a |\pi(a|s) - \pi'(a|s)| + \sum_a \pi'(a|s)\epsilon/2 \\
&\leq \epsilon.
\end{aligned}$$

It follows that $\mathcal{C}_v$ is an ϵ-cover of $\mathcal{V}$ with respect to $\|\cdot\|_\infty$ with $|\mathcal{C}_v| = |\mathcal{C}_\zeta||\mathcal{C}_\pi|$ and we are done. □

B.2 Concentration Inequalities

Lemma 43 (Matrix Bernstein). *Consider a finite sequence $\{S_k\}$ of independent, random matrices with common dimension $d_1 \times d_2$. Assume that $\mathbb{E}S_k = 0$ and $\|S_k\| \leq L$ for each index k. Let $Z = \sum_k S_k$ and define $v(Z) := \max\{\|\mathbb{E}[ZZ^T]\|, \|\mathbb{E}[Z^TZ]\|\} = \max\{\|\sum_k \mathbb{E}[S_kS_k^T]\|, \|\sum_k \mathbb{E}[S_k^TS_k]\|\}$. Then,*

$$P(\|Z\| \geq t) \leq (d_1 + d_2) \exp\left(\frac{-t^2/2}{v(Z) + Lt/3}\right)$$

and it follows that with probability at least $1 - \delta$, we have

$$\|Z\| \leq \frac{2L\log((d_1 + d_2)/\delta)}{3} + \sqrt{2v(Z)\log((d_1 + d_2)/\delta)}.$$

Lemma 44. *There exists an example where $\boldsymbol{\lambda}^* = \mathbb{E}_{\mu^*}[\boldsymbol{\varphi}(s,a)]$ is not in the span of $\boldsymbol{\varphi}(s_1,a_1),\ldots,\boldsymbol{\varphi}(s_n,a_n)$ with probability at least $1/2$.*

Proof. Consider the case where $\mathcal{S} = \{s\}$, $\mathcal{A} = \{a_1,a_2\}$, $d = 2$, and $\boldsymbol{\varphi}(s,a_1) = \boldsymbol{e}_1$ and $\boldsymbol{\varphi}(s,a_2) = \boldsymbol{e}_2$. Let $\mu = \mu^*$ and $\mu(s,a_1) = p$ and $\mu(s,a_2) = 1-p$. Let F be the event where $\boldsymbol{\lambda}^*$ is not in the span of $\boldsymbol{\varphi}(s_1,a_1),\ldots,\boldsymbol{\varphi}(s_n,a_n)$. Then,

$$P(F) = (1-p)^n \geq \frac{1}{2}$$

as long as we choose $p \leq 1 - 2^{-1/n}$. □

Proof of Lemma 13. Let $c_k = w^*(s_k,a_k) = \mu^*(s_k,a_k)/\mu(s_k,a_k)$, $k = 1,\ldots,n$. Note that $\mu(s_k,a_k) > 0$, $k = 1,\ldots,n$ must hold, otherwise such s_k, a_k cannot be sampled. By the concentrability assumption, we have $c_k \in [0, C^*]$, $k = 1,\ldots,n$. Let $\boldsymbol{z}_k = c_k\boldsymbol{\varphi}(s_k,a_k)$ for $k = 1,\ldots,n$. Then, $\|\boldsymbol{z}_k\| \leq C^*$ and

$$\mathbb{E}[\boldsymbol{z}_k] = \mathbb{E}_{(s,a)\sim\mu}[w^*(s,a)\boldsymbol{\varphi}(s,a)] = \mathbb{E}_{(s,a)\sim\mu}\left[\frac{\mu^*(s,a)}{\mu(s,a)}\boldsymbol{\varphi}(s,a)\right] = \mathbb{E}_{(s,a)\sim\mu^*}[\boldsymbol{\varphi}(s,a)] = \boldsymbol{\lambda}^*.$$

Define $\boldsymbol{S}_k = \boldsymbol{z}_k - \boldsymbol{\lambda}^*$, $k = 1,\ldots,n$ Then, $\mathbb{E}[\boldsymbol{S}_k] = 0$ and $\|\boldsymbol{S}_k\|_2 \leq \|\boldsymbol{z}_k\|_2 + \mathbb{E}[\|\boldsymbol{z}_k\|_2] \leq 2C^*$ and $\|\mathbb{E}[\boldsymbol{S}_k^T\boldsymbol{S}_k]\|_2 \leq \mathbb{E}[\boldsymbol{z}_k^T\boldsymbol{z}_k] \leq (C^*)^2$ and $\|\mathbb{E}[\boldsymbol{S}_k\boldsymbol{S}_k^T]\|_2 \leq (C^*)^2$. Applying matrix Bernstein inequality (Lemma 43) on $\{\boldsymbol{S}_k\}_{k=1}^n$, we have

$$\|\frac{1}{n}\sum_{k=1}^n \boldsymbol{S}_k\|_2 = \|\frac{1}{n}w^*(s_k,a_k)\boldsymbol{\varphi}(s_k,a_k) - \boldsymbol{\lambda}^*\|_2 \leq \frac{4C^*\log((d+1)/\delta)}{3n} + \sqrt{\frac{8(C^*)^2\log((d+1)/\delta)}{n}}$$

with probability at least $1-\delta$ and the result follows. □

Lemma 45. *Under the feature coverage assumption 6, there exists $\widehat{\boldsymbol{\lambda}}^* \in \mathbb{R}^d$ of the form $\widehat{\boldsymbol{\lambda}}^* = \frac{1}{n}\sum_{k=1}^n c_k\boldsymbol{\varphi}(s_k,a_k)$ with $c_k \in [0,C^*]$, $k = 1,\ldots,n$ such that*

$$\|\widehat{\boldsymbol{\lambda}}^* - \boldsymbol{\lambda}^*\|_2 \leq \mathcal{O}\left(C^*\sqrt{\frac{\log(d/\delta)}{n}}\right)$$

with probability at least $1-\delta$.

Proof. Since $\boldsymbol{\lambda}^* \in \mathrm{Col}(\boldsymbol{\Lambda})$, we have $\boldsymbol{\lambda}^* = \boldsymbol{\Lambda}\boldsymbol{\Lambda}^\dagger\boldsymbol{\lambda}^*$ and it follows that

$$\begin{aligned}\boldsymbol{\lambda}^* &= \boldsymbol{\Lambda}\boldsymbol{\Lambda}^\dagger\boldsymbol{\lambda}^* \\ &= \mathbb{E}[\boldsymbol{\varphi}(s,a)\boldsymbol{\varphi}(s,a)^T\boldsymbol{\Lambda}^\dagger\boldsymbol{\lambda}^*] \\ &= \mathbb{E}[\langle\boldsymbol{\varphi}(s,a), \boldsymbol{\Lambda}^\dagger\boldsymbol{\lambda}^*\rangle\boldsymbol{\varphi}(s,a)].\end{aligned}$$

Let $c_k = \langle \boldsymbol{\varphi}(s_k, a_k), \boldsymbol{\Lambda}^\dagger \boldsymbol{\lambda}^* \rangle$. Then, by Cauchy-Schwartz, we have

$$|c_k| \leq \|\boldsymbol{\varphi}(s_k, a_k)\|_2 \|\boldsymbol{\Lambda}^\dagger \boldsymbol{\lambda}^*\|_2 \leq C^*$$

where the last inequality follows by the feature coverage assumption. Using matrix Bernstein inequality as is done in the proof of Lemma 13, the result follows. □

B.2.1 Proof of Lemma 15

Proof of Lemma 15. Consider $\boldsymbol{\lambda} = \frac{1}{n}\sum_{k=1}^n c_k \boldsymbol{\varphi}(s_k, a_k)$ with $|c_k| \leq B$, $k = 1, \ldots, n$. Let $\widehat{\boldsymbol{\Lambda}}_n = \boldsymbol{U}\boldsymbol{D}\boldsymbol{U}^T$ be the eigendecomposition of $\widehat{\boldsymbol{\Lambda}}_n = \frac{1}{n}\sum_{k=1}^n \boldsymbol{\varphi}(s_k, a_k)\boldsymbol{\varphi}(s_k, a_k)^T$ where $\boldsymbol{D} = \operatorname{diag}(d_1, \ldots, d_d)$ with $d_1 \geq \cdots \geq d_d \geq 0$. Then, we have $\boldsymbol{D} = \boldsymbol{U}^T \widehat{\boldsymbol{\Lambda}}_n \boldsymbol{U} = \frac{1}{n}\sum_{k=1}^n \boldsymbol{U}^T \boldsymbol{\varphi}(s_k, a_k)\boldsymbol{\varphi}(s_k, a_k)^T \boldsymbol{U}$ and it follows that $d_i = \frac{1}{n}\sum_{k=1}^n \langle \boldsymbol{u}_i, \boldsymbol{\varphi}(s_k, a_k)\rangle^2$ where d_i is the ith diagonal entry of $\boldsymbol{D}$. Also,

$$\begin{aligned}
\boldsymbol{\lambda} &= \frac{1}{n}\sum_{k=1}^n c_k \boldsymbol{\varphi}(s_k, a_k) \\
&= \frac{1}{n}\sum_{k=1}^n c_k \sum_{i=1}^d \langle \boldsymbol{\varphi}(s_k, a_k), \boldsymbol{u}_i\rangle \boldsymbol{u}_i \\
&= \sum_{i=1}^d \left(\frac{1}{n}\sum_{k=1}^n c_k \langle \boldsymbol{\varphi}(s_k, a_k), \boldsymbol{u}_i\rangle \right) \boldsymbol{u}_i.
\end{aligned}$$

where the second equality follows by $\boldsymbol{x} = \boldsymbol{U}\boldsymbol{U}^T\boldsymbol{x} = \sum_{i=1}^d \langle \boldsymbol{x}, \boldsymbol{u}_i\rangle \boldsymbol{u}_i$ for any vector $\boldsymbol{x} \in \mathbb{R}^d$. So,

$$\begin{aligned}
\boldsymbol{\lambda}^T \widehat{\boldsymbol{\Lambda}}_n^\dagger \boldsymbol{\lambda} &= \sum_{i=1}^{d'} \left(\frac{1}{n}\sum_{k=1}^n c_k \langle \boldsymbol{\varphi}(s_k, a_k), \boldsymbol{u}_i\rangle \right)^2 / d_i \\
&= \frac{1}{n}\sum_{i=1}^{d'} \left(\sum_{k=1}^n c_k \langle \boldsymbol{\varphi}(s_k, a_k), \boldsymbol{u}_i\rangle \right)^2 \Big/ \left(\sum_{k=1}^n \langle \boldsymbol{\varphi}(s_k, a_k), \boldsymbol{u}_i\rangle^2 \right) \\
&\leq \frac{1}{n}\sum_{i=1}^{d'} \left(\sum_{k=1}^n c_k^2 \right) \\
&\leq dB^2
\end{aligned}$$

where d' is the number of strictly positive diagonal entries in $\boldsymbol{D}$ and the first inequality follows by Cauchy-Schwartz. □

B.2.2 Proof of Lemma 14

Lemma 46. *Let $v : \mathcal{S} \to [0, D_v]$. With probability at least $1-\delta$, we have*

$$\|\boldsymbol{\Psi v} - \widehat{\boldsymbol{\Psi v}}\|_{n\widehat{\boldsymbol{\Lambda}}_n+\boldsymbol{I}} \leq \mathcal{O}\left(D_v\sqrt{d\log(n/\delta)}\right)$$

where $\widehat{\boldsymbol{\Psi v}}$ is the least squares estimate defined in (3.13).

Proof. Note that $\|\boldsymbol{\Psi v}\|_2 \leq D_v\sqrt{d}$ by the boundedness assumption on $\boldsymbol{\Psi}$. The result follows directly from Theorem 2 in Abbasi-Yadkori et al. [2011]. □

Proof of Lemma 14. Let $\mathcal{C}$ be an $(1/n)$-cover on $\mathcal{V}$. By Lemma 42, such a cover with $\log|\mathcal{C}| \leq \mathcal{O}(d\log(D_\zeta D_\pi n))$ exists. Applying a union bound over $\mathcal{C}$ and using the concentration bound in Lemma 46, we get

$$\left\|\boldsymbol{\Psi v} - \widehat{\boldsymbol{\Psi v}}\right\|_{n\widehat{\boldsymbol{\Lambda}}_n+\boldsymbol{I}} \leq \mathcal{O}\left(D_v\sqrt{d\log(D_\zeta D_\pi n/\delta)}\right) \tag{B.1}$$

for all $\boldsymbol{v} \in \mathcal{C}$ with probability at least $1-\delta$. For any $\boldsymbol{v} \in \mathcal{V}$, we can find v' in the cover that satisfies $\|\boldsymbol{v} - \boldsymbol{v}'\|_\infty \leq 1/n$. Hence,

$$\begin{aligned}\left\|\boldsymbol{\Psi v} - \widehat{\boldsymbol{\Psi v}}\right\|_{n\widehat{\boldsymbol{\Lambda}}_n+\boldsymbol{I}} &\leq \|\boldsymbol{\Psi}(\boldsymbol{v}-\boldsymbol{v}') + \boldsymbol{\Psi v}' - \widehat{\boldsymbol{\Psi v}'} + \widehat{\boldsymbol{\Psi v}'} - \widehat{\boldsymbol{\Psi v}}\|_{n\widehat{\boldsymbol{\Lambda}}_n+\boldsymbol{I}} \\ &\leq \|\boldsymbol{\Psi}(\boldsymbol{v}-\boldsymbol{v}')\|_{n\widehat{\boldsymbol{\Lambda}}_n+\boldsymbol{I}} + \left\|(n\widehat{\boldsymbol{\Lambda}}_n + \boldsymbol{I})^{-1}\sum_{k=1}^n (v'(s'_k) - v(s'_k))\boldsymbol{\varphi}(s_k, a_k)\right\|_{n\widehat{\boldsymbol{\Lambda}}_n+\boldsymbol{I}} \\ &\quad + \mathcal{O}\left(D_v\sqrt{d\log(D_\zeta D_\pi n/\delta)}\right).\end{aligned}$$

The first term can be bounded using the boundedness assumption on $\boldsymbol{\Psi}$ by

$$\|\boldsymbol{\Psi}(\boldsymbol{v}-\boldsymbol{v}')\|^2_{n\widehat{\boldsymbol{\Lambda}}_n+\boldsymbol{I}} \leq \|\boldsymbol{\Psi}(\boldsymbol{v}-\boldsymbol{v}')\|_2^2\|n\widehat{\boldsymbol{\Lambda}}_n + \boldsymbol{I}\|_2 \leq d/n^2 \cdot (1+n) \leq \mathcal{O}(1)$$

as long as $n \geq d$. The second term can be bounded by

$$\begin{aligned}
&\left\| (n\widehat{\boldsymbol{\Lambda}} + \boldsymbol{I})^{-1} \sum_{k=1}^{n} (v'(s'_k) - v(s'_k))\boldsymbol{\varphi}(s_k, a_k) \right\|^2_{n\widehat{\boldsymbol{\Lambda}}+\boldsymbol{I}} \\
&\quad = \sum_{k=1}^{n} (v'(s'_k) - v(s'_k))\boldsymbol{\varphi}(s_k, a_k)^T (n\widehat{\boldsymbol{\Lambda}} + \boldsymbol{I})^{-1} \sum_{k=1}^{n} (v'(s'_k) - v(s'_k))\boldsymbol{\varphi}(s_k, a_k) \\
&\quad \leq \sum_{k=1}^{n} \|(v'(s'_k) - v(s'_k))\boldsymbol{\varphi}(s_k, a_k)\|_2^2 \\
&\quad \leq \sum_{k=1}^{n} (v'(s'_k) - v(s'_k))^2 \\
&\quad \leq 1
\end{aligned}$$

where the first inequality uses $n\widehat{\boldsymbol{\Lambda}} + \boldsymbol{I} \succcurlyeq \boldsymbol{I}$. The result follows. □

B.3 Computational Efficiency

In this section, we explain why our algorithms are computationally efficient by showing that the algorithms only require computing quantities for states that appear in the offline dataset to compute the policy π_t at each step. This is how we avoid computation complexity that scales with the size of the state space.

Recall that $\pi_t = \sigma(\alpha \sum_{i=1}^{t-1} \boldsymbol{\Phi}\boldsymbol{\zeta}_i)$ and by definition of $\sigma(\cdot)$,

$$\pi_t(a|s) = \frac{\exp(\alpha \sum_{i=1}^{t-1} \boldsymbol{\varphi}(s,a)^T \boldsymbol{\zeta}_i)}{\sum_{a'} \exp(\alpha \sum_{i=1}^{t-1} \boldsymbol{\varphi}(s,a')^T \boldsymbol{\zeta}_i)}.$$

We argue that the algorithm only needs to compute $\pi_t(a|s)$ for the states s that appear as the next state in the dataset $\mathcal{D}$. There are two parts where the object π_t is used in the algorithm:

Line 3 in Algorithm 5 and Line 4 in Algorithm 6 In these lines, the object π_t is used to compute

$$\widehat{\boldsymbol{\Psi}\boldsymbol{v}}_{\boldsymbol{\zeta}_t,\pi_t} = (n\widehat{\Lambda} + I)^{-1} \sum_{k=1}^{n} v_{\boldsymbol{\zeta}_t,\pi_t}(s'_k)\boldsymbol{\varphi}(s_k, a_k)$$

where $\boldsymbol{v}_{\boldsymbol{\zeta}_t,\pi_t}(s'_k) = (n\widehat{\Lambda} + I)^{-1} \sum_a \pi_t(a|s'_k)\langle \boldsymbol{\zeta}_t, \boldsymbol{\varphi}(s'_k, a)\rangle$. As we claimed, we only need to compute $\pi_t(\cdot|s'_k)$ for s'_k that appear in the dataset $\mathcal{D}$.

Line 2 in Algorithm 5 and Line 2 in Algorithm 6 In this lines, the object π_t is used to compute $\mathbf{\Phi}^T\widehat{\boldsymbol{\mu}}_{\boldsymbol{\lambda}_t,\pi_t}$ for $\boldsymbol{\lambda}_t$ of the form $\boldsymbol{\lambda}_t = \frac{1}{n}\sum_{k=1}^n c_k\boldsymbol{\varphi}(s_k, a_k)$. By definition,

$$\widehat{\mu}_{\lambda,\pi} = \pi \circ E[(1-\gamma)\nu_0 + \gamma\widehat{\Psi\lambda} = \pi \circ E[(1-\gamma)e_{s_0} + \gamma\frac{1}{n}\sum_{k=1}^n c_k e_{s'_k}]$$

and it follows that

$$\begin{aligned}\mathbf{\Phi}^T\widehat{\boldsymbol{\mu}}_{\boldsymbol{\lambda}_t,\pi_t} &= (1-\gamma)\mathbf{\Phi}^T(\pi_t \circ \boldsymbol{E}\boldsymbol{e}_{s_0}) + \gamma\frac{1}{n}\sum_{k=1}^n c_k\mathbf{\Phi}^T(\pi_t \circ \boldsymbol{E}\boldsymbol{e}_{s'_k}) \\ &= (1-\gamma)\sum_a \pi_t(s_0, a)\boldsymbol{\varphi}(s_0, a) + \gamma\frac{1}{n}\sum_{k=1}^n c_k \sum_a \pi_t(a|s'_k)\boldsymbol{\varphi}(s'_k, a).\end{aligned}$$

Again, we only need to compute $\pi_t(\cdot|s'_k)$ for s'_k that appears in $\mathcal{D}$.

B.4 Regret Analysis

B.4.1 Bounding the Regret of π-Player

Lemma 47 (Mirror Descent, Lemma D.2 in Gabbianelli et al. [2024a]). *Let $q_1, \ldots, q_T$ be a sequence of functions from $\mathcal{S} \times \mathcal{A}$ to $\mathbb{R}$ with $\|q_t\|_\infty \le D_q$ for $t = 1, \ldots, T$. Given an initial policy $\pi_1 : \mathcal{S} \to \Delta(\mathcal{A})$ and a learning rate $\alpha > 0$, define the sequence of policies $\pi_2, \ldots, \pi_{T+1}$ such that*

$$\pi_{t+1}(a|s) \propto \pi_t(a|s)\exp(\alpha q_t(s, a)).$$

Then, for any comparator policy π^*, we have

$$\sum_{t=1}^T \sum_{s\in\mathcal{S}} \nu^{\pi^*}(s)\langle\pi^*(\cdot|s) - \pi_t(\cdot|s), q_t(s, \cdot)\rangle \le \frac{\mathcal{H}(\pi^*\|\pi_1)}{\alpha} + \frac{\alpha T D_q^2}{2}$$

where $\mathcal{H}(\pi\|\pi') := \sum_{s\in\mathcal{S}} \nu^\pi(s)\mathcal{D}(\pi(\cdot|s)\|\pi'(\cdot|s))$ is the conditional entropy.

Lemma 48 (Lemma B.3 in Gabbianelli et al. [2024a]). *The sequence of policies $\pi_1, \ldots, \pi_T$ produced by an exponentiation algorithm $\pi_{t+1} = \sigma(\alpha\sum_{i=1}^t \mathbf{\Phi}\boldsymbol{\zeta}_i)$ satisfies*

$$\sum_{t=1}^T \sum_{s\in\mathcal{S}} \nu^{\pi^*}(s)\sum_{a\in\mathcal{A}}(\pi^*(a|s) - \pi_t(a|s))\langle\boldsymbol{\zeta}_t, \boldsymbol{\varphi}(s, a)\rangle \le \frac{\log|\mathcal{A}|}{\alpha} + \frac{\alpha T D_\varphi^2 D_\zeta^2}{2}$$

where $\|\boldsymbol{\zeta}_t\|_2 \le D_\zeta$, $t = 1, \ldots, T$ and $\|\boldsymbol{\varphi}(\cdot, \cdot)\|_2 \le D_\varphi$.

