RATE-OPTIMAL ALGORITHM FOR LEARNING ADVERSARIAL LINEAR CMDPS

Anonymous authors
Paper under double-blind review

ABSTRACT

We study episodic adversarial linear constrained Markov decision processes (CMDPs) with unknown transitions, where both the loss and constraint functions may vary adversarially across episodes. The best previous algorithm achieves $\widetilde{\mathcal{O}}(K^{3/4})$ regret and cumulative constraint violation, leaving a gap to the optimal $\widetilde{\mathcal{O}}(\sqrt{K})$ dependence on the number of episodes K. We close this gap by proposing a new primal dual algorithm that achieves $\widetilde{\mathcal{O}}(\sqrt{K})$ regret and cumulative constraint violation without assuming Slater's condition. The main challenge is that learning linear CMDPs requires uniform concentration over a value function class with a controlled covering number, whereas standard techniques in constrained online learning, such as policy mixing, can make this class more complex. Our algorithm combines adaptive Follow the Regularized Leader (FTRL), contracted value estimation, and an exponential Lyapunov function. An adaptive dual regularizer offsets the dependence on the dual weights in the primal regret bound, removing the need for policy mixing. We further show that the normalization in the FTRL update bounds the policy parameters independently of the magnitudes of the dual weights, which explains why the resulting policy class remains compatible with uniform concentration. Under feature access, the computational complexity is independent of the size of the state space.

1 INTRODUCTION

Safe reinforcement learning (RL) studies sequential decision-making under safety constraints through interaction with an unknown environment. Many safety-critical real-world applications have been studied under the safe RL framework, including autonomous driving (Isele et al., 2018), robotics (Achiam et al., 2017), health care (Coronato et al., 2020), and large language models (Dai et al., 2024). A standard framework for modeling safe RL is the constrained Markov decision process (CMDP), where an agent aims to optimize an expected cumulative loss while satisfying expected constraints that capture safety requirements (Altman, 2021).

Recently, algorithms for learning CMDPs have been extensively studied along several important directions. One direction is to incorporate function approximation in order to handle environments with large state spaces. In particular, linear CMDPs assume that the transition kernel, loss functions, and constraint functions admit linear representations with respect to a known feature mapping. Under stochastic loss and constraint functions, where the functions are sampled from fixed distributions, several algorithms have been developed for linear CMDPs with sublinear regret and constraint violation guarantees (Ghosh et al., 2022; 2024; Kitamura et al., 2025; Peng et al., 2026). However, these works focus only on fixed environments, in the sense that the underlying distributions of the functions remain fixed, and do not handle changing environments.

To capture changing environments, another important direction is to consider adversarial environments. In many real-world applications, loss and constraint functions may change arbitrarily over time rather than being sampled from fixed distributions. For example, safety requirements or environmental conditions may vary across episodes due to changes in operating conditions, user preferences, or external factors. To capture such scenarios, online adversarial CMDPs allow loss or constraint functions to be chosen adversarially across episodes. This setting has been extensively investigated for tabular CMDPs (Zhu et al., 2025; Stradi et al., 2024; Roy & Das, 2026). While these

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

works successfully handle adversarial environments, they are limited to the tabular setting, which cannot handle large state spaces.

A natural question is whether these two directions—linear function approximation and adversarially changing environments—can be handled simultaneously. Recently, Yu et al. (2026a;b) initiated the study of CMDPs with adversarial losses and stochastic constraints under linear function approximation. Specifically, while Yu et al. (2026a) consider adversarial linear mixture CMDPs with rate-optimal guarantees of $\widetilde{\mathcal{O}}(\sqrt{K})$, where K denotes the number of episodes, their algorithm becomes intractable when the state space is arbitrarily large. On the other hand, Yu et al. (2026b) propose a model-free algorithm for adversarial linear CMDPs that remains computationally efficient even when the state space is large, but their guarantees remain suboptimal in terms of K, i.e., $\widetilde{\mathcal{O}}(K^{3/4})$. In summary, existing algorithms for adversarial CMDPs under linear function approximation are either computationally inefficient or statistically suboptimal in terms of K.

To close this gap, we propose a model-free and rate-optimal algorithm for learning adversarial linear CMDPs, which is thus computationally efficient even when the state space is large and statistically optimal in terms of K. However, achieving this goal is nontrivial because learning linear CMDPs requires controlling the covering number of the value function class (Jin et al., 2020). In particular, while the covering number argument is an essential analytical tool for establishing concentration in linear MDPs, it also limits the direct use of other useful techniques from constrained online learning. For instance, Yu et al. (2026b) employ policy mixing, which is a common technique in the broad area of constrained online learning (Wei et al., 2020; Qiu et al., 2020; Yu et al., 2026a), but is problematic for linear CMDPs due to the covering number issue. Indeed, this induces a trade-off, which is a main reason why their guarantees remain suboptimal.

Main Contributions. As previously mentioned, learning adversarial linear CMDPs is particularly challenging due to the covering number issue, which does not appear in other constrained online learning problems. Nevertheless, we design a rate-optimal algorithm for learning adversarial linear CMDPs. We summarize our main contributions below and in Table 1.

- We propose an algorithm for learning adversarial linear CMDPs, where both loss and constraint functions are adversarially chosen in each episode under full-information feedback. In particular, our algorithm achieves the rate-optimal guarantees of $\widetilde{\mathcal{O}}(\sqrt{K})$ for both regret and constraint violation. For comparison, this result improves upon the previous state-of-the-art algorithm (Yu et al., 2026b), whose guarantees are limited to $\widetilde{\mathcal{O}}(K^{3/4})$ and which cannot handle adversarial constraints.
- Moreover, our algorithm does not require Slater's condition, which assumes the existence of a strictly feasible solution with some positive margin, called the Slater constant ξ. Specifically, our algorithm does not take ξ as an input, and our guarantees do not depend on ξ; thus, they do not deteriorate even when ξ becomes arbitrarily close to 0. Recently, in the linear CMDP literature, several algorithms have been proposed that do not require this condition, but they are either computationally inefficient (Ghosh et al., 2024) or limited to suboptimal guarantees (Peng et al., 2026). In contrast, our algorithm is computationally efficient and achieves rate-optimal guarantees without Slater's condition in the linear CMDP setting.
- The intuition behind our algorithm is to develop a new drift analysis for adversarial linear CMDPs that avoids policy mixing, thereby avoiding the conflict with the covering number argument. To this end, we incorporate an adaptive regularizer into the dual update together with an exponential Lyapunov function, inspired by Yu et al. (2026d), where such a technique was originally developed for convex optimization. This adaptive regularization leads to a drift upper bound that can be controlled without policy mixing. In contrast, the analysis of Yu et al. (2026b) relies on policy mixing to control the KL divergence terms appearing in the drift analysis.
- While adopting the adaptive regularizer in the dual update allows us to avoid policy mixing, there are additional technical challenges that we must address. The first challenge is the unbounded diameter of the solution space. In particular, a direct extension of Yu et al. (2026d) to the policy space would require a bounded KL-diameter, which does not hold for the probability simplex. To resolve this issue, we adopt adaptive Follow-the-Regularized-Leader (FTRL) for the policy update, whose regret analysis avoids such a diameter dependence. The second challenge is that a naïve bound on the dual variable can grow exponentially in K due to the use of an exponential Lyapunov function. Since such a bound enters the covering number analysis, we clip the

Table 1: Comparison of recent algorithms for adversarial CMDPs. Here, Sto. and Adv. denote stochastic and adversarial, respectively. The $|\mathcal{S}|$-free column indicates whether the computational complexity is independent of $|\mathcal{S}|$. Slater-free indicates that neither the algorithm nor its guarantees require Slater's condition.

Algorithm	MDP	Loss	Constr.	Rate	$\|\mathcal{S}\|$-free	Slater-free
Roy & Das (2026)	Tabular	Adv.	Adv.	$\sqrt{K}$	✗	✓
Yu et al. (2026a)	Linear Mixture	Adv.	Sto.	$\sqrt{K}$	✗	✗
Yu et al. (2026b)	Linear	Adv.	Sto.	$K^{3/4}$	✓	✗
Ours	Linear	Adv.	Adv.	$\sqrt{K}$	✓	✓

dual variable at a polynomial threshold and then show that this clipping does not affect the dual dynamics.

The additional discussion on related work is provided in Appendix B.

2 Problem Setting

Notation. Let $[n] = \{1, \dots, n\}$ for some positive integer n, and let $\Delta(\mathcal{A})$ denote the probability simplex over the set $\mathcal{A}$. Given a positive integer d and a vector $\mathbf{x} \in \mathbb{R}^d$, let $||\mathbf{x}||$ denote the ℓ_2-norm, and let $||\mathbf{x}||_1, ||\mathbf{x}||_\infty$ denote the ℓ_1- and ℓ_∞-norms, respectively. For a positive definite matrix $\mathbf{A} \in \mathbb{R}^{d\times d}$, let $||\mathbf{A}||$ denote the operator norm, let $||\mathbf{x}||_{\mathbf{A}} = \sqrt{\mathbf{x}^\top \mathbf{A}\mathbf{x}}$, and let $\lesssim$ denote the inequality that omits constants and polylogarithmic factors. The additional notation used throughout the paper is summarized in Appendix A.

Adversarial CMDPs. Let $\mathcal{M} = (\mathcal{S}, \mathcal{A}, H, K, s_1, b, \{f^k, g^k\}_{k=1}^K, \{\mathbb{P}_h\}_{h=1}^H)$ denote a CMDP, where $\mathcal{S}$ is a finite state space, $\mathcal{A}$ is a finite action space, H is a horizon length, K is the number of episodes, $s_1 \in \mathcal{S}$ is an initial state in each episode k, and $b \in [0, H]$ is a constraint threshold. While we assume a finite state space for simplicity, it can be extended to the arbitrarily large state case, as the statistical guarantees and computational complexity do not depend on $|\mathcal{S}|$. Additionally, $\{f^k, g^k\}_{k=1}^K$ denotes the sequence of loss and constraint functions, where $f_h^k, g_h^k : \mathcal{S} \times \mathcal{A} \to [0, 1]$ for each $(h, k) \in [H] \times [K]$, and $\mathbb{P}$ denotes a transition kernel, where $\mathbb{P}_h(s'|s, a)$ is the probability of transitioning from state s to s' by taking action a at step $h \in [H]$.

The interaction between the agent and the environment proceeds as follows. For each episode k and step h, after observing s_h^k, the agent takes an action $a_h^k \sim \pi_h^k(\cdot|s_h^k)$, where π_h^k denotes the agent's policy. Then, the next state s_{h+1}^k is sampled from $\mathbb{P}_h(\cdot|s_h^k, a_h^k)$. Once episode k ends, the full-information feedback of f^k, g^k is revealed, while the transition kernel $\{\mathbb{P}_h\}_{h=1}^H$ remains unknown.[1]

Given a CMDP $\mathcal{M}$, we define the value functions as follows. We define the value function as $V_{f^k,h}^\pi(s) = \mathbb{E}_{\mathbb{P},\pi}[\sum_{h'=h}^H f_{h'}^k(s_{h'}, a_{h'})|s_h = s]$ and the Q-function as $Q_{f^k,h}^\pi(s, a) = \mathbb{E}_{\mathbb{P},\pi}[\sum_{h'=h}^H f_{h'}^k(s_{h'}, a_{h'})|s_h = s, a_h = a]$, where the expectations are taken over the trajectory $\{s_h, a_h\}_{h=1}^H$ given $\mathbb{P}, \pi$. Moreover, we consider regret and constraint violation as performance metrics:

$$\text{Regret}(K) = \sum_{k=1}^K (V_{f^k,1}^{\pi^k}(s_1) - V_{f^k,1}^{\pi^\star}(s_1)) \tag{1}$$

$$\text{Violation}(K) = \sum_{k=1}^K (V_{g^k,1}^{\pi^k}(s_1) - b) \tag{2}$$

where $\pi^\star \in \arg\min_{\pi\in\Pi} \sum_{k=1}^K V_{f^k,1}^\pi(s_1)$ s.t. $V_{g^k,1}^\pi(s_1) \le b \ \forall k \in [K]$, under the assumption that the corresponding feasible policy region is non-empty. Here, Π denotes the set of Markovian policies, i.e., $\Pi = \{\{\pi_h\}_{h=1}^H : \pi_h : \mathcal{S} \to \Delta(\mathcal{A}), \ \forall h \in [H]\}$.

Remark 1. For adversarial losses and constraints, we may consider a different definition of $\pi^\star$, for example, $\arg\min_{\pi\in\Pi} \sum_{k=1}^K V_{f^k,1}^\pi(s_1)$ s.t. $\sum_{k=1}^K V_{g^k,1}^\pi(s_1) \le Kb$. However, under this defini-

[1] While we consider adversarial constraints under full-information feedback, our algorithm can be applied to stochastic constraints under bandit feedback with a few modifications to the algorithm, maintaining the main results. This is possible because, if the constraint function satisfies the linear structure, the cost of estimating the constraint function does not change the dominant terms in the regret and violation bounds.

tion of $\pi^\star$, there is no algorithm that makes both $\mathrm{Regret}(K)$ and $\mathrm{Violation}(K)$ sublinear (Mannor et al., 2009). Given this fact, our definition of $\pi^\star$ is standard in the constrained online learning literature (Sinha & Vaze, 2024; Zhu et al., 2025).

Linear CMDPs. Throughout the paper, we consider linear CMDPs. That is, given a dimension d and a known feature mapping $\boldsymbol{\phi} : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$, the loss and constraint functions admit the following linear representation: $f_h^k(s,a) = \boldsymbol{\phi}(s,a)^\top \boldsymbol{\theta}_{f,h}^k$ and $g_h^k(s,a) = \boldsymbol{\phi}(s,a)^\top \boldsymbol{\theta}_{g,h}^k$ for each $(s,a,h,k) \in \mathcal{S} \times \mathcal{A} \times [H] \times [K]$. Similarly, the transition kernel can be represented as $\mathbb{P}_h(s'|s,a) = \boldsymbol{\phi}(s,a)^\top \boldsymbol{\mu}_h(s')$ for some $\boldsymbol{\mu}_h : \mathcal{S} \to \mathbb{R}^d$. We note that $\{\boldsymbol{\mu}_h\}_{h=1}^H$ is unknown during the learning process, while $\{\boldsymbol{\theta}_{f,h}^k, \boldsymbol{\theta}_{g,h}^k\}_{h=1}^H$ are revealed at the end of episode k. For each $(s,a,h,k) \in \mathcal{S} \times \mathcal{A} \times [H] \times [K]$, we assume that $||\boldsymbol{\theta}_{f,h}^k||, ||\boldsymbol{\theta}_{g,h}^k|| \le \sqrt{d}$, $||\boldsymbol{\phi}(s,a)|| \le 1$, and $||\sum_{s' \in \mathcal{S}} |\boldsymbol{\mu}_h(s')|| \le \sqrt{d}$, where $|\boldsymbol{\mu}_h(s')|$ denotes the component-wise absolute value of a vector. Note that we follow the standard regularity assumptions for linear MDPs (Cassel & Rosenberg, 2024).

Remark 2. Unlike the previous work (Yu et al., 2026b), our problem setting does not impose Slater's condition. Slater's condition requires the existence of a policy that satisfies the constraint with a strictly positive margin ξ, referred to as the Slater constant. The guarantees of Yu et al. (2026b) depend inversely on ξ and therefore deteriorate as ξ becomes small. In contrast, our guarantees are independent of ξ and hold without assuming Slater's condition.

3 Difficulty and Our Approach

The fundamental difficulty of learning adversarial linear CMDPs arises from the need to control the covering number. Here, the covering number argument allows us to establish concentration properties for linear CMDPs, constituting an essential analytical tool (Jin et al., 2020). However, this requirement often creates an obstacle to directly applying useful techniques from the constrained online learning literature. This is why the state-of-the-art algorithm for adversarial linear CMDPs remains suboptimal (Yu et al., 2026b), while several optimal algorithms have been proposed in other related settings, such as constrained online convex optimization (Sinha & Vaze, 2024), tabular CMDPs (Qiu et al., 2020), and linear mixture CMDPs (Yu et al., 2026a), which do not require a covering number argument.

To elaborate on this difficulty, we investigate the bottleneck in the previous work toward achieving optimal guarantees. Yu et al. (2026b) proposed a primal-dual algorithm for linear CMDPs with adversarial losses, but their guarantees are limited to $\widetilde{\mathcal{O}}(K^{3/4})$, implying that there remains a gap from the optimal rate $\widetilde{\mathcal{O}}(\sqrt{K})$. Importantly, this gap stems from the policy mixing step in their algorithm, which slightly perturbs the policy before optimization. Indeed, this is a well-known technique in the broad area of constrained online learning, as it is helpful for the stability of the dual process by mitigating KL divergence terms that appear in the Lyapunov drift analysis (Wei et al., 2020; Qiu et al., 2020; Yu et al., 2026a); see also (3). However, in the linear CMDP setting, this causes additional problems, as it leads to a large covering number, resulting in looser concentration bounds. This is because the policy structure resulting from mixing is complex (Yu et al., 2026b). In summary, policy mixing is required for their Lyapunov analysis, whereas it causes additional problems particularly in the linear CMDP setting, leading to the suboptimal guarantees.

Motivated by this investigation, the key intuition behind our approach is to develop a new Lyapunov drift analysis for adversarial linear CMDPs that does not rely on mixing. Specifically, its particular advantage in this setting is that the corresponding policy structure remains simple, leading to a small covering number. To achieve the desired drift analysis, we adopt an adaptive regularizer in the dual update together with an exponential Lyapunov function, inspired by Yu et al. (2026d). For comparison, in the analysis of Yu et al. (2026b), the Lyapunov drift term is bounded by KL divergence terms, which can be expressed in the following simplified form, while the detailed version is provided in Appendix C: given the dual variable Y_k, a Slater constant ξ, and the corresponding policy $\bar{\pi}$,

$$Y_{k+1}^2 - Y_k^2 \lesssim -\xi Y_k + D_{\mathrm{KL}}(\bar{\pi}||\pi_h^k). \tag{3}$$

From (3), to ensure that the drift term is bounded, perturbing π_h^k is essential for bounding the KL divergence term on the right-hand side. Without perturbation, the KL divergence could be unbounded over the policy space. On the other hand, our method yields the following drift bound. Let R_k

denote the adaptive regularizer, Φ denote the exponential Lyapunov function, and $\hat{V}_{g,1}^k(s_1)$ denote the constraint value function estimate, whose precise formulations are given in Section 4. Then we have

$$\Phi(Y_{k+1}) - \Phi(Y_k) \lesssim -R_k\Phi'(Y_k) + \Phi'(Y_k)(\hat{V}_{g,1}^k(s_1) - b). \tag{4}$$

From (4), we observe that the right-hand side does not depend on such a KL divergence term. This suggests that our Lyapunov analysis does not require the mixing step to guarantee the stability of the dual process and is therefore more appropriate for the linear CMDP setting.

Remark 3. Another important distinction between (3) and (4) lies in the role of the Slater constant ξ. In particular, the negative drift term $-\xi Y_k$ in (3) relies on ξ, which plays a conceptual role in ensuring the stability of the dual process. On the other hand, the Lyapunov drift in (4) is not controlled by ξ, implying that stability can be guaranteed even without the presence of ξ. That said, our algorithm and analysis still work without Slater's condition, while those of Yu et al. (2026b) do not.

However, directly applying the approach of Yu et al. (2026d), which was originally proposed for convex optimization, does not lead to rate-optimal guarantees in the linear CMDP setting. In particular, we must address the following two technical challenges. The first challenge is the unbounded diameter. Specifically, they assume that the diameter of the primal space in terms of the ℓ_2-norm is bounded, whereas in our case, the diameter of the policy space in terms of KL divergence is unbounded. The second challenge is that, since we adopt an exponential Lyapunov function Φ, naïve bounds on Y_k can grow exponentially in K. Under our analysis, such a naïve bound enters the covering number as in (5) (see also Lemma 16), it could make the covering number large, leading to poor guarantees. That is, letting C_Y be the uniform naïve upper bound on Y_k,

$$\log \mathcal{N}_\epsilon \lesssim d^2 \log(C_Y) \lesssim d^2 K. \tag{5}$$

We address these technical challenges as follows. To resolve the first challenge, we apply adaptive FTRL (Orabona, 2019) to the MDP setting.[2] Specifically, Adaptive Online Mirror Descent (AdaOMD), which is the most natural extension of the Adaptive Gradient Descent used in Yu et al. (2026d) to the MDP setting, typically yields the following regret analysis: given the diameter D and the gradients ∇_k,[3]

$$\mathrm{Regret}_{\mathrm{AdaOMD}}(K) \lesssim \underbrace{D}_{\substack{\text{unbounded} \\ \text{for MDPs}}} \sqrt{\sum_{k=1}^{K} ||\nabla_k||^2}.$$

Here, $\mathrm{Regret}_{\mathrm{AdaOMD}}(K)$ requires the diameter parameter D, which is unbounded in the MDP setting. On the other hand, we observe that the regret bound of adaptive FTRL can be written as follows:

$$\mathrm{Regret}_{\mathrm{AdaFTRL}}(K) \lesssim \log|\mathcal{A}| \sqrt{\sum_{k=1}^{K} ||\nabla_k||^2}.$$

For comparison, $\mathrm{Regret}_{\mathrm{AdaFTRL}}$ depends on $\log|\mathcal{A}|$ instead of D. We then show that adaptive FTRL can be successfully applied to the linear CMDP setting. A detailed analysis is provided in Appendix D.2.

To resolve the second challenge, we clip the dual variable at a polynomial in K so that the naïve bound on Y_k becomes polynomial in K. Given this, the log covering number in (5) becomes $d^2 \log K$ rather than $d^2 K$. However, this modification leads to an additional issue: the clipping step may affect the analysis of constraint violation, since it may change the dual variable dynamics. We resolve this issue by showing that such clipping does not occur once we choose a sufficiently large clipping parameter (Lemma 8). This allows us to complete the violation analysis successfully.

[2] While in the convex optimization literature, Yu et al. (2026d) considers the Euclidean setting together with gradient descent algorithms, based on our observation, their work may be extended to the probability simplex case by adopting adaptive FTRL.

[3] For simplicity, the provided regret bounds are written for online convex optimization rather than MDPs.

Algorithm 1 Rate-Optimal Algorithm for Adversarial Linear CMDPs

Input: $\delta \in (0,1)$; $\beta_p = 187dH\sqrt{\log(2HK/\delta)}$; $\beta_\phi = 4\beta_p \log K$; $\beta_Q = 2H$; $\beta_R = 2H^2(\log|\mathcal{A}| + 20)\sqrt{dH\log(2K)}$; $\gamma = \min\{1/(36H\sqrt{K}), 1/(6\beta_R)\}$

Initialization: $m \leftarrow 0$; $\Phi(x) = e^{\gamma x} - 1$; $Z_1 \leftarrow 0$; $Y_1 \leftarrow \Phi'(Z_1)$; $\mathbf{\Lambda}_h^1 \leftarrow \mathbf{I}\ \forall h \in [H]$; $\hat{Q}_{f,H+1}^k(\cdot,\cdot), \hat{Q}_{g,H+1}^k(\cdot,\cdot), \hat{V}_{f,H+1}^k(\cdot), \hat{V}_{g,H+1}^k(\cdot) = 0, \forall k \in [K]$

1: **for** $k = 1, \ldots, K$ **do**
2: **if** $k = 1$ or $\exists h' \in [H]$ such that $\det(\mathbf{\Lambda}_{h'}^k) \geq 2\det(\mathbf{\Lambda}_{h'}^{k(m)})$ **then**
3: $m \leftarrow m + 1$ and $k(m) \leftarrow k$
4: $\pi_h^{k(m)}(\cdot|s) \leftarrow \pi_{\text{unif}}(\cdot|s),\ \forall h \in [H]$
5: $\bar{\boldsymbol{\phi}}_h^m(\cdot,\cdot) = \boldsymbol{\phi}(\cdot,\cdot)\sigma(-\beta_\phi||\boldsymbol{\phi}(\cdot,\cdot)||_{(\mathbf{\Lambda}_h^{k(m)})^{-1}} + \log K),\ \forall h \in [H]$
6: **end if**
7: Execute π^k and observe $\{s_h^k, a_h^k, \boldsymbol{\theta}_{f,h}^k, \boldsymbol{\theta}_{g,h}^k\}_{h=1}^H$
8: **for** $h = H, \ldots, 1$ **do**
9: $\mathbf{\Lambda}_h^{k+1} \leftarrow \mathbf{\Lambda}_h^k + \boldsymbol{\phi}(s_h^k, a_h^k)\boldsymbol{\phi}(s_h^k, a_h^k)^\top$
10: **for** $\ell = f, g$ **do**
11: $\hat{\boldsymbol{\mu}}_h^k \hat{V}_{\ell,h+1}^k \leftarrow (\mathbf{\Lambda}_h^k)^{-1} \sum_{\tau \in [k-1]} \boldsymbol{\phi}(s_h^\tau, a_h^\tau)\hat{V}_{\ell,h+1}^k(s_{h+1}^\tau)$
12: $\hat{Q}_{\ell,h}^k(\cdot,\cdot) \leftarrow \bar{\boldsymbol{\phi}}_h^m(\cdot,\cdot)^\top \left[\boldsymbol{\theta}_{\ell,h}^k + \hat{\boldsymbol{\mu}}_h^k \hat{V}_{\ell,h+1}^k\right] - \beta_p||\bar{\boldsymbol{\phi}}_h^m(\cdot,\cdot)||_{(\mathbf{\Lambda}_h^{k(m)})^{-1}}$
13: $\hat{V}_{\ell,h}^k(\cdot) \leftarrow \sum_{a \in \mathcal{A}} \pi_h^k(a|\cdot)\hat{Q}_{\ell,h}^k(\cdot,a)$
14: **end for**
15: Update the policy π_h^{k+1}:
16:

$$\pi_h^{k+1}(\cdot|s) \in \underset{\pi(\cdot|s) \in \Delta(\mathcal{A})}{\arg\min}\ \psi_k(\pi(\cdot|s)) + \sum_{\tau=k(m)}^{k} (\hat{Q}_{f,h}^\tau(s,\cdot) + Y_\tau \hat{Q}_{g,h}^\tau(s,\cdot))^\top \pi(\cdot|s)$$

17: where $\psi_k(\pi(\cdot|s)) = \alpha_k \sum_{a \in \mathcal{A}} \pi(a|s)\log\pi(a|s)$ and $\alpha_k = \frac{\beta_Q}{4}\sqrt{\sum_{\tau=k(m)}^k (1+Y_\tau)^2}$.
18: **end for**
19: $R_k = 108\gamma H^2 + \beta_R \frac{Y_k}{\sqrt{\sum_{\tau=1}^k Y_\tau^2}}$
20: $Z_{k+1} = Z_k + \hat{V}_{g,1}^k(s_1) - b - R_k$
21: $Y_{k+1} = \min\{\Phi'(Z_{k+1}), K^8\}$
22: **end for**

4 Algorithm

In this section, we present our rate-optimal algorithm for adversarial linear CMDPs (Algorithm 1). We first describe the main components of the algorithm and explain their roles, and then state the resulting theoretical guarantees. Finally, we compare our approach with prior work in terms of both guarantees and algorithmic design.

