Learning Weakly Communicating Average-Reward CMDPs: Strong Duality and Improved Regret

Kihyun Yu
Department of Industrial and Systems Engineering
KAIST
khyu99@kaist.ac.kr

Beomhan Baek
Department of Computer Science
UCLA
beomhanbaek@ucla.edu

Dabeen Lee*
Department of Mathematical Science
Seoul National University
dabeenl@snu.ac.kr

Abstract

We study infinite-horizon average-reward constrained Markov decision processes (CMDPs) under the weakly communicating assumption. Our contributions are twofold. First, we establish strong duality for weakly communicating average-reward CMDPs over stationary policies with finite state and action spaces. Despite the absence of a linear programming formulation and the resulting nonconvexity under the weakly communicating setting, we show that strong duality still holds by carefully exploiting the geometric structure of the occupation measure set. Second, building on this result, we propose a primal–dual clipped value iteration algorithm for learning weakly communicating average-reward linear CMDPs. Our algorithm achieves regret and constraint violation bounds of $\widetilde{\mathcal{O}}(T^{2/3})$, improving upon the best known bounds, where T denotes the number of interactions. Our approach extends clipped value iteration to the constrained setting and adapts it to a finite-horizon approximation, which stabilizes the dual variable and is crucial for achieving improved regret bounds. To analyze this, we develop a novel approach based on strong duality that enables the decomposition of the composite Lagrangian regret into separate bounds on regret and constraint violation.

1 Introduction

Safe reinforcement learning (RL) studies how to optimize performance while ensuring that safety constraints are satisfied during learning (García and Fernández, 2015). It is crucial in safety-critical applications such as autonomous driving, where unsafe actions may lead to collisions (Isele et al., 2018); robotics, where constraint violations can damage the environment (Achiam et al., 2017); healthcare, where unreliable decisions may harm patients (Coronato et al., 2020); and large language models, where unsafe outputs must be avoided (Dai et al., 2023). These problems can be naturally formulated as constrained Markov decision processes (CMDPs) (Altman, 1999).

To capture long-run optimal behavior, infinite-horizon average-reward CMDPs have been widely studied. In this setting, the long-run average is defined as the limit of the expected time-average of the reward and constraint functions, and the objective is to find a policy that maximizes the long-run average reward while ensuring that the long-run average constraint satisfies a prescribed threshold.

*Additional affiliations: Research Institute of Mathematics, Seoul National University; Interdisciplinary Program in Artificial Intelligence, Seoul National University; Korea Institute for Advanced Study.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

Accordingly, a number of learning algorithms have been developed for this problem (Wei et al., 2022; Singh et al., 2022; Chen et al., 2022; Ghosh et al., 2023).

A key feature of learning average-reward CMDPs is that both the difficulty of the problem and the resulting performance guarantees depend critically on the structural assumptions of the transition dynamics, such as ergodic, unichain, and weakly communicating assumptions (Puterman, 1994).[2] For instance, in the tabular setting, Chen et al. (2022) proposed computationally efficient algorithms under different assumptions, achieving regret bounds of $\widetilde{\mathcal{O}}(\sqrt{T})$ under the ergodic assumption and $\widetilde{\mathcal{O}}(T^{2/3})$ under the weakly communicating assumption, while also providing a computationally inefficient algorithm achieving $\widetilde{\mathcal{O}}(\sqrt{T})$ under the weakly communicating assumption, where T denotes the number of interactions. More recently, under a similar assumption, Ghosh et al. (2023) extended the algorithm to the linear setting, establishing a regret bound of $\widetilde{\mathcal{O}}(T^{3/4})$. These examples illustrate that weaker structural assumptions, while more general, lead to increased analytical challenges and weaker performance guarantees.

The fundamental reason for this gap is that weaker structural assumptions admit fewer technical tools. A prominent example is strong duality. Although it plays a central role in many CMDP algorithms (Paternain et al., 2019), its applicability in average-reward CMDPs depends on the underlying structural assumptions. In particular, in the average-reward CMDP literature, strong duality has been established under the unichain assumption (Altman, 1999). More recently, Bai et al. (2024a) established strong duality under general policy parameterization; however, their result relies on the ergodic assumption. In contrast, beyond these stronger structural assumptions, its validity under the weakly communicating assumption remains unclear. These examples highlight that weaker structural assumptions limit the available technical tools—such as strong duality—thereby motivating the need to extend these tools to enable improved learning algorithms.

Motivated by this limitation, we study weakly communicating average-reward CMDPs, starting with an investigation of strong duality under this assumption. The main challenge is that, unlike the unichain case, this problem does not admit an equivalent linear programming (LP) reformulation. Specifically, under the unichain assumption, the problem can be reformulated as a linear program (LP) via occupation measures (Altman, 1999), allowing strong duality to be directly applied. In contrast, under the weakly communicating assumption, the occupation measure formulation no longer yields an LP, and the resulting problem becomes nonconvex. This fundamental difference significantly complicates the analysis and necessitates new techniques to establish strong duality in this more general setting. Despite these challenges, our main contributions are summarized as follows.

- We establish strong duality for weakly communicating average-reward CMDPs. In particular, we consider the average-reward CMDP over stationary policies and show that, when the problem is feasible, an optimal dual variable exists and the primal supremum equals the attained dual minimum. The main challenge in establishing strong duality lies in the fact that the set of occupation measures does not form a polytope and is, in fact, nonconvex. To address this, we show that certain geometric properties—namely, that the set behaves like a polytope with some missing boundary points—are sufficient to invoke strong duality for LPs. To the best of our knowledge, this is the first strong duality result for weakly communicating average-reward CMDPs over stationary policies, and it admits broad applicability in the analysis of a wide class of learning algorithms.
- Building on our strong duality result, we propose a primal–dual clipped value iteration algorithm for learning weakly communicating average-reward linear CMDPs. Our algorithm achieves regret and constraint violation bounds of $\widetilde{\mathcal{O}}(T^{2/3})$, improving upon the best known bound of $\widetilde{\mathcal{O}}(T^{3/4})$ (Ghosh et al., 2023). To design the algorithm, we extend clipped value iteration (Hong et al., 2024) to the constrained setting and adapt it to the finite-horizon approximation, whereas prior work applies it to the discounted-reward approximation. The finite-horizon approximation stabilizes the dual variable, making it a crucial modification for achieving improved regret bounds in the constrained setting, together with value function clipping. To complete the analysis, we develop a novel approach that fundamentally relies on the strong duality result. This enables us to decompose the composite Lagrangian regret into separate bounds on regret and constraint violation.

[2] If an MDP is ergodic, then it is unichain; if it is unichain, then it is weakly communicating. For more details on MDP classifications, see Appendix C.

Further details on related work can be found in the appendix.

2 Preliminaries

Infinite-horizon average-reward CMDP Let $\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, r, g, b)$ denote a CMDP, where $\mathcal{S}$, $\mathcal{A}$ are the finite state and action spaces, P is the transition kernel, $r, g : \mathcal{S} \times \mathcal{A} \to [0, 1]$ are the reward and constraint functions, respectively, and $b \in [0, 1]$ is the constraint threshold. In addition, $P(s'|s, a)$ represents the transition probability from state $s \in \mathcal{S}$ to state $s' \in \mathcal{S}$ by taking action $a \in \mathcal{A}$. For simplicity, we assume that r, g are deterministic. Given $\mathcal{M}$ and initial state $s_1 \in \mathcal{S}$, we define an average-reward CMDP as

$$\sup_{\pi \in \Pi} J_r^\pi(s_1) \quad \text{s.t.} \quad J_g^\pi(s_1) \geq b, \tag{CMDP}$$

where $J_r^\pi(s_1)$, $J_g^\pi(s_1)$ denote the gains associated with the policy π, Π denotes the set of stationary randomized policies, and $J_r^*(s_1)$ denotes the optimal value of (CMDP). We define $J_r^\pi(s) \triangleq \lim_{T\to\infty} \frac{1}{T}\mathbb{E}_\pi\big[\sum_{t=1}^T r(s_t, a_t)|s_1 = s\big]$, where the expectation is taken over the trajectory $\{(s_t, a_t)\}_{t=1}^T$ generated by P and π, with $s_{t+1} \sim P(\cdot|s_t, a_t)$ and $a_t \sim \pi(\cdot|s_t)$ for all $t \in [T]$. We define the state bias functions as $v_r^\pi(s) \triangleq \lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^T \mathbb{E}_{P,\pi}\left[\sum_{i=1}^t (r(s_i, a_i) - J_r^\pi(s_i))|s_1 = s\right]$. Note that $J_g^\pi(s)$ and $v_g^\pi(s)$ can be defined in similar ways. Given any bounded function $V : \mathcal{S} \to \mathbb{R}$, we define the span of V as $\mathrm{sp}(V) \triangleq \max_{s\in\mathcal{S}} V(s) - \min_{s'\in\mathcal{S}} V(s')$, and $PV(s, a) \triangleq \sum_{s'} P(s'|s, a)V(s')$.

Weakly communicating setting Throughout the paper, we assume weakly communicating CMDPs, defined as follows. An MDP $\mathcal{M}$ is said to be *weakly communicating* if there exists a stationary randomized policy which induces a Markov chain with a single closed irreducible class and a set of states which is transient under all stationary policies (Proposition 8.3.1 in Puterman (1994)). For additional notions of Markov chains and MDP classifications, we refer the reader to Appendix C.

We introduce the notion of occupation measure, which represents the limiting expected frequency of visiting the state–action pair (s, a) under policy π starting from s_1. Given an MDP and an initial state s_1, we define the *occupation measure* associated with a stationary policy π as $q_{s_1}^\pi \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{A}|}$. For each $(s, a) \in \mathcal{S} \times \mathcal{A}$, it is given by $q_{s_1}^\pi(s, a) \triangleq \lim_{T\to\infty} \frac{1}{T}\mathbb{E}_\pi\left[\sum_{t=1}^T \mathbb{1}\{s_t = s, a_t = a\} \mid s_1\right]$.

For a stationary policy π, when the state and action spaces are finite, the above limit exists (Proposition 8.9.1 in (Puterman, 1994)), and $J_r^\pi(s_1)$ can be expressed as $r^\top q_{s_1}^\pi$, with an analogous expression for g. In the weakly communicating model, the occupation measure may depend on the initial state, which motivates explicitly including s_1 in the notation. Accordingly, let $\widehat{\mathcal{P}}(s_1)$ denote the set of occupation measures achievable by stationary policies from s_1, defined as

$$\widehat{\mathcal{P}}(s_1) \triangleq \left\{q_{s_1}^\pi \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{A}|} : \pi \in \Pi\right\}.$$

Using occupation measures, we rewrite (CMDP) as follows.

$$\sup_{q\in\widehat{\mathcal{P}}(s_1)} \quad r^\top q \quad \text{s.t.} \quad g^\top q \geq b. \tag{Occ-CMDP}$$

Additionally, we define relative interior and extreme points of a convex set as follows. Given a convex set $\mathcal{P} \subseteq \mathbb{R}^n$, a point $x \in \mathcal{P}$ is called *relative interior* if for all $y \in \mathcal{P}$, there exists $\epsilon > 1$ such that $\epsilon x + (1 - \epsilon)y \in \mathcal{P}$. A point $x \in \mathcal{P}$ is called an *extreme* point if $x = \epsilon y + (1 - \epsilon)z$ implies that $x = y = z$ for $y, z \in \mathcal{P}$ and $\epsilon \in (0, 1)$. Let $\mathrm{relint}(\mathcal{P})$ and $\mathrm{ext}(\mathcal{P})$ denote the set of relative interior and extreme points of $\mathcal{P}$, respectively.

Online linear CMDP We next formulate the online infinite-horizon average-reward linear CMDPs on which Section 4 mainly focuses. The agent learns an optimal policy for an unknown CMDP through interaction with the environment. The interaction starts from an arbitrary initial state $s_1 \in \mathcal{S}$. At each step $t \in [T]$, the agent observes s_t and takes an action $a_t \sim \pi_t(\cdot \mid s_t)$. The environment then samples $s_{t+1} \sim P(\cdot \mid s_t, a_t)$. The performance metrics are regret and constraint violation, defined as $\mathrm{Regret}(T) = \sum_{t=1}^T (J_r^*(s_1) - r(s_t, a_t))$ and $\mathrm{Violation}(T) = \sum_{t=1}^T (b - g(s_t, a_t))$, where $J_r^*(s_1)$ denotes the optimal value of (CMDP).

A CMDP $\mathcal{M}$ is called *linear* if the following conditions hold. For a known feature mapping $\phi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$, the reward and constraint functions, and transition kernel admit the following linear representations: for some $\theta_r, \theta_g \in \mathbb{R}^d$ and $\psi : \mathcal{S} \to \mathbb{R}^d_+$, $r(s,a) = \phi(s,a)^\top \theta_r$, $g(s,a) = \phi(s,a)^\top \theta_g$, and $P(s'|s,a) = \phi(s,a)^\top \psi(s')$ for all $(s,a,s') \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}$. We further assume that $\|\phi(s,a)\|_2 \le 1$ for all $(s,a) \in \mathcal{S} \times \mathcal{A}$, and $\|\theta_r\|_2, \|\theta_g\|_2 \le \sqrt{d}$, $\|\sum_{s' \in \mathcal{S}} \psi(s')\|_2 \le \sqrt{d}$. For simplicity, we assume that θ_r, θ_g are known, while ψ is unknown[3].

We next introduce the Slater condition, a standard assumption for learning CMDPs (Ghosh et al., 2023). While our strong duality theorem itself does not require the Slater condition, this condition guarantees an upper bound on the optimal dual variable (Lemma 3), which is useful for designing and analyzing our online algorithm.

Assumption 1 (Slater condition)**.** *There exists a Slater policy $\bar{\pi} \in \Pi$ such that $J_g^{\bar{\pi}}(s_1) \ge b + \gamma$ for some Slater constant $\gamma > 0$. Moreover, we assume that γ is known, while $\bar{\pi}$ is unknown.*

3 Strong duality

In this section, we establish strong duality and highlight the challenges that arise, particularly in the weakly communicating setting. We begin by introducing the Lagrangian formulation. For a dual variable $\lambda \ge 0$, the Lagrangian associated with the primal problem (CMDP) is defined as

$$\mathcal{L}(\pi, \lambda) = J_r^\pi(s_1) + \lambda\big(J_g^\pi(s_1) - b\big).$$

The corresponding dual function is defined as the supremum of $\mathcal{L}$ over stationary policies:

$$D(\lambda) = \sup_{\pi \in \Pi} \mathcal{L}(\pi, \lambda).$$

The dual problem is then given by $\inf_{\lambda \ge 0} D(\lambda)$. Let λ^* denote a minimizer of $D(\lambda)$, when it exists. We are now ready to state the strong duality result.

Theorem 1. *Let $\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, r, g, b)$ be a weakly communicating CMDP with $|\mathcal{S}|, |\mathcal{A}| < \infty$, and let $s_1 \in \mathcal{S}$ be any initial state. Suppose that (CMDP) is feasible. Then the following properties hold:*

1. *There exists $\lambda^* \ge 0$ such that $D(\lambda^*) = \inf_{\lambda \ge 0} D(\lambda)$.*
2. *Strong duality, i.e., $J_r^*(s_1) = D(\lambda^*)$.*

Even when the underlying MDP is only weakly communicating, this theorem guarantees that, with finite state and action spaces and feasibility, the dual problem attains its minimum at some nonnegative dual variable. It further establishes strong duality, so the optimal values of the primal and dual problems coincide. This property is central in many analyses, and in Section 4, we show how it can be leveraged to analyze an online algorithm and obtain improved regret upper bounds.

3.1 Challenges and techniques: strong duality

While strong duality for CMDPs may appear well established—suggesting that Theorem 1 should follow directly—this is not the case in our setting. The difficulty arises because we assume only a weakly communicating structure. In particular, under this assumption, (CMDP) cannot be easily reformulated as an equivalent LP, making the application of LP strong duality nontrivial.

To clarify this challenge, we first explain why establishing strong duality becomes more straightforward under stronger structural assumptions. For instance, Altman (1999) considers the setting where the MDP satisfies a unichain structure. In such a case, the problem admits an equivalent LP reformulation, and the corresponding duality follows directly from LP duality. Specifically, one can show that the set of occupation measures induced by stationary policies is indeed a polytope (Theorem 4.2 in Altman (1999)), i.e., $\widehat{\mathcal{P}}(s_1) = \mathcal{P}$, where the polytope $\mathcal{P}$ is defined as

$$\mathcal{P} \triangleq \left\{ q \in \mathbb{R}^{|\mathcal{S}| \times |\mathcal{A}|} : \begin{array}{l} \sum_{s,a} q(s,a) = 1,\ q(s,a) \ge 0\ \forall (s,a) \in \mathcal{S} \times \mathcal{A}, \\ \sum_{a} q(s,a) = \sum_{s',a'} q(s',a') P(s|s',a'),\ \forall s \in \mathcal{S} \end{array} \right\}.$$

[3] Our results for the online algorithm can be extended to the case of unknown stochastic rewards and constraints without affecting the main contributions up to constant factors.

Figure 1: Comparison of $\widehat{\mathcal{P}}(s_1)$ and $\mathcal{P}$ under the weakly communicating assumption.

As a consequence, (Occ-CMDP) admits the following equivalent LP reformulation, and this connection allows us to rely on LP strong duality under the unichain assumption:

$$\text{(Occ-CMDP)} \overset{\widehat{\mathcal{P}}(s_1)=\mathcal{P}}{=} \max_{q\in\mathcal{P}} r^\top q \text{ s.t. } g^\top q \geq b. \tag{1}$$

However, this is not applicable under the weakly communicating assumption, as $\widehat{\mathcal{P}}(s_1) \neq \mathcal{P}$, and $\widehat{\mathcal{P}}(s_1)$ may be nonconvex (Mannor and Tsitsiklis, 2005). This prevents an LP reformulation of the CMDP, in contrast to the unichain case as in (1). Consequently, even when expressed in terms of occupation measures, (Occ-CMDP) remains a nonconvex problem, which constitutes the main obstacle to establishing strong duality, particularly in the weakly communicating setting.

To overcome these limitations in the weakly communicating setting, our key observation is that $\widehat{\mathcal{P}}(s_1)$ still possesses several favorable geometric properties; perhaps surprisingly, these properties satisfy sufficient conditions for applying LP strong duality, even though $\widehat{\mathcal{P}}(s_1) \neq \mathcal{P}$ and may be nonconvex. These properties are formalized in the following lemma.[4]

Lemma 1. *Let $\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, r)$ be a weakly communicating MDP with $|\mathcal{S}|, |\mathcal{A}| < \infty$, and let $s_1 \in \mathcal{S}$ be any initial state. We have $\mathrm{relint}(\mathcal{P}) \subseteq \widehat{\mathcal{P}}(s_1)$, $\mathrm{ext}(\mathcal{P}) \subseteq \widehat{\mathcal{P}}(s_1)$, and $\widehat{\mathcal{P}}(s_1) \subseteq \mathcal{P}$.*

Intuitively, this lemma suggests that (Occ-CMDP) may retain key properties of LPs, even though it is not itself an LP. This is because $\widehat{\mathcal{P}}(s_1)$ has a structure similar to that of a polytope, in the sense that only certain boundary points are missing from $\mathcal{P}$, as illustrated in Figure 1. In particular, $\mathrm{relint}(\mathcal{P}) \subseteq \widehat{\mathcal{P}}(s_1)$ in Lemma 1 allows us to recover the same optimal value as follows:

$$\text{(Occ-CMDP)} \overset{\mathrm{relint}(\mathcal{P})\subseteq\widehat{\mathcal{P}}(s_1)}{=} \max_{q\in\mathcal{P}} r^\top q \text{ s.t. } g^\top q \geq b.$$

Thus, although $\widehat{\mathcal{P}}(s_1) \neq \mathcal{P}$, the two problems share the same optimal value, establishing a key connection between (Occ-CMDP) and an LP over $\mathcal{P}$. Now, since $\mathcal{P}$ is a polytope, strong duality for LPs applies, allowing us to express the right-hand side equivalently as

$$\max_{q\in\mathcal{P}} r^\top q \text{ s.t. } g^\top q \geq b \quad = \quad \min_{\lambda\geq 0} \max_{q\in\mathcal{P}} r^\top q + \lambda(g^\top q - b).$$

Since $\widehat{\mathcal{P}}(s_1) \neq \mathcal{P}$ in our case, it remains to show that $\mathcal{P}$ in the right-hand side can be replaced by $\widehat{\mathcal{P}}(s_1)$ so as to recover the dual function $D(\lambda)$. This follows from the inclusion $\mathrm{ext}(\mathcal{P}) \subseteq \widehat{\mathcal{P}}(s_1)$. Specifically, for any linear objective, an LP over $\mathcal{P}$ attains its optimum at an extreme point, which implies that the same optimal value is obtained when optimizing over $\widehat{\mathcal{P}}(s_1)$. Namely, for any $\lambda \geq 0$,

$$\max_{q\in\mathcal{P}} r^\top q + \lambda(g^\top q - b) \;=\; \sup_{q\in\widehat{\mathcal{P}}(s_1)} r^\top q + \lambda(g^\top q - b).$$

Note that the right-hand side coincides with the dual function $D(\lambda)$ expressed in terms of occupation measures. Taking $\min_{\lambda\geq 0}$ on both sides yields

$$\min_{\lambda\geq 0} \max_{q\in\mathcal{P}} r^\top q + \lambda(g^\top q - b) \;=\; \min_{\lambda\geq 0} D(\lambda).$$

Consequently, combining these, we obtain (CMDP) $= \min_{\lambda\geq 0} D(\lambda)$—strong duality for average-reward CMDP under the weakly communicating assumption.