B.4.2 Bounding the regret of ζ-player

Proof of Lemma 11. Recall $\mu_{\boldsymbol{\lambda}(\boldsymbol{c}),\pi}(s,a) = \pi(a|s)\left[(1-\gamma)\nu_0(s) + \gamma(\boldsymbol{\Psi}^T\boldsymbol{\lambda}(\boldsymbol{c}))_s\right]$ where we use the notation $(\boldsymbol{x})_s$ to denote the sth entry of vector $\boldsymbol{x}$. We can write

$$\begin{aligned}
\boldsymbol{\Phi}^T\boldsymbol{\mu}_{\boldsymbol{\lambda}(\boldsymbol{c}),\pi} &= \sum_{s,a}\mu_{\boldsymbol{\lambda}(\boldsymbol{c}),\pi}(s,a)\boldsymbol{\varphi}(s,a)\\
&= \sum_{s,a}\pi(a|s)[(1-\gamma)\nu_0(s) + \gamma(\boldsymbol{\Psi}^T\boldsymbol{\lambda}(\boldsymbol{c}))_s]\boldsymbol{\varphi}(s,a)\\
&= (1-\gamma)\sum_a \pi(a|s_0)\boldsymbol{\varphi}(s_0,a) + \gamma\sum_s(\boldsymbol{\Psi}^T\boldsymbol{\lambda}(\boldsymbol{c}))_s\sum_a\pi(a|s)\boldsymbol{\varphi}(s,a)\\
&= (1-\gamma)\boldsymbol{\varphi}(s_0,\pi) + \gamma\sum_s(\boldsymbol{\Psi}^T\boldsymbol{\lambda}(\boldsymbol{c}))_s\boldsymbol{\varphi}(s,\pi)
\end{aligned}$$

where we use the notation $\boldsymbol{\varphi}(s,\pi) = \sum_a\pi(a|s)\boldsymbol{\varphi}(s,a)$. Recall that $\boldsymbol{\lambda}(\boldsymbol{c}) = \frac{1}{n}\sum_{k=1}^n c_k\boldsymbol{\varphi}(s_k,a_k)$ where $c_k\in[-B,B]$, $k=1,\ldots,n$. Following the same argument for expanding $\boldsymbol{\Phi}^T\boldsymbol{\mu}_{\boldsymbol{\lambda}(\boldsymbol{c}),\pi}$, we get

$$\begin{aligned}
\boldsymbol{\Phi}^T\widehat{\boldsymbol{\mu}}_{\boldsymbol{\lambda}(\boldsymbol{c}),\pi} &= (1-\gamma)\boldsymbol{\varphi}(s_0,\pi) + \gamma\sum_s(\widehat{\boldsymbol{\Psi}^T\boldsymbol{\lambda}}(\boldsymbol{c}))_s\boldsymbol{\varphi}(s,\pi)\\
&= (1-\gamma)\boldsymbol{\varphi}(s_0,\pi) + \frac{\gamma}{n}\sum_{k=1}^n c_k\boldsymbol{\varphi}(s'_k,\pi).
\end{aligned}$$

Also, using $\boldsymbol{\Psi}^T\boldsymbol{\lambda}(\boldsymbol{c}) = \frac{1}{n}\sum_{k=1}^n c_k\boldsymbol{\Psi}^T\boldsymbol{\varphi}(s_k,a_k) = \frac{1}{n}\sum_{k=1}^n c_kP(\cdot|s_k,a_k) = \frac{1}{n}\sum_{k=1}^n c_k\mathbb{E}[\boldsymbol{e}_{s'_k}|s_k,a_k]$, we get

$$\begin{aligned}
\boldsymbol{\Phi}^T\boldsymbol{\mu}_{\boldsymbol{\lambda}(\boldsymbol{c}),\pi} &= (1-\gamma)\boldsymbol{\varphi}(s_0,\pi) + \gamma\sum_s(\boldsymbol{\Psi}^T\boldsymbol{\lambda}(\boldsymbol{c}))_s\boldsymbol{\varphi}(s,\pi)\\
&= (1-\gamma)\boldsymbol{\varphi}(s_0,\pi) + \frac{\gamma}{n}\sum_{k=1}^n c_k\mathbb{E}[\boldsymbol{\varphi}(s'_k,\pi)|s_k,a_k].
\end{aligned}$$

Hence,

$$\begin{aligned}
\|\boldsymbol{\Phi}^T(\boldsymbol{\mu}_{\boldsymbol{\lambda}(\boldsymbol{c}),\pi} - \widehat{\boldsymbol{\mu}}_{\boldsymbol{\lambda}(\boldsymbol{c}),\pi})\|_2 &= \gamma\left\|\frac{1}{n}\sum_{k=1}^n c_k(\boldsymbol{\varphi}(s'_k,\pi) - \mathbb{E}[\boldsymbol{\varphi}(s'_k,\pi)|s_k,a_k])\right\|_2\\
&\leq \mathcal{O}\left(B\sqrt{\frac{\log(d/\delta)}{n}}\right)
\end{aligned}$$

where the last inequality uses Matrix Bernstein inequality (Lemma 43) with $S_k = c_k\boldsymbol{\varphi}(s'_k,\pi) - c_k\mathbb{E}[\boldsymbol{\varphi}(s'_k,\pi)|s_k,a_k]$. □

Lemma 49. *Given a fixed* $\boldsymbol{\lambda} \in \mathcal{C}_n(B)$*, we have for all* $\pi \in \Pi(D_\pi)$ *that*

$$\|\boldsymbol{\Phi}^T \boldsymbol{\mu}_{\boldsymbol{\lambda},\pi} - \boldsymbol{\Phi}^T \widehat{\boldsymbol{\mu}}_{\boldsymbol{\lambda},\pi}\|_2 \leq \mathcal{O}\left(B\sqrt{\frac{\log(d/\delta) + d\log(D_\pi dn)}{n}}\right)$$

with probability at least $1-\delta$*.*

Proof. Consider an ε-cover of $\Pi(D_\pi)$ with covering balls when measuring distances with the norm $\|\pi - \pi'\|_{\infty,1} = \sup_{s\in\mathcal{S}} \sum_{a\in\mathcal{A}} |\pi(a|s) - \pi'(a|s)|$. By Lemma 41, there exists such a cover with log covering number

$$\log \mathcal{N}(\Pi(D_\pi), \|\cdot\|_{\infty,1}, \varepsilon) \leq d\log\left(1 + \frac{16D_\pi}{\varepsilon}\right).$$

Fix any $\pi \in \Pi(D_\pi)$ and consider its nearest cover center π' measuring distances by $\|\cdot\|_{\infty,1}$. Then,

$$\begin{aligned}\|\boldsymbol{\Phi}^T \boldsymbol{\mu}_{\boldsymbol{\lambda},\pi} - \boldsymbol{\Phi}^T \widehat{\boldsymbol{\mu}}_{\boldsymbol{\lambda},\pi}\|_2 \leq \| \underbrace{\boldsymbol{\Phi}^T \boldsymbol{\mu}_{\boldsymbol{\lambda},\pi} - \boldsymbol{\Phi}^T \boldsymbol{\mu}_{\boldsymbol{\lambda},\pi'}}_{(i)} \|_2 + \| \underbrace{\boldsymbol{\Phi}^T \boldsymbol{\mu}_{\boldsymbol{\lambda},\pi'} - \boldsymbol{\Phi}^T \widehat{\boldsymbol{\mu}}_{\boldsymbol{\lambda},\pi'}}_{(ii)} \|_2 \\ + \| \underbrace{\boldsymbol{\Phi}^T \widehat{\boldsymbol{\mu}}_{\boldsymbol{\lambda},\pi'} - \boldsymbol{\Phi}^T \widehat{\boldsymbol{\mu}}_{\boldsymbol{\lambda},\pi}}_{(iii)} \|_2.\end{aligned}$$

Note that

$$\begin{aligned}\boldsymbol{\Phi}^T \boldsymbol{\mu}_{\boldsymbol{\lambda},\pi} &= \sum_{s,a} \mu_{\boldsymbol{\lambda},\pi}(s,a)\boldsymbol{\varphi}(s,a) \\ &= \sum_{s,a} \pi(a|s)[(1-\gamma)\nu_0(s) + \gamma(\boldsymbol{\Psi}^T\boldsymbol{\lambda})_s]\boldsymbol{\varphi}(s,a) \\ &= (1-\gamma)\sum_a \pi(a|s_0)\boldsymbol{\varphi}(s_0,a) + \gamma\sum_s (\boldsymbol{\Psi}^T\boldsymbol{\lambda})_s \sum_a \pi(a|s)\boldsymbol{\varphi}(s,a) \\ &= (1-\gamma)\boldsymbol{\varphi}(s_0,\pi) + \gamma\sum_s (\boldsymbol{\Psi}^T\boldsymbol{\lambda})_s \boldsymbol{\varphi}(s,\pi), \qquad \text{(B.2)}\end{aligned}$$

where we use the notation $\boldsymbol{\varphi}(s,\pi) = \sum_a \pi(a|s)\boldsymbol{\varphi}(s,a)$. The first term (i) can be bounded by

$$\begin{aligned}
\|\boldsymbol{\Phi}^T\boldsymbol{\mu}_{\boldsymbol{\lambda},\pi} - \boldsymbol{\Phi}^T\boldsymbol{\mu}_{\boldsymbol{\lambda},\pi'}\|_2 &\leq (1-\gamma)\left\|\boldsymbol{\varphi}(s_0,\pi-\pi')\right\|_2 + \gamma\left\|\sum_s(\boldsymbol{\Psi}^T\boldsymbol{\lambda})_s\boldsymbol{\varphi}(s,\pi-\pi')\right\|_2 \\
&= (1-\gamma)\left\|\sum_a(\pi(a|s_0)-\pi'(a|s_0))\boldsymbol{\varphi}(s_0,a)\right\|_2 \\
&\quad + \gamma\left\|\sum_s(\boldsymbol{\Phi}^T\boldsymbol{\lambda})_s\sum_a(\pi(a|s)-\pi'(a|s))\boldsymbol{\varphi}(s,a)\right\|_2 \\
&\leq (1-\gamma)\sum_a|\pi(a|s_0)-\pi'(a|s_0)|\|\boldsymbol{\varphi}(s_0,a)\|_2 \\
&\quad + \gamma\sum_s(|\boldsymbol{\Psi}|^T|\boldsymbol{\lambda}|)_s\sum_a|\pi(a|s)-\pi'(a|s)|\|\boldsymbol{\varphi}(s,a)\|_2 \\
&\leq (1-\gamma)\varepsilon + \gamma\varepsilon\mathbf{1}_{\mathcal{S}}^T|\boldsymbol{\Psi}|^T|\boldsymbol{\lambda}| \\
&\leq (1-\gamma)\varepsilon + \gamma\varepsilon D_\psi\sqrt{d}\|\boldsymbol{\lambda}\|_2 \\
&\leq (1-\gamma)\varepsilon + \gamma\varepsilon\sqrt{d}D_\psi B
\end{aligned}$$

where the second to last inequality uses the boundedness assumption on $\boldsymbol{\Psi}$ and the last inequality follows by $\|\boldsymbol{\lambda}\|_2 \leq B$.

The second term (ii) can be bounded by a union bound of the concentration inequality in Lemma 11 across all π' in the cover, resulting in the following bound

$$\|\boldsymbol{\Phi}^T\boldsymbol{\mu}_{\boldsymbol{\lambda},\pi'} - \boldsymbol{\Phi}^T\widehat{\boldsymbol{\mu}}_{\boldsymbol{\lambda},\pi'}\|_2 \leq \mathcal{O}\left(B\sqrt{\frac{\log(d/\delta)+d\log(D_\pi/\varepsilon)}{n}}\right).$$

To bound the third term (iii), note that

$$\begin{aligned}
\boldsymbol{\Phi}^T\widehat{\boldsymbol{\mu}}_{\boldsymbol{\lambda},\pi} &= (1-\gamma)\boldsymbol{\varphi}(s_0,\pi) + \gamma\sum_s(\widehat{\boldsymbol{\Psi}^T\boldsymbol{\lambda}})_s\boldsymbol{\varphi}(s,\pi) \\
&= (1-\gamma)\boldsymbol{\varphi}(s_0,\pi) + \frac{\gamma}{n}\sum_{k=1}^n c_k\boldsymbol{\varphi}(s'_k,\pi). \qquad \text{(B.3)}
\end{aligned}$$

where $\boldsymbol{\lambda} = \frac{1}{n}\sum_{k=1}^n c_k\boldsymbol{\varphi}(s_k,a_k)$. Therefore,

$$\begin{aligned}
\|\boldsymbol{\Phi}^T\widehat{\boldsymbol{\mu}}_{\boldsymbol{\lambda},\pi'} - \boldsymbol{\Phi}^T\widehat{\boldsymbol{\mu}}_{\boldsymbol{\lambda},\pi}\|_2 &\leq (1-\gamma)\|\boldsymbol{\varphi}(s_0,\pi-\pi')\|_2 + \frac{\gamma}{n}\sum_{k=1}^n c_k\|\boldsymbol{\varphi}(s'_k,\pi-\pi')\|_2 \\
&\leq (1-\gamma)\varepsilon + \gamma B\varepsilon
\end{aligned}$$

where the last inequality uses $\|\boldsymbol{\varphi}(s, \pi - \pi')\|_2 = \|\sum_a (\pi(a|s) - \pi'(a|s))\boldsymbol{\varphi}(s,a)\|_2 \leq \sum_a |\pi(a|s) - \pi'(a|s)| \|\boldsymbol{\varphi}(s,a)\|_2 \leq \varepsilon$. Combining the bounds of the three terms, we get

$$\|\boldsymbol{\Phi}^T \boldsymbol{\mu}_{\boldsymbol{\lambda},\pi} - \boldsymbol{\Phi}^T \widehat{\boldsymbol{\mu}}_{\boldsymbol{\lambda},\pi}\|_2 \leq \mathcal{O}\left(B\sqrt{\frac{\log(d/\delta) + d\log(D_\pi/\varepsilon)}{n}} + \sqrt{d}B\varepsilon \right).$$

Choosing $\varepsilon = 1/\sqrt{dn}$, we get the desired result. □

Lemma 50. *The sequences $\{\pi_t\}, \{\boldsymbol{\lambda}(\boldsymbol{c}'_t)\}, \{\boldsymbol{\zeta}_t\}$ produced by Algorithm 5 satisfies*

$$\mathrm{REG}_t^\zeta = \langle \boldsymbol{\zeta}_t - \boldsymbol{\zeta}^{\pi_t}, \boldsymbol{\Phi}^T \boldsymbol{\mu}_{\boldsymbol{\lambda}(\boldsymbol{c}'_t),\pi_t} - \boldsymbol{\lambda}(\boldsymbol{c}'_t) \rangle \leq \mathcal{O}\left(\frac{C^* d}{1-\gamma} \sqrt{\frac{\log(dnT(\log|\mathcal{A}|)/\delta)}{n}} \right)$$

for $t = 1, \ldots, T$ with probability at least $1 - \delta$.

Proof. Recall that $\mathcal{I} \subseteq \{1, \ldots, n\}$ is an index set of size d such that $\{\boldsymbol{\varphi}(s_j, a_j)\}_{j \in \mathcal{I}}$ is a 2-approximate barycentric spanner for $\{\boldsymbol{\varphi}(s_k, a_k)\}_{k=1}^n$. Let $\mathcal{C}'_n(C^*) = \{\boldsymbol{c}' \in [-2C^*, 2C^*]^n : c'_j = 0 \text{ if } j \in \mathcal{I}\}$. Consider a ε-cover of $\mathcal{C}'_n(C^*)$ with respect to distance induced by $\|\cdot\|_\infty$ where ε is to be chosen later, and let $\boldsymbol{c}''_t$ be the closest covering center to $\boldsymbol{c}'_t$. There exists a cover with covering number $(1 + 4C^*/\varepsilon)^d$. We can decompose the regret of ζ-player at step t into

$$\begin{aligned}
\mathrm{REG}_t^\zeta &= \langle \boldsymbol{\zeta}_t - \boldsymbol{\zeta}^{\pi_t}, \boldsymbol{\Phi}^T \boldsymbol{\mu}_{\boldsymbol{\lambda}(\boldsymbol{c}'_t),\pi_t} - \boldsymbol{\lambda}(\boldsymbol{c}'_t) \rangle \\
&= \underbrace{\left\langle \boldsymbol{\zeta}_t - \boldsymbol{\zeta}^{\pi_t}, \boldsymbol{\Phi}^T \boldsymbol{\mu}_{\boldsymbol{\lambda}(\boldsymbol{c}'_t),\pi_t} - \boldsymbol{\Phi}^T \boldsymbol{\mu}_{\boldsymbol{\lambda}_t(\boldsymbol{c}''_t),\pi_t} \right\rangle}_{(a)} + \underbrace{\left\langle \boldsymbol{\zeta}_t - \boldsymbol{\zeta}^{\pi_t}, \boldsymbol{\Phi}^T \boldsymbol{\mu}_{\boldsymbol{\lambda}(\boldsymbol{c}''_t),\pi_t} - \boldsymbol{\Phi}^T \widehat{\boldsymbol{\mu}}_{\boldsymbol{\lambda}(\boldsymbol{c}''_t),\pi_t} \right\rangle}_{(b)} \\
&\quad + \underbrace{\left\langle \boldsymbol{\zeta}_t - \boldsymbol{\zeta}^{\pi_t}, \boldsymbol{\Phi}^T \widehat{\boldsymbol{\mu}}_{\boldsymbol{\lambda}(\boldsymbol{c}''_t),\pi_t} - \boldsymbol{\Phi}^T \widehat{\boldsymbol{\mu}}_{\boldsymbol{\lambda}(\boldsymbol{c}'_t),\pi_t} \right\rangle}_{(c)} + \underbrace{\left\langle \boldsymbol{\zeta}_t - \boldsymbol{\zeta}^{\pi_t}, \boldsymbol{\Phi}^T \widehat{\boldsymbol{\mu}}_{\boldsymbol{\lambda}(\boldsymbol{c}'_t),\pi_t} - \boldsymbol{\lambda}(\boldsymbol{c}'_t) \right\rangle}_{(d)}.
\end{aligned}$$

Bounding (a) Recall from equation (B.2) that

$$\boldsymbol{\Phi}^T \boldsymbol{\mu}_{\boldsymbol{\lambda},\pi} = (1-\gamma)\boldsymbol{\varphi}(s_0, \pi) + \gamma \sum_s (\boldsymbol{\Psi}^T \boldsymbol{\lambda})_s \boldsymbol{\varphi}(s, \pi)$$

where we use the notation $\boldsymbol{\varphi}(s, \pi) = \sum_a \pi(a|s)\boldsymbol{\varphi}(s,a)$. Also, since $\|\boldsymbol{c}'_t - \boldsymbol{c}''_t\|_\infty \leq \varepsilon$, we have

$$\|\boldsymbol{\lambda}(\boldsymbol{c}'_t) - \boldsymbol{\lambda}(\boldsymbol{c}''_t)\|_2 = \frac{1}{n}\left\| \sum_{k=1}^n (c'_{tk} - c''_{tk})\boldsymbol{\varphi}(s_k, a_k) \right\|_2 \leq \frac{1}{n}\sum_{k=1}^n |c'_{tk} - c''_{tk}| \|\boldsymbol{\varphi}(s_k, a_k)\|_2 \leq \varepsilon.$$

Hence,

$$\begin{aligned}\|\boldsymbol{\Phi}^T\boldsymbol{\mu}_{\boldsymbol{\lambda}(\boldsymbol{c}_t'),\pi_t} - \boldsymbol{\Phi}^T\boldsymbol{\mu}_{\boldsymbol{\lambda}(\boldsymbol{c}_t''),\pi_t}\|_2 &= \gamma\|\sum_{s\in\mathcal{S}}(\boldsymbol{\Psi}^T(\boldsymbol{\lambda}(\boldsymbol{c}_t') - \boldsymbol{\lambda}(\boldsymbol{c}_t'')))_s\boldsymbol{\varphi}(s,\pi)\| \\ &\leq \gamma\sum_{s\in\mathcal{S}}|(\boldsymbol{\Psi}^T(\boldsymbol{\lambda}(\boldsymbol{c}_t') - \boldsymbol{\lambda}(\boldsymbol{c}_t'')))_s|\|\boldsymbol{\varphi}(s,\pi)\|_2 \\ &\leq \gamma\sum_{s\in\mathcal{S}}|(\boldsymbol{\Psi}^T(\boldsymbol{\lambda}(\boldsymbol{c}_t') - \boldsymbol{\lambda}(\boldsymbol{c}_t'')))_s| \\ &\leq \gamma\varepsilon\mathbf{1}_{\mathcal{S}}^T|\boldsymbol{\Psi}|^T\mathbf{1}_d \\ &\leq \gamma\varepsilon D_\psi d\end{aligned}$$

where we use the notation $|\boldsymbol{\Psi}|$ for the matrix that takes element-wise absolute value of $\boldsymbol{\Psi}$. The second inequality follows since $\|\boldsymbol{\psi}(s,\pi)\|_2 = \|\sum_a \pi(a|s)\boldsymbol{\varphi}(s,a)\|_2 \leq \sum_a \pi(a|s)\|\boldsymbol{\varphi}(s,a)\|_2 \leq \sum_a \pi(a|s) = 1$. The last inequality follows by the boundedness assumption on $\boldsymbol{\Phi}$. Hence, choosing $\varepsilon = 1/\sqrt{dn}$, Term (a) can be bounded by

$$\begin{aligned}\langle\boldsymbol{\zeta}_t - \boldsymbol{\zeta}_{\boldsymbol{w}_t}^{\pi_t}, \boldsymbol{\Phi}^T\boldsymbol{\mu}_{\boldsymbol{\lambda}_t,\pi_t} - \boldsymbol{\Phi}^T\boldsymbol{\mu}_{\boldsymbol{\lambda}_t',\pi_t}\rangle &\leq \|\boldsymbol{\zeta}_t - \boldsymbol{\zeta}_{\boldsymbol{w}_t}^{\pi_t}\|_2\|\boldsymbol{\Phi}^T\boldsymbol{\mu}_{\boldsymbol{\lambda}_t,\pi_t} - \boldsymbol{\Phi}^T\boldsymbol{\mu}_{\boldsymbol{\lambda}_t',\pi_t}\|_2 \\ &\leq \frac{2\gamma D_\zeta D_\psi\sqrt{d}}{\sqrt{n}}.\end{aligned}$$