4.1 Description of Algorithm

From Lines 2-13, the algorithm proceeds through epoch-wise policy updates and value estimation. A new epoch begins when $\det(\mathbf{\Lambda}_{h'}^k) \geq 2\det(\mathbf{\Lambda}_{h'}^{k(m)})$ for some h', where $k(m)$ denotes the initial episode of the epoch m. At the beginning of each epoch, the policy is reset to the uniform policy π_{unif}, and the contracted feature $\bar{\boldsymbol{\phi}}_h^m$ is updated using the sigmoid function $\sigma(z) = 1/(1+e^{-z})$ for any $z \in \mathbb{R}$. This contracted feature is then used to define the Q-function and value function estimates, ensuring their boundedness without explicit clipping such as $\max\{\cdot, 0\}$. In each episode k, the agent executes the policy $\{\pi_h^k\}_{h=1}^H$, observes the trajectory $\{s_h^k, a_h^k\}_{h=1}^H$ and the full-information feedback $\{\boldsymbol{\theta}_{f,h}^k, \boldsymbol{\theta}_{g,h}^k\}_{h=1}^H$, and updates $\mathbf{\Lambda}_h^{k+1}$ and the estimates $\hat{Q}_{f,h}^k, \hat{Q}_{g,h}^k, \hat{V}_{f,h}^k, \hat{V}_{g,h}^k$. Importantly, avoiding explicit clipping yields a simpler policy parameterization, which is crucial for the covering number analysis and is a standard technique in adversarial linear MDPs in both the

unconstrained (Cassel & Rosenberg, 2024) and constrained (Yu et al., 2026b) settings. The precise parameterizations are provided in Appendix E.

Adaptive FTRL (Lines 15-17): To update the policy π_h^{k+1}, we use adaptive FTRL to accommodate the policy space. Here, adaptive FTRL minimizes the sum of the linearized losses $\sum_{\tau=k(m)}^{k}(\hat{Q}_{f,h}^{\tau}(s,\cdot) + Y_\tau \hat{Q}_{g,h}^{\tau}(s,\cdot))^\top \pi(\cdot|s)$ together with the adaptive regularizer ψ_k. This differs from the Euclidean setting of Yu et al. (2026d), whose analysis relies on a bounded ℓ_2-diameter of the feasible primal domain. This is a crucial step because the adaptive FTRL update does not require bounding the diameter of the policy space with respect to KL divergence. We then show that it can be combined with the feature contraction technique used in our value function estimation.

Adaptive Regularizer for Dual Update (Lines 19-20): To update the proxy dual variable Z_{k+1}, we incorporate the adaptive regularizer R_k by subtracting it in the update of the proxy dual variable Z_{k+1}. Intuitively, this subtraction helps stabilize the dual process. For comparison, Yu et al. (2026b) also introduces a regularization term in the dual update, but their resulting Lyapunov drift analysis contains KL divergence terms and therefore requires policy mixing to control them. In contrast, with our choice of R_k, the resulting drift upper bound does not involve KL divergence terms. This allows the stability analysis to be carried out without relying on policy mixing, which is crucial in the linear CMDP setting.

Clipping for Dual Update (Line 21): To update the dual variable Y_{k+1}, we evaluate the derivative of the Lyapunov function Φ' at the proxy dual variable Z_{k+1} and then clip the resulting value at K^8. Applying Φ' in the dual update is standard in the literature (Sinha & Vaze, 2024), while the clipping step is introduced to control the covering number in our linear CMDP setting. In our covering number analysis, a polynomial upper bound on Y_{k+1} in terms of K is necessary, whereas $\Phi'(Z_{k+1})$ can be exponentially large in K because Φ is an exponential function. Clipping $\Phi'(Z_{k+1})$ at K^8 therefore prevents this naïve exponential bound from entering the covering number. Importantly, we later show that the clipping is not activated with high probability, which allows the violation analysis to proceed with the original, unclipped dual dynamics.

The computational complexity of Algorithm 1 is $\mathcal{O}(d^3HK + d^2|\mathcal{A}|HK^2)$, which is comparable to that of Yu et al. (2026b) and importantly is independent of $|\mathcal{S}|$. Specifically, computing the determinants requires $\mathcal{O}(d^3HK)$ operations, while the feature contraction and value function estimation require $\mathcal{O}(d^2|\mathcal{A}|HK^2)$ operations. Moreover, since adaptive FTRL admits a closed-form update, we do not need to solve a convex optimization problem (Appendix D.2); its computational cost, as well as that of the dual update, is not dominant.

4.2 Main Result

We present the following theorem, which establishes the rate-optimal guarantees of Algorithm 1.

Theorem 4. *Let $H \geq 2$, and let $\delta \in (0,1)$. With probability at least $1-\delta$, we have*

$$\mathrm{Regret}(K) = \widetilde{\mathcal{O}}\left(\sqrt{d^3H^4K} + \sqrt{dH^5K} + d^3H^4\right),$$
$$\mathrm{Violation}(K) = \widetilde{\mathcal{O}}\left(\sqrt{d^3H^4K} + \sqrt{dH^5K} + d^3H^4\right).$$

where $\widetilde{\mathcal{O}}(\cdot)$ hides polylogarithmic factors in $d, H, |\mathcal{A}|, K, 1/\delta$.

Note that the regret and constraint violation guarantees in Theorem 4 have comparable orders in terms of d, H, K to those of Cassel & Rosenberg (2024), which studied adversarial linear (unconstrained) MDPs. Indeed, these orders are the best-known guarantees without applying variance-related techniques. However, even in the unconstrained setting, no algorithm for adversarial linear MDPs has been proposed that achieves near-optimal guarantees, i.e., $\widetilde{\mathcal{O}}(d\sqrt{H^3K})$ (He et al., 2023). Therefore, we leave achieving optimal dependence on d and H as future work in our setting.

Comparison From the perspectives of both results and algorithms, we compare our algorithm with Yu et al. (2026b), which is the previous state-of-the-art algorithm in a similar setting, i.e., adversarial losses and stochastic constraints. From the perspective of results, the improvements of Algorithm 1 can be summarized as follows: (i) improved guarantees of $\widetilde{\mathcal{O}}(\sqrt{K})$, (ii) no Slater con-

dition is required, and (iii) adversarial constraints can be handled. On the other hand, their algorithm can only attain $\widetilde{\mathcal{O}}(K^{3/4})$, requires Slater's condition, and can only handle stochastic constraints.

From the algorithmic perspective, the differences can be summarized as follows: (i) we adopt adaptive FTRL, while the policy update of Yu et al. (2026b) can be viewed as FTRL with a constant step size, which is easier to handle than in our case; (ii) we adopt an adaptive regularizer for the dual update, which plays a key role in deriving a KL divergence-free drift analysis; and (iii) we adopt upper clipping for the dual variable, while they do not. Hence, in our case, the violation analysis is more challenging, since it requires additional techniques to show that the clipping step does not affect the analysis.

5 Analysis

This section presents the analysis of Algorithm 1 for establishing the rate-optimal guarantees in Theorem 4. The overview of our analysis is as follows. We first introduce the key relation (6), inspired by Yu et al. (2026d), which bounds the sum of regret and $\Phi(Z_{K+1})$. To establish this relation, we provide two key ingredients, (7) and (8), and then explain how they can be obtained in the linear CMDP setting. The details of the proofs are provided in Appendix G.

Before presenting the main analysis, we introduce the good event $\mathcal{E}_g$. Intuitively, this event captures the accuracy and boundedness of the value function estimates and holds with high probability. Although establishing such a good event is a standard step in the online MDP literature, our dual variable clipping is useful here, as it ensures a polynomial upper bound on the dual variable and thereby keeps the resulting covering number under control. The formal definition and the proofs that $\mathcal{E}_g$ holds with high probability are given in Appendix F.

In our analysis, the key relation we need to establish is as follows: the sum of $\mathrm{Regret}(K)$ and the Lyapunov evaluation of Z_{K+1} is bounded by $\widetilde{\mathcal{O}}(\sqrt{K})$, i.e.,

$$\mathrm{Regret}(K) + \Phi(Z_{K+1}) \lesssim \sqrt{K}. \tag{6}$$

This relation is useful because the regret and violation upper bounds naturally follow from it. Specifically, since we take $\Phi(x) = e^{\gamma x} - 1 \geq -1$ for any x, it directly gives us that $\mathrm{Regret}(K) = \widetilde{\mathcal{O}}(\sqrt{K})$. To obtain an upper bound on the violation, we observe that

$$\begin{aligned}\mathrm{Violation}(K) &= \sum_{k=1}^{K}(V_{g^k,1}^{\pi^k}(s_1) - \hat{V}_{g,1}^{k}(s_1)) + \sum_{k=1}^{K}(\hat{V}_{g,1}^{k}(s_1) - b) \\ &= \sum_{k=1}^{K}(V_{g^k,1}^{\pi^k}(s_1) - \hat{V}_{g,1}^{k}(s_1)) + Z_{K+1} + \sum_{k=1}^{K} R_k.\end{aligned}$$

The first sum, which consists of the bias terms, can be bounded using a standard argument. On the other hand, an upper bound on Z_{K+1} can be obtained from (6), while an upper bound on the sum of R_k can be obtained from the following lemma.

Lemma 5. *Assume that the good event $\mathcal{E}_g$ holds. Let $H \geq 2$, and let $K \geq cdH(\log|\mathcal{A}| + 20)\log(2H/\delta)$ for some absolute constant c. We have $\sum_{k=1}^{K} R_k = \widetilde{\mathcal{O}}(\sqrt{K})$.*

So far, we have shown why establishing (6) is crucial in our analysis. Next, to explain how to obtain this relation, we introduce two ingredients. The first is a bound on the sum of regret and the optimistic constraint violation weighted by the dual variable as follows:

$$\sum_{k=1}^{K}\left(V_{f^k,1}^{\pi^k}(s_1) - V_{f^k,1}^{\pi^\star}(s_1)\right) + \sum_{k=1}^{K} Y_k\left(\hat{V}_{g,1}^{k}(s_1) - b\right) \lesssim \sqrt{K} + \sqrt{\sum_{k=1}^{K} Y_k^2}. \tag{7}$$

The second is an upper bound on the Lyapunov drift as follows:

$$\Phi(Z_{k+1}) - \Phi(Z_k) \lesssim Y_k\left(\hat{V}_{g,1}^{k}(s_1) - b - R_k\right). \tag{8}$$

Indeed, (7) and (8) can be combined to derive (6) as follows. Summing (8) over k gives $\Phi(Z_{K+1}) \lesssim \sum_{k=1}^{K} Y_k(\hat{V}_{g,1}^k(s_1) - b) - \sum_{k=1}^{K} Y_k R_k$. Combining this with (7) then yields (6). This is because the choice of R_k allows us to show that $\sum_{k=1}^{K} Y_k R_k \approx \sqrt{\sum_{k=1}^{K} Y_k^2}$, and then the two corresponding terms then cancel each other out.

Now, we explain how to derive (7) and (8) in the linear CMDP setting. To show (7), we prove the adaptive FTRL lemma for the policy space as below. Importantly, we observe that the regret bound for adaptive FTRL does not require a uniform bound on the feasible set of the primal solution. This observation is essential, since the corresponding term in our case is the diameter of the policy space in terms of KL divergence, which is unbounded.

Lemma 6 (Informal[4])**.** *Assume that the good event $\mathcal{E}_g$ holds. Then we have for all $k \in [K]$,*

$$\sum_{k=1}^{K}\sum_{h=1}^{H}\left(\hat{Q}_{f,h}^k(s_h^k,\cdot) + Y_k\hat{Q}_{g,h}^k(s_h^k,\cdot)\right)^\top(\pi_h^k(\cdot|s_h^k) - \pi_h^\star(\cdot|s_h^k)) \lesssim \sqrt{K} + \sqrt{\sum_{k=1}^{K} Y_k^2}.$$

For comparison, a naïve extension of the previous work requires additional techniques to bound the KL divergence, e.g., policy mixing, which is not a promising method in the linear CMDP setting, due to the covering number issue. Roughly, combining this lemma with the standard value difference lemma (Lemma 10) then yields (7).

To show (8), we may start with the following drift bound, which follows from basic algebraic properties of the exponential function together with the choice of R_k.

Lemma 7. *Assume that the good event $\mathcal{E}_g$ holds. Then,*

$$\Phi(Z_{k+1}) - \Phi(Z_k) \le \Phi'(Z_k)\left(\hat{V}_{g,1}^k(s_1) - b - R_k/2\right).$$

Although this lemma shares a similar formulation with (8), there is an important distinction due to the discrepancy between Y_k and $\Phi'(Z_k)$. In particular, since Y_k is obtained by clipping $\Phi'(Z_k)$ at K[8], Lemma 7 does not directly imply (8). Recall that clipping is introduced to bound the covering number in the linear CMDP setting, whereas it does not appear in other constrained online learning settings. To address this issue, we introduce the following lemma, which shows that such clipping does not occur. Consequently, by combining Lemmas 7 and 8, we deduce (8).

Lemma 8. *Assume that the good event $\mathcal{E}_g$ holds. Let $H \ge 2$, and let $K \ge cdH(\log|\mathcal{A}| + 20)\log(2H/\delta)$ for some absolute constant c. Then, for all $k \in [K+1]$, we have*

$$\Phi'(Z_k) = Y_k.$$

6 Conclusion

This paper studies adversarial linear CMDPs, where both loss and constraint functions are adversarially chosen under full-information feedback. We propose a computationally efficient primal-dual algorithm that achieves rate-optimal $\widetilde{\mathcal{O}}(\sqrt{K})$ regret and constraint violation guarantees without requiring Slater's condition. The key idea is to develop a new drift analysis that avoids policy mixing, thereby resolving the conflict between stabilizing the dual process and controlling the covering number in linear CMDPs. To make this approach applicable to the linear CMDP setting, we further employ adaptive FTRL for the policy update and a clipping mechanism for the dual variable. One limitation of this work is that we consider cumulative constraint violation rather than hard constraint violation, which does not allow cancellation across episodes. Developing computationally efficient algorithms for adversarial linear CMDPs with hard constraint violation guarantees is an important direction for future work.

References

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

[4]For simplicity, we focus on the formulation while omitting expectations with respect to the trajectory on the left-hand side. For the precise statement, we refer to Lemma 21.

Joshua Achiam, David Held, Aviv Tamar, and Pieter Abbeel. Constrained policy optimization. In *International conference on machine learning*, pp. 22–31. PMLR, 2017.

Eitan Altman. *Constrained Markov decision processes*. Routledge, 2021.

Sanae Amani, Christos Thrampoulidis, and Lin Yang. Safe reinforcement learning with linear function approximation. In *International Conference on Machine Learning*, pp. 243–253. PMLR, 2021.

Archana Bura, Aria HasanzadeZonuzy, Dileep Kalathil, Srinivas Shakkottai, and Jean-Francois Chamberland. Dope: Doubly optimistic and pessimistic exploration for safe reinforcement learning. *Advances in neural information processing systems*, 35:1047–1059, 2022.

Asaf Cassel and Aviv Rosenberg. Warm-up free policy optimization: Improved regret in linear markov decision processes. *Advances in Neural Information Processing Systems*, 37:3275–3303, 2024.

Liyu Chen, Rahul Jain, and Haipeng Luo. Learning infinite-horizon average-reward markov decision process with constraints. In *International Conference on Machine Learning*, pp. 3246–3270. PMLR, 2022.

Antonio Coronato, Muddasar Naeem, Giuseppe De Pietro, and Giovanni Paragliola. Reinforcement learning for intelligent healthcare applications: A survey. *Artificial intelligence in medicine*, 109: 101964, 2020.

Juntao Dai, Xuehai Pan, Ruiyang Sun, Jiaming Ji, Xinbo Xu, Mickel Liu, Yizhou Wang, and Yaodong Yang. Safe rlhf: Safe reinforcement learning from human feedback. In *International Conference on Learning Representations*, volume 2024, pp. 50750–50777, 2024.

Dongsheng Ding, Xiaohan Wei, Zhuoran Yang, Zhaoran Wang, and Mihailo Jovanovic. Provably efficient safe exploration via primal-dual policy optimization. In *International conference on artificial intelligence and statistics*, pp. 3304–3312. PMLR, 2021.

Yuhao Ding and Javad Lavaei. Provably efficient primal-dual reinforcement learning for cmdps with non-stationary objectives and constraints. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 7396–7404, 2023.

Yonathan Efroni, Shie Mannor, and Matteo Pirotta. Exploration-exploitation in constrained mdps. *arXiv preprint arXiv:2003.02189*, 2020.

Arnob Ghosh, Xingyu Zhou, and Ness Shroff. Provably efficient model-free constrained rl with linear function approximation. *Advances in Neural Information Processing Systems*, 35:13303–13315, 2022.

Arnob Ghosh, Xingyu Zhou, and Ness Shroff. Achieving sub-linear regret in infinite horizon average reward constrained MDP with linear function approximation. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=zZhX4eYNeeh.

Arnob Ghosh, Xingyu Zhou, and Ness Shroff. Towards achieving sub-linear regret and hard constraint violation in model-free rl. In *International Conference on Artificial Intelligence and Statistics*, pp. 1054–1062. PMLR, 2024.

Jiafan He, Heyang Zhao, Dongruo Zhou, and Quanquan Gu. Nearly minimax optimal reinforcement learning for linear markov decision processes. In *International Conference on Machine Learning*, pp. 12790–12822. PMLR, 2023.

David Isele, Alireza Nakhaei, and Kikuo Fujimura. Safe reinforcement learning on autonomous vehicles. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 1–6. IEEE, 2018.

Chi Jin, Zhuoran Yang, Zhaoran Wang, and Michael I Jordan. Provably efficient reinforcement learning with linear function approximation. In *Conference on learning theory*, pp. 2137–2143. PMLR, 2020.

Toshinori Kitamura, Arnob Ghosh, Tadashi Kozuno, Wataru Kumagai, Kazumi Kasaura, Kenta Hoshino, Yohei Hosoe, and Yutaka Matsuo. Provably efficient rl under episode-wise safety in constrained mdps with linear function approximation. *arXiv preprint arXiv:2502.10138*, 2025.

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

Tao Liu, Ruida Zhou, Dileep Kalathil, Panganamala Kumar, and Chao Tian. Learning policies with zero or bounded constraint violation for constrained mdps. *Advances in Neural Information Processing Systems*, 34:17183–17193, 2021.

Shie Mannor, John N Tsitsiklis, et al. Online learning with sample path constraints. *Journal of Machine Learning Research*, 10(3), 2009.

Adrian Müller, Pragnya Alatur, Volkan Cevher, Giorgia Ramponi, and Niao He. Truly no-regret learning in constrained mdps. In *International Conference on Machine Learning*, pp. 36605–36653. PMLR, 2024.

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

Xiyue Peng, Lingkai Zu, Ziyu Shao, and Xin Liu. Towards achieving optimal strong regret and constraint violation via computationally efficient model-free RL. In *Forty-third International Conference on Machine Learning*, 2026.

Shuang Qiu, Xiaohan Wei, Zhuoran Yang, Jieping Ye, and Zhaoran Wang. Upper confidence primal-dual reinforcement learning for cmdp with adversarial loss. *Advances in Neural Information Processing Systems*, 33:15277–15287, 2020.

Amirhossein Roknilamouki, Arnob Ghosh, Ming Shi, Fatemeh Nourzad, Eylem Ekici, and Ness Shroff. Provably efficient RL for linear MDPs under instantaneous safety constraints in non-convex feature spaces. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=sElAqKsJrQ.

Aviv Rosenberg, Alon Cohen, Yishay Mansour, and Haim Kaplan. Near-optimal regret bounds for stochastic shortest path. In *International Conference on Machine Learning*, pp. 8210–8219. PMLR, 2020.

Srinjoy Roy and Swagatam Das. Optimal regret and hard violation for constrained markov decision processes with adversarial losses and constraints. *Transactions on Machine Learning Research*, 2026. ISSN 2835-8856.

Lior Shani, Yonathan Efroni, Aviv Rosenberg, and Shie Mannor. Optimistic policy optimization with bandit feedback. In *International Conference on Machine Learning*, pp. 8604–8613. PMLR, 2020.

Ming Shi, Yingbin Liang, and Ness Shroff. A near-optimal algorithm for safe reinforcement learning under instantaneous hard constraints. In *International Conference on Machine Learning*, pp. 31243–31268. PMLR, 2023.

Abhishek Sinha and Rahul Vaze. Optimal algorithms for online convex optimization with adversarial constraints. *Advances in Neural Information Processing Systems*, 37:41274–41302, 2024.

Francesco Emanuele Stradi, Jacopo Germano, Gianmarco Genalti, Matteo Castiglioni, Alberto Marchesi, and Nicola Gatti. Online learning in cmdps: Handling stochastic and adversarial constraints. In *Forty-first International Conference on Machine Learning*, 2024.

Francesco Emanuele Stradi, Matteo Castiglioni, Alberto Marchesi, and Nicola Gatti. Learning adversarial MDPs with stochastic hard constraints. In *Forty-second International Conference on Machine Learning*, 2025a. URL https://openreview.net/forum?id=s0AwKb1dAW.

Francesco Emanuele Stradi, Matteo Castiglioni, Alberto Marchesi, and Nicola Gatti. Optimal strong regret and violation in constrained MDPs via policy optimization. In *The Thirteenth International Conference on Learning Representations*, 2025b. URL https://openreview.net/forum?id=8eNLKk5by4.

Francesco Emanuele Stradi, Anna Lunghi, Matteo Castiglioni, Alberto Marchesi, and Nicola Gatti. Policy optimization for cmdps with bandit feedback: Learning stochastic and adversarial constraints. In *Forty-second International Conference on Machine Learning*, 2025c.

Tian Tian, Lin Yang, and Csaba Szepesvári. Confident natural policy gradient for local planning in q_π-realizable constrained mdps. *Advances in Neural Information Processing Systems*, 37:76139–76176, 2024.

Honghao Wei, Xin Liu, and Lei Ying. 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, pp. 3868–3876, 2022a.

Honghao Wei, Xin Liu, and Lei Ying. Triple-q: A model-free algorithm for constrained reinforcement learning with sublinear regret and zero constraint violation. In *International Conference on Artificial Intelligence and Statistics*, pp. 3274–3307. PMLR, 2022b.

Honghao Wei, Arnob Ghosh, Ness Shroff, Lei Ying, and Xingyu Zhou. Provably efficient model-free algorithms for non-stationary cmdps. In *International Conference on Artificial Intelligence and Statistics*, pp. 6527–6570. PMLR, 2023.

Honghao Wei, Xin Liu, and Lei Ying. Safe reinforcement learning with instantaneous constraints: the role of aggressive exploration. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 21708–21716, 2024.

Xiaohan Wei, Hao Yu, and Michael J Neely. Online primal-dual mirror descent under stochastic constraints. *Proceedings of the ACM on Measurement and Analysis of Computing Systems*, 4(2): 1–36, 2020.

Kihyun Yu, Duksang Lee, William Overman, and Dabeen Lee. Improved regret bound for safe reinforcement learning via tighter cost pessimism and reward optimism. *Reinforcement Learning Journal*, 2025.

Kihyun Yu, Seoungbin Bae, and Dabeen Lee. Near-optimal primal-dual algorithm for learning linear mixture cmdps with adversarial rewards. *arXiv preprint arXiv:2603.27884*, 2026a.

Kihyun Yu, Seoungbin Bae, and Dabeen Lee. Primal-dual policy optimization for linear CMDPs with adversarial losses. In *The Fourteenth International Conference on Learning Representations*, 2026b.