[4]This observation was noted in Remark 3.2 of Mannor and Tsitsiklis (2005) without a rigorous proof. For completeness, we provide a proof in Appendix E.

Discussion on multichain CMDPs A natural further question is whether strong duality holds under more general structural assumptions, such as multichain CMDPs (see Appendix C for the formal definition). This appears more challenging, since the key ingredient, Lemma 1, may no longer hold. For comparison, under the weakly communicating assumption, the inclusion $\mathrm{relint}(\mathcal{P}) \subseteq \widehat{\mathcal{P}}(s_1)$ heavily relies on the existence of a stationary policy that induces a single closed irreducible class, together with properties of the relative interior. However, such a policy is not guaranteed for multichain CMDPs. Therefore, the desired inclusion may fail, and so may $\mathrm{ext}(\mathcal{P}) \subseteq \widehat{\mathcal{P}}(s_1)$, making strong duality over stationary policies unclear.

4 Improved algorithm for average-reward linear CMDPs

In this section, we propose an algorithm for learning weakly communicating average-reward linear CMDPs, establishing improved regret bounds. We emphasize that the improvement stems from a novel primal-dual algorithm, together with an analysis enabled by Theorem 1.

Finite-horizon approximation As a preliminary, we first introduce a method for learning average-reward MDP problems—*episodic finite-horizon approximation* framework (Wei et al., 2021; Chen et al., 2022; Ghosh et al., 2023). This approach partitions the total iterations T into K episodes of length H, such that $T = HK$, and applies algorithms designed for finite-horizon CMDPs. Rather than directly estimating the average-reward quantities J_r^π or v_r^π, the method focuses on learning $Q_{r,h}^\pi$ and $V_{r,h}^\pi$, referred to as Q-function and value function under finite-horizon approximation. For a fixed horizon H and any reward function r, these functions are defined as $Q_{r,h}^\pi(s,a) = \mathbb{E}_\pi\left[\sum_{j=h}^H r(s_j,a_j)|s_h=s,a_h=a\right]$, $V_{r,h}^\pi(s) = \mathbb{E}_\pi\left[\sum_{j=h}^H r(s_j,a_j)|s_h=s\right]$. The validity of the finite-horizon approximation is supported by the following lemma, which establishes a connection between two quantities, $V_{r,h}^\pi$ and J_r^π, arising from different settings.

Lemma 2 (Lemma 4 in Chen et al. (2022))**.** *For any $\pi \in \Pi$ and reward function $r \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{A}|}$ such that $J_r^\pi(s) = J_r^\pi$ for all $s \in \mathcal{S}$, we have $|V_{r,h}^\pi(s) - (H-h+1)J_r^\pi| \le \mathrm{sp}(v_r^\pi)$ for $(s,h) \in \mathcal{S}\times[H]$.*

4.1 Challenges and techniques: average-reward linear CMDPs

Under the finite-horizon approximation framework, the main challenge in the linear setting can be summarized as follows: obtaining a refined regret bound in the finite-horizon regime remains nontrivial even for time-homogeneous P. More specifically, since directly estimating P is intractable in the linear setting, one instead estimates the optimal value functions $\{V_h^*\}_{h=1}^H$. However, estimating this sequence introduces an additional multiplicative factor in H, since $\{V_h^*\}_{h=1}^H$ varies with h, even when P is time-homogeneous (Zhang et al., 2024). As H may depend on T, this leads to an unfavorable regret bound in the linear setting (Ghosh et al., 2023), highlighting a key challenge compared to the tabular setting (Chen et al., 2022). Hence, this necessitates a different approach tailored to this setting.

To improve the regret bound, our technique—inspired by Hong et al. (2024)—focuses on the boundedness condition of V_h^*, while still estimating the sequence $\{V_h^*\}_{h=1}^H$. In particular, by leveraging Lemma 2, we have $\mathrm{sp}(V_h^*) \le 2\zeta$ for some ζ that is independent of H, whereas Ghosh et al. (2023) relies on the trivial bound $\|V_h^*\|_\infty \le H$. Here, ζ is a quantity related to the span of bias functions, which will be specified later in Section 4.2. This suggests replacing the $\mathcal{O}(H)$ dependence in the regret bound with $\mathcal{O}(\zeta)$, yielding a significant improvement, since H may depend on T, whereas ζ does not. Consequently, our refined regret bound arises from exploiting the bound $\mathrm{sp}(V_h^*) \le 2\zeta$, while still estimating the optimal value functions $\{V_h^*\}_{h=1}^H$.

To incorporate the described technique into the constrained setting, a natural approach is to consider a primal–dual extension of the algorithm for the unconstrained setting (Hong et al., 2024). However, analyzing the resulting algorithm becomes nontrivial, as it requires additional tools. Importantly, our strong duality result plays a key role. To be more specific, our analysis based on Theorem 1 provides a new tool by establishing a lower bound on $\mathrm{Regret}(T)$. This, in turn, allows us to derive separate bounds on $\mathrm{Regret}(T)$ and $\mathrm{Violation}(T)$ from the composite Lagrangian regret $\mathrm{Regret}(T) + \lambda\mathrm{Violation}(T)$. In Section 4.3, we provide additional details.

Algorithm 1 PD-LSCVI-UCB

Input: horizon $H = T^{1/3}$; $K = T^{2/3}$; clipping parameter ζ; bonus factor β; Slater constant γ; step size η;

Initialize: $\mathbf{\Sigma}_1 = I$; $\lambda_1 = 0$; $V_{H+1}^k(\cdot) \leftarrow 0$ for all $k \in [K]$;

1: Receive state s_1.
2: **for** $k = 1, \ldots, K$ **do**
3: **for** $h = H, \ldots, 1$ **do**
4: $\mathbf{w}_{h+1}^k \leftarrow \mathbf{\Sigma}_k^{-1} \sum_{\tau=1}^{k-1} \sum_{j=1}^{H} \phi(s_j^\tau, a_j^\tau)(V_{h+1}^k(s_{j+1}^\tau) - \min_{s'} V_{h+1}^k(s'))$
5: $Q_h^k(\cdot,\cdot) \leftarrow \left(r(\cdot,\cdot) + \lambda_k g(\cdot,\cdot) + \phi(\cdot,\cdot)^\top \mathbf{w}_{h+1}^k + \min_{s'} V_{h+1}^k(s') + \beta \|\phi(\cdot,\cdot)\|_{\mathbf{\Sigma}_k^{-1}}\right) \wedge \left(1 + \frac{2}{\gamma}\right) H$
6: $\widetilde{V}_h^k(\cdot) \leftarrow \max_a Q_h^k(\cdot, a)$
7: $V_h^k(\cdot) \leftarrow \widetilde{V}_h^k(\cdot) \wedge (\min_{s'} \widetilde{V}_h^k(s') + 2\zeta)$
8: **end for**
9: **for** $h = 1, \ldots, H$ **do**
10: Take $a_h^k \in \arg\max_a Q_h^k(s_h^k, a)$ and observe $s_{h+1}^k \sim P(\cdot|s_h^k, a_h^k)$
11: **end for**
12: $s_1^{k+1} \leftarrow s_{H+1}^k$
13: $\mathbf{\Sigma}_{k+1} \leftarrow \mathbf{\Sigma}_k + \sum_{h=1}^{H} \phi(s_h^k, a_h^k)\phi(s_h^k, a_h^k)^\top$
14: $\lambda_{k+1} \leftarrow \left[\lambda_k + \eta\left(Hb - \sum_{h=1}^{H} g(s_h^k, a_h^k)\right)\right]_{[0, \frac{2}{\gamma}]}$
15: **end for**

4.2 Proposed algorithm

We present *primal–dual least-squares clipped value iteration with upper confidence bound* (PD-LSCVI-UCB, Algorithm 1). The algorithm adopts the finite-horizon approximation framework with K episodes and the horizon H. For brevity, let $s_h^k = s_{(k-1)H+h}$, $a_h^k = a_{(k-1)H+h}$, and let $H = T^{1/3},\ K = T^{2/3}$ be an integer. In each episode $k \in [K]$, $Q_h^k, \widetilde{V}_h^k$ are computed via backward induction from $h = H$ to 1, providing optimistic estimates under a composite reward function $r + \lambda_k g$, where λ_k denotes the dual variable in episode k. The clipped value function estimate V_h^k is then computed. The dual variable is updated based on the cumulative constraint violation in the episode, i.e., $Hb - \sum_{h=1}^{H} g(s_h^k, a_h^k)$, and clipped to the interval $[0, 2/\gamma]$. Here, the clipping parameter $2/\gamma$ is chosen because it is a strict upper bound on the optimal dual variable, as shown in the following lemma.

Lemma 3. *Suppose that Assumption 1 holds. Then* $\lambda^* \leq 1/\gamma$.

Value function clipping (Line 7) This step is crucial for exploiting the boundedness of the span of optimal value functions. The key intuition is that if the optimal value functions have span bounded by 2ζ for some ζ, then the value function estimates can also be restricted to this range. Hence, even when the estimates are clipped at 2ζ, they continue to overestimate the optimal value functions, ensuring sufficient yet tight exploration.

To successfully implement value function clipping, we assume the existence of an optimal stationary policy with constant gain. This assumption is required to apply Lemma 2, and has also been adopted in prior works (Chen et al., 2022; Ghosh et al., 2023). On the other hand, weakly communicating average-reward CMDPs need not admit an optimal policy; a detailed discussion is provided in Appendix B. This highlights a key difference from the unconstrained setting, where an optimal stationary policy with constant gain always exists. For clarity, we formalize this via the following assumption.

Assumption 2. *An optimal policy* π^* *for* (CMDP) *exists. Its gains* $J_r^*(s)$ *and* $J_g^*(s)$ *are constant across* $s \in \mathcal{S}$.[5] *Moreover, the spans of its state bias functions,* $\mathrm{sp}(v_r^*)$ *and* $\mathrm{sp}(v_g^*)$*, are known.*

[5]While we assume the existence of such π^*, as in the previous works, we introduce an alternative approach in Appendix B for cases where such a policy does not exist.

Building on this assumption, we define the clipping parameter ζ as

$$\zeta \triangleq \mathrm{sp}(v_r^*) + \frac{2}{\gamma}\mathrm{sp}(v_g^*). \tag{2}$$

Combining this with Lemma 2, we obtain

$$\mathrm{sp}(V_{r,h}^*) \le \max_{s,s'} \left|V_{r,h}^*(s) - (H-h+1)J_r^*\right| + \left|V_{r,h}^*(s') - (H-h+1)J_r^*\right| \le 2\mathrm{sp}(v_r^*).$$

The same argument applies to g. As a consequence, since $\lambda_k \in [0, 2/\gamma]$, we have

$$\mathrm{sp}\left(V_{r,h}^* + \lambda_k V_{g,h}^*\right) \le \mathrm{sp}(V_{r,h}^*) + \lambda_k \mathrm{sp}(V_{g,h}^*) \le 2\zeta. \tag{3}$$

As required, 2ζ serves as an upper bound on the span of the optimal value functions, motivating us to clip $\widetilde{V}_h^k$ at this level. Finally, PD-LSCVI-UCB achieves the following results.

Theorem 2. *Let $\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, r, g, b)$ be a weakly communicating linear CMDP with $|\mathcal{S}|, |\mathcal{A}| < \infty$, and let $s_1 \in \mathcal{S}$ be any initial state. Suppose that Assumptions 1, 2 hold. Let $K = T^{2/3}, H = T^{1/3}, \beta = \widetilde{\mathcal{O}}((\zeta+1)d), \eta = 2/(\gamma\sqrt{KH^2})$, and let $\mathrm{sp}(v_{\lambda^*}^*)$ be defined in Lemma 7. With probability at least $1-3\delta$, PD-LSCVI-UCB (Algorithm 1) guarantees*

$$\mathrm{Regret}(T) = \widetilde{\mathcal{O}}\left(\left(\zeta + \frac{2}{\gamma}\right) T^{2/3} + (\zeta+1)(\sqrt{d^3T} + d^2T^{1/3})\right),$$

$$\mathrm{Violation}(T) = \widetilde{\mathcal{O}}\left((\gamma\zeta + 2)\, T^{2/3} + \gamma(\zeta+1)(\sqrt{d^3T} + d^2T^{1/3}) + \gamma\mathrm{sp}(v_{\lambda^*}^*)\sqrt{T}\right),$$

where $\widetilde{\mathcal{O}}(\cdot)$ hides polynomial factors in $\log(dT|\mathcal{A}|\zeta/(\delta\gamma))$.

Comparisons with prior works We provide a comparison with prior algorithms for learning average-reward linear MDPs. Hong et al. (2024) proposed a clipped value iteration algorithm for learning average-reward unconstrained linear MDPs. Beyond the primal–dual extension, there is another key difference. Their work employs a discounted approximation framework, whereas ours adopts a finite-horizon approximation. Under the discounted approximation, the horizon can increase up to $H = \mathcal{O}(T)$, as it is determined by the doubling trick (see Line 5 of their algorithm). However, this becomes problematic in the constrained setting, since $\lambda_{k+1} - \lambda_k$ can also grow up to $\mathcal{O}(\eta T)$, which makes the dual variable unstable. To overcome this, we adopt a finite-horizon approximation with $H = T^{1/3}$ and show that their technique extends to this framework.

Ghosh et al. (2023) proposed a primal–dual algorithm for learning average-reward linear CMDPs. While their algorithm is also based on the finite-horizon approximation, it does not employ value function clipping (Line 7), and thus the tighter span bound (3) is not fully leveraged. Moreover, their analysis relies on strong duality for finite-horizon CMDPs, resulting in regret bounds of $\widetilde{\mathcal{O}}(T^{3/4})$. In contrast, our method employs value function clipping, and its analysis is enabled by strong duality for average-reward CMDPs (Theorem 1), yielding improved regret bounds of $\widetilde{\mathcal{O}}(T^{2/3})$. Indeed, these bounds match those of the tabular setting in terms of T (Chen et al., 2022).

4.3 Outline of analysis

The overall strategy for the analysis consists of the following two steps. In the first step, we derive an upper bound Δ on the composite Lagrangian regret $\mathrm{Regret}(T) + \lambda\mathrm{Violation}(T)$ that holds for any $\lambda \in [0, 2/\gamma]$. This immediately yields an upper bound on $\mathrm{Regret}(T)$ by taking $\lambda = 0$. On the other hand, bounding $\mathrm{Violation}(T)$ requires a lower bound on $\mathrm{Regret}(T)$, since $\mathrm{Violation}(T) \le \frac{\gamma}{2}(\Delta - \mathrm{Regret}(T))$. Therefore, in the second step, we obtain such a lower bound by leveraging strong duality.

Step 1: Upper bound on $\mathrm{Regret}(T) + \lambda\mathrm{Violation}(T)$ To upper bound the composite Lagrangian regret, we first provide the following useful lemmas. In particular, Lemma 4 captures the estimation error and characterizes the optimistic bonus. Then, Lemma 5 shows that the clipped value function estimates optimistically the composite value functions.

Lemma 4. *Let $\beta = \widetilde{\mathcal{O}}((\zeta+1)d)$. With probability at least $1-\delta$, for $(s,a,h,k) \in \mathcal{S}\times\mathcal{A}\times[H]\times[K]$,*

$$|\phi(s,a)^\top(\mathbf{w}_{h+1}^k - \mathbf{w}_{h+1}^{k,*})| \le \beta\|\phi(s,a)\|_{\mathbf{\Sigma}_k^{-1}}.$$

where $\mathbf{w}_{h+1}^{k,*} = \sum_s \psi(s)(V_{h+1}^k(s) - \min_{s'} V_{h+1}^k(s'))$.

Lemma 5. *Suppose that Lemma 4 holds. Then for all* $(s, h, k) \in \mathcal{S} \times [H] \times [K]$,

$$V^*_{r,h}(s) + \lambda_k V^*_{g,h}(s) \leq V^k_h(s).$$

Given these lemmas, we can decompose $\mathrm{Regret}(T) + \lambda \mathrm{Violation}(T)$ as follows.

$$\begin{aligned}\mathrm{Regret}(T) + \lambda \mathrm{Violation}(T) \leq K\zeta + \sum_{k=1}^{K}\sum_{h=1}^{H} 2\beta \|\phi(s^k_h, a^k_h)\|_{\mathbf{\Sigma}_k^{-1}} \\ + \sum_{k=1}^{K}\sum_{h=1}^{H} \left(PV^k_{h+1}(s^k_h, a^k_h) - V^k_{h+1}(s^k_{h+1})\right) + \frac{\lambda^2}{2\eta} + \frac{\eta K H^2}{2}.\end{aligned} \tag{4}$$

Here, the second term can be bounded by the elliptical potential lemma, and the third term is bounded, as it is a sum of martingale differences. Finally, with the parameter choice given in Theorem 2, we derive the desired bound, as stated in the following lemma.

Lemma 6. *Suppose that Lemma 4 holds. With probability at least* $1 - \delta$*, for any* $\lambda \in [0, 2/\gamma]$,

$$\mathrm{Regret}(T) + \lambda \mathrm{Violation}(T) = \widetilde{\mathcal{O}}\left(\left(\zeta + \frac{2}{\gamma}\right) T^{2/3} + (\zeta + 1)(\sqrt{d^3 T} + d^2 T^{1/3})\right).$$

Step 2: Lower bound on $\mathrm{Regret}(T)$ We achieve this by leveraging Theorem 1, together with the Bellman optimality equation for weakly communicating average-reward unconstrained MDPs (Puterman, 1994). First, we consider a bounded optimal dual variable λ^*, which is guaranteed by Theorem 1 and Lemma 3. For a reward $r + \lambda^*(g - b)$, the Bellman optimality condition for weakly communicating unconstrained MDPs implies that there exists $J^*_{\lambda^*}(s_1) \in \mathbb{R}$, $v^*_{\lambda^*} : \mathcal{S} \to \mathbb{R}$ such that for any $(s, a) \in \mathcal{S} \times \mathcal{A}$,

$$J^*_{\lambda^*}(s_1) - r(s, a) \geq -\lambda^*(b - g(s, a)) + Pv^*_{\lambda^*}(s, a) - v^*_{\lambda^*}(s). \tag{5}$$

Given $\mathrm{sp}(v^*_{\lambda^*})$, we are now ready to state a lemma about a lower bound on $\mathrm{Regret}(T)$.

Lemma 7. *Let* $v^*_{\lambda^*}$ *be defined in* (5)*. With probability at least* $1 - \delta$,

$$\mathrm{Regret}(T) \geq -\lambda^* \mathrm{Violation}(T) - \mathrm{sp}(v^*_{\lambda^*})\sqrt{2T \log(1/\delta)} - \mathrm{sp}(v^*_{\lambda^*}).$$

The proof proceeds as follows. When the state and action spaces are finite, $J^\pi_{r+\lambda^*(g-b)}(s_1)$ can be decomposed as $J^\pi_r(s_1) + \lambda^*(J^\pi_g(s_1) - b)$. Therefore, maximizing $J^\pi_{r+\lambda^*(g-b)}(s_1)$ over $\pi \in \Pi$ coincides with $D(\lambda^*)$. By strong duality, we deduce that $J^*_r(s_1) = D(\lambda^*) = J^*_{\lambda^*}(s_1)$.

Now, recall the Bellman optimality condition (5). Since we showed that $J^*_r(s_1) = J^*_{\lambda^*}(s_1)$ by strong duality, summing the left-hand side along the trajectory recovers $\mathrm{Regret}(T)$, while the right-hand side provides its lower bound. To formalize this, given the trajectory $\{(s_t, a_t)\}_{t=1}^T$, we substitute $s = s_t, a = a_t$, and sum over $t = 1$ to T. After further computation, we obtain the desired lower bound on $\mathrm{Regret}(T)$.

5 Conclusion

This paper studied weakly communicating average-reward CMDPs and established strong duality under this general setting, despite the absence of an LP formulation and the resulting nonconvexity. The key insight is that the geometric structure of the occupation measure set is sufficient to recover strong duality. Building on this, we proposed a primal–dual algorithm for linear CMDPs that achieves $\widetilde{\mathcal{O}}(T^{2/3})$ regret and constraint violation bounds by combining clipped value iteration with a strong duality–based analysis to control the composite Lagrangian regret and its decomposition. Despite these results, several important directions remain open. First, it is of interest to extend the strong duality result to infinite state and action spaces, as the current approach relies on finite-dimensional LP arguments, e.g., Lemmas 12 and 13 in the appendix. Second, it remains open whether the regret and constraint violation bounds can be further improved to $\widetilde{\mathcal{O}}(\sqrt{T})$, which remains unresolved for both tabular and linear average-reward CMDPs under the weakly communicating assumption.

Acknowledgments and Disclosure of Funding

This work was supported by the New Faculty Startup Fund from Seoul National University. The authors declare no competing interests.

References

Achiam, J., Held, D., Tamar, A., and Abbeel, P. (2017). Constrained policy optimization. In *International conference on machine learning*, pages 22–31. PMLR.

Agarwal, M., Bai, Q., and Aggarwal, V. (2021). Markov decision processes with long-term average constraints. *arXiv preprint arXiv:2106.06680*.

Agarwal, M., Bai, Q., and Aggarwal, V. (2022a). Concave utility reinforcement learning with zero-constraint violations. *Transactions on Machine Learning Research*.

Agarwal, M., Bai, Q., and Aggarwal, V. (2022b). Regret guarantees for model-based reinforcement learning with long-term average constraints. In *Uncertainty in Artificial Intelligence*, volume 180 of *Proceedings of Machine Learning Research*, pages 22–31. PMLR.