Bounding (b) The second term can be bounded by a union bound of the concentration inequality in Lemma 49 over a $(1/\sqrt{dn})$-cover of $\mathcal{C}_n'(C^*) = \{\boldsymbol{c}' \in [-2C^*, 2C^*]^n : c_j' = 0 \text{ if } j \in \mathcal{I}\}$, which gives

$$\begin{aligned}\langle\boldsymbol{\zeta}_t - \boldsymbol{\zeta}^{\pi_t}, \boldsymbol{\Phi}^T\boldsymbol{\mu}_{\boldsymbol{\lambda}_t',\pi_t} - \boldsymbol{\Phi}^T\widehat{\boldsymbol{\mu}}_{\boldsymbol{\lambda}_t',\pi_t}\rangle &\leq \|\boldsymbol{\zeta}_t - \boldsymbol{\zeta}^{\pi_t}\|_2\|\boldsymbol{\Phi}^T\boldsymbol{\mu}_{\boldsymbol{\lambda}_t',\pi_t} - \boldsymbol{\Phi}^T\widehat{\boldsymbol{\mu}}_{\boldsymbol{\lambda}_t',\pi_t}\|_2 \\ &\leq \mathcal{O}\left(D_\zeta C^*\sqrt{\frac{d\log(D_\pi dn/\delta)}{n}}\right)\end{aligned}$$

Bounding (c) Recall from (B.3) that $\boldsymbol{\Phi}^T\widehat{\boldsymbol{\mu}}_{\boldsymbol{\lambda},\pi} = (1-\gamma)\boldsymbol{\varphi}(s_0,\pi) + \frac{\gamma}{n}\sum_{k=1}^n c_k\boldsymbol{\varphi}(s_k',\pi)$. Since $\|c_t' - c_t''\|_\infty \leq 1/\sqrt{dn}$, we have

$$\|\boldsymbol{\Phi}^T\widehat{\boldsymbol{\mu}}_{\boldsymbol{\lambda}(\boldsymbol{c}_t''),\pi_t} - \boldsymbol{\Phi}^T\widehat{\boldsymbol{\mu}}_{\boldsymbol{\lambda}(\boldsymbol{c}_t'),\pi_t}\|_2 = \frac{\gamma}{n}\left\|\sum_{k=1}^n(c_{tk}'' - c_{tk}')\boldsymbol{\varphi}(s_k',\pi_t)\right\|_2 \leq \gamma/\sqrt{dn}.$$

It follows by Cauchy-Schwartz that

$$\langle\boldsymbol{\zeta}_t - \boldsymbol{\zeta}^{\pi_t}, \boldsymbol{\Phi}^T\widehat{\boldsymbol{\mu}}_{\boldsymbol{\lambda}(\boldsymbol{c}_t''),\pi_t} - \boldsymbol{\Phi}^T\widehat{\boldsymbol{\mu}}_{\boldsymbol{\lambda}(\boldsymbol{c}_t'),\pi_t}\rangle \leq \mathcal{O}(D_\zeta/\sqrt{dn}).$$

Bounding (d) Recall that ζ-player chooses $\boldsymbol{\zeta}_t \in \mathbb{B}_d(D_\zeta)$ greedily that minimizes $\langle \cdot, \boldsymbol{\Phi}^T \widehat{\boldsymbol{\mu}}_{\boldsymbol{\lambda}(\boldsymbol{c}'_t),\pi_t} - \boldsymbol{\lambda}(\boldsymbol{c}'_t)\rangle$ and that $\boldsymbol{\zeta}^{\pi_t}_{\boldsymbol{w}_t} \in \mathbb{B}_d(D_\zeta)$. Hence, the term (d) can be bounded by

$$\langle \boldsymbol{\zeta}_t - \boldsymbol{\zeta}^{\pi_t}, \boldsymbol{\Phi}^T \widehat{\boldsymbol{\mu}}_{\boldsymbol{\lambda}(\boldsymbol{c}'_t),\pi_t} - \boldsymbol{\lambda}(\boldsymbol{c}'_t)\rangle \leq 0.$$

Combining all the bounds, and using $D_\zeta := \sqrt{d} + \frac{\gamma D_\psi \sqrt{d}}{1-\gamma} \leq \mathcal{O}(\frac{\sqrt{d}}{1-\gamma})$ and $D_\pi = \alpha T D_\zeta \leq \mathcal{O}(\sqrt{\log|\mathcal{A}|T})$, we get

$$\textsc{Reg}^\zeta_t \leq \mathcal{O}\left(\frac{C^* d}{1-\gamma}\sqrt{\frac{\log(dnT(\log|\mathcal{A}|)/\delta)}{n}}\right).$$

□

B.4.3 Bounding the Regret of λ-Player

Lemma 51. *The sequences* $\{\pi_t\}, \{\boldsymbol{\lambda}(\boldsymbol{c}_t)\}, \{\boldsymbol{\zeta}_t\}$ *produced by Algorithm 5 satisfies*

$$\frac{1}{T}\sum_{t=1}^{T} \textsc{Reg}^\lambda_t \leq \mathcal{O}\left(\frac{C^* d^{3/2}}{1-\gamma}\sqrt{\frac{\log(dnT(\log|\mathcal{A}|)/\delta)}{n}}\right) + \epsilon^\lambda_{opt}(T)$$

with probability at least $1-\delta$.

Proof. The regret of λ-player at step t can be bounded by

$$\begin{aligned}
\textsc{Reg}^\lambda_t &= f(\boldsymbol{\zeta}_t, \boldsymbol{\lambda}^*, \pi_t) - f(\boldsymbol{\zeta}_t, \boldsymbol{\lambda}(\boldsymbol{c}_t), \pi_t) \\
&= \langle \boldsymbol{\lambda}^* - \boldsymbol{\lambda}(\boldsymbol{c}_t), \boldsymbol{\theta} + \gamma \boldsymbol{\Psi}\boldsymbol{v}_{\boldsymbol{\zeta}_t,\pi_t} - \boldsymbol{\zeta}_t\rangle \\
&= \langle \widehat{\boldsymbol{\lambda}}^* - \boldsymbol{\lambda}(\boldsymbol{c}_t), \boldsymbol{\theta} + \gamma \widehat{\boldsymbol{\Psi}\boldsymbol{v}}_{\boldsymbol{\zeta}_t,\pi_t} - \boldsymbol{\zeta}_t\rangle + \gamma\langle \widehat{\boldsymbol{\lambda}}^* - \boldsymbol{\lambda}(\boldsymbol{c}_t), \boldsymbol{\Psi}\boldsymbol{v}_{\boldsymbol{\zeta}_t,\pi_t} - \widehat{\boldsymbol{\Psi}\boldsymbol{v}}_{\boldsymbol{\zeta}_t,\pi_t}\rangle \\
&\qquad + \langle \boldsymbol{\lambda}^* - \widehat{\boldsymbol{\lambda}}^*, \boldsymbol{\theta} + \gamma \boldsymbol{\Psi}\boldsymbol{v}_{\boldsymbol{\zeta}_t,\pi_t} - \boldsymbol{\zeta}_t\rangle.
\end{aligned}$$

The average of the first term over $t = 1, \ldots, T$ is $\epsilon^\lambda_{\text{opt}}(T)$ which vanishes as T increases since the λ-player employs a no-regret online convex optimization oracle (Definition 2) on the sequence of functions $\langle \cdot, \boldsymbol{\theta} + \gamma \widehat{\boldsymbol{\Psi v}}_{\boldsymbol{\zeta}_t,\pi_t} - \boldsymbol{\zeta}_t \rangle$. The second term can be bounded as follows.

$$\begin{aligned}
\langle \widehat{\boldsymbol{\lambda}}^* - \boldsymbol{\lambda}(\boldsymbol{c}_t), \boldsymbol{\Psi v}_{\boldsymbol{\zeta}_t,\pi_t} - \widehat{\boldsymbol{\Psi v}}_{\boldsymbol{\zeta}_t,\pi_t} \rangle &\leq \|\widehat{\boldsymbol{\lambda}}^* - \boldsymbol{\lambda}(\boldsymbol{c}_t)\|_{(n\widehat{\boldsymbol{\Lambda}}_n + \boldsymbol{I})^{-1}} \|\boldsymbol{\Psi v}_{\boldsymbol{\zeta}_t,\pi_t} - \widehat{\boldsymbol{\Psi v}}_{\boldsymbol{\zeta}_t,\pi_t}\|_{n\widehat{\boldsymbol{\Lambda}}_n + \boldsymbol{I}} \\
&\leq \frac{1}{\sqrt{n}} \|\widehat{\boldsymbol{\lambda}}^* - \boldsymbol{\lambda}(\boldsymbol{c}_t)\|_{\widehat{\boldsymbol{\Lambda}}_n^\dagger} \|\boldsymbol{\Psi v}_{\boldsymbol{\zeta}_t,\pi_t} - \widehat{\boldsymbol{\Psi v}}_{\boldsymbol{\zeta}_t,\pi_t}\|_{n\widehat{\boldsymbol{\Lambda}}_n + \boldsymbol{I}} \\
&\leq \frac{2}{\sqrt{n}} C^* \sqrt{d} \cdot \mathcal{O}(D_v \sqrt{d \log(D_\zeta D_\pi n/\delta)}) \\
&\leq \mathcal{O}\left(\frac{C^* d^{3/2}}{1-\gamma} \sqrt{\frac{\log(dnT(\log|\mathcal{A}|)/\delta)}{n}} \right)
\end{aligned}$$

where the second inequality follows since $n\widehat{\boldsymbol{\Lambda}}_n + \boldsymbol{I} \succcurlyeq n\widehat{\boldsymbol{\Lambda}}_n$ and the fact that both $\widehat{\boldsymbol{\lambda}}^*$ and $\boldsymbol{\lambda}(\boldsymbol{c}_t)$ are in the column space of $\widehat{\boldsymbol{\Lambda}}$; the third inequality follows by Lemma 15 and Lemma 14 and the fact that the range of $\boldsymbol{v}_{\boldsymbol{\zeta}_t,\pi_t}$ is $[0, D_\zeta]$ so that we can set $D_v = D_\zeta$; the last inequality follows by $D_\zeta \leq \mathcal{O}(\frac{\sqrt{d}}{1-\gamma})$ and $D_\pi = \alpha T D_\zeta = \mathcal{O}(\sqrt{T \log|\mathcal{A}|})$. The third term can be bounded by

$$\begin{aligned}
\langle \boldsymbol{\lambda}^* - \widehat{\boldsymbol{\lambda}}^*, \boldsymbol{\theta} + \gamma \boldsymbol{\Psi v}_{\boldsymbol{\zeta}_t,\pi_t} - \boldsymbol{\zeta}_t \rangle &\leq \|\boldsymbol{\lambda}^* - \widehat{\boldsymbol{\lambda}}^*\|_2 \|\boldsymbol{\theta} + \gamma \boldsymbol{\Psi v}_{\boldsymbol{\zeta}_t,\pi_t} - \boldsymbol{\zeta}_t\|_2 \\
&\leq \mathcal{O}\left(C^* \sqrt{\frac{\log(d/\delta)}{n}} \right) \cdot \mathcal{O}(D_\zeta \sqrt{d}) \\
&\leq \mathcal{O}\left(\frac{C^* d}{1-\gamma} \sqrt{\frac{\log(d/\delta)}{n}} \right)
\end{aligned}$$

where the second inequality follows by Lemma 13 and the last inequality follows by the bound $D_\zeta \leq \mathcal{O}(\frac{\sqrt{d}}{1-\gamma})$ and the boundedness assumption on $\boldsymbol{\Psi}$. Combining the three bounds completes the proof. □

B.5 Details in Offline Constrained RL Setting

B.5.1 Lagrangian Formulation

Recall that in the linear CMDP setting, the optimization problem of interest is

$$\begin{aligned}
\max_{\pi} \quad & J_0(\pi) \\
\text{subject to} \quad & J_i(\pi) \geq \tau_i, \quad i = 1, \ldots, I.
\end{aligned}$$

which we denote by $\mathcal{P}(\boldsymbol{\tau})$ parameterized by the thresholds $\boldsymbol{\tau} \in \mathbb{R}^I$. We write the Lagrangian function corresponding to the optimization problem $\mathcal{P}(\boldsymbol{\tau})$ as

$$L(\pi, \boldsymbol{w}) := J(\pi) + \boldsymbol{w} \cdot (\boldsymbol{J}(\pi) - \boldsymbol{\tau})$$

where $\boldsymbol{J}(\cdot) = (J_1(\cdot), \ldots, J_I(\cdot))$, $\boldsymbol{\tau} = (\tau_1, \ldots, \tau_I)$ and $\boldsymbol{w} \in \mathbb{R}^I$ is the Lagrangian multipliers corresponding to the constraints. The linear programming formulation of the constrained reinforcement learning problem is:

$$\begin{aligned} \max_{\boldsymbol{\mu} \geq \boldsymbol{0}} \quad & \langle \boldsymbol{r}_0, \boldsymbol{\mu} \rangle \\ \text{subject to} \quad & \langle \boldsymbol{r}_i, \boldsymbol{\mu} \rangle \geq \tau_i, \quad i = 1, \ldots, I, \\ & \boldsymbol{E}^T \boldsymbol{\mu} = (1-\gamma)\boldsymbol{\nu}_0 + \gamma \boldsymbol{P}^T \boldsymbol{\mu}. \end{aligned}$$

Using $\boldsymbol{r}_i = \boldsymbol{\Phi}\boldsymbol{\theta}_i$, $i = 0, \ldots, I$, and $\boldsymbol{P} = \boldsymbol{\Phi}\boldsymbol{\Psi}$, which holds by the linear CMDP assumption (Assumption 5), the linear program can be written as

$$\begin{aligned} \max_{\boldsymbol{\mu} \geq 0} \quad & \langle \boldsymbol{\theta}_0, \boldsymbol{\Phi}^T \boldsymbol{\mu} \rangle \\ \text{subject to} \quad & \langle \boldsymbol{\theta}_i, \boldsymbol{\Phi}^T \boldsymbol{\mu} \rangle \geq \tau_i, \quad i = 1, \ldots, I, \\ & \boldsymbol{E}^T \boldsymbol{\mu} = (1-\gamma)\boldsymbol{\nu}_0 + \gamma \boldsymbol{\Psi}^T \boldsymbol{\Phi}^T \boldsymbol{\mu} \end{aligned}$$

Note that the optimization variable $\boldsymbol{\mu} \in \mathbb{R}^{|\mathcal{S} \times \mathcal{A}|}$ is high-dimensional that depends on the size of $\mathcal{S}$. With the goal of computational and statistical efficiency, we introduce a low-dimensional optimization variable $\boldsymbol{\lambda} = \boldsymbol{\Phi}^T \boldsymbol{\mu} \in \mathbb{R}^d$, which has the interpretation of the average occupancy in the feature space. With the reparametrization, the optimization problem becomes

$$\begin{aligned} \max_{\boldsymbol{\mu} \geq \boldsymbol{0}, \boldsymbol{\lambda}} \quad & \langle \boldsymbol{\theta}_0, \boldsymbol{\lambda} \rangle \\ \text{subject to} \quad & \langle \boldsymbol{\theta}_i, \boldsymbol{\lambda} \rangle \geq \tau_i, \quad i = 1, \ldots, I, \\ & \boldsymbol{E}^T \boldsymbol{\mu} = (1-\gamma)\boldsymbol{\nu}_0 + \gamma \boldsymbol{\Psi}^T \boldsymbol{\lambda} \\ & \boldsymbol{\lambda} = \boldsymbol{\Phi}^T \boldsymbol{\mu}. \end{aligned}$$

The dual of the linear program above is

$$\begin{aligned} \min_{\boldsymbol{w} \geq \boldsymbol{0}, \boldsymbol{v}, \boldsymbol{\zeta}} \quad & (1-\gamma)\langle \boldsymbol{\nu}_0, \boldsymbol{v} \rangle - \langle \boldsymbol{w}, \boldsymbol{\tau} \rangle \\ \text{subject to} \quad & \boldsymbol{\zeta} = \boldsymbol{\theta}_0 + \boldsymbol{\Theta}\boldsymbol{w} + \gamma \boldsymbol{\Psi} \boldsymbol{v} \\ & \boldsymbol{E}\boldsymbol{v} \geq \boldsymbol{\Phi}\boldsymbol{\zeta}. \end{aligned}$$

where we write $\mathbf{\Theta} = \begin{bmatrix} \boldsymbol{\theta}_1 & \cdots & \boldsymbol{\theta}_I \end{bmatrix} \in \mathbb{R}^{d \times I}$. The Lagrangian associated to this pair of linear programs is

$$\begin{aligned} L(\boldsymbol{\lambda}, \boldsymbol{\mu}; \boldsymbol{v}, \boldsymbol{w}, \boldsymbol{\zeta}) &= (1-\gamma)\langle \boldsymbol{\nu}_0, \boldsymbol{v} \rangle + \langle \boldsymbol{\lambda}, \boldsymbol{\theta}_0 + \gamma \mathbf{\Psi} \boldsymbol{v} - \boldsymbol{\zeta} \rangle + \langle \boldsymbol{\mu}, \mathbf{\Phi} \boldsymbol{\zeta} - \boldsymbol{E} \boldsymbol{v} \rangle - \langle \boldsymbol{w}, \boldsymbol{\tau} - \mathbf{\Theta}^T \boldsymbol{\lambda} \rangle \\ &= \langle \boldsymbol{\lambda}, \boldsymbol{\theta}_0 \rangle + \langle \boldsymbol{v}, (1-\gamma)\boldsymbol{\nu}_0 + \gamma \mathbf{\Psi}^T \boldsymbol{\lambda} - \boldsymbol{E}^T \boldsymbol{\mu} \rangle + \langle \boldsymbol{\zeta}, \mathbf{\Phi}^T \boldsymbol{\mu} - \boldsymbol{\lambda} \rangle - \langle \boldsymbol{w}, \boldsymbol{\tau} - \mathbf{\Theta}^T \boldsymbol{\lambda} \rangle. \end{aligned}$$

Note that the optimization variables $\boldsymbol{\lambda}, \boldsymbol{\zeta} \in \mathbb{R}^d$ and $\boldsymbol{w} \in \mathbb{R}^I$ are low-dimensional, but $\boldsymbol{\mu} \in \mathbb{R}^{|\mathcal{S} \times \mathcal{A}|}$ and $\boldsymbol{v} \in \mathbb{R}^{|\mathcal{S}|}$ are not. With the goal of running a primal-dual algorithm on the Lagrangian using only low-dimensional variables, we introduce policy variable π and parameterize $\boldsymbol{\mu}$ and $\boldsymbol{v}$, as was done for the unconstrained RL setting, by

$$\begin{aligned} \mu_{\boldsymbol{\lambda},\pi}(s,a) &= \pi(a|s)\left[(1-\gamma)\nu_0(s) + \gamma\langle \psi(s), \boldsymbol{\lambda} \rangle\right] \\ v_{\boldsymbol{\zeta},\pi}(s) &= \sum_a \pi(a|s) \langle \boldsymbol{\zeta}, \boldsymbol{\varphi}(s,a) \rangle. \end{aligned}$$

Note that the choice of $\boldsymbol{\mu}_{\boldsymbol{\lambda},\pi}$ makes the Bellman flow constraint $\boldsymbol{E}^T \boldsymbol{\mu} = (1-\gamma)\boldsymbol{\nu}_0 + \gamma \mathbf{\Psi}^T \boldsymbol{\lambda}$ of the primal problem satisfied. Also, the choice of $\boldsymbol{v}_{\boldsymbol{\zeta},\pi}$ makes $\langle \boldsymbol{\mu}_{\boldsymbol{\lambda},\pi}, \mathbf{\Phi}\boldsymbol{\zeta} - \boldsymbol{E}\boldsymbol{v}_{\boldsymbol{\zeta},\pi} \rangle = 0$. Using the above parameterization, the Lagrangian can be rewritten in terms of $\boldsymbol{\zeta}, \boldsymbol{\lambda}, \boldsymbol{w}, \pi$ as follows:

$$\begin{aligned} g(\boldsymbol{\lambda}, \boldsymbol{\zeta}, \boldsymbol{w}, \pi) &= \langle \boldsymbol{\lambda}, \boldsymbol{\theta}_0 \rangle + \langle \boldsymbol{\zeta}, \mathbf{\Phi}^T \boldsymbol{\mu}_{\boldsymbol{\lambda},\pi} - \boldsymbol{\lambda} \rangle - \langle \boldsymbol{w}, \boldsymbol{\tau} - \mathbf{\Theta}^T \boldsymbol{\lambda} \rangle \quad \text{(B.4)} \\ &= (1-\gamma)\langle \boldsymbol{\nu}_0, \boldsymbol{v}_{\boldsymbol{\zeta},\pi} \rangle + \langle \boldsymbol{\lambda}, \boldsymbol{\theta}_0 + \gamma \mathbf{\Psi} \boldsymbol{v}_{\boldsymbol{\zeta},\pi} - \boldsymbol{\zeta} \rangle - \langle \boldsymbol{w}, \boldsymbol{\tau} - \mathbf{\Theta}^T \boldsymbol{\lambda} \rangle. \quad \text{(B.5)} \end{aligned}$$

At the cost of having to keep track of π, we can now run a primal-dual algorithm on the low-dimensional variables $\boldsymbol{\zeta}$, $\boldsymbol{\lambda}$ and $\boldsymbol{w}$. As is the case for the unconstrained RL setting, the introduction of π in the equation does not make the algorithm inefficient because we can only keep track of the distribution $\pi(s|a)$ for state-action pairs that appear in the dataset.

B.5.2 Technical Lemmas on Lagrangian

For a linearized reward function $u = r_0 + \boldsymbol{w} \cdot \boldsymbol{r}$ where we use the notation $\boldsymbol{r}$ to denote the vector of reward functions $r_1, \ldots, r_I$ such that $u(s,a) = r_0(s,a) + \sum_{i=1}^{I} w_i r_i(s,a)$, the Bellman equation becomes

$$\boldsymbol{Q}_{\boldsymbol{w}}^{\pi} = \mathbf{\Phi}(\boldsymbol{\theta}_0 + \mathbf{\Theta}\boldsymbol{w} + \gamma \mathbf{\Psi} \boldsymbol{V}_{\boldsymbol{w}}^{\pi}) = \mathbf{\Phi}\boldsymbol{\zeta}_{\boldsymbol{w}}^{\pi} \quad \text{(B.6)}$$

where we write $\boldsymbol{Q}_{\boldsymbol{w}}^{\pi}$ and $\boldsymbol{V}_{\boldsymbol{w}}^{\pi}$ as the value functions of the policy π with respect to the linearized reward function $r_0 + \boldsymbol{w} \cdot \boldsymbol{r}$ and define

$$\boldsymbol{\zeta}_{\boldsymbol{w}}^{\pi} := \boldsymbol{\theta}_0 + \boldsymbol{\Theta}\boldsymbol{w} + \gamma \boldsymbol{\Psi} \boldsymbol{V}_{\boldsymbol{w}}^{\pi}.$$

Note that if $\boldsymbol{w} \in D_w \boldsymbol{\Delta}^I$, we have $\|\boldsymbol{\zeta}_{\boldsymbol{w}}^{\pi}\|_2 \leq 1 + D_w + \frac{\gamma\sqrt{d}(1+D_w)}{1-\gamma} = \mathcal{O}\left(\frac{D_w\sqrt{d}}{1-\gamma}\right)$. We define $D_\zeta := 1 + D_w + \frac{\gamma\sqrt{d}(1+D_w)}{1-\gamma}$.