Kihyun Yu, Beomhan Baek, and Dabeen Lee. Learning weakly communicating average-reward cmdps: Strong duality and improved regret. *arXiv preprint arXiv:2605.11586*, 2026c.

Kihyun Yu, Junehee Lee, and Dabeen Lee. Constrained online convex optimization without slater's condition. *arXiv preprint arXiv:2606.31480*, 2026d.

Jiahui Zhu, Kihyun Yu, Dabeen Lee, Xin Liu, and Honghao Wei. An optimistic algorithm for online CMDPS with anytime adversarial constraints. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=fFgiXamW8E.

CONTENTS

1 Introduction **1**

2 Problem Setting **3**

3 Difficulty and Our Approach **4**

4 Algorithm **6**

4.1 Description of Algorithm 6

4.2 Main Result 7

5 Analysis **8**

6 Conclusion **9**

A Notation **14**

B Related Work **14**

C Additional Discussions **15**

D Additional Preliminaries **16**

D.1 Contracted MDP 16

D.2 Adaptive FTRL 17

D.2.1 Proof of Lemma 11 18

D.2.2 Proof Overview of Eq. (15) 21

E Parameterization and Covering Number **22**

F Good Event **23**

G Deferred Proofs for Section 5 and Theorem 4 **27**

G.1 Proof of Lemma 5 27

G.2 Proofs of Lemma 6 and Eq. (7) 28

G.3 Proofs of Lemmas 7 and 8 33

G.4 Proof of Theorem 4 36

H Numerical Experiments **38**

I Auxiliary Lemmas **38**

A NOTATION

We summarize the notation used throughout the paper.

Table 2: Summary of Notation

$\boldsymbol{\mu}_h$	the parameters for transition kernel
$\boldsymbol{\theta}^k_{f,h}, \boldsymbol{\theta}^k_{g,h}$	the parameters for loss and constraint functions
$\boldsymbol{\phi}$	the known feature map
$\bar{\boldsymbol{\phi}}^m_h$	the contracted feature
$V^{\pi}_{f^k,h}, \hat{V}^k_{f,h}$	the value function and its estimate
$Q^{\pi}_{f^k,h}, \hat{Q}^k_{f,h}$	the Q-function and its estimate
$\bar{\mathbb{P}}^m_h$	the transition kernel of a contracted CMDP in epoch $m \in E$
$\bar{f}^k_h, \bar{g}^k_h$	the loss and constraint functions of a contracted CMDP
E	the set of all epochs
$k(m)$	the first episode of epoch $m \in E$
$K(m)$	the set of all episodes of epoch $m \in E$, i.e., $K(m) = \{k(m), \dots, k(m+1) - 1\}$
$E(k)$	the set of epochs up to and including episode $k \in [K]$, i.e., $E(k) = \{m \in E : k(m) \le k\}$
γ	the parameter for Φ
Φ	the Lyapunov function $\Phi(x) = e^{\gamma x} - 1$
α_k	the adaptive coefficient for ψ_k
ψ_k	the negative entropy regularizer for FTRL with the adaptive coefficient α_k
R_k	the adaptive regularizer for the dual update
Z_k	the proxy dual variable
Y_k	the dual variable
c_{bias}	the absolute constant given by Lemma 23
$\mathcal{F}_{k,h}$	the σ-algebra generated by information up to k, h, i.e., $\mathcal{F}_{k,h} = \begin{cases} \sigma(\{s^1_1, a^1_1, \dots, s^k_h, a^k_h\} \cup \{\boldsymbol{\theta}^1_{f,1}, \boldsymbol{\theta}^1_{g,1}, \dots, \boldsymbol{\theta}^{k-1}_{f,H}, \boldsymbol{\theta}^{k-1}_{g,H}\}), & h < H, \\ \sigma(\{s^1_1, a^1_1, \dots, s^k_H, a^k_H\} \cup \{\boldsymbol{\theta}^1_{f,1}, \boldsymbol{\theta}^1_{g,1}, \dots, \boldsymbol{\theta}^k_{f,H}, \boldsymbol{\theta}^k_{g,H}\}), & h = H. \end{cases}$

B RELATED WORK

Tabular CMDPs Extensive work has studied learning in tabular CMDPs, in which the state and action spaces are finite. Efroni et al. (2020) study exploration–exploitation in unknown tabular CMDPs using optimistic planning and primal–dual methods. With a known strictly safe baseline policy, Liu et al. (2021); Bura et al. (2022); Yu et al. (2025) obtain sublinear regret while ensuring that every deployed policy satisfies its expected cumulative-cost constraint with high probability. Müller et al. (2024); Stradi et al. (2025b) study strong regret and strong constraint violation, which disallow cancellation across episodes. Model-free learning is studied by Wei et al. (2022b) in finite-horizon CMDPs, while Wei et al. (2022a); Chen et al. (2022) consider the infinite-horizon average-reward setting. For adversarial objectives, Qiu et al. (2020) study adversarial losses and stochastic constraints. Stradi et al. (2024; 2025c) allow both objectives and constraints to be adversarial under full-information and bandit feedback, respectively, with adversarial-constraint guarantees relative to a fraction of the optimal reward. Stradi et al. (2025a) study adversarial losses and stochastic constraints under bandit feedback, obtaining cancellation-free violation guarantees. More recently, Zhu et al. (2025); Roy & Das (2026) establish sublinear regret and strong constraint violation for adversarial constraints without Slater's condition. However, these works are limited to the tabular setting, and their statistical guarantees typically depend on $|\mathcal{S}|$, making them inappropriate for problems with large state spaces.

Stochastic CMDPs with Linear Function Approximation Several works have studied CMDPs with linear function approximation under stochastic settings. Ding et al. (2021) develop primal–dual policy optimization for linear mixture CMDPs, while Ghosh et al. (2022) propose a model-free primal–dual algorithm for linear CMDPs. For the latter model, Ghosh et al. (2024) achieve $\widetilde{O}(\sqrt{K})$ regret and hard constraint violation without Slater's condition. Peng et al. (2026) provide a

computationally efficient algorithm achieving $\widetilde{O}(\sqrt{K})$ strong regret and strong constraint violation under Slater's condition, and $\widetilde{O}(K^{3/4})$ bounds without it. Under a known strictly safe baseline policy, Kitamura et al. (2025) achieve sublinear regret and zero episode-wise constraint violation. Instantaneous safety, which constrains individual decisions rather than the expected total cost of a policy, is investigated for linear MDPs by Amani et al. (2021); Roknilamouki et al. (2025) and for linear mixture MDPs by Shi et al. (2023). These works pursue zero instantaneous violation under suitable safety and structural assumptions. In contrast, Wei et al. (2024) allow sublinear cumulative hard violation without requiring a known safe action at every state. Beyond episodic problems, Ghosh et al. (2023); Yu et al. (2026c) develop algorithms for average-reward linear CMDPs. Under a different interaction model, Tian et al. (2024) study local planning in q_π-realizable CMDPs with local simulator access, providing sample-complexity rather than online-regret guarantees.

Adversarial and Non-Stationary CMDPs with Linear Function Approximation In contrast to stochastic settings, several works consider CMDPs whose components can vary over time. In particular, a related line of work studies non-stationary CMDPs under linear mixture and linear MDP assumptions (Ding & Lavaei, 2023; Wei et al., 2023). These works allow changes in transitions, objectives, and constraints, and provide dynamic-regret and constraint-violation guarantees that depend on variation budgets. This differs from unrestricted adversarial objectives under fixed dynamics, where performance is compared against a fixed feasible policy. More closely related to our setting, Yu et al. (2026a) study linear mixture CMDPs with adversarial rewards under full-information feedback and a fixed, known constraint function, achieving $\widetilde{O}(\sqrt{K})$ regret and constraint violation through regularized dual updates and variance-aware estimation. Yu et al. (2026b) study linear CMDPs with unknown transitions, adversarial losses under full-information feedback, and stochastic constraints under bandit feedback. Their primal–dual policy-optimization algorithm combines periodic policy mixing with a regularized dual update, achieving $\widetilde{O}(K^{3/4})$ regret and constraint violation. We emphasize that the algorithm of Yu et al. (2026a) can become computationally intractable when the state space is large, as its computational complexity depends on $|\mathcal{S}|$, while Yu et al. (2026b) is statistically suboptimal in terms of K. In contrast, our algorithm is model-free, its computational complexity is independent of $|\mathcal{S}|$, and it achieves optimal guarantees in terms of K, i.e., $\widetilde{\mathcal{O}}(\sqrt{K})$.

C Additional Discussions

Why is policy mixing inevitable in Yu et al. (2026b)? Previously, we described policy mixing as a bottleneck in Yu et al. (2026b) that leads to suboptimal guarantees. Here, we investigate why policy mixing is inevitable under their analysis. The main reason is that the uniform upper bound on Y_k across k relies on bounding the KL divergence between policies, which can only be guaranteed when mixing is applied.

To elaborate, we first see how policy mixing works. Specifically, the policy mixing used in their work can be expressed as follows: for some small constant $\theta \in (0, 1)$,

$$\pi_h^k(\cdot|s) \leftarrow (1-\theta)\pi_h^k(\cdot|s) + \theta\pi_{\text{unif}} \quad \text{and then} \quad \pi_h^{k+1}(\cdot|s) \propto \pi_h^k(\cdot|s)\exp(-\alpha Q_h^k(s,\cdot)).$$

This step is useful for ensuring that the KL divergence with respect to π_h^k is small, which is essential for stabilizing the dual process. Indeed, the following simplified drift analysis supports this:

$$Y_{k+1}^2 - Y_k^2 \lesssim -\eta\gamma Y_k + \frac{\eta}{\alpha}\mathbb{E}\left[\sum_{h=1}^{H}\left(D_{\text{KL}}(\bar{\pi}||\pi_h^k) - D_{\text{KL}}(\bar{\pi}||\pi_h^{k+1})\right)\right] + C, \tag{9}$$

where η denotes the step size for the dual update, γ and $\bar{\pi}$ denote the Slater constant and Slater policy, respectively, $D_{\text{KL}}(\cdot||\cdot)$ denotes the KL divergence, and C absorbs additional positive terms.

We emphasize that the mixing step provides a uniform bound on the KL divergence terms in (9), as $D_{\text{KL}}(\bar{\pi}||\pi_h^k) \le \log(|\mathcal{A}|/\theta)$, leading to a uniform upper bound on $Y_{k+1}^2 - Y_k^2$. This is a key step in their analysis for obtaining sublinear $\text{Violation}(K)$. On the other hand, without the policy mixing step, the KL divergence term in (9) can become arbitrarily large, making the dual process unstable and leading to a linear $\text{Violation}(K)$ bound.

Some readers may wonder whether we can utilize the telescoping structure between the two KL divergence terms in (9). In particular, if we sum (9) over k, then the KL divergence terms take the form $D_{\mathrm{KL}}(\bar{\pi}||\pi_h^1) - D_{\mathrm{KL}}(\bar{\pi}||\pi_h^{K+1})$, which is small since we set π_h^1 to the uniform policy and KL divergence is nonnegative. Then, we may obtain a compact bound on Y_{K+1}. While this approach sounds reasonable, as $\mathrm{Violation}(K)$ highly depends on Y_{K+1}, a uniform bound on Y_k over k is still required, e.g., in Lemma 5 of Yu et al. (2026b), which bounds an online mirror descent (OMD) regret term. Specifically, if we use the naïve upper bound $Y_k = \widetilde{\mathcal{O}}(\sqrt{K})$, then the OMD regret term becomes $\widetilde{\mathcal{O}}(K)$, implying that such a modified approach still yields suboptimal guarantees. For these reasons, the policy mixing step is unavoidable within their analysis.

Limitations and Future Work We discuss a limitation of our work and explain why addressing it is challenging. In particular, Algorithm 1 can only handle cumulative constraint violation, where violations can cancel each other out across episodes. On the other hand, several algorithms for linear CMDPs can handle hard constraint violation, which does not allow such cancellation (Kitamura et al., 2025; Ghosh et al., 2024; Peng et al., 2026). In our setting, the hard constraint violation can be formulated as follows.

$$\mathrm{Violation}_{\mathrm{Hard}}(K) = \sum_{k=1}^{K} \left[V_{g^k,1}^{\pi^k}(s_1) - b \right]_+ .$$

Next, we explain why achieving sublinear guarantees for hard constraint violation is difficult in our setting. To elaborate, we first introduce the following modified strategy to obtain hard constraint violation guarantees. Let $[\cdot]_+$ denote $\max\{\cdot, 0\}$. In particular, inspired by Sinha & Vaze (2024), the proxy dual variable may be updated as $Z_{k+1} \leftarrow Z_k + [\hat{V}_{g,1}^k(s_1) - b]_+ - R_k$, and then the following modified equation of (7) is desired:

$$\sum_{k=1}^{K} \left(V_{f^k,1}^{\pi^k}(s_1) - V_{f^k,1}^{\pi^\star}(s_1) \right) + \sum_{k=1}^{K} Y_k \left[\hat{V}_{g,1}^k(s_1) - b \right]_+ \lesssim \sqrt{K} + \sqrt{\sum_{k=1}^{K} Y_k^2}. \tag{10}$$

Note that the only difference from (7) is that (10) takes $[\cdot]_+$ in the second term on the left-hand side. By combining Lemma 7 with a few modifications, (10) naturally yields a hard constraint violation guarantee, as it eventually provides an upper bound on $\sum_{k=1}^K [\hat{V}_{g,1}^k(s_1) - b]_+$. Moreover, we can bound $\sum_{k=1}^K [V_{g^k,1}^{\pi^k}(s_1) - \hat{V}_{g,1}^k(s_1)]_+$ by using Lemma 23 and an optimism property, e.g., $\hat{V}_{g,h}^k(s) \le V_{g^k,h}^{\pi^k}(s)$. Combining these, we may obtain a hard constraint violation guarantee.

However, the main challenge arises in deriving (10). For comparison, to obtain (7), we leverage the value difference lemma. On the other hand, when $[\cdot]_+$ is imposed, such a strategy is no longer applicable, since $[\cdot]_+$ destroys the linearity of the left-hand side. Therefore, it is nontrivial to obtain (10), which is a modified version of (7) designed to ensure hard constraint violation. We leave this direction as future work.

D Additional Preliminaries

D.1 Contracted MDP

Let $\bar{\mathcal{M}} = (\mathcal{S}, \mathcal{A}, H, \{\bar{\ell}_h\}_{h=1}^H, s_1, \{\bar{\mathbb{P}}_h\}_{h=1}^H, \rho)$ denote a contracted MDP. Basically, while it shares a similar definition with an MDP introduced in Section 2, the key difference is ρ. Specifically, $\rho : \mathcal{S} \times \mathcal{A} \times [H] \to [0,1]$ controls how much an MDP is contracted due to the following definition:

$$\bar{\ell}_h(s,a) = \rho(s,a,h)\ell_h(s,a), \quad \bar{\mathbb{P}}_h(s' \mid s,a) = \rho(s,a,h)\mathbb{P}_h(s' \mid s,a)$$

where $\bar{\ell}_h(s,a)$ and $\bar{\mathbb{P}}_h$ denote the contracted loss and transition kernel. Note that since we may have $\sum_{s'} \bar{\mathbb{P}}_h(s'|s,a) \le 1$, we call it a sub-probability measure. Nevertheless, under these definitions, we have the following useful lemmas.

Lemma 9 (Lemma 2 of Cassel & Rosenberg (2024)). *For any $\rho : \mathcal{S} \times \mathcal{A} \times [H] \to [0,1]$, $\pi \in \Pi$, $h \in [H]$, $s \in \mathcal{S}$, and $\ell : \mathcal{S} \times \mathcal{A} \times [H] \to [0,1]$, we have $\bar{V}_{\ell,h}^{\pi}(s;\rho) \le V_{\ell,h}^{\pi}(s)$.*

Lemma 10 (Lemma 1 of Shani et al. (2020) and Lemma 14 of Cassel & Rosenberg (2024)). *Let $\pi, \hat{\pi}$ be two policies, and let $\mathcal{M} = (\mathcal{S}, \mathcal{A}, H, s_1, \{\ell_h\}_{h=1}^H, \{\mathbb{P}_h\}_{h=1}^H)$ be a (possibly sub) MDP. For all $h \in [H]$, let $\hat{Q}_{\ell,h} : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ be an arbitrary function, and let $\hat{V}_{\ell,h}(s) = \hat{Q}_{\ell,h}(s,\cdot)^\top \hat{\pi}_h(\cdot \mid s)$ for all $s \in \mathcal{S}$. Then,*

$$V_{\ell,1}^{\pi}(s_1) - \hat{V}_{\ell,1}(s_1) = \mathbb{E}_{\mathbb{P},\pi}\left[\sum_{h=1}^{H} \hat{Q}_{\ell,h}(s_h,\cdot)^\top(\pi_h(\cdot \mid s_h) - \hat{\pi}_h(\cdot \mid s_h)) \,\middle|\, s_1\right] + \mathbb{E}_{\mathbb{P},\pi}\left[\sum_{h=1}^{H} \ell_h(s_h,a_h) + \mathbb{P}_h\hat{V}_{\ell,h+1}(s_h,a_h) - \hat{Q}_{\ell,h}(s_h,a_h) \,\middle|\, s_1\right],$$

where $V_{\ell,1}^{\pi}$ is the value function of π, and $\mathbb{P}_h\hat{V}_{\ell,h+1}(s,a) = \sum_{s' \in \mathcal{S}} \mathbb{P}_h(s' \mid s,a)\hat{V}_{\ell,h+1}(s')$.

D.2 Adaptive FTRL

In this subsection, we review adaptive FTRL for online convex optimization when the solution space is the probability simplex $\Delta_{n-1} = \{\mathbf{x} \in \mathbb{R}^n : \mathbf{x}_i \geq 0 \text{ for all } i \text{ and } \sum_{i=1}^n \mathbf{x}_i = 1\}$. While these results are standard in the literature, we include them for completeness. Most of the proofs in this section are taken from Chapter 7 in Orabona (2019) with a few modifications. We consider the following algorithm.

Algorithm 2 Adaptive FTRL for online convex optimization with the probability simplex

Initialize: $\mathbf{x}_1 = (1/n)\mathbf{1}$

for $k = 1, \ldots, K$ **do**

Take $\mathbf{x}_k$ and observe $\ell_k(\mathbf{x}) = \nabla_k^\top \mathbf{x}$

Update $\psi_k(\mathbf{x}) = \alpha_k \sum_{i=1}^n \mathbf{x}_i \log \mathbf{x}_i$ and α_k is defined as follows:

$$\alpha_k = \frac{1}{4}\sqrt{\sum_{\tau=1}^{k} L_\tau^2} \quad \text{where} \quad ||\nabla_\tau||_\infty \leq L_\tau \ \forall \tau \in [k]$$

Update $\mathbf{x}_{k+1}$:

$$\mathbf{x}_{k+1} \in \operatorname*{arg\,min}_{\mathbf{x} \in \Delta_{n-1}} \ \psi_k(\mathbf{x}) + \sum_{\tau=1}^{k} \nabla_\tau^\top \mathbf{x} \tag{11}$$

end for

Note that the $\mathbf{x}_{k+1}$ defined by (11) is uniquely determined and its closed form is written as follows (see Example 28.3 in Lattimore & Szepesvári (2020)): for each $i \in [n]$,

$$\mathbf{x}_{k+1,i} = \frac{\exp\left(-\frac{1}{\alpha_k}\sum_{\tau=1}^{k} \nabla_{\tau,i}\right)}{\sum_{j=1}^{n} \exp\left(-\frac{1}{\alpha_k}\sum_{\tau=1}^{k} \nabla_{\tau,j}\right)}. \tag{12}$$

The following lemma establishes a regret bound under Algorithm 2.

Lemma 11. *Suppose that we run Algorithm 2 and $L_1 > 0$. For any $\mathbf{x}^\star \in \Delta_{n-1}$ and $k \in [K]$, we have*

$$\mathrm{Regret}_{\textit{AdaFTRL}}(k) \leq \left(\frac{\log n}{4} + 4\sqrt{2}\right)\sqrt{\sum_{\tau=1}^{k} L_\tau^2}. \tag{13}$$

Comparison with AdaOMD For comparison, we briefly introduce the analysis of AdaOMD, which is the most natural extension of Adaptive Gradient Descent to the probability simplex. In particular, the update of AdaOMD can be written as follows:

$$\mathbf{x}_{k+1} \in \operatorname*{arg\,min}_{\mathbf{x} \in \Delta_{n-1}} \quad \frac{1}{\eta_k} D_{\mathrm{KL}}(\mathbf{x}||\mathbf{x}_k) + \nabla_k^\top \mathbf{x}. \tag{14}$$

where $\eta_k \asymp 1/\sqrt{\sum_{\tau=1}^{k} ||\nabla_\tau||_\infty^2}$. Its regret bound is given as follows:

$$\mathrm{Regret}_{\mathrm{AdaOMD}}(K) \lesssim \sup_{\mathbf{x},\mathbf{y}\in\Delta_{n-1}} D_{\mathrm{KL}}(\mathbf{x}||\mathbf{y}) \sqrt{\sum_{k=1}^{K} ||\nabla_k||_\infty^2}. \tag{15}$$

Note that since $\sup_{\mathbf{x},\mathbf{y}\in\Delta_{n-1}} D_{\mathrm{KL}}(\mathbf{x}||\mathbf{y})$ is unbounded, the regert bound of AdaOMD with over the probability simplex is unbounded. That said, while the right-hand side of (15) is indeed invalid, we show it for comparison with (13). The regret bound of AdaFTRL as in (13) takes $\log n$ instead of $\sup_{\mathbf{x},\mathbf{y}\in\Delta_{n-1}} D_{\mathrm{KL}}(\mathbf{x}||\mathbf{y})$. Therefore, AdaFTRL is appropriate for MDPs in which we must handle the policy space.

D.2.1 Proof of Lemma 11

Before providing the regret analysis of Algorithm 2, we first introduce several notations and lemmas. Throughout the section, for $k = 1, \dots, K$, we define

$$\bar{\psi}_k(\mathbf{x}) = \alpha_k \left(\sum_{i=1}^{n} \mathbf{x}_i \log \mathbf{x}_i + \log n \right),$$

$$F_{k+1}(\mathbf{x}) = \bar{\psi}_k(\mathbf{x}) + \sum_{\tau=1}^{k} \nabla_\tau^\top \mathbf{x} \quad \text{and} \quad F_1(\mathbf{x}) = 0$$

Note that the minimum of $\sum_{i=1}^{n} \mathbf{x}_i \log \mathbf{x}_i$ is $-\log n$ when $\mathbf{x} = (1/n)\mathbf{1}$, and thus $\bar{\psi}_k$ is nonnegative. Under this definition, we have the following useful relation, which follows from $\{\alpha_k\}_{k=1}^{K}$ is a nondecreasing sequence:

$$\bar{\psi}_{k-1}(\mathbf{x}) - \bar{\psi}_k(\mathbf{x}) \le 0, \quad \forall \mathbf{x} \in \Delta_{n-1},\ k = 2, \dots, K$$

Moreover, since the only difference between $\bar{\psi}_k$ and ψ_k is the term $\alpha_k \log n$, which is independent of $\mathbf{x}$, we have $\mathbf{x}_{k+1} \in \arg\min_{\mathbf{x}\in\Delta_{n-1}} F_{k+1}(\mathbf{x})$ even when F_{k+1} is defined in $\bar{\psi}_k$. Now we introduce the following lemma as a first step of regret analysis for Algorithm 2.