Agrawal, P. and Agrawal, S. (2024). Optimistic q-learning for average reward and episodic reinforcement learning. *arXiv preprint arXiv:2407.13743*.

Agrawal, S. and Jia, R. (2017). Optimistic posterior sampling for reinforcement learning: Worst-case regret bounds. In *Advances in Neural Information Processing Systems*, volume 30, pages 1184–1194.

Altman, E. (1999). *Constrained Markov Decision Processes*, volume 7. CRC Press.

Amani, S., Thrampoulidis, C., and Yang, L. (2021). Safe reinforcement learning with linear function approximation. In *International Conference on Machine Learning*, pages 243–253. PMLR.

Bai, Q., Mondal, W. U., and Aggarwal, V. (2024a). Learning general parameterized policies for infinite horizon average reward constrained mdps via primal-dual policy gradient algorithm. *Advances in Neural Information Processing Systems*, 37:108566–108599.

Bai, Q., Mondal, W. U., and Aggarwal, V. (2024b). Regret analysis of policy gradient algorithm for infinite horizon average reward markov decision processes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 10980–10988.

Bartlett, P. L. and Tewari, A. (2012). Regal: A regularization based algorithm for reinforcement learning in weakly communicating mdps. *arXiv preprint arXiv:1205.2661*.

Bertsekas, D. (2009). *Convex optimization theory*, volume 1. Athena Scientific.

Bura, A., HasanzadeZonuzy, A., Kalathil, D., Shakkottai, S., and Chamberland, J.-F. (2022). Dope: Doubly optimistic and pessimistic exploration for safe reinforcement learning. *Advances in neural information processing systems*, 35:1047–1059.

Chen, L., Jain, R., and Luo, H. (2022). Learning infinite-horizon average-reward markov decision process with constraints. In *International Conference on Machine Learning*, pages 3246–3270. PMLR.

Coronato, A., Naeem, M., De Pietro, G., and Paragliola, G. (2020). Reinforcement learning for intelligent healthcare applications: A survey. *Artificial intelligence in medicine*, 109:101964.

Dai, J., Pan, X., Sun, R., Ji, J., Xu, X., Liu, M., Wang, Y., and Yang, Y. (2023). Safe rlhf: Safe reinforcement learning from human feedback. *arXiv preprint arXiv:2310.12773*.

Ding, D., Wei, X., Yang, Z., Wang, Z., and Jovanovic, M. (2021). Provably efficient safe exploration via primal-dual policy optimization. In *International conference on artificial intelligence and statistics*, pages 3304–3312. PMLR.

Durrett, R. (2019). *Probability: theory and examples*, volume 49. Cambridge university press.

Efroni, Y., Mannor, S., and Pirotta, M. (2020). Exploration-exploitation in constrained mdps. *arXiv preprint arXiv:2003.02189*.

Fruit, R., Pirotta, M., Lazaric, A., and Ortner, R. (2018). Efficient bias-span-constrained exploration-exploitation in reinforcement learning. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 1578–1586. PMLR.

Ganesh, S. and Aggarwal, V. (2025). Regret analysis of average-reward unichain mdps via an actor-critic approach. *arXiv preprint arXiv:2505.19986*.

Ganesh, S., Mondal, W. U., and Aggarwal, V. (2024). Order-optimal regret with novel policy gradient approaches in infinite-horizon average reward mdps. *arXiv preprint arXiv:2404.02108*.

García, J. and Fernández, F. (2015). A comprehensive survey on safe reinforcement learning. *Journal of Machine Learning Research*, 16(42):1437–1480.

Ghosh, A., Zhou, X., and Shroff, N. (2022). Provably efficient model-free constrained rl with linear function approximation. *Advances in Neural Information Processing Systems*, 35:13303–13315.

Ghosh, A., Zhou, X., and Shroff, N. (2023). Achieving sub-linear regret in infinite horizon average reward constrained mdp with linear function approximation. In *The Eleventh International Conference on Learning Representations*.

He, J., Zhou, D., and Gu, Q. (2022). Near-optimal policy optimization algorithms for learning adversarial linear mixture mdps. In *International Conference on Artificial Intelligence and Statistics*, pages 4259–4280. PMLR.

Hong, K., Chae, W., Zhang, Y., Lee, D., and Tewari, A. (2024). Reinforcement learning for infinite-horizon average-reward linear mdps via approximation by discounted-reward mdps. *arXiv preprint arXiv:2405.15050*.

Hong, K. and Tewari, A. (2025). A computationally efficient algorithm for infinite-horizon average-reward linear mdps. *arXiv preprint arXiv:2504.11997*.

Isele, D., Nakhaei, A., and Fujimura, K. (2018). Safe reinforcement learning on autonomous vehicles. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1–6. IEEE.

Jaksch, T., Ortner, R., and Auer, P. (2010). Near-optimal regret bounds for reinforcement learning. *Journal of Machine Learning Research*, 11(51):1563–1600.

Jin, C., Yang, Z., Wang, Z., and Jordan, M. I. (2020). Provably efficient reinforcement learning with linear function approximation. In *Conference on learning theory*, pages 2137–2143. PMLR.

Kitamura, T., Ghosh, A., Kozuno, T., Kumagai, W., Kasaura, K., Hoshino, K., Hosoe, Y., and Matsuo, Y. (2025). Provably efficient rl under episode-wise safety in constrained mdps with linear function approximation. *arXiv preprint arXiv:2502.10138*.

Lee, J., Bravo, M., and Cominetti, R. (2025). Near-optimal sample complexity for mdps via anchoring. *arXiv preprint arXiv:2502.04477*.

Lee, J. and Ryu, E. (2023). Accelerating value iteration with anchoring. *Advances in Neural Information Processing Systems*, 36:53924–53963.

Lee, J. and Ryu, E. K. (2025). Optimal non-asymptotic rates of value iteration for average-reward markov decision processes. *arXiv preprint arXiv:2504.09913*.

Lee, J. and Ryu, E. K. (2026). Policy gradient algorithms in average-reward multichain mdps. *arXiv preprint arXiv:2602.18003*.

Liu, T., Zhou, R., Kalathil, D., Kumar, P., and Tian, C. (2021). Learning policies with zero or bounded constraint violation for constrained mdps. *Advances in Neural Information Processing Systems*, 34:17183–17193.

Mannor, S. and Tsitsiklis, J. N. (2005). On the empirical state-action frequencies in markov decision processes under general policies. *Mathematics of Operations Research*, 30(3):545–561.

Müller, A., Alatur, P., Cevher, V., Ramponi, G., and He, N. (2024). Truly no-regret learning in constrained mdps. In *International Conference on Machine Learning*, pages 36605–36653. PMLR.

Norris, J. R. (1998). *Markov chains*. Cambridge university press.

Ouyang, Y., Gagrani, M., Nayyar, A., and Jain, R. (2017). Learning unknown markov decision processes: A thompson sampling approach. In *Advances in Neural Information Processing Systems*, volume 30, pages 1333–1342.

Paternain, S., Chamon, L., Calvo-Fullana, M., and Ribeiro, A. (2019). Constrained reinforcement learning has zero duality gap. *Advances in Neural Information Processing Systems*, 32.

Provodin, D., Kaptein, M. C., and Pechenizkiy, M. (2024). Efficient exploration in average-reward constrained reinforcement learning: Achieving near-optimal regret with posterior sampling. In Salakhutdinov, R., Kolter, Z., Heller, K., Weller, A., Oliver, N., Scarlett, J., and Berkenkamp, F., editors, *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 41144–41162. PMLR.

Puterman, M. L. (1994). Markov decision processes: Discrete stochastic dynamic programming.

Qiu, S., Wei, X., Yang, Z., Ye, J., and Wang, Z. (2020). Upper confidence primal-dual reinforcement learning for cmdp with adversarial loss. *Advances in Neural Information Processing Systems*, 33:15277–15287.

Rockafellar, R. T. (1970). *Convex Analysis*. Princeton University Press.

Roknilamouki, A., Ghosh, A., Shi, M., Nourzad, F., Ekici, E., and Shroff, N. (2025). Provably efficient RL for linear MDPs under instantaneous safety constraints in non-convex feature spaces. In *Forty-second International Conference on Machine Learning*.

Satheesh, A. and Aggarwal, V. (2026). Regret analysis of unichain average reward constrained mdps with general parameterization. *arXiv preprint arXiv:2602.08000*.

Satheesh, A., Barman, P. K., Mondal, W. U., and Aggarwal, V. (2026). Global convergence of average reward constrained mdps with neural critic and general policy parameterization. *arXiv preprint arXiv:2603.07698*.

Satheesh, A., Sathish, S., Ganesh, S., Powell, K., and Aggarwal, V. (2025). Primal-only actor critic algorithm for robust constrained average cost mdps. *arXiv preprint arXiv:2511.05758*.

Singh, R., Gupta, A., and Shroff, N. B. (2022). Learning in constrained markov decision processes. *IEEE Transactions on Control of Network Systems*, 10(1):441–453.

Stradi, F. E., Castiglioni, M., Marchesi, A., and Gatti, N. (2025a). Learning adversarial MDPs with stochastic hard constraints. In *Forty-second International Conference on Machine Learning*.

Stradi, F. E., Castiglioni, M., Marchesi, A., and Gatti, N. (2025b). Optimal strong regret and violation in constrained MDPs via policy optimization. In *The Thirteenth International Conference on Learning Representations*.

Stradi, F. E., Germano, J., Genalti, G., Castiglioni, M., Marchesi, A., and Gatti, N. (2024). Online learning in cmdps: Handling stochastic and adversarial constraints. In *Forty-first International Conference on Machine Learning*.

Stradi, F. E., Lunghi, A., Castiglioni, M., Marchesi, A., and Gatti, N. (2025c). Policy optimization for cmdps with bandit feedback: Learning stochastic and adversarial constraints. In *Forty-second International Conference on Machine Learning*.

Wei, C.-Y., Jahromi, M. J., Luo, H., and Jain, R. (2021). Learning infinite-horizon average-reward mdps with linear function approximation. In *International Conference on Artificial Intelligence and Statistics*, pages 3007–3015. PMLR.

Wei, C.-Y., Jahromi, M. J., Luo, H., Sharma, H., and Jain, R. (2020). Model-free reinforcement learning in infinite-horizon average-reward markov decision processes. In *International conference on machine learning*, pages 10170–10180. PMLR.

Wei, H., Ghosh, A., Shroff, N., Ying, L., and Zhou, X. (2023). Provably efficient model-free algorithms for non-stationary cmdps. In *International Conference on Artificial Intelligence and Statistics*, pages 6527–6570. PMLR.

Wei, H., Liu, X., and Ying, L. (2022). A provably-efficient model-free algorithm for infinite-horizon average-reward constrained markov decision processes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 3868–3876.

Wei, H., Liu, X., and Ying, L. (2024). Safe reinforcement learning with instantaneous constraints: the role of aggressive exploration. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 21708–21716.

Wei, Y., Li, X., and Yang, L. F. (2026). Near-optimal sample complexity bounds for constrained average-reward MDPs. In *The Fourteenth International Conference on Learning Representations*.

Xu, Y., Ganesh, S., Mondal, W. U., Bai, Q., and Aggarwal, V. (2025). Global convergence for average reward constrained mdps with primal-dual actor critic algorithm. *arXiv preprint arXiv:2505.15138*.

Yu, K., Bae, S., and Lee, D. (2026a). Near-optimal primal-dual algorithm for learning linear mixture cmdps with adversarial rewards. *arXiv preprint arXiv:2603.27884*.

Yu, K., Bae, S., and Lee, D. (2026b). Primal-dual policy optimization for linear CMDPs with adversarial losses. In *The Fourteenth International Conference on Learning Representations*.

Yu, K., Lee, D., Overman, W., and Lee, D. (2025). Improved regret bound for safe reinforcement learning via tighter cost pessimism and reward optimism. *Reinforcement Learning Journal*, 6:493–546.

Zhang, Z., Lee, J. D., Chen, Y., and Du, S. S. (2024). Horizon-free regret for linear markov decision processes. In *The Twelfth International Conference on Learning Representations*.

Zhang, Z. and Xie, Q. (2023). Sharper model-free reinforcement learning for average-reward markov decision processes. In *The Thirty Sixth Annual Conference on Learning Theory*, pages 5476–5477. PMLR.

Zhu, J., Yu, K., Lee, D., Liu, X., and Wei, H. (2025). An optimistic algorithm for online CMDPS with anytime adversarial constraints. In *Forty-second International Conference on Machine Learning*.

Zurek, M. and Chen, Y. (2024). Span-based optimal sample complexity for weakly communicating and general average reward mdps. In *Advances in Neural Information Processing Systems*, volume 37, pages 33455–33504.

A Related work

Average-reward CMDPs Learning in average-reward CMDPs has been studied under several algorithmic frameworks and structural assumptions. Early works considered tabular CMDPs with long-term average constraints and established regret and constraint-violation guarantees (Agarwal et al., 2021, 2022b; Wei et al., 2022; Singh et al., 2022; Chen et al., 2022). In particular, Chen et al. (2022) provided algorithms under both ergodic and weakly communicating assumptions, while Ghosh et al. (2023) extended this line to linear function approximation. More recently, Provodin et al. (2024) studied posterior-sampling-based exploration for average-reward CMDPs, and Wei et al. (2026) investigated near-optimal sample complexity bounds under a generative model. Related constrained formulations have also been studied through concave utility objectives and zero-violation guarantees (Agarwal et al., 2022a). Another line of work studies average-reward CMDPs through policy-gradient and actor–critic methods. Bai et al. (2024a) considered general policy parameterization and developed a primal–dual policy-gradient algorithm. Subsequent works further studied convergence, regret, and actor–critic variants under different parameterization and structural assumptions (Xu et al., 2025; Satheesh et al., 2025; Satheesh and Aggarwal, 2026; Satheesh et al., 2026). These works provide important progress for average-reward CMDPs, but they typically rely on stronger assumptions such as ergodicity or unichain structure, or focus on different learning models. In contrast, our work establishes strong duality under the weakly communicating assumption and uses it to obtain improved regret and constraint violation bounds for average-reward linear CMDPs.

Average-reward unconstrained MDPs There is also a large literature on learning unconstrained average-reward MDPs. In the tabular setting, Jaksch et al. (2010) proposed UCRL2 for communicating MDPs, while posterior-sampling approaches were studied in communicating and weakly communicating settings (Agrawal and Jia, 2017; Ouyang et al., 2017). For weakly communicating MDPs, Bartlett and Tewari (2012) introduced REGAL using bias-span regularization, and Fruit et al. (2018) developed a computationally efficient bias-span-constrained algorithm. Model-free algorithms for average-reward MDPs were studied by Wei et al. (2020); Zhang and Xie (2023), and recent works further investigate planning and sample-complexity aspects of average-reward MDPs (Agrawal and Agrawal, 2024; Zurek and Chen, 2024; Lee and Ryu, 2023; Lee et al., 2025; Lee and Ryu, 2025). For linear function approximation, Wei et al. (2021) proposed algorithms for infinite-horizon average-reward linear MDPs, including both statistically optimal and computationally efficient variants under different assumptions. More recently, Hong et al. (2024) studied average-reward linear MDPs via approximation by discounted-reward MDPs and introduced value function clipping to obtain sharp regret guarantees, while Hong and Tewari (2025) further improved computational efficiency by avoiding dependence on the state-space size in the clipping step. Policy-gradient methods have also been studied for unconstrained average-reward MDPs (Bai et al., 2024b; Ganesh et al., 2024; Ganesh and Aggarwal, 2025), as well as extensions beyond the unichain setting (Lee and Ryu, 2026). Our algorithm is inspired by the value-function clipping technique developed for unconstrained average-reward linear MDPs (Hong et al., 2024), but extending this technique to CMDPs requires a primal–dual formulation and a strong-duality-based analysis.

Finite-horizon CMDPs Compared with the average-reward setting, finite-horizon CMDPs have been studied under a broader variety of models, feedback structures, and safety requirements. In the tabular setting, Efroni et al. (2020) studied the exploration–exploitation tradeoff in CMDPs through both LP-based and primal–dual approaches. Subsequent works refined the guarantees by considering zero or bounded constraint violation (Liu et al., 2021; Bura et al., 2022; Yu et al., 2025), and optimal strong regret and violation (Müller et al., 2024; Stradi et al., 2025b). Another line of work studies online CMDPs with stochastic or adversarial losses and constraints (Qiu et al., 2020; Stradi et al., 2024, 2025c,a; Zhu et al., 2025). Under function approximation, Ding et al. (2021) and Ghosh et al. (2022) developed provably efficient algorithms for finite-horizon linear CMDPs, and subsequent works further studied episode-wise safety, instantaneous constraints, non-stationarity, and adversarial losses in linear or related CMDP models (Amani et al., 2021; Wei et al., 2023, 2024; Kitamura et al., 2025; Roknilamouki et al., 2025; Yu et al., 2026b,a). These works focus on finite-horizon or episodic CMDPs, whereas our work studies infinite-horizon average-reward CMDPs under the weakly communicating assumption.

B Additional discussions

Absence of optimal stationary policies We explain why (CMDP) may fail to admit an optimal stationary policy under the weakly communicating assumption for a fixed initial state s_1. Equivalently, it suffices to show that (Occ-CMDP) may have a finite supremum that is not attained. The main reason is that $\widehat{\mathcal{P}}(s_1)$ may miss some boundary points (Lemma 1). As a result, the feasible set $\widehat{\mathcal{P}}(s_1) \cap \{q : g^\top q \geq b\}$ may exclude some extreme points for certain $g \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{A}|}$ and $b \in \mathbb{R}$. Consequently, for such g and b, the optimization problem may fail to attain its optimum for some $r \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{A}|}$. This phenomenon is illustrated in Figure 2.

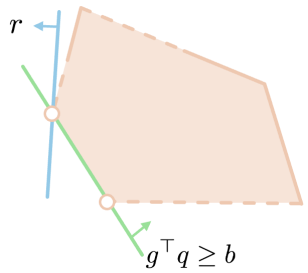
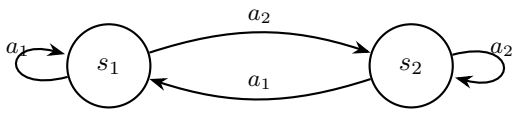


Figure 2: Illustration of an average-reward CMDP over the occupation measure space. The green line represents the constraint $g^\top q \geq b$, the red region represents the feasible region $\widehat{\mathcal{P}}(s_1) \cap \{q : g^\top q \geq b\}$, and the blue arrow indicates the direction of r.

We provide a more concrete example, inspired by Example 3.1 in Mannor and Tsitsiklis (2005). Let $\mathcal{S} = \{s_1, s_2\}$, $\mathcal{A} = \{a_1, a_2\}$, $b = 1/2$, and let $r(s_2, a_2) = 1$, $g(s_1, a_1) = 1$, otherwise 0. The transition kernel is deterministic, which is illustrated as follows.



Note that this MDP is weakly communicating because s_1, s_2 become recurrent under a policy taking a_2 at s_1, and a_1 at s_2. Given these, we consider the following average-reward CMDPs:

$$\sup_{\pi\in\Pi} J_r^\pi(s_1) \quad \text{s.t.} \quad J_g^\pi(s_1) \geq 1/2. \tag{6}$$

To find an optimal stationary policy, consider $\pi(a_2|s_1) = \alpha$ and $\pi(a_1|s_2) = \beta$ for some $\alpha, \beta \in [0, 1]$. Suppose that $\alpha, \beta > 0$. In this case, the resulting Markov chain is

$$P_\pi = \begin{bmatrix} 1-\alpha & \alpha \\ \beta & 1-\beta \end{bmatrix}.$$

Since it is an irreducible chain, its (unique) stationary distribution satisfies $d_\pi(s_1) = \beta/(\alpha+\beta)$ and $d_\pi(s_2) = \alpha/(\alpha+\beta)$. By Proposition 8.9.1 in Puterman (1994), we can deduce that $q_{s_1}^\pi(s, a) = d_\pi(s)\pi(a|s)$ for $(s, a) \in \mathcal{S} \times \mathcal{A}$. It follows that

$$q_{s_1}^\pi(s_1, a_1) = \frac{\beta(1-\alpha)}{\alpha+\beta}, \; q_{s_1}^\pi(s_2, a_2) = \frac{\alpha(1-\beta)}{\alpha+\beta}.$$

To achieve the supremum, for $\varepsilon \in (0, 1/3)$, take

$$\alpha = \varepsilon, \quad \beta = \frac{\varepsilon}{1-2\varepsilon}.$$

Then, π is feasible, as $g^\top q_{s_1}^\pi = 1/2$ for any $\varepsilon \in (0, 1/3)$, and

$$r^\top q_{s_1}^\pi = \frac{1}{2} - \frac{\varepsilon}{1-\varepsilon}.$$

As $\varepsilon \to 0+$, the supremum of (6) becomes $1/2$. However, there is no fixed $\alpha, \beta > 0$ that achieves $r^\top q^\pi_{s_1} = 1/2$ and $g^\top q^\pi_{s_1} \geq 1/2$, because $q^\pi_{s_1}(s_1, a_1) + q^\pi_{s_1}(s_2, a_2) < 1$ for any $\alpha, \beta > 0$. Moreover, this can be easily shown when $\alpha = 0$ or $\beta = 0$. Consequently, this example demonstrates that weakly communicating average-reward CMDPs can have no optimal stationary solutions.

Alternative approach when no π^* exists We provide an alternative approach for cases where (CMDP) admits no optimal stationary policy π^* whose gains are constant across all $s \in \mathcal{S}$, as assumed in Assumption 2. For clarification, as demonstrated earlier in this section, the weakly communicating assumption does not guarantee the existence of such a policy in the CMDP setting. In contrast, the weakly communicating assumption guarantees this property in the unconstrained MDP setting.