Lemma 52. *Let $\boldsymbol{\zeta}_{\boldsymbol{w}}^{\pi}$ be the parameter that satisfies $\boldsymbol{Q}_{\boldsymbol{w}}^{\pi} = \boldsymbol{\Phi}\boldsymbol{\zeta}_{\boldsymbol{w}}^{\pi}$ for a given $\boldsymbol{w} \in \mathbb{R}^I$ and a policy π. Then,*

$$L(\pi, \boldsymbol{w}) = g(\boldsymbol{\zeta}_{\boldsymbol{w}}^{\pi}, \boldsymbol{\lambda}, \pi, \boldsymbol{w})$$

for all $\boldsymbol{\lambda} \in \mathbb{R}^d$ in the span of $\{\boldsymbol{\varphi}(s,a)\}_{(s,a)\in\mathcal{S}\times\mathcal{A}}$.

Proof. For convenience, define the reward function $u(s,a) = r_0(s,a) + \sum_{i=1}^{I} w_i r_i(s,a)$. By the linear CMDP assumption, we have $\boldsymbol{u} = \boldsymbol{\Phi}(\boldsymbol{\theta}_0 + \boldsymbol{\Theta}\boldsymbol{w})$ where $\boldsymbol{u} \in \mathbb{R}^{|\mathcal{S}\times\mathcal{A}|}$ is the vector representation of the reward function u. Also, by the definition of $v_{\boldsymbol{\zeta},\pi}$ in (3.6), we have

$$v_{\boldsymbol{\zeta}_{\boldsymbol{w}}^{\pi},\pi}(s) = \sum_a \pi(a|s)\langle \boldsymbol{\zeta}_{\boldsymbol{w}}^{\pi}, \boldsymbol{\varphi}(s,a)\rangle = \sum_a \pi(a|s) Q_{\boldsymbol{w}}^{\pi}(s,a) = V_{\boldsymbol{w}}^{\pi}(s).$$

Since we assume $\boldsymbol{\lambda} \in \mathbb{R}^d$ is in the span of $\{\boldsymbol{\varphi}(s,a)\}_{(s,a)\in\mathcal{S}\times\mathcal{A}}$, there exists $\boldsymbol{\alpha} \in \mathbb{R}^{|\mathcal{S}\times\mathcal{A}|}$ such that $\boldsymbol{\lambda} = \boldsymbol{\Phi}^T\boldsymbol{\alpha}$. Hence, using the form (B.5) of the Lagrangian function, we have

$$\begin{aligned}
g(\boldsymbol{\zeta}_{\boldsymbol{w}}^{\pi}, \boldsymbol{\lambda}, \pi, \boldsymbol{w}) &= (1-\gamma)\langle \boldsymbol{\nu}_0, \boldsymbol{v}_{\boldsymbol{\zeta}_{\boldsymbol{w}}^{\pi}}\rangle + \langle \boldsymbol{\lambda}, \boldsymbol{\theta}_0 + \gamma\boldsymbol{\Psi}\boldsymbol{v}_{\boldsymbol{\zeta}_{\boldsymbol{w}}^{\pi}} - \boldsymbol{\zeta}_{\boldsymbol{w}}^{\pi}\rangle - \langle \boldsymbol{w}, \boldsymbol{\tau} - \boldsymbol{\Theta}^T\boldsymbol{\lambda}\rangle \\
&= (1-\gamma)\langle \boldsymbol{\nu}_0, \boldsymbol{V}_{\boldsymbol{w}}^{\pi}\rangle + \langle \boldsymbol{\lambda}, \boldsymbol{\theta}_0 + \boldsymbol{\Theta}\boldsymbol{w} + \gamma\boldsymbol{\Psi}\boldsymbol{V}_{\boldsymbol{w}}^{\pi} - \boldsymbol{\zeta}_{\boldsymbol{w}}^{\pi}\rangle - \langle \boldsymbol{w}, \boldsymbol{\tau}\rangle \\
&= (1-\gamma)\langle \boldsymbol{\nu}_0, \boldsymbol{V}_{\boldsymbol{w}}^{\pi}\rangle + \langle \boldsymbol{\alpha}, \boldsymbol{\Phi}(\boldsymbol{\theta}_0 + \boldsymbol{\Theta}\boldsymbol{w} + \gamma\boldsymbol{\Psi}\boldsymbol{V}_{\boldsymbol{w}}^{\pi} - \boldsymbol{\zeta}_{\boldsymbol{w}}^{\pi})\rangle - \langle \boldsymbol{w}, \boldsymbol{\tau}\rangle \\
&= (1-\gamma)\langle \boldsymbol{\nu}_0, \boldsymbol{V}_{\boldsymbol{w}}^{\pi}\rangle + \langle \boldsymbol{\alpha}, \boldsymbol{u} + \gamma\boldsymbol{P}\boldsymbol{V}_{\boldsymbol{w}}^{\pi} - \boldsymbol{Q}_{\boldsymbol{w}}^{\pi}\rangle - \langle \boldsymbol{w}, \boldsymbol{\tau}\rangle \\
&= (1-\gamma)\langle \boldsymbol{\nu}_0, \boldsymbol{V}_{\boldsymbol{w}}^{\pi}\rangle - \langle \boldsymbol{w}, \boldsymbol{\tau}\rangle \\
&= L(\pi, \boldsymbol{w})
\end{aligned}$$

where the second to last equality uses the Bellman equation (3.9) and the last equality is by $L(\pi, \boldsymbol{w}) = J_0(\pi) + \boldsymbol{w} \cdot (\boldsymbol{J}(\pi) - \boldsymbol{\tau})$ and the fact that $J_0(\pi) + \boldsymbol{w} \cdot \boldsymbol{J}(\pi)$ is the value of π with respect to the linearized value function $r_0 + \boldsymbol{w} \cdot \boldsymbol{r}$. □

Lemma 53. *Under the linear MDP setting, let $\boldsymbol{\zeta}^{\pi}$ be the parameter that satisfies $\boldsymbol{Q}^{\pi} = \boldsymbol{\Phi}\boldsymbol{\zeta}^{\pi}$ for a policy π. Then,*

$$J(\pi) = f(\boldsymbol{\zeta}^{\pi}, \boldsymbol{\lambda}, \pi)$$

for all $\boldsymbol{\lambda} \in \mathbb{R}^d$ in the span of $\{\boldsymbol{\varphi}(s,a)\}_{(s,a)\in\mathcal{S}\times\mathcal{A}}$.

Proof. This is a direct corollary of Lemma 52, which can be seen by setting $\boldsymbol{w} = \boldsymbol{0}$. □

Lemma 54. *Given a policy π, let μ^π be the occupancy measure induced by π and let $\boldsymbol{\lambda}^\pi = \boldsymbol{\Phi}^T \boldsymbol{\mu}^\pi$. Then, for any $\boldsymbol{\zeta} \in \mathbb{R}^d$ and any $\boldsymbol{w} \in \mathbb{R}^I$, we have*

$$L(\pi, \boldsymbol{w}) = g(\boldsymbol{\zeta}, \boldsymbol{\lambda}^\pi, \boldsymbol{w}, \pi).$$

Proof. By the definition of $\mu_{\boldsymbol{\lambda},\pi}$ in (3.5), we have

$$\begin{aligned} \mu_{\boldsymbol{\lambda}^\pi,\pi}(s,a) &= \pi(a|s)\left[(1-\gamma)\nu_0(s) + \gamma\langle \boldsymbol{\psi}(s), \boldsymbol{\lambda}^\pi\rangle\right] \\ &= \pi(a|s)\left[(1-\gamma)\nu_0(s) + \gamma\langle \boldsymbol{\psi}(s), \boldsymbol{\Phi}^T\boldsymbol{\mu}^\pi\rangle\right] \\ &= \mu^\pi(s,a). \end{aligned}$$

Using the form (B.4) of the Lagrangian function, we have

$$\begin{aligned} g(\boldsymbol{\zeta}, \boldsymbol{\lambda}^\pi, \boldsymbol{w}, \pi) &= \langle \boldsymbol{\lambda}^\pi, \boldsymbol{\theta}_0\rangle + \langle \boldsymbol{\zeta}, \boldsymbol{\Phi}^T\boldsymbol{\mu}_{\boldsymbol{\lambda}^\pi,\pi} - \boldsymbol{\lambda}^\pi\rangle - \langle \boldsymbol{w}, \boldsymbol{\tau} - \boldsymbol{\Theta}^T\boldsymbol{\lambda}^\pi\rangle \\ &= \langle \boldsymbol{\Phi}^T\boldsymbol{\mu}^\pi, \boldsymbol{\theta}_0\rangle - \langle \boldsymbol{w}, \boldsymbol{\tau} - \boldsymbol{\Theta}^T\boldsymbol{\Phi}^T\boldsymbol{\mu}^\pi\rangle \\ &= \langle \boldsymbol{\mu}^\pi, \boldsymbol{r}_0\rangle - \langle \boldsymbol{w}, \boldsymbol{\tau} - \boldsymbol{R}^T\boldsymbol{\mu}^\pi\rangle \\ &= L(\pi, \boldsymbol{w}) \end{aligned}$$

where the second equality uses $\boldsymbol{\mu}_{\boldsymbol{\lambda}^\pi,\pi} = \boldsymbol{\mu}^\pi$ and $\boldsymbol{\lambda}^\pi = \boldsymbol{\Phi}^T\boldsymbol{\mu}^\pi$; and the third equality uses the matrix notation for the reward functions $\boldsymbol{R} = \{r_i(s,a)\}_{(s,a)\in\mathbb{R}^{|\mathcal{S}\times\mathcal{A}|}, i\in[I]}$; the last equality uses $J_i(\pi) = \langle \boldsymbol{\mu}^\pi, \boldsymbol{r}_i\rangle$, $i = 1, \ldots, I$. □

Lemma 55. *Under the linear MDP setting, let μ^π be the occupancy measure induced by a policy π and let $\boldsymbol{\lambda}^\pi = \boldsymbol{\Phi}^T\boldsymbol{\mu}^\pi$. Then, for any $\boldsymbol{\zeta} \in \mathbb{R}^d$, we have*

$$J(\pi) = f(\boldsymbol{\zeta}, \boldsymbol{\lambda}^\pi, \pi).$$

Proof. This is a direct corollary of Lemma 54, which can be seen by setting $\boldsymbol{w} = \boldsymbol{0}$. □

B.5.3 Proof of Theorem 5

For a given $\boldsymbol{w} \in \mathbb{R}^I$ and a policy π, define $\boldsymbol{\zeta}^\pi_{\boldsymbol{w}} \in \mathbb{R}^d$ to be the parameter that satisfies $\boldsymbol{Q}^\pi_{\boldsymbol{w}} = \boldsymbol{\Phi}\boldsymbol{\zeta}^\pi_{\boldsymbol{w}}$ where $\boldsymbol{Q}^\pi_{\boldsymbol{w}}$ is the state-action value function of the policy π with respect to the reward function $r_0 + \boldsymbol{w}\cdot\boldsymbol{r}$. Using $g(\boldsymbol{\zeta}^\pi_{\boldsymbol{w}}, \boldsymbol{\lambda}, \boldsymbol{w}, \pi) = L(\pi, \boldsymbol{w})$ for any $\boldsymbol{\lambda}$ that is a linear combination of $\{\boldsymbol{\varphi}(s,a)\}_{(s,a)\in\mathcal{S}\times\mathcal{A}}$ (Lemma B.4) and $g(\boldsymbol{\zeta}, \boldsymbol{\lambda}^\pi, \boldsymbol{w}, \pi) = L(\pi, \boldsymbol{w})$ for any $\boldsymbol{\zeta} \in \mathbb{R}^d$

where $\boldsymbol{\lambda}^\pi = \boldsymbol{\Phi}^T \boldsymbol{\mu}^\pi$ (Lemma B.5), we have

$$\begin{aligned}
L(\pi^*, \boldsymbol{w}_t) - L(\pi_t, \boldsymbol{w}) &= g(\boldsymbol{\zeta}_t, \boldsymbol{\lambda}^*, \boldsymbol{w}_t, \pi) - g(\boldsymbol{\zeta}^{\pi_t}_{\boldsymbol{w}_t}, \boldsymbol{\lambda}(\boldsymbol{c}'_t), \boldsymbol{w}, \pi_t)) \\
&= (\underbrace{g(\boldsymbol{\zeta}_t, \boldsymbol{\lambda}^*, \boldsymbol{w}_t, \pi^*) - g(\boldsymbol{\zeta}_t, \boldsymbol{\lambda}^*, \boldsymbol{w}_t, \pi_t)}_{\textsc{Reg}_t^\pi}) \\
&\qquad + (\underbrace{g(\boldsymbol{\zeta}_t, \boldsymbol{\lambda}^*, \boldsymbol{w}_t, \pi_t) - g(\boldsymbol{\zeta}_t, \boldsymbol{\lambda}(\boldsymbol{c}'_t), \boldsymbol{w}_t, \pi_t)}_{\textsc{Reg}_t^\lambda}) \\
&\qquad + (\underbrace{g(\boldsymbol{\zeta}_t, \boldsymbol{\lambda}(\boldsymbol{c}'_t), \boldsymbol{w}_t, \pi_t) - g(\boldsymbol{\zeta}_t, \boldsymbol{\lambda}(\boldsymbol{c}'_t), \boldsymbol{w}, \pi_t)}_{\textsc{Reg}_t^w}) \\
&\qquad + (\underbrace{g(\boldsymbol{\zeta}_t, \boldsymbol{\lambda}(\boldsymbol{c}'_t), \boldsymbol{w}, \pi_t) - g(\boldsymbol{\zeta}^{\pi_t}_{\boldsymbol{w}_t}, \boldsymbol{\lambda}(\boldsymbol{c}'_t), \boldsymbol{w}, \pi_t)}_{\textsc{Reg}_t^\zeta})
\end{aligned}$$

where π^* is an optimal policy for the optimization problem $\mathcal{P}(\boldsymbol{\tau})$ and we use the notation $\boldsymbol{\lambda}^* = \boldsymbol{\lambda}^{\pi^*}$. Note that the suboptimality $L(\pi^*, \boldsymbol{w}_t) - L(\pi_t, \boldsymbol{w})$ is decomposed into regret terms of the four players. As long as we show that the sum of the four regrets over $t = 1, \ldots, T$ are sublinear in T and the dataset size n, we obtain $\frac{1}{T}\sum_{t=1}^T L(\pi^*, \boldsymbol{w}_t) - L(\pi_t, \boldsymbol{w}) = L(\pi^*, \bar{\boldsymbol{w}}) - L(\bar{\pi}, \boldsymbol{w}) = o(1)$ where $\bar{\boldsymbol{w}} = \frac{1}{T}\sum_{t=1}^T \boldsymbol{w}_t$ and $\bar{\pi} = \mathrm{Unif}(\pi_1, \ldots, \pi_T)$ is the mixture policy that chooses a policy among $\pi_1, \ldots, \pi_T$ uniformly at random and runs the chosen policy for the entire trajectory. Then, for large enough T and n, we get $L(\pi^*, \bar{\boldsymbol{w}}) \leq L(\bar{\pi}, \boldsymbol{w}) + \epsilon$ where ϵ vanishes as T and n increase. Such a pair is a near saddle point of the Lagrangian function $L(\cdot, \cdot)$ and it can be shown that the mixture policy $\bar{\pi}$ is a near-optimal solution of the optimization problem (3.1). Specifically, we can show that if the Slater's condition (Assumption 3) holds, then a near saddle point $(\bar{\pi}, \bar{\boldsymbol{w}})$ of $L(\cdot, \cdot)$ with $L(\pi, \bar{\boldsymbol{w}}) \leq L(\bar{\pi}, \boldsymbol{w}) + \mathcal{O}(\epsilon)$ for all policies π and $\boldsymbol{w} \in (1 + \frac{1}{\phi})\boldsymbol{\Delta}^I$ satisfies

$$\begin{aligned}
&J_0(\bar{\pi}) \geq J_0(\pi^*) - \epsilon \\
&J_i(\bar{\pi}) \geq \tau_i - \epsilon, \quad i = 1, \ldots, I
\end{aligned}$$

where π^* is the optimal policy for $\mathcal{P}(\boldsymbol{\tau})$, implying that $\bar{\pi}$ is a nearly optimal solution for the optimization problem.

In the rest of the section, we sketch the analysis that shows that $L(\pi, \bar{\boldsymbol{w}}) - L(\bar{\pi}, \boldsymbol{w}) \leq \epsilon$ for large enough T and $n = \mathcal{O}(\epsilon^{-2})$. With the decomposition of $L(\pi, \boldsymbol{w}_t) - L(\pi_t, \boldsymbol{w})$ into regrets of the four players discussed previously, we study how the four regrets can be bounded in the next four subsections.

B.5.3.1 Bounding Regret of π-Player

Using the expression (B.5), the regret of π-player simplifies to

$$\begin{aligned}\mathrm{Reg}_t^{\pi} &= g(\boldsymbol{\zeta}_t, \boldsymbol{\lambda}^*, \boldsymbol{w}_t, \pi^*) - g(\boldsymbol{\zeta}_t, \boldsymbol{\lambda}^*, \boldsymbol{w}_t, \pi_t)\\ &= \langle \boldsymbol{\nu}^*, \boldsymbol{v}_{\zeta_t,\pi} - \boldsymbol{v}_{\zeta_t,\pi_t}\rangle\\ &= \langle \boldsymbol{\nu}^*, \textstyle\sum_a (\pi(a|\cdot) - \pi_t(a|\cdot))\langle \boldsymbol{\zeta}_t, \boldsymbol{\varphi}(\cdot, a)\rangle\rangle.\end{aligned}$$

where $\boldsymbol{\nu}^* = (1-\gamma)\boldsymbol{\nu}_0 + \gamma\boldsymbol{\Psi}^T\boldsymbol{\lambda}^*$ is the state occupancy measure induced by π^*. This is identical to the regret term for the π-player in the unconstrained RL setting. As is done in the unconstrained RL setting, choosing $\alpha = \mathcal{O}((1-\gamma)\sqrt{\log|\mathcal{A}|/(dT)})$, we get

$$\frac{1}{T}\sum_{t=1}^{T} \mathrm{REG}_t^{\pi} \le \mathcal{O}\left(\frac{1}{1-\gamma}\sqrt{(d\log|\mathcal{A}|)/T}\right)$$

which is sublinear in T. Consequently, choosing T to be at least $\Omega(\frac{d\log|\mathcal{A}|}{(1-\gamma)^2\epsilon^2})$ gives $\frac{1}{T}\sum_{t=1}^{T}\mathrm{Reg}_t^{\pi} \le \epsilon$.

B.5.3.2 Bounding Regret of ζ-Player

Note that the regret for the ζ-player simplifies to

$$\begin{aligned}\mathrm{REG}_t^{\zeta} &= g(\boldsymbol{\zeta}_t, \boldsymbol{\lambda}(\boldsymbol{c}_t'), \boldsymbol{w}, \pi_t) - g(\boldsymbol{\zeta}_{\boldsymbol{w}_t}^{\pi_t}, \boldsymbol{\lambda}(\boldsymbol{c}_t'), \boldsymbol{w}, \pi_t)\\ &= \langle \boldsymbol{\zeta}_t - \boldsymbol{\zeta}_{\boldsymbol{w}_t}^{\pi_t}, \boldsymbol{\Phi}^T\boldsymbol{\mu}_{\boldsymbol{\lambda}_t,\pi_t} - \boldsymbol{\lambda}_t\rangle\end{aligned}$$

which has the same form as in the unconstrained case. The proof is essentially the same as the proof in Section B.4.2 for the unconstrained setting. The only difference is that $D_\zeta \le \mathcal{O}(\frac{D_w\sqrt{d}}{1-\gamma})$ where $D_w = 1 + \frac{1}{\phi}$. Following the proof, we get

$$\mathrm{REG}_t^{\zeta} \le \mathcal{O}\left(\frac{C^* d}{(1-\gamma)\phi}\sqrt{\frac{\log(dnT(\log|\mathcal{A}|)/(\delta\phi))}{n}}\right).$$

B.5.3.3 Bounding Regret of λ-Player

Using the expression (B.5), the regret of λ-player simplifies to

$$\begin{aligned}\mathrm{REG}_t^{\lambda} &= f(\boldsymbol{\zeta}_t, \boldsymbol{\lambda}^*, \boldsymbol{w}_t, \pi_t) - f(\boldsymbol{\zeta}_t, \boldsymbol{\lambda}_t, \boldsymbol{w}_t, \pi_t)\\ &= \langle \boldsymbol{\lambda}^* - \boldsymbol{\lambda}_t, \underbrace{\boldsymbol{\theta}_0 + \gamma\boldsymbol{\Psi}\boldsymbol{v}_{\zeta_t,\pi_t} - \boldsymbol{\zeta}_t + \boldsymbol{\Theta}\boldsymbol{w}_t}_{=\boldsymbol{\xi}_t}\rangle\end{aligned}$$

Following the analysis for the unconstrained setting in Section B.4.3, we get

$$\frac{1}{T}\sum_{t=1}^{T}\textsc{Reg}_t^{\lambda} \leq \mathcal{O}\left(\frac{C^* d^{3/2}}{(1-\gamma)\phi}\sqrt{\frac{\log(dnT(\log|\mathcal{A}|)/(\delta\phi))}{n}}\right) + \epsilon_{\text{opt}}^{\lambda}(T)$$

B.5.3.4 Bounding Regret of w-Player

Using expression (B.4), the regret of w-player simplifies to

$$g(\boldsymbol{\zeta}_t, \boldsymbol{\lambda}(\boldsymbol{c}_t'), \boldsymbol{w}_t, \pi_t) - g(\boldsymbol{\zeta}_t, \boldsymbol{\lambda}(\boldsymbol{c}_t'), \boldsymbol{w}, \pi_t) = \langle \boldsymbol{w}_t - \boldsymbol{w}, \boldsymbol{\tau} - \boldsymbol{\Theta}^T \boldsymbol{\lambda}_t \rangle$$

which is bounded by 0 since the w-player choose $\boldsymbol{w}_t \in D_w\Delta^I$ that minimizes $\langle \cdot, \boldsymbol{\tau} - \boldsymbol{\Theta}^T \boldsymbol{\lambda}_t \rangle$.

