Provably Efficient Safe Exploration via Primal-Dual Policy Optimization

Dongsheng Ding
University of Southern California
dongshed@usc.edu

Xiaohan Wei
Facebook, Inc.
ubimeteor@fb.com

Zhuoran Yang
Princeton University
zy6@princeton.edu

Zhaoran Wang
Northwestern University
zhaoranwang@gmail.com

Mihailo R. Jovanović
University of Southern California
mihailo@usc.edu

Abstract

We study the safe reinforcement learning problem using the constrained Markov decision processes in which an agent aims to maximize the expected total reward subject to a safety constraint on the expected total value of a utility function. We focus on an episodic setting with the function approximation where the Markov transition kernels have a linear structure but do not impose any additional assumptions on the sampling model. Designing safe reinforcement learning algorithms with provable computational and statistical efficiency is particularly challenging under this setting because of the need to incorporate both the safety constraint and the function approximation into the fundamental exploitation/exploration tradeoff. To this end, we present an Optimistic Primal-Dual Proximal Policy OPtimization (OPDOP) algorithm where the value function is estimated by combining the least-squares policy evaluation and an additional bonus term for safe exploration. We prove that the proposed algorithm achieves an $\widetilde{O}(dH^{2.5}\sqrt{T})$ regret and an $\widetilde{O}(dH^{2.5}\sqrt{T})$ constraint violation, where d is the dimension of the feature mapping, H is the horizon of each episode, and T is the total number of steps. These bounds hold when the reward/utility functions are fixed but the feedback after each episode is bandit. Our bounds depend on the capacity of the state-action space only through the dimension of the feature mapping and thus our results hold even when the number of states goes to infinity. To the best of our knowledge, we provide the first provably efficient online policy optimization algorithm for constrained Markov decision processes in the function approximation setting, with safe exploration.

Proceedings of the 24[th] International Conference on Artificial Intelligence and Statistics (AISTATS) 2021, San Diego, California, USA. PMLR: Volume 130.

1 Introduction

Reinforcement Learning (RL) studies how an agent learns to maximize its expected total reward by interacting with an unknown environment over time [60]. Safe RL augments RL with a practical consideration of safety to deal with restrictions/constraints arising from real-world problems [33, 6, 28]. Examples include collision-avoidance in autonomous robots [31, 32], cost limitations in medical applications [34, 11], and legal and business restrictions in financial management [2]. A standard environment model for safe RL is the Constrained Markov Decision Processes (CMDPs) [5] that generalize the classical MDPs to maximizing the expected total reward under a safety-related constraint on the expected total utility [3, 65]. The presence of constraints makes the fundamental exploration/exploitation trade-off more challenging.

There is considerable growth in safe RL, especially those studies on CMDPs, e.g., constrained policy gradient [63, 59], Lagrangian-based actor-critic [15, 14, 61, 46, 74], constrained policy optimization [3, 72, 78], primal-dual policy optimization [53, 52]. A key highlight of their developments is the successful integration of the constrained optimization and the policy-based RL for addressing constraints. Notwithstanding many successes, these safe RL algorithms either do not have a convergence theory or are limited to asymptotic convergence. In practice, only a finite amount of data is available. Hence, it is imperative to design safe RL algorithms with computational and statistical efficiency guarantees. For this purpose, we must address the exploration/exploitation trade-off under constraints.

In this work, we look at the challenging problem of finding a sequence of policies in response to online streaming samples of transition, reward functions, and utility functions. We attempt to provide theoretical guarantees on the regret of an algorithm approaching the best policy in hindsight, and feasibility region determined by constraints. The task of *safe exploration* is to explore the unknown environment and learn to adapt the policy to the constraint set. Our problem setting deviates from existing scenarios, where good priors on constraints or transition models are more focused, e.g., references [62, 13, 25, 66, 23, 24, 65]. Recent policy-based safe RL algorithms for CMDPs, e.g., constrained policy optimization [3, 72, 78] and primal-dual policy optimization [53, 52], seek a single safe policy via the constrained policy optimization whose sample efficiency guarantees do not have a theory.

In this paper, we aim to answer a theoretical question:

Can we design a provably sample efficient online policy optimization algorithm for CMDPs in the function approximation setting?

Contribution. We propose a provably efficient safe RL algorithm for CMDPs with an unknown transition model in the linear episodic setting – an Optimistic Primal-Dual Proximal Policy OPtimization (OPDOP) algorithm – where the value function is estimated by combining the least-squares policy evaluation and an additional bonus term for safe exploration. Theoretically, we prove that the proposed algorithm achieves an $\widetilde{O}(dH^{2.5}\sqrt{T})$ regret and the same $\widetilde{O}(dH^{2.5}\sqrt{T})$ constraint violation, where d is the dimension of the feature mapping, H is the horizon of each episode, and T is the total number of steps. We establish these bounds in the setting where the reward/utility functions are fixed but the feedback after each episode is bandit. Our bounds depend on the capacity of the state space only through the dimension of the feature mapping and thus hold even when the number of states goes to infinity. To the best of our knowledge, our result is the first provably efficient online policy optimization for CMDPs in the function approximation setting, with safe exploration.

Related Work. Our work is related to a line of provably efficient RL algorithms based on the linear function approximation, e.g., references [70, 71, 37, 20, 76]. Using the optimism in the face of uncertainty [7, 19], these references address the exploration/exploitation trade-off by adding the Upper Confidence Bound (UCB) bonus, and proposed algorithms are provably sample-efficient. A closely-related reference [20] connects policy-based RL with optimism, and proposes an optimistic proximal policy optimization with UCB exploration. However, all these references only study some particular MDPs in unconstrained RL problems. Additional efforts need to pay for making them work for CMDPs. Our work seeks to design an optimistic variant of proximal policy optimization for CMDPs. For the large CMDPs with unknown transition models, there is a line of literature that is related to the policy optimization under constraints, e.g., references [63, 3, 72, 61, 48, 78]. However, the exploration under constraints is less studied and their theoretical guarantees are unknown. Our work fills in this gap.

The study of RL algorithms for CMDPs has received growing attention, especially those on learning CMDPs with unknown transitions and rewards. As we know, most of them are model-based and only apply to finite state-action spaces. References [58, 29] leverage upper confidence bound (UCB) on fixed reward, utility, and transition probability to propose sample-efficient algorithms for tabular CMDPs; reference [58] establishes an $\widetilde{O}(\sqrt{|\mathcal{A}|T^{1.5}\log T})$ regret and constraint violation via linear program in the average-cost case in time T; reference [29] achieves an $\widetilde{O}(|\mathcal{S}|\sqrt{H^3T})$ regret and constraint violation in the episodic setting via linear program and primal-dual policy optimization, where $\mathcal{S}$ is a state-space, $\mathcal{A}$ is an action space, and H is a fixed horizon of episode. In reference [55], the authors study an adversarial stochastic shortest path problem under constraints and unknown transitions with $\widetilde{O}(|\mathcal{S}|\sqrt{|\mathcal{A}|H^2T})$ regret and constraint violation. Reference [10] extends Q-learning with optimism for finite state-action CMDPs with peak constraints. Reference [18] proposes UCB-based convex planning for episodic tabular CMDPs in dealing with convex or hard constraints. References [40, 35] establish probably approximately correct (PAC) guarantees that enjoy better problem-dependent sample-complexity. In contrast, our proposed algorithm can potentially apply to scenarios with infinite state-space, and our provided sublinear regret and constraint violation bounds only depend on the implicit dimension instead of the true dimension of the state space. Compared to more recent references [26, 69, 21, 77], our development attacks the exploration directly and does not rely on any policy 'simulators' (or generative models).

2 Problem Setup

We consider an episodic Markov decision process (MDP) – $\text{MDP}(\mathcal{S},\mathcal{A},H,\mathbb{P},r)$ – where $\mathcal{S}$ is a state space, $\mathcal{A}$ is an action space, H is a fixed length of each episode, $\mathbb{P} = \{\mathbb{P}_h\}_{h=1}^H$ is a collection of transition probability measures, and $r = \{r_h\}_{h=1}^H$ is a collection of reward functions. We assume that $\mathcal{S}$ is a measurable space with possibly infinite number of elements. Moreover, for each step $h \in [H]$, $\mathbb{P}_h(\,\cdot\,|\,x,a)$ is a transition kernel over next state if action a is taken for state x and r_h: $\mathcal{S}\times\mathcal{A}\to[0,1]$ is a reward function. The constrained MDP – CMDP$(\mathcal{S},\mathcal{A},H,\mathbb{P},r,g)$ – additionally contains utility functions $g = \{g_h\}_{h=1}^H$ where g_h: $\mathcal{S}\times\mathcal{A}\to[0,1]$. We assume that reward/utility functions are deterministic. Our analysis readily generalizes to the setting where reward/utility are random.

Let the policy space $\Delta(\mathcal{A}\,|\,\mathcal{S},H)$ be $\{\{\pi_h(\,\cdot\,|\,\cdot\,)\}_{h=1}^H$: $\pi_h(\,\cdot\,|\,x)\in\Delta(\mathcal{A})$, $\forall x\in\mathcal{S}$ and $h\in[H]\}$, where $\Delta(\mathcal{A})$ denotes a probability simplex over the action space. Let $\pi^k\in\Delta(\mathcal{A}\,|\,\mathcal{S},H)$ be a policy taken by the agent at episode k, where $\pi_h^k(\,\cdot\,|\,x_h^k)$: $\mathcal{S}\to\mathcal{A}$ is the action that the agent takes at state x_h^k. For simplicity, we assume the initial state x_1^k to be fixed as x_1 in different episodes for brevity. The agent interacts with the environment in the kth episode as follows. At the beginning, the agent determines a policy π^k. Then, at each step $h\in[H]$, the agent observes the state $x_h^k\in\mathcal{S}$, determines an action a_h^k following the policy $\pi_h^k(\,\cdot\,|\,x_h^k)$, and receives a reward $r_h(x_h^k,a_h^k)$ together with an utility $g_h(x_h^k,a_h^k)$. Meanwhile, the MDP evolves into next state x_{h+1}^k drawing from the probability $\mathbb{P}_h(\,\cdot\,|\,x_h^k,a_h^k)$. The episode terminates at state x_H^k in which no control action is taken and both reward and utility functions are equal to zero. In this paper, we focus a bandit setting where the agent only observes the values of reward/utility functions, $r_h(x_h^k,a_h^k)$, $g_h(x_h^k,a_h^k)$, at visited state-action pair (x_h^k,a_h^k). We assume that reward/utility functions are fixed over episodes.

Given a policy $\pi\in\Delta(\mathcal{A}\,|\,\mathcal{S},H)$, the value function $V_{r,h}^\pi$ associated with the reward function r at each step h are the expected values of total rewards,

$$V_{r,h}^\pi(x) \;=\; \mathbb{E}_\pi\left[\,\sum_{i=h}^H r_i(x_i,a_i)\,\middle|\,x_h = x\right]$$

for all $x\in\mathcal{S}$, $h\in[H]$, where the expectation $\mathbb{E}_\pi$ is taken over the random state-action sequence $\{(x_h,a_h)\}_{h=i}^H$; the action a_h follows the policy $\pi_h(\,\cdot\,|x_h)$ at the state x_h and the next state x_{h+1} follows the transition dynamics $\mathbb{P}_h(\,\cdot\,|x_h,a_h)$. Thus, the action-value function $Q_{r,h}^\pi(x,a)$: $\mathcal{S}\times\mathcal{A}\to\mathbb{R}$ associated with the reward function r is the expected value of total rewards when the agent starts from state-action pair (x,a) at step h and follows policy π,

$$Q_{r,h}^\pi(x,a) \;=\; \mathbb{E}_\pi\left[\,\sum_{i=h}^H r_i(x_i,a_i)\,\middle|\,x_h = x, a_h = a\right]$$

for all $(x,a)\in\mathcal{S}\times\mathcal{A}$ and $h\in[H]$. Similarly, we define the value function $V_{g,h}^\pi$: $\mathcal{S}\to\mathbb{R}$ and the action-value function $Q_{g,h}^\pi(x,a)$: $\mathcal{S}\times\mathcal{A}\to\mathbb{R}$ associated with the utility function g. Denote symbol $\diamond = r$ or g. For brevity, we take the shorthand $\mathbb{P}_h V_{\diamond,h+1}^\pi(x,a) := \mathbb{E}_{x'\sim\mathbb{P}_h(\,\cdot\,|\,x,a)}V_{\diamond,h+1}^\pi(x')$. The Bellman equations associated with a policy π are given by

$$Q_{\diamond,h}^\pi(x,a) \;=\; \big(\diamond_h + \mathbb{P}_h V_{\diamond,h+1}^\pi\big)(x,a) \tag{1}$$

where $V_{\diamond,h}^\pi(x) = \big\langle Q_{\diamond,h}^\pi(x,\,\cdot\,),\pi_h(\,\cdot\,|\,x)\big\rangle_{\mathcal{A}}$, for all $(x,a)\in\mathcal{S}\times\mathcal{A}$. Here, the inner product of a function f: $\mathcal{S}\times\mathcal{A}\to\mathbb{R}$ with $\pi(\,\cdot\,|\,x)\in\Delta(\mathcal{A})$ at fixed $x\in\mathcal{S}$ represents $\langle f(x,\,\cdot\,),\pi(\,\cdot\,|\,x)\rangle_{\mathcal{A}} := \sum_{a\in\mathcal{A}}\langle f(x,a),\pi(a\,|\,x)\rangle$.

2.1 Learning Performance

The learning agent aims to find a solution of a constrained problem in which the objective function is the expected total rewards and the constraint is on the expected total utilities,

$$\underset{\pi\,\in\,\Delta(\mathcal{A}\,|\,\mathcal{S},H)}{\text{maximize}}\;\; V_{r,1}^\pi(x_1)\;\;\text{subject to}\;\; V_{g,1}^\pi(x_1)\;\geq\; b \tag{2}$$

where we take $b\in(0,H]$ to avoid triviality. It is readily generalized to the problem with multiple constraints. Let $\pi^\star\in\Delta(\mathcal{A}\,|\,\mathcal{S},H)$ be a solution to problem (2). Since the policy $\pi^\star$ is computed from knowing the transition model and all reward and utility functions, we refer it as an optimal policy in-hindsight.

The associated Lagrangian of problem (2) is given by

$$\mathcal{L}(\pi,Y) \;:=\; V_{r,1}^\pi(x_1) \,+\, Y\,(V_{g,1}^\pi(x_1) - b)$$

where π is the primal variable and $Y\geq 0$ is the dual variable. We can cast (2) into a saddle-point problem,

$$\underset{\pi\,\in\,\Delta(\mathcal{A}\,|\,\mathcal{S},H)}{\text{maximize}}\;\;\underset{Y\,\geq\,0}{\text{minimize}}\;\;\mathcal{L}(\pi,\,Y)$$

where $\mathcal{L}(\pi,Y)$ is convex in Y and is non-concave in π in general. To address the non-concavity, we will exploit the structure of value functions to propose a variant of Lagrange multipliers method for constrained RL problems in Section 3, which warrants a new line of primal-dual mirror descent type analysis in sequel. This distinguishes from unconstrained RL, e.g., [4, 20].

Another key feature of constrained RL is the safe exploration under constraints [33]. Without any constraint information *a priori*, it is infeasible for each policy to satisfy the constraint since utility information on constraints is only revealed after a policy is decided. Instead, we allow each policy to violate the constraint in each episode and minimize regret while minimizing total constraint violations for safe exploration over K episodes. We define the regret as the difference between the total reward value of policy $\pi^\star$ in hindsight and that of the agent's policy π^k over K episodes, and the constraint violation as a difference between the offset Kb and the total utility value of the agent's policy π^k over K episodes,

$$\begin{aligned} \text{Regret}(K) &= \sum_{k=1}^{K} \left(V_{r,1}^{\pi^\star}(x_1) - V_{r,1}^{\pi^k}(x_1) \right) \\ \text{Violation}(K) &= \sum_{k=1}^{K} \left(b - V_{g,1}^{\pi^k}(x_1) \right). \end{aligned} \tag{3}$$

In this paper, we design algorithms, taking bandit feedback of the reward/utility functions, with both regret and constraint violation being sublinear in the total number of steps $T := HK$. Put differently, the algorithm should ensure that given $\epsilon > 0$, if $T = O(1/\epsilon^2)$, then both $\text{Regret}(K) = O(\epsilon)$ and $\text{Violation}(K) = O(\epsilon)$ hold with high probability.

Let $\mathcal{D}(Y) := \text{maximize}_\pi \mathcal{L}(\pi, Y)$ be the dual function and $Y^\star := \text{argmin}_{Y \geq 0} \mathcal{D}(Y)$ be the optimal dual variable. We assume feasibility for problem (2) in Assumption 1 that is known as the Slater condition [53, 29, 55]. It is convenient to establish the strong duality [53] and the boundedness of the optimal dual variable $Y^\star$ that can be found in Appendix E.