Next, we explain the alternative approach. The key idea is to use an ϵ-optimal feasible policy $\pi^{*,\epsilon}$ satisfying $J^*_r(s_1) - \epsilon \leq J^{\pi^{*,\epsilon}}_r(s_1)$ and $J^{\pi^{*,\epsilon}}_g(s_1) \geq b$, instead of an optimal policy π^*. Here, $\epsilon > 0$ is sufficiently small, such as $\epsilon \asymp 1/T$. The existence of such a policy is guaranteed for any $\epsilon > 0$ by considering a point in $\mathrm{relint}(\mathcal{P}) \subseteq \mathrm{relint}(\widehat{\mathcal{P}}(s_1))$ (see Figure 2). Moreover, Lemma 15 implies that the gains of $\pi^{*,\epsilon}$ are constant across all $s \in \mathcal{S}$. This observation allows us to modify the analysis by using $\pi^{*,\epsilon}$ as the benchmark instead of π^*. The resulting additional regret term is negligible due to the choice of ϵ. However, this approach introduces another requirement: the span $\mathrm{sp}(v^{\pi^{*,\epsilon}})$ must be known and bounded independently of T, since this quantity appears in the final regret and constraint violation bounds. In summary, the alternative approach removes the assumption that an optimal stationary policy with constant gains exists, at the cost of requiring knowledge of and a bound on $\mathrm{sp}(v^{\pi^{*,\epsilon}})$ for $\epsilon \asymp 1/T$.

Importance of strong duality over stationary policies We explain why strong duality over stationary policies is particularly useful for learning problems. Specifically, one may ask why we do not instead consider (CMDP) over the history-dependent randomized policy class, which is a more general formulation:

$$J^{*,\mathrm{HR}}_r(s_1) \quad \triangleq \quad \max_{\pi \in \Pi^{\mathrm{HR}}} \; J^\pi_r(s_1) \text{ s.t. } J^\pi_g(s_1) \geq b, \tag{HR-CMDP}$$

where Π^{HR} denotes the set of history-dependent policies whose occupation measures exist. One may then analyze regret and constraint violation defined with respect to $J^{*,\mathrm{HR}}_r(s_1)$ using strong duality over Π^{HR}. While this approach is possible, it prevents us from leveraging favorable properties specific to stationary policies, such as Lemma 2, thereby hindering tighter regret bounds. To resolve this issue, we must answer the following questions: (i) can we restrict the policy class to the stationary policy class Π without loss of optimality, and (ii) does strong duality hold over Π?

Our results in Section 3 answer these questions. In particular, we can show that $J^{*,\mathrm{HR}}_r(s_1) = J^*_r(s_1)$, where $J^*_r(s_1)$ denotes the supremum of (CMDP) over Π. This follows from Lemma 1 together with the fact that the occupation measure set induced by Π^{HR} coincides with $\mathcal{P}$ (Theorem 3.1 in Mannor and Tsitsiklis (2005)). Consequently, this justifies restricting the policy class from Π^{HR} to Π without loss of optimality, which in turn allows us to leverage favorable properties of stationary policies. Finally, Theorem 1 fills the remaining gap by establishing strong duality over Π.

Computational complexity We provide the computational complexity of Algorithm 1. Since our algorithm is a primal–dual extension of Algorithm 2 in Hong et al. (2024), its complexity is comparable to that of the original algorithm. The main difference arises from our use of the finite-horizon approximation. Specifically, in each episode $k \in [K]$, the algorithm performs $\mathcal{O}(H)$ value iterations. In each step $h \in [H]$, it computes $\min_{s'} \widetilde{V}^k_h(s')$, which requires $\mathcal{O}(d^2|\mathcal{S}||\mathcal{A}|)$ operations. In addition, other operations require $\mathcal{O}(d^2|\mathcal{A}|T)$ operations. Therefore, each episode requires $\mathcal{O}(Hd^2|\mathcal{S}||\mathcal{A}|T)$ operations, and the total computational complexity is $\mathcal{O}(d^2KH|\mathcal{S}||\mathcal{A}|T)$. Since $HK = T$, this simplifies to $\mathcal{O}(d^2|\mathcal{S}||\mathcal{A}|T^2)$.

Limitations of Algorithm 1 We discuss the limitations of Algorithm 1. First, its computational complexity depends on $|\mathcal{S}|$. In particular, the value function clipping step computes $\min_{s'} \widetilde{V}^k_h(s')$, requiring at least $|\mathcal{S}|$ operations, which can be prohibitive when the state space is large. Second, its regret bound is suboptimal, as it does not achieve $\widetilde{\mathcal{O}}(\sqrt{T})$, whereas optimal regret rates are attainable in the unconstrained setting (Hong et al., 2024; Hong and Tewari, 2025). In our case, the

suboptimal $\widetilde{\mathcal{O}}(T^{2/3})$ regret arises from controlling the dual variable. More precisely, in the regret decomposition (4), the relevant terms are

$$\frac{\lambda^2}{2\eta} + \frac{\eta K H^2}{2}.$$

Under the finite-horizon approximation framework, we have $KH = T$, and thus these terms can be written as $\frac{\lambda^2}{2\eta} + \frac{\eta HT}{2}$. Since H depends on T, obtaining a $\mathcal{O}(\sqrt{T})$ bound for these terms is nontrivial.

For the constrained setting, we illustrate the challenges in designing an optimal algorithm, while its computational complexity is independent of $|\mathcal{S}|$. A natural approach is to extend algorithms designed for the unconstrained setting to the constrained setting. Indeed, in the unconstrained setting, Hong and Tewari (2025) proposed an optimal algorithm. In particular, their algorithm proceeds as follows. To avoid $\mathcal{O}(|\mathcal{S}|)$ computational complexity, the algorithm clips the value function estimates using only the visited states, replacing $\min_{s'} \widetilde{V}_h^k(s')$ with $\min_{s \in \mathcal{S}_{\text{visit}}} \widetilde{V}_h^k(s)$, where $\mathcal{S}_{\text{visit}}$ denotes the set of visited states. While this modification reduces the computational cost, it introduces additional errors. To control this error, it applies a deviation control technique to ensure that $\|V_{h+1}^k - V_{h+1}^{k+1}\|_\infty$ remains small.

However, extending this approach to the constrained setting introduces additional challenges, primarily due to the presence of the dual variable. In particular, when adapting their algorithm to a primal–dual framework, we may introduce a dual variable λ_k, for example through the modified reward $r + \lambda_k g$. Since λ_k varies across iterations k, controlling the deviation term $\|V_{h+1}^k - V_{h+1}^{k+1}\|_\infty$ becomes significantly more difficult. Therefore, designing an optimal algorithm with $|\mathcal{S}|$-independent computational complexity in our setting remains an open problem.

C Additional preliminaries

We introduce some additional notions that are useful for proving strong duality and regret analysis. For more details, we refer the reader to Puterman (1994); Norris (1998).

Definitions for Markov chains Consider a Markov chain P_π on a finite state $\mathcal{S}$. Let $P_\pi^{(n)}(s'|s)$ denote $\mathbb{P}(s_{t+n} = s'|s_t = s)$. We say s' is *accessible* from s if $P_\pi^{(n)}(s'|s) > 0$ for some $n \geq 0$, denoted by $s \to s'$, and s *communicates* with s' if $s \to s',\ s' \to s$, denoted by $s \leftrightarrow s'$.

A set $C \subseteq \mathcal{S}$ is *closed* if there is no state in $\mathcal{S} \backslash C$ that is accessible from any state in C. A closed set C is *irreducible* if there is no proper subset that is closed. Let τ_s denote the time of first visit (return to s starting from s). A state s is *recurrent* if $\mathbb{P}(\tau_s < \infty) = 1$, and *transient* if $\mathbb{P}(\tau_s < \infty) < 1$. We say $d_\pi : \mathcal{S} \to \mathbb{R}_+$ is a *stationary distribution* for P_π if $\sum_s d_\pi(s) = 1$ and $\sum_{s'} P_\pi(s|s') d_\pi(s') = d_\pi(s)$ for all s. Equivalently, we can write it as $\mathbf{1}^\top d_\pi = 1$ and $P_\pi^\top d_\pi = d_\pi$ using matrix notation.

Since $\leftrightarrow$ is an equivalence relation, $\mathcal{S}$ can be partitioned into *communicating classes*. Moreover, for each communicating class C, all states in C are either recurrent or transient. A communicating class C is *recurrent class* if all states in C are recurrent, and *transient class* if all states in C are transient. Let $\gcd\{n > 0 : P_\pi^{(n)}(s|s) > 0\}$ denote the period of state s, and all states in a recurrent class share the same period. We call a class *periodic* if the period of states exceeds 1, and *aperiodic* if 1.

A Markov chain P_π is *irreducible* if it consists of a single closed class, *unichain* if it consists of a single set of closed irreducible states and a set of (possibly empty) transient states, and otherwise *multichain*.

MDP classifications An MDP $\mathcal{M}$ is called *ergodic* if, for every stationary policy, the induced transition matrix is irreducible and aperiodic. An MDP $\mathcal{M}$ is called *unichain* if, for every deterministic stationary policy, the induced transition matrix is unichain. An MDP $\mathcal{M}$ is called *(general) multichain* if there exists at least one stationary policy whose induced transition matrix contains one or more closed irreducible recurrent classes. These classes satisfy ergodic$\subseteq$ unichain $\subseteq$ weakly communicating $\subseteq$ multichain.

Next, from Lemmas 8 to 10, we provide standard results for average-reward MDPs with finite $\mathcal{S}, \mathcal{A}$.

Lemma 8 (Proposition 8.9.1 in Puterman (1994))**.** *Let $\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, r)$ be an MDP with $|\mathcal{S}|, |\mathcal{A}| < \infty$, and let $s_1 \in \mathcal{S}$ be any initial state. For any $\pi \in \Pi$, the occupation measure $q_{s_1}^\pi$ exists.*

Corollary 1. *Let $\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, r)$ be an MDP with $|\mathcal{S}|, |\mathcal{A}| < \infty$, and let $s_1 \in \mathcal{S}$ be any initial state. Let $r : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ be any bounded reward function. For any $\pi \in \Pi$, we have $J_r^\pi(s_1) = r^\top q_{s_1}^\pi$.*

Proof. Note that

$$J_r^\pi(s_1) = \lim_{T\to\infty} \frac{1}{T}\mathbb{E}_{P,\pi}\left[\sum_{t=1}^{T} r(s_t, a_t)|s_1\right] = \lim_{T\to\infty} \frac{1}{T}\mathbb{E}_{P,\pi}\left[\sum_{t=1}^{T}\sum_{s,a} r(s,a)\mathbb{1}\{s_t = s, a_t = a\}|s_1\right].$$

By Lemma 8, since $q_{s_1}^\pi(s,a) = \lim_{T\to\infty} \frac{1}{T}\mathbb{E}_{P,\pi}\left[\sum_{t=1}^{T} \mathbb{1}\{s_t = s, a_t = a\}|s_1\right]$ exists and $\mathcal{S}, \mathcal{A}$ are finite, we have

$$J_r^\pi(s_1) = \sum_{s,a} r(s,a) \lim_{T\to\infty} \frac{1}{T}\mathbb{E}_{P,\pi}\left[\sum_{t=1}^{T} \mathbb{1}\{s_t = s, a_t = a\}|s_1\right] = r^\top q_{s_1}^\pi.$$

□

Lemma 9 (Theorem 8.9.3 in Puterman (1994))**.** *Let $\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, r)$ denote an MDP with $|\mathcal{S}| < \infty$ and $|\mathcal{A}| < \infty$, and let $s_1 \in \mathcal{S}$ be any initial state. Then we have*

$$\mathrm{conv}(\widehat{\mathcal{P}}^{\mathrm{SD}}(s_1)) = \overline{\mathrm{conv}}(\widehat{\mathcal{P}}^{\mathrm{SD}}(s_1)) = \overline{\mathrm{conv}}(\widehat{\mathcal{P}}(s_1)).$$

where $\mathrm{conv}(\cdot), \overline{\mathrm{conv}}(\cdot)$ denote the convex hull and closed convex hull, respectively, Π^{SD} denotes the set of stationary deterministic policies, and $\widehat{\mathcal{P}}^{\mathrm{SD}}(s_1) = \{q_{s_1}^\pi : \pi \in \Pi^{\mathrm{SD}}\}$.

Proof. Since $\overline{\mathrm{conv}}(\widehat{\mathcal{P}}^{\mathrm{SD}}(s_1)) = \overline{\mathrm{conv}}(\widehat{\mathcal{P}}(s_1))$ directly follows from Theorem 8.9.3 in Puterman (1994), it is sufficient to show that $\mathrm{conv}(\widehat{\mathcal{P}}^{\mathrm{SD}}(s_1)) = \overline{\mathrm{conv}}(\widehat{\mathcal{P}}^{\mathrm{SD}}(s_1))$. Since $|\mathcal{S}|, |\mathcal{A}| < \infty$, the number of stationary deterministic policies is at most $|\mathcal{A}|^{|\mathcal{S}|}$, which is finite. Moreover, $\widehat{\mathcal{P}}^{\mathrm{SD}}(s_1)$ is finite, since it is defined as occupation measures with respect to stationary deterministic policies. Note that a closed convex hull of a finite set is equal to its convex hull. This concludes the proof. □

Lemma 10. *Let $\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, r)$ denote an MDP with $|\mathcal{S}| < \infty$ and $|\mathcal{A}| < \infty$, and let $s_1 \in \mathcal{S}$ be any initial state. Let $r : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ be any bounded reward function. There exists $\pi^* \in \Pi^{\mathrm{SD}}$ such that $J_r^{\pi^*}(s_1) = \sup_{\pi\in\Pi} J_r^\pi(s_1)$.*

Proof. Due to the definition of $\widehat{\mathcal{P}}(s_1)$ and Corollary 1, $\sup_{\pi\in\Pi} J_r^\pi(s_1)$ can be rewritten as

$$\sup_{q\in\widehat{\mathcal{P}}(s_1)} r^\top q.$$

Since $\widehat{\mathcal{P}}(s_1) \subseteq \overline{\mathrm{conv}}(\widehat{\mathcal{P}}(s_1))$ and, by Lemma 9, $\overline{\mathrm{conv}}(\widehat{\mathcal{P}}(s_1)) = \mathrm{conv}(\widehat{\mathcal{P}}^{\mathrm{SD}}(s_1))$, it follows that

$$\sup_{q\in\widehat{\mathcal{P}}(s_1)} r^\top q \le \sup_{q\in\overline{\mathrm{conv}}(\widehat{\mathcal{P}}(s_1))} r^\top q = \sup_{q\in\mathrm{conv}(\widehat{\mathcal{P}}^{\mathrm{SD}}(s_1))} r^\top q.$$

Note that $\widehat{\mathcal{P}}^{\mathrm{SD}}(s_1)$ is finite. Then $\mathrm{conv}(\widehat{\mathcal{P}}^{\mathrm{SD}}(s_1))$ is a non-empty polytope, and the right-hand side becomes an LP. Therefore, the maximizer is attained at an extreme point of $\mathrm{conv}(\widehat{\mathcal{P}}^{\mathrm{SD}}(s_1))$—denoted by q^*—and it is clear that $q^* \in \widehat{\mathcal{P}}^{\mathrm{SD}}(s_1)$, i.e.,

$$\exists q^* \in \widehat{\mathcal{P}}^{\mathrm{SD}}(s_1) : \; r^\top q^* = \sup_{q\in\mathrm{conv}(\widehat{\mathcal{P}}^{\mathrm{SD}}(s_1))} r^\top q.$$

This implies that

$$\sup_{q\in\widehat{\mathcal{P}}(s_1)} r^\top q \le r^\top q^*.$$

Moreover, note that $q^* \in \widehat{\mathcal{P}}^{\mathrm{SD}}(s_1) \subseteq \widehat{\mathcal{P}}(s_1)$. This implies that

$$r^\top q^* \le \sup_{q\in\widehat{\mathcal{P}}(s_1)} r^\top q.$$

Finally, we show that there exists $q^* \in \widehat{\mathcal{P}}^{\mathrm{SD}}(s_1)$ such that $r^\top q^* = \sup_{q\in\widehat{\mathcal{P}}(s_1)} r^\top q$. By the definition of $\widehat{\mathcal{P}}^{\mathrm{SD}}(s_1)$, there exists some stationary deterministic policy π^* that induces q^*. By Corollary 1, $J_r^{\pi^*}(s_1) = r^\top q^*$, and this completes the proof. □

D Proof roadmap

Figure 3 summarizes the main dependencies in the proofs of Theorems 1 and 2. The upper part establishes strong duality through the geometry of stationary occupation measures. The lower part shows how strong duality complements the composite upper bound to control regret and constraint violation separately.

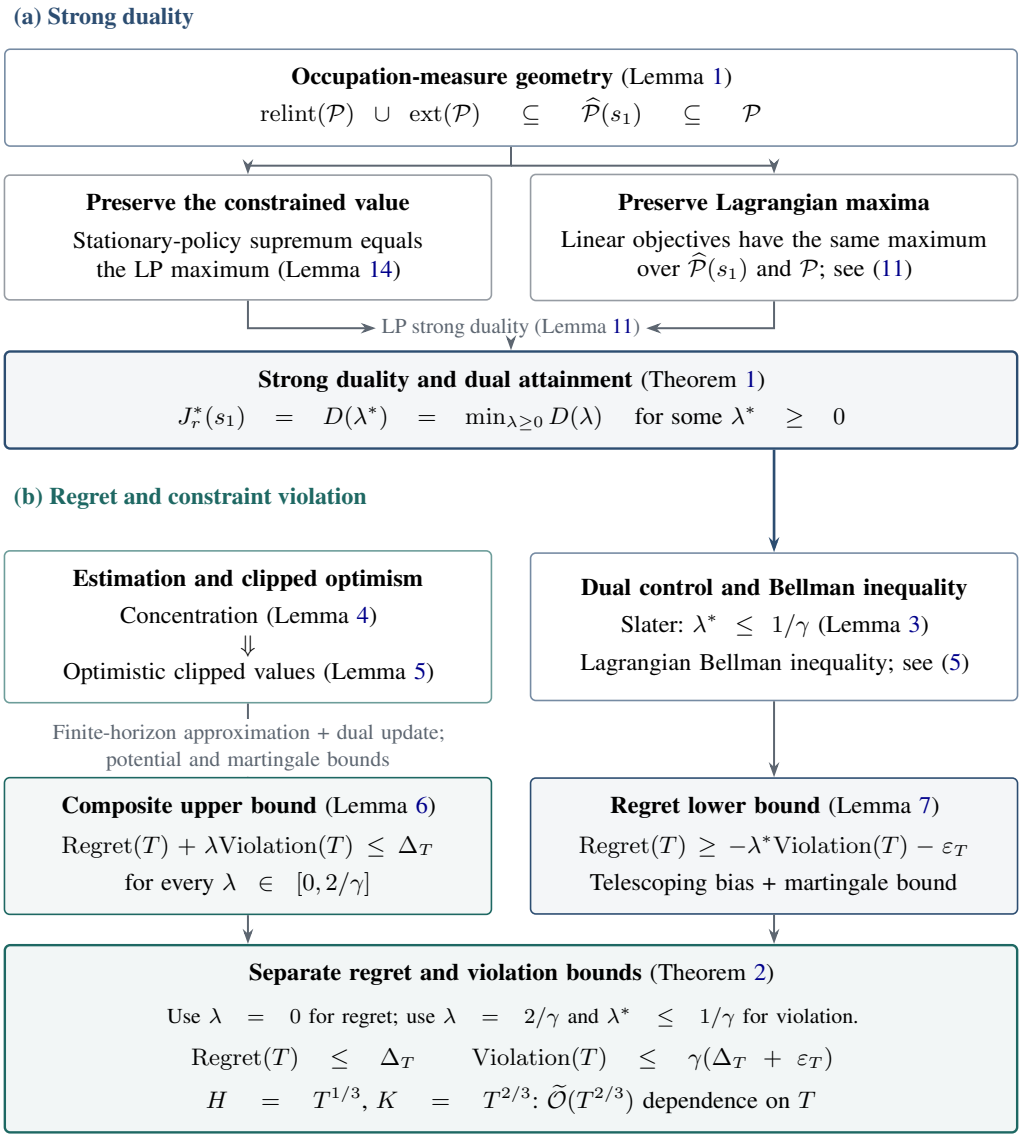


Figure 3: Proof roadmap for the two main theorems. Arrows indicate the principal dependencies. Here Δ_T denotes the composite upper bound in Lemma 6, and $\varepsilon_T = \mathrm{sp}(v^*_{\lambda^*})\big(\sqrt{2T\log(1/\delta)} + 1\big)$ is the error in Lemma 7. The final inequalities hold jointly with probability at least $1 - 3\delta$; the full dependence on dimension, Slater constant, and bias spans is given in Theorem 2.

E Missing proofs for Section 3

In this section, we provide a rigorous proof for strong duality when the underlying MDP structure is weakly communicating. Recall that the strong duality result we are interested in is formulated as

$$\begin{array}{ll} \sup\limits_{\pi\in\Pi} & J_r^\pi(s_1) \\ \text{s.t.} & J_g^\pi(s_1) \geq b \end{array} \quad = \quad \min_{\lambda\geq 0} \sup_{\pi\in\Pi} J_r^\pi(s_1) + \lambda(J_g^\pi(s_1) - b).$$

To show this, we first introduce Lemma 11, which states strong duality property for LPs, and Lemmas 12 - 15, which are useful for proving strong duality for average-reward CMDPs. Subsequently, in Section E.1, we provide a proof for Lemma 1, which characterizes the structure of $\widehat{\mathcal{P}}(s_1)$ with arbitrarily chosen s_1. Finally, based on these results, we prove Theorem 1 in Section E.2.