APPENDIX C

Details of Offline Constrained RL with General Function Approximation

C.1 Analysis for Offline Unconstrained RL

C.1.1 Invariance of Saddle Points

Lemma 56 (Lemma 31 in Zhan et al. [2022])**.** *Suppose (x^*, y^*) is a saddle point of $f(x, y)$ over $\mathcal{X} \times \mathcal{Y}$, then for any $\mathcal{X}' \subseteq \mathcal{X}$ and $\mathcal{Y}' \subseteq \mathcal{Y}$, if $(x^*, y^*) \in \mathcal{X}' \times \mathcal{Y}'$, we have:*

$$
\begin{aligned}
(x^*, y^*) &\in \operatorname*{argmin}_{x \in \mathcal{X}'} \operatorname*{argmax}_{y \in \mathcal{Y}'} f(x, y), \\
(x^*, y^*) &\in \operatorname*{argmax}_{y \in \mathcal{Y}'} \operatorname*{argmin}_{x \in \mathcal{X}'} f(x, y).
\end{aligned}
$$

C.1.2 Proof of Proposition 1

Recall the linear programming formulation of MDP $\mathcal{M} = (\mathcal{S}, \mathcal{A}, r, \gamma, d_0)$:

$$
\begin{aligned}
\max_{\boldsymbol{\mu} \geq 0} \quad & \langle \boldsymbol{\mu}, \boldsymbol{r} \rangle \\
\text{subject to} \quad & \boldsymbol{E}^\top \boldsymbol{\mu} = (1 - \gamma) \boldsymbol{d}_0 + \gamma \boldsymbol{P}^\top \boldsymbol{\mu}
\end{aligned}
$$

where $\boldsymbol{\mu} \in \mathbb{R}_+^{|\mathcal{S} \times \mathcal{A}|}$. Consider the associated Lagrangian function

$$
L(\boldsymbol{\mu}; \boldsymbol{V}) = \langle \boldsymbol{\mu}, \boldsymbol{r} \rangle + \langle \boldsymbol{V}, (1 - \gamma) \boldsymbol{d}_0 + \gamma \boldsymbol{P}^\top \boldsymbol{\mu} - \boldsymbol{E}^\top \boldsymbol{\mu} \rangle
$$

where $\boldsymbol{V} \in \mathbb{R}^{|\mathcal{S}|}$ is the Lagrangian multiplier. Let $(\boldsymbol{\mu}^*, \boldsymbol{V}^*)$ be a saddle point of L over $\mathbb{R}_+^{|\mathcal{S} \times \mathcal{A}|} \times \mathbb{R}^{|\mathcal{S}|}$ such that the policy $\pi(\mu^*)$ extracted from μ^* is optimal. Consider having access to function classes $\mathcal{U} \subseteq \mathbb{R}_+^{|\mathcal{S} \times \mathcal{A}|}$ and $\mathcal{V} \subseteq \mathbb{R}^{|\mathcal{S}|}$ that capture the saddle point so that $\boldsymbol{\mu}^* \in \mathcal{U}$ and $\boldsymbol{V}^* \in \mathcal{V}$. A natural procedure for solving the linear program with access to the function classes is to find a saddle point $(\widehat{\mu}, \widehat{V})$ of L over $\mathcal{U} \times \mathcal{V}$ and extract the policy $\pi(\widehat{\mu})$. However, we show that such a procedure does not necessarily give an optimal policy.

Proof of Proposition 1. We explicitly construct a MDP $(\mathcal{S}, \mathcal{A}, r, \gamma, d_0)$ and function classes $\mathcal{U}$, $\mathcal{V}$ as follows (see Figure C.1). Let $\mathcal{S} = \{s_0, l_1, r_1, l_2, r_2\}$ and $\mathcal{A} = \{L, R\}$. Let the reward function be defined as $r(l_1, a) = 1$, $r(l_2, a) = 2$ for all $a \in \mathcal{A}$, and all other entries of the reward function are 0. Let $\gamma = 1/2$ and $d_0(s_0) = 1$ such that the MDP starts from s_0 deterministically. The states l_1, l_2 and r_2 are absorbing, such that both actions L and R taken at these states do not change states, i.e., $P(s|s, L) = P(s|s, R) = 1$ for $s = l_1, l_2, r_2$. In other states s_0 and r_2, the action L has equal probability of transitioning to the states in the level below, i.e., $P(l_1|s_0, L) = P(r_1|s_0, L) = 1/2$ and $P(l_2|r_1, L) = P(r_2|r_1, L) = 1/2$. On the other hand, action R has probability 1/4 of transitioning to the left state and probability 3/4 of transitioning to the right state, i.e., $P(l_1|s_0, R) = P(l_2|r_1, R) = 1/4$ and $P(r_1|s_0, R) = P(r_2|r_1, R) = 3/4$.

It can be seen that the optimal state value function V^* has values $V^*(s_0) = 1, V^*(l_1) = 2, V^*(r_1) = 2, V^*(l_2) = 8, V^*(r_2) = 0$ and an optimal occupancy measure μ^* is the one induced by the policy that always chooses the action L: $\mu^*(s_0, L) = 1/2, \mu^*(l_1, L) = 1/4, \mu^*(r_1, L) = 1/8, \mu^*(l_2, L) = \mu^*(r_2, L) = 1/16$, and all other entries 0. By straightforward calculation, we get

$$L(\boldsymbol{\mu}, \boldsymbol{V}^*) = \frac{1}{2} V^*(s_0) - \frac{1}{4}\mu(r_1, R).$$

Note that $L(\boldsymbol{\mu}, \boldsymbol{V}^*)$ depends on μ only through the value $\mu(r_1, R)$. Hence, any $\boldsymbol{\mu}$ is a best response to $\boldsymbol{V}^*$ that maximizes $L(\boldsymbol{\mu}, \boldsymbol{V}^*)$ as long as $\mu(r_1, R) = 0$. Consider function classes $\mathcal{V} = \{V^*\}$ and $\mathcal{U} = \{\mu^*, \widetilde{\mu}\}$ where $\widetilde{\mu}$ is a function with $\widetilde{\mu}(s_0, R) = 1/2$ and $\widetilde{\mu}(l_1, L) = \widetilde{\mu}(l_1, R) = 1/4$, and the function values for other entries are 0. The policy $\widetilde{\pi}$ extracted from $\widetilde{\mu}$ chooses the action R at the initial state s_0. Since $\widetilde{\mu}(r_1, R) = 0$, it is a saddle point of L. However, $\widetilde{\mu}$ is not an optimal solution of the linear program, since it is not feasible. □

C.1.3 Proof of Lemma 16

Proof of Lemma 16. Recall the linear programming formulation

$$\begin{aligned} \max_{\boldsymbol{\mu} \geq 0} \quad & \langle \boldsymbol{\mu}, \boldsymbol{r} \rangle \\ \text{subject to} \quad & E^\top \boldsymbol{\mu} = (1-\gamma)\boldsymbol{d}_0 + \gamma \boldsymbol{P}^\top \boldsymbol{\mu} \end{aligned}$$

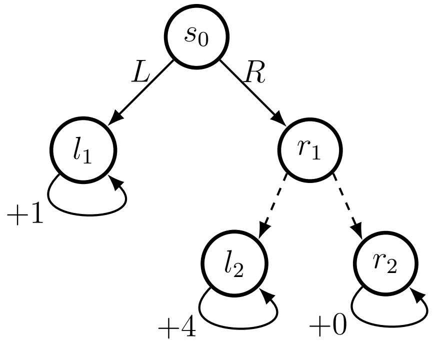

Figure C.1: An example MDP with spurious saddle points.

and its Lagrangian function

$$L(\boldsymbol{\mu}; \boldsymbol{V}) = \langle \boldsymbol{\mu}, \boldsymbol{r} \rangle + \langle \boldsymbol{V}, (1-\gamma)\boldsymbol{d}_0 + \gamma \boldsymbol{P}^\top \boldsymbol{\mu} - E^\top \boldsymbol{\mu} \rangle.$$

Let (μ^*, V^*) be a saddle point of the Lagrangian function L over $\mathbb{R}_+^{|\mathcal{S}\times\mathcal{A}|} \times \mathbb{R}^{|\mathcal{S}|}$. Then, by the saddle point theorem for convex optimization, $\boldsymbol{\mu}^*$ solves the linear program. In particular, $\boldsymbol{\mu}^*$ satisfies the Bellman flow constraint and the extracted policy $\pi^* = \pi(\boldsymbol{\mu}^*)$ induces $\boldsymbol{\mu}^*$. It follows that

$$L(\boldsymbol{\mu}^*, \boldsymbol{V}) = \langle \boldsymbol{\mu}^*, \boldsymbol{r} \rangle = (1-\gamma) J(\pi^*)$$

for all $\boldsymbol{V} \in \mathbb{R}^{|\mathcal{S}|}$. Also, rewriting the Lagrangian function as

$$L(\boldsymbol{\mu}; \boldsymbol{V}) = \langle \boldsymbol{\mu}, \boldsymbol{r} + \gamma \boldsymbol{P}\boldsymbol{V} - \boldsymbol{E}\boldsymbol{V} \rangle + \langle \boldsymbol{V}, (1-\gamma)\boldsymbol{d}_0 \rangle,$$

it can be seen by the Bellman equation $\boldsymbol{r} + \gamma \boldsymbol{P}\boldsymbol{V}^\pi = \boldsymbol{Q}^\pi$ that

$$L(\boldsymbol{\mu}; \boldsymbol{V}^{\pi(\mu)}) = \langle \boldsymbol{V}^{\pi(\mu)}, (1-\gamma)\boldsymbol{d}_0 \rangle = (1-\gamma) J(\pi(\mu))$$

for all μ.

Consider any pair of function classes $\mathcal{U} \subseteq \mathbb{R}^{|\mathcal{S}\times\mathcal{A}|}$ and $\mathcal{V} \subseteq \mathbb{R}^{|\mathcal{S}|}$ such that $\mu^* \in \mathcal{U}$ and $V^* \in \mathcal{V}$, and that $V^\pi \in \mathcal{V}$ for all policy π. Let $(\widehat{\mu}, \widehat{V})$ be any saddle point of L over $\mathcal{U} \times \mathcal{V}$, and let $\widehat{\pi} = \pi(\widehat{\mu})$ be the policy extracted from $\widehat{\mu}$. Then, since $\widehat{\mu}$ is the best response among functions in $\mathcal{U}$ to $\widehat{V}$, and $\mu^* \in \mathcal{U}$, we have

$$L(\widehat{\mu}, \widehat{V}) \geq L(\mu^*, \widehat{V}) = (1-\gamma) J(\pi^*).$$

On the other hand, since $\widehat{V}$ is the best response to $\widehat{\mu}$ and $V^{\widehat{\pi}} \in \mathcal{V}$ where $\widehat{\pi} = \pi(\widehat{\mu})$ is the policy extracted from $\widehat{\mu}$, we have

$$L(\widehat{\mu}, \widehat{V}) \leq L(\widehat{\mu}, V^{\widehat{\pi}}) = (1-\gamma)J(\widehat{\pi}).$$

It follows that $J(\widehat{\pi}) \geq J(\pi^*)$ and that $\widehat{\pi} = \pi(\widehat{\mu})$ is optimal. □

Proof of Lemma 17. Suppose $(\widehat{\mu}, \widehat{\nu}; \widehat{Q})$ is a saddle point of L over $\mathcal{F}(\mathcal{U}, \Pi) \times \mathcal{Q}$. First, we show that if $(\mu, \nu) \in \mathcal{F}(\mathcal{U}, \Pi)$ and $\pi = \pi(\nu)$ is the policy extracted from ν, then $L(\mu, \nu; Q^\pi) = J(\pi)$. It can be seen by

$$\begin{aligned}
L(\mu, \nu; Q^\pi) &= \langle r, \mu \rangle + \langle Q^\pi, \nu - \mu \rangle \\
&= \langle r, \mu \rangle + \sum_{s,a} Q^\pi(s,a)(\pi(a|s)((1-\gamma)d_0(s) + \gamma[P^\top \mu](s))) - \langle Q^\pi, \mu \rangle \\
&= \langle r, \mu \rangle + \sum_{s} V^\pi(s)((1-\gamma)d_0(s) + \gamma[P^\top \mu](s)) - \langle Q^\pi, \mu \rangle \\
&= \langle r, \mu \rangle + \langle V^\pi, (1-\gamma)d_0 + \gamma P^\top \mu \rangle - \langle Q^\pi, \mu \rangle \\
&= J(\pi) + \langle \mu, r + \gamma P V^\pi - Q^\pi \rangle \\
&= J(\pi)
\end{aligned}$$

where the last equality is by the Bellman equation $Q^\pi(s,a) = r(s,a) + \gamma[PV^\pi](s,a)$. Also, $L(\mu^*, \nu^*; \widehat{Q}) = \langle r, \mu^* \rangle = (1-\gamma)J(\pi^*)$. Since $(\mu^*, \nu^*) \in \mathcal{F}(\mathcal{U}, \Pi)$, we have

$$J(\pi^*) = L(\mu^*, \nu^*; \widehat{Q}) \leq L(\widehat{\mu}, \widehat{\nu}; \widehat{Q}) \leq L(\widehat{\mu}, \widehat{\nu}; Q^{\widehat{\pi}}) = J(\widehat{\pi})$$

where we denote by $\widehat{\pi} = \pi(\widehat{\nu})$ the policy extracted from $\widehat{\nu}$. The inequalities hold since $(\widehat{\nu}, \widehat{\nu}; \widehat{Q})$ is a saddle point. It follows that $\widehat{\pi}$ is optimal. □

C.1.4 Proof of Theorem 6

We first show the following concentration bound.

Lemma 57. *Consider a Lagrangian function estimate $\widehat{L}$:*

$$\widehat{L}(w, V) = (1-\gamma)V(s_0) + \frac{1}{n}\sum_{i=1}^{n} w(s_i, a_i)(r(s_i, a_i) + \gamma V(s_i') - V(s_i)).$$

Given a function class $\mathcal{W}$ and $\mathcal{V}$ for w and V respectively, with boundedness condition $\|w\|_\infty \leq C$ for all $w \in \mathcal{W}$ and $\|V\|_\infty \leq \frac{1}{1-\gamma}$ for all $V \in \mathcal{V}$, we have for all $w \in \mathcal{W}$

and $V \in \mathcal{V}$ that

$$|L(w,V) - \widehat{L}(w,V)| \leq \mathcal{O}((1+\frac{C}{1-\gamma})\sqrt{\log(|\mathcal{W}||\mathcal{V}|/\delta)/n})$$

with probability at least $1-\delta$.

Proof. Fix $w \in \mathcal{W}$ and $V \in \mathcal{V}$. Let $Z_i = (1-\gamma)V(s_0) + w(s_i,a_i)(r(s_i,a_i) + \gamma V(s_i') - V(s_i))$. Then, $\mathbb{E}[Z_i] = L(w,V)$, and $|Z_i| \leq \mathcal{O}(1+\frac{C}{1-\gamma})$. By Hoeffding's inequality, we have $\frac{1}{n}\sum_{i=1}^n \mathbb{E}[Z_i] - Z_i \leq \mathcal{O}((1+\frac{C}{1-\gamma})\sqrt{\log(1/\delta)/n})$. The desired uniform concentration bound follows by a union bound over $\mathcal{W} \times \mathcal{V}$. □

Now, we are ready to prove the theorem.

Proof of Theorem 6. Let $(\widehat{w}, \widehat{V})$ be a saddle point that solves

$$\max_{w\in\mathcal{W}} \min_{V\in\mathcal{V}} \widehat{L}(w,V),$$

and let $\widehat{\pi}$ be the policy extracted from $\widehat{w}$. We have for all policy π that

$$L(w^\pi, V^\pi) = (1-\gamma)V^\pi(s_0) + \langle \mu^\pi, r + \gamma P V^\pi - V^\pi \rangle = (1-\gamma)J(\pi).$$

Let π^* be an optimal policy with $w^{\pi^*} \in \mathcal{W}$. Then,

$$\begin{aligned}
(1-\gamma)J(\pi^*) &= L(w^{\pi^*}, V^{\pi^*}) \\
&\leq L(w^{\pi^*}, \widehat{V}) \\
&\leq \widehat{L}(w^{\pi^*}, \widehat{V}) + \varepsilon \\
&\leq \widehat{L}(\widehat{w}, \widehat{V}) + \varepsilon \\
&\leq \widehat{L}(\widehat{w}, V^{\widehat{\pi}}) + \varepsilon \\
&\leq L(\widehat{w}, V^{\widehat{\pi}}) + 2\varepsilon \\
&= (1-\gamma)J(\widehat{\pi}) + 2\varepsilon,
\end{aligned}$$

where $\varepsilon_n = \mathcal{O}((1+C^*/(1-\gamma))\sqrt{\log(|\mathcal{W}||\mathcal{V}|/\delta)/n})$. □

C.1.5 Proof of Proposition 2

Proof of Proposition 2. We use the same MDP constructed for proving Proposition 1. Let $\mathcal{S} = \{s_0, l_1, r_1, l_2, r_2\}$ and $\mathcal{A} = \{L, R\}$. Let the reward function be defined as $r(l_1, a) = 1$, $r(l_2, a) = 2$ for all $a \in \mathcal{A}$, and all other entries of the reward function are 0. Let $\gamma = 1/2$ and

$d_0(s_0) = 1$ such that the MDP starts from s_0 deterministically. The states l_1, l_2 and r_2 are absorbing, such that both actions L and R taken at these states do not change states, i.e., $P(s|s, L) = P(s|s, R) = 1$ for $s = l_1, l_2, r_2$. In other states s_0 and r_2, the action L has equal probability of transitioning to the states in the level below, i.e., $P(l_1|s_0, L) = P(r_1|s_0, L) = 1/2$ and $P(l_2|r_1, L) = P(r_2|r_1, L) = 1/2$. On the other hand, action R has probability $1/4$ of transitioning to the left state and probability $3/4$ of transitioning to the right state, i.e., $P(l_1|s_0, R) = P(l_2|r_1, R) = 1/4$ and $P(r_1|s_0, R) = P(r_2|r_1, R) = 3/4$.

It can be seen that the optimal state-action value function Q^* has values $Q^*(s_0, L) = Q^*(s_0, R) = 1, Q^*(l_1, R) = Q^*(l_1, L) = 2, Q^*(r_1, L) = 2, Q^*(r_1, R) = 0, Q^*(l_2, L) = Q^*(l_2, R) = 8, Q^*(r_2, L) = Q^*(r_2, R) = 0$ and an optimal occupancy measure μ^* is the one induced by the policy that always chooses the action L: $\mu^*(s_0, L) = 1/2, \mu^*(l_1, L) = 1/4, \mu^*(r_1, L) = 1/8, \mu^*(l_2, L) = \mu^*(r_2, L) = 1/16$, and all other entries 0. By straightforward calculation, we get

$$\begin{aligned}
L(\mu^*, \nu_{\mu^*,\pi}; Q^*) &= (1-\gamma)Q^*(s_0, \pi) + \langle \boldsymbol{\mu}^*, \boldsymbol{r} + \gamma \boldsymbol{P} Q^*(\cdot, \pi) - \boldsymbol{Q}^* \rangle \\
&= \frac{1}{2} + \mu^*(s_0, L)(r(s_0, L) + \gamma[PQ^*(\cdot, \pi)](s_0, L) - Q^*(s_0, L)) \\
&\quad + \mu^*(l_1, L)(r(l_1, L) + \gamma[PQ^*(\cdot, \pi)](l_1, L) - Q^*(l_1, L)) \\
&\quad + \mu^*(r_1, L)(r(r_1, L) + \gamma[PQ^*(\cdot, \pi)](r_1, L) - Q^*(r_1, L)) \\
&\quad + \mu^*(l_2, L)(r(l_2, L) + \gamma[PQ^*(\cdot, \pi)](l_2, L) - Q^*(l_2, L)) \\
&\quad + \mu^*(r_2, L)(r(l_1, L) + \gamma[PQ^*(\cdot, \pi)](r_2, L) - Q^*(r_2, L)) \\
&= \frac{1}{2}
\end{aligned}$$

Note that $L(\boldsymbol{\mu}^*, \nu_{\mu^*,\pi}; \boldsymbol{Q}^*)$ does not depend on π. Hence, any $(\mu^*, \nu_{\mu^*,\pi})$ is a best response to $\boldsymbol{Q}^*$ that maximizes $L(\mu^*, \nu_{\mu^*,\pi}; \boldsymbol{Q}^*)$.

Consider function classes $\mathcal{Q} = \{Q^*\}$ and $\mathcal{U} = \{\mu^*\}$ and $\Pi = \{\pi^*, \widetilde{\pi}\}$ where $\widetilde{\pi}$ always chooses the action R. Then, $\widetilde{\pi}$ is not optimal, but both $(\mu^*, \nu_{\mu^*,\pi^*}; Q^*)$ and $(\mu^*, \nu_{\mu^*,\widetilde{\pi}}; Q^*)$ are saddle points of L over $\mathcal{F}(\mathcal{U}, \Pi) \times \mathcal{Q}$, and the learner cannot identify the optimal policy π^*. □

C.1.6 Proof of Theorem 7

We first show the following concentration bound.

Lemma 58. *Consider a Lagrangian function estimate*

$$\widehat{L}(w,\pi;Q) = (1-\gamma)Q(s_0,\pi) + \frac{1}{n}\sum_{i=1}^{n} w(s_i,a_i)(r(s_i,a_i) + \gamma Q(s_i',\pi) - Q(s_i,a_i))$$

Given a function class $\mathcal{W}$, Π and $\mathcal{Q}$ for w, π and Q, respectively, with boundedness condition $\|w\|_\infty \leq C$ for all $w \in \mathcal{W}$ and $\|Q\|_\infty \leq \frac{1}{1-\gamma}$ for all $Q \in \mathcal{Q}$, we have for all $w \in \mathcal{W}$, $\pi \in \Pi$ and $Q \in \mathcal{Q}$ that

$$|L(w,\pi;Q) - \widehat{L}(w,\pi;Q)| \leq \varepsilon_n.$$

with probability at least $1-\delta$ where $\varepsilon_n = \mathcal{O}(\frac{C}{1-\gamma}\sqrt{\log(|\mathcal{W}||\Pi||\mathcal{Q}|/\delta)/n})$.