Assumption 1 (Feasibility)**.** *There exists $\gamma > 0$ and $\bar{\pi} \in \Delta(\mathcal{A} \,|\, \mathcal{S}, H)$ such that $V_{g,1}^{\bar{\pi}}(x_1) \geq b + \gamma$.*

Lemma 1 (Strong Duality, Boundedness of $Y^\star$)**.** *Let Assumption 1 hold. Then $V_{r,1}^{\pi^\star}(x_1) = \mathcal{D}(Y^\star)$. Moreover, $0 \leq Y^\star \leq (V_{r,1}^{\pi^\star}(x_1) - V_{r,1}^{\bar{\pi}}(x_1))/\gamma$.*

Lemma 1 provides useful optimization properties of (2) for our algorithm design and analysis.

2.2 Linear Function Approximation

We focus on a class of CMDPs, where transition kernels are linear in feature maps.

Assumption 2. *The* $\text{CMDP}(\mathcal{S}, \mathcal{A}, H, \mathbb{P}, r, g)$ *is a linear MDP with a kernel feature map ψ: $\mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}^{d_1}$, if for any $h \in [H]$, there exists a vector $\theta_h \in \mathbb{R}^{d_1}$ with $\|\theta_h\|_2 \leq \sqrt{d_1}$ such that for any $(x, a, x') \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}$,*

$$\mathbb{P}_h(x' \,|\, x, a) = \langle \psi(x, a, x'), \theta_h \rangle;$$

there exists a feature map φ: $\mathcal{S} \times \mathcal{A} \to \mathbb{R}^{d_2}$ and vectors $\theta_{r,h}, \theta_{g,h} \in \mathbb{R}^{d_2}$ such that for any $(x, a) \in \mathcal{S} \times \mathcal{A}$,

$$r_h(x, a) = \langle \varphi(x, a), \theta_{r,h} \rangle \text{ and } g_h(x, a) = \langle \varphi(x, a), \theta_{g,h} \rangle$$

where $\max(\|\theta_{r,h}\|_2, \|\theta_{g,h}\|_2) \leq \sqrt{d_2}$. Moreover, we assume that for any function V: $\mathcal{S} \to [0, H]$, $\left\| \int_{\mathcal{S}} \psi(x, a, x') V(x') dx' \right\|_2 \leq \sqrt{d_1} H$ for all $(x, a) \in \mathcal{S} \times \mathcal{A}$, and $\max(d_1, d_2) \leq d$.

Assumption 2 adapts the definition of linear kernel MDP [8, 79, 20] for CMDPs. Linear kernel MDP examples include tabular MDPs [79], feature embedded transition models [71], and linear combinations of base models [50]. We can construct related examples of CMDPs with linear structure by adding adding proper constraints. For usefulness of linear structure, see discussions in the literature [27, 64, 43]. For more general transition dynamics, see factored MDPs [54].

Although our definition in Assumption 2 and linear MDPs [70, 37] all contain tabular MDPs as special cases, they define transition dynamics using different feature maps. They are not comparable since one cannot be implied by the other [79]. We provide more details on the tabular case of Assumption 2 in Section 5.

3 Proposed Algorithm

In Algorithm 1, we present a new variant of proximal policy optimization [57] – an Optimistic Primal-Dual Proximal Policy OPtimization (OPDOP) algorithm. We effectuate the optimism through the Upper-Confidence Bounds (UCB) [70, 71, 37], and address the constraints via the union of the Lagrange multipliers method with the value function structure that is captured by the performance difference lemma [38, 20].

Lemma 2 (Performance Difference Lemma)**.** *For any two policies $\pi, \pi' \in \Delta(\mathcal{A} \,|\, \mathcal{S}, H)$, $\diamond = r$ or g,*

$$\begin{aligned} & V_{\diamond,1}^{\pi'}(x_1) - V_{\diamond,1}^{\pi}(x_1) \\ &= \mathbb{E}_{\pi'} \left[\sum_{h=1}^{H} \left\langle Q_{\diamond,h}^{\pi}(x_h, \cdot), \pi'_h(\cdot \,|\, x_h) - \pi_h(\cdot \,|\, x_h) \right\rangle \Big| \, x_1 \right]. \end{aligned}$$

In each episode, our algorithm consists of three main stages. The first stage (lines 4–8) is *Policy Improvement*: we receive a new policy π^k by improving previous π^{k-1} via a mirror descent type optimization; The second stage (line 9) is *Dual Update*: we update dual variable Y^k based on the constraint violation induced by previous policy π^k; The third stage (line 10) is *Policy Evaluation*: we optimistically evaluate newly obtained policy via the least-squares policy evaluation with an additional UCB bonus term for exploration.

3.1 Policy Improvement

In the k-th episode, a natural attempt of obtaining a policy π^k is to solve a Lagrangian-based policy optimization problem,

$$\underset{\pi \in \Delta(\mathcal{A}|\mathcal{S},H)}{\text{maximize}} \; \mathcal{L}(\pi, Y^{k-1}) := V_{r,1}^{\pi}(x_1) - Y^{k-1}(b - V_{g,1}^{\pi}(x_1))$$

where $\mathcal{L}(\pi, Y)$ is the Lagrangian and the dual variable $Y^{k-1} \geq 0$ is from the last episode; we show that Y^{k-1}

can be updated efficiently in Section 3.2. This type update also finds in references [45, 53, 52, 61]. They rely on an oracle solver, e.g., Q-learning [30], proximal policy optimization [57], or trust region policy optimization [56], to deliver a near-optimal policy, making overall algorithmic complexity expensive. Hence, they are not suitable for online use. In contrast, this work utilizes RL problem structure and shows that only an easily-computable proximal step is sufficient for efficiently achieving near-optimal performance.

Recall symbol $\diamond = r$ or g. Via the performance difference lemma, we can expand value functions $V_{\diamond,1}^{\pi}(x_1)$ at the previously known policy π^{k-1},

$$V_{\diamond,1}^{\pi}(x_1) \;=\; V_{\diamond,1}^{\pi^{k-1}}(x_1^k) + \mathbb{E}_{\pi^{k-1}}\left[\sum_{h=1}^{H}\left\langle Q_{\diamond,h}^{\pi}(x_h,\cdot),(\pi_h-\pi_h^{k-1})(\,\cdot\,|\,x_h)\right\rangle\right]$$

where $\mathbb{E}_{\pi^{k-1}}$ is taken over the random state-action sequence $\{(x_h,a_h)\}_{h=1}^H$. Thus, we introduce an approximation of $V_{\diamond,1}^{\pi}(x_1)$ for any state-action sequence $\{(x_h,a_h)\}_{h=1}^H$ induced by π,

$$L_{\diamond}^{k-1}(\pi) \;=\; V_{\diamond,1}^{k-1}(x_1) + \sum_{h=1}^{H}\left\langle Q_{\diamond,h}^{k-1}(x_h,\cdot),(\pi_h-\pi_h^{k-1})(\,\cdot\,|\,x_h)\right\rangle$$

where $V_{\diamond,h}^{k-1}$ and $Q_{\diamond,h}^{k-1}$ can be estimated from an optimistic policy evaluation that will be discussed in Section 3.3. With this notion, in each episode, instead of solving a Lagrangian-based policy optimization, we perform a simple policy update in online mirror descent fashion,

$$\underset{\pi\,\in\,\Delta(\mathcal{A}|\mathcal{S},H)}{\text{maximize}}\;\; L_r^{k-1}(\pi) \,-\, Y^{k-1}\big(b-L_g^{k-1}(\pi)\big) - \frac{1}{\alpha}\sum_{h=1}^{H} D\big(\pi_h(\,\cdot\,|\,x_h)\,|\,\widetilde{\pi}_h^{k-1}(\,\cdot\,|\,x_h)\big)$$

where $\widetilde{\pi}_h^{k-1}(\,\cdot\,|\,x_h) \,=\, (1-\theta)\,\pi_h^{k-1}(\,\cdot\,|\,x_h)+\theta\,\mathrm{Unif}(\mathcal{A})$ is a mixed policy of the previous one and the uniform distribution $\mathrm{Unif}(\mathcal{A})$ with $\theta\in(0,1]$. The constant $\alpha>0$ is a trade-off parameter, $D(\pi\,|\,\widetilde{\pi}^{k-1})$ is the KL divergence between π and $\widetilde{\pi}^{k-1}$ in which π is absolutely continuous in $\widetilde{\pi}^{k-1}$. The policy mixing step ensures such absolute continuity and implies uniformly bounded KL divergence; see Lemma 15 in Appendix F. Ignoring other π-irrelevant terms, we update π^k in terms of previous policy π^{k-1} by

$$\underset{\pi\in\Delta(\mathcal{A}|\mathcal{S},H)}{\text{argmax}}\sum_{h=1}^{H}\big\langle(Q_{r,h}^{k-1}+Y^{k-1}Q_{g,h}^{k-1})(x_h,\,\cdot\,),\pi_h(\,\cdot\,|\,x_h)\big\rangle - \frac{1}{\alpha}\sum_{h=1}^{H}D\big(\pi_h(\,\cdot\,|\,x_h)\,|\,\widetilde{\pi}_h^{k-1}(\,\cdot\,|\,x_h)\big).$$

Since the above update is separable over H steps, we can update the policy π^k as line 6 in Algorithm 1, a closed-form solution for any step $h\in[H]$. If we set $Y^{k-1}=0$ and $\theta=0$, the above update reduces to one step in an optimistic proximal policy optimization [20]. The idea of KL-divergence regularization in policy optimization has been widely used in many unconstrained scenarios [39, 57, 56, 67, 47]. Our method is distinct in that it is based on the performance difference lemma and the optimistically estimated value functions.

Algorithm 1 Optimistic Primal-Dual Proximal Policy OPtimization (OPDOP)

1: **Initialization**: Let $\{Q_{r,h}^0, Q_{g,h}^0\}_{h=1}^H$ be zero functions, $\{\pi_h^0\}_{h\,\in\,[H]}$ be uniform distributions on $\mathcal{A}$, $V_{g,1}^0$ be b, Y^0 be 0, χ be $2H/\gamma$, $\alpha,\eta>0,\theta\in(0,1]$.
2: **for** episode $k=1,\ldots,K+1$ **do**
3: Set the initial state $x_1^k = x_1$.
4: **for** step $h=1,2,\ldots,H$ **do**
5: Mix the policy

$$\widetilde{\pi}_h^{k-1}(\cdot|\cdot) \;=\; (1-\theta)\pi_h^{k-1}(\cdot|\cdot) \,+\, \theta\,\mathrm{Unif}(\mathcal{A}).$$

6: Update the policy

$$\pi_h^k(\cdot|\cdot) \propto \widetilde{\pi}_h^{k-1}(\cdot|\cdot)\,\mathrm{e}^{\left(\alpha\,\left(Q_{r,h}^{k-1}+Y^{k-1}Q_{g,h}^{k-1}\right)(\cdot,\cdot)\right)}.$$

7: Take an action $a_h^k \sim \pi_h^k(\,\cdot\,|\,x_h^k\,)$ and recieve reward/utility, $r_h(x_h^k,a_h^k)$, $g_h(x_h^k,a_h^k)$.
8: Observe the next state x_{h+1}^k.
9: Update the dual variable Y^k by

$$Y^k \;=\; \mathrm{Proj}_{[\,0,\,\chi\,]}\big(\,Y^{k-1} \,+\, \eta\,(b-V_{g,1}^{k-1}(x_1))\,\big).$$

10: Estimate the action-value or value functions $\{Q_{r,h}^k(\cdot,\cdot),Q_{g,h}^k(\cdot,\cdot),V_{g,h}^k(\cdot)\}_{h=1}^H$ via

$$\mathrm{LSTD}\Big(\big\{x_h^\tau,a_h^\tau,r_h(x_h^\tau,a_h^\tau),g_h(x_h^\tau,a_h^\tau)\big\}_{h,\tau=1}^{H,k}\Big).$$

3.2 Dual Update

To infer the constraint violation for the dual update, we estimate $V_{g,1}^{\pi^k}(x_1)$ via an optimistic policy evaluation by $V_{g,1}^{k-1}(x_1)$ that is discussed in Section 3.3. We update the Lagrange multiplier Y by moving Y^k to the direction of minimizing the Lagrangian $\mathcal{L}(\pi,Y)$ over $Y\geq 0$ in line 9 of Algorithm 1, where $\eta>0$ is a stepsize and $\mathrm{Proj}_{[\,0,\,\chi\,]}$ is a projection onto $[0,\chi]$ with an upper bound χ on Y^k. By Lemma 1, we choose $\chi=2H/\gamma\geq 2Y^\star$ so that projection interval $[\,0,\,\chi\,]$ includes the optimal dual variable $Y^\star$. This type design also finds in references [29, 51].

The dual update works as a trade-off between the reward maximization and the constraint violation reduction. If the current policy π^k satisfies the approximated constraint, i.e., $b - L_g^{k-1}(\pi^k) \leq 0$, we put less weight on the action-value function associated with the utility and maximize the reward; otherwise, we sacrifice the reward a bit to satisfy the constraint. The dual update has a similar use in dealing with constraints in CMDPs, e.g., Lagrangian-based actor-critic [22, 46], and online constrained optimization [73, 68, 75]. In contrast, we handle the dual update via the optimistic policy evaluation, yielding a simple, but efficient estimation on the constraint violation.

Algorithm 2 Least-Squares Temporal Difference (LSTD) with UCB exploration

1: **Input**: $\{x_h^\tau, a_h^\tau, r_h(x_h^\tau, a_h^\tau), g_h(x_h^\tau, a_h^\tau)\}_{h,\tau=1}^{H,k}$.

2: **Initialization**: Set $\{V_{r,H+1}^k, V_{g,H+1}^k\}$ be zero functions and $\lambda = 1, \beta = O(\sqrt{dH^2 \log(dT/p)})$.

3: **for** step $h = H, H-1, \cdots, 1$ **do** $\quad \triangleright\ \diamond = r,\ g$

4: $\quad \Lambda_{\diamond,h}^k = \sum_{\tau=1}^{k-1} \phi_{\diamond,h}^\tau(x_h^\tau, a_h^\tau)\phi_{\diamond,h}^\tau(x_h^\tau, a_h^\tau)^\top + \lambda I.$

5: $\quad w_{\diamond,h}^k = (\Lambda_{\diamond,h}^k)^{-1} \sum_{\tau=1}^{k-1} \phi_{\diamond,h}^\tau(x_h^\tau, a_h^\tau) V_{\diamond,h+1}^\tau(x_{h+1}^\tau).$

6: $\quad \phi_{\diamond,h}^k(\cdot,\cdot) = \int_{\mathcal{S}} \psi(\cdot,\cdot,x') V_{\diamond,h+1}^k(x')dx'.$

7: $\quad \Gamma_{\diamond,h}^k(\cdot,\cdot) = \beta(\phi_{\diamond,h}^k(\cdot,\cdot)^\top (\Lambda_{\diamond,h}^k)^{-1} \phi_{\diamond,h}^k(\cdot,\cdot))^{1/2}.$

8: $\quad \Lambda_h^k = \sum_{\tau=1}^{k-1} \varphi(x_h^\tau, a_h^\tau)\varphi(x_h^\tau, a_h^\tau)^\top + \lambda I.$

9: $\quad u_{\diamond,h}^k = (\Lambda_h^k)^{-1} \sum_{\tau=1}^{k-1} \varphi(x_h^\tau, a_h^\tau) \diamond_h (x_h^\tau, a_h^\tau).$

10: $\quad \Gamma_h^k(\cdot,\cdot) = \beta(\varphi(\cdot,\cdot)^\top (\Lambda_h^k)^{-1} \varphi(\cdot,\cdot))^{1/2}.$

11: $\quad Q_{\diamond,h}^k(\cdot,\cdot) = \min\left(H-h+1, \varphi(\cdot,\cdot)^\top u_{\diamond,h}^k + \phi_{\diamond,h}^k(\cdot,\cdot)^\top w_{\diamond,h}^k + (\Gamma_h^k + \Gamma_{\diamond,h}^k)(\cdot,\cdot)\right)^+.$

12: $\quad V_{\diamond,h}^k(\cdot) = \left\langle Q_{\diamond,h}^k(\cdot,\cdot), \pi_h^k(\cdot|\cdot)\right\rangle_{\mathcal{A}}.$

13: **Return**: $\{Q_{\diamond,h}^k(\cdot,\cdot), V_{\diamond,h}^k(\cdot,\cdot)\}_{h=1}^H$.