Lemma 11 (LP Strong Duality)**.** *Let $\mathcal{X} \subseteq \mathbb{R}^n$ be a nonempty polytope, and let $r, g \in \mathbb{R}^n$ and $b \in \mathbb{R}$. Consider a primal LP $P^* = \max_{x\in\mathcal{X}} r^\top x$ s.t. $g^\top x \geq b$. Define the corresponding dual functions as $D(\lambda) = \max_{x\in\mathcal{X}} r^\top x + \lambda(g^\top x - b)$. If the primal LP is feasible, then there exists $\lambda^* \geq 0$ such that $\lambda^* \in \arg\min_{\lambda\geq 0} D(\lambda)$, and moreover, $P^* = D(\lambda^*)$.*

Proof. This is a special case of Proposition 5.3.6 in Bertsekas (2009). □

Lemma 12. *Let $\mathcal{X} = \{x \in \mathbb{R}^n : x \geq 0, Ax = b\}$ be a non-empty polyhedron, and let $I_+ = \{i \in [n] : \exists x \in \mathcal{X} \text{ s.t. } x_i > 0\}$. Suppose that I_+ is non-empty. For $x \in \mathcal{X}$, we have $x \in \text{relint}(\mathcal{X})$ if and only if $x_i > 0$ for all $i \in I_+$.*

Proof. ($\Rightarrow$) Let $x \in \text{relint}(\mathcal{X})$. Fix $i \in I_+$. By the definition of I_+, we can always find $y \in \mathcal{X}$ such that $y_i > 0$. By the definition of $\text{relint}(\mathcal{X})$, there exists $\epsilon > 1$ such that $\epsilon x + (1-\epsilon)y \in \mathcal{X}$. By the definition of $\mathcal{X}$, it follows that $\epsilon x_i + (1-\epsilon)y_i \geq 0$. Since $\epsilon > 1$ and $y_i > 0$, we have $x_i \geq \frac{(\epsilon-1)y_i}{\epsilon} > 0$. This implies that $x_i > 0$ for all $i \in I_+$.

($\Leftarrow$) Let $x \in \mathcal{X}$ and $x_i > 0$ for all $i \in I_+$. Fix any $y \in \mathcal{X}$. By the definition of I_+, we know that $y_j = 0$ for all $j \in I_+^c$. Take $\epsilon = 1 + \frac{\min_{i\in I_+} x_i}{\max\{1, \|y\|_\infty\}}$, so $\epsilon > 1$. We claim that $\epsilon x + (1-\epsilon)y \in \mathcal{X}$. To show this, first it is clear that $A(\epsilon x + (1-\epsilon)y) = b$. Moreover, for $i \in I_+$, $\epsilon x_i + (1-\epsilon)y_i \geq 0$, due to our choice of ϵ. For $j \in I_+^c$, $\epsilon x_j + (1-\epsilon)y_j = 0$. Therefore, by the definition of $\text{relint}(\mathcal{X})$, $x \in \text{relint}(\mathcal{X})$. □

Lemma 13. *Let P_π be a Markov chain over finite states $\mathcal{S}$. Suppose that P_π induces a single set of closed irreducible states $\mathcal{S}_C$. Let $\mathcal{S}_T = \mathcal{S}\backslash\mathcal{S}_C$ be the set of transient states. Then the following properties hold:*

1. *There exists a unique stationary distribution d_π for P_π such that $d_\pi(s) > 0$ for $s \in \mathcal{S}_C$ and $d_\pi(s) = 0$ for $s \in \mathcal{S}_T$.*
2. *Let $\mu_t(s) = \mathbb{E}_{P_\pi}[\mathbb{1}\{s_t = s\}|s_1 \sim \mu_1]$, where μ_1 is the given initial state distribution. Then we have $\lim_{n\to\infty} \frac{1}{n}\sum_{t=1}^n \mu_t(s) = d_\pi(s)$ for all $s \in \mathcal{S}$.*

Proof. For simplicity, with some abuse of notation, let $P_\pi \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{S}|}$ denote the transition matrix, where $P_{\pi,ij}$ denotes the probability of transition from state i to state j. Also, we use $d_\pi, \mu_t \in \mathbb{R}^{|\mathcal{S}|}$ for the vector notation of the stationary distribution and the expected state distribution at t, respectively. Moreover, let $d_{\pi,\mathcal{S}_C}, \mu_{t,\mathcal{S}_C} \in \mathbb{R}^{|\mathcal{S}_C|}$ denote subvectors such that $d_{\pi,\mathcal{S}_C}(s) = d_\pi(s)$ and $\mu_{t,\mathcal{S}_C} = \mu_t(s)$ for all $s \in \mathcal{S}_C$. A similar rule is applied to $d_{\pi,\mathcal{S}_T}, \mu_{t,\mathcal{S}_T} \in \mathbb{R}^{|\mathcal{S}_T|}$. Let $\mathbf{0}_\mathcal{S}, \mathbf{1}_\mathcal{S} \in \mathbb{R}^{|\mathcal{S}|}$, $\mathbf{0}_{\mathcal{S}_C}, \mathbf{1}_{\mathcal{S}_C} \in \mathbb{R}^{|\mathcal{S}_C|}$, and $\mathbf{0}_{\mathcal{S}_T}, \mathbf{1}_{\mathcal{S}_T} \in \mathbb{R}^{|\mathcal{S}_T|}$ denote all-0 and all-1 vectors, respectively.

Proof of the first statement. Without loss of generality, we can express P_π as the following block matrix:

$$P_\pi = \begin{bmatrix} M_C & 0 \\ M & M_T \end{bmatrix}$$

where M_T is the transition matrix between states in $\mathcal{S}_T$, M is from $\mathcal{S}_T$ to $\mathcal{S}_C$, and M_C is between states in $\mathcal{S}_C$. Note that a stationary distribution $d_\pi \in \mathbb{R}_+^{|\mathcal{S}|}$ satisfies $P_\pi^\top d_\pi = d_\pi$ and $\mathbf{1}_{\mathcal{S}}^\top d_\pi = 1$. Here, the condition $P_\pi^\top d_\pi = d_\pi$ can be rewritten as

$$(I - M_T^\top) d_{\pi,\mathcal{S}_T} = \mathbf{0}_{\mathcal{S}_T}, \quad M^\top d_{\pi,\mathcal{S}_T} + M_C^\top d_{\pi,\mathcal{S}_C} = d_{\pi,\mathcal{S}_C}.$$

It is known that $I - M_T$ is invertible, when M_T is a transition matrix for transient states (Proposition A.3 in Puterman (1994)). Therefore, we have $d_{\pi,\mathcal{S}_T} = \mathbf{0}_{\mathcal{S}_T}$. Now, we focus on $d_{\pi,\mathcal{S}_C}$ such that $M_C^\top d_{\pi,\mathcal{S}_C} = d_{\pi,\mathcal{S}_C}$ and $\mathbf{1}_{\mathcal{S}_C}^\top d_{\pi,\mathcal{S}_C} = 1$. Note that it has a unique solution that is $d_{\pi,\mathcal{S}_C}(s) > 0$ for all $s \in \mathcal{S}_C$ (Theorem 1.7.7 in Norris (1998)). This concludes the proof of the first statement.

Proof of the second statement. First, we show that $\lim_{n\to\infty} \frac{1}{n} \sum_{t=1}^n \mu_t(s) = 0$ for $s \in \mathcal{S}_T$. Note that $\mu_t = (P_\pi^\top)^{t-1} \mu_1$. Since M_T is for transient states, its spectral radius is strictly less than 1 (Proposition A.3 in Puterman (1994)). This leads to, as $t \to \infty$,

$$\lim_{t\to\infty} \mu_{t,\mathcal{S}_T} = \lim_{t\to\infty} (M_T^\top)^{t-1} \mu_{1,\mathcal{S}_T} = \mathbf{0}_{\mathcal{S}_T}. \tag{7}$$

This implies that the limit of its average satisfies $\lim_{n\to\infty} \frac{1}{n} \sum_{t=1}^n \mu_t(s) = 0$ for all $s \in \mathcal{S}_T$. Next, we consider states in $\mathcal{S}_C$. Let $\bar{\mu}_t(s) = \mathbb{1}\{s_t = s\}$. Let $\tau = \inf\{t \geq 1 : s_t \in \mathcal{S}_C\}$ denote the time of the first visit $\mathcal{S}_C$. Since $\mathcal{S}_T$ is a transient set and $\mathcal{S}_C$ is the unique recurrent class, the chain visits $\mathcal{S}_T$ only finite times a.s., thus, $\tau < \infty$ a.s. For $t \geq \tau$, since the transition only occurs in $\mathcal{S}_C$, we can consider the Markov chain restricted to $\mathcal{S}_C$, which is irreducible. Note that Theorem 1.10.2 in Norris (1998) implies that for $s \in \mathcal{S}_C$,

$$\lim_{n\to\infty} \frac{1}{n-\tau+1} \sum_{t=\tau}^n \bar{\mu}_t(s) = d_\pi(s) \quad \text{a.s.}$$

This leads to

$$\lim_{n\to\infty} \frac{1}{n} \sum_{t=1}^n \bar{\mu}_t(s) = \lim_{n\to\infty} \left(\frac{1}{n} \sum_{t=1}^{\tau-1} \bar{\mu}_t(s) + \frac{n-\tau+1}{n} \frac{1}{n-\tau+1} \sum_{t=\tau}^n \bar{\mu}_t(s) \right) = d_\pi(s) \quad \text{a.s.}$$

Moreover, we know that $|\frac{1}{n} \sum_{t=1}^n \bar{\mu}_{t,\mathcal{S}_C}| \leq 1$. Then by the bounded convergence theorem (Theorem 1.5.3 in Durrett (2019)), it follows that for $s \in \mathcal{S}_C$,

$$\lim_{n\to\infty} \frac{1}{n} \sum_{t=1}^n \mathbb{E}_{P_\pi} [\bar{\mu}_t(s) | s_1 \sim \mu_1] = d_\pi(s).$$

Since $\mathbb{E}_{P_\pi}[\bar{\mu}_t(s)|s_1 \sim \mu_1] = \mu_t(s)$, this implies that $\lim_{n\to\infty} \frac{1}{n} \sum_{t=1}^n \mu_{t,\mathcal{S}_C} = d_{\pi,\mathcal{S}_C}$. By combining this with (7), we complete the proof of the second statement. □

Lemma 14. *For $n \in \mathbb{N}$, let $r, g \in \mathbb{R}^n$ and $b \in \mathbb{R}$. Let $\mathcal{P} \subseteq \mathbb{R}^n$ be a non-empty polytope. Let $\widehat{\mathcal{P}} \subseteq \mathcal{P}$ denote a subset of $\mathcal{P}$, which is possibly not a polytope. Suppose that $\{q \in \widehat{\mathcal{P}} : g^\top q \geq b\} \neq \varnothing$ and $\mathrm{relint}(\mathcal{P}) \cup \mathrm{ext}(\mathcal{P}) \subseteq \widehat{\mathcal{P}}$. Then we have*

$$\begin{array}{ll} \sup\limits_{q \in \widehat{\mathcal{P}}} & r^\top q \\ \text{s.t.} & g^\top q \geq b \end{array} = \begin{array}{ll} \max\limits_{q \in \mathcal{P}} & r^\top q \\ \text{s.t.} & g^\top q \geq b. \end{array} \tag{8}$$

Proof. Let $F_{\mathcal{P}}, F_{\widehat{\mathcal{P}}}$ denote the feasible sets of each optimization problems, i.e.,

$$F_{\mathcal{P}} = \{q \in \mathcal{P} : g^\top q \geq b\}, \; F_{\widehat{\mathcal{P}}} = \{q \in \widehat{\mathcal{P}} : g^\top q \geq b\}.$$

Since $F_{\widehat{\mathcal{P}}} \subseteq F_{\mathcal{P}}$ and $F_{\widehat{\mathcal{P}}}$ is non-empty by the assumption, we know that $F_{\mathcal{P}}$ is a non-empty polytope. Then the right-hand side of (8) becomes an LP, and its maximum is attainable by some $q^* \in \mathcal{P}$. We define P^* as

$$P^* = r^\top q^* = \max_{q \in \mathcal{P}} r^\top q \quad \text{s.t.} \quad g^\top q \geq b.$$

Since $F_{\widehat{\mathcal{P}}} \subseteq F_{\mathcal{P}}$, it is clear that

$$P^* \geq \sup_{q \in \widehat{\mathcal{P}}} r^\top q \quad \text{s.t.} \quad g^\top q \geq b.$$

Now, it is sufficient to prove that $P^* \leq \sup_{q \in \widehat{\mathcal{P}}} r^\top q \quad \text{s.t.} \quad g^\top q \geq b$.

Case 1: $\exists \bar{q} \in \{q \in \mathrm{relint}(\mathcal{P}) : g^\top q > b\}$**.** In this case, for $\theta \in (0,1)$, we define

$$\bar{q}_\theta = \theta \bar{q} + (1-\theta) q^*.$$

Since $q^* \in \mathcal{P}$, $\bar{q} \in \mathrm{relint}(\mathcal{P})$, we have $\bar{q}_\theta \in \mathrm{relint}(\mathcal{P}) \subseteq \widehat{\mathcal{P}}$ for $\theta \in (0,1)$ (Theorem 6.1 in Rockafellar (1970)). Moreover, it is clear that $g^\top \bar{q}_\theta > b$ for $\theta \in (0,1)$. Therefore, $\bar{q}_\theta \in F_{\widehat{\mathcal{P}}}$ for $\theta \in (0,1)$. By taking $\theta \to 0+$, we have

$$\sup_{q \in F_{\widehat{\mathcal{P}}}} r^\top q \geq \lim_{\theta \to 0+} r^\top \bar{q}_\theta = r^\top q^*.$$

Case 2: $\{q \in \mathrm{relint}(\mathcal{P}) : g^\top q > b\} = \varnothing$**.** Note that this implies that $\mathrm{relint}(\mathcal{P}) = \{q \in \mathrm{relint}(\mathcal{P}) : g^\top q \leq b\}$. Based on this fact,

$$\mathcal{P} = \overline{\mathrm{relint}(\mathcal{P})} = \overline{\{q \in \mathrm{relint}(\mathcal{P}) : g^\top q \leq b\}} \subseteq \{q \in \overline{\mathrm{relint}(\mathcal{P})} : g^\top q \leq b\} = \{q \in \mathcal{P} : g^\top q \leq b\}$$

where the second last relation follows from $\overline{A \cap B} \subseteq \overline{A} \cap \overline{B}$. This implies that $\mathcal{P} = \{q \in \mathcal{P} : g^\top q \leq b\}$. Moreover, it follows that

$$F_{\mathcal{P}} = \mathcal{P} \cap \{q \in \mathbb{R}^n : g^\top q \geq b\} = \{q \in \mathcal{P} : g^\top q = b\}.$$

Thus, $F_{\mathcal{P}}$ is a face of the polytope $\mathcal{P}$, which is also a polytope. Then the maximizer of $\max_{q \in F_{\mathcal{P}}} r^\top q$, denoted by q^*, is attained at an extreme point of $F_{\mathcal{P}}$, i.e., $q^* \in \mathrm{ext}(F_{\mathcal{P}})$. Additionally, we know that extreme points of a face of a polytope are also extreme points of that polytope, i.e., $\mathrm{ext}(F_{\mathcal{P}}) \subseteq \mathrm{ext}(\mathcal{P})$. These imply that $q^* \in \mathrm{ext}(\mathcal{P}) \subseteq \widehat{\mathcal{P}}$. Note that $g^\top q^* = b$. Therefore, we have

$$\sup_{q \in F_{\widehat{\mathcal{P}}}} r^\top q \geq r^\top q^* = P^*.$$

In both cases, we prove $\sup_{q \in F_{\widehat{\mathcal{P}}}} r^\top q \geq P^*$. This completes the proof. □

Lemma 15. *Let $\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, r)$ be a weakly communicating MDP, and let $s_1 \in \mathcal{S}$ be any initial state. For any $q^* \in \mathrm{relint}(\mathcal{P})$, there exists $\pi^* \in \Pi$ that satisfies the following: (1) P_{π^*} is unichain, (2) $q^* = q^{\pi^*}_{s_1}$.*

Proof. The outline of the proof is as follows. Given $q^* \in \mathrm{relint}(\mathcal{P})$, we construct the stationary randomized policy π^*. Next, based on properties of relative interior and weakly communicating MDPs, we show P_{π^*} is unichain. Finally, we prove that the occupation measure π^* is equal to q^*, i.e., $q^{\pi^*}_{s_1} = q^*$.

Step 1: Construct a stationary policy π^*. We define π^* as

$$\pi^*(a|s) = \frac{q^*(s,a)}{\sum_a q^*(s,a)} \quad \text{if} \quad \sum_a q^*(s,a) > 0.$$

Otherwise, for s such that $\sum_a q^*(s,a) = 0$, take an arbitrary action. Then, we define

$$d^*(s) = \sum_a q^*(s,a).$$

Next, we introduce a useful property that $d^*(s)$ is a stationary distribution under P_{π^*}. This is because for any $s \in \mathcal{S}$,

$$\begin{aligned}
\sum_{s'} P_{\pi^*}(s|s')d^*(s') &= \sum_{s'}\sum_{a'} P(s|s',a')\pi^*(a'|s')d^*(s') \\
&= \sum_{s':d^*(s')>0}\sum_{a'} P(s|s',a')\pi^*(a'|s')d^*(s') + \sum_{s':d^*(s')=0}\sum_{a'} P(s|s',a')\pi^*(a'|s')d^*(s') \\
&= \sum_{s':d^*(s')>0}\sum_{a'} P(s|s',a')\frac{q^*(s',a')}{d^*(s')}d^*(s') + \sum_{s':d^*(s')=0}\sum_{a'} P(s|s',a')\pi^*(a'|s')d^*(s') \\
&= \sum_{s':d^*(s')>0}\sum_{a'} P(s|s',a')\frac{q^*(s',a')}{d^*(s')}d^*(s') + \sum_{s':d^*(s')=0}\sum_{a'} P(s|s',a')q^*(s',a') \\
&= \sum_{s'}\sum_{a'} P(s|s',a')q^*(s',a') \\
&= \sum_{a} q^*(s,a) \\
&= d^*(s)
\end{aligned}$$

where the third equality is due to the definition of π^*, the fourth equality follows from $\pi^*(a'|s')d^*(s') = q^*(s',a') = 0$ for all s' such that $d^*(s') = 0$, and the second last equality follows from $q^* \in \mathcal{P}$.

Step 2: Show P_{π^*} is unichain. Now, we show P_{π^*} has a set of single closed irreducible states with transient states.

Since $\mathcal{M}$ is a weakly communicating MDP, there exists a stationary randomized policy $\check{\pi}$ whose induced Markov chain $P_{\check{\pi}}$ has a set of single closed irreducible states, denoted by $\mathcal{S}_C$. Moreover, $\mathcal{S}_T = \mathcal{S}\backslash\mathcal{S}_C$ is the set of states that are transient under all stationary policies. By Lemma 13, there exists a unique stationary distribution $d_{\check{\pi}}$ such that $d_{\check{\pi}}(s) > 0$ for $s \in \mathcal{S}_C$ and $d_{\check{\pi}}(s) = 0$ for $s \in \mathcal{S}_T$. Based on this, we define

$$q^{\check{\pi}}(s,a) = \check{\pi}(a|s)d_{\check{\pi}}(s).$$

Note that $q^{\check{\pi}} \in \mathcal{P}$ can be shown as follows. We have $\sum_{s,a} q^{\check{\pi}}(s,a) = \sum_s d_{\check{\pi}}(s)\sum_a \check{\pi}(a|s) = 1$ and $q^{\check{\pi}}(s,a) \geq 0$. Moreover,

$$\sum_{s',a'} P(s|s',a')q^{\check{\pi}}(s',a') = \sum_{s',a'} P(s|s',a')\check{\pi}(a'|s')d_{\check{\pi}}(s') = \sum_{s'} P(s|s')d_{\check{\pi}}(s') = d_{\check{\pi}}(s)$$

where the last equality follows from that $d_{\check{\pi}}$ is a stationary distribution. Furthermore, it is trivial that $d_{\check{\pi}}(s) = \sum_a q^{\check{\pi}}(s,a)$. These imply that $q^{\check{\pi}} \in \mathcal{P}$.