Lemma 12. *Suppose that we run Algorithm 2 and $L_1 > 0$. For any $\mathbf{x}^\star \in \Delta_{n-1}$ and $k \in [K]$, we have*

$$\sum_{\tau=1}^{k} \nabla_\tau^\top (\mathbf{x}_\tau - \mathbf{x}^\star) \le \bar{\psi}_k(\mathbf{x}^\star) + \sum_{\tau=1}^{k} \left(F_\tau(\mathbf{x}_\tau) - F_{\tau+1}(\mathbf{x}_{\tau+1}) + \nabla_\tau^\top \mathbf{x}_\tau \right)$$

Proof. By the definition of F_{k+1},

$$\sum_{\tau=1}^{k} \nabla_\tau^\top \mathbf{x}^\star = F_{k+1}(\mathbf{x}^\star) - \bar{\psi}_k(\mathbf{x}^\star)$$

Moreover, by telescoping, we have

$$0 = -F_1(\mathbf{x}_1) + \sum_{\tau=1}^{k} (F_\tau(\mathbf{x}_\tau) - F_{\tau+1}(\mathbf{x}_{\tau+1})) + F_{k+1}(\mathbf{x}_{k+1}).$$

Then it follows that

$$\begin{aligned}
&\sum_{\tau=1}^{k} \left(\nabla_\tau^\top \mathbf{x}_\tau - \nabla_\tau^\top \mathbf{x}^\star \right) \\
&= \bar{\psi}_k(\mathbf{x}^\star) - F_1(\mathbf{x}_1) + \sum_{\tau=1}^{k} \left(F_\tau(\mathbf{x}_\tau) - F_{\tau+1}(\mathbf{x}_{\tau+1}) + \nabla_\tau^\top \mathbf{x}_\tau \right) \\
&\quad + F_{k+1}(\mathbf{x}_{k+1}) - F_{k+1}(\mathbf{x}^\star) \\
&\le \bar{\psi}_k(\mathbf{x}^\star) + \sum_{\tau=1}^{k} \left(F_\tau(\mathbf{x}_\tau) - F_{\tau+1}(\mathbf{x}_{\tau+1}) + \nabla_\tau^\top \mathbf{x}_\tau \right)
\end{aligned}$$

where the inequality follows from the fact that $\mathbf{x}_{k+1} \in \arg\min F_{k+1}(\mathbf{x})$ and $F_1(\mathbf{x}_1) = 0$. □

Lemma 13. *Suppose that we run Algorithm 2 and $L_1 > 0$. We have*

1. *For $k = 2, \ldots, K$, we have*
$$F_k(\mathbf{x}_k) - F_{k+1}(\mathbf{x}_{k+1}) + \nabla_k^\top \mathbf{x}_k \leq \frac{||\nabla_k||_\infty^2}{2\alpha_{k-1}}.$$
2. *For $k = 1, \ldots, K$, we have*
$$F_k(\mathbf{x}_k) - F_{k+1}(\mathbf{x}_{k+1}) + \nabla_k^\top \mathbf{x}_k \leq 2||\nabla_k||_\infty$$
3. *For $k = 1, \ldots, K$, we have*
$$F_k(\mathbf{x}_k) - F_{k+1}(\mathbf{x}_{k+1}) + \nabla_k^\top \mathbf{x}_k \leq \frac{2\sqrt{2}L_k^2}{\sqrt{\sum_{\tau=1}^k L_\tau^2}}.$$

Proof. **Proof of the first statement:** Note that

$$F_{k+1}(\mathbf{x}_{k+1}) = \bar{\psi}_k(\mathbf{x}_{k+1}) + \sum_{\tau=1}^{k} \nabla_\tau^\top \mathbf{x}_{k+1},$$
$$F_k(\mathbf{x}_{k+1}) = \bar{\psi}_{k-1}(\mathbf{x}_{k+1}) + \sum_{\tau=1}^{k-1} \nabla_\tau^\top \mathbf{x}_{k+1}.$$

This implies that for $k = 2, \ldots, K$,

$$F_{k+1}(\mathbf{x}_{k+1}) - F_k(\mathbf{x}_{k+1}) = \bar{\psi}_k(\mathbf{x}_{k+1}) - \bar{\psi}_{k-1}(\mathbf{x}_{k+1}) + \nabla_k^\top \mathbf{x}_{k+1}.$$

Equivalently, we have $F_{k+1}(\mathbf{x}_{k+1}) = F_k(\mathbf{x}_{k+1}) + \bar{\psi}_k(\mathbf{x}_{k+1}) - \bar{\psi}_{k-1}(\mathbf{x}_{k+1}) + \nabla_k^\top \mathbf{x}_{k+1}$. Then it follows that for $k = 2, \ldots, K$,

$$\begin{aligned}
&F_k(\mathbf{x}_k) - F_{k+1}(\mathbf{x}_{k+1}) + \nabla_k^\top \mathbf{x}_k \\
&= F_k(\mathbf{x}_k) + \nabla_k^\top \mathbf{x}_k - (F_k(\mathbf{x}_{k+1}) + \nabla_k^\top \mathbf{x}_{k+1}) + \bar{\psi}_{k-1}(\mathbf{x}_{k+1}) - \bar{\psi}_k(\mathbf{x}_{k+1}) \\
&\leq F_k(\mathbf{x}_k) + \nabla_k^\top \mathbf{x}_k - (F_k(\mathbf{x}_{k+1}) + \nabla_k^\top \mathbf{x}_{k+1})
\end{aligned}$$

where the inequality follows from the fact that $\{\alpha_k\}_{k=1}^K$ is a nondecreasing sequence, implying that $\bar{\psi}_{k-1}(\mathbf{x}) - \bar{\psi}_k(\mathbf{x}) \leq 0$ for any $\mathbf{x}$ and $k = 2, \ldots, K$.

To further deduce, recall that $\bar{\psi}_{k-1}$ is α_{k-1}-strongly convex. This implies that F_k is α_{k-1}-strongly convex and thus $F_k(\mathbf{x}) + \nabla_k^\top \mathbf{x} + \mathbb{1}_{\Delta_{n-1}}(\mathbf{x})$ is also α_{k-1}-strongly convex, where $\mathbb{1}_{\Delta_{n-1}}$ denotes the indicator function with respect to Δ_{n-1}. Moreover, as $\mathbf{x}_k \in \arg\min_{\mathbf{x}\in\Delta_{n-1}} F_k(\mathbf{x})$, we have $\mathbf{0} \in \partial(F_k + \mathbb{1}_{\Delta_{n-1}})(\mathbf{x}_k)$. Then this implies that ∇_k is a subgradient of $F_k(\mathbf{x}) + \nabla_k^\top \mathbf{x} + \mathbb{1}_{\Delta_{n-1}}(\mathbf{x})$ at $\mathbf{x}_k$. It follows that

$$\begin{aligned}
&F_k(\mathbf{x}_k) + \nabla_k^\top \mathbf{x}_k - \big(F_k(\mathbf{x}_{k+1}) + \nabla_k^\top \mathbf{x}_{k+1}\big) \\
&= F_k(\mathbf{x}_k) + \nabla_k^\top \mathbf{x}_k + \mathbb{1}_{\Delta_{n-1}}(\mathbf{x}_k) - \big(F_k(\mathbf{x}_{k+1}) + \nabla_k^\top \mathbf{x}_{k+1} + \mathbb{1}_{\Delta_{n-1}}(\mathbf{x}_{k+1})\big) \\
&\leq \nabla_k^\top (\mathbf{x}_k - \mathbf{x}_{k+1}) - \frac{\alpha_{k-1}}{2}||\mathbf{x}_{k+1} - \mathbf{x}_k||_1^2 \\
&\leq ||\nabla_k||_\infty ||\mathbf{x}_k - \mathbf{x}_{k+1}||_1 - \frac{\alpha_{k-1}}{2}||\mathbf{x}_{k+1} - \mathbf{x}_k||_1^2 \\
&\leq \frac{||\nabla_k||_\infty^2}{2\alpha_{k-1}}
\end{aligned}$$

where the equality follows from the fact that $\mathbf{x}_k, \mathbf{x}_{k+1} \in \Delta_{n-1}$, the first inequality follows from the first-order property of α_{k-1}-strongly convex function, the second inequality is due to Hölder's inequality, and the last inequality follows from the fact that $bx - ax^2 \leq b^2/(4a)$. We deduce that for $k = 2, \ldots, K$,

$$F_k(\mathbf{x}_k) - F_{k+1}(\mathbf{x}_{k+1}) + \nabla_k^\top \mathbf{x}_k \leq \frac{||\nabla_k||_\infty^2}{2\alpha_{k-1}}.$$

Proof of the second statement: Since $\mathbf{x}_k \in \arg\min_{\mathbf{x}\in\Delta_{n-1}} F_k(\mathbf{x})$, we have $F_k(\mathbf{x}_k) \le F_k(\mathbf{x}_{k+1})$. It follows that for $k = 2, \ldots, K$,

$$\begin{aligned} F_k(\mathbf{x}_k) - F_{k+1}(\mathbf{x}_{k+1}) + \nabla_k^\top \mathbf{x}_k &\le F_k(\mathbf{x}_{k+1}) - F_{k+1}(\mathbf{x}_{k+1}) + \nabla_k^\top \mathbf{x}_k \\ &= \bar{\psi}_{k-1}(\mathbf{x}_{k+1}) - \bar{\psi}_k(\mathbf{x}_{k+1}) - \nabla_k^\top \mathbf{x}_{k+1} + \nabla_k^\top \mathbf{x}_k \\ &\le -\nabla_k^\top \mathbf{x}_{k+1} + \nabla_k^\top \mathbf{x}_k \\ &\le 2||\nabla_k||_\infty \end{aligned}$$

where the second inequality follows from the fact that α_k is nondecreasing, and the last inequality follows from Hölder's inequality and the fact that $||\mathbf{x}_{k+1} - \mathbf{x}_k||_1 \le 2$.

Next, we consider $k = 1$. In this case,

$$\begin{aligned} F_1(\mathbf{x}_1) - F_2(\mathbf{x}_2) + \nabla_1^\top \mathbf{x}_1 &= -\bar{\psi}_1(\mathbf{x}_2) + \nabla_1^\top (\mathbf{x}_1 - \mathbf{x}_2) \\ &\le \nabla_1^\top (\mathbf{x}_1 - \mathbf{x}_2) \\ &\le 2||\nabla_1||_\infty \end{aligned}$$

where the first inequality follows from the fact that $\bar{\psi}_1(\mathbf{x}_2) \ge 0$, and the second inequality follows from Hölder's inequality and the fact that $||\mathbf{x}_1 - \mathbf{x}_2||_1 \le 2$. Hence, the second statement holds for all $k = 1, \ldots, K$.

Proof of the third statement: By the first and second statements, for $k = 2, \ldots, K$, we have

$$\begin{aligned} F_k(\mathbf{x}_k) - F_{k+1}(\mathbf{x}_{k+1}) + \nabla_k^\top \mathbf{x}_k &\le \min\left\{ \frac{||\nabla_k||_\infty^2}{2\alpha_{k-1}}, 2||\nabla_k||_\infty \right\} \\ &= \min\left\{ \frac{2||\nabla_k||_\infty^2}{\sqrt{\sum_{\tau=1}^{k-1} L_\tau^2}}, 2||\nabla_k||_\infty \right\} \\ &\le \min\left\{ \frac{2L_k^2}{\sqrt{\sum_{\tau=1}^{k-1} L_\tau^2}}, 2L_k \right\} \\ &= 2\sqrt{\min\left\{ \frac{L_k^4}{\sum_{\tau=1}^{k-1} L_\tau^2}, L_k^2 \right\}} \\ &\le 2\sqrt{\frac{2}{\frac{\sum_{\tau=1}^{k-1} L_\tau^2}{L_k^4} + \frac{1}{L_k^2}}} \\ &= \frac{2\sqrt{2}L_k^2}{\sqrt{\sum_{\tau=1}^{k} L_\tau^2}} \end{aligned}$$

where the second inequality follows from $||\nabla_k||_\infty \le L_k$, and the third inequality follows from the fact that $\min\{a, b\} \le \frac{2}{(1/a)+(1/b)}$, assuming $L_k > 0$. Note that if $L_k = 0$, then the desired relation holds trivially. For $k = 1$, by the second statement, we have

$$F_1(\mathbf{x}_1) - F_2(\mathbf{x}_2) + \nabla_1^\top \mathbf{x}_1 \le 2||\nabla_1||_\infty \le 2L_1 \le \frac{2\sqrt{2}L_1^2}{\sqrt{\sum_{\tau=1}^{1} L_\tau^2}}.$$

Hence, the third statement holds for all $k = 1, \ldots, K$. □

Proof of Lemma 11. By Lemma 12 and the third statement of Lemma 13, we have

$$\begin{aligned} \sum_{\tau=1}^{k} \nabla_\tau^\top (\mathbf{x}_\tau - \mathbf{x}^\star) &\le \bar{\psi}_k(\mathbf{x}^\star) + \sum_{\tau=1}^{k} \frac{2\sqrt{2}L_\tau^2}{\sqrt{\sum_{\tau'=1}^{\tau} L_{\tau'}^2}} \\ &= \alpha_k \left(\sum_{i=1}^{n} \mathbf{x}_i^\star \log \mathbf{x}_i^\star + \log n \right) + \sum_{\tau=1}^{k} \frac{2\sqrt{2}L_\tau^2}{\sqrt{\sum_{\tau'=1}^{\tau} L_{\tau'}^2}} \end{aligned}$$

$$\leq \alpha_k \log n + \sum_{\tau=1}^{k} \frac{2\sqrt{2}L_\tau^2}{\sqrt{\sum_{\tau'=1}^{\tau} L_{\tau'}^2}}$$
$$\leq \alpha_k \log n + 4\sqrt{2}\sqrt{\sum_{\tau=1}^{k} L_\tau^2}$$
$$= \frac{\log n}{4}\sqrt{\sum_{\tau=1}^{k} L_\tau^2} + 4\sqrt{2}\sqrt{\sum_{\tau=1}^{k} L_\tau^2}$$

where the second inequality follows from the fact that $\sum_{i=1}^{n} \mathbf{x}_i^\star \log \mathbf{x}_i^\star \leq 0$, and the third inequality follows from $\sum_{\tau=1}^{k} y_\tau / \sqrt{\sum_{\tau'=1}^{\tau} y_{\tau'}} \leq 2\sqrt{\sum_{\tau=1}^{k} y_\tau}$ for any nonnegative sequence $\{y_\tau\}_{\tau \geq 1}$ with $y_1 > 0$. This completes the proof. □

D.2.2 Proof Overview of Eq. (15)

Under the update of (14), we can use the following standard fact (e.g., Lemma 14 in Wei et al. (2020)):

$$\nabla_k^\top \mathbf{x}_{k+1} + \frac{1}{\eta_k} D_{\mathrm{KL}}(\mathbf{x}_{k+1} || \mathbf{x}_k) \leq \nabla_k^\top \mathbf{x}^\star + \frac{1}{\eta_k} D_{\mathrm{KL}}(\mathbf{x}^\star || \mathbf{x}_k) - \frac{1}{\eta_k} D_{\mathrm{KL}}(\mathbf{x}^\star || \mathbf{x}_{k+1}).$$

It follows that

$$\begin{aligned}
&\nabla_k^\top (\mathbf{x}_k - \mathbf{x}^\star) \\
&\leq \nabla_k^\top (\mathbf{x}_k - \mathbf{x}_{k+1}) + \frac{1}{\eta_k} D_{\mathrm{KL}}(\mathbf{x}^\star || \mathbf{x}_k) - \frac{1}{\eta_k} D_{\mathrm{KL}}(\mathbf{x}^\star || \mathbf{x}_{k+1}) - \frac{1}{\eta_k} D_{\mathrm{KL}}(\mathbf{x}_{k+1} || \mathbf{x}_k) \\
&\leq ||\nabla_k||_\infty ||\mathbf{x}_k - \mathbf{x}_{k+1}||_1 - \frac{1}{2\eta_k} ||\mathbf{x}_{k+1} - \mathbf{x}_k||_1^2 + \frac{1}{\eta_k} D_{\mathrm{KL}}(\mathbf{x}^\star || \mathbf{x}_k) - \frac{1}{\eta_k} D_{\mathrm{KL}}(\mathbf{x}^\star || \mathbf{x}_{k+1}) \\
&\leq 8||\nabla_k||_\infty^2 \eta_k + \frac{1}{\eta_k} D_{\mathrm{KL}}(\mathbf{x}^\star || \mathbf{x}_k) - \frac{1}{\eta_k} D_{\mathrm{KL}}(\mathbf{x}^\star || \mathbf{x}_{k+1}).
\end{aligned}$$

where the second inequality follows from Hölder's inequality and Pinsker's inequality, and the second inequality follows from the fact that $bx - ax^2 \leq 4b^2/a$. By summing over $k = 1, \ldots, K$, we have

$$\begin{aligned}
\mathrm{Regret}_{\mathrm{AdaOMD}}(K) &= \sum_{k=1}^{K} \nabla_k^\top (\mathbf{x}_k - \mathbf{x}^\star) \\
&\lesssim \sum_{k=1}^{K} 8||\nabla_k||_\infty^2 \eta_k + \sum_{k=1}^{K} \left(\frac{1}{\eta_k} D_{\mathrm{KL}}(\mathbf{x}^\star || \mathbf{x}_k) - \frac{1}{\eta_k} D_{\mathrm{KL}}(\mathbf{x}^\star || \mathbf{x}_{k+1}) \right) \\
&\lesssim \sum_{k=1}^{K} 8||\nabla_k||_\infty^2 \eta_k + \sum_{k=1}^{K-1} \left(\frac{1}{\eta_{k+1}} - \frac{1}{\eta_k} \right) D_{\mathrm{KL}}(\mathbf{x}^\star || \mathbf{x}_{k+1}) \\
&\lesssim \sqrt{\sum_{k=1}^{K} ||\nabla_k||_\infty^2} + \sup_{\mathbf{x},\mathbf{y} \in \Delta_{n-1}} D_{\mathrm{KL}}(\mathbf{x}||\mathbf{y}) \sum_{k=1}^{K-1} \left(\frac{1}{\eta_{k+1}} - \frac{1}{\eta_k} \right) \\
&\lesssim \sqrt{\sum_{k=1}^{K} ||\nabla_k||_\infty^2} + \frac{\sup_{\mathbf{x},\mathbf{y} \in \Delta_{n-1}} D_{\mathrm{KL}}(\mathbf{x}||\mathbf{y})}{\eta_K} \\
&\lesssim \sup_{\mathbf{x},\mathbf{y} \in \Delta_{n-1}} D_{\mathrm{KL}}(\mathbf{x}||\mathbf{y}) \sqrt{\sum_{k=1}^{K} ||\nabla_k||_\infty^2}
\end{aligned}$$

where the third inequality follows from that fact that $\sum_{k=1}^{K} y_k / \sqrt{\sum_{\tau=1}^{k} y_k} \leq 2\sqrt{\sum_{k=1}^{K} y_k}$, and the fourth inequality is due to the telescoping sum. □

E Parameterization and Covering Number

In this subsection, we introduce parameterizations for Q-function estimates $\hat{Q}$, policies π, and value function estimates $\hat{V}$, which are useful for analyzing covering number (Cassel & Rosenberg, 2024; Yu et al., 2026b). For the Q-function estimate, given $\beta \in \mathbb{R}$, $\mathbf{w} \in \mathbb{R}^d$, and $\mathbf{\Lambda} \in \mathbb{R}^{d\times d}$, we define $\hat{Q}(\cdot,\cdot;\beta,\mathbf{w},\mathbf{\Lambda})$ as

$$\hat{Q}(\cdot,\cdot;\beta,\mathbf{w},\mathbf{\Lambda}) = (\boldsymbol{\phi}(\cdot,\cdot)^\top \mathbf{w} - \beta||\boldsymbol{\phi}(\cdot,\cdot)||_{\mathbf{\Lambda}}) \cdot \sigma(-\beta_\phi||\boldsymbol{\phi}(\cdot,\cdot)||_{\mathbf{\Lambda}} + \log K).$$

Note that $\hat{Q}(\cdot,\cdot;\beta,\mathbf{w},\mathbf{\Lambda})$ captures $\{\hat{Q}^k_{f,h}, \hat{Q}^k_{g,h}\}_{h\in[H],k\in[K]}$ in Algorithm 1, since these can be written as for each $\ell = f, g$ and $(h,k) \in [H]\times[K]$,

$$\begin{aligned}\hat{Q}^k_{\ell,h}(\cdot,\cdot) &= \bar{\boldsymbol{\phi}}^m_h(\cdot,\cdot)^\top \left[\boldsymbol{\theta}^k_{\ell,h} + \hat{\boldsymbol{\mu}}^k_h \hat{V}^k_{\ell,h+1}\right] - \beta_p||\bar{\boldsymbol{\phi}}^m_h(\cdot,\cdot)||_{(\mathbf{\Lambda}^{k(m)}_h)^{-1}} \\ &= \left(\boldsymbol{\phi}(\cdot,\cdot)^\top \left[\boldsymbol{\theta}^k_{\ell,h} + \hat{\boldsymbol{\mu}}^k_h \hat{V}^k_{\ell,h+1}\right] - \beta_p||\boldsymbol{\phi}(\cdot,\cdot)||_{(\mathbf{\Lambda}^{k(m)}_h)^{-1}}\right) \cdot \sigma(-\beta_\phi||\boldsymbol{\phi}(\cdot,\cdot)||_{(\mathbf{\Lambda}^{k(m)}_h)^{-1}} + \log K).\end{aligned}$$

Given some nonnegative boundedness constants $C_\beta, C_{\mathbf{w}}, C_Q$, we define the function class $\mathcal{Q}(C_\beta, C_{\mathbf{w}}, C_Q)$ as

$$\mathcal{Q}(C_\beta, C_{\mathbf{w}}, C_Q) = \left\{ \hat{Q}(\cdot,\cdot;\beta,\mathbf{w},\mathbf{\Lambda}) : \begin{array}{l} |\beta| \le C_\beta,\ ||\mathbf{w}|| \le C_{\mathbf{w}},\ (2K)^{-1}\mathbf{I} \preceq \mathbf{\Lambda} \preceq \mathbf{I}, \\ ||\hat{Q}(\cdot,\cdot;\beta,\mathbf{w},\mathbf{\Lambda})||_\infty \le C_Q \end{array} \right\}.$$

Next, we define the parameterization for π, which is defined, for each $(s,a) \in \mathcal{S}\times\mathcal{A}$,

$$\pi(a|s;\beta,\mathbf{w},\mathbf{\Lambda}) = \frac{\exp(\hat{Q}(s,a;\beta,\mathbf{w},\mathbf{\Lambda}))}{\sum_{a'\in\mathcal{A}} \exp(\hat{Q}(s,a';\beta,\mathbf{w},\mathbf{\Lambda}))}.$$

By (12), we know that $\pi(\cdot|\cdot;\beta,\mathbf{w},\mathbf{\Lambda})$ captures $\{\pi^{k+1}_h\}_{h\in[H],k\in[K]}$ in Algorithm 1. Subsequently, we define the policy class $\Pi(C_\beta, C_{\mathbf{w}}, C_Q, C_Y)$ as

$$\Pi(C_\beta, C_{\mathbf{w}}, C_Q, C_Y) = \left\{ \pi(\cdot|\cdot;\beta,\mathbf{w},\mathbf{\Lambda}) : \begin{array}{l} |\beta| \le (1+C_Y)C_\beta,\ ||\mathbf{w}|| \le (1+C_Y)C_{\mathbf{w}}, \\ (2K)^{-1}\mathbf{I} \preceq \mathbf{\Lambda} \preceq \mathbf{I}, \\ ||\hat{Q}(\cdot,\cdot;\beta,\mathbf{w},\mathbf{\Lambda})||_\infty \le (1+C_Y)C_Q \end{array} \right\}.$$

Finally, we define the function class of value function estimates $\mathcal{V}(C_\beta, C_{\mathbf{w}}, C_Q, C_Y)$ as follows.

$$\mathcal{V}(C_\beta, C_{\mathbf{w}}, C_Q, C_Y) = \left\{ \hat{V}(\cdot) : \hat{V}(\cdot) = \sum_{a\in\mathcal{A}} \pi(a|\cdot)\hat{Q}(\cdot,a), \begin{array}{l} \hat{Q} \in \mathcal{Q}(C_\beta, C_{\mathbf{w}}, C_Q), \\ \pi \in \Pi(KC_\beta, KC_{\mathbf{w}}, KC_Q, C_Y) \end{array} \right\}.$$

Lemma 14 (Lemma 10 in Yu et al. (2026b)). *Let $1 \le \beta_\phi, C_{\mathbf{w}}, C_\beta$. For any $\hat{Q}(\cdot,\cdot;\beta_1,\mathbf{w}_1,\mathbf{\Lambda}_1)$, $\hat{Q}(\cdot,\cdot;\beta_2,\mathbf{w}_2,\mathbf{\Lambda}_2) \in \mathcal{Q}(C_\beta, C_{\mathbf{w}}, C_Q)$, we have*

$$\begin{aligned}&\left\|\hat{Q}(\cdot,\cdot;\beta_1,\mathbf{w}_1,\mathbf{\Lambda}_1) - \hat{Q}(\cdot,\cdot;\beta_2,\mathbf{w}_2,\mathbf{\Lambda}_2)\right\|_\infty \\ &\le 4\sqrt{K}\beta_\phi \max\{C_{\mathbf{w}}, C_\beta\} \|(\beta_1,\mathbf{w}_1,\mathbf{\Lambda}_1) - (\beta_2,\mathbf{w}_2,\mathbf{\Lambda}_2)\|\end{aligned}$$

where $\|(\beta_1,\mathbf{w}_1,\mathbf{\Lambda}_1) - (\beta_2,\mathbf{w}_2,\mathbf{\Lambda}_2)\| = \sqrt{(\beta_1-\beta_2)^2 + ||\mathbf{w}_1-\mathbf{w}_2||^2 + ||\mathbf{\Lambda}_1-\mathbf{\Lambda}_2||_F^2}$.

The following lemmas are modifications of Lemmas 11 and 12 in Yu et al. (2026b). These are obtained by taking $n = 0$ for their lemmas. Here, n denotes the number of mixing steps that have been applied to the policy, where $n = 0$ in our case, as we do not apply the mixing step. For further details on the mixing step, we refer the reader to Yu et al. (2026b).