3.3 Policy Evaluation

The last stage of the kth episode takes the Least-Squares Temporal Difference (LSTD) [17, 16, 44, 42] to evaluate the policy π^k based on previous $k-1$ historical trajectories. For each step $h \in [H]$, instead of $\mathbb{P}_h V_{r,h+1}^{\pi^k}$ in the Bellman equations (1), we estimate $\mathbb{P}_h V_{r,h+1}^k$ by $(\phi_{r,h}^k)^\top w_{r,h}^k$ where $w_{r,h}^k$ is updated by the minimizer of the regularized least-squares problem over w,

$$\sum_{\tau=1}^{k-1} \left(V_{r,h+1}^\tau(x_{h+1}^\tau) - \phi_{r,h}^\tau(x_h^\tau, a_h^\tau)^\top w\right)^2 + \lambda \|w\|_2^2 \quad (4)$$

where $\phi_{r,h}^\tau(\cdot,\cdot) := \int_{\mathcal{S}} \psi(\cdot,\cdot,x') V_{r,h+1}^\tau(x')dx'$, $V_{r,h+1}^\tau(\cdot) = \langle Q_{r,h+1}^\tau(\cdot,\cdot), \pi_{h+1}^\tau(\cdot\,|\,\cdot)\rangle_{\mathcal{A}}$ for $h \in [H-1]$ and $V_{H+1}^\tau = 0$, and $\lambda > 0$ is the regularization parameter. Similarly, we estimate $\mathbb{P}_h V_{g,h+1}^k$ by $(\phi_{g,h}^k)^\top w_{g,h}^k$. We display the least-squares solution in line 4–6 of Algorithm 2, where symbol $\diamond = r$ or g. We also estimate $r_h(\cdot,\cdot)$ by $\varphi^\top u_{r,h}^k$, where $u_{r,h}^k$ is updated by the minimizer of another regularized least-squares problem,

$$\sum_{\tau=1}^{k-1} \left(r_h(x_h^\tau, a_h^\tau) - \varphi(x_h^\tau, a_h^\tau)^\top u\right)^2 + \lambda \|u\|_2^2 \quad (5)$$

where $\lambda > 0$ is the regularization parameter. Similarly, we estimate $g_h(\cdot,\cdot)$ by $\varphi^\top u_{g,h}^k$. The least-squares solutions lead to line 8–9 of Algorithm 2.

After obtaining estimates of $\mathbb{P}_h V_{\diamond,h+1}^k$ and $\diamond_h(\cdot,\cdot)$ for $\diamond = r$ or g, we update the estimated action-value function $\{Q_{\diamond,h}^k\}_{h=1}^H$ iteratively in line 11 of Algorithm 2, where $\varphi^\top u_{\diamond,h}^k$ is an estimate of $\diamond_h$ and $(\phi_{\diamond,h}^k)^\top w_{\diamond,h}^k$ is an estimate of $\mathbb{P}_h V_{\diamond,h+1}^k$; we add UCB bonus terms $\Gamma_h^k(\cdot,\cdot), \Gamma_{\diamond,h}^k(\cdot,\cdot)$: $\mathcal{S} \times \mathcal{A} \to \mathbb{R}^+$ so that

$$\varphi^\top u_{\diamond,h}^k + \Gamma_h^k \quad \text{and} \quad (\phi_{\diamond,h}^k)^\top w_{\diamond,h}^k + \Gamma_{\diamond,h}^k$$

all become their upper confidence bounds. Here, the bonus terms take $\Gamma_h^k = \beta(\varphi^\top(\Lambda_h^k)^{-1}\varphi)^{1/2}$ and $\Gamma_{\diamond,h}^k = \beta((\phi_{\diamond,h}^k)^\top(\Lambda_{\diamond,h}^k)^{-1}\phi_{\diamond,h}^k)^{1/2}$ and we leave the parameter $\beta > 0$ to be tuned later. Moreover, the bounded reward/utility $\diamond_h \in [0,1]$ implies $Q_{\diamond,h}^k \in [0, H-h+1]$.

We remark the computational efficiency of Algorithm 1. For the time complexity, since line 6 is a scalar update, they need $O(d|\mathcal{A}|T)$ time. A dominating calculation is from lines 5/9 in Algorithm 2. If we use the Sherman–Morrison formula for computing $(\Lambda_h^k)^{-1}$, it takes $O(d^2T)$ time. Another important calculation is the integration from line 6 in Algorithm 2. We can either compute it analytically if it is tractable or approximate it via the Monte Carlo integration [79] that assumes polynomial time. Therefore, the time complexity is $O(\text{poly}(d)|\mathcal{A}|T)$ in total. For the space complexity, we don't need to store policy since it is recursively calculated via line 6 of Algorithm 1. By updating Y^k, Λ_h^k, $\Lambda_{\diamond,h}^k$, $w_{\diamond,h}^k$, $u_{\diamond,h}^k$, and $\diamond_h(x_h^k, a_h^k)$ recursively, it takes $O((d^2 + |\mathcal{A}|)H)$ space.

4 Regret and Constraint Violation Analysis

We now prove that the regret and the constraint violation for Algorithm 1 are sublinear in $T := KH$, the total number of steps taken by the algorithm, where K is the total number of episodes and H is the episode horizon. We recall that $|\mathcal{A}|$ is the cardinality of action space $\mathcal{A}$ and d is the dimension of the feature map.

Theorem 1 (Linear Kernal MDP: Regret and Constraint Violation)**.** *Let Assumptions 1 and 2 hold. Fix $p \in (0,1)$. We set $\alpha = \sqrt{\log|\mathcal{A}|/(H^2K)}$, $\beta = C_1\sqrt{dH^2\log(dT/p)}$, $\eta = 1/\sqrt{K}$, $\theta = 1/K$, and $\lambda = 1$ in Algorithm 1, where C_1 is an absolute constant. Suppose $\log|\mathcal{A}| = O\left(d^2\log^2(dT/p)\right)$. Then, with probability $1-p$, the regret and the constraint violation in (3) satisfy*

$$
\begin{aligned}
\text{Regret}(K) \;&\leq\; C\, dH^{2.5}\sqrt{T}\,\log\left(\tfrac{dT}{p}\right)\\
[\text{Violation}(K)]_+ \;&\leq\; C'\, dH^{2.5}\sqrt{T}\log\left(\tfrac{dT}{p}\right)
\end{aligned}
$$

where C and C' are absolute constants.

The above result establishes that Algorithm 1 enjoys an $\widetilde{O}(dH^{2.5}\sqrt{T})$ regret and an $\widetilde{O}(dH^{2.5}\sqrt{T})$ constraint violation if we set algorithm parameters $\{\alpha,\beta,\eta,\theta,\lambda\}$ properly. Our results have the optimal dependence on the total number of steps T up to some logarithmic factors. The d dependence occurs due to the uniform concentration for controlling the fluctuations in the least-squares policy evaluation. This matches the existing bounds in the linear MDP setting without any constraints [20, 8, 79]. Our bounds differ from them only by H dependence, which is a price introduced by the uniform bound on the constraint violation. It is noticed that our algorithm works for bandit feedback of reward/utility functions after each episode.

Regarding safe exploration, our violation bound provides finite-time convergence to the feasibility region defined by constraints. In the interaction with an unknown environment, the UCB exploration in the utility value function adds optimism towards constraint satisfaction. The dual update regularizes the policy improvement for governing actual constraint violation. Our regret and violation bounds readily lead to PAC guarantees [36]. Compared to most recent references [26, 69, 21, 77], our algorithm is sample-efficient in exploration and does not need simulations of policy.

We remark the tabular setting for Algorithm 1; see Appendix C for details. The tabular CMDP is a special case of Assumption 2 by taking canonical bases as feature mappings; see them in Section 5. The feature map has dimension $d = |\mathcal{S}|^2|\mathcal{A}|$ and thus Theorem 1 automatically provides $O(|\mathcal{S}|^2|\mathcal{A}|H^{2.5}\sqrt{T})$ regret and constraint violation for the tabular CMDPs. The $d = |\mathcal{S}|^2|\mathcal{A}|$ dependence relies on the least-squares policy evaluation and it can be improved via other optimistic policy evolution methods if we are limited to the tabular case. We provide such results in Section 5.

4.1 Proof Outline of Theorem 1

We sketch the proof for Theorem 1. We state key lemmas and delay their full versions and proofs to Appendix B. In what follows, we fix $p \in (0,1)$ and use the shorthand w.p. for with probability.

Regret Analysis. We take a regret decomposition,

$$
\begin{aligned}
\text{Regret}(K) \;=\;& \underbrace{\sum_{k=1}^{K}\left(V_{r,1}^{\pi^\star}(x_1) - V_{r,1}^{k}(x_1)\right)}_{\text{(R.I)}}\\
&+ \underbrace{\sum_{k=1}^{K}\left(V_{r,1}^{k}(x_1) - V_{r,1}^{\pi^k}(x_1)\right)}_{\text{(R.II)}}
\end{aligned}
$$

where $\pi^\star$ is an optimal policy in hindsight, and $V_{r,1}^k(x_1)$ is estimated via our optimistic policy evaluation given by Algorithm 2. Since we use $V_{r,h+1}^k$ to estimate $V_{r,h+1}^{\pi^k}$, it leads a model prediction error in the Bellman equations, $\iota_{r,h}^k := r_h^k + \mathbb{P}_h V_{r,h+1}^k - Q_{r,h}^k$; similarly define $\iota_{g,h}^k$. In Appendix D.3, the UCB optimism of $\iota_{\diamond,h}^k$ with $\diamond = r$ or g, shows that or any $(k,h) \in [K]\times[H]$ and $(x,a) \in \mathcal{S}\times\mathcal{A}$, w.p. $1-p/2$, we have

$$
-2(\Gamma_h^k + \Gamma_{\diamond,h}^k)(x,a) \;\leq\; \iota_{\diamond,h}^k(x,a) \;\leq\; 0.
$$

By assumptions of Theorem 1, the policy improvement in line 6 of Algorithm 1 yields Lemma 1, depicting weighted total differences of estimates $V_{r,1}^k(x_1)$, $V_{g,1}^k(x_1)$ to the optimal ones.

Lemma 1 (Policy Improvement: Primal-Dual Mirror Descent Step)**.** *Let assumptions of Theorem 1 hold. Then,*

$$
\begin{aligned}
\text{(R.I)} \;\leq\; & -\sum_{k=1}^{K} Y^k\left(V_{g,1}^{\pi^\star}(x_1) - V_{g,1}^k(x_1)\right)\\
&+ \sum_{k=1}^{K}\sum_{h=1}^{H}\mathbb{E}_{\pi^\star}\left[\iota_{r,h}^k(x_h,a_h) + Y^k\iota_{g,h}^k(x_h,a_h)\right]\\
&+ O\left(H^{2.5}\sqrt{T\log|\mathcal{A}|}\right).
\end{aligned}
$$

Lemma 1 displays coupling between the regret (R.I) and the constraint. This coupling also finds in online convex optimization [49, 75, 68, 41] and CMDP problems [29]. The proof of Lemma 1 takes a primal-dual mirror descent type analysis of line 6 of Algorithm 1, using the performance difference lemma.

Via the dual update in line 9 of Algorithm 1, we can verify that the second total differences $-\sum_{k=1}^{K} Y^k\left(V_{g,1}^{\pi^\star}(x_1) - V_{g,1}^k(x_1)\right)$ scales $O(\sqrt{K})$. Together with a decomposition of (R.II),

$$
\text{(R.II)} \;=\; -\sum_{k=1}^{K}\sum_{h=1}^{H}\iota_{r,h}^k(x_h^k,a_h^k) \;+\; M_{r,H,2}^K
$$

where $M_{r,H,2}^K$ is a martingale, we now have Lemma 2.

Lemma 2. *Let assumptions of Theorem 1 hold. Then,*

$$\begin{aligned}\text{Regret}(K) \;\leq\; & \sum_{k=1}^{K}\sum_{h=1}^{H}\big(\mathbb{E}_{\pi^\star}[\iota_{r,h}^k(x_h,a_h)] - \iota_{r,h}^k(x_h^k,a_h^k)\big)\\ & + M_{r,H,2}^K \,+\, O\big(H^{2.5}\sqrt{T\log|\mathcal{A}|}\big).\end{aligned}$$

Finally, we note that $M_{r,H,2}^K$ is a martingale that scales as $O(H\sqrt{T})$ via the Azuma-Hoeffding inequality. For the model prediction error, we use the UCB optimism and apply the elliptical potential lemma.

Lemma 3. *Let assumptions of Theorem 1 hold. Then,*

$$\begin{aligned}&\sum_{k=1}^{K}\sum_{h=1}^{H}\big(\mathbb{E}_{\pi^\star}[\iota_{r,h}^k(x_h,a_h)] - \iota_{r,h}^k(x_h^k,a_h^k)\big)\\ &\leq\; O\big(dH^{1.5}\sqrt{T\log{(K)}\log{(dT/p)}}\big),\ \text{w.p.}\,1-p/2.\end{aligned}$$

Lemma 4. *Let assumptions of Theorem 1 hold. Then,*

$$\big|M_{r,H,2}^K\big| \;\leq\; 4H\sqrt{T\log{(4/p)}},\ \text{w.p.}\,1-p/2.$$

Finally, we apply probability bounds from Lemmas 3 and 4 to Lemma 2 to get our regret bound.

Constraint Violation Analysis. We take a violation decomposition,

$$\begin{aligned}\text{Violation}(K) \;=\; & \sum_{k=1}^{K}\big(b - V_{g,1}^k(x_1)\big)\\ & + \underbrace{\sum_{k=1}^{K}\big(V_{g,1}^k(x_1) - V_{g,1}^{\pi^k}(x_1)\big)}_{\text{(V.II)}}.\end{aligned}$$

We begin with the policy improvement in line 6 of Algorithm 1 to refine Lemma 1 as Lemma 5.

Lemma 5 (Policy Improvement: Refined Primal-Dual Mirror Descent Step)**.** *Let assumptions of Theorem 1 hold. Then, for any $Y \in [0,\chi]$,*

$$\text{(R.I)} + Y\sum_{k=1}^{K}\big(b - V_{g,1}^k(x_1)\big) \;\leq\; O\big(H^{2.5}\sqrt{T\log|\mathcal{A}|}\big).$$

Lemma 5 removes the dual update Y^k in the second total differences in Lemma 1. We prove Lemma 5 by combining Lemma 1 with the UCB optimism and a change of variable of Y^k for the dual update.

Similar to (R.II), we also have

$$\text{(V.II)} \;=\; -\sum_{k=1}^{K}\sum_{h=1}^{H}\iota_{g,h}^k(x_h^k,a_h^k) \,+\, M_{g,H,2}^K$$

where $M_{g,H,2}^K$ is a martingale. By adding (V.II) to the inequality in Lemma 5 with multiplier $Y \geq 0$, and also adding (R.II) to it,

$$\begin{aligned}&\sum_{k=1}^{K}\big(V_{r,1}^{\pi^\star}(x_1) - V_{r,1}^{\pi^k}(x_1)\big) \,+\, Y\sum_{k=1}^{K}\big(b - V_{g,1}^{\pi^k}(x_1)\big)\\ &\leq\; -\sum_{k=1}^{K}\sum_{h=1}^{H}\big(\iota_{r,h}^k(x_h^k,a_h^k) + Y\iota_{g,h}^k(x_h^k,a_h^k)\big)\\ &\quad + O\big(H^{2.5}\sqrt{T\log|\mathcal{A}|}\big) \,+\, M_{r,H,2}^K \,+\, YM_{g,H,2}^K\end{aligned}$$

Then, we take $Y = 0$ if $\sum_{k=1}^{K}\big(b - V_{g,1}^{\pi^k}(x_1)\big) \leq 0$; otherwise $Y = \chi$, w.p. $1-p$, we have,

$$\begin{aligned}&\big(V_{r,1}^{\pi^\star}(x_1) - V_{r,1}^{\pi'}(x_1)\big) + \chi\big[b - V_{g,1}^{\pi'}(x_1)\big]_+\\ &\qquad\leq\; O\big(dH^{2.5}\sqrt{T}\log{(dT/p)}\,/K\big)\end{aligned}$$

where $V_{r,1}^{\pi'}(x_1) = \frac{1}{K}\sum_{k=1}^{K}V_{r,1}^{\pi^k}(x_1)$ and $V_{g,1}^{\pi'}(x_1) = \frac{1}{K}\sum_{k=1}^{K}V_{g,1}^{\pi^k}(x_1)$ for some existing policy π'. Here, we bound $\Gamma_h^k + \Gamma_{\diamond,h}^k$ and $M_{\diamond,H,2}^K$ as done in Lemmas 3 and 4.