Consider a Markov chain P_{π^*} induced by π^*. First, we prove that $\mathcal{S}_C$ is closed irreducible under P_{π^*}. Since $\mathcal{S}_C$ is closed irreducible under $P_{\check{\pi}}$, for any $s', s'' \in \mathcal{S}_C$, there exists a sequence of states in $\mathcal{S}_C$, denoted by $(s_1, s_2, \ldots, s_m)$ for some $m \in \mathbb{N}$, such that $s' = s_1$, $s'' = s_m$, and $P_{\check{\pi}}(s_{i+1}|s_i) > 0$ for all $i = 1, \ldots, m-1$. It can be rewritten as $\sum_a P(s_{i+1}|s_i,a)\check{\pi}(a|s_i) > 0$ for $i = 1, \ldots, m-1$. This implies that $\check{\pi}(a_{s_i}|s_i) > 0$ and $P(s_{i+1}|s_i,a_{s_i}) > 0$ for some a_{s_i} for all $i = 1, \ldots, m-1$. By the definition of $q^{\check{\pi}}$, for each $i = 1, \ldots, m-1$, we have $q^{\check{\pi}}(s_i,a_{s_i}) > 0$. By Lemma 12, $q^*(s_i,a_{s_i}) > 0$ for all $i = 1, \ldots, m-1$. This implies that

$$P_{\pi^*}(s_{i+1}|s_i) = \sum_a P(s_{i+1}|s_i,a)\pi^*(a|s_i) = \sum_a P(s_{i+1}|s_i,a)\frac{q^*(s_i,a)}{\sum_a q^*(s_i,a)} > 0.$$

Since $P_{\pi^*}(s_{i+1}|s_i) > 0$ for all $i = 1, \ldots, m-1$, we have any $s', s'' \in \mathcal{S}_C$ communicate under P_{π^*}. Moreover, $\mathcal{S}_C$ is closed. If not, there exist $s \in \mathcal{S}_C,\ s' \in \mathcal{S}_T$ such that $s \to s'$. Since s' is transient, so is s, as it is known that a state s is transient if $s \to s'$ for some transient state s' (Appendix A.3 in Puterman (1994)). Since all states in $\mathcal{S}_C$ communicate, $\mathcal{S}_C$ becomes transient. This implies that all states in $\mathcal{S}$ are transient, which contradicts the fact that any finite Markov chain has at least one positive recurrent class. Finally, this implies that P_{π^*} induces a single closed irreducible set $\mathcal{S}_C$ and transient set $\mathcal{S}_T$.

Step 3: Show $q_{s_1}^{\pi^*} = q^*$. Let d_{π^*} denote the stationary distribution of P_{π^*}. Since P_{π^*} induces a single recurrent class and a transient set, by Lemma 13, for all $s \in \mathcal{S}$,

$$\lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}_{P_{\pi^*}}[\mathbb{1}\{s_t = s\}|s_1] = d_{\pi^*}(s).$$

Multiplying both sides by $\pi^*(a|s)$ yields

$$\lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}_{\pi^*}[\mathbb{1}\{s_t = s, a_t = a\}|s_1] = d_{\pi^*}(s)\pi^*(a|s).$$

To further deduce, we derive the following property: for any $s \in \mathcal{S}$,

$$d_{\pi^*}(s) = d^*(s).$$

Note that this property follows from that (i) by Lemma 13, P_{π^*} has the unique stationary distribution $d_{\pi^*}(s)$ and (ii) by the argument in Step 1, $d^*(s) = \sum_a q^*(s,a)$ is also a stationary distribution of P_{π^*}. This leads to for all $(s,a) \in \mathcal{S} \times \mathcal{A}$,

$$d^*(s)\pi^*(a|s) = q^*(s,a),$$

since for s such that $d^*(s) > 0$ we have $d^*(s)\pi^*(a|s) = d^*(s)\frac{q^*(s,a)}{d^*(s)} = q^*(s,a)$, and for s such that $d^*(s) = 0$ we have $d^*(s)\pi^*(a|s) = q^*(s,a) = 0$.

Finally, we show that

$$\lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}_{\pi^*}[\mathbb{1}\{s_t = s, a_t = a\}|s_1] = q^*(s,a).$$

Recall that the left-hand side is the definition of $q_{s_1}^{\pi^*}(s,a)$. This implies that for any $q^* \in \mathrm{relint}(\mathcal{P})$, there exists a stationary policy π^* whose occupation measure satisfies $q_{s_1}^{\pi^*} = q^*$. □

E.1 Proof of Lemma 1

Proof of $\widehat{\mathcal{P}}(s_1) \subseteq \mathcal{P}$. The statement is a direct consequence of Proposition 3.1 in Mannor and Tsitsiklis (2005).

Proof of $\mathrm{relint}(\mathcal{P}) \subseteq \widehat{\mathcal{P}}(s_1)$. By Lemma 15, for any $q^* \in \mathrm{relint}(\mathcal{P})$, there exists $\pi^* \in \Pi$ such that $q^* = q_{s_1}^{\pi^*}$. Since $q_{s_1}^{\pi^*}$ is an occupation measure induced by stationary policy π^* starting from s_1, $q_{s_1}^{\pi^*} \in \widehat{\mathcal{P}}(s_1)$. This concludes the proof.

Proof of $\mathrm{ext}(\mathcal{P}) \subseteq \widehat{\mathcal{P}}(s_1)$. Fix $q^* \in \mathrm{ext}(\mathcal{P})$. Then there exists a vector $r \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{A}|}$ such that $\arg\max_{q\in\mathcal{P}} r^\top q = \{q^*\}$. For such r, we consider the following MDP:

$$\sup_{\pi\in\Pi} J_r^\pi(s_1).$$

It can be rewritten as

$$\sup_{q\in\widehat{\mathcal{P}}(s_1)} r^\top q.$$

Note that since $\mathrm{relint}(\mathcal{P})$ is the relative interior of $\mathcal{P}$,

$$\sup_{q\in\mathrm{relint}(\mathcal{P})} r^\top q = \max_{q\in\mathcal{P}} r^\top q = r^\top q^*.$$

We know that $\mathrm{relint}(\mathcal{P}) \subseteq \widehat{\mathcal{P}}(s_1)$ and $\widehat{\mathcal{P}}(s_1) \subseteq \mathcal{P}$. This implies that

$$r^\top q^* = \sup_{q\in\mathrm{relint}(\mathcal{P})} r^\top q \le \sup_{q\in\widehat{\mathcal{P}}(s_1)} r^\top q \le \sup_{q\in\mathcal{P}} r^\top q = r^\top q^*.$$

This implies that

$$\sup_{\pi\in\Pi} J_r^\pi(s_1) = \sup_{q\in\widehat{\mathcal{P}}(s_1)} r^\top q = r^\top q^*.$$

By Lemma 10, there exists $\pi^* \in \Pi^{\text{SD}}$ such that $J_r^{\pi^*}(s_1) = \sup_{\pi\in\Pi} J_r^{\pi}(s_1)$. This implies that $r^\top q_{s_1}^{\pi^*} = J_r^{\pi^*}(s_1) = r^\top q^*$. Moreover, we have $q_{s_1}^{\pi^*} \in \widehat{\mathcal{P}}(s_1) \subseteq \mathcal{P}$. Then $q_{s_1}^{\pi^*}$ is also an optimal solution to $\max_{q\in\mathcal{P}} r^\top q$. However, we pick r that induces a unique optimal solution, it follows that $q_{s_1}^{\pi^*} = q^*$. Therefore, any $q^* \in \text{ext}(\mathcal{P})$ is achievable by some $\pi^* \in \Pi$, i.e.,

$$\text{ext}(\mathcal{P}) \subseteq \widehat{\mathcal{P}}(s_1). \tag{9}$$

E.2 Proof of Theorem 1

Recall that

$$\mathcal{P} = \left\{ q \in \mathbb{R}_+^{|\mathcal{S}|\times|\mathcal{A}|} : \sum_{s,a} q(s,a) = 1,\ \sum_{s',a'} P(s|s',a')q(s',a') = \sum_a q(s,a)\ \forall s \right\}$$

and $\widehat{\mathcal{P}}(s_1)$ is the set of occupation measures induced by stationary policies. By Lemma 1,

$$\text{relint}(\mathcal{P}) \cup \text{ext}(\mathcal{P}) \subseteq \widehat{\mathcal{P}}(s_1) \subseteq \mathcal{P}. \tag{10}$$

Based on this, we introduce a well-known fact from the linear programming literature. Since $\mathcal{P}$ is a non-empty polytope and $\text{ext}(\mathcal{P}) \subseteq \widehat{\mathcal{P}}(s_1)$, for any bounded function $\ell : \mathcal{S}\times\mathcal{A} \to \mathbb{R}$

$$\sup_{q\in\widehat{\mathcal{P}}(s_1)} \ell^\top q = \max_{q\in\widehat{\mathcal{P}}(s_1)} \ell^\top q = \max_{q\in\mathcal{P}} \ell^\top q. \tag{11}$$

By the definition of $\widehat{\mathcal{P}}(s_1)$, the primal problem (CMDP) can be rewritten as

$$\begin{array}{ll} \sup\limits_{\pi\in\Pi} & J_r^\pi(s_1) \\ \text{s.t.} & J_g^\pi(s_1) \geq b \end{array} \quad = \quad \begin{array}{ll} \sup\limits_{q\in\widehat{\mathcal{P}}(s_1)} & r^\top q \\ \text{s.t.} & g^\top q \geq b \end{array} \tag{12}$$

Note that (10) and that $\{q \in \widehat{\mathcal{P}}(s_1) : g^\top q \geq b\}$ is non-empty by the assumption. Then by Lemma 14,

$$\begin{array}{ll} \sup\limits_{q\in\widehat{\mathcal{P}}(s_1)} & r^\top q \\ \text{s.t.} & g^\top q \geq b \end{array} \quad = \quad \begin{array}{ll} \max\limits_{q\in\mathcal{P}} & r^\top q \\ \text{s.t.} & g^\top q \geq b \end{array}$$

Then we observe that the right-hand side is an LP, as $\mathcal{P}$ is a polytope. Moreover, we know that its feasible region $\{q \in \mathcal{P} : g^\top q \geq b\}$ is non-empty, because $\{q \in \widehat{\mathcal{P}}(s_1) : g^\top q \geq b\}$ is non-empty and $\{q \in \widehat{\mathcal{P}}(s_1) : g^\top q \geq b\} \subseteq \{q \in \mathcal{P} : g^\top q \geq b\}$. Then by the strong duality of LP (Lemma 11), $\lambda^* \in \arg\min_{\lambda\geq 0} \max_{q\in\mathcal{P}} r^\top q + \lambda(g^\top q - b)$ and

$$\begin{array}{ll} \max\limits_{q\in\mathcal{P}} & r^\top q \\ \text{s.t.} & g^\top q \geq b \end{array} \quad = \quad \min_{\lambda\geq 0} \max_{q\in\mathcal{P}} \quad r^\top q + \lambda(g^\top q - b). \tag{13}$$

By (11), for any $\lambda \geq 0$, we have $\max_{q\in\mathcal{P}} r^\top q + \lambda(g^\top q - b) = \sup_{q\in\widehat{\mathcal{P}}(s_1)} r^\top q + \lambda(g^\top q - b)$. Thus, the minimizer λ^* to the left-hand side is also a minimizer to the right-hand side. Taking $\min_{\lambda\geq 0}$ on both sides yields

$$\min_{\lambda\geq 0} \max_{q\in\mathcal{P}} r^\top q + \lambda(g^\top q - b) = \min_{\lambda\geq 0} \sup_{q\in\widehat{\mathcal{P}}(s_1)} r^\top q + \lambda(g^\top q - b) \tag{14}$$

and

$$\lambda^* \in \arg\min_{\lambda\geq 0} \sup_{q\in\widehat{\mathcal{P}}(s_1)} r^\top q + \lambda(g^\top q - b). \tag{15}$$

By the definition of $\widehat{\mathcal{P}}(s_1)$, we have

$$\min_{\lambda\geq 0} \sup_{\widehat{q}\in\widehat{\mathcal{P}}(s_1)} r^\top \widehat{q} + \lambda(g^\top \widehat{q} - b) \quad = \quad \min_{\lambda\geq 0} \sup_{\pi\in\Pi} J_r^\pi(s_1) + \lambda(J_g^\pi(s_1) - b) \tag{16}$$

The existence of optimal dual variable is guaranteed by (15). Moreover, by (12) - (16), we prove that

$$\begin{array}{ll} \sup\limits_{\pi\in\Pi} & J_r^\pi(s_1) \\ \text{s.t.} & J_g^\pi(s_1) \geq b \end{array} \quad = \quad \min_{\lambda\geq 0} \sup_{\pi\in\Pi} J_r^\pi(s_1) + \lambda(J_g^\pi(s_1) - b).$$

F Missing proofs for Section 4

F.1 Proof of Lemma 3

Since the primal problem (CMDP) is assumed to be feasible, we can use Theorem 1. Thus, by strong duality, there exists $\lambda^* \in \arg\min_{\lambda \geq 0} D(\lambda)$, where $D(\lambda) = \sup_{\pi \in \Pi} J_r^\pi(s_1) + \lambda(J_g^\pi(s_1) - b)$. Moreover, $J_r^* = D(\lambda^*)$, where $J_r^*(s_1) = \sup_{\pi \in \Pi} J_r^\pi(s_1)$ s.t $J_g^\pi(s_1) \geq b$. Then it follows that

$$\begin{aligned} J_r^*(s_1) = D(\lambda^*) &= \sup_{\pi \in \Pi} J_r^\pi(s_1) + \lambda^*(J_g^\pi(s_1) - b) \\ &\geq J_r^{\bar{\pi}}(s_1) + \lambda^*(J_g^{\bar{\pi}}(s_1) - b) \\ &\geq J_r^{\bar{\pi}}(s_1) + \lambda^* \gamma \end{aligned}$$

where $\bar{\pi}$ is the Slater policy satisfying $J_g^{\bar{\pi}}(s_1) \geq b + \gamma$. Since $0 \leq J_r^*(s_1), J_r^{\bar{\pi}}(s_1) \leq 1$, it can be rewritten as

$$\lambda^* \leq \frac{J_r^*(s_1) - J_r^{\bar{\pi}}(s_1)}{\gamma} \leq \frac{1}{\gamma}.$$

F.2 Proof of Lemma 4

Recall that

$$\begin{aligned} \mathbf{w}_{h+1}^k &= \mathbf{\Sigma}_k^{-1} \sum_{\tau=1}^{k-1} \sum_{j=1}^{H} \phi(s_j^\tau, a_j^\tau)(V_{h+1}^k(s_{j+1}^\tau) - \min_{s'} V_{h+1}^k(s')) \\ \mathbf{w}_{h+1}^{k,*} &= \sum_{s \in \mathcal{S}} \psi(s)(V_{h+1}^k(s) - \min_{s'} V_{h+1}^k(s')). \end{aligned}$$

By the definition of linear MDP, we have $\phi(s,a)^\top \mathbf{w}_{h+1}^{k,*} = PV_{h+1}^k(s,a) - \min_{s'} V_{h+1}^k(s')$, as $\sum_{s' \in \mathcal{S}} \phi(s,a)^\top \psi(s') = \sum_{s'} P(s'|s,a) = 1$. Note that $V_{h+1}^k(s) - \min_{s'} V_{h+1}^k(s') \leq \mathrm{sp}(V_{h+1}^k) \leq 2\zeta$ for all $s \in \mathcal{S}$. Then, by Lemmas 17 and 18, we have

$$\|\mathbf{w}_{h+1}^k\|_2 \leq 2\zeta\sqrt{dT}, \quad \|\mathbf{w}_{h+1}^{k,*}\|_2 \leq 2\zeta\sqrt{d}. \tag{17}$$

Fix $(h,k) \in [H] \times [K]$. For simplicity, let $\bar{V}_{h+1}^k(s) = V_{h+1}^k(s) - \min_{s'} V_{h+1}^k(s')$. Note that

$$\begin{aligned} &\left|\phi(s,a)^\top(\mathbf{w}_{h+1}^k - \mathbf{w}_{h+1}^{k,*})\right| \\ &= \left|\phi(s,a)^\top \mathbf{\Sigma}_k^{-1} \left(\sum_{\tau=1}^{k-1} \sum_{j=1}^{H} \phi(s_j^\tau, a_j^\tau)(V_{h+1}^k(s_{j+1}^\tau) - \min_{s'} V_{h+1}^k(s')) - \mathbf{\Sigma}_k \mathbf{w}_{h+1}^{k,*}\right)\right| \\ &\leq \|\phi(s,a)\|_{\mathbf{\Sigma}_k^{-1}} \left\|\sum_{\tau=1}^{k-1} \sum_{j=1}^{H} \phi(s_j^\tau, a_j^\tau) \left((V_{h+1}^k(s_{j+1}^\tau) - \min_{s'} V_{h+1}^k(s')) - \phi(s_j^\tau, a_j^\tau)^\top \mathbf{w}_{h+1}^{k,*}\right)\right\|_{\mathbf{\Sigma}_k^{-1}} \\ &\quad + \left|\phi(s,a)^\top \mathbf{\Sigma}_k^{-1} \mathbf{w}_{h+1}^{k,*}\right| \\ &= \|\phi(s,a)\|_{\mathbf{\Sigma}_k^{-1}} \left\|\sum_{\tau=1}^{k-1} \sum_{j=1}^{H} \phi(s_j^\tau, a_j^\tau) \left(\bar{V}_{h+1}^k(s_{j+1}^\tau) - P\bar{V}_{h+1}^k(s_j^\tau, a_j^\tau)\right)\right\|_{\mathbf{\Sigma}_k^{-1}} \\ &\quad + \left|\phi(s,a)^\top \mathbf{\Sigma}_k^{-1} \mathbf{w}_{h+1}^{k,*}\right| \end{aligned}$$

where the inequality follows from the fact that $\mathbf{\Sigma}_k = I + \sum_{\tau=1}^{k-1} \sum_{j=1}^{H} \phi(s_j^\tau, a_j^\tau)\phi(s_j^\tau, a_j^\tau)^\top$ and the triangle inequality. By (17), we have

$$\left|\phi(s,a)^\top \mathbf{\Sigma}_k^{-1} \mathbf{w}_{h+1}^{k,*}\right| \leq \|\mathbf{w}_{h+1}^{k,*}\|_{\mathbf{\Sigma}_k^{-1}} \|\phi(s,a)\|_{\mathbf{\Sigma}_k^{-1}} \leq 2\zeta\sqrt{d}\|\phi(s,a)\|_{\mathbf{\Sigma}_k^{-1}}.$$

By Lemma 19, with probability at least $1-\delta/T$,

$$\left\|\sum_{\tau=1}^{k-1}\sum_{j=1}^{H}\phi(s_j^\tau,a_j^\tau)\left(\bar{V}_{h+1}^k(s_{j+1}^\tau)-P\bar{V}_{h+1}^k(s_j^\tau,a_j^\tau)\right)\right\|_{\mathbf{\Sigma}_k^{-1}}^2$$
$$\leq 16\zeta^2\left[\frac{d}{2}\log(T+1)+\log\frac{\mathcal{N}_\epsilon T}{\delta}\right]+8T^2\epsilon^2.$$

where $\mathcal{N}_\epsilon$ denotes the ϵ-covering number of the value function class, which is the set of all possible $\bar{V}_{h+1}^k$. Note that the centering mapping $V(\cdot)\mapsto V(\cdot)-\min_{s'}V(s')$ is 2-Lipschitz. Therefore, this transformation affects the covering number only up to constant factors. Then, to bound $\log(\mathcal{N}_\epsilon)$, we first deduce that

$$r(s,a)+\lambda_k g(s,a)+\phi(s,a)^\top\mathbf{w}_{h+1}^k=\phi(s,a)^\top\left(\theta_r+\lambda_k\theta_g+\mathbf{w}_{h+1}^k\right).$$

It follows that

$$\|\theta_r+\lambda_k\theta_g+\mathbf{w}_{h+1}^k\|_2\leq\sqrt{d}+\frac{2}{\gamma}\sqrt{d}+2\zeta\sqrt{dT}\leq\frac{3}{\gamma}\sqrt{d}+2\zeta\sqrt{dT}.$$

Then, by Lemma 20,

$$\log(\mathcal{N}_\epsilon)\leq d\log\left(1+\frac{8\sqrt{d}(3/\gamma+2\zeta\sqrt{T})}{\epsilon}\right)+\log(1+4H(1+2/\gamma)/\epsilon)$$
$$+d^2\log\left(1+8d^{1/2}\beta^2/\epsilon^2\right).$$

It follows that

$$\left\|\sum_{\tau=1}^{k-1}\sum_{j=1}^{H}\phi(s_j^\tau,a_j^\tau)\left(\bar{V}_{h+1}^k(s_{j+1}^\tau)-P\bar{V}_{h+1}^k(s_j^\tau,a_j^\tau)\right)\right\|_{\mathbf{\Sigma}_k^{-1}}^2$$
$$\leq 16\zeta^2\left[\frac{d}{2}\log(T+1)+d\log\left(1+\frac{8\sqrt{d}(3/\gamma+2\zeta\sqrt{T})}{\epsilon}\right)\right.$$
$$\left.+\log(1+4H(1+2/\gamma)/\epsilon)+d^2\log\left(1+8d^{1/2}\beta^2/\epsilon^2\right)+\log\frac{T}{\delta}\right]+8T^2\epsilon^2.$$

By taking $\epsilon=d/T$, it follows that

$$\left\|\sum_{\tau=1}^{k-1}\sum_{j=1}^{H}\phi(s_j^\tau,a_j^\tau)\left(\bar{V}_{h+1}^k(s_{j+1}^\tau)-P\bar{V}_{h+1}^k(s_j^\tau,a_j^\tau)\right)\right\|_{\mathbf{\Sigma}_k^{-1}}^2$$
$$\leq 16\zeta^2\left[\frac{d}{2}\log(2T)+d\log\left(16T^2(3/\gamma+2\zeta)\right)+\log(8H(1+2/\gamma)T)+d^2\log\left(16\beta^2T^2\right)+\log\frac{T}{\delta}\right]+8d^2$$
$$\leq 72\zeta^2d^2\log\left(16HT^2\beta^2(3/\gamma+2\zeta)/\delta\right)+8d^2.$$

Consider $\beta=c_\beta(\zeta+1)d\log(16HT^2(1/\gamma+2\zeta)/\delta)$. Then there exists $c_\beta=\widetilde{\mathcal{O}}(1)$ such that

$$\sqrt{72\zeta^2d^2\log\left(16HT^2\beta^2(3/\gamma+2\zeta)/\delta\right)+8d^2}+2\zeta\sqrt{d}\leq\beta.$$

This implies that by taking $\beta=\widetilde{\mathcal{O}}((\zeta+1)d)$, where $\widetilde{\mathcal{O}}(\cdot)$ hides polynomial factors in $\log(dT|\mathcal{A}|\zeta/(\delta\gamma))$, we have with probability at least $1-\delta/T$, for any $(s,a)\in\mathcal{S}\times\mathcal{A}$,

$$|\phi(s,a)^\top(\mathbf{w}_{h+1}^k-\mathbf{w}_{h+1}^{k,*})|\leq\beta\|\phi(s,a)\|_{\mathbf{\Sigma}_k^{-1}}.$$

By taking union bound over all $(h,k)\in[H]\times[K]$, the statement holds for all (h,k), with probability at least $1-\delta$.