Proof. Fix $w \in \mathcal{W}$, $\pi \in \Pi$ and $Q \in \mathcal{Q}$. Let

$$Z_i = (1-\gamma)Q(s_0,\pi) + w(s_i,a_i)(r(s_i,a_i) + \gamma Q(s_i',\pi) - Q(s_i,a_i)).$$

Then, $\mathbb{E}[Z_i] = L(w,\pi;Q)$ and $|Z_i - \mathbb{E}[Z_i]| \leq \mathcal{O}(\frac{C}{1-\gamma})$. By the Hoeffding's inequality, we have

$$|L(w,\pi;Q) - \widehat{L}(w,\pi;Q)| \leq \mathcal{O}(\frac{C}{1-\gamma}\sqrt{\log(1/\delta)/n}).$$

By a union bound over $(w,\pi,Q) \in \mathcal{W}\times\Pi\times\mathcal{Q}$, the proof is complete. □

Now, we are ready to prove the theorem.

Proof of Theorem 7. Let $(\widehat{w},\widehat{\pi};\widehat{Q})$ be a saddle point that solves

$$\max_{w\in\mathcal{W},\pi\in\Pi}\min_{Q\in\mathcal{Q}} \widehat{L}(w,\pi;Q),$$

We have for all policy π that

$$\begin{aligned} L(w,\pi;Q^\pi) &= \langle \mu, r\rangle + \langle Q^\pi, \nu - \mu\rangle \\ &= \langle \mu, r\rangle + \sum_s Q^\pi(s,\pi)((1-\gamma)d_0(s) + \gamma[P^\top \mu](s)) - \langle Q^\pi, \mu\rangle \\ &= (1-\gamma)\langle Q^\pi(\cdot,\pi), d_0\rangle + \langle \mu, r + \gamma[PQ^\pi(\cdot,\pi)] - Q^\pi\rangle \\ &= (1-\gamma)J(\pi). \end{aligned}$$

Also,

$$L(w^*,\pi^*;Q) = (1-\gamma)J(\pi^*)$$

for all Q. Hence,

$$\begin{aligned}
(1-\gamma)J(\pi^*) &= L(w^{\pi^*}, \pi^*; Q^{\pi^*}) \\
&\leq L(w^{\pi^*}, \pi^*; \widehat{Q}) \\
&\leq \widehat{L}(w^{\pi^*}, \pi^*; \widehat{Q}) + \varepsilon_n \\
&\leq \widehat{L}(\widehat{w}, \widehat{\pi}; \widehat{Q}) + \varepsilon_n \\
&\leq \widehat{L}(\widehat{w}, \widehat{\pi}; Q^{\widehat{\pi}}) + \varepsilon_n \\
&\leq L(\widehat{w}, \widehat{\pi}; Q^{\widehat{\pi}}) + 2\varepsilon_n \\
&= (1-\gamma)J(\widehat{\pi}) + 2\varepsilon_n,
\end{aligned}$$

where $\varepsilon_n = \mathcal{O}((1+\frac{C}{1-\gamma})\sqrt{\log(|\mathcal{W}||\Pi||\mathcal{Q}|/\delta)/n})$. □

C.1.7 Proof of Theorem 8

We first show that the softmax policy class that captures the policies encountered by the algorithm has low covering number.

Lemma 59. *Consider two value functions $Q, Q' : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ with $\|Q' - Q\|_\infty \leq 1$. Then, for all $s \in \mathcal{S}$, we have*

$$\sum_{a \in \mathcal{A}} |\pi_{Q'}(a|s) - \pi_Q(a|s)| \leq 8\|Q - Q'\|_\infty$$

where $\pi_Q(\cdot|s) \propto \exp(Q(s,\cdot))$ denotes the softmax policy induced by Q.

Proof. Straight from the definition, we have

$$\begin{aligned}
T &:= \frac{\pi_{Q'}(a|s)}{\pi_Q(a|s)} \\
&= \frac{\exp(Q'(s,a))}{\exp(Q(s,a))} \cdot \frac{\sum_{a''} \exp(Q(s,a''))}{\sum_{a'} \exp(Q'(s,a'))} \\
&= \exp(Q'(s,a) - Q(s,a)) \cdot \sum_{a''} (\exp(Q(s,a'') - Q'(s,a'')) \cdot \frac{\exp(Q'(s,a''))}{\sum_{a'} \exp(Q'(s,a'))}) \\
&= \exp(Q'(s,a) - Q(s,a)) \sum_{a''} \pi_{Q'}(a''|s) \exp(Q(s,a'') - Q'(s,a'')) \\
&\leq \exp(2\Delta) \\
&\leq 1 + 4\Delta
\end{aligned}$$

where we use the notation $\Delta = \|Q' - Q\|_\infty$. The second to last inequality follows since $\sum_{a''} \pi_{Q'}(a''|s) = 1$. The last inequality follows by the identity $\exp(x) \leq 1 + 2x$ for $x \in [0,1]$.

Rearranging, we get

$$\pi_{Q'}(a|s) - \pi_Q(a|s) \leq 4\pi_Q(a|s)\Delta.$$

Switching the roles of Q and Q', we get $\pi_Q(a|s) - \pi_{Q'}(a|s) \leq 4\pi_{Q'}(a|s)\Delta$, and it follows that

$$|\pi_Q(a|s) - \pi_{Q'}(a|s)| \leq 4\Delta \max\{\pi_Q(a|s), \pi_{Q'}(a|s)\} \leq 4\Delta(\pi_Q(a|s) + \pi_{Q'}(a|s)).$$

Summing over $a \in \mathcal{A}$, we get the desired inequality. □

Lemma 60. *For any $\epsilon \in (0, 1)$, we have*

$$\log \mathcal{N}(\Pi(\mathcal{Q}, C), \|\cdot\|_{\infty,1}, \epsilon) \leq \log \mathcal{N}(\mathcal{Q}, \|\cdot\|_{\infty}, \epsilon/(8C))$$

where the norm $\|\cdot\|_{\infty,1}$ is defined by

$$\|\pi - \pi'\|_{\infty,1} := \sup_{s\in\mathcal{S}} \sum_{a\in\mathcal{A}} |\pi(a|s) - \pi'(a|s)|.$$

Proof. The result is a direct consequence of Lemma 59. □

With the bound on the covering number of the softmax policy class, we show a uniform concentration bound on the Lagrangian estimate.

Lemma 61. *Consider a Lagrangian function estimate*

$$\widehat{L}(w, \pi; Q) = (1-\gamma)Q(s_0, \pi) + \frac{1}{n}\sum_{i=1}^{n} w(s_i, a_i)(r(s_i, a_i) + \gamma Q(s_i', \pi) - Q(s_i, a_i))$$

Given a function class $\mathcal{W}$, Π and $\mathcal{Q}$ for w, π and Q, respectively, with boundedness condition $\|w\|_\infty \leq C$ for all $w \in \mathcal{W}$ and $\|Q\|_\infty \leq \frac{1}{1-\gamma}$ for all $Q \in \mathcal{Q}$, we have with probability at least $1-\delta$ that, for all $w \in \mathcal{W}$, $\pi \in \Pi$ and $Q \in \mathcal{Q}$,

$$|L(w, \pi; Q) - \widehat{L}(w, \pi; Q)| \leq \varepsilon_n.$$

where $\varepsilon_n = \mathcal{O}(\frac{C}{1-\gamma}\sqrt{\log(\mathcal{N}_{1/\sqrt{n}}(\mathcal{W}, \|\cdot\|_\infty)\mathcal{N}_{1/\sqrt{n}}(\Pi, \|\cdot\|_{1,\infty})\mathcal{N}_{1/\sqrt{n}}(\mathcal{Q}, \|\cdot\|_\infty)/\delta)/n})$.

Proof. Consider coverings $\mathcal{C}(\mathcal{W}, \|\cdot\|_\infty, \varepsilon)$, $\mathcal{C}(\Pi, \|\cdot\|_{1,\infty}, \varepsilon)$ and $\mathcal{C}(\mathcal{Q}, \|\cdot\|_\infty, \varepsilon)$ for $\mathcal{W}$, Π and $\mathcal{Q}$, respectively, where $\varepsilon = 1/\sqrt{n}$ with cardinalities bounded by $\mathcal{N}_\varepsilon(\mathcal{W}, \|\cdot\|_\infty)$, $\mathcal{N}_\varepsilon(\Pi, \|\cdot\|_{1,\infty})$ and $\mathcal{N}_\varepsilon(\mathcal{Q}, \|\cdot\|_\infty)$, respectively. Following the proof of Lemma 58 and applying a union bound over the coverings, we get with probability at least $1-\delta$ that for all $\bar{w} \in \mathcal{C}(\mathcal{W}, \|\cdot\|_\infty, \varepsilon)$,

$\bar{\pi} \in \mathcal{C}(\Pi, \|\cdot\|_{1,\infty}, \varepsilon)$ and $\bar{Q} \in \mathcal{C}(\mathcal{Q}, \|\cdot\|_\infty, \varepsilon)$ that

$$\begin{aligned}
&|L(\bar{w}, \bar{\pi}; \bar{Q}) - \widehat{L}(\bar{w}, \bar{\pi}; \bar{Q})| \\
&\qquad \leq \mathcal{O}(\frac{C}{1-\gamma}\sqrt{\log(\mathcal{N}_\varepsilon(\mathcal{W}, \|\cdot\|_\infty)\mathcal{N}_\varepsilon(\Pi, \|\cdot\|_{1,\infty})\mathcal{N}_\varepsilon(\mathcal{Q}, \|\cdot\|_\infty)/\delta)/n}).
\end{aligned}$$

Given any $(w, \pi, Q) \in \mathcal{W} \times \Pi \times \mathcal{Q}$ consider $\bar{w}, \bar{\pi}, \bar{Q}$ in respective coverings such that $\|w - \bar{w}\|_\infty \leq \varepsilon$, $\max_s \|\pi(\cdot|s) - \bar{\pi}(\cdot|s)\|_1 \leq \varepsilon$ and $\|Q - \bar{Q}\|_\infty \leq \varepsilon$. Then, we have the following inequalities:

$$\begin{aligned}
\widehat{L}(w, \pi; Q) - \widehat{L}(w, \pi; \bar{Q}) &= (1-\gamma)(Q - \bar{Q})(s_0, \pi) \\
&\qquad + \frac{1}{n}\sum_{i=1}^n w(s_i, a_i)(\gamma(Q - \bar{Q})(s_i', \pi) - (Q - \bar{Q})(s_i, a_i)) \\
&\leq (1-\gamma)\varepsilon + 2C\varepsilon,
\end{aligned}$$

and

$$\begin{aligned}
\widehat{L}(w, \pi; \bar{Q}) - \widehat{L}(w, \bar{\pi}; \bar{Q}) &= (1-\gamma)\bar{Q}(s_0, \pi - \bar{\pi}) + \frac{\gamma}{n}\sum_{i=1}^n w(s_i, a_i)\bar{Q}(s_i', \pi - \bar{\pi}) \\
&\leq \varepsilon + \frac{C\varepsilon}{1-\gamma},
\end{aligned}$$

and

$$\begin{aligned}
\widehat{L}(w, \bar{\pi}; \bar{Q}) - \widehat{L}(\bar{w}, \bar{\pi}; \bar{Q}) &= \frac{1}{n}\sum_{i=1}^n (w - \bar{w})(s_i, a_i)(r(s_i, a_i) + \gamma Q(s_i', \pi) - Q(s_i, a_i)) \\
&\leq \frac{2\varepsilon}{1-\gamma}.
\end{aligned}$$

Summing up, we get $\widehat{L}(w, \pi; Q) - \widehat{L}(\bar{w}, \bar{\pi}; \bar{Q}) \leq \mathcal{O}(\frac{C\varepsilon}{1-\gamma})$. Similarly, we get the reverse inequality $\widehat{L}(\bar{w}, \bar{\pi}; \bar{Q}) - \widehat{L}(w, \pi; Q) \leq \mathcal{O}(\frac{C\varepsilon}{1-\gamma})$. Using the fact that $\widehat{L}$ is an unbiased estimate of L, and taking expectations on both sides of the inequalities, we get $|L(w, \pi; Q) - L(\bar{w}, \bar{\pi}; \bar{Q})| \leq \mathcal{O}(\frac{C\varepsilon}{1-\gamma})$. It follows by the choice $\varepsilon = 1/\sqrt{n}$ that

$$\begin{aligned}
|L(w, \pi; Q) - \widehat{L}(w, \pi; Q)| &\leq |L(w, \pi; Q) - L(\bar{w}, \bar{\pi}; \bar{Q})| + |L(\bar{w}, \bar{\pi}; \bar{Q}) - \widehat{L}(\bar{w}, \bar{\pi}; \bar{Q})| \\
&\qquad + |\widehat{L}(\bar{w}, \bar{\pi}; \bar{Q}) - \widehat{L}(w, \pi; Q)| \\
&\leq \varepsilon_n
\end{aligned}$$

where ε_n is defined in the theorem statement. □

Before showing the theorem, we provide a guarantee for the strategy employed by the

π-player.

Lemma 62 (Lemma D.2 in Gabbianelli et al. [2024b])**.** *Let $Q_1, \dots, Q_T : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ be a sequence of functions so that $\|Q_t\|_\infty \leq D$ for $t = 1, \dots, T$. Given an initial policy π_1 and a learning rate $\alpha > 0$, let*

$$\pi_{t+1}(\cdot|s) \propto \pi_t(\cdot|s) \exp(\alpha Q_t(s, \cdot))$$

for all $s \in \mathcal{S}$ for $t = 1, \dots, T$. Then, for any comparator policy π^*, we have

$$\sum_{t=1}^{T} \sum_{s \in \mathcal{S}} \mu^{\pi^*}(s) \langle \pi^*(\cdot|s) - \pi_t(\cdot|s), Q_t(s, \cdot) \rangle \leq \frac{\mathcal{H}(\pi^* \| \pi_1)}{\alpha} + \frac{\alpha T D^2}{2}$$

where $\mathcal{H}(\pi \| \pi') := \sum_{s \in \mathcal{S}} \mu^\pi(s) KL(\pi(\cdot|s) \| \pi'(\cdot|s))$.

We are now ready to show the theorem.

Proof of Theorem 8. Let $\bar{\pi} = \text{Unif}(\pi_1, \dots, \pi_T)$ be the policy returned by Algorithm 7. It is a policy that first uniformly randomly draws from $\{\pi_1, \dots, \pi_T\}$, then follows the policy for the entire trajectory. It follows that $J(\bar{\pi}) = \frac{1}{T} \sum_{t=1}^{T} J(\pi_t)$, and

$$\begin{aligned}(1-\gamma)J(\pi^*) - (1-\gamma)J(\bar{\pi}) &= \frac{1}{T} \sum_{t=1}^{T} L(w^*, \pi^*; Q_t) - L(w_t, \pi_t; Q^{\pi_t}) \\ &= \frac{1}{T} \sum_{t=1}^{T} L(w^*, \pi^*; Q_t) - L(w^*, \pi_t; Q_t) \\ &\qquad + \frac{1}{T} \sum_{t=1}^{T} L(w^*, \pi_t; Q_t) - L(w_t, \pi_t; Q_t) \\ &\qquad + \frac{1}{T} \sum_{t=1}^{T} L(w_t, \pi_t; Q_t) - L(w_t, \pi_t; Q^{\pi_t}).\end{aligned}$$

Note that the suboptimality of the policy $\bar{\pi}$ is decomposed into average regret terms of the three players. We bound each of the regret terms as follows.

Regret of π-player The one-step regret of π-player can be written as

$$\begin{aligned}
L(w^*,\pi^*;Q_t)-L(w^*,\pi_t;Q_t) &= (1-\gamma)\langle Q_t(\cdot,\pi^*-\pi_t),d_0\rangle+\gamma\langle\mu^*,PQ_t(\cdot,\pi^*-\pi_t)\rangle\\
&= (1-\gamma)\langle Q_t(\cdot,\pi^*-\pi_t),d_0\rangle+\gamma\langle Q_t(\cdot,\pi^*-\pi_t),P^\top\mu^*\rangle\\
&= \langle Q_t(\cdot,\pi^*-\pi_t),(1-\gamma)d_0+\gamma P^\top\mu^*\rangle\\
&= \langle \boldsymbol{E}^\top\mu^*,Q_t(\cdot,\pi^*-\pi_t)\rangle\\
&= \sum_{s\in\mathcal{S}}\mu^*(s)Q_t(s,\pi^*-\pi_t)\\
&= \sum_{s\in\mathcal{S}}\mu^*(s)\langle\pi^*(\cdot|s)-\pi_t(\cdot|s),Q_t(s,\cdot)\rangle.
\end{aligned}$$

Hence, the regret of π-player can be bounded as

$$\begin{aligned}
\sum_{t=1}^T(L(w^*,\pi^*;Q_t)-L(w^*,\pi_t;Q_t)) &= \sum_{t=1}^T\sum_{s\in\mathcal{S}}\mu^*(s)\langle\pi^*(\cdot|s)-\pi_t(\cdot|s),Q_t(s,\cdot)\rangle\\
&\le \frac{\mathcal{H}(\pi^*\|\pi_1)}{\alpha}+\frac{\alpha TD^2}{2}\\
&\le \frac{\log|\mathcal{A}|}{\alpha}+\frac{\alpha T}{2(1-\gamma)^2}\\
&\le \frac{2}{1-\gamma}\sqrt{T\log|\mathcal{A}|}
\end{aligned}$$

where the first inequality is by Lemma 62 and the second inequality is by a trivial bound of relative entropy: $\mathcal{H}(\pi^*\|\pi_1)=\mathbb{E}_{s\sim\mu^\pi}[KL(\pi^*(\cdot|s)\|\pi_1(\cdot|s))]\le\log|\mathcal{A}|$. The last inequality is by the choice of $\alpha=(1-\gamma)\sqrt{\log|\mathcal{A}|}/\sqrt{T}$.

Regret of w-player Note that the policies encountered by the algorithm is in the function class $\Pi(\mathcal{Q},T)$. By the concentration bound in Lemma 58 and the covering number bound of $\Pi(\mathcal{Q},T)$ provided by Lemma 60, the one-step regret of w-player can be bounded by

$$L(w^*,\pi_t;Q_t)-L(w_t,\pi_t;Q_t)\le\widehat{L}(w^*,\pi_t;Q_t)-\widehat{L}(w_t,\pi_t;Q_t)+2\varepsilon_n$$

where $\varepsilon_n=\mathcal{O}((\frac{C}{1-\gamma})\sqrt{\log(\mathcal{N}_{1/\sqrt{n}}(\mathcal{W},\|\cdot\|_\infty)\mathcal{N}_{1/(8\sqrt{n}T)}(\mathcal{Q},\|\cdot\|_\infty)/\delta)/n})$. Hence, the regret of w-player can be bounded by

$$\begin{aligned}
\sum_{t=1}^T L(w^*,\pi_t;Q_t)-L(w_t,\pi_t;Q_t) &\le \sum_{t=1}^T\widehat{L}(w^*,\pi_t;Q_t)-\widehat{L}(w_t,\pi_t;Q_t)+2\varepsilon_nT\\
&\le \mathrm{Reg}_T+2\varepsilon_nT
\end{aligned}$$

where the second inequality holds by the fact that the w-player employs a no-regret oracle.

Regret of Q-player The one-step regret of Q-player can be bounded as

$$\begin{aligned} L(w_t, \pi_t; Q_t) - L(w_t, \pi_t; Q^{\pi_t}) &\leq \widehat{L}(w_t, \pi_t; Q_t) - \widehat{L}(w_t, \pi_t; Q^{\pi_t}) + 2\varepsilon_n \\ &\leq 2\varepsilon_n \end{aligned}$$

where the first inequality is by the concentration inequality in Lemma 58 and the second inequality is by the fact that Q-player chooses $Q_t \in \mathcal{Q}$ greedily and that $Q^{\pi_t} \in \mathcal{Q}$ by the all-policy realizability assumption. Hence, the regret of Q-player can be bounded by

$$\sum_{t=1}^{T} L(w_t, \pi_t; Q_t) - L(w_t, \pi_t; Q^{\pi_t}) \leq 2\varepsilon_n T.$$

Combining the regret bounds, we get

$$(1-\gamma)J(\pi^*) - (1-\gamma)J(\bar{\pi}) \leq \frac{2}{1-\gamma}\sqrt{\log|\mathcal{A}|}/\sqrt{T} + \mathrm{Reg}_T/T + 4\varepsilon_n$$

□

C.2 Analysis for Offline Constrained RL

C.2.1 Convex Optimization

Proof of Lemma 18. Let $\pi^\star$ be an optimal policy of the optimization problem $\mathcal{P}(\boldsymbol{\tau})$. Define the dual function $f(Q, \boldsymbol{\lambda}) = \max_{\mu \in \mathbb{R}_+^{\mathcal{S}\times\mathcal{A}}, \nu \in \mathbb{R}_+^{\mathcal{S}\times\mathcal{A}}} L(\mu, \nu; Q, \lambda)$. Let $(\boldsymbol{Q}^*, \boldsymbol{\lambda}^\star) = \mathrm{argmin}_{\boldsymbol{Q} \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}}, \boldsymbol{\lambda} \in \mathbb{R}_+^I} f(\boldsymbol{Q}, \boldsymbol{\lambda})$. Trivially, $\lambda_i^\star \geq 0$ for all $i = 1, \ldots, I$. Also, by strong duality, we have $f(Q^*, \boldsymbol{\lambda}^\star) = J_0(\pi^\star)$.