Last, by the strong duality in Lemma 1, we apply the constraint violation bound from constrained optimization that is stated in Lemma 10 in Appendix E,

$$\big[b - V_{g,1}^{\pi'}(x_1)\big]_+ \;\leq\; O\big(dH^{2.5}\sqrt{T}\log{(dT/p)}\,/(\chi K)\big)$$

which gives our desired violation bound.

Algorithm 3 Optimistic Policy Evaluation (OPE)

Input: $\{x_h^\tau, a_h^\tau, r_h(x_h^\tau,a_h^\tau), g_h(x_h^\tau,a_h^\tau)\}_{h,\tau=1}^{H,k}$.
Initialization: Set $\{V_{r,H+1}^k, V_{g,H+1}^k\}$ be zero functions, and $\lambda = 1$, $\beta = C_1 H\sqrt{|\mathcal{S}|\log(|\mathcal{S}||\mathcal{A}|T/p)}$.
for step $h = H, H-1, \cdots, 1$ **do** ▷ $\diamond = r,\, g$
  Compute counters $n_h^k(x,a,x')$ and $n_h^k(x,a)$ via (7) for all $(x,a,x') \in \mathcal{S}\times\mathcal{A}\times\mathcal{S}$ and $(x,a) \in \mathcal{S}\times\mathcal{A}$.
  Estimate reward/utility functions $\widehat{r}_h^k$, $\widehat{g}_h^k$ via (8) for all $(x,a) \in \mathcal{S}\times\mathcal{A}$.
  Estimate transition $\widehat{\mathbb{P}}_h^k$ via (9) for all $(x,a,x') \in \mathcal{S}\times\mathcal{A}\times\mathcal{S}$, and take bonus $\Gamma_h^k = \beta\,\big(n_h^k(x,a)+\lambda\big)^{-1/2}$ for all $(x,a) \in \mathcal{S}\times\mathcal{A}$.
  $Q_{\diamond,h}^k(\cdot,\cdot) \;=\; \min\big(H-h+1,\, \widehat{\diamond}_h^k(\cdot,\cdot) + \sum_{x'\in\mathcal{S}}\widehat{\mathbb{P}}_h(x'\,|\,\cdot,\cdot)V_{\diamond,h+1}^k(x') + 2\Gamma_h^k(\cdot,\cdot),\ \big)^+$.
  $V_{\diamond,h}^k(\cdot) \;=\; \big\langle Q_{\diamond,h}^k(\cdot,\cdot), \pi_h^k(\cdot\,|\,\cdot)\big\rangle_{\mathcal{A}}$.
Return: $\{Q_{r,h}^k(\cdot,\cdot), Q_{g,h}^k(\cdot,\cdot)\}_{h=1}^H$.

5 Further Results on Tabular Case

The tabular CMDP$(\mathcal{S},\mathcal{A},H,\mathbb{P},r,g)$ is a special case of Assumption 2 with $|\mathcal{S}| < \infty$ and $|\mathcal{A}| < \infty$. Let $d_1 = |\mathcal{S}|^2|\mathcal{A}|$ and $d_2 = |\mathcal{S}||\mathcal{A}|$. We take the following feature maps $\psi(x,a,x') \in \mathbb{R}^{d_1}$, $\varphi(x,a) \in \mathbb{R}^{d_2}$, and parameter vectors,

$$\begin{aligned} &\psi(x,a,x') = \mathbf{e}_{(x,a,x')},\, \theta_h = \mathbb{P}_h(\,\cdot\,,\cdot\,,\cdot\,) \\ &\varphi(x,a) = \mathbf{e}_{(x,a)},\, \theta_{r,h} = r_h(\,\cdot\,,\cdot\,),\, \theta_{g,h} = g_h(\,\cdot\,,\cdot\,) \end{aligned} \tag{6}$$

where $\mathbf{e}_{(x,a,x')}$ is a canonical basis of $\mathbb{R}^{d_1}$ associated with (x,a,x') and $\theta_h = \mathbb{P}_h(\,\cdot\,,\cdot\,,\cdot\,)$ reads that for any $(x,a,x') \in \mathcal{S}\times\mathcal{A}\times\mathcal{S}$, the (x,a,x')th entry of θ_h is $\mathbb{P}(x'\,|\,x,a)$; similarly we define $\mathbf{e}_{(x,a)}$, $\theta_{r,h}$, and $\theta_{g,h}$. We can verify that $\|\theta_h\| \leq \sqrt{d_1}$, $\|\theta_{r,h}\| \leq \sqrt{d_2}$, $\|\theta_{g,h}\| \leq \sqrt{d_2}$, and for any V: $\mathcal{S} \to [0,H]$ and any $(x,a) \in \mathcal{S}\times\mathcal{A}$, we have $\left\|\sum_{x'\in\mathcal{S}} \psi(x,a,x')V(x')\right\| \leq \sqrt{|\mathcal{S}|}H \leq \sqrt{d_1}H$. Therefore, we take $d := \max(d_1,d_2) = |\mathcal{S}|^2|\mathcal{A}|$ in Assumption (2) for the tabular case.

The proof of Theorem 1 is generic, since it is ready to achieve sublinear regret and constraint violation bounds as long as the policy evaluation is sample-efficient, e.g., the UCB design of 'optimism in the face of uncertainty.' In what follows, we introduce another efficient policy evaluation for line 10 of Algorithm 1 in the tabular case. Let us first introduce some notation. For any $(h,k) \in [H]\times[K]$, any $(x,a,x') \in \mathcal{S}\times\mathcal{A}\times\mathcal{S}$, and any $(x,a) \in \mathcal{S}\times\mathcal{A}$, we define two visitation counters $n_h^k(x,a,x')$ and $n_h^k(x,a)$ at step h in episode k,

$$\begin{aligned} n_h^k(x,a,x') &= \sum_{\tau=1}^{k-1} \mathbf{1}\{(x,a,x') = (x_h^\tau, a_h^\tau, a_{h+1}^\tau)\} \\ n_h^k(x,a) &= \sum_{\tau=1}^{k-1} \mathbf{1}\{(x,a) = (x_h^\tau, a_h^\tau)\}. \end{aligned} \tag{7}$$

This allows us to estimate reward function r, utility function g, and transition kernel $\mathbb{P}_h$ for episode k by

$$\begin{aligned} \widehat{r}_h^k(x,a) &= \sum_{\tau=1}^{k-1} \frac{\mathbf{1}\{(x,a) = (x_h^\tau, a_h^\tau)\} r_h(x_h^\tau, a_h^\tau)}{n_h^k(x,a)+\lambda} \\ \widehat{g}_h^k(x,a) &= \sum_{\tau=1}^{k-1} \frac{\mathbf{1}\{(x,a) = (x_h^\tau, a_h^\tau)\} g_h(x_h^\tau, a_h^\tau)}{n_h^k(x,a)+\lambda} \end{aligned} \tag{8}$$

$$\widehat{\mathbb{P}}_h^k(x'\,|\,x,a) = \frac{n_h^k(x,a,x')}{n_h^k(x,a)+\lambda} \tag{9}$$

for all $(x,a,x') \in \mathcal{S}\times\mathcal{A}\times\mathcal{S}$, $(x,a) \in \mathcal{S}\times\mathcal{A}$ where $\lambda > 0$ is the regularization parameter. Moreover, we introduce the bonus term Γ_h^k: $\mathcal{S}\times\mathcal{A} \to \mathbb{R}$, $\Gamma_h^k(x,a) = \beta\left(n_h^k(x,a)+\lambda\right)^{-1/2}$ which adapts the counter-based bonus terms in the literature [9, 36], where $\beta > 0$ is to be determined later.

Using the estimated transition kernels $\{\widehat{\mathbb{P}}_h^k\}_{h=1}^H$, the estimated reward/utility functions $\{\widehat{r}_h^k, \widehat{g}_h^k\}_{h=1}^H$, and the bonus terms $\{\Gamma_h^k\}_{h=1}^H$, we now can estimate the action-value function via line 7 of Algorithm 3 for any $(x,a) \in \mathcal{S}\times\mathcal{A}$, where $\diamond = r$ or g. Thus, $V_{\diamond,h}^k(x) = \langle Q_{\diamond,h}^k(x,\cdot), \pi_h^k(\cdot\,|\,x)\rangle_{\mathcal{A}}$. We summarize the above procedure in Algorithm 3. Using already estimated $\{Q_{r,h}^k(\cdot,\cdot), Q_{g,h}^k(\cdot,\cdot)\}_{h=1}^H$, we execute the policy improvement and the dual update in Algorithm 1.

As in Theorem 1, we provide theoretical guarantees in Theorem 2; see Appendix C.2 for the proof. Theorem 2 improves $(|\mathcal{S}|,|\mathcal{A}|)$ dependence in Theorem 1 for the tabular case and also matches $|\mathcal{S}|$ dependence in references [29, 55]. It is worthy mentioning our Algorithm 1 is generic in handling an infinite state space.

Theorem 2 (Tabular Case: Regret and Constraint Violation). *Let Assumption 1 hold and let Assumption 2 hold with feature maps (6). Fix $p \in (0,1)$. In Algorithm 1, we set $\alpha = \sqrt{\log|\mathcal{A}|/(H^2K)}$, $\beta = C_1 H\sqrt{|\mathcal{S}|\log(|\mathcal{S}||\mathcal{A}|T/p)}$, $\eta = 1/\sqrt{K}$, $\theta = 1/K$, and $\lambda = 1$ where C_1 is an absolute constant. Then, with probability $1-p$, the regret and the constraint violation in (3) satisfy*

$$\begin{aligned} \text{Regret}(K) &\leq C|\mathcal{S}|\sqrt{|\mathcal{A}|H^5T}\log\left(\tfrac{|\mathcal{S}||\mathcal{A}|T}{p}\right) \\ [\text{Violation}(K)]_+ &\leq C'|\mathcal{S}|\sqrt{|\mathcal{A}|H^5T}\log\left(\tfrac{|\mathcal{S}||\mathcal{A}|T}{p}\right) \end{aligned}$$

where C and C' are absolute constants.

6 Concluding Remarks

We have developed a provably efficient safe reinforcement learning algorithm in the linear MDP setting. The algorithm extends the proximal policy optimization to CMDPs by incorporating the UCB exploration. We prove that the proposed algorithm achieves an $\widetilde{O}(\sqrt{T})$ regret and an $\widetilde{O}(\sqrt{T})$ constraint violation under mild conditions, where T is the total number of steps taken by the algorithm. Our algorithm works in the setting where reward/utility functions are given by bandit feedback. To the best of our knowledge, our algorithm is the first provably efficient online policy optimization algorithm for CMDPs in the function approximation setting.

Mathematically, our algorithm framework allows reward/utility functions to be adversarial. We believe that approaches from the adversarial MDP literature allow us to derive similar regret and constraint violation bounds, although we leave it as future work. Beyond linear kernel MDPs, the UCB exploration has previously been applied for other types of MDPs, e.g., factored MDPs, or infinite-horizon MDPs. It remains to be seen if these are extendable for CMDPs in similar settings. In practice, we often encounter general function approximation beyond linear functions, e.g., neural nets. It would be useful to design provably efficient exploration algorithms for CMDPs with general function approximation.

Acknowledgements

D. Ding and M. R. Jovanović were supported by the National Science Foundation under Awards ECCS-1708906 and ECCS-1809833.

References

[1] Y. Abbasi-Yadkori, D. Pál, and C. Szepesvári. Improved algorithms for linear stochastic bandits. In *Advances in Neural Information Processing Systems*, volume 24, pages 2312–2320, 2011.

[2] N. Abe, P. Melville, C. Pendus, C. K. Reddy, D. L. Jensen, V. P. Thomas, J. J. Bennett, G. F. Anderson, B. R. Cooley, M. Kowalczyk, et al. Optimizing debt collections using constrained reinforcement learning. In *International Conference on Knowledge Discovery and Data Mining*, pages 75–84, 2010.

[3] J. Achiam, D. Held, A. Tamar, and P. Abbeel. Constrained policy optimization. In *International Conference on Machine Learning*, volume 70, pages 22–31, 2017.

[4] A. Agarwal, S. M. Kakade, J. D. Lee, and G. Mahajan. Optimality and approximation with policy gradient methods in Markov decision processes. *arXiv preprint arXiv:1908.00261*, 2019.

[5] E. Altman. *Constrained Markov Decision Processes*, volume 7. CRC Press, 1999.

[6] D. Amodei, C. Olah, J. Steinhardt, P. Christiano, J. Schulman, and D. Mané. Concrete problems in AI safety. *arXiv preprint arXiv:1606.06565*, 2016.

[7] P. Auer, N. Cesa-Bianchi, and P. Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47(2):235–256, 2002.

[8] A. Ayoub, Z. Jia, C. Szepesvari, M. Wang, and L. Yang. Model-based reinforcement learning with value-targeted regression. In *International Conference on Machine Learning*, pages 463–474, 2020.

[9] M. G. Azar, I. Osband, and R. Munos. Minimax regret bounds for reinforcement learning. In *International Conference on Machine Learning*, pages 263–272, 2017.

[10] Q. Bai, A. Gattami, and V. Aggarwal. Model-free algorithm and regret analysis for MDPs with peak constraints. *arXiv preprint arXiv:2003.05555*, 2020.

[11] Y. Bai, T. Xie, N. Jiang, and Y.-X. Wang. Provably efficient Q-learning with low switching cost. In *Advances in Neural Information Processing Systems*, volume 32, pages 8002–8011, 2019.

[12] A. Beck. *First-order methods in optimization*. SIAM, 2017.

[13] F. Berkenkamp, M. Turchetta, A. Schoellig, and A. Krause. Safe model-based reinforcement learning with stability guarantees. In *Advances in Neural Information Processing Systems*, volume 30, pages 908–918, 2017.

[14] S. Bhatnagar and K. Lakshmanan. An online actor–critic algorithm with function approximation for constrained Markov decision processes. *Journal of Optimization Theory and Applications*, 153(3):688–708, 2012.

[15] V. S. Borkar. An actor-critic algorithm for constrained Markov decision processes. *Systems & Control Letters*, 54(3):207–213, 2005.

[16] J. A. Boyan. Least-squares temporal difference learning. In *International Conference on Machine Learning*, pages 49–56, 1999.

[17] S. J. Bradtke and A. G. Barto. Linear least-squares algorithms for temporal difference learning. *Machine Learning*, 22(1-3):33–57, 1996.

[18] K. Brantley, M. Dudik, T. Lykouris, S. Miryoosefi, M. Simchowitz, A. Slivkins, and W. Sun. Constrained episodic reinforcement learning in concave-convex and knapsack settings. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

[19] S. Bubeck, N. Cesa-Bianchi, et al. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends® in Machine Learning*, 5(1):1–122, 2012.

[20] Q. Cai, Z. Yang, C. Jin, and Z. Wang. Provably efficient exploration in policy optimization. In *International Conference on Machine Learning*, pages 1283–1294, 2020.

[21] Y. Chen, J. Dong, and Z. Wang. A primal-dual approach to constrained Markov decision processes. *arXiv preprint arXiv:2101.10895*, 2021.

[22] Y. Chow, M. Ghavamzadeh, L. Janson, and M. Pavone. Risk-constrained reinforcement learning with percentile risk criteria. *The Journal of Machine Learning Research*, 18(1):6070–6120, 2017.

[23] Y. Chow, O. Nachum, E. Duenez-Guzman, and M. Ghavamzadeh. A lyapunov-based approach to safe reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 31, pages 8092–8101, 2018.

[24] Y. Chow, O. Nachum, A. Faust, M. Ghavamzadeh, and E. Duenez-Guzman. Lyapunov-based safe policy optimization for continuous control. *arXiv preprint arXiv:1901.10031*, 2019.

[25] G. Dalal, K. Dvijotham, M. Vecerik, T. Hester, C. Paduraru, and Y. Tassa. Safe exploration in continuous action spaces. *arXiv preprint arXiv:1801.08757*, 2018.

[26] D. Ding, K. Zhang, T. Basar, and M. Jovanovic. Natural policy gradient primal-dual method for constrained Markov decision processes. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

[27] S. S. Du, S. M. Kakade, R. Wang, and L. F. Yang. Is a good representation sufficient for sample efficient reinforcement learning? In *International Conference on Learning Representations*, 2019.

[28] G. Dulac-Arnold, D. Mankowitz, and T. Hester. Challenges of real-world reinforcement learning. *arXiv preprint arXiv:1904.12901*, 2019.

[29] Y. Efroni, S. Mannor, and M. Pirotta. Exploration-exploitation in constrained MDPs. *arXiv preprint arXiv:2003.02189*, 2020.

[30] D. Ernst, P. Geurts, and L. Wehenkel. Tree-based batch mode reinforcement learning. *Journal of Machine Learning Research*, 6(Apr):503–556, 2005.

[31] S. Feyzabadi. *Robot Planning with Constrained Markov Decision Processes*. PhD thesis, UC Merced, 2017.