F.3 Proof of Lemma 5

Fix k. For all $s \in \mathcal{S}$, we have

$$V^*_{r,H+1}(s) + \lambda_k V^*_{g,H+1}(s) = V^k_{H+1}(s) = 0.$$

Assume that $V^*_{r,h+1}(s) + \lambda_k V^*_{g,h+1}(s) \le V^k_{h+1}(s)$. Note that for any (s,a)

$$\begin{aligned}
&Q^*_{r,h}(s,a) + \lambda_k Q^*_{g,h}(s,a) \\
&= r(s,a) + \lambda_k g(s,a) + P\left(V^*_{r,h+1} + \lambda_k V^*_{g,h+1}\right)(s,a) \\
&\le r(s,a) + \lambda_k g(s,a) + PV^k_{h+1}(s,a) \\
&\le r(s,a) + \lambda_k g(s,a) + \phi(s,a)^\top \mathbf{w}^k_{h+1} + \min_{s'} V^k_{h+1}(s') \\
&\quad + |PV^k_{h+1}(s,a) - \min_{s'} V^k_{h+1}(s') - \phi(s,a)^\top \mathbf{w}^k_{h+1}| \\
&= r(s,a) + \lambda_k g(s,a) + \phi(s,a)^\top \mathbf{w}^k_{h+1} + \min_{s'} V^k_{h+1}(s') \\
&\quad + |\phi(s,a)^\top (\mathbf{w}^{k,*}_{h+1} - \mathbf{w}^k_{h+1})| \\
&\le r(s,a) + \lambda_k g(s,a) + \phi(s,a)^\top \mathbf{w}^k_{h+1} + \min_{s'} V^k_{h+1}(s') + \beta \|\phi(s,a)\|_{\mathbf{\Sigma}_k^{-1}} \\
&\le \left(r(s,a) + \lambda_k g(s,a) + \phi(s,a)^\top \mathbf{w}^k_{h+1} + \min_{s'} V^k_{h+1}(s') + \beta \|\phi(s,a)\|_{\mathbf{\Sigma}_k^{-1}}\right) \wedge H(1 + \frac{2}{\gamma}) \\
&= Q^k_h(s,a)
\end{aligned}$$

where the second inequality is due to the triangle inequality, the second equality follows from the definition of $\mathbf{w}^{k,*}_{h+1}$, the third inequality follows from Lemma 4, and the fourth inequality follows from the fact that $Q^*_{r,h}(s,a) + \lambda_k Q^*_{g,h}(s,a) \le H(1 + 2/\gamma)$. Moreover,

$$V^*_{r,h}(s) + \lambda_k V^*_{g,h}(s) \le \max_a \; Q^*_{r,h}(s,a) + \lambda_k Q^*_{g,h}(s,a) \le \max_a \; Q^k_h(s,a) = \widetilde{V}^k_h(s).$$

Note that $J^*_r(s), J^*_g(s)$ are constants in $s \in \mathcal{S}$. Then, by Lemma 2, we have for any $s, s' \in \mathcal{S}$ and $h \in [H]$,

$$|V^*_{r,h}(s) - V^*_{r,h}(s')| \le |V^*_{r,h}(s) - (H-h+1)J^*_r| + |V^*_{r,h}(s') - (H-h+1)J^*_r| \le 2\mathrm{sp}(v^*_r).$$

This implies that, for all $h \in [H]$, $\mathrm{sp}(V^*_{r,h}) \le 2\mathrm{sp}(v^*_r)$, and similarly, we have $\mathrm{sp}(V^*_{g,h}) \le 2\mathrm{sp}(v^*_g)$. Moreover, it follows that $\mathrm{sp}(V^*_{r,h} + \lambda_k V^*_{g,h}) \le \mathrm{sp}(V^*_{r,h}) + \lambda_k \mathrm{sp}(V^*_{g,h}) \le 2\mathrm{sp}(v^*_r) + \frac{4}{\gamma}\mathrm{sp}(v^*_g) = 2\zeta$. Together with the fact that $V^*_{r,h}(s) + \lambda_k V^*_{g,h}(s) \le \widetilde{V}^k_h(s)$, we have

$$V^*_{r,h}(s) + \lambda_k V^*_{g,h}(s) \le \widetilde{V}^k_h(s) \wedge (\min_{s'} \widetilde{V}^k_h(s') + 2\zeta) = V^k_h(s).$$

By induction, we have for all (s,h,k),

$$V^*_{r,h}(s) + \lambda_k V^*_{g,h}(s) \le V^k_h(s)$$

as desired.

F.4 Decomposition

In this subsection, we provide the detailed decomposition (4) of $\mathrm{Regret}(T) + \lambda \mathrm{Violation}(T)$, which is introduced in Section 4.3. Throughout the decomposition procedure, we assume Lemma 4, which holds with high probability. Then $\mathrm{Regret}(T) + \lambda \mathrm{Violation}(T)$ can be decomposed as follows.

$$\begin{aligned}
&\sum_{t=1}^T (J^*_r(s_1) - r(s_t,a_t)) + \lambda \sum_{t=1}^T (b - g(s_t,a_t)) \qquad (18) \\
&= \sum_{k=1}^K \sum_{h=1}^H \left(J^*_r(s_1) - r(s^k_h, a^k_h)\right) + \lambda \sum_{k=1}^K \sum_{h=1}^H (b - g(s^k_h, a^k_h)) \\
&\le \underbrace{\sum_{k=1}^K \sum_{h=1}^H \left[\left(J^*_r(s_1) - r(s^k_h,a^k_h)\right) + \lambda_k \left(J^*_g(s_1) - g(s^k_h,a^k_h)\right)\right]}_{\text{(I)}} + \underbrace{\sum_{k=1}^K (\lambda - \lambda_k) \left(Hb - \sum_{h=1}^H g(s^k_h, a^k_h)\right)}_{\text{(II)}}
\end{aligned}$$

where the inequality follows from the fact that $J_g^*(s_1) \geq b$ and $\lambda_k \geq 0$. Note that (II) can be simply bounded as follows. Due to the dual update rule, for any $\lambda \in [0, \frac{2}{\gamma}]$,

$$\begin{aligned}(\lambda_{k+1} - \lambda)^2 &\leq \left(\lambda_k + \eta\left(\sum_{h=1}^H (b - g(s_h^k, a_h^k))\right) - \lambda\right)^2 \\ &= (\lambda_k - \lambda)^2 + \eta^2\left(\sum_{h=1}^H (b - g(s_h^k, a_h^k))\right)^2 + 2\eta(\lambda_k - \lambda)\left(\sum_{h=1}^H (b - g(s_h^k, a_h^k))\right).\end{aligned}$$

where the first inequality is due to the fact that $[\cdot]_{[0,\frac{2}{\gamma}]}$ is a contraction operator. By rearranging it and summing over $k = 1, \ldots, K$, we have

$$\begin{aligned}\text{(II)} &\leq \frac{\lambda^2}{2\eta} + \frac{\eta\sum_{k=1}^K\left(\sum_{h=1}^H (b - g(s_h^k, a_h^k))\right)^2}{2} \\ &\leq \frac{\lambda^2}{2\eta} + \frac{\eta K H^2}{2}.\end{aligned} \tag{19}$$

where the second inequality follows from the fact that $|b - g(s_h^k, a_h^k)| \leq 1$. Next, we focus on bounding (I). Note that our design of Q_h^k satisfies that

$$Q_h^k(s_h^k, a_h^k) \leq r(s_h^k, a_h^k) + \lambda_k g(s_h^k, a_h^k) + \phi(s_h^k, a_h^k)^\top \mathbf{w}_{h+1}^k + \min_{s'} V_{h+1}^k(s') + \beta\|\phi(s_h^k, a_h^k)\|_{\mathbf{\Sigma}_k^{-1}}. \tag{20}$$

Let $V_{r,h}^*, V_{g,h}^*$ denote the value functions with respect to π^*. Then (I) can be bounded as follows.

$$\begin{aligned}\text{(I)} &= \sum_{h=1}^H \left(J_r^*(s_1) - r(s_h^k, a_h^k)\right) + \sum_{h=1}^H \lambda_k \left(J_g^*(s_1) - g(s_h^k, a_h^k)\right) \\ &\leq \sum_{h=1}^H \left(J_r^*(s_1) + \lambda_k J_g^*(s_1) - Q_h^k(s_h^k, a_h^k) + \phi(s_h^k, a_h^k)^\top \mathbf{w}_{h+1}^k + \min_{s'} V_{h+1}^k(s') + \beta\|\phi(s_h^k, a_h^k)\|_{\mathbf{\Sigma}_k^{-1}}\right) \\ &\leq H J_r^*(s_1) - V_{r,1}^*(s_1^k) + \lambda_k\left(H J_g^*(s_1) - V_{g,1}^*(s_1^k)\right) \\ &\quad + V_{r,1}^*(s_1^k) + \lambda_k V_{g,1}^*(s_1^k) + \sum_{h=1}^H \left(-Q_h^k(s_h^k, a_h^k) + P V_{h+1}^k(s_h^k, a_h^k)\right) + \sum_{h=1}^H 2\beta\|\phi(s_h^k, a_h^k)\|_{\mathbf{\Sigma}_k^{-1}}\end{aligned}$$

where the first inequality follows from (20), and the last inequality is because Lemma 4 implies that $\phi(s_h^k, a_h^k)^\top \mathbf{w}_{h+1}^k + \min_{s'} V_{h+1}^k(s') \leq P V_{h+1}^k(s_h^k, a_h^k) + \beta\|\phi(s_h^k, a_h^k)\|_{\mathbf{\Sigma}_k^{-1}}$. Note that J_r^*, J_g^* are assumed to be constant across $s \in \mathcal{S}$ by Assumption 2. Then, by Lemma 2, we have for all $h \in [H]$,

$$|(H - h + 1)J_r^* - V_{r,h}^*(s)| \leq \mathrm{sp}(v_r^*),\ |(H - h + 1)J_g^* - V_{g,h}^*(s)| \leq \mathrm{sp}(v_g^*). \tag{21}$$

By taking $h = 1$ for (21), it leads to

$$\begin{aligned}&H J_r^*(s_1) - V_{r,1}^*(s_1^k) + \lambda_k\left(H J_g^*(s_1) - V_{g,1}^*(s_1^k)\right) \\ &\quad + V_{r,1}^*(s_1^k) + \lambda_k V_{g,1}^*(s_1^k) + \sum_{h=1}^H \left(-Q_h^k(s_h^k, a_h^k) + P V_{h+1}^k(s_h^k, a_h^k)\right) + \sum_{h=1}^H 2\beta\|\phi(s_h^k, a_h^k)\|_{\mathbf{\Sigma}_k^{-1}} \\ &\leq \mathrm{sp}(v_r^*) + \lambda_k \mathrm{sp}(v_g^*) \\ &\quad + V_{r,1}^*(s_1^k) + \lambda_k V_{g,1}^*(s_1^k) + \sum_{h=1}^H \left(-Q_h^k(s_h^k, a_h^k) + P V_{h+1}^k(s_h^k, a_h^k)\right) + \sum_{h=1}^H 2\beta\|\phi(s_h^k, a_h^k)\|_{\mathbf{\Sigma}_k^{-1}} \\ &\leq \mathrm{sp}(v_r^*) + \frac{2}{\gamma}\mathrm{sp}(v_g^*) \\ &\quad + \underbrace{V_{r,1}^*(s_1^k) + \lambda_k V_{g,1}^*(s_1^k) + \sum_{h=1}^H \left(-Q_h^k(s_h^k, a_h^k) + P V_{h+1}^k(s_h^k, a_h^k)\right)}_{\text{(III)}} + \sum_{h=1}^H 2\beta\|\phi(s_h^k, a_h^k)\|_{\mathbf{\Sigma}_k^{-1}}.\end{aligned}$$

where the last inequality follows from $\lambda_k \leq \frac{2}{\gamma}$ for all k. Term (III) can be further bounded as follows.

$$
\begin{aligned}
\text{(III)} &= V^*_{r,1}(s^k_1) + \lambda_k V^*_{g,1}(s^k_1) + \sum_{h=1}^{H} \left(-Q^k_h(s^k_h, a^k_h) + PV^k_{h+1}(s^k_h, a^k_h)\right) \\
&= V^*_{r,1}(s^k_1) + \lambda_k V^*_{g,1}(s^k_1) - V^k_1(s^k_1) + V^k_1(s^k_1) + \sum_{h=1}^{H} \left(-Q^k_h(s^k_h, a^k_h) + PV^k_{h+1}(s^k_h, a^k_h)\right) \\
&= V^*_{r,1}(s^k_1) + \lambda_k V^*_{g,1}(s^k_1) - V^k_1(s^k_1) \\
&\quad + V^k_1(s^k_1) - Q^k_1(s^k_1, a^k_1) + PV^k_2(s^k_1, a^k_1) - V^k_2(s^k_2) + V^k_2(s^k_2) + \sum_{h=2}^{H} \left(-Q^k_h(s^k_h, a^k_h) + PV^k_{h+1}(s^k_h, a^k_h)\right) \\
&\quad \vdots \\
&= V^*_{r,1}(s^k_1) + \lambda_k V^*_{g,1}(s^k_1) - V^k_1(s^k_1) + \sum_{h=1}^{H} \left(V^k_h(s^k_h) - Q^k_h(s^k_h, a^k_h)\right) + \sum_{h=1}^{H} \left(PV^k_{h+1}(s^k_h, a^k_h) - V^k_{h+1}(s^k_{h+1})\right)
\end{aligned}
$$

This implies that for each $k \in [K]$,

$$
\begin{aligned}
\text{(I)} &\leq \mathrm{sp}(v^*_r) + \frac{2}{\gamma}\mathrm{sp}(v^*_g) \\
&\quad + V^*_{r,1}(s^k_1) + \lambda_k V^*_{g,1}(s^k_1) - V^k_1(s^k_1) + \sum_{h=1}^{H} \left(V^k_h(s^k_h) - Q^k_h(s^k_h, a^k_h)\right) + \sum_{h=1}^{H} \left(PV^k_{h+1}(s^k_h, a^k_h) - V^k_{h+1}(s^k_{h+1})\right) \\
&\quad + \sum_{h=1}^{H} 2\beta \|\phi(s^k_h, a^k_h)\|_{\mathbf{\Sigma}_k^{-1}}.
\end{aligned} \tag{22}
$$

Note that taking sum over $k = 1, \ldots, K$ to (22) yields the first term in (18). In addition to this, by applying (19), we have for any $\lambda \in [0, \frac{2}{\gamma}]$,

$$
\begin{aligned}
&\sum_{t=1}^{T} \left(J^*_r(s_1) - r(s_t, a_t)\right) + \lambda \sum_{t=1}^{T} (b - g(s_t, a_t)) \\
&\leq K \left(\mathrm{sp}(v^*_r) + \frac{2}{\gamma}\mathrm{sp}(v^*_g)\right) + \sum_{k=1}^{K}\sum_{h=1}^{H} 2\beta \|\phi(s^k_h, a^k_h)\|_{\mathbf{\Sigma}_k^{-1}} \\
&\quad + \sum_{k=1}^{K} \left(V^*_{r,1}(s^k_1) + \lambda_k V^*_{g,1}(s^k_1) - V^k_1(s^k_1)\right) \\
&\quad + \sum_{k=1}^{K}\sum_{h=1}^{H} \left(V^k_h(s^k_h) - Q^k_h(s^k_h, a^k_h)\right) + \sum_{k=1}^{K}\sum_{h=1}^{H} \left(PV^k_{h+1}(s^k_h, a^k_h) - V^k_{h+1}(s^k_{h+1})\right) \\
&\quad + \frac{\lambda^2}{2\eta} + \frac{\eta K H^2}{2}.
\end{aligned}
$$

Note that $V^k_h(s^k_h) \leq \widetilde{V}^k_h(s^k_h) = Q^k_h(s^k_h, a^k_h)$. It follows that $\sum_{k=1}^{K}\sum_{h=1}^{H} \left(V^k_h(s^k_h) - Q^k_h(s^k_h, a^k_h)\right) \leq 0$. This concludes the proof.

F.5 Proof of Lemma 6

Assume that the statement of Lemma 4 holds, which occurs with probability at least $1-\delta$. By (4), we have

$$\begin{aligned}\mathrm{Regret}(T)+\lambda\mathrm{Violation}(T) \le K\zeta &+ \underbrace{\sum_{k=1}^{K}\sum_{h=1}^{H} 2\beta\|\phi(s_h^k,a_h^k)\|_{\mathbf{\Sigma}_k^{-1}}}_{\text{(I)}} \\ &+ \underbrace{\sum_{k=1}^{K}\left(V_{r,1}^*(s_1^k)+\lambda_k V_{g,1}^*(s_1^k)-V_1^k(s_1^k)\right)}_{\text{(II)}} \\ &+ \underbrace{\sum_{k=1}^{K}\sum_{h=1}^{H}\left(PV_{h+1}^k(s_h^k,a_h^k)-V_{h+1}^k(s_{h+1}^k)\right)}_{\text{(III)}} + \frac{\lambda^2}{2\eta}+\frac{\eta KH^2}{2}.\end{aligned} \tag{23}$$

By Lemma 5, we have

$$\text{(II)} \le 0. \tag{24}$$

Moreover, since (III) is the sum of a martingale difference sequence, it can be bounded using the Azuma-Hoeffding inequality. Note that $\mathrm{sp}(V_{h+1}^k) \le 2\mathrm{sp}(v_r^*)+\frac{4}{\gamma}\mathrm{sp}(v_g^*) = 2\zeta$. Then, with probability at least $1-\delta$,

$$\text{(III)} \le 2\zeta\sqrt{2KH\log(2/\delta)}. \tag{25}$$

Now, we bound (I). The standard elliptical potential lemma cannot be directly applied due to the definition of $\mathbf{\Sigma}_k$, i.e., $\mathbf{\Sigma}_{k+1} \neq \mathbf{\Sigma}_k + \phi(s_h^k,a_h^k)\phi(s_h^k,a_h^k)^\top$. Therefore, our proof uses a similar technique from the proof of Theorem 5 in Wei et al. (2021), which decomposes the desired sum into two terms as follows.

$$\begin{aligned}&\sum_{k=1}^{K}\sum_{h=1}^{H}\|\phi(s_h^k,a_h^k)\|_{\mathbf{\Sigma}_k^{-1}} \\ &= \sum_{k:2\det(\mathbf{\Sigma}_k)\ge\det(\mathbf{\Sigma}_{k+1})}\sum_{h=1}^{H}\|\phi(s_h^k,a_h^k)\|_{\mathbf{\Sigma}_k^{-1}} + \sum_{k:2\det(\mathbf{\Sigma}_k)<\det(\mathbf{\Sigma}_{k+1})}\sum_{h=1}^{H}\|\phi(s_h^k,a_h^k)\|_{\mathbf{\Sigma}_k^{-1}}.\end{aligned}$$

Note that

$$\begin{aligned}\sum_{k:2\det(\mathbf{\Sigma}_k)\ge\det(\mathbf{\Sigma}_{k+1})}\sum_{h=1}^{H}\|\phi(s_h^k,a_h^k)\|_{\mathbf{\Sigma}_k^{-1}} &\le \sqrt{2}\sum_{k:2\det(\mathbf{\Sigma}_k)\ge\det(\mathbf{\Sigma}_{k+1})}\sum_{h=1}^{H}\|\phi(s_h^k,a_h^k)\|_{\mathbf{\Sigma}_{k+1}^{-1}} \\ &\le \sqrt{2T}\sqrt{\sum_{k=1}^{K}\sum_{h=1}^{H}\|\phi(s_h^k,a_h^k)\|^2_{\mathbf{\Sigma}_{k+1}^{-1}}}.\end{aligned}$$

Since $\mathbf{\Sigma}_{k+1} = I+\sum_{\tau=1}^{k}\sum_{h=1}^{H}\phi(s_h^\tau,a_h^\tau)\phi(s_h^\tau,a_h^\tau)^\top \succeq \bar{\mathbf{\Sigma}}_{k,h}$, where $\bar{\mathbf{\Sigma}}_{k,h}$ is defined as $\bar{\mathbf{\Sigma}}_{k,h} = I+\sum_{\tau=1}^{k-1}\sum_{h'=1}^{H}\phi(s_{h'}^\tau,a_{h'}^\tau)\phi(s_{h'}^\tau,a_{h'}^\tau)^\top+\sum_{h''=1}^{h}\phi(s_{h''}^k,a_{h''}^k)\phi(s_{h''}^k,a_{h''}^k)^\top$. Then it follows that

$$\sum_{k=1}^{K}\sum_{h=1}^{H}\|\phi(s_h^k,a_h^k)\|^2_{\mathbf{\Sigma}_{k+1}^{-1}} \le \sum_{k=1}^{K}\sum_{h=1}^{H}\|\phi(s_h^k,a_h^k)\|^2_{\bar{\mathbf{\Sigma}}_{k,h}^{-1}} \le 2\log\frac{\det(\mathbf{\Sigma}_{K+1})}{\det(\mathbf{\Sigma}_1)} \le 2d\log(T+1).$$

where the first inequality follows from that $\|\cdot\|_{\mathbf{\Sigma}_{k+1}^{-1}} \le \|\cdot\|_{\bar{\mathbf{\Sigma}}_{k,h}^{-1}}$, the second inequality follows from Lemma 21, and the last inequality follows from that $\mathbf{\Sigma}_1 = I$ and

$$\det(\mathbf{\Sigma}_{K+1}) \le \|\mathbf{\Sigma}_{K+1}\|_2^d \le \left(1+\sum_{k,h}\|\phi(s_h^k,a_h^k)\|_2\right)^d \le (T+1)^d.$$