By the definition of the dual function, for any policy π, we have

$$\begin{aligned} (1-\gamma)J_0(\pi^*) = f(Q^*, \lambda^*) &\geq L(\mu^\pi, \mu^\pi; Q^*, \lambda^*) \\ &= \langle \mu^\pi, r_0 \rangle + \langle \mu^\pi, R^\top \lambda^* \rangle - \langle \lambda^*, \tau \rangle \\ &= (1-\gamma)J_0(\pi) + (1-\gamma)\sum_{i=1}^{I} \lambda_i^*(J_i(\pi) - \frac{\tau_i}{1-\gamma}). \end{aligned} \tag{C.1}$$

Let $\widehat{\pi}$ be a feasible policy with $J_i(\widehat{\pi}) \geq (\tau_i + \varphi)/(1-\gamma)$ for all $i = 1, \ldots, I$. Such a policy exists by the assumption of this lemma. Then, using the display above, we get

$$\begin{aligned}(1-\gamma)J_0(\pi^*) &\geq (1-\gamma)J_0(\widehat{\pi}) + (1-\gamma)\sum_{i=1}^{I}\lambda_i^*(J_i(\widehat{\pi}) - \frac{\tau_i}{1-\gamma}) \\ &\geq (1-\gamma)J_0(\widehat{\pi}) + \varphi\sum_{i=1}^{I}\lambda_i^* \\ &= (1-\gamma)J_0(\widehat{\pi}) + \varphi\|\lambda^*\|_1.\end{aligned}$$

Rearranging and using $1/(1-\gamma) \geq J_0(\pi^\star) \geq J_0(\widehat{\pi}) \geq 0$ completes the proof:

$$\|\boldsymbol{\lambda}^\star\|_1 \leq \frac{(1-\gamma)J_0(\pi^\star) - (1-\gamma)J_0(\widehat{\pi})}{\varphi} \leq \frac{1}{\varphi}.$$

□

C.2.2 Proof of Theorem 9

We first show a concentration bound on the Lagrangian estimate.

Lemma 63. *Consider a Lagrangian function estimate*

$$\begin{aligned}&\widehat{L}(w,\pi;Q,\lambda) \\ &= (1-\gamma)Q(s_0,\pi) + \frac{1}{n}\sum_{j=1}^{n}w(s_j,a_j)((r_0 + \sum_{i=1}^{I}\lambda_i r_i)(s_j,a_j) + \gamma Q(s_j',\pi) - Q(s_j,a_j)) - \langle\lambda,\tau\rangle\end{aligned}$$

Given a function class $\mathcal{W}$, Π and $\mathcal{Q}$ for w, π and Q, respectively, with boundedness condition $\|w\|_\infty \leq C$ for all $w \in \mathcal{W}$ and $\|Q\|_\infty \leq \frac{1}{1-\gamma}$ for all $Q \in \mathcal{Q}$, we have with probability at least $1-\delta$ that, for all $w \in \mathcal{W}$, $\pi \in \Pi$, $Q \in \mathcal{Q}$ and $\lambda \in \frac{1}{\varphi}\Delta^I$,

$$|L(w,\pi;Q,\lambda) - \widehat{L}(w,\pi;Q,\lambda)| \leq \varepsilon_n.$$

where $\varepsilon_n = \mathcal{O}((\frac{1}{\varphi} + \frac{1}{1-\gamma})C\sqrt{\log(\mathcal{N}_{1/\sqrt{n}}(\mathcal{W},\|\cdot\|_\infty)\mathcal{N}_{1/\sqrt{n}}(\Pi,\|\cdot\|_{1,\infty})\mathcal{N}_{1/\sqrt{n}}(\mathcal{Q},\|\cdot\|_\infty)/\delta)/n})$.

Proof. Fix $w \in \mathcal{W}$, $\pi \in \Pi$, $Q \in \mathcal{Q}$ and $\lambda \in \frac{1}{\varphi}\Delta^I$. Let

$$Z_j = (1-\gamma)Q(s_0,\pi) + w(s_j,a_j)((r_0 + \sum_{i=1}^{I}\lambda_i r_i)(s_j,a_j) + \gamma Q(s_j',\pi) - Q(s_j,a_j)) - \langle\lambda,\tau\rangle.$$

Then, $\mathbb{E}[Z_j] = L(w,\pi;Q,\lambda)$ and $|Z_j - \mathbb{E}[Z_j]| \leq \mathcal{O}(\frac{C}{\varphi} + \frac{C}{1-\gamma})$. By the Hoeffding's inequality, we have

$$|L(w,\pi;Q,\lambda)-\widehat{L}(w,\pi;Q,\lambda)|\le \mathcal{O}((\frac{C}{\varphi}+\frac{C}{1-\gamma})\sqrt{\log(1/\delta)/n}).$$

Given any $(w,\pi,Q,\lambda)\in\mathcal{W}\times\Pi\times\mathcal{Q}\times\frac{1}{\varphi}\Delta^I$ consider $\bar{w},\bar{\pi},\bar{Q},\bar{\lambda}$ in respective coverings such that $\|w-\bar{w}\|_\infty\le\varepsilon$, $\max_s\|\pi(\cdot|s)-\bar{\pi}(\cdot|s)\|_1\le\varepsilon$, $\|Q-\bar{Q}\|_\infty\le\varepsilon$ and $\|\lambda-\bar{\lambda}\|_1\le\varepsilon$. Then, following a similar calculation as in the proof of Lemma 58, we have

$$\begin{aligned}
&\widehat{L}(w,\pi;Q,\lambda)-\widehat{L}(w,\pi;Q,\bar{\lambda})\le 2C\varepsilon\\
&\widehat{L}(w,\pi;Q,\bar{\lambda})-\widehat{L}(w,\pi;\bar{Q},\bar{\lambda})\le (1-\gamma)\varepsilon+2C\varepsilon\\
&\widehat{L}(w,\pi;\bar{Q},\bar{\lambda})-\widehat{L}(w,\bar{\pi};\bar{Q},\bar{\lambda})\le \varepsilon+\frac{C\varepsilon}{1-\gamma}\\
&\widehat{L}(w,\bar{\pi};\bar{Q},\bar{\lambda})-\widehat{L}(\bar{w},\bar{\pi};\bar{Q},\bar{\lambda})\le \frac{2\varepsilon}{1-\gamma}+\frac{\varepsilon}{\varphi}.
\end{aligned}$$

Summing up, we get $\widehat{L}(w,\pi;Q,\lambda)-\widehat{L}(\bar{w},\bar{\pi};\bar{Q},\bar{\lambda})\le\mathcal{O}(\frac{C\varepsilon}{1-\gamma}+\frac{\varepsilon}{\varphi})$. Similarly, we get the reverse inequality $\widehat{L}(\bar{w},\bar{\pi};\bar{Q},\bar{\lambda})-\widehat{L}(w,\pi;Q,\lambda)\le\mathcal{O}(\frac{C\varepsilon}{1-\gamma}+\frac{\varepsilon}{\varphi})$. Using the fact that $\widehat{L}$ is an unbiased estimate of L, and taking expectations on both sides of the inequalities, we get $|L(w,\pi;Q,\lambda)-L(\bar{w},\bar{\pi};\bar{Q},\bar{\lambda})|\le\mathcal{O}(\frac{C\varepsilon}{1-\gamma}+\frac{\varepsilon}{\varphi})$. It follows by the choice $\varepsilon=1/\sqrt{n}$ that

$$\begin{aligned}
&|L(w,\pi;Q,\lambda)-\widehat{L}(w,\pi;Q,\lambda)|\\
&\qquad\le|L(w,\pi;Q,\lambda)-L(\bar{w},\bar{\pi};\bar{Q},\bar{\lambda})|+|L(\bar{w},\bar{\pi};\bar{Q},\bar{\lambda})-\widehat{L}(\bar{w},\bar{\pi};\bar{Q},\bar{\lambda})|\\
&\qquad\qquad+|\widehat{L}(\bar{w},\bar{\pi};\bar{Q},\bar{\lambda})-\widehat{L}(w,\pi;Q,\lambda)|\\
&\qquad\le\varepsilon_n
\end{aligned}$$

where ε_n is defined in the theorem statement. □

Definition 4. *We say $(\bar{x},\bar{y})$ is a ξ-near saddle point for a function $L(\cdot,\cdot)$ with respect to the input space $\mathcal{X}\times\mathcal{Y}$ if $L(x,\bar{y})\le L(\bar{x},y)+\xi$ for all $x\in\mathcal{X}$ and $y\in\mathcal{Y}$.*

Lemma 64. *If $(\widehat{x},\widehat{y})$ is a saddle point of $\widehat{L}(\cdot,\cdot)$ over $\mathcal{X}\times\mathcal{Y}$ and $|L(x,y)-\widehat{L}(x,y)|\le\xi/2$ for all $(x,y)\in\mathcal{X}\times\mathcal{Y}$, then $(\widehat{x},\widehat{y})$ is a ξ-near saddle point of $L(\cdot,\cdot)$.*

Proof. Since $(\widehat{x},\widehat{y}$ is a saddle point of $\widehat{L}(\cdot,\cdot)$, we have for all $(x,y)\in\mathcal{X}\times\mathcal{A}$ that

$$L(x,\widehat{y})-L(\widehat{x},y)\le\widehat{L}(x,\widehat{y})-\widehat{L}(\widehat{x},y)+\xi\le\xi,$$

as required. □

Lemma 65. *Assume that Slater's condition (Assumption 3) holds. Suppose $(\widehat{w},\widehat{\pi};\widehat{Q},\widehat{\lambda})$ is a ξ-near saddle point for the Lagrangian function $L(\cdot,\cdot;\cdot,\cdot)$ with respect to $(\mathcal{W}\times\Pi)\times(\mathcal{Q}\times$*

$(1+\frac{1}{\varphi})\Delta^I)$. *Then, we have*

$$J_0(\widehat{\pi}) \geq J_0(\pi^*) - \frac{\xi}{1-\gamma} \tag{Optimality}$$

$$J_i(\widehat{\pi}) \geq \frac{\tau_i}{1-\gamma} - \frac{\xi}{1-\gamma}, \quad \textit{for all } i = 1, \dots I \tag{Feasibility}$$

where π^* *is an optimal policy in* Π.

Proof. We first prove $J_0(\bar{\pi}) \geq J_0(\pi^*) - \xi$. Since $(\widehat{w}, \widehat{\pi}; \widehat{Q}, \widehat{\lambda})$ is a ξ-near saddle point for $L(\cdot, \cdot; \cdot, \cdot)$ with respect to $(\mathcal{W} \times \Pi) \times (\mathcal{Q} \times (1+\frac{1}{\varphi})\Delta^I)$ and $\pi^* \in \Pi$, we have $L(w^*, \pi^*; \widehat{Q}, \widehat{\lambda}) \leq L(w^*, \pi^*; \widehat{Q}, \lambda) + \xi$ for all $\lambda \in (1+\frac{1}{\varphi})\Delta^I$. Choosing $\lambda = \mathbf{0}$, we get

$$L(w^*, \pi^*; \widehat{Q}, \widehat{\lambda}) \leq L(\widehat{w}, \widehat{\pi}; Q_0^{\widehat{\pi}}, 0) + \xi = (1-\gamma)J_0(\widehat{\pi}) + \xi.$$

Rearranging, we get

$$\begin{aligned} J_0(\widehat{\pi}) &\geq J_0(\pi^*) + \sum_{i=1}^{I} \widehat{\lambda}_i (J_i(\pi^*) - \frac{\tau_i}{1-\gamma}) - \frac{\xi}{1-\gamma} \\ &\geq J_0(\pi^*) - \frac{\xi}{1-\gamma} \end{aligned}$$

where the second inequality uses the feasibility of π^*. This proves the near optimality of $\widehat{\pi}$ with respect to π^*.

Now, we prove near feasibility of $\widehat{\pi}$. Consider a saddle point $(w^*, \pi^*; Q^*, \lambda^*)$ of L. Then, by (C.1), we have

$$J_0(\pi^*) - J_0(\widehat{\pi}) \geq \sum_{i=1}^{I} \lambda_i^* (J_i(\widehat{\pi}) - \frac{\tau_i}{1-\gamma}) \geq m \sum_{i=1}^{I} \lambda_i^* = m\|\lambda^*\|_1 \tag{C.2}$$

where we define $m = \min_{i \in [I]}(J_i(\widehat{\pi}) - \frac{\tau_i}{1-\gamma})$.

Recall from the proof of the near optimality that $L(w^*, \pi^*; \widehat{Q}, \widehat{\lambda}) \leq L(\widehat{w}, \widehat{\pi}; Q, \lambda) + \xi$ for all $Q \in \mathcal{Q}$ and $\lambda \in (1+\frac{1}{\varphi})\Delta^I$ holds since $(\widehat{w}, \widehat{\pi}; \widehat{Q}, \widehat{\lambda})$ is a ξ-near saddle point. Choosing λ such that $\lambda_j = 1 + \frac{1}{\varphi}$ for $j = \operatorname{argmin}_{i \in [I]}(J_i(\widehat{\pi}) - \frac{\tau_i}{1-\gamma})$ and $\lambda_j = 0$ for other j's, we get

$$L(w^*, \pi^*; \widehat{Q}, \widehat{\lambda}) \leq L(\widehat{w}, \widehat{\pi}; Q_\lambda^{\widehat{\pi}}, \lambda) + \xi = (1-\gamma)J_0(\widehat{\pi}) + (1-\gamma)(1+1/\varphi)m + \xi,$$

where we define $Q_\lambda^\pi = Q_0^\pi + \sum_{i=1}^{I} \lambda_i Q_i^\pi$, which is the action value function of π with respect to the reward $r_0 + \sum_{i=1}^{I} \lambda_i r_i$. On the other hand, the feasibility of π^* gives

$$L(w^*, \pi^*; \widehat{Q}, \widehat{\lambda}) = (1-\gamma)J_0(\pi^*) + (1-\gamma)\sum_{i=1}^{I} \widehat{\lambda}_i (J_i(\pi^*) - \frac{\tau_i}{1-\gamma}) \geq (1-\gamma)J_0(\pi^*).$$

Combining the previous two inequalities, and (C.2), we get

$$(1+1/\varphi)m + \frac{\xi}{1-\gamma} \geq J_0(\pi^*) - J_0(\widehat{\pi}) \geq m\|\lambda^*\|_1.$$

Rearranging, we get

$$\frac{-\xi}{1+1/\varphi - \|\lambda^*\|_1} \leq (1-\gamma)m \leq (1-\gamma)J_i(\widehat{\pi}) - \tau_i$$

for all $i = 1, \ldots, I$. It follows that

$$(1-\gamma)J_i(\widehat{\pi}) \geq \tau_i - \frac{\xi}{1+1/\varphi - \|\lambda^*\|_1} \geq \tau_i - \xi$$

where the last inequality uses the fact that $\|\lambda^*\|_1 \leq 1/\varphi$. This completes the proof. □

Proof of Theorem 9. By Lemma 63, we have with probability at least $1-\delta$ that for all $w \in \mathcal{W}$, $\pi \in \Pi$, $Q \in \mathcal{Q}$ and $\lambda \in (1+\frac{1}{\varphi})\Delta^I$,

$$|L(w, \pi; Q, \lambda) - \widehat{L}(w, \pi; Q, \lambda)| \leq \varepsilon_n$$

where $\varepsilon_n = \mathcal{O}((\frac{1}{\varphi} + \frac{1}{1-\gamma})C\sqrt{\log(\mathcal{N}_{1/\sqrt{n}}(\mathcal{W}, \|\cdot\|_\infty)\mathcal{N}_{1/\sqrt{n}}(\Pi, \|\cdot\|_{1,\infty})\mathcal{N}_{1/\sqrt{n}}(\mathcal{Q}, \|\cdot\|_\infty)/\delta)/n})$. Let $(\widehat{w}, \widehat{\pi}; \widehat{Q}, \widehat{\lambda})$ be a saddle point of $\widehat{L}$. Then, by Lemma 64, it is a $(2\varepsilon_n)$-near saddle point of L, and the result follows by Lemma 65. □

C.2.3 Proof of Theorem 10

Lemma 66. *Assume that Slater's condition (Assumption 3) holds. Define* $L(\pi, \lambda) = (1-\gamma)J_0(\pi) + (1-\gamma)\sum_{i=1}^{I} \lambda_i (J_i(\pi) - \frac{\tau_i}{1-\gamma})$. *If* $L(\pi^*, \widehat{\lambda}) \leq L(\widehat{\pi}, \lambda) + \xi$ *for all* $\lambda \in (1+\frac{1}{\varphi})\Delta^I$, *then we have*

$$J_0(\widehat{\pi}) \geq J_0(\pi^*) - \frac{\xi}{1-\gamma} \qquad \text{(Optimality)}$$

$$J_i(\widehat{\pi}) \geq \frac{\tau_i}{1-\gamma} - \frac{\xi}{1-\gamma}, \quad \textit{for all } i = 1, \ldots I \qquad \text{(Feasibility)}$$

Proof. To show the near optimality, observe that

$$L(\pi^*, \widehat{\lambda}) \leq L(\widehat{\pi}, 0) + \xi = (1-\gamma)J_0(\widehat{\pi}) + \xi.$$

Rearranging, we get

$$\begin{aligned}(1-\gamma)J_0(\widehat{\pi}) &\geq L(\pi^*, \widehat{\lambda}) - \xi \\ &= (1-\gamma)J_0(\pi^*) + (1-\gamma)\sum_{i=1}^{I} \lambda_i(J_i(\pi^*) - \frac{\tau_i}{1-\gamma}) - \xi \\ &\geq (1-\gamma)J_0(\pi^*) - \xi,\end{aligned}$$

as desired.

To show near feasibility, consider a saddle point $(\mu^*, \pi^*; Q^*, \lambda^*)$ of $L(\cdot, \cdot; \cdot, \cdot)$ defined in Section 3.4.3. Then, π^* is an optimal policy, and for all $\lambda \in \frac{1}{\varphi}\Delta^I$, we have

$$L(\widehat{\pi}, \lambda^*) = L(w^{\widehat{\pi}}, \widehat{\pi}; Q^*, \lambda^*) \leq L(w^*, \pi^*; Q, \lambda) = L(\pi^*, \lambda).$$

Hence, choosing $\lambda = \lambda^*$, we have $L(\widehat{\pi}, \lambda^*) \leq L(\pi^*, \lambda^*) = (1-\gamma)J_0(\pi^*)$, and rearranging, we get

$$(1-\gamma)J_0(\pi^*) \geq (1-\gamma)J_0(\widehat{\pi}) + (1-\gamma)\sum_{i=1}^{I} \lambda_i^*(J_i(\widehat{\pi}) - \frac{\tau_i}{1-\gamma}) \geq (1-\gamma)J_0(\widehat{\pi}) + (1-\gamma)m\|\lambda^*\|_1$$

where we define $m = \min_{i\in[I]}(J_i(\widehat{\pi}) - \frac{\tau_i}{1-\gamma})$. Also, choosing λ to be $\lambda_i = 1 + \frac{1}{\varphi}$ where $i = \operatorname{argmin}_{i\in[I]}(J_i(\widehat{\pi}) - \frac{\tau_i}{1-\gamma})$, we get

$$\begin{aligned}(1-\gamma)J_0(\pi^*) &\leq (1-\gamma)J_0(\pi^*) + (1-\gamma)\sum_{i=1}^{I} \widehat{\lambda}_i(J_i(\pi^*) - \frac{\tau_i}{1-\gamma}) \\ &= L(\pi^*, \widehat{\lambda}) \leq L(\widehat{\pi}, \lambda) + \xi \\ &= (1-\gamma)J_0(\widehat{\pi}) + (1-\gamma)m(1 + \frac{1}{\varphi}) + \xi\end{aligned}$$

where the first inequality follows by the feasibility of π^*. Rearranging, we get

$$J_0(\pi^*) - J_0(\widehat{\pi}) \leq m(1 + \frac{1}{\varphi}) + \frac{\xi}{1-\gamma},$$

and combining with the previous inequality, we get

$$m\|\lambda^*\|_1 \leq m(1+\frac{1}{\varphi}) + \frac{\xi}{1-\gamma}.$$

Following the same calculation as in the proof of Lemma 65, we get the desired result. □

Proof of Theorem 10. Define $L(\pi,\lambda) = (1-\gamma)J_0(\pi) + (1-\gamma)\sum_{i=1}^I \lambda_i(J_i(\pi) - \frac{\tau_i}{1-\gamma})$. Let $\widehat{\pi} = \mathrm{Unif}\{\pi_1,\dots,\pi_T\}$ be the policy returned by the algorithm and let $\widehat{\lambda} = \frac{1}{T}\sum_{t=1}^T \lambda_t$. By Lemma 66, it is enough to show that $L(\pi^*,\widehat{\lambda}) \leq L(\widehat{\pi},\lambda) + \varepsilon_n$ for all $\lambda \in (1+\frac{1}{\varphi})\Delta^I$. Consider the following decomposition:

$$\begin{aligned}
L(\pi^*,\widehat{\lambda}) - L(\widehat{\pi},\lambda) &= \frac{1}{T}\sum_{t=1}^T L(\pi^*,\lambda_t) - L(\pi_t,\lambda) \\
&= \frac{1}{T}\sum_{t=1}^T L(w^*,\pi^*;Q_t,\lambda_t) - L(w_t,\pi_t;Q_\lambda^{\pi_t},\lambda) \\
&= \frac{1}{T}\sum_{t=1}^T L(w^*,\pi^*;Q_t,\lambda_t) - L(w^*,\pi_t;Q_t,\lambda_t) \\
&\qquad + \frac{1}{T}\sum_{t=1}^T L(w^*,\pi_t;Q_t,\lambda_t) - L(w_t,\pi_t;Q_t,\lambda_t) \\
&\qquad + \frac{1}{T}\sum_{t=1}^T L(w_t,\pi_t;Q_t,\lambda_t) - L(w_t,\pi_t;Q_\lambda^{\pi_t},\lambda_t) \\
&\qquad + \frac{1}{T}\sum_{t=1}^T L(w_t,\pi_t;Q_\lambda^{\pi_t},\lambda_t) - L(w_t,\pi_t;Q_\lambda^{\pi_t},\lambda).
\end{aligned}$$

Regret of π-player Using the uniform concentration bound on the Lagrangian estimate provided in Lemma 63, the summand in the first term can be bounded by

$$\begin{aligned}
L(w^*,\pi^*;Q_t,\lambda_t) - L(w^*,\pi_t;Q_t,\lambda_t) &\leq \widehat{L}(w^*,\pi^*;Q_t,\lambda_t) - \widehat{L}(w^*,\pi_t;Q_t,\lambda_t) + 2\varepsilon_n \\
&= (1-\gamma)Q_t(s_0,\pi^*-\pi_t) + \gamma\langle\mu^*, PQ_t(\cdot,\pi^*-\pi_t)\rangle,
\end{aligned}$$

which coincides with the regret of π-player in the analysis of the primal dual algorithm for the unconstrained RL provided in the proof of Theorem 8. Following the same proof, using the fact that the π player employs , we get

$$\sum_{t=1}^T L(w^*,\pi^*;Q_t,\lambda_t) - L(w^*,\pi_t;Q_t,\lambda_t) \leq \frac{2}{1-\gamma}\sqrt{T\log|\mathcal{A}|}.$$

Regret of w-player Similarly, the summand in the second term can be bounded by

$$L(w^*, \pi_t; Q_t, \lambda_t) - L(w_t, \pi_t; Q_t, \lambda_t) \leq \widehat{L}(w^*, \pi_t; Q_t, \lambda_t) - \widehat{L}(w_t, \pi_t; Q_t, \lambda_t) + 2\varepsilon_n,$$

and the no-regret oracle employed by the w-player ensures the summation is bounded by $\mathrm{Reg}_T + 2\varepsilon_n$.