Lemma 15 (Lemma 11 in Yu et al. (2026b)). *Let $1 \le \beta_\phi, C_{\mathbf{w}}, C_\beta$, and $0 \le C_Y$. For any $\pi(\cdot|\cdot;\beta_1,\mathbf{w}_1,\mathbf{\Lambda}_1)$, $\pi(\cdot|\cdot;\beta_2,\mathbf{w}_2,\mathbf{\Lambda}_2) \in \Pi(C_\beta, C_{\mathbf{w}}, C_Q, C_Y)$, we have for any $s \in \mathcal{S}$,*

$$\begin{aligned}&||\pi(\cdot|s;\beta_1,\mathbf{w}_1,\mathbf{\Lambda}_1) - \pi(\cdot|s;\beta_2,\mathbf{w}_2,\mathbf{\Lambda}_2)||_1 \\ &\le 32\sqrt{K}\beta_\phi(1+C_Y)\max\{C_{\mathbf{w}}, C_\beta\}||(\beta_1,\mathbf{w}_1,\mathbf{\Lambda}_1) - (\beta_2,\mathbf{w}_2,\mathbf{\Lambda}_2)||\end{aligned}$$

Lemma 16 (Lemma 12 in Yu et al. (2026b)). *Let $1 \le \beta_\phi, C_{\mathbf{w}}, C_\beta, C_Q$, and $0 \le C_Y$. Given $\epsilon > 0$, let $\mathcal{N}_\epsilon(\mathcal{V}(C_\beta, C_{\mathbf{w}}, C_Q, C_Y))$ denote the ϵ-covering number of $\mathcal{V}(C_\beta, C_{\mathbf{w}}, C_Q, C_Y)$ with respect to ℓ_∞-norm. Then we have*

$$\log \mathcal{N}_\epsilon(\mathcal{V}(C_\beta, C_{\mathbf{w}}, C_Q, C_Y)) \le 12d^2 \log\left(1 + \frac{2C_1C_2}{\epsilon}\right)$$

where $C_1 = 33\sqrt{K^3}\beta_\phi \max\{C_{\mathbf{w}}, C_\beta\}C_Q(1+C_Y)$, $C_2 = 2(1+C_Y)K(C_\beta + C_{\mathbf{w}}) + 2\sqrt{d}$.

F GOOD EVENT

In this section, we introduce the good event $\mathcal{E}_g$. The definition of $\mathcal{E}_g$ is similar to that of Cassel & Rosenberg (2024); Yu et al. (2026b). The main differences stem from its analysis, since our algorithm considers the adaptive FTRL for π_h^{k+1}, whereas those works mainly focus on the OMD with a fixed step size, which is equivalent to the non-adaptive FTRL. Moreover, we also have to consider the sophisticated update rule for Y_{k+1}.

Now, we introduce the formal definition of $\mathcal{E}_g$ as follows. We define $\mathcal{E}_g = \mathcal{E}_1 \cap \mathcal{E}_2$, where

$$\mathcal{E}_1 = \left\{ \forall (k,h) \in [K] \times [H] : \begin{array}{c} ||(\boldsymbol{\mu}_h - \hat{\boldsymbol{\mu}}_h^k)\hat{V}_{f,h+1}^k||_{\boldsymbol{\Lambda}_h^k} \le \beta_p,\ ||\hat{Q}_{f,h}^k||_\infty \le \beta_Q, \\ ||(\boldsymbol{\mu}_h - \hat{\boldsymbol{\mu}}_h^k)\hat{V}_{g,h+1}^k||_{\boldsymbol{\Lambda}_h^k} \le \beta_p,\ ||\hat{Q}_{g,h}^k||_\infty \le \beta_Q \end{array} \right\},$$

$$\mathcal{E}_2 = \left\{ \sum_{\tau=1}^{k} \mathbb{E}_{\mathbb{P},\pi^\tau}[W_\tau] \le 2\sum_{\tau=1}^{k} W_\tau + 4H(3\beta_p + 8\beta_Q\beta_\phi^2)\log\frac{4K^2}{\delta}, \quad \forall k \in [K] \right\}$$

Here, we define $\beta_p, \beta_Q, \beta_\phi$, and W_k as follows:

$$\begin{aligned} \beta_p &= 187dH\sqrt{\log(2HK/\delta)}, \\ \beta_Q &= 2H, \\ \beta_\phi &= 4\beta_p \log K, \\ W_\tau &= \sum_{h=1}^{H} \left(3\beta_p ||\boldsymbol{\phi}(s_h^\tau, a_h^\tau)||_{(\boldsymbol{\Lambda}_h^\tau)^{-1}} + 8\beta_Q\beta_\phi^2 ||\boldsymbol{\phi}(s_h^\tau, a_h^\tau)||^2_{(\boldsymbol{\Lambda}_h^\tau)^{-1}} \right). \end{aligned}$$

Before proving that $\mathcal{E}_g$ holds with high probability, we first define the proxy good event $\bar{\mathcal{E}}_g = \bar{\mathcal{E}}_1 \cap \mathcal{E}_2$, where

$$\bar{\mathcal{E}}_1 = \left\{ \forall (k,h,\hat{V}) \in [K] \times [H] \times \mathcal{V}(\beta_p, \sqrt{d} + K\beta_Q, \beta_Q, K^8) : ||(\boldsymbol{\mu}_h - \hat{\boldsymbol{\mu}}_h^k)\hat{V}||_{\boldsymbol{\Lambda}_h^k} \le \beta_p \right\}.$$

We first prove that the proxy good event $\bar{\mathcal{E}}_g$ holds with high probability.

Lemma 17. *Let* $1 \le \beta_\phi \le K$. *Then* $\Pr(\bar{\mathcal{E}}_g) \ge 1 - \delta$ *for any* $\delta \in (0,1)$.

Proof. We closely follow the proof of Lemma 14 in Yu et al. (2026b). We first prove that $\Pr(\bar{\mathcal{E}}_1) \ge 1 - \delta/2$ and $\Pr(\mathcal{E}_2) \ge 1 - \delta/2$ separately, and then apply a union bound to show $\Pr(\bar{\mathcal{E}}_g) \ge 1 - \delta$.

To show that $\Pr(\bar{\mathcal{E}}_1) \ge 1 - \delta/2$, we bound the ϵ-covering number of the value function class $\mathcal{V}(\beta_p, \sqrt{d} + K\beta_Q, \beta_Q, K^8)$ for some $\epsilon > 0$, where ϵ will be specified later. Note that $1 \le \beta_\phi,\ \sqrt{d} + K\beta_Q,\ \beta_p$, and $0 \le K^8$, meaning that the conditions of Lemma 16 are satisfied. Then, since we have $||\hat{V}||_\infty \le \beta_Q$ for any $\hat{V} \in \mathcal{V}(\beta_p, \sqrt{d} + K\beta_Q, \beta_Q, K^8)$, by Lemma 16, we have

$$\log \mathcal{N}_\epsilon \left(\mathcal{V}(\beta_p, \sqrt{d} + K\beta_Q, \beta_Q, K^8) \right) \le 12d^2 \log\left(1 + \frac{2C_1C_2}{\epsilon}\right).$$

Here, C_1, C_2 can be bounded as follows. Since $\sqrt{d} \le \beta_p \le \beta_\phi \le K$ and $\beta_Q = 2H$, we have

$$\begin{aligned} C_1 &= 33\sqrt{K^3}\beta_\phi \max\{\sqrt{d} + K\beta_Q, \beta_p\}\beta_Q(1 + K^8) \le 528H^2K^{11.5}, \\ C_2 &= 2(1 + K^8)K(\beta_p + \sqrt{d} + K\beta_Q) + 2\sqrt{d} \le 22\sqrt{d}HK^{10}. \end{aligned}$$

Then it follows that

$$\begin{aligned} \log \mathcal{N}_\epsilon \left(\mathcal{V}(\beta_p, \sqrt{d} + K\beta_Q, \beta_Q, K^8) \right) &\le 12d^2 \log\left(1 + \frac{23232\sqrt{d}H^3K^{21.5}}{\epsilon}\right) \\ &= 12d^2 \log\left(1 + 23232H^2K^{22.5}\right) \\ &\le 270d^2 \log(2HK). \end{aligned}$$

where the equality follows by taking $\epsilon = \beta_Q\sqrt{d}/(2K) = H\sqrt{d}/K$. Then by Lemma 28, we have with probability at least $1-\delta/2$, for all $h \in [H], k \in [K], \hat{V} \in \mathcal{V}(\beta_p, \sqrt{d}+K\beta_Q, \beta_Q, K^8)$,

$$\begin{aligned}||(\boldsymbol{\mu}_h - \hat{\boldsymbol{\mu}}_h^k)\hat{V}||_{\boldsymbol{\Lambda}_h^k} &\le 4\beta_Q\sqrt{d\log(K+1) + 2\log\mathcal{N}_\epsilon\left(\mathcal{V}(\beta_p, \sqrt{d}+K\beta_Q, \beta_Q, K^8)\right) + 2\log(2H^2/\delta)} \\ &\le 4\beta_Q\sqrt{d\log(K+1) + 540d^2\log(2HK) + 2\log(2H^2/\delta)} \\ &\le 187dH\sqrt{\log(2HK/\delta)}.\end{aligned}$$

Since we set $\beta_p = 187dH\sqrt{\log(2HK/\delta)}$, this implies that $\Pr(\bar{\mathcal{E}}_1) \ge 1-\delta/2$.

Next, we prove that $\mathcal{E}_2$ holds with probability at least $1-\delta/2$. Note that for any $(s,a) \in \mathcal{S}\times\mathcal{A}$,

$$\sum_{h=1}^H \left(3\beta_p||\boldsymbol{\phi}(s,a)||_{(\boldsymbol{\Lambda}_h^\tau)^{-1}} + 8\beta_Q\beta_\phi^2||\boldsymbol{\phi}(s,a)||^2_{(\boldsymbol{\Lambda}_h^\tau)^{-1}}\right) \le H(3\beta_p + 8\beta_Q\beta_\phi^2).$$

This implies that $W_\tau \le H(3\beta_p + 8\beta_Q\beta_\phi^2)$ almost surely. By the algorithm, the trajectory $\{(s_h^\tau, a_h^\tau)\}_{h\in[H]}$ are generated under $\mathbb{P}$ and π^τ for all $\tau \in [k]$. Moreover, since π^τ is $\mathcal{F}_{\tau-1,H}$-measurable, $\{W_\tau\}_{\tau\ge1}$ is adapted to the filtration $\{\mathcal{F}_{\tau,H}\}_{\tau\ge1}$. Then by Lemma 31 with probability at least $1-\delta/(2K)$, we have

$$\begin{aligned}\sum_{\tau=1}^k \mathbb{E}_{\mathbb{P},\pi^\tau}[W_\tau] &= \sum_{\tau=1}^k \mathbb{E}\left[\sum_{h=1}^H \left(3\beta_p||\boldsymbol{\phi}(s_h^\tau,a_h^\tau)||_{(\boldsymbol{\Lambda}_h^\tau)^{-1}} + 8\beta_Q\beta_\phi^2||\boldsymbol{\phi}(s_h^\tau,a_h^\tau)||^2_{(\boldsymbol{\Lambda}_h^\tau)^{-1}}\right) \mid \mathcal{F}_{\tau-1,H}\right] \\ &\le 2\sum_{\tau=1}^k W_\tau + 4H(3\beta_p + 8\beta_Q\beta_\phi^2)\log\frac{4K^2}{\delta}.\end{aligned}$$

By taking a union bound over k, this implies that $\Pr(\mathcal{E}_2) \ge 1-\delta/2$. Moreover, since $\Pr(\bar{\mathcal{E}}_1) \ge 1-\delta/2$ and $\Pr(\mathcal{E}_2) \ge 1-\delta/2$, by a union bound, we have $\Pr(\bar{\mathcal{E}}_g) \ge 1-\delta$, which concludes the proof. □

Lemma 18. *Suppose that $\bar{\mathcal{E}}_g$ holds. If $\pi_h^k \in \Pi(K\beta_p, K\sqrt{d}+K^2\beta_Q, K\beta_{Q,h}, K^8)$ for all $h\in[H]$, then we have $\hat{Q}_{f,h}^k, \hat{Q}_{g,h}^k \in \mathcal{Q}(\beta_p, \sqrt{d}+K\beta_Q, \beta_{Q,h})$ and $\hat{V}_{f,h}^k, \hat{V}_{g,h}^k \in \mathcal{V}(\beta_p, \sqrt{d}+K\beta_Q, \beta_{Q,h}, K^8)$ for all $h\in[H+1]$, where $\beta_{Q,h} = 2(H-h+1)$.*

Proof. We closely follow the proof of Lemma 15 in Yu et al. (2026b). We prove the lemma by backward induction on $h = H+1, \dots, 1$. For the base case, since Algorithm 1 sets $\hat{Q}_{f,H+1}^k(\cdot,\cdot), \hat{Q}_{g,H+1}^k(\cdot,\cdot) = 0$ and $\hat{V}_{f,H+1}^k(\cdot), \hat{V}_{g,H+1}^k(\cdot) = 0$, it is trivial that $\hat{Q}_{f,H+1}^k, \hat{Q}_{g,H+1}^k \in \mathcal{Q}(\beta_p, \sqrt{d}+K\beta_Q, \beta_{Q,H+1})$ and $\hat{V}_{f,H+1}^k, \hat{V}_{g,H+1}^k \in \mathcal{V}(\beta_p, \sqrt{d}+K\beta_Q, \beta_{Q,H+1}, K^8)$.

As an induction hypothesis, we assume that $\hat{Q}_{f,h+1}^k, \hat{Q}_{g,h+1}^k \in \mathcal{Q}(\beta_p, \sqrt{d}+K\beta_Q, \beta_{Q,h+1})$ and $\hat{V}_{f,h+1}^k, \hat{V}_{g,h+1}^k \in \mathcal{V}(\beta_p, \sqrt{d}+K\beta_Q, \beta_{Q,h+1}, K^8)$. Given this, we have to show that $\hat{Q}_{f,h}^k, \hat{Q}_{g,h}^k \in \mathcal{Q}(\beta_p, \sqrt{d}+K\beta_Q, \beta_{Q,h})$ and $\hat{V}_{f,h}^k, \hat{V}_{g,h}^k \in \mathcal{V}(\beta_p, \sqrt{d}+K\beta_Q, \beta_{Q,h}, K^8)$. First, for the case of f, we have

$$\begin{aligned}|\hat{Q}_{f,h}^k(s,a)| &= \left|\bar{\boldsymbol{\phi}}_h^m(s,a)^\top(\boldsymbol{\theta}_{f,h}^k + \hat{\boldsymbol{\mu}}_h^k\hat{V}_{f,h+1}^k) - \beta_p\left\|\bar{\boldsymbol{\phi}}_h^m(s,a)\right\|_{(\boldsymbol{\Lambda}_h^{k(m)})^{-1}}\right| \\ &\le \left|\bar{\boldsymbol{\phi}}_h^m(s,a)^\top(\boldsymbol{\theta}_{f,h}^k + \boldsymbol{\mu}_h\hat{V}_{f,h+1}^k)\right| + \left|\bar{\boldsymbol{\phi}}_h^m(s,a)^\top\left(\hat{\boldsymbol{\mu}}_h^k\hat{V}_{f,h+1}^k - \boldsymbol{\mu}_h\hat{V}_{f,h+1}^k\right)\right| \\ &\quad + \beta_p\left\|\bar{\boldsymbol{\phi}}_h^m(s,a)\right\|_{(\boldsymbol{\Lambda}_h^{k(m)})^{-1}} \\ &\le \left|\bar{\boldsymbol{\phi}}_h^m(s,a)^\top(\boldsymbol{\theta}_{f,h}^k + \boldsymbol{\mu}_h\hat{V}_{f,h+1}^k)\right| + \left\|(\hat{\boldsymbol{\mu}}_h^k - \boldsymbol{\mu}_h)\hat{V}_{f,h+1}^k\right\|_{\boldsymbol{\Lambda}_h^{k(m)}}\left\|\bar{\boldsymbol{\phi}}_h^m(s,a)\right\|_{(\boldsymbol{\Lambda}_h^{k(m)})^{-1}} \\ &\quad + \beta_p\left\|\bar{\boldsymbol{\phi}}_h^m(s,a)\right\|_{(\boldsymbol{\Lambda}_h^{k(m)})^{-1}}\end{aligned}$$

where the first inequality follows from the triangle inequality, and the second inequality follows from the Cauchy-Schwarz inequality. Note that $\hat{V}_{f,h+1} \in \mathcal{V}(\beta_p, \sqrt{d}+K\beta_Q, \beta_{Q,h+1}, K^8)$ by the induction hypothesis. Furthermore, we have

$$||(\hat{\boldsymbol{\mu}}_h^k - \boldsymbol{\mu}_h)\hat{V}_{f,h+1}^k||_{\boldsymbol{\Lambda}_h^{k(m)}} \le ||(\hat{\boldsymbol{\mu}}_h^k - \boldsymbol{\mu}_h)\hat{V}_{f,h+1}^k||_{\boldsymbol{\Lambda}_h^k} \le \beta_p$$

where the first inequality follows from the fact that $\mathbf{\Lambda}_h^{k(m)} \preceq \mathbf{\Lambda}_h^k$ and $\bar{\mathcal{E}}_1$. It leads to

$$\begin{aligned}
&|\hat{Q}_{f,h}^k(s,a)| \\
&\leq \left|\bar{\boldsymbol{\phi}}_h^m(s,a)^\top(\boldsymbol{\theta}_{f,h}^k + \boldsymbol{\mu}_h \hat{V}_{f,h+1}^k)\right| + 2\beta_p \left\|\bar{\boldsymbol{\phi}}_h^m\right\|_{(\mathbf{\Lambda}_h^{k(m)})^{-1}} \\
&= \sigma\left(-\beta_\phi ||\boldsymbol{\phi}(s,a)||_{(\mathbf{\Lambda}_h^{k(m)})^{-1}} + \log K\right)\left|f_h^k(s,a) + \mathbb{P}_h \hat{V}_{f,h+1}^k\right| + 2\beta_p \left\|\bar{\boldsymbol{\phi}}_h^m\right\|_{(\mathbf{\Lambda}_h^{k(m)})^{-1}} \\
&\leq \left|f_h^k(s,a) + \mathbb{P}_h \hat{V}_{f,h+1}^k(s,a)\right| + 2\beta_p \left\|\bar{\boldsymbol{\phi}}_h^m\right\|_{(\mathbf{\Lambda}_h^{k(m)})^{-1}} \\
&\leq 1 + ||\hat{V}_{f,h+1}^k||_\infty + 2\beta_p \left\|\bar{\boldsymbol{\phi}}_h^m\right\|_{(\mathbf{\Lambda}_h^{k(m)})^{-1}}
\end{aligned}$$

where the second inequality follows from the fact that $\sigma(z) \leq 1$ for any $z \in \mathbb{R}$, the last inequality follows from the fact that $|\mathbb{P}_h \hat{V}_{f,h+1}^k| = |\sum_{s'} \mathbb{P}_h(s'|s,a)\hat{V}_{f,h+1}^k(s')| \leq ||\hat{V}_{f,h+1}^k||_\infty$. Moreover, we have

$$\begin{aligned}
2\beta_p \left\|\bar{\boldsymbol{\phi}}_h^m\right\|_{(\mathbf{\Lambda}_h^{k(m)})^{-1}} &= 2\beta_p \sigma\left(-\beta_\phi ||\boldsymbol{\phi}(s,a)||_{(\mathbf{\Lambda}_h^{k(m)})^{-1}} + \log K\right) \|\boldsymbol{\phi}(s,a)\|_{(\mathbf{\Lambda}_h^{k(m)})^{-1}} \\
&\leq 2\beta_p \cdot \frac{2\log K}{\beta_\phi} \\
&\leq 1
\end{aligned}$$

where the first inequality follows from Lemma 32, and the second inequality is true due to $\beta_\phi = 4\beta_p \log K$. Thus, we have

$$|\hat{Q}_{f,h}^k(s,a)| \leq 2 + ||\hat{V}_{f,h+1}^k||_\infty \leq 2(H-h+1) = \beta_{Q,h}. \tag{16}$$

Moreover, note that

$$\begin{aligned}
\left\|\boldsymbol{\theta}_{f,h}^k + \hat{\boldsymbol{\mu}}_h^k \hat{V}_{f,h+1}^k\right\| &\leq \sqrt{d} + \left\|\hat{\boldsymbol{\mu}}_h^k \hat{V}_{f,h+1}^k\right\| \\
&= \sqrt{d} + \left\|(\mathbf{\Lambda}_h^k)^{-1} \sum_{\tau=1}^{k-1} \boldsymbol{\phi}(s_h^\tau, a_h^\tau)\hat{V}_{f,h+1}^k(s_{h+1}^\tau)\right\| \\
&\leq \sqrt{d} + \left\|(\mathbf{\Lambda}_h^k)^{-1}\right\| \left\|\sum_{\tau=1}^{k-1} \boldsymbol{\phi}(s_h^\tau, a_h^\tau)\hat{V}_{f,h+1}^k(s_{h+1}^\tau)\right\| \\
&\leq \sqrt{d} + K\beta_Q
\end{aligned} \tag{17}$$

where the first inequality is due to $||\boldsymbol{\theta}_{f,h}^k|| \leq \sqrt{d}$ followed from the linear CMDP assumptions, and the last inequality is true because $||(\mathbf{\Lambda}_h^k)^{-1}|| \leq 1$ and $||\sum_{\tau=1}^{k-1} \boldsymbol{\phi}(s_h^\tau, a_h^\tau)\hat{V}_{f,h+1}^k(s_{h+1}^\tau)|| \leq K\beta_Q$. Hence, by (16) and (17), together with $(2K)^{-1}\mathbf{I} \preceq (\mathbf{\Lambda}_h^{k(m)})^{-1} \preceq \mathbf{I}$, we have $\hat{Q}_{f,h}^k \in \mathcal{Q}(\beta_p, \sqrt{d} + K\beta_Q, \beta_{Q,h})$. Moreover, we have $\hat{V}_{f,h}^k \in \mathcal{V}(\beta_p, \sqrt{d} + K\beta_Q, \beta_{Q,h}, K^8)$, since $\pi_h^k \in \Pi(K\beta_p, K\sqrt{d} + K^2\beta_Q, K\beta_{Q,h}, K^8)$ and $\hat{V}_{f,h}^k$ is given by the inner product of π_h^k and $\hat{Q}_{f,h}^k$. For the case of g, by applying the same argument, we have $\hat{Q}_{g,h}^k \in \mathcal{Q}(\beta_p, \sqrt{d} + K\beta_Q, \beta_{Q,h})$ and $\hat{V}_{g,h}^k \in \mathcal{V}(\beta_p, \sqrt{d} + K\beta_Q, \beta_{Q,h}, K^8)$. This completes the induction. □

Lemma 19. *Let $1 \leq \beta_\phi \leq K$, and let $H \geq 2$. Then $\Pr(\mathcal{E}_g) \geq 1 - \delta$ for any $\delta \in (0,1)$.*

Proof. Suppose that $\bar{\mathcal{E}}_g = \bar{\mathcal{E}}_1 \cap \mathcal{E}_2$ holds, which occurs with probability at least $1-\delta$ by Lemma 17. To show that $\Pr(\mathcal{E}_g) \geq 1 - \delta$, it is sufficient to show that $\bar{\mathcal{E}}_g$ implies $\mathcal{E}_1$.

Before showing this implication, we first show that $\pi_h^k \in \Pi(K\beta_p, K\sqrt{d} + K^2\beta_Q, K\beta_{Q,h}, K^8)$, where $\beta_{Q,h} = 2(H-h+1)$ for all $(h,k) \in [H] \times [K]$. We prove this by induction on $k \in K(m)$ for each fixed $m \in E$. For the base case, we have $\pi_h^{k(m)} \in \Pi(K\beta_p, K\sqrt{d} + K^2\beta_Q, K\beta_{Q,h}, K^8)$ for all $h \in [H]$, since we set $\pi_h^{k(m)} = \pi_{\text{unif}}$ for all h at the beginning of each epoch m.

Then the induction can be written as follows. Given $k \in K(m)$, we assume that $\pi_h^{k'} \in \Pi(K\beta_p, K\sqrt{d} + K^2\beta_Q, K\beta_{Q,h}, K^8)$ for all $h \in [H]$ and k' such that $k(m) \leq k' < k$. To complete the induction, we show that $\pi_h^k \in \Pi(K\beta_p, K\sqrt{d} + K^2\beta_Q, K\beta_{Q,h}, K^8)$.