[32] J. F. Fisac, A. K. Akametalu, M. N. Zeilinger, S. Kaynama, J. Gillula, and C. J. Tomlin. A general safety framework for learning-based control in uncertain robotic systems. *IEEE Transactions on Automatic Control*, 64(7):2737–2752, 2018.

[33] J. Garcıa and F. Fernández. A comprehensive survey on safe reinforcement learning. *Journal of Machine Learning Research*, 16(1):1437–1480, 2015.

[34] C. J. Girard. *Structural Results for Constrained Markov Decision Processes*. PhD thesis, Cornell University, 2018.

[35] A. HasanzadeZonuzy, D. Kalathil, and S. Shakkottai. Learning with safety constraints: Sample complexity of reinforcement learning for constrained mdps. *arXiv preprint arXiv:2008.00311*, 2020.

[36] C. Jin, Z. Allen-Zhu, S. Bubeck, and M. I. Jordan. Is Q-learning provably efficient? In *Advances in Neural Information Processing Systems*, volume 31, pages 4863–4873, 2018.

[37] C. Jin, Z. Yang, Z. Wang, and M. I. Jordan. Provably efficient reinforcement learning with linear function approximation. In *Conference on Learning Theory*, pages 2137–2143, 2020.

[38] S. Kakade and J. Langford. Approximately optimal approximate reinforcement learning. In *Proceedings of the Nineteenth International Conference on Machine Learning*, page 267–274, 2002.

[39] S. M. Kakade. A natural policy gradient. In *Advances in Neural Information Processing Systems*, volume 15, pages 1531–1538, 2002.

[40] K. C. Kalagarla, R. Jain, and P. Nuzzo. A sample-efficient algorithm for episodic finite-horizon MDP with constraints. *arXiv preprint arXiv:2009.11348*, 2020.

[41] A. Koppel, K. Zhang, H. Zhu, and T. Başar. Projected stochastic primal-dual method for constrained online learning with kernels. *IEEE Transactions on Signal Processing*, 67(10):2528–2542, 2019.

[42] M. G. Lagoudakis and R. Parr. Least-squares policy iteration. *Journal of Machine Learning Research*, 4(Dec):1107–1149, 2003.

[43] T. Lattimore, C. Szepesvari, and G. Weisz. Learning with good feature representations in bandits and in RL with a generative model. In *International Conference on Machine Learning*, pages 5662–5670, 2020.

[44] A. Lazaric, M. Ghavamzadeh, and R. Munos. Finite-sample analysis of LSTD. In *International Conference on Machine Learning*, pages 615–622, 2010.

[45] H. Le, C. Voloshin, and Y. Yue. Batch policy learning under constraints. In *International Conference on Machine Learning*, pages 3703–3712, 2019.

[46] Q. Liang, F. Que, and E. Modiano. Accelerated primal-dual policy optimization for safe reinforcement learning. *arXiv preprint arXiv:1802.06480*, 2018.

[47] B. Liu, Q. Cai, Z. Yang, and Z. Wang. Neural trust region/proximal policy optimization attains globally optimal policy. In *Advances in Neural Information Processing Systems*, volume 32, pages 10564–10575, 2019.

[48] Y. Liu, J. Ding, and X. Liu. IPO: Interior-point policy optimization under constraints. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 4940–4947, 2020.

[49] M. Mahdavi, R. Jin, and T. Yang. Trading regret for efficiency: online convex optimization with long term constraints. *The Journal of Machine Learning Research*, 13(1):2503–2528, 2012.

[50] A. Modi, N. Jiang, A. Tewari, and S. Singh. Sample complexity of reinforcement learning using linearly combined model ensembles. In *International Conference on Artificial Intelligence and Statistics*, pages 2010–2020, 2020.

[51] A. Nedić and A. Ozdaglar. Subgradient methods for saddle-point problems. *Journal of optimization theory and applications*, 142(1):205–228, 2009.

[52] S. Paternain, M. Calvo-Fullana, L. F. Chamon, and A. Ribeiro. Safe policies for reinforcement learning via primal-dual methods. *arXiv preprint arXiv:1911.09101*, 2019.

[53] S. Paternain, L. Chamon, M. Calvo-Fullana, and A. Ribeiro. Constrained reinforcement learning has zero duality gap. In *Advances in Neural Information Processing Systems*, volume 32, pages 7553–7563, 2019.

[54] B. Á. Pires and C. Szepesvári. Policy error bounds for model-based reinforcement learning with factored linear models. In *Conference on Learning Theory*, pages 121–151, 2016.

[55] S. Qiu, X. Wei, Z. Yang, J. Ye, and Z. Wang. Upper confidence primal-dual optimization: Stochastically constrained Markov decision processes with adversarial losses and unknown transitions. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

[56] J. Schulman, S. Levine, P. Abbeel, M. Jordan, and P. Moritz. Trust region policy optimization. In *International Conference on Machine Learning*, pages 1889–1897, 2015.

[57] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

[58] R. Singh, A. Gupta, and N. B. Shroff. Learning in Markov decision processes under constraints. *arXiv preprint arXiv:2002.12435*, 2020.

[59] A. Stooke, J. Achiam, and P. Abbeel. Responsive safety in reinforcement learning by PID Lagrangian methods. In *International Conference on Machine Learning*, pages 9133–9143, 2020.

[60] R. S. Sutton and A. G. Barto. *Reinforcement Learning: An Introduction*. MIT press, 2018.

[61] C. Tessler, D. J. Mankowitz, and S. Mannor. Reward constrained policy optimization. In *International Conference on Learning Representations*, 2019.

[62] M. Turchetta, F. Berkenkamp, and A. Krause. Safe exploration in finite Markov decision processes with Gaussian processes. In *Advances in Neural Information Processing Systems*, volume 29, pages 4312–4320, 2016.

[63] E. Uchibe and K. Doya. Constrained reinforcement learning from intrinsic and extrinsic rewards. In *International Conference on Development and Learning*, pages 163–168, 2007.

[64] B. Van Roy and S. Dong. Comments on the Du-Kakade-Wang-Yang lower bounds. *arXiv preprint arXiv:1911.07910*, 2019.

[65] A. Wachi and Y. Sui. Safe reinforcement learning in constrained Markov decision processes. In *International Conference on Machine Learning*, pages 9797–9806, 2020.

[66] A. Wachi, Y. Sui, Y. Yue, and M. Ono. Safe exploration and optimization of constrained MDPs using Gaussian processes. In *AAAI Conference on Artificial Intelligence*, 2018.

[67] L. Wang, Q. Cai, Z. Yang, and Z. Wang. Neural policy gradient methods: Global optimality and rates of convergence. In *International Conference on Learning Representations*, 2019.

[68] X. Wei, H. Yu, and M. 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.

[69] T. Xu, Y. Liang, and G. Lan. A primal approach to constrained policy optimization: Global optimality and finite-time analysis. *arXiv preprint arXiv:2011.05869*, 2020.

[70] L. Yang and M. Wang. Sample-optimal parametric Q-learning using linearly additive features. In *International Conference on Machine Learning*, pages 6995–7004, 2019.

[71] L. Yang and M. Wang. Reinforcement learning in feature space: Matrix bandit, kernels, and regret bound. In *International Conference on Machine Learning*, pages 10746–10756, 2020.

[72] T.-Y. Yang, J. Rosca, K. Narasimhan, and P. J. Ramadge. Projection-based constrained policy optimization. In *International Conference on Learning Representations*, 2019.

[73] H. Yu, M. Neely, and X. Wei. Online convex optimization with stochastic constraints. In *Advances in Neural Information Processing Systems*, volume 30, pages 1428–1438, 2017.

[74] M. Yu, Z. Yang, M. Kolar, and Z. Wang. Convergent policy optimization for safe reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 32, pages 3121–3133, 2019.

[75] J. Yuan and A. Lamperski. Online convex optimization for cumulative constraints. In *Advances in Neural Information Processing Systems*, volume 31, pages 6137–6146, 2018.

[76] A. Zanette, A. Lazaric, M. Kochenderfer, and E. Brunskill. Learning near optimal policies with low inherent bellman error. In *International Conference on Machine Learning*, pages 10978–10989, 2020.

[77] J. Zhang, A. Koppel, A. S. Bedi, C. Szepesvari, and M. Wang. Variational policy gradient method for reinforcement learning with general utilities. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

[78] Y. Zhang, Q. Vuong, and K. Ross. First order constrained optimization in policy space. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

[79] D. Zhou, J. He, and Q. Gu. Provably efficient reinforcement learning for discounted MDPs with feature mapping. *arXiv preprint arXiv:2006.13165*, 2020.

Supplementary Materials for "Provably Efficient Safe Exploration via Primal-Dual Policy Optimization"

A Preliminaries

Our analysis begins with decomposition of the regret given in (3).

$$\text{Regret}(K) \;=\; \underbrace{\sum_{k=1}^{K} \left(V_{r,1}^{\pi^\star}(x_1) - V_{r,1}^{k}(x_1)\right)}_{\text{(R.I)}} + \underbrace{\sum_{k=1}^{K} \left(V_{r,1}^{k}(x_1) - V_{r,1}^{\pi^k}(x_1)\right)}_{\text{(R.II)}} \tag{10}$$

where we add and subtract the value $V_{r,1}^k(x_1)$ estimated from an optimistic policy evaluation by Algorithm 2; the policy $\pi^\star$ in hindsight is the best policy in hindsight for problem (2). To bound the total regret (10), we would like to analyze (R.I) and (R.II) separately.

First, we define the model prediction error for the reward as

$$\iota_{r,h}^k \;:=\; r_h \,+\, \mathbb{P}_h V_{r,h+1}^k \,-\, Q_{r,h}^k \tag{11}$$

for all $(k,h) \in [K] \times [H]$, which describes the prediction error in the Bellman equations (1) using $V_{r,h+1}^k$ instead of $V_{r,h+1}^{\pi^k}$. With this notation, we expand (R.I) into

$$\sum_{k=1}^{K}\sum_{h=1}^{H} \mathbb{E}_{\pi^\star}\left[\left\langle Q_{r,h}^k(x_h,\,\cdot\,), \pi_h^\star(\,\cdot\,|\,x_h) - \pi_h^k(\,\cdot\,|\,x_h)\right\rangle\right] \,+\, \sum_{k=1}^{K}\sum_{h=1}^{H} \mathbb{E}_{\pi^\star}\left[\iota_{r,h}^k(x_h,a_h)\right] \tag{12}$$

where the first double sum is linear in terms of the policy difference and the second one describes the total model prediction error. The above expansion is based on the standard performance difference lemma (see Lemma 2) and we provide a proof in Section D.1 for readers' convenience. Meanwhile, if we define the model prediction error for the utility as

$$\iota_{g,h}^k \;:=\; g_h \,+\, \mathbb{P}_h V_{g,h+1}^k \,-\, Q_{g,h}^k \tag{13}$$

then, similarly, we can expand $\sum_{k=1}^K \left(V_{g,1}^{\pi^\star}(x_1) - V_{g,1}^k(x_1)\right)$ into

$$\sum_{k=1}^{K}\sum_{h=1}^{H} \mathbb{E}_{\pi^\star}\left[\left\langle Q_{g,h}^k(x_h,\,\cdot\,), \pi_h^\star(\,\cdot\,|\,x_h) - \pi_h^k(\,\cdot\,|\,x_h)\right\rangle\right] \,+\, \sum_{k=1}^{K}\sum_{h=1}^{H} \mathbb{E}_{\pi^\star}\left[\iota_{g,h}^k(x_h,a_h)\right]. \tag{14}$$

To analyze the constraint violation, we also introduce a useful decomposition,

$$\text{Violation}(K) \;=\; \sum_{k=1}^{K} \left(b - V_{g,1}^{k}(x_1)\right) \,+\, \underbrace{\sum_{k=1}^{K} \left(V_{g,1}^{k}(x_1) - V_{g,1}^{\pi^k}(x_1)\right)}_{\text{(V.II)}} \tag{15}$$

which the inserted value $V_{g,1}^k(x_1)$ is estimated from an optimistic policy evaluation by Algorithm 2.

For notational simplicity, we introduce the underlying probability structure as follows. For any $(k,h) \in [K]\times[H]$, we define $\mathcal{F}_{h,1}^k$ as a σ-algebra generated by state-action sequences, reward and utility functions,

$$\{(x_i^\tau, a_i^\tau)\}_{(\tau,i)\,\in\,[k-1]\times[H]} \bigcup \{(x_i^k, a_i^k)\}_{i\,\in\,[h]}.$$

Similarly, we define $\mathcal{F}_{h,2}^k$ as an σ-algebra generated by

$$\{(x_i^\tau, a_i^\tau)\}_{(\tau,i)\,\in\,[k-1]\times[H]} \bigcup \{(x_i^k, a_i^k)\}_{i\,\in\,[h]} \bigcup \{x_{h+1}^k\}.$$

Here, x_{H+1}^k is a null state for any $k \in [K]$. A filtration is a sequence of σ-algebras $\{\mathcal{F}_{h,m}^k\}_{(k,h,m)\,\in\,[K]\times[H]\times[2]}$ in terms of time index

$$t(k,h,m) \;:=\; 2(k-1)H \,+\, 2(h-1) \,+\, m \tag{16}$$

which holds that $\mathcal{F}_{h,m}^k \subset \mathcal{F}_{h',m'}^{k'}$ for any $t \leq t'$. The estimated reward/utility value functions, $V_{r,h}^k, V_{g,h}^k$, and the associated Q-functions, $Q_{r,h}^k, Q_{g,h}^k$ are $\mathcal{F}_{1,1}^k$-measurable since they are obtained from previous $k-1$ historical trajectories. With these notations, we can expand (R.II) in (10) into

$$\text{(R.II)} \;=\; -\sum_{k\,=\,1}^{K}\sum_{h\,=\,1}^{H} \iota_{r,h}^k(x_h^k,a_h^k) \,+\, M_{r,H,2}^K \tag{17}$$

where $\{M_{r,h,m}^k\}_{(k,h,m)\in[K]\times[H]\times[2]}$ is a martingale adapted to the filtration $\{\mathcal{F}_{h,m}^k\}_{(k,h,m)\in[K]\times[H]\times[2]}$ in terms of time index t. Similarly, we have it for (V.II),

$$\text{(V.II)} \;=\; -\sum_{k\,=\,1}^{K}\sum_{h\,=\,1}^{H} \iota_{g,h}^k(x_h^k,a_h^k) \,+\, M_{g,H,2}^K \tag{18}$$

where $\{M_{g,h,m}^k\}_{(k,h,m)\in[K]\times[H]\times[2]}$ is a martingale adapted to the filtration $\{\mathcal{F}_{h,m}^k\}_{(k,h,m)\in[K]\times[H]\times[2]}$ in terms of time index t. We prove (17) in Section D.2 for completeness (also see [20, Lemma 4.2]); (18) is similar.

We recall two UCB bonus terms $\Gamma_{\diamond,h}^k := \beta((\phi_{\diamond,h}^k)^\top(\Lambda_{\diamond,h}^k)^{-1}\phi_{\diamond,h}^k)^{1/2}$ and $\Gamma_h^k := \beta((\varphi)^\top(\Lambda_h^k)^{-1}\varphi)^{1/2}$ in the action-value function estimation of Algorithm 2. By the UCB argument, if we set $\lambda = 1$ and $\beta = C_1\sqrt{dH^2\log(dT/p)}$ where C_1 is an absolute constant, then for any $(k,h) \in [K]\times[H]$ and $(x,a) \in \mathcal{S}\times\mathcal{A}$, we have

$$-\,2(\Gamma_h^k + \Gamma_{\diamond,h}^k)(x,a) \;\leq\; \iota_{\diamond,h}^k\,(x,a) \;\leq\; 0 \tag{19}$$

with probability $1-p/2$ where the symbol $\diamond = r$ or g. We prove (19) in Section D.3 for completeness.

In what follows we delve into the analysis of the regret and the constraint violation.

B Proof of Regret and Constraint Violation

The goal is to prove that the regret and the constraint violation for Algorithm 1 are sublinear in the total number of steps $T := KH$, taken by the algorithm. Here, K is the total number of episodes and H is the horizon length. We recall that $|\mathcal{A}|$ is the size of action space $\mathcal{A}$ and d is the feature map's dimension. We repeat Theorem 1 here for readers' convenience.