Regret of Q-player The summand in the third term can be bounded by

$$L(w_t, \pi_t; Q_t, \lambda_t) - L(w_t, \pi_t; Q_\lambda^{\pi_t}, \lambda_t) \leq \widehat{L}(w_t, \pi_t; Q_t, \lambda_t) - \widehat{L}(w_t, \pi_t; Q_\lambda^{\pi_t}, \lambda_t) + 2\varepsilon_n,$$

which is bounded by $2\varepsilon_n$ since the Q-player employs a greedy strategy.

Regret of λ-player The summand in the fourth term can be bounded by

$$L(w_t, \pi_t; Q_\lambda^{\pi_t}, \lambda_t) - L(w_t, \pi_t; Q_\lambda^{\pi_t}, \lambda) \leq \widehat{L}(w_t, \pi_t; Q_\lambda^{\pi_t}, \lambda_t) - \widehat{L}(w_t, \pi_t; Q_\lambda^{\pi_t}, \lambda) + 2\varepsilon_n,$$

which is bounded by $2\varepsilon_n$ since the λ-player employs a greedy strategy.

Combining the four bounds, we get

$$L(\pi^*, \widehat{\lambda}) - L(\widehat{\pi}, \lambda) \leq \mathcal{O}(\varepsilon_n),$$

and invoking Lemma 66 completes the proof. □

C.3 Online Gradient Descent

We show a guarantee on the online gradient descent algorithm where the sequence of functions are linear:

Lemma 67. *Given $x_1 \in \mathcal{X} \subseteq \mathbb{R}^d$ where $\mathcal{X}$ is convex and $\eta > 0$, define the sequences $x_2, \ldots, x_{n+1} \in \mathcal{X}$ and $h_1, \ldots, h_n \in \mathbb{R}^d$ such that for $k = 1, \ldots, n$,*

$$x_{k+1} = \Pi_{\mathcal{X}}(x_k + \eta h_k)$$

where $\Pi_{\mathcal{X}}(\cdot)$ is a projection onto $\mathcal{X}$. Then, with the assumption that $\|h_k\|_2 \leq G$ for all

$k = 1, \dots, n$, we have for any $x^* \in \mathcal{X}$:

$$\sum_{k=1}^{n} \langle x^* - x_k, h_k \rangle \leq \frac{\|x_1 - x^*\|_2^2}{2\eta} + \frac{\eta n G^2}{2}.$$

Proof. We start by bounding following term:

$$\begin{aligned}\|x_{k+1} - x^*\|_2^2 &= \|\Pi_{\mathcal{X}}(x_k + \eta h_k) - x^*\|_2^2 \\ &\leq \|x_k + \eta h_k - x^*\|_2^2 \\ &= \|x_k - x^*\|_2^2 + 2\eta \langle x_k - x^*, h_k \rangle + \eta^2 \|h_k\|_2^2.\end{aligned}$$

Rearranging, we get

$$2\eta \langle x^* - x_k, h_k \rangle \leq \|x_k - x^*\|_2^2 - \|x_{k+1} - x^*\|_2^2 + \eta^2 \|h_k\|_2^2.$$

Summing over $k = 1, \dots, n$, we get

$$2\eta \sum_{k=1}^{n} \langle x^* - x_k, h_k \rangle \leq \|x_1 - x^*\|_2^2 - \|x_{n+1} - x^*\|_2^2 + \eta^2 \sum_{k=1}^{n} \|h_k\|_2^2 \leq \|x_1 - x^*\|_2^2 + \eta^2 n G^2.$$

Rearranging completes the proof. □

Using the lemma above, we can get see that the online gradient descent algorithm can be used as a no-regret oracle for the w-player in Algorithm 7. To see this, recall from the proof of Theorem 8 that the w-player aims to minimize

$$\begin{aligned}&\sum_{t=1}^{T} \widehat{L}(w^*, \pi_t; Q_t) - \widehat{L}(w_t, \pi_t; Q_t) \\ &\qquad = \sum_{t=1}^{T} \left(\frac{1}{n} \sum_{i=1}^{n} (w^*(s_i, a_i) - w_t(s_i, a_i))(r(s_i, a_i) + \gamma Q_t(s_i', \pi_t) - Q_t(s_i, a_i)) \right) \\ &\qquad = \sum_{t=1}^{T} \langle \widetilde{w}^* - \widetilde{w}_t, h_t \rangle,\end{aligned}$$

where $\widetilde{w}^*$ and $\widetilde{w}_t$ are n-dimensional vectors with i-th entry corresponding to $w^*(s_i, a_i)$ and $w_t(s_i, a_i)$, respectively. The n-vector h_t represents the vector with i-entry $r(s_i, a_i) + \gamma Q_t(s_i', \pi_t) - Q_t(s_i, a_i)$. Note that $\|h_t\|_2 \leq \sqrt{n}(1 + \frac{1}{1-\gamma})$ for all $t = 1, \dots, T$. Applying the lemma, we get

$$\sum_{t=1}^{T} \widehat{L}(w^*, \pi_t; Q_t) - \widehat{L}(w_t, \pi_t, ; Q_t) \leq \frac{C^2 n}{\eta} + \frac{\eta n T B^2}{2},$$

where $B = 1 + \frac{1}{1-\gamma}$. Choosing $\eta = C/(B\sqrt{T})$, we get the upper bound $\mathcal{O}(nCB\sqrt{T})$, which is sublinear in T.

The online gradient descent algorithm gives a similar guarantee when employed in Algorithm 8 for the constrained RL setting.

BIBLIOGRAPHY

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

Yasin Abbasi-Yadkori, David Pal, and Csaba Szepesvari. Online-to-confidence-set conversions and application to sparse stochastic bandits. In *Artificial Intelligence and Statistics*, pages 1–9. PMLR, 2012.

Priyank Agrawal and Shipra Agrawal. Optimistic q-learning for average reward and episodic reinforcement learning. *arXiv preprint arXiv:2407.13743*, 2024.

Eitan Altman. *Constrained Markov Decision Processes*, volume 7. CRC Press, 1999.

András Antos, Csaba Szepesvári, and Rémi Munos. Fitted q-iteration in continuous action-space mdps. *Advances in neural information processing systems*, 20, 2007.

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

Baruch Awerbuch and Robert Kleinberg. Online linear optimization and adaptive routing. *Journal of Computer and System Sciences*, 74(1):97–114, 2008.

Alex Ayoub, Zeyu Jia, Csaba Szepesvari, Mengdi Wang, and Lin Yang. Model-based reinforcement learning with value-targeted regression. In *International Conference on Machine Learning*, pages 463–474. PMLR, 2020.

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

Qinbo Bai, Amrit Singh Bedi, Mridul Agarwal, Alec Koppel, and Vaneet Aggarwal. Achieving zero constraint violation for concave utility constrained reinforcement learning via primal-dual approach. *Journal of Artificial Intelligence Research*, 78:975–1016, 2023.

Peter Bartlett and Ambuj Tewari. Regal: a regularization based algorithm for reinforcement learning in weakly communicating mdps. In *Uncertainty in Artificial Intelligence: Proceedings of the 25th Conference*, pages 35–42. AUAI Press, 2009.

Victor Boone and Zihan Zhang. Achieving tractable minimax optimal regret in average reward mdps. *Advances in Neural Information Processing Systems*, 2024.

Stephen P Boyd and Lieven Vandenberghe. *Convex optimization.* Cambridge university press, 2004.

Woojin Chae, Kihyuk Hong, Yufan Zhang, Ambuj Tewari, and Dabeen Lee. Learning infinite-horizon average-reward linear mixture mdps of bounded span. In *International Conference on Artificial Intelligence and Statistics*, 2025.

Jinglin Chen and Nan Jiang. Offline reinforcement learning under value and density-ratio realizability: the power of gaps. In *Uncertainty in Artificial Intelligence*, pages 378–388. PMLR, 2022.

Liyu Chen, Rahul Jain, and Haipeng Luo. Learning infinite-horizon average-reward markov decision process with constraints. In *International Conference on Machine Learning*, pages 3246–3270. PMLR, 2022.

Yi Chen, Jing Dong, and Zhaoran Wang. A primal-dual approach to constrained markov decision processes. *arXiv preprint arXiv:2101.10895*, 2021.

Ching-An Cheng, Tengyang Xie, Nan Jiang, and Alekh Agarwal. Adversarially trained actor critic for offline reinforcement learning. In *International Conference on Machine Learning*, pages 3852–3878. PMLR, 2022.

Dongsheng Ding, Kaiqing Zhang, Tamer Basar, and Mihailo Jovanovic. Natural policy gradient primal-dual method for constrained markov decision processes. *Advances in Neural Information Processing Systems*, 33:8378–8390, 2020.

Ronan Fruit, Matteo Pirotta, Alessandro Lazaric, and Ronald Ortner. Efficient bias-span-constrained exploration-exploitation in reinforcement learning. In *International Conference on Machine Learning*, pages 1578–1586. PMLR, 2018.

Ronan Fruit, Matteo Pirotta, and Alessandro Lazaric. Improved analysis of ucrl2 with empirical bernstein inequality. *arXiv preprint arXiv:2007.05456*, 2020.

Germano Gabbianelli, Gergely Neu, Matteo Papini, and Nneka M Okolo. Offline primal-dual reinforcement learning for linear mdps. In *International Conference on Artificial Intelligence and Statistics*, pages 3169–3177. PMLR, 2024a.

Germano Gabbianelli, Gergely Neu, Matteo Papini, and Nneka M Okolo. Offline primal-dual reinforcement learning for linear mdps. In *International Conference on Artificial Intelligence and Statistics*, pages 3169–3177. PMLR, 2024b.

Arnob Ghosh, Xingyu Zhou, and Ness Shroff. Achieving sub-linear regret in infinite horizon average reward constrained mdp with linear function approximation. In *The Eleventh International Conference on Learning Representations*, 2023.

Ilaria Giannoccaro and Pierpaolo Pontrandolfo. Inventory management in supply chains: a reinforcement learning approach. *International Journal of Production Economics*, 78(2): 153–161, 2002.

Joren Gijsbrechts, Robert N Boute, Jan A Van Mieghem, and Dennis J Zhang. Can deep reinforcement learning improve inventory management? performance on lost sales, dual-sourcing, and multi-echelon problems. *Manufacturing & Service Operations Management*, 24(3):1349–1368, 2022.

Abhijit Gosavi. Reinforcement learning for long-run average cost. *European journal of operational research*, 155(3):654–674, 2004.

Monique Guignard and Siwhan Kim. Lagrangean decomposition: A model yielding stronger lagrangean bounds. *Mathematical programming*, 39(2):215–228, 1987.

Elad Hazan et al. Introduction to online convex optimization. *Foundations and Trends® in Optimization*, 2(3-4):157–325, 2016.

Jiafan He, Dongruo Zhou, and Quanquan Gu. Nearly minimax optimal reinforcement learning for discounted mdps. *Advances in Neural Information Processing Systems*, 34:22288–22300, 2021.

Jiafan He, Heyang Zhao, Dongruo Zhou, and Quanquan Gu. Nearly minimax optimal reinforcement learning for linear markov decision processes. In *International Conference on Machine Learning*, pages 12790–12822. PMLR, 2023.

Jianliang He, Han Zhong, and Zhuoran Yang. Sample-efficient learning of infinite-horizon average-reward mdps with general function approximation. In *The Twelfth International Conference on Learning Representations*, 2024.

Kihyuk Hong and Ambuj Tewari. A primal-dual algorithm for offline constrained reinforcement learning with linear mdps. In *International Conference on Machine Learning*, pages 18711–18737. PMLR, 2024.

Kihyuk Hong and Ambuj Tewari. A computationally efficient algorithm for infinite-horizon average-reward linear mdps. In *International Conference on Machine Learning*. PMLR, 2025a.

Kihyuk Hong and Ambuj Tewari. Offline constrained reinforcement learning under partial data coverage. *arXiv preprint arXiv:2505.17506*, 2025b.

Kihyuk Hong, Yuhang Li, and Ambuj Tewari. A primal-dual-critic algorithm for offline constrained reinforcement learning. In *International Conference on Artificial Intelligence and Statistics*, pages 280–288. PMLR, 2024.

Kihyuk Hong, Woojin Chae, Yufan Zhang, Dabeen Lee, and Ambuj Tewari. Reinforcement learning for infinite-horizon average-reward linear mdps via approximation by discounted-reward mdps. In *International Conference on Artificial Intelligence and Statistics*, 2025.

Xiang Ji and Gen Li. Regret-optimal model-free reinforcement learning for discounted mdps with short burn-in time. *Advances in Neural Information Processing Systems*, 36, 2024.

Chi Jin, Zhuoran Yang, Zhaoran Wang, and Michael I Jordan. Provably efficient reinforcement learning with linear function approximation. In *Conference on learning theory*, pages 2137–2143. PMLR, 2020.

Ying Jin, Zhuoran Yang, and Zhaoran Wang. Is pessimism provably efficient for offline rl? In *International Conference on Machine Learning*, pages 5084–5096. PMLR, 2021.

Yujia Jin and Aaron Sidford. Towards tight bounds on the sample complexity of average-reward mdps. In *International Conference on Machine Learning*, pages 5055–5064. PMLR, 2021.

Aviral Kumar, Anikait Singh, Stephen Tian, Chelsea Finn, and Sergey Levine. A workflow for offline model-free robotic reinforcement learning. *arXiv preprint arXiv:2109.10813*, 2021.

Hoang Le, Cameron Voloshin, and Yisong Yue. Batch policy learning under constraints. In *International Conference on Machine Learning*, pages 3703–3712. PMLR, 2019.

Sergey Levine, Peter Pastor, Alex Krizhevsky, Julian Ibarz, and Deirdre Quillen. Learning hand-eye coordination for robotic grasping with deep learning and large-scale data collection. *The International journal of robotics research*, 37(4-5):421–436, 2018.

Sergey Levine, Aviral Kumar, George Tucker, and Justin Fu. Offline reinforcement learning: Tutorial, review, and perspectives on open problems. *arXiv preprint arXiv:2005.01643*, 2020.

Shuang Liu and Hao Su. Regret bounds for discounted mdps. *arXiv preprint arXiv:2002.05138*, 2020.

Zoubir Mammeri. Reinforcement learning based routing in networks: Review and classification of approaches. *Ieee Access*, 7:55916–55950, 2019.

Alan S Manne. Linear programming and sequential decisions. *Management Science*, 6(3): 259–267, 1960.

Prashant Mehta and Sean Meyn. Q-learning and pontryagin's minimum principle. In *Proceedings of the 48h IEEE Conference on Decision and Control (CDC) held jointly with 2009 28th Chinese Control Conference*, pages 3598–3605. IEEE, 2009.

Aditya Modi, Nan Jiang, Ambuj Tewari, and Satinder Singh. Sample complexity of reinforcement learning using linearly combined model ensembles. In *International Conference on Artificial Intelligence and Statistics*, pages 2010–2020. PMLR, 2020.

Rémi Munos. Error bounds for approximate policy iteration. In *Proceedings of the Twentieth International Conference on International Conference on Machine Learning*, pages 560–567, 2003.

Rémi Munos. Error bounds for approximate value iteration. In *Proceedings of the National Conference on Artificial Intelligence*, page 1006. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999, 2005.

Rémi Munos and Csaba Szepesvári. Finite-time bounds for fitted value iteration. *Journal of Machine Learning Research*, 9(5), 2008.

Gergely Neu and Nneka Okolo. Efficient global planning in large mdps via stochastic primal-dual optimization. In *International Conference on Algorithmic Learning Theory*, pages 1101–1123. PMLR, 2023.

Gergely Neu and Nneka Okolo. Offline rl via feature-occupancy gradient ascent. In *International Conference on Artificial Intelligence and Statistics*. PMLR, 2025a.

Gergely Neu and Nneka Okolo. Offline rl via feature-occupancy gradient ascent. In *International Conference on Artificial Intelligence and Statistics*. PMLR, 2025b.

Ronald Ortner. Regret bounds for reinforcement learning via markov chain concentration. *Journal of Artificial Intelligence Research*, 67:115–128, 2020.

Yi Ouyang, Mukul Gagrani, Ashutosh Nayyar, and Rahul Jain. Learning unknown markov decision processes: A thompson sampling approach. *Advances in neural information processing systems*, 30, 2017.

Martin L Puterman. *Markov decision processes: discrete stochastic dynamic programming.* John Wiley & Sons, 2014.

Paria Rashidinejad, Banghua Zhu, Cong Ma, Jiantao Jiao, and Stuart Russell. Bridging offline reinforcement learning and imitation learning: A tale of pessimism. *Advances in Neural Information Processing Systems*, 34:11702–11716, 2021.

Paria Rashidinejad, Hanlin Zhu, Kunhe Yang, Stuart Russell, and Jiantao Jiao. Optimal conservative offline rl with general function approximation via augmented lagrangian. *arXiv preprint arXiv:2211.00716*, 2022.

Fred Shepardson and Roy E Marsten. A lagrangean relaxation algorithm for the two duty period scheduling problem. *Management Science*, 26(3):274–281, 1980.

Shengpu Tang and Jenna Wiens. Model selection for offline reinforcement learning: Practical considerations for healthcare settings. In *Machine Learning for Healthcare Conference*, pages 2–35. PMLR, 2021.

Masatoshi Uehara and Wen Sun. Pessimistic model-based offline reinforcement learning under partial coverage. In *International Conference on Learning Representations*, 2022.

Andrew J Wagenmaker, Yifang Chen, Max Simchowitz, Simon Du, and Kevin Jamieson. First-order regret in reinforcement learning with linear function approximation: A robust estimation approach. In *International Conference on Machine Learning*, pages 22384–22429. PMLR, 2022.

Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge university press, 2019.

Jinghan Wang, Mengdi Wang, and Lin F Yang. Near sample-optimal reduction-based policy learning for average reward mdp. *arXiv preprint arXiv:2212.00603*, 2022.

Shengbo Wang, Jose Blanchet, and Peter Glynn. Optimal sample complexity for average reward markov decision processes. *arXiv preprint arXiv:2310.08833*, 2023.

Tianhao Wang, Dongruo Zhou, and Quanquan Gu. Provably efficient reinforcement learning with linear function approximation under adaptivity constraints. *Advances in Neural Information Processing Systems*, 34:13524–13536, 2021.

Chen-Yu Wei, Mehdi Jafarnia Jahromi, Haipeng Luo, Hiteshi Sharma, and Rahul Jain. Model-free reinforcement learning in infinite-horizon average-reward markov decision processes. In *International conference on machine learning*, pages 10170–10180. PMLR, 2020.

Chen-Yu Wei, Mehdi Jafarnia Jahromi, Haipeng Luo, and Rahul Jain. Learning infinite-horizon average-reward mdps with linear function approximation. In *International Conference on Artificial Intelligence and Statistics*, pages 3007–3015. PMLR, 2021.

Runzhe Wu, Yufeng Zhang, Zhuoran Yang, and Zhaoran Wang. Offline constrained multi-objective reinforcement learning via pessimistic dual value iteration. *Advances in Neural Information Processing Systems*, 34:25439–25451, 2021.

Yue Wu, Dongruo Zhou, and Quanquan Gu. Nearly minimax optimal regret for learning infinite-horizon average-reward mdps with linear function approximation. In *International Conference on Artificial Intelligence and Statistics*, pages 3883–3913. PMLR, 2022.

Tengyang Xie and Nan Jiang. Q* approximation schemes for batch reinforcement learning: A theoretical comparison. In *Conference on Uncertainty in Artificial Intelligence*, pages 550–559. PMLR, 2020.

Tengyang Xie, Ching-An Cheng, Nan Jiang, Paul Mineiro, and Alekh Agarwal. Bellman-consistent pessimism for offline reinforcement learning. *Advances in neural information processing systems*, 34:6683–6694, 2021.

Hongbing Yang, Wenchao Li, and Bin Wang. Joint optimization of preventive maintenance and production scheduling for multi-state production systems based on reinforcement learning. *Reliability Engineering & System Safety*, 214:107713, 2021.

Andrea Zanette. When is realizability sufficient for off-policy reinforcement learning? In *International Conference on Machine Learning*, pages 40637–40668. PMLR, 2023.

Andrea Zanette, Martin J Wainwright, and Emma Brunskill. Provable benefits of actor-critic methods for offline reinforcement learning. *Advances in neural information processing systems*, 34:13626–13640, 2021.

Wenhao Zhan, Baihe Huang, Audrey Huang, Nan Jiang, and Jason Lee. Offline reinforcement learning with realizability and single-policy concentrability. In *Conference on Learning Theory*, pages 2730–2775. PMLR, 2022.

Haobo Zhang, Xiyue Peng, Honghao Wei, and Xin Liu. Safe and efficient: A primal-dual method for offline convex cmdps under partial data coverage. *Advances in Neural Information Processing Systems*, 37:34239–34269, 2024.

Zihan Zhang and Xiangyang Ji. Regret minimization for reinforcement learning by evaluating the optimal bias function. *Advances in Neural Information Processing Systems*, 32, 2019.

Zihan Zhang and Qiaomin Xie. Sharper model-free reinforcement learning for average-reward markov decision processes. In *The Thirty Sixth Annual Conference on Learning Theory*, pages 5476–5477. PMLR, 2023.

Hanlin Zhu, Paria Rashidinejad, and Jiantao Jiao. Importance weighted actor-critic for optimal conservative offline reinforcement learning. *arXiv preprint arXiv:2301.12714*, 2023.

Matthew Zurek and Yudong Chen. Span-based optimal sample complexity for average reward mdps. *arXiv preprint arXiv:2311.13469*, 2023.