For all $k(m) \le k' < k$, we observe that

$$\begin{aligned}
&\hat{Q}^{k'}_{\ell,h}(\cdot,\cdot) \\
&= \bar{\boldsymbol{\phi}}^m_h(\cdot,\cdot)^\top \left[\boldsymbol{\theta}^{k'}_{\ell,h} + \hat{\boldsymbol{\mu}}^{k'}_h \hat{V}^{k'}_{\ell,h+1}\right] - \beta_p ||\bar{\boldsymbol{\phi}}^m_h(\cdot,\cdot)||_{(\boldsymbol{\Lambda}^{k(m)}_h)^{-1}} \\
&= \left(\boldsymbol{\phi}(\cdot,\cdot)^\top \mathbf{w}^{k'}_{\ell,h} - \beta_p ||\boldsymbol{\phi}(\cdot,\cdot)||_{(\boldsymbol{\Lambda}^{k(m)}_h)^{-1}}\right) \cdot \sigma\left(-\beta_\phi ||\boldsymbol{\phi}(\cdot,\cdot)||_{(\boldsymbol{\Lambda}^{k(m)}_h)^{-1}} + \log K\right)
\end{aligned}$$

where

$$\mathbf{w}^{k'}_{\ell,h} = \boldsymbol{\theta}^{k'}_{\ell,h} + \hat{\boldsymbol{\mu}}^{k'}_h \hat{V}^{k'}_{\ell,h+1}.$$

Moreover, since $\pi^{k'}_h \in \Pi(K\beta_p, K\sqrt{d} + K^2\beta_Q, K\beta_{Q,h}, K^8)$ by the induction hypothesis, then by Lemma 18, we have $\hat{Q}^{k'}_{f,h}, \hat{Q}^{k'}_{g,h} \in \mathcal{Q}(\beta_p, \sqrt{d} + K\beta_Q, \beta_{Q,h})$ and $\hat{V}^{k'}_{f,h}, \hat{V}^{k'}_{g,h} \in \mathcal{V}(\beta_p, \sqrt{d} + K\beta_Q, \beta_{Q,h}, K^8)$. This implies that for each $\ell \in \{f, g\}$,

$$||\mathbf{w}^{k'}_{\ell,h}|| \le \sqrt{d} + K\beta_Q, \quad ||\hat{Q}^{k'}_{\ell,h}(\cdot,\cdot)||_\infty \le \beta_{Q,h}. \tag{18}$$

Recall that

$$\begin{aligned}
&\pi^k_h(\cdot|s) \\
&\propto \exp\left(-\frac{1}{\alpha_{k-1}} \sum_{\tau=k(m)}^{k-1} \left(\hat{Q}^\tau_{f,h}(s,\cdot) + Y_\tau \hat{Q}^\tau_{g,h}(s,\cdot)\right)\right) \\
&\propto \exp\left(\hat{Q}(s,\cdot; \beta^k_h, \mathbf{w}^k_h, (\boldsymbol{\Lambda}^{k(m)}_h)^{-1})\right)
\end{aligned}$$

where

$$\beta^k_h = -\frac{1}{\alpha_{k-1}} \sum_{\tau=k(m)}^{k-1} (1 + Y_\tau)\beta_p, \quad \mathbf{w}^k_h = -\frac{1}{\alpha_{k-1}} \sum_{\tau=k(m)}^{k-1} (\mathbf{w}^\tau_{f,h} + Y_\tau \mathbf{w}^\tau_{g,h}).$$

To show that $\pi^k_h(\cdot|s) \propto \exp(\hat{Q}(s,\cdot; \beta^k_h, \mathbf{w}^k_h, (\boldsymbol{\Lambda}^{k(m)}_h)^{-1})) \in \Pi(K\beta_p, K\sqrt{d} + K^2\beta_Q, K\beta_{Q,h}, K^8)$, we observe that

$$|\beta^k_h| = \left|\sum_{\tau=k(m)}^{k-1} \frac{(1 + Y_\tau)\beta_p}{\alpha_{k-1}}\right| \le \left|\sum_{\tau=k(m)}^{k-1} (1 + Y_\tau)\beta_p\right| \le (1 + K^8)K\beta_p$$

where the first inequality follows from $\alpha_{k-1} = H/2\sqrt{\sum_{\tau=k(m)}^{k-1}(1 + Y_\tau)^2} \ge 1$, as $H \ge 2$ and $Y_\tau \ge 0$, and the second inequality is due to $Y_\tau \le K^8$. Moreover,

$$\begin{aligned}
||\mathbf{w}^k_h|| &= \left\|\sum_{\tau=k(m)}^{k-1} \frac{\mathbf{w}^\tau_{f,h} + Y_\tau \mathbf{w}^\tau_{g,h}}{\alpha_{k-1}}\right\| \\
&\le \left\|\sum_{\tau=k(m)}^{k-1} (\mathbf{w}^\tau_{f,h} + Y_\tau \mathbf{w}^\tau_{g,h})\right\| \\
&\le (1 + K^8)(K\sqrt{d} + K^2\beta_Q)
\end{aligned}$$

where the first inequality follows from $\alpha_{k-1} \ge 1$, and the second inequality follows from the fact that $Y_\tau \le K^8$ and (18). Again, by (18), we have

$$\begin{aligned}
\left\|\hat{Q}(\cdot,\cdot; \beta^k_h, \mathbf{w}^k_h, (\boldsymbol{\Lambda}^{k(m)}_h)^{-1})\right\|_\infty &= \left\|\frac{1}{\alpha_{k-1}} \sum_{\tau=k(m)}^{k-1} \left(\hat{Q}^\tau_{f,h}(\cdot,\cdot) + Y_\tau \hat{Q}^\tau_{g,h}(\cdot,\cdot)\right)\right\|_\infty \\
&\le K(1 + K^8)\beta_{Q,h}.
\end{aligned}$$

Finally, under the induction hypothesis, we show that $\pi_h^k(\cdot|s) \propto \exp(\hat{Q}(s,\cdot;\beta_h^k,\mathbf{w}_h^k,(\mathbf{\Lambda}_h^{k(m)})^{-1}))$, where $|\beta_h^k| \leq (1+K^8)K\beta_p$, $||\mathbf{w}_h^k|| \leq (1+K^8)(K\sqrt{d}+K^2\beta_Q)$, and $||\hat{Q}(\cdot,\cdot;\beta_h^k,\mathbf{w}_h^k,(\mathbf{\Lambda}_h^{k(m)})^{-1})||_\infty \leq K(1+K^8)\beta_{Q,h}$. Moreover, it is clear that $(2K)^{-1}\mathbf{I} \preceq (\mathbf{\Lambda}_h^{k(m)})^{-1} \preceq \mathbf{I}$. These imply that $\pi_h^k(\cdot|s) \propto \exp(\hat{Q}(s,\cdot;\beta_h^k,\mathbf{w}_h^k,(\mathbf{\Lambda}_h^{k(m)})^{-1})) \in \Pi(K\beta_p, K\sqrt{d}+K^2\beta_Q, K\beta_{Q,h}, K^8)$ as desired. This concludes the induction, showing that $\pi_h^k \in \Pi(K\beta_p, K\sqrt{d}+K^2\beta_Q, K\beta_{Q,h}, K^8)$.

We conclude the proof as follows. Since we have shown that $\pi_h^k(\cdot|s) \in \Pi(K\beta_p, K\sqrt{d}+K^2\beta_Q, K\beta_{Q,h}, K^8)$ for all $(h,k)\in[H]\times[K]$, by Lemma 18, we have $\hat{Q}_{f,h}^k, \hat{Q}_{g,h}^k \in \mathcal{Q}(\beta_p, \sqrt{d}+K\beta_Q, \beta_{Q,h})$ and $\hat{V}_{f,h}^k, \hat{V}_{g,h}^k \in \mathcal{V}(\beta_p, \sqrt{d}+K\beta_Q, \beta_{Q,h}, K^8)$ for all (h,k). These imply that

$$||\hat{Q}_{f,h}^k||_\infty, ||\hat{Q}_{g,h}^k||_\infty \leq \beta_Q.$$

Moreover, since $\bar{\mathcal{E}}_1$ is assumed, we have

$$||(\boldsymbol{\mu}_h - \hat{\boldsymbol{\mu}}_h^k)\hat{V}_{f,h+1}^k||_{\mathbf{\Lambda}_h^k},\ ||(\boldsymbol{\mu}_h - \hat{\boldsymbol{\mu}}_h^k)\hat{V}_{g,h+1}^k||_{\mathbf{\Lambda}_h^k} \leq \beta_p.$$

These imply that $\mathcal{E}_1$ holds. As a consequence, we showed that if $\bar{\mathcal{E}}_g$ holds, then $\mathcal{E}_1$ holds; thus $\mathcal{E}_g$ holds. Again, since $\Pr(\bar{\mathcal{E}}_g) \geq 1-\delta$, we have $\Pr(\mathcal{E}_g) \geq 1-\delta$. □

G Deferred Proofs for Section 5 and Theorem 4

G.1 Proof of Lemma 5

Lemma 20 (Restatement of Lemma 5)**.** *Assume that the good event $\mathcal{E}_g$ holds. Let $H \geq 2$, and let $K \geq c_{bias} dH(\log|\mathcal{A}|+20)\log(2H/\delta)$, where c_{bias} is the absolute constant given by Lemma 23. We have*

$$\sum_{k=1}^K R_k \leq 12\sqrt{dH^5K}(\log|\mathcal{A}|+20)\log(3dH(\log|\mathcal{A}|+20)K).$$

Proof. Note that

$$\begin{aligned}
\sum_{k=1}^K R_k &= \sum_{k=1}^K \left(108\gamma H^2 + \beta_R \frac{Y_k}{\sqrt{\sum_{\tau=1}^k Y_\tau^2}}\right) \\
&= 108K\gamma H^2 + \beta_R \sum_{k=1}^K \frac{Y_k}{\sqrt{\sum_{\tau=1}^k Y_\tau^2}} \\
&\leq 3H\sqrt{K} + \beta_R \sum_{k=1}^K \frac{Y_k}{\sqrt{\sum_{\tau=1}^k Y_\tau^2}}
\end{aligned}$$

where the inequality follows from the fact that $108K\gamma H^2 \leq 3H\sqrt{K}$, as $\gamma \leq 1/(36H\sqrt{K})$. Moreover, we have

$$\begin{aligned}
\sum_{k=1}^K \frac{Y_k}{\sqrt{\sum_{\tau=1}^k Y_\tau^2}} &\leq \sqrt{K}\sqrt{\sum_{k=1}^K \frac{Y_k^2}{\sum_{\tau=1}^k Y_\tau^2}} \\
&\leq \sqrt{K}\sqrt{1+\log\left(\frac{1}{\gamma^2}\sum_{k=1}^K Y_k^2\right)} \\
&\leq \sqrt{K}\sqrt{1+\log\frac{K^{17}}{\gamma^2}}
\end{aligned}$$

$$\leq \sqrt{K}\sqrt{1+\log(144dH^5(\log|\mathcal{A}|+20)^2K^{18})}$$
$$\leq \sqrt{18K\log(3dH(\log|\mathcal{A}|+20)K)}$$

where the first inequality follows from the Cauchy-Schwarz inequality, the second inequality follows from the fact that $\sum_{k=1}^{K} y_k/(\sum_{\tau=1}^{k} y_\tau) \leq 1+\log(\sum_{k=1}^{K} y_k/y_1)$ for any positive sequence $\{y_k\}_{k\geq 1}$ and that $Y_1 = \gamma$, and the third inequality follows from $Y_k \leq K^8$, and the fourth inequality is due to

$$\frac{1}{\gamma} = \max\{36H\sqrt{K}, 12H^2(\log|\mathcal{A}|+20)\sqrt{dH\log(2K)}\}$$
$$\leq 12H^2(\log|\mathcal{A}|+20)\sqrt{dHK}.$$

Finally, we deduce that

$$\sum_{k=1}^{K} R_k \leq 3H\sqrt{K} + 2H^2(\log|\mathcal{A}|+20)\sqrt{dH\log(2K)}\sqrt{18K\log(3dH(\log|\mathcal{A}|+20)K)}$$
$$\leq 12\sqrt{dH^5}(\log|\mathcal{A}|+20)\sqrt{K}\log(3dH(\log|\mathcal{A}|+20)K).$$

This completes the proof. □

G.2 Proofs of Lemma 6 and Eq. (7)

In this subsection, we first prove Lemma 6, which establishes the regret bound for adaptive FTRL over the policy space. We then prove the lemmas that bound the bias and optimism terms, following standard arguments in the online MDP literature. Finally, by combining these results with the value difference lemma (Lemma 10), we establish (7).

Lemma 21 (Restatement of Lemma 6)**.** *Assume that the good event $\mathcal{E}_g$ holds. Then we have for any $k \in [K]$,*

$$\sum_{m\in E(k)}\sum_{\tau\in K(m),\tau\leq k} \mathbb{E}_{\bar{\mathbb{P}}^m,\pi^\star}\left[\sum_{h=1}^{H}\left(\hat{Q}^\tau_{f,h}(s^\tau_h,\cdot)+Y_\tau\hat{Q}^\tau_{g,h}(s^\tau_h,\cdot)\right)^\top(\pi^\tau_h(\cdot|s^\tau_h)-\pi^\star_h(\cdot|s^\tau_h))\right]$$
$$\leq H^2(\log|\mathcal{A}|+20)\sqrt{dH\log(2K)}\sqrt{k+\sum_{\tau=1}^{k}Y_\tau^2}$$

Proof. Recall that

$$\pi^{k+1}_h(\cdot|s) \in \operatorname*{arg\,min}_{\pi(\cdot|s)\in\Delta(\mathcal{A})} \psi_k(\pi(\cdot|s)) + \sum_{\tau=k(m)}^{k}\left(\hat{Q}^\tau_{f,h}(s,\cdot)+Y_\tau\hat{Q}^\tau_{g,h}(s,\cdot)\right)^\top\pi(\cdot|s)$$

where $\psi_k(\pi(\cdot|s)) = \alpha_k\sum_{a\in\mathcal{A}}\pi(a|s)\log\pi(a|s)$ and $\alpha_k = (\beta_Q/4)\sqrt{\sum_{\tau=k(m)}^{k}(1+Y_\tau)^2}$. Note that under the good event $\mathcal{E}_g$, we have $||\hat{Q}^k_{f,h}(s,\cdot)||_\infty, ||\hat{Q}^k_{g,h}(s,\cdot)||_\infty \leq \beta_Q = 2H$. Now we apply Lemma 11. In particular, we have $n \leftarrow |\mathcal{A}|$, $\nabla_k \leftarrow \hat{Q}^k_{f,h}(s,\cdot)+Y_k\hat{Q}^k_{g,h}(s,\cdot)$, $\mathbf{x}_1 \leftarrow \pi_h^{k(m)}(\cdot|s)$, $\mathbf{x}_k \leftarrow \pi^k_h(\cdot|s)$, $L_k \leftarrow \beta_Q(1+Y_k)$ $\mathbf{x}^\star \leftarrow \pi^\star_h(\cdot|s)$. Moreover, note that $L_k > 0$ for all $k \in [K]$, as $\beta_Q(1+Y_k) > 0$. Then by Lemma 11, we have

$$\sum_{\tau\in k(m),\tau\leq k}\left(\hat{Q}^\tau_{f,h}(s,\cdot)+Y_\tau\hat{Q}^\tau_{g,h}(s,\cdot)\right)^\top(\pi^\tau_h(\cdot|s)-\pi^\star_h(\cdot|s))$$
$$\leq \left(\frac{\log|\mathcal{A}|}{4}+4\sqrt{2}\right)\beta_Q\sqrt{\sum_{\tau\in k(m),\tau\leq k}(1+Y_\tau)^2}.$$

By summing over $h = 1,\ldots,H$ and taking $\mathbb{E}_{\bar{\mathbb{P}}^m,\pi^\star}$ on both sides, we have

$$\sum_{\tau\in k(m),\tau\leq k}\mathbb{E}_{\bar{\mathbb{P}}^m,\pi^\star}\left[\sum_{h=1}^{H}\left(\hat{Q}^\tau_{f,h}(s^\tau_h,\cdot)+Y_\tau\hat{Q}^\tau_{g,h}(s^\tau_h,\cdot)\right)^\top(\pi^\tau_h(\cdot|s^\tau_h)-\pi^\star_h(\cdot|s^\tau_h))\right]$$

$$\leq H\left(\frac{\log|\mathcal{A}|}{4}+4\sqrt{2}\right)\beta_Q\sqrt{\sum_{\tau\in k(m),\tau\leq k}(1+Y_\tau)^2}.$$

Next, we take the summation over $m \in E(k)$. Then it follows that

$$\begin{aligned}
&\sum_{m\in E(k)}\sum_{\tau\in k(m),\tau\leq k}\mathbb{E}_{\bar{\mathbb{P}}^m,\pi^\star}\left[\sum_{h=1}^{H}\left(\hat{Q}^\tau_{f,h}(s^\tau_h,\cdot)+Y_\tau\hat{Q}^\tau_{g,h}(s^\tau_h,\cdot)\right)^\top(\pi^\tau_h(\cdot|s^\tau_h)-\pi^\star_h(\cdot|s^\tau_h))\right]\\
&\leq H\left(\frac{\log|\mathcal{A}|}{4}+4\sqrt{2}\right)\beta_Q\sum_{m\in E(k)}\sqrt{\sum_{\tau\in k(m),\tau\leq k}(1+Y_\tau)^2}\\
&\leq H\left(\frac{\log|\mathcal{A}|}{4}+4\sqrt{2}\right)\beta_Q\sqrt{|E|}\sqrt{\sum_{\tau=1}^{k}(1+Y_\tau)^2}\\
&\leq H\left(\frac{\sqrt{2}\log|\mathcal{A}|}{4}+8\right)\beta_Q\sqrt{|E|}\sqrt{k+\sum_{\tau=1}^{k}Y_\tau^2}
\end{aligned}$$

where the second inequality follows from the Cauchy-Schwarz inequality and $|E(k)| \leq |E|$, and the last inequality follows from $(a+b)^2 \leq 2(a^2+b^2)$. Note that $|E| \leq (3/2)dH\log(2K)$ by Lemma 30 and $\beta_Q = 2H$. It leads to

$$\begin{aligned}
&\sum_{m\in E(k)}\sum_{\tau\in k(m),\tau\leq k}\mathbb{E}_{\bar{\mathbb{P}}^m,\pi^\star}\left[\sum_{h=1}^{H}\left(\hat{Q}^\tau_{f,h}(s^\tau_h,\cdot)+Y_\tau\hat{Q}^\tau_{g,h}(s^\tau_h,\cdot)\right)^\top(\pi^\tau_h(\cdot|s^\tau_h)-\pi^\star_h(\cdot|s^\tau_h))\right]\\
&\leq H^2\left(\frac{\sqrt{2}\log|\mathcal{A}|}{2}+16\right)\sqrt{(3/2)dH\log(2K)}\sqrt{k+\sum_{\tau=1}^{k}Y_\tau^2}\\
&\leq H^2\left(\log|\mathcal{A}|+20\right)\sqrt{dH\log(2K)}\sqrt{k+\sum_{\tau=1}^{k}Y_\tau^2}
\end{aligned}$$

where the last inequality follows from $\beta_Q = 2H$, $(\sqrt{2}/2)\sqrt{3/2} \leq 1$, and $16\sqrt{3/2} \leq 20$. □

Lemma 22. *Assume that the good event $\mathcal{E}_g$ holds. For all $(s,a,h,k,\ell) \in \mathcal{S}\times\mathcal{A}\times[H]\times[K]\times\{f,g\}$, we have*

$$\hat{Q}^k_{\ell,h}(s,a) - \bar{\boldsymbol{\phi}}^m_h(s,a)^\top(\boldsymbol{\theta}^k_{\ell,h}+\boldsymbol{\mu}_h\hat{V}^k_{\ell,h+1}) \leq 0$$

Proof. For any $(s,a,h,k) \in \mathcal{S}\times\mathcal{A}\times[H]\times[K]$, we have

$$\begin{aligned}
&\hat{Q}^k_{f,h}(s,a) - \bar{\boldsymbol{\phi}}^m_h(s,a)^\top(\boldsymbol{\theta}^k_{f,h}+\boldsymbol{\mu}_h\hat{V}^k_{f,h+1})\\
&= \bar{\boldsymbol{\phi}}^m_h(s,a)^\top\left((\hat{\boldsymbol{\mu}}^k_h-\boldsymbol{\mu}_h)\hat{V}^k_{f,h+1}\right) - \beta_p||\bar{\boldsymbol{\phi}}^m_h(s,a)||_{(\boldsymbol{\Lambda}^{k(m)}_h)^{-1}}\\
&\leq ||\bar{\boldsymbol{\phi}}^m_h(s,a)||_{(\boldsymbol{\Lambda}^k_h)^{-1}}\left\|(\hat{\boldsymbol{\mu}}^k_h-\boldsymbol{\mu}_h)\hat{V}_{f,h+1}\right\|_{\boldsymbol{\Lambda}^k_h} - \beta_p||\bar{\boldsymbol{\phi}}^m_h(s,a)||_{(\boldsymbol{\Lambda}^{k(m)}_h)^{-1}}\\
&\leq \beta_p||\bar{\boldsymbol{\phi}}^m_h(s,a)||_{(\boldsymbol{\Lambda}^k_h)^{-1}} - \beta_p||\bar{\boldsymbol{\phi}}^m_h(s,a)||_{(\boldsymbol{\Lambda}^{k(m)}_h)^{-1}}\\
&\leq 0
\end{aligned}$$

where the first inequality follows from the Cauchy-Schwarz inequality, the second inequality follows from $\mathcal{E}_g$, and the last inequality follows from the fact that $\boldsymbol{\Lambda}^{k(m)}_h \preceq \boldsymbol{\Lambda}^k_h$. Moreover, the same argument applies to the case of g. □

Lemma 23. *Assume that the good event $\mathcal{E}_g$ holds. For all $(k,\ell) \in [K]\times\{f,g\}$, there exists an absolute constant c_{bias} such that*

$$\sum_{\tau=1}^{k}(V^{\pi^\tau}_{\ell^\tau,1}(s_1) - \hat{V}^\tau_{\ell,1}(s_1)) \leq c_{bias}\left(d^{3/2}H^2\sqrt{K}\log(2HK/\delta) + d^3H^4\log^4(2HK/\delta)\right)$$

Proof. Consider the case of f. By Lemma 10, we have

$$\begin{aligned}
&\sum_{\tau=1}^{k}(V^{\pi^\tau}_{f^\tau,1}(s_1)-\hat{V}^\tau_{f,1}(s_1))\\
&=\sum_{\tau=1}^{k}\mathbb{E}_{\mathbb{P},\pi^\tau}\left[\sum_{h=1}^{H}\left(f^\tau_h(s_h,a_h)+\mathbb{P}_h\hat{V}^\tau_{f,h+1}(s_h,a_h)-\hat{Q}^\tau_{f,h}(s_h,a_h)\right)\mid s_1\right] \qquad (19)\\
&=\sum_{\tau=1}^{k}\mathbb{E}_{\mathbb{P},\pi^\tau}\left[\sum_{h=1}^{H}\left(\boldsymbol{\phi}(s_h,a_h)^\top(\boldsymbol{\theta}^\tau_{f,h}+\boldsymbol{\mu}_h\hat{V}^\tau_{f,h+1})-\hat{Q}^\tau_{f,h}(s_h,a_h)\right)\mid s_1\right].
\end{aligned}$$

Note that

$$\begin{aligned}
&\boldsymbol{\phi}(s,a)^\top(\boldsymbol{\theta}^\tau_{f,h}+\boldsymbol{\mu}_h\hat{V}^\tau_{f,h+1})-\hat{Q}^\tau_{f,h}(s,a)\\
&=\boldsymbol{\phi}(s,a)^\top(\boldsymbol{\theta}^\tau_{f,h}+\boldsymbol{\mu}_h\hat{V}^\tau_{f,h+1})-\bar{\boldsymbol{\phi}}^m_h(s,a)^\top(\boldsymbol{\theta}^\tau_{f,h}+\hat{\boldsymbol{\mu}}^\tau_h\hat{V}^\tau_{f,h+1})+\beta_p||\bar{\boldsymbol{\phi}}^m_h(s,a)||_{(\boldsymbol{\Lambda}_h^{k(m)})^{-1}}\\
&=\boldsymbol{\phi}(s,a)^\top(\boldsymbol{\theta}^\tau_{f,h}+\boldsymbol{\mu}_h\hat{V}^\tau_{f,h+1})-\bar{\boldsymbol{\phi}}^m_h(s,a)^\top(\boldsymbol{\theta}^\tau_{f,h}+\boldsymbol{\mu}_h\hat{V}^\tau_{f,h+1})\\
&\quad+\bar{\boldsymbol{\phi}}^m_h(s,a)^\top\left((\boldsymbol{\mu}_h-\hat{\boldsymbol{\mu}}^\tau_h)\hat{V}^\tau_{f,h+1}\right)+\beta_p||\bar{\boldsymbol{\phi}}^m_h(s,a)||_{(\boldsymbol{\Lambda}_h^{k(m)})^{-1}}\\
&\le\boldsymbol{\phi}(s,a)^\top(\boldsymbol{\theta}^\tau_{f,h}+\boldsymbol{\mu}_h\hat{V}^\tau_{f,h+1})-\bar{\boldsymbol{\phi}}^m_h(s,a)^\top(\boldsymbol{\theta}^\tau_{f,h}+\boldsymbol{\mu}_h\hat{V}^\tau_{f,h+1})\\
&\quad+||\bar{\boldsymbol{\phi}}^m_h(s,a)||_{(\boldsymbol{\Lambda}^\tau_h)^{-1}}\left\|(\boldsymbol{\mu}_h-\hat{\boldsymbol{\mu}}^\tau_h)\hat{V}^\tau_{f,h+1}\right\|_{\boldsymbol{\Lambda}^\tau_h}+\beta_p||\bar{\boldsymbol{\phi}}^m_h(s,a)||_{(\boldsymbol{\Lambda}_h^{k(m)})^{-1}}\\
&\le\left(\boldsymbol{\phi}(s,a)-\bar{\boldsymbol{\phi}}^m_h(s,a)\right)^\top(\boldsymbol{\theta}^\tau_{f,h}+\boldsymbol{\mu}_h\hat{V}^\tau_{f,h+1})+\beta_p||\boldsymbol{\phi}(s,a)||_{(\boldsymbol{\Lambda}^\tau_h)^{-1}}+\beta_p||\boldsymbol{\phi}(s,a)||_{(\boldsymbol{\Lambda}_h^{k(m)})^{-1}} \qquad (20)
\end{aligned}$$

where the last inequality follows from $\mathcal{E}_g$ and that $\sigma(z)\le 1$ for all $z\in\mathbb{R}$. By Lemma 34, we have

$$\begin{aligned}
&\left(\boldsymbol{\phi}(s,a)-\bar{\boldsymbol{\phi}}^m_h(s,a)\right)^\top(\boldsymbol{\theta}^\tau_{f,h}+\boldsymbol{\mu}_h\hat{V}^\tau_{f,h+1})\\
&\le\left(4\beta_\phi^2||\boldsymbol{\phi}(s,a)||^2_{(\boldsymbol{\Lambda}^\tau_h)^{-1}}+2K^{-1}\right)|\boldsymbol{\phi}(s,a)^\top(\boldsymbol{\theta}^\tau_{f,h}+\boldsymbol{\mu}_h\hat{V}^\tau_{f,h+1})|\\
&=\left(4\beta_\phi^2||\boldsymbol{\phi}(s,a)||^2_{(\boldsymbol{\Lambda}^\tau_h)^{-1}}+2K^{-1}\right)|f^\tau_h(s,a)+\mathbb{P}_h\hat{V}^\tau_{f,h+1}(s,a)| \qquad (21)\\
&\le 16H\beta_\phi^2||\boldsymbol{\phi}(s,a)||^2_{(\boldsymbol{\Lambda}^\tau_h)^{-1}}+8HK^{-1}
\end{aligned}$$

where the second inequality follows from $|f^\tau_h(s,a)+\mathbb{P}_h\hat{V}^\tau_{f,h+1}(s,a)|\le 4H$ under $\mathcal{E}_g$. Moreover, we have

$$\begin{aligned}
\beta_p||\boldsymbol{\phi}(s,a)||_{(\boldsymbol{\Lambda}_h^{k(m)})^{-1}}&\le\beta_p\sqrt{\frac{\det(\boldsymbol{\Lambda}^\tau_h)}{\det(\boldsymbol{\Lambda}_h^{k(m)})}}||\boldsymbol{\phi}(s,a)||_{(\boldsymbol{\Lambda}^\tau_h)^{-1}} \qquad (22)\\
&\le 2\beta_p||\boldsymbol{\phi}(s,a)||_{(\boldsymbol{\Lambda}^\tau_h)^{-1}}
\end{aligned}$$

where the first inequality follows from Lemma 35, and the second inequality follows from $\det(\boldsymbol{\Lambda}^\tau_h)\le 2\det(\boldsymbol{\Lambda}_h^{k(m)})$. Applying (21) and (22) to (20) leads to

$$\begin{aligned}
&\boldsymbol{\phi}(s,a)^\top(\boldsymbol{\theta}^\tau_{f,h}+\boldsymbol{\mu}_h\hat{V}^\tau_{f,h+1})-\hat{Q}^\tau_{f,h}(s,a)\\
&\le 16H\beta_\phi^2||\boldsymbol{\phi}(s,a)||^2_{(\boldsymbol{\Lambda}^\tau_h)^{-1}}+8HK^{-1}+3\beta_p||\boldsymbol{\phi}(s,a)||_{(\boldsymbol{\Lambda}^\tau_h)^{-1}}.
\end{aligned}$$