Theorem 1 (Linear Kernal MDP: Regret and Constraint Violation)**.** *Let Assumptions 1 and 2 hold. Fix* $p \in (0,1)$*. We set* $\alpha = \sqrt{\log|\mathcal{A}|}/(H^2K)$*,* $\beta = C_1\sqrt{dH^2\log(dT/p)}$*,* $\eta = 1/\sqrt{K}$*,* $\theta = 1/K$*, and* $\lambda = 1$ *in Algorithm 1 with the full-information setting, where* C_1 *is an absolute constant. Suppose* $\log|\mathcal{A}| = O\left(d^2\log^2(dT/p)\right)$*. Then, the regret and the constraint violation in (3) satisfy*

$$\text{Regret}(K) \;\leq\; C\,dH^{2.5}\sqrt{T}\,\log\left(\frac{dT}{p}\right) \quad \textit{and} \quad [\text{Violation}(K)]_+ \;\leq\; C'\,dH^{2.5}\sqrt{T}\log\left(\frac{dT}{p}\right)$$

with probability $1-p$ *where* C *and* C' *are absolute constants.*

We divide the proof into two parts for the regret bound and the constraint violation, respectively, in Section B.1 and Section B.2.

B.1 Proof of Regret Bound

Our analysis begins with a primal-dual mirror descent type analysis for the policy update in line 6 of Algorithm 1. In Lemma 3, we present a key upper bound on the total differences of estimated values $V_{r,1}^k(x_1)$ and $V_{g,1}^k(x_1)$ given by Algorithm 2 to the optimal ones.

Lemma 3 (Policy Improvement: Primal-Dual Mirror Descent Step)**.** *Let Assumption 1 and Assumption 2 hold. In Algorithm 1, if we set* $\alpha = \sqrt{\log|\mathcal{A}|}/(H^2\sqrt{K})$ *and* $\theta = 1/K$*, then*

$$
\begin{aligned}
&\sum_{k=1}^{K} \big(V_{r,1}^{\pi^\star}(x_1) - V_{r,1}^{k}(x_1)\big) \,+\, \sum_{k=1}^{K} Y^k \big(V_{g,1}^{\pi^\star}(x_1) - V_{g,1}^{k}(x_1)\big) \\
&\leq\; C_2 H^{2.5}\sqrt{T\log|\mathcal{A}|} \,+\, \sum_{k=1}^{K}\sum_{h=1}^{H} \mathbb{E}_{\pi^\star}\Big[\iota_{r,h}^{k}(x_h,a_h)\Big] \,+\, \sum_{k=1}^{K}\sum_{h=1}^{H} Y^k \mathbb{E}_{\pi^\star}\Big[\iota_{g,h}^{k}(x_h,a_h)\Big].
\end{aligned} \tag{20}
$$

where C_2 *is an absolute constant and* $T = HK$*.*

Proof. We recall that line 6 of Algorithm 1 follows a solution π^k to the following subproblem,

$$
\underset{\pi\,\in\,\Delta(\mathcal{A}\,|\,\mathcal{S},H)}{\text{maximize}} \; \sum_{h=1}^{H} \big\langle Q_{r,h}^{k-1} + Y^{k-1} Q_{g,h}^{k-1}, \pi_h \big\rangle \,-\, \frac{1}{\alpha}\sum_{h=1}^{H} D\big(\pi_h \,|\, \widetilde{\pi}_h^{k-1}\big) \tag{21}
$$

where we use the shorthand $\langle Q_{r,h}^{k-1} + Y^{k-1}Q_{g,h}^{k-1}, \pi_h\rangle$ for $\big\langle (Q_{r,h}^{k-1} + Y^{k-1}Q_{g,h}^{k-1})(x_h,\cdot), \pi_h(\cdot\,|\,x_h)\big\rangle$ and the shorthand $D(\pi_h\,|\,\widetilde{\pi}_h^{k-1})$ for $D(\pi_h(\cdot\,|\,x_h)\,|\,\widetilde{\pi}_h^{k-1}(\cdot\,|\,x_h))$ if dependence on the state-action sequence $\{x_h, a_h\}_{h=1}^H$ is clear from context. We note that (21) is in form of a mirror descent subproblem in Lemma 14. We can apply the pushback property with $x^\star = \pi_h^k$, $y = \widetilde{\pi}_h^{k-1}$ and $z = \pi_h^\star$,

$$
\begin{aligned}
&\sum_{h=1}^{H} \big\langle Q_{r,h}^{k-1} + Y^{k-1}Q_{g,h}^{k-1}, \pi_h^k \big\rangle \,-\, \frac{1}{\alpha}\sum_{h=1}^{H} D\big(\pi_h^k\,|\,\widetilde{\pi}_h^{k-1}\big) \\
&\geq\; \sum_{h=1}^{H} \big\langle Q_{r,h}^{k-1} + Y^{k-1}Q_{g,h}^{k-1}, \pi_h^\star \big\rangle \,-\, \frac{1}{\alpha}\sum_{h=1}^{H} D\big(\pi_h^\star\,|\,\widetilde{\pi}_h^{k-1}\big) \,+\, \frac{1}{\alpha}\sum_{h=1}^{H} D\big(\pi_h^\star\,|\,\pi_h^k\big).
\end{aligned}
$$

Equivalently, we write the above inequality as follows,

$$
\begin{aligned}
&\sum_{h=1}^{H} \big\langle Q_{r,h}^{k-1}, \pi_h^\star - \pi_h^{k-1}\big\rangle \,+\, Y^{k-1}\sum_{h=1}^{H} \big\langle Q_{g,h}^{k-1}, \pi_h^\star - \pi_h^{k-1}\big\rangle \\
&\leq\; \sum_{h=1}^{H} \big\langle Q_{r,h}^{k-1} + Y^{k-1}Q_{g,h}^{k-1}, \pi_h^k - \pi_h^{k-1}\big\rangle \,-\, \frac{1}{\alpha}\sum_{h=1}^{H} D\big(\pi_h^k\,|\,\widetilde{\pi}_h^{k-1}\big) \\
&\quad +\, \frac{1}{\alpha}\sum_{h=1}^{H} D\big(\pi_h^\star\,|\,\widetilde{\pi}_h^{k-1}\big) \,-\, \frac{1}{\alpha}\sum_{h=1}^{H} D\big(\pi_h^\star\,|\,\pi_h^k\big).
\end{aligned} \tag{22}
$$

By taking expectation $\mathbb{E}_{\pi^\star}$ on both sides of (22) over the random state-action sequence $\{(x_h,a_h)\}_{h=1}^H$ starting from x_1, and applying decompositions (12) and (14), we have

$$
\begin{aligned}
&\big(V_{r,1}^{\pi^\star}(x_1) - V_{r,1}^{k-1}(x_1)\big) \,+\, Y^{k-1}\big(V_{g,1}^{\pi^\star}(x_1) - V_{g,1}^{k-1}(x_1)\big) \\
&\leq\; \sum_{h=1}^{H} \mathbb{E}_{\pi^\star}\Big[\big\langle Q_{r,h}^{k-1} + Y^{k-1}Q_{g,h}^{k-1}, \pi_h^k - \pi_h^{k-1}\big\rangle\Big] \,-\, \frac{1}{\alpha}\sum_{h=1}^{H} \mathbb{E}_{\pi^\star}\big[D\big(\pi_h^k\,|\,\widetilde{\pi}_h^{k-1}\big)\big] \\
&\quad +\, \frac{1}{\alpha}\sum_{h=1}^{H} \mathbb{E}_{\pi^\star}\big[D\big(\pi_h^\star\,|\,\widetilde{\pi}_h^{k-1}\big) - D\big(\pi_h^\star\,|\,\pi_h^k\big)\big] \\
&\quad +\, \sum_{h=1}^{H} \mathbb{E}_{\pi^\star}\Big[\iota_{r,h}^{k-1}(x_h,a_h)\Big] \,+\, Y^{k-1}\sum_{h=1}^{H} \mathbb{E}_{\pi^\star}\Big[\iota_{g,h}^{k-1}(x_h,a_h)\Big]
\end{aligned} \tag{23}
$$

The rest is to bound the right-hand side of the above inequality (23). By the Hölder's inequality and the Pinsker's inequality, we first have

$$
\begin{aligned}
& \sum_{h=1}^{H} \left\langle Q_{r,h}^{k-1} + Y^{k-1} Q_{g,h}^{k-1}, \pi_h^k - \pi_h^{k-1} \right\rangle - \frac{1}{\alpha} \sum_{h=1}^{H} D\big(\pi_h^k \,|\, \widetilde{\pi}_h^{k-1}\big) \\
=\; & \sum_{h=1}^{H} \left\langle Q_{r,h}^{k-1} + Y^{k-1} Q_{g,h}^{k-1}, \pi_h^k - \widetilde{\pi}_h^{k-1} \right\rangle - \frac{1}{\alpha} \sum_{h=1}^{H} D\big(\pi_h^k \,|\, \widetilde{\pi}_h^{k-1}\big) \\
& + \sum_{h=1}^{H} \left\langle Q_{r,h}^{k-1} + Y^{k-1} Q_{g,h}^{k-1}, \widetilde{\pi}_h^{k-1} - \pi_h^{k-1} \right\rangle \\
\leq\; & \sum_{h=1}^{H} \left(\big\| Q_{r,h}^{k-1} + Y^{k-1} Q_{g,h}^{k-1} \big\|_\infty \big\| \pi_h^k - \widetilde{\pi}_h^{k-1} \big\|_1 - \frac{1}{2\alpha} \big\| \pi_h^k - \widetilde{\pi}_h^{k-1} \big\|_1^2 \right) \\
& + \sum_{h=1}^{H} \big\| Q_{r,h}^{k-1} + Y^{k-1} Q_{g,h}^{k-1} \big\|_\infty \big\| \widetilde{\pi}_h^{k-1} - \pi_h^{k-1} \big\|_1.
\end{aligned}
$$

Then, using the square completion,

$$
\begin{aligned}
& \big\| Q_{r,h}^{k-1} + Y^{k-1} Q_{g,h}^{k-1} \big\|_\infty \big\| \pi_h^k - \widetilde{\pi}_h^{k-1} \big\|_1 - \frac{1}{2\alpha} \big\| \pi_h^k - \widetilde{\pi}_h^{k-1} \big\|_1^2 \\
=\; & -\frac{1}{2\alpha} \left(\alpha \big\| Q_{r,h}^{k-1} + Y^{k-1} Q_{g,h}^{k-1} \big\|_\infty - \big\| \pi_h^k - \widetilde{\pi}_h^{k-1} \big\|_1 \right)^2 + \frac{\alpha}{2} \big\| Q_{r,h}^{k-1} + Y^{k-1} Q_{g,h}^{k-1} \big\|_\infty^2 \\
\leq\; & \frac{\alpha}{2} \big\| Q_{r,h}^{k-1} + Y^{k-1} Q_{g,h}^{k-1} \big\|_\infty^2
\end{aligned}
$$

where we dropoff the first quadratic term for the inequality, and $\big\| \widetilde{\pi}_h^{k-1} - \pi_h^{k-1} \big\|_1 \leq \theta$, we have

$$
\begin{aligned}
& \sum_{h=1}^{H} \left\langle Q_{r,h}^{k-1} + Y^{k-1} Q_{g,h}^{k-1}, \pi_h^k - \pi_h^{k-1} \right\rangle - \frac{1}{\alpha} \sum_{h=1}^{H} D\big(\pi_h^k \,|\, \widetilde{\pi}_h^{k-1}\big) \\
\leq\; & \frac{\alpha}{2} \sum_{h=1}^{H} \big\| Q_{r,h}^{k-1} + Y^{k-1} Q_{g,h}^{k-1} \big\|_\infty^2 + \theta \sum_{h=1}^{H} \big\| Q_{r,h}^{k-1} + Y^{k-1} Q_{g,h}^{k-1} \big\|_\infty \\
\leq\; & \frac{\alpha (1+\chi)^2 H^3}{2} + \theta\, (1+\chi)\, H^2
\end{aligned}
\tag{24}
$$

where the last inequality is due to $\big\| Q_{r,h}^{k-1} \big\|_\infty \leq H$, a fact from line 12 in Algorithm 2, and $0 \leq Y^{k-1} \leq \chi$. Taking the same expectation $\mathbb{E}_{\pi^\star}$ as previously on both sides of (24) and substituting it into the left-hand side of (23) yield,

$$
\begin{aligned}
& \big(V_{r,1}^{\pi^\star}(x_1) - V_{r,1}^{k-1}(x_1) \big) + Y^{k-1} \big(V_{g,1}^{\pi^\star}(x_1) - V_{g,1}^{k-1}(x_1) \big) \\
\leq\; & \frac{\alpha (1+\chi)^2 H^3}{2} + \theta\, (1+\chi)\, H^2 + \frac{1}{\alpha} \sum_{h=1}^{H} \mathbb{E}_{\pi^\star} \left[D\big(\pi_h^\star \,|\, \widetilde{\pi}_h^{k-1}\big) - D\big(\pi_h^\star \,|\, \pi_h^k\big) \right] \\
& + \sum_{h=1}^{H} \mathbb{E}_{\pi^\star} \left[\iota_{r,h}^{k-1}(x_h, a_h) \right] + Y^{k-1} \sum_{h=1}^{H} \mathbb{E}_{\pi^\star} \left[\iota_{g,h}^{k-1}(x_h, a_h) \right] \\
\leq\; & \frac{\alpha (1+\chi)^2 H^3}{2} + \theta\, (1+\chi)\, H^2 + \frac{\theta H \log |\mathcal{A}|}{\alpha} + \frac{1}{\alpha} \sum_{h=1}^{H} \mathbb{E}_{\pi^\star} \left[D\big(\pi_h^\star \,|\, \pi_h^{k-1}\big) - D\big(\pi_h^\star \,|\, \pi_h^k\big) \right] \\
& + \sum_{h=1}^{H} \mathbb{E}_{\pi^\star} \left[\iota_{r,h}^{k-1}(x_h, a_h) \right] + Y^{k-1} \sum_{h=1}^{H} \mathbb{E}_{\pi^\star} \left[\iota_{g,h}^{k-1}(x_h, a_h) \right].
\end{aligned}
\tag{25}
$$

where in the second inequality we note the fact that $D(\pi_h^\star \,|\, \widetilde{\pi}_h^{k-1}) - D(\pi_h^\star \,|\, \pi_h^{k-1}) \leq \theta \log |\mathcal{A}|$ from Lemma 15.

We note that Y^0 is initialized to be zero. By taking a telescoping sum of both sides of (25) from $k = 1$ to

$k = K+1$ and shifting the index k by one, we have

$$\sum_{k=1}^{K} \left(V_{r,1}^{\pi^\star}(x_1) - V_{r,1}^k(x_1)\right) + \sum_{k=1}^{K} Y^k \left(V_{g,1}^{\pi^\star}(x_1) - V_{g,1}^k(x_1)\right)$$
$$\leq \frac{\alpha(1+\chi)^2 H^3(K+1)}{2} + \theta\,(1+\chi)\,H^2(K+1) + \frac{\theta H(K+1)\log|\mathcal{A}|}{\alpha} + \frac{H\log|\mathcal{A}|}{\alpha} \tag{26}$$
$$+ \sum_{k=1}^{K}\sum_{h=1}^{H} \mathbb{E}_{\pi^\star}\left[\iota_{r,h}^k(x_h,a_h)\right] + \sum_{k=1}^{K}\sum_{h=1}^{H} Y^k \mathbb{E}_{\pi^\star}\left[\iota_{g,h}^k(x_h,a_h)\right].$$

where we ignore $-\alpha^{-1}\sum_{h=1}^{H}\mathbb{E}_{\pi^\star}[D(\pi_h^\star\,|\,\pi_h^{K+1})]$ and utilize

$$D\left(\pi_h^\star\,|\,\pi_h^0\right) = \sum_{a\in\mathcal{A}} \pi_h^\star(a\,|\,x_h)\log\left(|\mathcal{A}|\,\pi_h^\star(a\,|\,x_h)\right) \leq \log|\mathcal{A}|$$

where π_h^0 is uniform over $\mathcal{A}$ and we ignore $\sum_{a\in\mathcal{A}}\pi_h^\star(a\,|\,x_h)\log\left(\pi_h^\star(a\,|\,x_h)\right)$ that is nonpositive.

Finally, we take $\chi := H/\gamma$ and α, θ in the lemma to complete the proof. □

By the dual update of Algorithm 1, we can simplify the result in Lemma 3 and return back to the regret (10).