It follows that

$$\sum_{k:2\det(\mathbf{\Sigma}_k)\geq\det(\mathbf{\Sigma}_{k+1})}\sum_{h=1}^{H}\|\phi(s_h^k,a_h^k)\|_{\mathbf{\Sigma}_k^{-1}}\leq 2\sqrt{dT\log(T+1)}.$$

Moreover, by Lemma 22, $|\{k\in[K]:2\det(\mathbf{\Sigma}_k)<\det(\mathbf{\Sigma}_{k+1})\}|\leq\frac{3}{2}d\log(1+T)$. Since $\|\phi(s_h^k,a_h^k)\|_{\mathbf{\Sigma}_k^{-1}}\leq 1$, it follows that

$$\sum_{k:2\det(\mathbf{\Sigma}_k)<\det(\mathbf{\Sigma}_{k+1})}\sum_{h=1}^{H}\|\phi(s_h^k,a_h^k)\|_{\mathbf{\Sigma}_k^{-1}}\leq\frac{3}{2}dH\log(1+T).$$

Then we have

$$\sum_{k=1}^{K}\sum_{h=1}^{H}\|\phi(s_h^k,a_h^k)\|_{\mathbf{\Sigma}_k^{-1}}\leq 2\sqrt{dT\log(T+1)}+\frac{3}{2}dH\log(1+T).$$

Thus, we have

$$\text{(I)}\leq 4\beta\sqrt{dT\log(T+1)}+3\beta dH\log(1+T). \tag{26}$$

By applying (24), (25), and (26) to (23), for all $\lambda\in[0,\frac{2}{\gamma}]$,

$$\begin{aligned}\text{Regret}(T)+\lambda\text{Violation}(T)&\leq 1+K\zeta+4\beta\sqrt{dT\log(T+1)}+3\beta dH\log(1+T)\\&\quad+2\zeta\sqrt{2KH\log(2/\delta)}+\frac{\lambda^2}{2\eta}+\frac{\eta KH^2}{2}\\&\leq 1+K\zeta+4\beta\sqrt{dT\log(T+1)}+3\beta dH\log(1+T)\\&\quad+2\zeta\sqrt{2KH\log(2/\delta)}+\frac{(2/\gamma)^2}{2\eta}+\frac{\eta KH^2}{2}.\end{aligned}$$

Recall that $K=T^{2/3}$, $H=T^{1/3}$, $\beta=\widetilde{\mathcal{O}}((\zeta+1)d)$, $\eta=\frac{2}{\gamma\sqrt{KH^2}}$. It follows that for all $\lambda\in[0,\frac{2}{\gamma}]$

$$\text{Regret}(T)+\lambda\text{Violation}(T)=\widetilde{\mathcal{O}}\left(\left(\zeta+\frac{2}{\gamma}\right)T^{2/3}+(\zeta+1)(\sqrt{d^3T}+d^2T^{1/3})\right).$$

F.6 Proof of Lemma 7

By Theorem 1 and Lemma 3, there exists a bounded optimal dual variable λ^*. Given λ^*, consider a reward function $r+\lambda^*(g-b):\mathcal{S}\times\mathcal{A}\to\mathbb{R}$ such that $[r+\lambda^*(g-b)](s,a)=r(s,a)+\lambda^*(g(s,a)-b)$ for each $(s,a)\in\mathcal{S}\times\mathcal{A}$. Then, by the Bellman optimality condition for weakly communicating unconstrained MDPs, there exists $J_{\lambda^*}^*(s_1)\in\mathbb{R}$, $v_{\lambda^*}^*:\mathcal{S}\to\mathbb{R}$ such that for any $(s,a)\in\mathcal{S}\times\mathcal{A}$,

$$r(s,a)+\lambda^*(g(s,a)-b)+Pv_{\lambda^*}^*(s,a)-J_{\lambda^*}^*(s_1)\leq v_{\lambda^*}^*(s). \tag{27}$$

Next, we show that $J_r^*(s_1)=J_{\lambda^*}^*(s_1)$. Note that

$$\begin{aligned}D(\lambda^*)&=\sup_{\pi\in\Pi}J_r^\pi(s_1)+\lambda^*(J_g^\pi(s_1)-b)\\&=\sup_{\pi\in\Pi}J_{r+\lambda^*(g-b)}^\pi(s_1)\\&=\max_{\pi\in\Pi}J_{r+\lambda^*(g-b)}^\pi(s_1)\\&=J_{\lambda^*}^*(s_1)\end{aligned}$$

where the second equality follows from that for any stationary policy $J_{r+\lambda^*(g-b)}^\pi(s_1)=(r+\lambda^*(g-b\mathbf{1}))^\top q_{s_1}^\pi=r^\top q_{s_1}^\pi+\lambda^*(g^\top q_{s_1}^\pi-b)=J_r^\pi(s_1)+\lambda^*(J_g^\pi(s_1)-b)$ with the all-1 vector $\mathbf{1}\in\mathbb{R}^{|\mathcal{S}|\times|\mathcal{A}|}$, and the third equality is due to the fact that an optimal stationary policy is attainable for weakly communicating MDPs. Moreover, by Theorem 1, we have $J_r^*(s_1)=D(\lambda^*)$. Therefore, it follows that $J_r^*(s_1)=J_{\lambda^*}^*(s_1)$ as desired. Applying this to (27), it can be written as

$$r(s,a)+\lambda^*(g(s,a)-b)+Pv_{\lambda^*}^*(s,a)-J_r^*(s_1)\leq v_{\lambda^*}^*(s).$$

Given the trajectory $\{(s_t, a_t)\}_{t=1}^T$ generated by our algorithm, taking $s = s_t, a = a_t$ yields

$$\begin{aligned} r(s_t, a_t) - J_r^*(s_1) &\le \lambda^*(b - g(s_t, a_t)) + v_{\lambda^*}^*(s_t) - P v_{\lambda^*}^*(s_t, a_t) \\ &= \lambda^*(b - g(s_t, a_t)) + v_{\lambda^*}^*(s_{t+1}) - P v_{\lambda^*}^*(s_t, a_t) + v_{\lambda^*}^*(s_t) - v_{\lambda^*}^*(s_{t+1}) \end{aligned}$$

Summing over $t = 1, \dots, T$ yields

$$\begin{aligned} \sum_{t=1}^T (r(s_t, a_t) - J_r^*(s_1)) \le \lambda^* \sum_{t=1}^T (b - g(s_t, a_t)) \\ + \underbrace{\sum_{t=1}^T v_{\lambda^*}^*(s_{t+1}) - P v_{\lambda^*}^*(s_t, a_t)}_{\text{(I)}} + \underbrace{\sum_{t=1}^T v_{\lambda^*}^*(s_t) - v_{\lambda^*}^*(s_{t+1})}_{\text{(II)}}. \end{aligned}$$

For (I), since it is a martingale difference sequence, by the Azuma-Hoeffding inequality, we have (I) $\le \mathrm{sp}(v_{\lambda^*}^*)\sqrt{T \log(1/\delta)}$ with probability at least $1 - \delta$. For (II), since it is a telescoping sum, we have (II) $\le v_{\lambda^*}^*(s_1) - v_{\lambda^*}^*(s_{T+1}) \le \mathrm{sp}(v_{\lambda^*}^*)$. Multiplying both sides by (-1), with probability at least $1 - \delta$,

$$\sum_{t=1}^T (J_r^*(s_1) - r(s_t, a_t)) \ge -\lambda^* \sum_{t=1}^T (b - g(s_t, a_t)) - \mathrm{sp}(v_{\lambda^*}^*)\sqrt{2T \log(1/\delta)} - \mathrm{sp}(v_{\lambda^*}^*).$$

Since $\mathrm{Regret}(T) = \sum_{t=1}^T (J_r^*(s_1) - r(s_t, a_t))$ and $\mathrm{Violation}(T) = \sum_{t=1}^T (b - g(s_t, a_t))$, this completes the proof.

F.7 Proof of Theorem 2

Suppose that Lemma 4 holds, which occurs with probability at least $1 - \delta$. Then, by Lemma 6, for any $\lambda \in [0, \frac{2}{\gamma}]$, with probability at least $1 - \delta$,

$$\sum_{t=1}^T (J_r^* - r(s_t, a_t)) + \lambda \sum_{t=1}^T (b - g(s_t, a_t)) = \widetilde{\mathcal{O}}\left(\left(\zeta + \frac{2}{\gamma}\right) T^{2/3} + (\zeta + 1)(\sqrt{d^3 T} + d^2 T^{1/3})\right). \tag{28}$$

By taking $\lambda = 0$,

$$\sum_{t=1}^T (J_r^* - r(s_t, a_t)) = \widetilde{\mathcal{O}}\left(\left(\zeta + \frac{2}{\gamma}\right) T^{2/3} + (\zeta + 1)(\sqrt{d^3 T} + d^2 T^{1/3})\right).$$

If $\sum_{t=1}^T (b - g(s_t, a_t)) \le 0$, then the proof is finished. Otherwise, we show a violation upper bound as follows. With probability at least $1 - \delta$, by Lemma 7,

$$\begin{aligned} &\sum_{t=1}^T (J_r^* - r(s_t, a_t)) + \frac{2}{\gamma} \sum_{t=1}^T (b - g(s_t, a_t)) \\ &\ge \left(\frac{2}{\gamma} - \lambda^*\right) \sum_{t=1}^T (b - g(s_t, a_t)) - \mathrm{sp}(v_{\lambda^*}^*)\sqrt{T \log(1/\delta)} - \mathrm{sp}(v_{\lambda^*}^*) \qquad (29) \\ &\ge \frac{1}{\gamma} \sum_{t=1}^T (b - g(s_t, a_t)) - \mathrm{sp}(v_{\lambda^*}^*)\sqrt{T \log(1/\delta)} - \mathrm{sp}(v_{\lambda^*}^*) \end{aligned}$$

where the second inequality follows from $\lambda^* \le 1/\gamma$ (Lemma 3) and taking $\lambda = \frac{2}{\gamma}$. By union bound, Lemma 4, (28), and (29) hold with probability at least $1 - 3\delta$. Finally, with probability at least $1 - 3\delta$, we have

$$\sum_{t=1}^T (b - g(s_t, a_t)) = \widetilde{\mathcal{O}}\left((\gamma\zeta + 2) T^{2/3} + \gamma(\zeta + 1)(\sqrt{d^3 T} + d^2 T^{1/3}) + \gamma \mathrm{sp}(v_{\lambda^*}^*)\sqrt{T}\right).$$

G Numerical experiments

We evaluate Algorithm 1 on a tabular CMDP, constructed by modifying the MDP of He et al. (2022) to incorporate constraints and an infinite-horizon setting. The state space is $\mathcal{S} = \{0, \dots, H+1\}$, and the action space is $\mathcal{A} = \{-1, 1\}^{m-1}$. For each state $s \in \{0, \dots, H-1\}$ and action $a \in \mathcal{A}$, the reward and constraint functions are defined as $r(s,a) = 0.4 \cdot \text{score}(a)$ and $g(s,a) = \text{score}(a)$, where $\text{score}(a)$ denotes the fraction of positive components in a. State H yields reward 0 and state $H+1$ yields reward 1; both states incur zero constraint value. For each $s \in \{0, \cdots, H-1\}$, the process transitions to state $s+1$ with probability $0.95 - 0.01 \cdot \mathbf{1}_{m-1}^\top a$ and transitions to state $H+1$ otherwise. The terminal/reset states H and $H+1$ transition back to the initial state, thereby inducing an infinite-horizon process.

In the experiment, we choose $H = 6, m = 2$, $d = |\mathcal{S}||\mathcal{A}| = 12$, and the constraint threshold to $b = 0.7$. We compare Algorithm 1 with a random agent and Algorithm 2 of Ghosh et al. (2023), which is a canonical primal-dual algorithm based on a finite-horizon approximation. For Algorithm 1, we set the hyperparameters according to the choices specified in Theorem 2, while we use the adaptive clipping parameter $\zeta_k = \text{sp}(v_r^*) + \lambda_k \text{sp}(v_g^*)$ and set $\beta_k = (\zeta_k + 1)d$ to improve learning stability. For Algorithm 2 in Ghosh et al. (2023), we follow the default parameter choices, except that we choose $H = T^{1/4}$, omitting the factor $d^{3/4}$ in the denominator, and set $\beta = dH$, omitting the constant and logarithmic terms. We conduct 4 simulations with $K = 1000$ under different random seeds.

As shown in Figure 4, Algorithm 1 achieves sublinear regret and constraint violation, providing empirical support for the main claim of Theorem 2. In contrast, Algorithm 2 of Ghosh et al. (2023) exhibits near-random performance. This behavior can be attributed to the choice of β, which induces a large bonus term and consequently leads to overly optimistic value-function estimates. These results suggest that the clipping mechanism in Algorithm 1, characterized by the clipping parameter ζ and its dependence on $\text{sp}(v_r^*)$ and $\text{sp}(v_g^*)$, provides improved adaptivity in this CMDP. Overall, the results highlight the empirical efficiency of Algorithm 1.

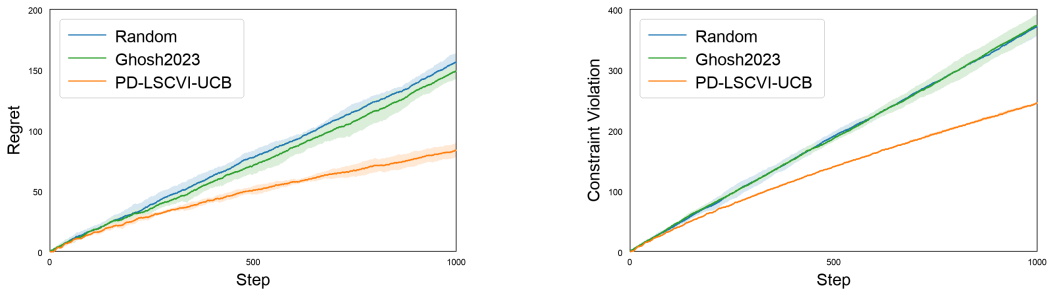


Figure 4: Comparison of random agent, Algorithm 2 in Ghosh et al. (2023), and PD-LSCVI-UCB (Algorithm 1). Each plot displays regret and constraint violation for $T = 1000$ steps. The shaded regions indicate 95% confidence intervals.

H Auxiliary lemmas

Lemma 16 (Lemma D.1 in Jin et al. (2020)). *Let $\mathbf{\Sigma}_t = I + \sum_{i=1}^t \phi_i \phi_i^\top$ where $\phi_i \in \mathbb{R}^d$. Then $\sum_{i=1}^t \phi_i^\top \mathbf{\Sigma}_t^{-1} \phi_i \le d$.*

Lemma 17 (Lemma 12 in Hong et al. (2024)). *Let $V : \mathcal{S} \to [0, B]$ be a bounded function. Then, there exists $\mathbf{w}^* \in \mathbb{R}^d$ such that $PV(s,a) = \phi(s,a)^\top \mathbf{w}^*$ for all $(s,a) \in \mathcal{S} \times \mathcal{A}$ and $\|\mathbf{w}^*\|_2 \le B\sqrt{d}$.*

Lemma 18 (Lemma 13 in Hong et al. (2024)). *Let $\mathbf{w}$ be a ridge regression coefficient obtained by regressing $y \in [0, B]$ on $\mathbf{x} \in \mathbb{R}^d$ using the dataset $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$ so that $\mathbf{w} = \mathbf{\Sigma}^{-1} \sum_{i=1}^n \mathbf{x}_i y_i$, where $\mathbf{\Sigma} = I + \sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i^\top$. Then, $\|\mathbf{w}\|_2 \le B\sqrt{dn}$.*

Lemma 19 (Lemma 14 in Hong et al. (2024)). *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 $\|\phi_t\|_2 \le 1$. Let $\mathbf{\Sigma}_n = I + \sum_{t=1}^n \phi_t \phi_t^\top$. 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 $\mathrm{sp}(V) \leq 2\zeta$, we have

$$\left\| \sum_{t=1}^{n} \phi_t (V(x_t) - \mathbb{E}[V(x_t)|\mathcal{F}_{t-1}]) \right\|^2_{\mathbf{\Sigma}_n^{-1}} \leq 16\zeta^2 \left[\frac{d}{2} \log(n+1) + \log \frac{\mathcal{N}_\epsilon}{\delta} \right] + 8n^2\epsilon^2$$

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

Lemma 20 (Lemma 15 in Hong et al. (2024))**.** *Let $\mathcal{V}$ be a class of functions mapping from $\mathcal{S}$ to $\mathbb{R}$ with the following parametric form:*

$$V(\cdot) = \left(\max_a \mathbf{w}^\top \phi(\cdot, a) + v + \beta \|\phi(\cdot, a)\|_{\mathbf{\Sigma}^{-1}} \right) \wedge M$$

where the parameters $(\mathbf{w}, \beta, v, \mathbf{\Sigma}, M)$ satisfy $\|\mathbf{w}\|_2 \leq L$, $\beta \in [0, B]$, $v \in [0, D]$, $M \geq 0$ and the minimum eigenvalue satisfies $\lambda_{\min}(\mathbf{\Sigma}) \geq 1$. Assume $\|\phi(s,a)\|_2 \leq 1$ for all (s,a), and let $\mathcal{N}_\epsilon$ be the ϵ-covering number of $\mathcal{V}$ with respect to $\mathrm{dist}(V, V') \leq \sup_s |V(s) - V'(s)|$. *Then*

$$\log(\mathcal{N}_\epsilon) \leq d \log(1 + 8L/\epsilon) + \log(1 + 4D/\epsilon) + d^2 \log\left(1 + 8d^{1/2}B^2/\epsilon^2\right).$$

Lemma 21 (Lemma D.2 in Jin et al. (2020))**.** *Let $\{\phi_t\}_{t \geq 0}$ be a bounded sequence in $\mathbb{R}^d$ satisfying $\sup_{t \geq 0} \|\phi_t\| \leq 1$. Let $\mathbf{\Sigma}_0 \in \mathbb{R}^{d \times d}$ be a positive definite matrix. For any $t \geq 0$, we define $\mathbf{\Sigma}_t = \mathbf{\Sigma}_0 + \sum_{j=1}^{t} \phi_j \phi_j^\top$. Then if the smallest eigenvalue of $\mathbf{\Sigma}_0$ satisfies $\lambda_{\min}(\mathbf{\Sigma}_0) \geq 1$, we have*

$$\log \frac{\det(\mathbf{\Sigma}_t)}{\det(\mathbf{\Sigma}_0)} \leq \sum_{j=1}^{t} \phi_j^\top \mathbf{\Sigma}_{j-1}^{-1} \phi_j \leq 2 \log \frac{\det(\mathbf{\Sigma}_t)}{\det(\mathbf{\Sigma}_0)}.$$

Lemma 22. $|\{k \in [K] : 2\det(\mathbf{\Sigma}_k) < \det(\mathbf{\Sigma}_{k+1})\}| \leq \frac{3}{2} d \log(1+T)$.

Proof. For simplicity, let $N = |\{k \in [K] : 2\det(\mathbf{\Sigma}_k) < \det(\mathbf{\Sigma}_{k+1})\}|$. Note that since $\det(\mathbf{\Sigma}_{k+1}) \geq \det(\mathbf{\Sigma}_k)$, we have

$$\det(\mathbf{\Sigma}_{K+1}) = \prod_{k=1}^{K} \frac{\det(\mathbf{\Sigma}_{k+1})}{\det(\mathbf{\Sigma}_k)} \geq 2^N.$$

Moreover,

$$\det(\mathbf{\Sigma}_{K+1}) = \det\left(I + \sum_{k,h} \phi(s_h^k, a_h^k) \phi(s_h^k, a_h^k)^\top \right) \leq (1+T)^d.$$

Finally, we have $N \leq d \log_2(1+T) \leq \frac{3}{2} d \log(1+T)$. □

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We have included the main claims in the abstract and introduction.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We have discussed the limitations in Section 5 and Section B.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: The assumptions are provided in Section 2, and Assumptions 1 and 2. The proof of Theorems 1 and 2 are provided in the appendix.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The main contributions of the paper are theoretical. The numerical experiments are supplementary, and all details necessary for reproduction are provided in Section G.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: The entire code required to reproduce the experiments is included in the supplementary material.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: The main contributions of the paper are theoretical. The numerical experiments are supplementary, and all details necessary for reproduction are provided in Section G.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Figure 4 reports 95% confidence intervals for all experimental results.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [No]

   Justification: Our numerical experiments do not require GPU resources and can be reproduced using standard CPU computation.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We have reviewed the NeurIPS Code of Ethics and confirm that the research presented in this paper complies with these guidelines.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: The paper presents theoretical contributions.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper does not release datasets, pretrained models, or other assets that pose a high risk for misuse.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [N/A]

    Justification: The paper does not use external datasets, pretrained models, or third-party code assets.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: The paper does not release datasets, pretrained models, or other assets that pose a high risk for misuse.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.

- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: No LLMs are used in the core methods.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.