By summing over h, taking $\mathbb{E}_{\mathbb{P},\pi^\tau}$, and summing over τ, we have

$$\begin{aligned}
&\sum_{\tau=1}^{k}\mathbb{E}_{\mathbb{P},\pi^\tau}\left[\sum_{h=1}^{H}\left(\boldsymbol{\phi}(s_h,a_h)^\top(\boldsymbol{\theta}^\tau_{f,h}+\boldsymbol{\mu}_h\hat{V}^\tau_{f,h+1})-\hat{Q}^\tau_{f,h}(s_h,a_h)\right)\mid s_1\right]\\
&\le\sum_{\tau=1}^{k}\mathbb{E}_{\mathbb{P},\pi^\tau}\left[\sum_{h=1}^{H}\left(16H\beta_\phi^2||\boldsymbol{\phi}(s_h,a_h)||^2_{(\boldsymbol{\Lambda}^\tau_h)^{-1}}+8HK^{-1}+3\beta_p||\boldsymbol{\phi}(s_h,a_h)||_{(\boldsymbol{\Lambda}^\tau_h)^{-1}}\right)\mid s_1\right]\\
&\le 2\sum_{\tau=1}^{k}\sum_{h=1}^{H}\left(3\beta_p||\boldsymbol{\phi}(s^\tau_h,a^\tau_h)||_{(\boldsymbol{\Lambda}^\tau_h)^{-1}}+8\beta_Q\beta_\phi^2||\boldsymbol{\phi}(s^\tau_h,a^\tau_h)||^2_{(\boldsymbol{\Lambda}^\tau_h)^{-1}}\right)
\end{aligned}$$

$$+ 4H(3\beta_p + 8\beta_Q\beta_\phi^2)\log\frac{4K^2}{\delta} + 8H^2$$

where the second inequality follows from $\mathcal{E}_g$. Applying this to (19) yields

$$\begin{aligned}
&\sum_{\tau=1}^{k}(V_{f^\tau,1}^{\pi^\tau}(s_1) - \hat{V}_{f,1}^{\tau}(s_1)) \\
&\leq 2\sum_{\tau=1}^{k}\sum_{h=1}^{H}\left(3\beta_p||\boldsymbol{\phi}(s_h^\tau,a_h^\tau)||_{(\boldsymbol{\Lambda}_h^\tau)^{-1}} + 8\beta_Q\beta_\phi^2||\boldsymbol{\phi}(s_h^\tau,a_h^\tau)||^2_{(\boldsymbol{\Lambda}_h^\tau)^{-1}}\right) \\
&\quad + 4H(3\beta_p + 8\beta_Q\beta_\phi^2)\log\frac{4K^2}{\delta} + 8H^2.
\end{aligned} \tag{23}$$

Note that for each $h \in [H]$,

$$\sum_{\tau=1}^{k}||\boldsymbol{\phi}(s_h^\tau,a_h^\tau)||^2_{(\boldsymbol{\Lambda}_h^\tau)^{-1}} \leq 2\log\det(\boldsymbol{\Lambda}_h^{k+1}) \leq 2d\log(k+1)$$

where the first inequality is due to Lemma 36. This implies that

$$\begin{aligned}
\sum_{\tau=1}^{k}\sum_{h=1}^{H}||\boldsymbol{\phi}(s_h^\tau,a_h^\tau)||_{(\boldsymbol{\Lambda}_h^\tau)^{-1}} &\leq \sqrt{kH}\sqrt{\sum_{\tau=1}^{k}\sum_{h=1}^{H}||\boldsymbol{\phi}(s_h^\tau,a_h^\tau)||^2_{(\boldsymbol{\Lambda}_h^\tau)^{-1}}} \\
&\leq \sqrt{2dH^2k\log(k+1)}
\end{aligned}$$

where the first inequality follows from the Cauchy-Schwarz inequality. Note that $k \leq K$. Applying this to (23) leads to the following. For all $k \in [K]$,

$$\sum_{\tau=1}^{k}(V_{f^\tau,1}^{\pi^\tau}(s_1) - \hat{V}_{f,1}^{\tau}(s_1)) \leq c_{\text{bias}}\left(d^{3/2}H^2\sqrt{K}\log(2HK/\delta) + d^3H^4\log^4(2HK/\delta)\right)$$

where c_{bias} is an absolute constant. Note that the same argument applies to the case of g. This completes the proof. □

Lemma 24 (Formal statement of (7))**.** *Assume that the good event $\mathcal{E}_g$ holds. For all $k \in [K]$, we have*

$$\begin{aligned}
&\sum_{\tau=1}^{k}\left(V_{f^\tau,1}^{\pi^\tau}(s_1) - V_{f^\tau,1}^{\pi^\star}(s_1)\right) + \sum_{\tau=1}^{k}Y_\tau\left(\hat{V}_{g,1}^{\tau}(s_1) - b\right) \\
&\leq H^2\left(\log|\mathcal{A}| + 20\right)\sqrt{dH\log(2K)}\sqrt{k + \sum_{\tau=1}^{k}Y_\tau^2} \\
&\quad + c_{bias}d^{3/2}H^2\sqrt{K}\log(2HK/\delta) + c_{bias}d^3H^4\log^4(2HK/\delta)
\end{aligned}$$

where c_{bias} is the absolute constant given by Lemma 23.

Proof. For a fixed $k \in [K]$, we first decompose the desired term. Note that

$$\begin{aligned}
&\sum_{\tau=1}^{k}\left(V_{f^\tau,1}^{\pi^\tau}(s_1) - V_{f^\tau,1}^{\pi^\star}(s_1)\right) + \sum_{\tau=1}^{k}Y_\tau\left(\hat{V}_{g,1}^{\tau}(s_1) - b\right) \\
&= \sum_{\tau=1}^{k}\left(\hat{V}_{f,1}^{\tau}(s_1) - V_{f^\tau,1}^{\pi^\star}(s_1)\right) + \sum_{\tau=1}^{k}\left(V_{f^\tau,1}^{\pi^\tau}(s_1) - \hat{V}_{f,1}^{\tau}(s_1)\right) + \sum_{\tau=1}^{k}Y_\tau\left(\hat{V}_{g,1}^{\tau}(s_1) - b\right) \\
&\leq \sum_{\tau=1}^{k}\left(\hat{V}_{f,1}^{\tau}(s_1) - V_{f^\tau,1}^{\pi^\star}(s_1)\right) + \sum_{\tau=1}^{k}Y_\tau\left(\hat{V}_{g,1}^{\tau}(s_1) - V_{g^\tau,1}^{\pi^\star}(s_1)\right)
\end{aligned}$$

$$+\sum_{\tau=1}^{k}\left(V_{f^\tau,1}^{\pi^\tau}(s_1)-\hat{V}_{f,1}^{\tau}(s_1)\right)$$

where the inequality follows from the fact that $V_{g^\tau,1}^{\pi^\star}(s_1)\le b$ and that Y_τ is nonnegative for all $\tau\ge 1$. To bound the first two sums, recall that $\hat{V}_{f,h}^{\tau}(s)+Y_\tau\hat{V}_{g,h}^{\tau}(s)=\sum_{a\in\mathcal{A}}\pi_h^\tau(a|s)(\hat{Q}_{f,h}^{\tau}(s,a)+Y_\tau\hat{Q}_{g,h}^{\tau}(s,a))$ for all $(s,h)\in\mathcal{S}\times[H]$. Let $E(k)=\{m\in E: k(m)\le k\}$, i.e., the set of epochs up to and including the epoch containing episode k. Then we have

$$\begin{aligned}
&\sum_{\tau=1}^{k}\left(\hat{V}_{f,1}^{\tau}(s_1)-V_{f^\tau,1}^{\pi^\star}(s_1)\right)+\sum_{\tau=1}^{k}Y_\tau\left(\hat{V}_{g,1}^{\tau}(s_1)-V_{g^\tau,1}^{\pi^\star}(s_1)\right)\\
&\le\sum_{\tau=1}^{k}\left(\hat{V}_{f,1}^{\tau}(s_1)-\bar{V}_{f^\tau,1}^{\pi^\star}(s_1)\right)+\sum_{\tau=1}^{k}Y_\tau\left(\hat{V}_{g,1}^{\tau}(s_1)-\bar{V}_{g^\tau,1}^{\pi^\star}(s_1)\right)\\
&=\sum_{m\in E(k)}\sum_{\tau\in k(m),\tau\le k}\mathbb{E}_{\bar{\mathbb{P}}^m,\pi^\star}\left[\sum_{h=1}^{H}\left(\hat{Q}_{f,h}^{\tau}(s_h^\tau,\cdot)+Y_\tau\hat{Q}_{g,h}^{\tau}(s_h^\tau,\cdot)\right)^\top(\pi_h^\tau(\cdot|s_h^\tau)-\pi_h^\star(\cdot|s_h^\tau))\right]\\
&\quad+\sum_{m\in E(k)}\sum_{\tau\in k(m),\tau\le k}\mathbb{E}_{\bar{\mathbb{P}}^m,\pi^\star}\left[\sum_{h=1}^{H}\left(\hat{Q}_{f,h}^{\tau}(s_h^\tau,a_h^\tau)-\bar{f}_h^\tau(s_h^\tau,a_h^\tau)-\sum_{s'\in\mathcal{S}}\bar{\mathbb{P}}_h^m(s'|s_h^\tau,a_h^\tau)\hat{V}_{f,h+1}^{\tau}(s')\right)\right]\\
&\quad+\sum_{m\in E(k)}\sum_{\tau\in k(m),\tau\le k}\mathbb{E}_{\bar{\mathbb{P}}^m,\pi^\star}\left[\sum_{h=1}^{H}Y_\tau\left(\hat{Q}_{g,h}^{\tau}(s_h^\tau,a_h^\tau)-\bar{g}_h(s_h^\tau,a_h^\tau)-\sum_{s'\in\mathcal{S}}\bar{\mathbb{P}}_h^m(s'|s_h^\tau,a_h^\tau)\hat{V}_{g,h+1}^{\tau}(s')\right)\right]
\end{aligned}$$

where the inequality follows from Lemma 9, and the equality is due to Lemma 10. Moreover, by the definition of contracted MDPs, we have for $\ell=f,g$,

$$\begin{aligned}
&\hat{Q}_{\ell,h}^{\tau}(s_h^\tau,a_h^\tau)-\bar{\ell}_h^\tau(s_h^\tau,a_h^\tau)-\sum_{s'\in\mathcal{S}}\bar{\mathbb{P}}_h^m(s'|s_h^\tau,a_h^\tau)\hat{V}_{\ell,h+1}^{\tau}(s')\\
&=\hat{Q}_{\ell,h}^{\tau}(s,a)-\bar{\boldsymbol{\phi}}_h^m(s,a)^\top(\boldsymbol{\theta}_{\ell,h}^\tau+\boldsymbol{\mu}_h\hat{V}_{\ell,h+1}^{\tau}).
\end{aligned}$$

This leads to

$$\begin{aligned}
&\sum_{\tau=1}^{k}\left(V_{f^\tau,1}^{\pi^\tau}(s_1)-V_{f^\tau,1}^{\pi^\star}(s_1)\right)+\sum_{\tau=1}^{k}Y_\tau\left(\hat{V}_{g,1}^{\tau}(s_1)-b\right)\\
&\le\underbrace{\sum_{m\in E(k)}\sum_{\tau\in k(m),\tau\le k}\mathbb{E}_{\bar{\mathbb{P}}^m,\pi^\star}\left[\sum_{h=1}^{H}\left(\hat{Q}_{f,h}^{\tau}(s_h^\tau,\cdot)+Y_\tau\hat{Q}_{g,h}^{\tau}(s_h^\tau,\cdot)\right)^\top(\pi_h^\tau(\cdot|s_h^\tau)-\pi_h^\star(\cdot|s_h^\tau))\right]}_{\text{(I)}}\\
&\quad+\underbrace{\sum_{m\in E(k)}\sum_{\tau\in k(m),\tau\le k}\mathbb{E}_{\bar{\mathbb{P}}^m,\pi^\star}\left[\sum_{h=1}^{H}\left(\hat{Q}_{f,h}^{\tau}(s_h^\tau,a_h^\tau)-\bar{\boldsymbol{\phi}}_h^m(s_h^\tau,a_h^\tau)^\top(\boldsymbol{\theta}_{f,h}^\tau+\boldsymbol{\mu}_h\hat{V}_{f,h+1}^{\tau})\right)\right]}_{\text{(II)}}\\
&\quad+\underbrace{\sum_{m\in E(k)}\sum_{\tau\in k(m),\tau\le k}\mathbb{E}_{\bar{\mathbb{P}}^m,\pi^\star}\left[\sum_{h=1}^{H}Y_\tau\left(\hat{Q}_{g,h}^{\tau}(s_h^\tau,a_h^\tau)-\bar{\boldsymbol{\phi}}_h^m(s_h^\tau,a_h^\tau)^\top(\boldsymbol{\theta}_{g,h}^\tau+\boldsymbol{\mu}_h\hat{V}_{g,h+1}^{\tau})\right)\right]}_{\text{(III)}}\\
&\quad+\underbrace{\sum_{\tau=1}^{k}\left(V_{f^\tau,1}^{\pi^\tau}(s_1)-\hat{V}_{f,1}^{\tau}(s_1)\right)}_{\text{(IV)}}.
\end{aligned}$$

By Lemma 21, we have

$$\text{(I)}\le H^2\left(\log|\mathcal{A}|+20\right)\sqrt{dH\log(2K)}\sqrt{k+\sum_{\tau=1}^{k}Y_\tau^2}.$$

By Lemma 22, we have

$$\text{(II)}, \text{(III)} \leq 0.$$

By Lemma 23, we have

$$\text{(IV)} \leq c_{\text{bias}} \left(d^{3/2} H^2 \sqrt{K} \log(2HK/\delta) + d^3 H^4 \log^4(2HK/\delta) \right)$$

where c_{bias} is the absolute constant given by Lemma 23. Combining these yields

$$\begin{aligned}
&\sum_{\tau=1}^{k} \left(V^{\pi^\tau}_{f^\tau,1}(s_1) - V^{\pi^\star}_{f^\tau,1}(s_1) \right) + \sum_{\tau=1}^{k} Y_\tau \left(\hat{V}^{\tau}_{g,1}(s_1) - b \right) \\
&\leq H^2 \left(\log |\mathcal{A}| + 20 \right) \sqrt{dH \log(2K)} \sqrt{k + \sum_{\tau=1}^{k} Y_\tau^2} \\
&\quad + c_{\text{bias}} d^{3/2} H^2 \sqrt{K} \log(2HK/\delta) + c_{\text{bias}} d^3 H^4 \log^4(2HK/\delta).
\end{aligned}$$

□

G.3 Proofs of Lemmas 7 and 8

Lemma 25 (Restatement of Lemma 7)**.** *Assume that the good event $\mathcal{E}_g$ holds. Then we have*

$$\Phi(Z_{k+1}) - \Phi(Z_k) \leq \Phi'(Z_k) \left(\hat{V}^{k}_{g,1}(s_1) - b - \frac{1}{2} R_k \right).$$

Proof. We closely follow the proof of Lemma 7 in Yu et al. (2026d). Note that

$$\begin{aligned}
\Phi(Z_{k+1}) - \Phi(Z_k) &= e^{\gamma Z_{k+1}} - e^{\gamma Z_k} \\
&= e^{\gamma Z_k} \left(e^{\gamma (Z_{k+1} - Z_k)} - 1 \right) \\
&= e^{\gamma Z_k} \left(e^{\gamma (\hat{V}^{k}_{g,1}(s_1) - b - R_k)} - 1 \right) \\
&\leq e^{\gamma Z_k} \left(\gamma (\hat{V}^{k}_{g,1}(s_1) - b - R_k) + \gamma^2 (\hat{V}^{k}_{g,1}(s_1) - b - R_k)^2 \right) \\
&= \Phi'(Z_k) \left(\hat{V}^{k}_{g,1}(s_1) - b - R_k + \gamma (\hat{V}^{k}_{g,1}(s_1) - b - R_k)^2 \right)
\end{aligned} \tag{24}$$

where the inequality follows from the fact that $e^z \leq 1 + z + z^2$ for all $z \leq 1$ and that $\gamma(\hat{V}^{k}_{g,1}(s_1) - b - R_k) \leq \gamma(\hat{V}^{k}_{g,1}(s_1) - b) \leq 2\gamma H \leq 1$. This is true because $\gamma, R_k, b \geq 0$, $\hat{V}^{k}_{g,1}(s_1) \leq 2H$ under $\mathcal{E}_g$, and $\gamma \leq 1/(2H)$. The last equality follows from $e^{\gamma Z_k} = \Phi'(Z_k)/\gamma$. Moreover, we have

$$\begin{aligned}
\gamma (\hat{V}^{k}_{g,1}(s_1) - b - R_k)^2 &\leq 3\gamma (\hat{V}^{k}_{g,1}(s_1) - b)^2 + 3 \cdot 108^2 \cdot \gamma^3 H^4 + 3\gamma \beta_R^2 \frac{Y_k^2}{\sum_{\tau=1}^{k} Y_\tau^2} \\
&\leq 27\gamma H^2 + 3 \cdot 108^2 \cdot \gamma^3 H^4 + 3\gamma \beta_R^2 \frac{Y_k^2}{\sum_{\tau=1}^{k} Y_\tau^2} \\
&\leq 27\gamma H^2 + 27\gamma H^2 + 3\gamma \beta_R^2 \frac{Y_k^2}{\sum_{\tau=1}^{k} Y_\tau^2} \\
&\leq 54\gamma H^2 + \frac{1}{2} \beta_R \frac{Y_k}{\sqrt{\sum_{\tau=1}^{k} Y_\tau^2}} \\
&= \frac{1}{2} R_k
\end{aligned} \tag{25}$$

where the first inequality follows from $(x + y + z)^2 \leq 3(x^2 + y^2 + z^2)$ for any x, y, z, the second inequality follows from $-3H \leq \hat{V}^{k}_{g,1}(s_1) - b \leq 2H$ under $\mathcal{E}_g$, the third inequality follows from

$\gamma \le 1/(36H)$, and the last inequality follows from $\gamma \le 1/(6\beta_R)$ and that $Y_k/\sqrt{\sum_{\tau=1}^{k} Y_\tau^2} \le 1$. Since $\Phi'(Z_k) \ge 0$, by applying (25) to (24), we have

$$\Phi(Z_{k+1}) - \Phi(Z_k) \le \Phi'(Z_k)\left(\hat{V}_{g,1}^{k}(s_1) - b - \frac{1}{2}R_k\right).$$

This concludes the proof. □

Lemma 26 (Restatement of Lemma 8)**.** *Assume that the good event $\mathcal{E}_g$ holds. Let $H \ge 2$, and let $K \ge c_{bias} dH(\log|\mathcal{A}| + 20)\log(2H/\delta)$, where c_{bias} is the absolute constant given by Lemma 23. Then we have*

$$\Phi'(Z_k) = Y_k, \quad \forall k \in [K+1]$$

Proof. We prove the statement by contradiction. Suppose that there exists some k such that $\Phi'(Z_k) \ne Y_k$, i.e., $\{k \in [K+1] : \Phi'(Z_k) \ne Y_k\} \ne \varnothing$. Given this, let k_{hit} denote the first episode in which this event occurs, i.e., $k_{\text{hit}} = \min\{k \in [K+1] : \Phi'(Z_k) \ne Y_k\}$. To conclude the proof, it is sufficient to show that $\Phi'(Z_{k_{\text{hit}}}) = Y_{k_{\text{hit}}}$, which contradicts the assumption.