Lemma 4. *Let Assumption 1 and Assumption 2 hold. In Algorithm 1, if we set* $\alpha = \sqrt{\log|\mathcal{A}|}/(H^2\sqrt{K})$, $\eta = 1/\sqrt{K}$, *and* $\theta = 1/K$, *then with probability* $1-p/2$,

$$\text{Regret}(K) = C_3 H^{2.5}\sqrt{T\log|\mathcal{A}|} + \sum_{k=1}^{K}\sum_{h=1}^{H}\left(\mathbb{E}_{\pi^\star}\left[\iota_{r,h}^k(x_h,a_h)\right] - \iota_{r,h}^k(x_h^k,a_h^k)\right) + M_{r,H,2}^K \tag{27}$$

where C_3 *is an absolute constant.*

Proof. By the dual update in line 9 in Algorithm 1, we have

$$\begin{aligned}
0 &\leq \left(Y^{K+1}\right)^2 \\
&= \sum_{k=1}^{K+1}\left(\left(Y^k\right)^2 - \left(Y^{k-1}\right)^2\right) \\
&= \sum_{k=1}^{K+1}\left(\text{Proj}_{[\,0,\,\chi\,]}\left(Y^{k-1} + \eta(b - V_{g,1}^{k-1}(x_1))\right)\right)^2 - \left(Y^{k-1}\right)^2 \\
&\leq \sum_{k=1}^{K+1}\left(Y^{k-1} + \eta(b - V_{g,1}^{k-1}(x_1))\right)^2 - \left(Y^{k-1}\right)^2 \\
&\leq \sum_{k=1}^{K+1} 2\eta Y^{k-1}\left(V_{g,1}^{\pi^\star}(x_1) - V_{g,1}^{k-1}(x_1)\right) + \eta^2\left(b - V_{g,1}^{k-1}(x_1)\right)^2.
\end{aligned}$$

where we use the feasibility of $\pi^\star$ in the last inequality. Since $Y^0 = 0$ and $|b - V_{g,1}^{k-1}(x_1)| \leq H$, the above inequality implies that

$$-\sum_{k=1}^{K} Y^k\left(V_{g,1}^{\pi^\star}(x_1) - V_{g,1}^k(x_1)\right) \leq \sum_{k=1}^{K+1}\frac{\eta}{2}\left(b - V_{g,1}^{k-1}(x_1)\right)^2 \leq \frac{\eta H^2(K+1)}{2}. \tag{28}$$

By noting the UCB result (19) and $Y^k \geq 0$, the inequality (20) implies that

$$\sum_{k=1}^{K}\left(V_{r,1}^{\pi^\star}(x_1) - V_{r,1}^k(x_1)\right) + \sum_{k=1}^{K} Y^k\left(V_{g,1}^{\pi^\star}(x_1) - V_{g,1}^k(x_1)\right) \leq C_2 H^{2.5}\sqrt{T\log|\mathcal{A}|} + \sum_{k=1}^{K}\sum_{h=1}^{H}\mathbb{E}_{\pi^\star}\left[\iota_{r,h}^k(x_h,a_h)\right].$$

If we add (28) to the above inequality and take $\eta = 1/\sqrt{K}$, then,

$$\sum_{k=1}^{K}\left(V_{r,1}^{\pi^\star}(x_1) - V_{r,1}^k(x_1)\right) \leq C_3 H^{2.5}\sqrt{T\log|\mathcal{A}|} + \sum_{k=1}^{K}\sum_{h=1}^{H}\mathbb{E}_{\pi^\star}\left[\iota_{r,h}^k(x_h,a_h)\right] \tag{29}$$

where C_3 is an absolute constant. Finally, we combine (17) and (29) to complete the proof. □

By Lemma 4, the rest is to bound the last two terms in the right-hand side of (27). We next show two probability bounds for them in Lemma 5 and Lemma 6, separately.

Lemma 5 (Model Prediction Error Bound)**.** *Let Assumption 2 hold. Fix $p \in (0,1)$. If we set $\beta = C_1\sqrt{dH^2\log(dT/p)}$ in Algorithm 1, then with probability $1-p/2$ it holds that*

$$\sum_{k=1}^{K}\sum_{h=1}^{H}\left(\mathbb{E}_{\pi^\star}\left[\iota_{r,h}^k(x_h,a_h)\right]-\iota_{r,h}^k(x_h^k,a_h^k)\right) \;\leq\; 4C_1\sqrt{2d^2H^3T\log(K+1)\log\left(\frac{dT}{p}\right)} \tag{30}$$

where C_1 is an absolute constant and $T = HK$.

Proof. By the UCB result (19), with probability $1-p/2$ for any $(k,h)\in[K]\times[H]$ and $(x,a)\in\mathcal{S}\times\mathcal{A}$, we have

$$-2(\Gamma_h^k+\Gamma_{r,h}^k)(x,a) \;\leq\; \iota_{r,h}^k(x,a) \;\leq\; 0.$$

By the definition of $\iota_{r,h}^k(x,a)$, $|\iota_{r,h}^k(x,a)|\leq 2H$. Hence, it holds with probability $1-p/2$ that

$$-\iota_{r,h}^k(x,a) \;\leq\; 2\min\big(H,(\Gamma_h^k+\Gamma_{r,h}^k)(x,a)\big)$$

for any $(k,h)\in[K]\times[H]$ and $(x,a)\in\mathcal{S}\times\mathcal{A}$. Therefore, we have

$$\sum_{k=1}^{K}\sum_{h=1}^{H}\left(\mathbb{E}_{\pi^\star}\left[\iota_{r,h}^k(x_h,a_h)\,|\,x_1\right]-\iota_{r,h}^k(x_h^k,a_h^k)\right) \;\leq\; 2\sum_{k=1}^{K}\sum_{h=1}^{H}\min\left(H,(\Gamma_h^k+\Gamma_{r,h}^k)(x_h^k,a_h^k)\right)$$

where $\Gamma_h^k(\cdot,\cdot)=\beta(\varphi(\cdot,\cdot)^\top(\Lambda_h^k)^{-1}\varphi(\cdot,\cdot))^{1/2}$ and $\Gamma_{r,h}^k(\cdot,\cdot)=\beta(\phi_{r,h}^k(\cdot,\cdot)^\top(\Lambda_{r,h}^k)^{-1}\phi_{r,h}^k(\cdot,\cdot))^{1/2}$. Application of the Cauchy-Schwartz inequality shows that

$$\begin{aligned}
&\sum_{k=1}^{K}\sum_{h=1}^{H}\min\left(H,(\Gamma_h^k+\Gamma_{r,h}^k)(x_h^k,a_h^k)\right)\\
&\leq\;\beta\sum_{k=1}^{K}\sum_{h=1}^{H}\min\left(H/\beta,\left(\varphi(x_h^k,a_h^k)^\top(\Lambda_h^k)^{-1}\varphi(x_h^k,a_h^k)\right)^{1/2}+\left(\phi_{r,h}^k(x_h^k,a_h^k)^\top(\Lambda_{r,h}^k)^{-1}\phi_{r,h}^k(x_h^k,a_h^k)\right)^{1/2}\right)
\end{aligned} \tag{31}$$

Since we take $\beta = C_1\sqrt{dH^2\log(dT/p)}$ with $C_1>1$, we have $H/\beta\leq 1$. The rest is to apply Lemma 13. First, for any $h\in[H]$ it holds that

$$\sum_{k=1}^{K}\phi_{r,h}^k\left(x_h^k,a_h^k\right)^\top\left(\Lambda_{r,h}^k\right)^{-1}\phi_{r,h}^k\left(x_h^k,a_h^k\right) \;\leq\; 2\log\left(\frac{\det\left(\Lambda_{r,h}^{K+1}\right)}{\det\left(\Lambda_{r,h}^1\right)}\right).$$

Due to $\|\phi_{r,h}^k\|\leq\sqrt{d}H$ in Assumption 2 and $\Lambda_{r,h}^1=\lambda I$ in Algorithm 2, it is clear that for any $h\in[H]$,

$$\Lambda_{r,h}^{K+1} \;=\; \sum_{k=1}^{K}\phi_{r,h}^k\left(x_h^k,a_h^k\right)\phi_{r,h}^k\left(x_h^k,a_h^k\right)^\top+\lambda I \;\preceq\; (dH^2K+\lambda)I.$$

Thus,

$$\log\left(\frac{\det\left(\Lambda_{r,h}^{K+1}\right)}{\det\left(\Lambda_{r,h}^1\right)}\right) \;\leq\; \log\left(\frac{\det\left((dH^2K+\lambda)I\right)}{\det(\lambda I)}\right) \;\leq\; d\log\left(\frac{dH^2K+\lambda}{\lambda}\right).$$

Therefore,

$$\sum_{k=1}^{K}\phi_{r,h}^k\left(x_h^k,a_h^k\right)^\top\left(\Lambda_{r,h}^k\right)^{-1}\phi_{r,h}^k\left(x_h^k,a_h^k\right) \;\leq\; 2d\log\left(\frac{dH^2K+\lambda}{\lambda}\right). \tag{32}$$

Similarly, we can show that

$$\sum_{k=1}^{K}\varphi\left(x_h^k,a_h^k\right)^\top\left(\Lambda_h^k\right)^{-1}\varphi\left(x_h^k,a_h^k\right) \;\leq\; 2d\log\left(\frac{dK+\lambda}{\lambda}\right). \tag{33}$$

Applying the above inequalities (32) and (33) to (31) leads to

$$
\begin{aligned}
& \sum_{k=1}^{K}\sum_{h=1}^{H} \min\left(H, (\Gamma_h^k + \Gamma_{r,h}^k)(x_h^k, a_h^k)\right) \\
\leq\; & \beta \sum_{h=1}^{H} \min\left(K, \sum_{k=1}^{K} \left(\varphi(x_h^k, a_h^k)^\top (\Lambda_h^k)^{-1}\varphi(x_h^k, a_h^k)\right)^{1/2} + \left(\phi_{r,h}^k(x_h^k, a_h^k)^\top (\Lambda_{r,h}^k)^{-1}\phi_{r,h}^k(x_h^k, a_h^k)\right)^{1/2}\right) \\
\leq\; & \beta \sum_{h=1}^{H} \left(\left(K\sum_{k=1}^{K} \varphi(x_h^k, a_h^k)^\top (\Lambda_h^k)^{-1}\varphi(x_h^k, a_h^k)\right)^{1/2} + \left(K\sum_{k=1}^{K} \phi_{r,h}^k(x_h^k, a_h^k)^\top (\Lambda_{r,h}^k)^{-1}\phi_{r,h}^k(x_h^k, a_h^k)\right)^{1/2}\right) \\
\leq\; & \beta \sum_{h=1}^{H} \sqrt{K}\left(\left(2d\log\left(\frac{dK+\lambda}{\lambda}\right)\right)^{1/2} + \left(2d\log\left(\frac{dH^2K+\lambda}{\lambda}\right)\right)^{1/2}\right)
\end{aligned}
$$

Finally, we set $\beta = C_1\sqrt{dH^2\log(dT/p)}$ and $\lambda = 1$ to obtain (30). □

Lemma 6 (Matingale Bound)**.** *Fix $p \in (0,1)$. In Algorithm 1, it holds with probability $1-p/2$ that*

$$
\left|M_{r,H,2}^K\right| \;\leq\; 4\sqrt{H^2T\log\left(\frac{4}{p}\right)} \tag{34}
$$

where $T = HK$.

Proof. In the verification of (17) (see Section D.2), we introduce the following martingale,

$$
M_{r,H,2}^K \;=\; \sum_{k=1}^{K}\sum_{h=1}^{H}\left(D_{r,h,1}^k + D_{r,h,2}^k\right)
$$

where

$$
\begin{aligned}
D_{r,h,1}^k \;&=\; \left(\mathcal{I}_h^k\left(Q_{r,h}^k - Q_{r,h}^{\pi^k,k}\right)\right)(x_h^k) - \left(Q_{r,h}^k - Q_{r,h}^{\pi^k,k}\right)(x_h^k, a_h^k) \\
D_{r,h,2}^k \;&=\; \left(\mathbb{P}_h V_{r,h+1}^k - \mathbb{P}_h V_{r,h+1}^{\pi^k,k}\right)(x_h^k, a_h^k) - \left(V_{r,h+1}^k - V_{r,h+1}^{\pi^k,k}\right)(x_{h+1}^k)
\end{aligned}
$$

where $\left(\mathcal{I}_h^k f\right)(x) := \left\langle f(x,\cdot), \pi_h^k(\cdot|x)\right\rangle$.

Due to the truncation in line 11 of Algorithm 2, we know that $Q_{r,h}^k, Q_{r,h}^{\pi^k}, V_{r,h+1}^k, V_{r,h+1}^{\pi^k} \in [0,H]$. This shows that $|D_{r,h,1}^k|, |D_{r,h,2}^k| \leq 2H$ for all $(k,h) \in [K]\times[H]$. Application of the Azuma-Hoeffding inequality yields,

$$
P\left(\left|M_{r,H,2}^K\right| \geq s\right) \;\leq\; 2\exp\left(\frac{-s^2}{16H^2T}\right).
$$

For $p\in(0,1)$, if we set $s = 4H\sqrt{T\log(4/p)}$, then the inequality (34) holds with probability at least $1-p/2$. □

We now are ready to show the desired regret bound. Applying (30) and (34) to the right-hand side of the inequality (27), we have

$$
\text{Regret}(K) \;\leq\; C_3 H^{2.5}\sqrt{T\log|\mathcal{A}|} \;+\; 2C_1\sqrt{2d^2H^3T\log(K+1)\log\left(\frac{dT}{p}\right)} \;+\; 4\sqrt{H^2T\log\left(\frac{4}{p}\right)}
$$

with probability $1-p$ where C_1, C_3 are absolute constants. Then, with probability $1-p$ it holds that

$$
\text{Regret}(K) \;\leq\; CdH^{2.5}\sqrt{T}\log\left(\frac{dT}{p}\right)
$$

where C is an absolute constant.

B.2 Proof of Constraint Violation

In Lemma 3, we have provided a useful upper bound on the total differences that are weighted by the dual update Y^k. To extract the constraint violation, we first refine Lemma 3 as follows.

Lemma 7 ([Policy Improvement: Refined Primal-Dual Mirror Descent Step)**.** *Let Assumptions 1 and 2 hold. In Algorithm 1, if we set $\alpha = \sqrt{\log|\mathcal{A}|}/(H^2\sqrt{K})$, $\theta = 1/K$, and $\eta = 1/\sqrt{K}$, then Then, for any $Y \in [0, \chi]$, with probability $1 - p/2$,*

$$\sum_{k=1}^{K}\left(V_{r,1}^{\pi^\star}(x_1) - V_{r,1}^{k}(x_1)\right) + Y\sum_{k=1}^{K}\left(b - V_{g,1}^{k}(x_1)\right) \;\leq\; C_4 H^{2.5}\sqrt{T\log|\mathcal{A}|} \tag{35}$$

where C_4 is an absolute constant, $T = HK$, and $\chi := H/\gamma$.