Note that $k_{\text{hit}} - 1 \le K$. Then by Lemma 24, we have

$$\begin{aligned}
&\sum_{\tau=1}^{k_{\text{hit}}-1}\left(V_{f^\tau,1}^{\pi^\tau}(s_1) - V_{f^\tau,1}^{\pi^\star}(s_1)\right) + \sum_{\tau=1}^{k_{\text{hit}}-1} Y_\tau\left(\hat{V}_{g,1}^{\tau}(s_1) - b\right) \\
&\le H^2(\log|\mathcal{A}| + 20)\sqrt{dH\log(2K)}\sqrt{k_{\text{hit}} - 1 + \sum_{\tau=1}^{k_{\text{hit}}-1} Y_\tau^2} \\
&\quad + c_{\text{bias}} d^{3/2}H^2\sqrt{K}\log(2HK/\delta) + c_{\text{bias}} d^3 H^4 \log^4(2HK/\delta).
\end{aligned}$$

Note that for each $\tau \le k_{\text{hit}} - 1$, we assumed that $\Phi'(Z_\tau) = Y_\tau$. Then the above inequality can be rewritten as

$$\begin{aligned}
&\sum_{\tau=1}^{k_{\text{hit}}-1}\left(V_{f^\tau,1}^{\pi^\tau}(s_1) - V_{f^\tau,1}^{\pi^\star}(s_1)\right) + \sum_{\tau=1}^{k_{\text{hit}}-1} \Phi'(Z_\tau)\left(\hat{V}_{g,1}^{\tau}(s_1) - b\right) \\
&\le H^2(\log|\mathcal{A}| + 20)\sqrt{dH\log(2K)}\sqrt{k_{\text{hit}} - 1 + \sum_{\tau=1}^{k_{\text{hit}}-1} Y_\tau^2} \\
&\quad + c_{\text{bias}} d^{3/2}H^2\sqrt{K}\log(2HK/\delta) + c_{\text{bias}} d^3 H^4 \log^4(2HK/\delta).
\end{aligned} \tag{26}$$

Moreover, by summing the result of Lemma 7 over $\tau = 1, \dots, k_{\text{hit}} - 1$, we have

$$\sum_{\tau=1}^{k_{\text{hit}}-1}(\Phi(Z_{\tau+1}) - \Phi(Z_\tau)) \le \sum_{\tau=1}^{k_{\text{hit}}-1} \Phi'(Z_\tau)\left(\hat{V}_{g,1}^{\tau}(s_1) - b - \frac{1}{2}R_\tau\right). \tag{27}$$

Since the left-hand side is a telescoping sum and $\Phi(Z_1) = 0$, it can be written as $\Phi(Z_{k_{\text{hit}}})$. Then combining (26) and (27) yields

$$\begin{aligned}
&\sum_{\tau=1}^{k_{\text{hit}}-1} \left(V_{f^\tau,1}^{\pi^\tau}(s_1) - V_{f^\tau,1}^{\pi^\star}(s_1) \right) + \Phi(Z_{k_{\text{hit}}}) \\
&\leq H^2 \left(\log|\mathcal{A}| + 20\right) \sqrt{dH\log(2K)} \sqrt{k_{\text{hit}} - 1 + \sum_{\tau=1}^{k_{\text{hit}}-1} Y_\tau^2} \\
&\quad + c_{\text{bias}} d^{3/2} H^2 \sqrt{K} \log(2HK/\delta) + c_{\text{bias}} d^3 H^4 \log^4(2HK/\delta) \\
&\quad - \sum_{\tau=1}^{k_{\text{hit}}-1} \frac{\Phi'(Z_\tau)}{2} R_\tau \\
&\leq H^2 \left(\log|\mathcal{A}| + 20\right) \sqrt{dH\log(2K)} \sqrt{k_{\text{hit}} - 1} \\
&\quad + H^2 \left(\log|\mathcal{A}| + 20\right) \sqrt{dH\log(2K)} \sqrt{\sum_{\tau=1}^{k_{\text{hit}}-1} Y_\tau^2} \\
&\quad + c_{\text{bias}} d^{3/2} H^2 \sqrt{K} \log(2HK/\delta) + c_{\text{bias}} d^3 H^4 \log^4(2HK/\delta) \\
&\quad - \sum_{\tau=1}^{k_{\text{hit}}-1} \frac{\Phi'(Z_\tau)}{2} R_\tau
\end{aligned} \tag{28}$$

where the second inequality follows from the fact that $\sqrt{x+y} \leq \sqrt{x} + \sqrt{y}$. Again, since $\Phi'(Z_\tau) = Y_\tau$ for all $\tau \leq k_{\text{hit}} - 1$, we have

$$\begin{aligned}
\sum_{\tau=1}^{k_{\text{hit}}-1} \frac{\Phi'(Z_\tau)}{2} R_\tau &= \sum_{\tau=1}^{k_{\text{hit}}-1} \frac{\Phi'(Z_\tau)}{2} \left(108\gamma H^2 + \beta_R \frac{\Phi'(Z_\tau)}{\sqrt{\sum_{\tau'=1}^{\tau} \Phi'(Z_{\tau'})^2}} \right) \\
&\geq \frac{\beta_R}{2} \sum_{\tau=1}^{k_{\text{hit}}-1} \frac{\Phi'(Z_\tau)^2}{\sqrt{\sum_{\tau'=1}^{\tau} \Phi'(Z_{\tau'})^2}} \\
&= \frac{\beta_R}{2} \sum_{\tau=1}^{k_{\text{hit}}-1} \frac{\sum_{\tau'=1}^{\tau} \Phi'(Z_{\tau'})^2 - \sum_{\tau'=1}^{\tau-1} \Phi'(Z_{\tau'})^2}{\sqrt{\sum_{\tau'=1}^{\tau} \Phi'(Z_{\tau'})^2}} \\
&\geq \frac{\beta_R}{2} \sum_{\tau=1}^{k_{\text{hit}}-1} \left(\sqrt{\sum_{\tau'=1}^{\tau} \Phi'(Z_{\tau'})^2} - \sqrt{\sum_{\tau'=1}^{\tau-1} \Phi'(Z_{\tau'})^2} \right) \\
&\geq \frac{\beta_R}{2} \sqrt{\sum_{\tau=1}^{k_{\text{hit}}-1} \Phi'(Z_\tau)^2} \\
&= H^2 (\log|\mathcal{A}| + 20) \sqrt{dH\log(2K)} \sqrt{\sum_{\tau=1}^{k_{\text{hit}}-1} \Phi'(Z_\tau)^2}.
\end{aligned} \tag{29}$$

By combining (28) and (29), we have

$$\begin{aligned}
&\sum_{\tau=1}^{k_{\text{hit}}-1} \left(V_{f^\tau,1}^{\pi^\tau}(s_1) - V_{f^\tau,1}^{\pi^\star}(s_1) \right) + \Phi(Z_{k_{\text{hit}}}) \\
&\leq H^2 \left(\log|\mathcal{A}| + 20\right) \sqrt{dH\log(2K)} \sqrt{k_{\text{hit}} - 1} \\
&\quad + c_{\text{bias}} d^{3/2} H^2 \sqrt{K} \log(2HK/\delta) + c_{\text{bias}} d^3 H^4 \log^4(2HK/\delta).
\end{aligned} \tag{30}$$

Since $-H \leq V_{f^\tau,1}^{\pi^\tau}(s_1) - V_{f^\tau,1}^{\pi^\star}(s_1)$, it leads to

$$\Phi(Z_{k_{\text{hit}}}) + 1 \leq H(k_{\text{hit}} - 1) + H^2 \left(\log|\mathcal{A}| + 20\right) \sqrt{dH\log(2K)} \sqrt{k_{\text{hit}} - 1}$$

$$+ c_{\text{bias}} d^{3/2} H^2 \sqrt{K} \log(2HK/\delta) + c_{\text{bias}} d^3 H^4 \log^4(2HK/\delta) + 1.$$

To simplify the above relation, we use the following algebraic facts: $k_{\text{hit}} - 1 \leq K$; $\log K \leq \log(2K) \leq K$; $\log(2HK/\delta) \leq K \log(2H/\delta)$. Then it results in

$$\Phi(Z_{k_{\text{hit}}}) + 1 \leq 4 c_{\text{bias}} d^3 H^4 K^4 \left(\log|\mathcal{A}| + 20\right) \log^4\left(2H/\delta\right).$$

Moreover, by the choice of γ, we have $\gamma \leq 1/(36H\sqrt{K})$. Then it follows that

$$\begin{aligned}
\Phi'(Z_{k_{\text{hit}}}) &\leq \gamma(\Phi(Z_{k_{\text{hit}}}) + 1) \\
&\leq \frac{1}{9} c_{\text{bias}} d^3 H^3 K^{7/2} (\log|\mathcal{A}| + 20) \log^4(2H/\delta) \\
&< K^8
\end{aligned}$$

where the last inequality follows from $K \geq c_{\text{bias}} dH(\log|\mathcal{A}|+20)\log(2H/\delta)$. Note that this implies that $Y_{k_{\text{hit}}} = \min\{\Phi'(Z_{k_{\text{hit}}}), K^8\} = \Phi'(Z_{k_{\text{hit}}})$, which contradicts the assumption that k_{hit} is the first episode k that $\Phi'(Z_k) \neq Y_k$. Hence, this implies that $\{k \in [K+1] : \Phi'(Z_k) \neq Y_k\} = \varnothing$, which is equivalent to $\Phi'(Z_k) = Y_k$ for all $k \in [K+1]$. This concludes the proof. □

G.4 Proof of Theorem 4

In this section, we provide the proof of Theorem 4. Before presenting the proof, we first establish the following lemma, which bounds $\mathrm{Regret}(k)$, Y_{k+1}, and Z_{k+1} for all $k \in [K]$ by combining the results obtained so far.

Lemma 27. *Assume that the good event $\mathcal{E}_g$ holds. Let $H \geq 2$, and let $K \geq c_{bias} dH(\log|\mathcal{A}| + 20)\log(2H/\delta)$, where c_{bias} is the absolute constant given by Lemma 23. Then the following statements hold.*

1. *We have*

$$\begin{aligned}
\sum_{k=1}^{K} (V_{f^k,1}^{\pi^k}(s_1) - V_{f^k,1}^{\pi^\star}(s_1)) &\leq H^2 \left(\log|\mathcal{A}| + 20\right) \sqrt{dH\log(2K)}\sqrt{K} + 1 \\
&\quad + c_{bias} d^{3/2} H^2 \sqrt{K} \log(2HK/\delta) + c_{bias} d^3 H^4 \log^4(2HK/\delta).
\end{aligned}$$

2. *For any $k \in [K]$,*

$$Y_{k+1} \leq \frac{1}{9} c_{bias} d^3 H^3 K^{7/2} (\log|\mathcal{A}| + 20) \log^4(2H/\delta).$$

3. *For any $k \in [K]$,*

$$Z_{k+1} \leq 12H^2(\log|\mathcal{A}| + 20)\sqrt{dHK} \log\left(4 c_{bias} d^3 H^4 K^4 (\log|A| + 20) \log^4(2H/\delta)\right).$$

Proof. By Lemma 26, we know that $\Phi'(Z_k) = Y_k$ for all $k \in [K+1]$. By applying the same argument for deriving (30) in Lemma 26 but for any $k \in [K]$, instead of $k_{\text{hit}} - 1$, we have for each k,

$$\begin{aligned}
&\sum_{\tau=1}^{k} \left(V_{f^\tau,1}^{\pi^\tau}(s_1) - V_{f^\tau,1}^{\pi^\star}(s_1)\right) + \Phi(Z_{k+1}) \\
&\leq H^2 \left(\log|\mathcal{A}| + 20\right) \sqrt{dH\log(2K)}\sqrt{K} \\
&\quad + c_{\text{bias}} d^{3/2} H^2 \sqrt{K} \log(2HK/\delta) + c_{\text{bias}} d^3 H^4 \log^4(2HK/\delta).
\end{aligned} \tag{31}$$

Since $\Phi(Z_{k+1}) \geq -1$, we have the first statement.

Moreover, to derive the second statement, since $\sum_{\tau=1}^{k} \left(V_{f^\tau,1}^{\pi^\tau}(s_1) - V_{f^\tau,1}^{\pi^\star}(s_1)\right) \geq -HK$, we have

$$\begin{aligned}
\Phi(Z_{k+1}) + 1 &\leq HK + H^2 \left(\log|\mathcal{A}| + 20\right) \sqrt{dH\log(2K)}\sqrt{K} \\
&\quad + c_{\text{bias}} d^{3/2} H^2 \sqrt{K} \log(2HK/\delta) + c_{\text{bias}} d^3 H^4 \log^4(2HK/\delta) \\
&\leq 4 c_{\text{bias}} d^3 H^4 K^4 (\log|\mathcal{A}| + 20) \log^4(2H/\delta)
\end{aligned} \tag{32}$$

Since $\gamma \leq 1/(36H\sqrt{K})$, we have

$$\begin{aligned} Y_{k+1} = \Phi'(Z_{k+1}) &= \gamma(\Phi(Z_{k+1}) + 1) \\ &\leq \frac{1}{9} c_{\text{bias}} d^3 H^3 K^{7/2} (\log|\mathcal{A}| + 20) \log^4(2H/\delta). \end{aligned}$$

Finally, the third statement can be obtained as follows. Again, from (32), we derive that

$$\begin{aligned} Z_{k+1} &\leq \frac{1}{\gamma} \log\left(4c_{\text{bias}} d^3 H^4 K^4 (\log|A| + 20) \log^4(2H/\delta)\right) \\ &\leq 12H^2(\log|\mathcal{A}| + 20)\sqrt{dHK} \log\left(4c_{\text{bias}} d^3 H^4 K^4 (\log|A| + 20) \log^4(2H/\delta)\right). \end{aligned}$$

□

Proof of Theorem 4. To prove the theorem, we consider the following two cases: (1) $K < \beta_\phi$ or $K < c_{\text{bias}} dH(\log|\mathcal{A}|+20)\log(2H/\delta)$ and (2) $\beta_\phi \leq K$ and $c_{\text{bias}} dH(\log|\mathcal{A}|+20)\log(2H/\delta) \leq K$.

Case 1: Suppose that $K < \beta_\phi$ or $K < c_{\text{bias}} dH(\log|\mathcal{A}| + 20)\log(2H/\delta)$. In this case, we have

$$\begin{aligned} K &< \max\{\beta_\phi, c_{\text{bias}} dH(\log|\mathcal{A}| + 20)\log(2H/\delta)\} \\ &\leq \beta_\phi + c_{\text{bias}} dH(\log|\mathcal{A}| + 20)\log(2H/\delta) \\ &\leq 748dH\sqrt{\log(2HK/\delta)}\log K + c_{\text{bias}} dH(\log|\mathcal{A}| + 20)\log(2H/\delta). \end{aligned}$$

Then the trivial bounds for $\text{Regret}(K)$ and $\text{Violation}(K)$, i.e., HK, satisfy

$$\text{Regret}(K), \text{Violation}(K) \leq HK = \widetilde{\mathcal{O}}\left(dH^3\right).$$

Case 2: Suppose that $\beta_\phi \leq K$ and $c_{\text{bias}} dH(\log|\mathcal{A}|+20)\log(2H/\delta) \leq K$. It is clear that $1 \leq \beta_\phi$. In this case, we can assume $\mathcal{E}_g$, which occurs with probability at least $1-\delta$ by Lemma 19. Moreover, Lemmas 26, 27, and 20 hold.

To bound $\text{Regret}(K)$, by the first statement of Lemma 27, we have

$$\begin{aligned} \text{Regret}(K) &= \sum_{k=1}^{K} (V_{f^k,1}^{\pi^k}(s_1) - V_{f^k,1}^{\pi^\star}(s_1)) \\ &\leq H^2 (\log|\mathcal{A}| + 20)\sqrt{dH\log(2K)}\sqrt{K} + 1 \\ &\quad + c_{\text{bias}} d^{3/2} H^2 \sqrt{K}\log(2HK/\delta) + c_{\text{bias}} d^3 H^4 \log^4(2HK/\delta) \\ &= \widetilde{\mathcal{O}}\left(\sqrt{d^3H^4K} + \sqrt{dH^5K} + d^3H^4\right). \end{aligned}$$

To bound $\text{Violation}(K)$, we observe that

$$\begin{aligned} \text{Violation}(K) &= \sum_{k=1}^{K} \left(V_{g^k,1}^{\pi^k}(s_1) - b\right) \\ &= \sum_{k=1}^{K} \left(V_{g^k,1}^{\pi^k}(s_1) - \hat{V}_{g,1}^{k}(s_1)\right) + \sum_{k=1}^{K} \left(\hat{V}_{g,1}^{k}(s_1) - b\right) \\ &= \sum_{k=1}^{K} \left(V_{g^k,1}^{\pi^k}(s_1) - \hat{V}_{g,1}^{k}(s_1)\right) + Z_{K+1} + \sum_{k=1}^{K} R_k \end{aligned}$$

where the last equality can be obtained by summing $\hat{V}_{g,1}^{k}(s_1) - b = Z_{k+1} - Z_k + R_k$ over k, and the fact that $Z_1 = 0$. Then by Lemma 23, the third statement of Lemma 27, and Lemma 20, we have

$$\begin{aligned} &\text{Violation}(K) \\ &\leq c_{\text{bias}} \left(d^{3/2}H^2\sqrt{K}\log(2HK/\delta) + d^3H^4\log^4(2HK/\delta)\right) \\ &\quad + 12H^2(\log|\mathcal{A}| + 20)\sqrt{dHK}\log\left(4c_{\text{bias}} d^3 H^4 K^4 (\log|A| + 20)\log^4(2H/\delta)\right) \\ &\quad + 12\sqrt{dH^5}(\log|\mathcal{A}| + 20)\sqrt{K}\log(3dH(\log|\mathcal{A}| + 20)K) \\ &= \widetilde{\mathcal{O}}\left(\sqrt{d^3H^4K} + \sqrt{dH^5K} + d^3H^4\right). \end{aligned}$$

□

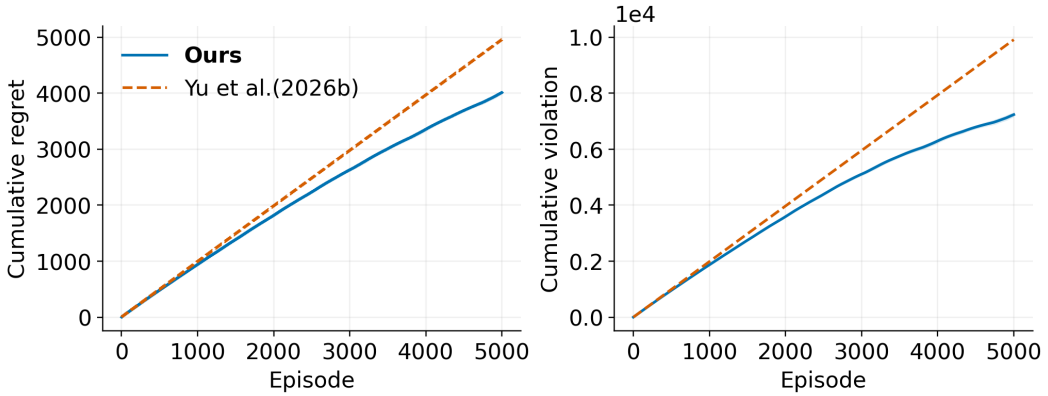

Figure 1: Cumulative regret (left) and the positive part of cumulative constraint violation (right) for $K = 5{,}000$ episodes. Curves show averages over five seeds; shaded regions indicate 95% confidence intervals.

H Numerical Experiments

We evaluate Algorithm 1 on the chain CMDP of Yu et al. (2026a), with periodically alternating losses and a fixed constraint, and compare it with the algorithm of Yu et al. (2026b). We use $H = 10$, state space $\mathcal{S} = \{0, \dots, H+1\}$, action space $\mathcal{A} = \{-1, 1\}^4$, and an exact linear-MDP representation with feature dimension $d = 22$. Each episode starts at $s_1 = 0$. For every step h and state $s < H$, we set $P_h(s+1 \mid s, a) = 0.95 - 0.01 \sum_{i=1}^{4} a_i$ and $P_h(H+1 \mid s, a) = 1 - P_h(s+1 \mid s, a)$, with all other transition probabilities equal to zero. States H and $H+1$ are absorbing: $P_h(s \mid s, a) = 1$ for $s \in \{H, H+1\}$. Each episode lasts H steps, including steps after reaching an absorbing state. Both methods receive full-information loss feedback and the same known constraint function, while transitions remain unknown. We conduct five simulations with paired random seeds, each running for $K = 5{,}000$ episodes.

Let $q(a) = \frac{1}{8} \sum_{i=1}^{4} (a_i + 1)$. For $s < H$, we set $f_h^k(s, a) = 1 - 0.4q(a)$ when $\lfloor k/10 \rfloor$ is even and $f_h^k(s, a) = 1 - 0.4(1 - q(a))$ otherwise. At the absorbing states, we set $f_h^k(H, a) = 1$ and $f_h^k(H+1, a) = 0$. The constraint function is fixed: $g_h^k(s, a) = 1 - q(a)$ for $s < H$ and $g_h^k(s, a) = 1$ otherwise, with budget $b = 4$. These definitions apply to every step h and episode k, with episode indices starting at $k = 1$.

To promote learning, we reduce the optimistic bonus coefficients while preserving their different dependence on K: we set $\beta_p = 0.01$ and $\beta_b = 0.01K^{1/4}$, with contraction coefficients $\beta_\phi = 4\beta_p \log K$ and $\beta_w = 4\beta_b \log K$. For our dual update, we set $\beta_R = 0.1$ and $\gamma = \min\{0.1/\sqrt{K}, 1/(6\beta_R)\}$, and replace the coefficient 108 in R_k with 0.1. These practical parameter settings differ from those covered by Theorem 4. All remaining algorithmic parameters follow the respective papers, with $\delta = 0.05$.

Figure 1 summarizes the results. Our method achieves lower cumulative regret and constraint violation than the baseline over the later episodes. The baseline of Yu et al. (2026b) exhibits approximately linear growth in both quantities. These results illustrate an empirical advantage of our method on this instance under the tested parameterizations.

I Auxiliary Lemmas

Lemma 28 (Lemma 22 of Cassel & Rosenberg (2024)). *Let $\hat{\boldsymbol{\mu}}_h^k : \mathbb{R}^{|\mathcal{S}|} \to \mathbb{R}^d$ be the linear operator defined in Line 11 of Algorithm 1. For all $h \in [H]$, let $\mathcal{V}_h \subseteq \mathbb{R}^{|\mathcal{S}|}$ be a set of mappings $\hat{V} : \mathcal{S} \to \mathbb{R}$ such that $||\hat{V}||_\infty \le \beta_Q$ and $\beta_Q \ge 1$. With probability at least $1 - \delta$, for all $h \in [H]$, $\hat{V} \in \mathcal{V}_{h+1}$, and $k \ge 1$,*

$$||(\boldsymbol{\mu}_h - \hat{\boldsymbol{\mu}}_h^k)\hat{V}||_{\boldsymbol{\Lambda}_h^k} \le 4\beta_Q \sqrt{d \log(K+1) + 2 \log(H\mathcal{N}_\epsilon/\delta)}$$

where $\epsilon \le \beta_Q\sqrt{d}/(2K)$, $\mathcal{N}_\epsilon = \sum_{h=1}^H \mathcal{N}_\epsilon(\mathcal{V}_h)$, *and* $\mathcal{N}_\epsilon(\mathcal{V}_h)$ *is the* ϵ*-covering number of* $\mathcal{V}_h$ *with respect to the* ℓ_∞*-norm.*

Lemma 29 (Lemma 24 of Cassel & Rosenberg (2024)). *Let* $\mathcal{V} = \{V(\cdot;\boldsymbol{\theta}) : ||\boldsymbol{\theta}|| \le W\}$ *denote a class of functions* $V : \mathcal{S} \to \mathbb{R}$. *Suppose that any* $V \in \mathcal{V}$ *is* L*-Lipschitz with respect to* $\boldsymbol{\theta}$ *and the supremum distance, i.e.,*

$$||V(\cdot;\boldsymbol{\theta}_1) - V(\cdot;\boldsymbol{\theta}_2)||_\infty \le L||\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2||, \quad ||\boldsymbol{\theta}_1||, ||\boldsymbol{\theta}_2|| \le W.$$

Let $\mathcal{N}_\epsilon(\mathcal{V})$ *be the* ϵ*-covering number of* $\mathcal{V}$ *with respect to the* ℓ_∞*-norm. Then*

$$\log \mathcal{N}_\epsilon(\mathcal{V}) \le d\log(1 + 2WL/\epsilon).$$

Lemma 30 (Lemma 8 in Cassel & Rosenberg (2024)). *The number of epochs* $|E|$ *is bounded by* $(3/2)dH\log(2K)$.

Lemma 31 (Lemma D.4 of Rosenberg et al. (2020)). *Let* $\{X_t\}_{t\ge1}$ *be a sequence of random variables with expectation adapted to a filtration* $\mathcal{F}_t$. *Suppose that* $0 \le X_t \le C$ *almost surely. Then with probability at least* $1-\delta$,

$$\sum_{t=1}^T \mathbb{E}[X_t \mid \mathcal{F}_{t-1}] \le 2\sum_{t=1}^T X_t + 4C\log\frac{2T}{\delta}.$$

Lemma 32 (Lemma 18 in Cassel & Rosenberg (2024)). *For any* $K \ge 2$, $z \ge 0$, *we have* $\max_{z\ge0}[z\cdot\sigma(-\beta z + \log K)] \le 2(\log K)/\beta$.

Lemma 33 (Lemma 19 in Cassel & Rosenberg (2024)). *For any* $K \ge 1$, $z \ge 0$, *we have* $\sigma(z - \log K) \le 2(z^2 + K^{-1})$.

Lemma 34 (Lemma 3 in Cassel & Rosenberg (2024)). *For any* $m \in E$ *and* $(s,a,h,k) \in \mathcal{S}\times\mathcal{A}\times[H]\times K(m)$, *and* $v \in \mathbb{R}^d$, *we have*

$$\left(\boldsymbol{\phi}(s,a) - \bar{\boldsymbol{\phi}}_h^m(s,a)\right)^\top \mathbf{v} \le \left(4\beta_\phi^2||\boldsymbol{\phi}(s,a)||^2_{(\boldsymbol{\Lambda}_h^k)^{-1}} + 2K^{-1}\right)|\boldsymbol{\phi}(s,a)^\top\mathbf{v}|$$

Lemma 35 (Lemma 12 in Abbasi-Yadkori et al. (2011)). *Let* $\mathbf{A}, \mathbf{B}$ *be positive definite matrices, and let* $\mathbf{C}$ *be a positive semidefinite matrix such that* $\mathbf{A} = \mathbf{B} + \mathbf{C}$. *Then, we have*

$$\sup_{\mathbf{x}\neq\mathbf{0}} \frac{\mathbf{x}^\top\mathbf{A}\mathbf{x}}{\mathbf{x}^\top\mathbf{B}\mathbf{x}} \le \frac{\det(\mathbf{A})}{\det(\mathbf{B})}.$$

Lemma 36 (Lemma D.2 of Jin et al. (2020)). *Let* $\{\boldsymbol{\phi}_t\}_{t\ge0}$ *be a bounded sequence in* $\mathbb{R}^d$ *satisfying* $\sup_{t\ge0}||\boldsymbol{\phi}_t|| \le 1$. *Let* $\boldsymbol{\Lambda}_0 \in \mathbb{R}^{d\times d}$ *be a positive definite matrix. For any* $t \ge 0$, *define*

$$\boldsymbol{\Lambda}_t = \boldsymbol{\Lambda}_0 + \sum_{j=1}^t \boldsymbol{\phi}_j\boldsymbol{\phi}_j^\top.$$

Then, if the smallest eigenvalue of $\boldsymbol{\Lambda}_0$ *satisfies* $\lambda_{\min}(\boldsymbol{\Lambda}_0) \ge 1$, *we have*

$$\log\left[\frac{\det(\boldsymbol{\Lambda}_t)}{\det(\boldsymbol{\Lambda}_0)}\right] \le \sum_{j=1}^t \boldsymbol{\phi}_j^\top\boldsymbol{\Lambda}_{j-1}^{-1}\boldsymbol{\phi}_j \le 2\log\left[\frac{\det(\boldsymbol{\Lambda}_t)}{\det(\boldsymbol{\Lambda}_0)}\right].$$

AI USE STATEMENT

In this work, we used generative AI tools for polishing paragraphs, checking the correctness of mathematical arguments, and obtaining assistance with the implementation of numerical experiments. We have not used generative AI tools to write complete mathematical proofs or complete drafts of the manuscript, and other tasks requiring disclosure are not applicable to this work. We have reviewed all AI-assisted work. All AI-generated experimental code was reviewed by the authors. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.