Proof. By the dual update in line 9 in Algorithm 1, for any $Y \in [0, \chi]$ we have

$$\begin{aligned}
|Y^{k+1} - Y|^2 \;&=\; \left|\mathrm{Proj}_{[0,\chi]}\left(Y^k + \eta(b - V_{g,1}^k(x_1))\right) - \mathrm{Proj}_{[0,\chi]}(Y)\right|^2 \\
&\leq\; \left|Y^k + \eta(b - V_{g,1}^k(x_1)) - Y\right|^2 \\
&\leq\; \left(Y^k - Y\right)^2 + 2\eta\left(b - V_{g,1}^k(x_1)\right)\left(Y^k - Y\right) + \eta^2 H^2
\end{aligned}$$

where we apply the non-expansiveness of projection in the first inequality and $|b - V_{g,1}^k(x_1)| \leq H$ for the last inequality. By summing the above inequality from $k = 1$ to $k = K$, we have

$$0 \;\leq\; |Y^{K+1} - Y|^2 \;=\; |Y^1 - Y|^2 + 2\eta\sum_{k=1}^{K}\left(b - V_{g,1}^k(x_1)\right)\left(Y^k - Y\right) + \eta^2 H^2 K$$

which implies that

$$\sum_{k=1}^{K}\left(b - V_{g,1}^k(x_1)\right)\left(Y - Y^k\right) \;\leq\; \frac{1}{2\eta}|Y^1 - Y|^2 + \frac{\eta}{2}H^2 K.$$

By adding the above inequality to (26) in Lemma 3 and noting that $V_{g,1}^{\pi^\star,k}(x_1) \geq b$ and the UCB result (19), we have

$$\begin{aligned}
&\sum_{k=1}^{K}\left(V_{r,1}^{\pi^\star}(x_1) - V_{r,1}^{k}(x_1)\right) + Y\sum_{k=1}^{K}\left(b - V_{g,1}^{k}(x_1)\right) \\
&\leq\; \frac{\alpha(1+\chi)^2 H^3(K+1)}{2} + \theta\,(1+\chi)\,H^2(K+1) + \frac{\theta H(K+1)\log|\mathcal{A}|}{\alpha} + \frac{H\log|\mathcal{A}|}{\alpha} \\
&\quad + \frac{1}{2\eta}|Y^1 - Y|^2 + \frac{\eta}{2}H^2 K.
\end{aligned}$$

By taking $\chi = H/\gamma$, and α, θ, η in the lemma, we complete the proof. ∎

According to Lemma 7, we can multiply (18) by $Y \geq 0$ and add it, together with (17), to (35),

$$\begin{aligned}
&\sum_{k=1}^{K}\left(V_{r,1}^{\pi^\star}(x_1) - V_{r,1}^{\pi^k}(x_1)\right) + Y\sum_{k=1}^{K}\left(b - V_{g,1}^{\pi^k}(x_1)\right) \\
&\leq\; C_4 H^{2.5}\sqrt{T\log|\mathcal{A}|} - \sum_{k=1}^{K}\sum_{h=1}^{H}\iota_{r,h}^k(x_h^k, a_h^k) - Y\sum_{k=1}^{K}\sum_{h=1}^{H}\iota_{g,h}^k(x_h^k, a_h^k) + M_{r,H,2}^K + Y M_{g,H,2}^K.
\end{aligned} \tag{36}$$

We now are ready to show the desired constraint violation bound. We note that there exists a policy π' such that $V_{r,1}^{\pi'}(x_1) = \frac{1}{K}\sum_{k=1}^{K}V_{r,1}^{\pi^k}(x_1)$ and $V_{g,1}^{\pi'}(x_1) = \frac{1}{K}\sum_{k=1}^{K}V_{g,1}^{\pi^k}(x_1)$. By the occupancy measure method [5], $V_{r,1}^{\pi^k}(x_1)$ and $V_{g,1}^{\pi^k}(x_1)$ are linear in terms of an occupancy measure induced by policy π^k and initial state x_1. Thus, an average of K occupancy measures is still an occupancy measure that produces policy π' with values $V_{r,1}^{\pi'}(x_1)$ and $V_{g,1}^{\pi'}(x_1)$. Particularly, we take $Y=0$ when $\sum_{k=1}^{K}\left(b-V_{g,1}^{\pi^k}(x_1)\right)<0$; otherwise $Y=\chi$. Therefore, we have

$$
\begin{aligned}
& V_{r,1}^{\pi^\star}(x_1)-\frac{1}{K}\sum_{k=1}^{K}V_{r,1}^{\pi^k}(x_1)\,+\,\chi\left[b-\frac{1}{K}\sum_{k=1}^{K}V_{g,1}^{\pi^k}(x_1)\right]_+\\
&=\,V_{r,1}^{\pi^\star}(x_1)-V_{r,1}^{\pi'}(x_1)\,+\,\chi\left[b-V_{g,1}^{\pi'}(x_1)\right]_+\\
&\leq\,\frac{C_4H^{2.5}\sqrt{T\log|\mathcal{A}|}}{K}\,-\,\frac{1}{K}\sum_{k=1}^{K}\sum_{h=1}^{H}\iota_{r,h}^k(x_h^k,a_h^k)\,-\,\frac{\chi}{K}\sum_{k=1}^{K}\sum_{h=1}^{H}\iota_{g,h}^k(x_h^k,a_h^k)\\
&\quad+\frac{1}{K}M_{r,H,2}^K\,+\,\frac{\chi}{K}\left|M_{g,H,2}^K\right|\\
&\leq\,\frac{C_4H^{2.5}\sqrt{T\log|\mathcal{A}|}}{K}\,+\,\frac{1}{K}\sum_{k=1}^{K}\sum_{h=1}^{H}(\Gamma_h^k+\Gamma_{r,h}^k)(x_h^k,a_h^k)\,+\,\frac{\chi}{K}\sum_{k=1}^{K}\sum_{h=1}^{H}(\Gamma_h^k+\Gamma_{g,h}^k)(x_h^k,a_h^k)\\
&\quad+\frac{1}{K}M_{r,H,2}^K\,+\,\frac{\chi}{K}\left|M_{g,H,2}^K\right|
\end{aligned} \tag{37}
$$

where we apply the UCB result (19) for the last inequality.

Finally, we recall two immediate results of Lemma 5 and Lemma 6. Fix $p\in(0,1)$, the proof of Lemma 5 also shows that with probability $1-p/2$,

$$
\sum_{k=1}^{K}\sum_{h=1}^{H}(\Gamma_h^k+\Gamma_{\diamond,h}^k)(x_h^k,a_h^k)\;\leq\;C_1\sqrt{2d^2H^3T\log\left(K+1\right)\log\left(\frac{dT}{p}\right)} \tag{38}
$$

and the proof of Lemma 6 shows that with probability $1-p/2$,

$$
\left|M_{g,H,2}^K\right|\;\leq\;4\sqrt{H^2T\log\left(\frac{4}{p}\right)}.
$$

If we take $\log|\mathcal{A}|=O(d^2\log^2(dT/p))$, (37) implies that with probability $1-p$ we have

$$
V_{r,1}^{\pi^\star}(x_1)-V_{r,1}^{\pi'}(x_1)\,+\,\chi\left[b-V_{g,1}^{\pi'}(x_1)\right]_+\;\leq\;C_5\,dH^{2.5}\sqrt{T}\log\left(\frac{dT}{p}\right).
$$

where C_5 is an absolute constant. Finally, by noting our choice of $\chi\geq 2Y^\star$, we can apply Lemma 10 to conclude that

$$
\left[\text{Violation}(K)\right]_+\;\leq\;C'\,dH^{2.5}\sqrt{T}\log\left(\frac{dT}{p}\right).
$$

with probability $1-p$, where C' is an absolute constant.

C Further Results on Tabular Case

A special case of Assumption 2 is the tabular CMDP$(\mathcal{S},\mathcal{A},H,\mathbb{P},r,g)$ with $|\mathcal{S}|<\infty$ and $|\mathcal{A}|<\infty$. We take the following feature maps and parameter vectors,

$$
d_1\;=\;|\mathcal{S}|^2|\mathcal{A}|,\;\psi(x,a,x')\;=\;\mathbf{e}_{(x,a,x')}\,\in\,\mathbb{R}^{d_1},\;\theta_h\;=\;\mathbb{P}_h(\,\cdot\,,\cdot\,,\cdot\,)\,\in\,\mathbb{R}^{d_1} \tag{39a}
$$

$$
d_2\;=\;|\mathcal{S}||\mathcal{A}|,\;\varphi(x,a)\;=\;\mathbf{e}_{(x,a)}\,\in\,\mathbb{R}^{d_2},\;\theta_{r,h}\;=\;r_h(\,\cdot\,,\cdot\,)\,\in\,\mathbb{R}^{d_2},\;\theta_{g,h}\;=\;g_h(\,\cdot\,,\cdot\,)\,\in\,\mathbb{R}^{d_2}. \tag{39b}
$$

where $\mathbf{e}_{(x,a,x')}$ is a canonical basis of $\mathbb{R}^{d_1}$ associated with (x,a,x') and $\theta_h\;=\;\mathbb{P}_h(\,\cdot\,,\cdot\,,\cdot\,)$ reads that for any $(x,a,x')\in\mathcal{S}\times\mathcal{A}\times\mathcal{S}$, the (x,a,x')th entry of θ_h is $\mathbb{P}(x'\,|\,x,a)$; similarly we define $\mathbf{e}_{(x,a)}$, $\theta_{r,h}$, and $\theta_{g,h}$. Thus, we can see that

$$
\mathbb{P}_h\left(x'\,|\,x,a\right)\;=\;\left\langle\psi\left(x,a,x'\right),\theta_h\right\rangle,\quad\text{for any }(x,a,x')\,\in\,\mathcal{S}\times\mathcal{A}\times\mathcal{S}
$$

$$
r_h(x,a)\;=\;\langle\varphi(x,a),\theta_{r,h}\rangle\;\;\text{and}\;\;g_h(x,a)\;=\;\langle\varphi(x,a),\theta_{g,h}\rangle,\;\;\text{for any }(x,a)\,\in\,\mathcal{S}\times\mathcal{A}.
$$

We can also verify that

$$\|\theta_h\| \;=\; \left(\sum_{(x,a,x')} |\mathbb{P}_h(x' \,|\, x,a)|^2\right)^{1/2} \;\leq\; \sqrt{|\mathcal{S}|^2|\mathcal{A}|} \;=\; \sqrt{d_1}$$

$$\|\theta_{r,h}\| \;=\; \left(\sum_{(x,a)} (r_h(x,a))^2\right)^{1/2} \;\leq\; \sqrt{|\mathcal{S}||\mathcal{A}|} \;=\; \sqrt{d_2}$$

$$\|\theta_{g,h}\| \;=\; \left(\sum_{(x,a)} (g_h(x,a))^2\right)^{1/2} \;\leq\; \sqrt{|\mathcal{S}||\mathcal{A}|} \;=\; \sqrt{d_2}$$

and for any $V\colon \mathcal{S}\to[0,H]$ and any $(x,a)\in\mathcal{S}\times\mathcal{A}$, we have

$$\left\|\sum_{x'\in\mathcal{S}} \psi(x,a,x')V(x')\right\| \;=\; \left(\sum_{x'\in\mathcal{S}} (V(x'))^2\right)^{1/2} \;\leq\; \sqrt{|\mathcal{S}|}\,H \;\leq\; \sqrt{d_1}\,H.$$

Therefore, the tabular CMDP is a special case of Assumption (2) with $d := \max(d_1, d_2) = |\mathcal{S}|^2|\mathcal{A}|$.

C.1 Tabular Case of Algorithm 1

We now detail Algorithm 1 for the tabular case as follows. Our policy evaluation works with regression feature $\phi_{\diamond,h}^\tau\colon \mathcal{S}\times\mathcal{A}\to\mathbb{R}^{d_2}$,

$$\phi_{\diamond,h}^\tau(x,a) \;=\; \sum_{x'} \psi(x,a,x')V_{\diamond,h+1}^\tau(x'), \quad \text{for any } (x,a)\in\mathcal{S}\times\mathcal{A}$$

where $\diamond = r$ or g. Thus, for any $(\bar{x},\bar{a},\bar{x}')\in\mathcal{S}\times\mathcal{A}\times\mathcal{S}$, the $(\bar{x},\bar{a},\bar{x}')$th entry of $\phi_{\diamond,h}^\tau(x,a)$ is given by

$$\left[\phi_{\diamond,h}^\tau(x,a)\right]_{(\bar{x},\bar{a},\bar{x}')} \;=\; \mathbf{1}\{(x,a) = (\bar{x},\bar{a})\}V_{\diamond,h+1}^\tau(\bar{x}')$$

which shows that $\phi_{\diamond,h}^\tau(x,a)$ is a sparse vector with $|\mathcal{S}|$ nonzero elements at $\{(x,a,x'), x'\in\mathcal{S}\}$ and the (x,a,x')th entry of $\phi_{\diamond,h}^\tau(x,a)$ is $V_{\diamond,h+1}^\tau(x')$. For instance of $\diamond = r$, the regularized least-squares problem (4) becomes

$$\sum_{\tau=1}^{k-1}\left(V_{r,h+1}^\tau(x_{h+1}^\tau) \;-\; \sum_{(x,a,x')} \mathbf{1}\{(x,a) = (x_h^\tau,a_h^\tau)\}V_{r,h+1}^\tau(x')[w]_{(x,a,x')}\right)^2 \;+\; \lambda\,\|w\|_2^2$$

where $[w]_{(x,a,x')}$ is the (x,a,x')th entry of w, and the solution $w_{r,h}^k$ serves as an estimator of the transition kernel $\mathbb{P}_h(\cdot\,|\,\cdot,\cdot)$. On the other hand, since $\varphi(x_h^\tau,a_h^\tau) = \mathbf{e}_{(x_h^\tau,a_h^\tau)}$, the regularized least-squares problem (5) becomes

$$\sum_{\tau=1}^{k-1}\left(r_h(x_h^\tau,a_h^\tau) \;-\; [u]_{(x_h^\tau,a_h^\tau)}\right)^2 \;+\; \lambda\,\|u\|_2^2$$

where $[u]_{(x,a)}$ is the (x,a)th entry of u, the solution $u_{r,h}^k$ gives an estimate of $r_h(x,a)$ as $\varphi(x,a)^\top u_{r,h}^k$. By adding similar UCB bonus terms Γ_h^k, $\Gamma_{r,h}^k\colon \mathcal{S}\times\mathcal{A}\to\mathbb{R}$ given in Algorithm 2, we estimate the action-value function as follows,

$$\begin{aligned}
Q_{r,h}^k(x,a) \;&=\; \min\left([u_{r,h}^k]_{(x,a)} + \phi_{r,h}^k(x,a)^\top w_{r,h}^k + (\Gamma_h^k + \Gamma_{r,h}^k)(x,a),\; H-h+1\right)^+ \\
&=\; \min\left([u_{r,h}^k]_{(x,a)} + \sum_{x'\in\mathcal{S}} V_{r,h+1}^k(x')[w_{r,h}^k]_{(x,a,x')} + (\Gamma_h^k + \Gamma_{r,h}^k)(x,a),\; H-h+1\right)^+
\end{aligned}$$

for any $(x,a)\in\mathcal{S}\times\mathcal{A}$. Thus, $V_{r,h}^k(x) = \langle Q_{r,h}^k(x,\cdot), \pi_h^k(\cdot\,|\,x)\rangle_{\mathcal{A}}$. Similarly, we estimate $g_h(x,a)$ and thus $Q_{g,h}^k(x,a)$ and $V_{g,h}^k(x)$. Using already estimated $\{Q_{r,h}^k(\cdot,\cdot), Q_{g,h}^k(\cdot,\cdot), V_{r,h}^k(\cdot), Q_{g,h}^k(\cdot)\}_{h=1}^H$, we execute the policy improvement and the dual update in Algorithm 1.

We restate the result of Theorem 1 for the tabular case as follows.

Corollary 1 (Regret and Constraint Violation). *For the tabular CMDP with feature maps* (39)*, let Assumption 1 hold. Fix* $p \in (0,1)$*. In Algorithm 1, we set* $\alpha = \sqrt{\log|\mathcal{A}|/(H^2K)}$*,* $\beta = C_1\sqrt{|\mathcal{S}|^2|\mathcal{A}|H^2\log(|\mathcal{S}||\mathcal{A}|T/p)}$*,* $\eta = 1/\sqrt{K}$*,* $\theta = 1/K$*, and* $\lambda = 1$ *where* C_1 *is an absolute constant. Then, the regret and the constraint violation in* (3) *satisfy*

$$\text{Regret}(K) \leq C|\mathcal{S}|^2|\mathcal{A}|H^{2.5}\sqrt{T}\log\left(\frac{|\mathcal{S}||\mathcal{A}|T}{p}\right) \ \textit{and}\ \text{Violation}(K) \leq C'|\mathcal{S}|^2|\mathcal{A}|H^{2.5}\sqrt{T}\log\left(\frac{|\mathcal{S}||\mathcal{A}|T}{p}\right)$$

with probability $1-p$ *where* C *and* C' *are absolute constants.*

Proof. It follows the proof of Theorem 1 by noting that the tabular CMDP is a special linear MDP in Assumption 2, with $d = |\mathcal{S}|^2|\mathcal{A}|$, and we have $\log|\mathcal{A}| \leq O\left(d^2\log(dT/p)\right)$ automatically. ☐

C.2 Further Results: Proof of Theorem 2

As we see in the proof of Theorem 1, our final regret or constraint violation bounds are dominated by the accumulated bonus terms, which come from the design of 'optimism in the face of uncertainty.' This framework provides a powerful flexibility for Algorithm 1 to incorporate other optimistic policy evaluation methods. In what follows, we introduce Algorithm 1 with a variant of optimistic policy evaluation.

We repeat notation for readers' convenience. For any $(h,k) \in [H]\times[K]$, any $(x,a,x') \in \mathcal{S}\times\mathcal{A}\times\mathcal{S}$, and any $(x,a)\in\mathcal{S}\times\mathcal{A}$, we define two visitation counters $n_h^k(x,a,x')$ and $n_h^k(x,a)$ at step h in episode k,

$$n_h^k(x,a,x') \;=\; \sum_{\tau=1}^{k-1}\mathbf{1}\{(x,a,x') = (x_h^\tau,a_h^\tau,a_{h+1}^\tau)\} \;\text{ and }\; n_h^k(x,a) \;=\; \sum_{\tau=1}^{k-1}\mathbf{1}\{(x,a) = (x_h^\tau,a_h^\tau)\}.$$

This allows us to estimate transition kernel $\mathbb{P}_h$, reward function r, and utility function g for episode k by

$$\widehat{\mathbb{P}}_h^k(x'\,|\,x,a) \;=\; \frac{n_h^k(x,a,x')}{n_h^k(x,a)+\lambda}, \;\text{ for all } (x,a,x') \in \mathcal{S}\times\mathcal{A}\times\mathcal{S}$$

$$\widehat{r}_h^k(x,a) \;=\; \frac{1}{n_h^k(x,a)+\lambda}\sum_{\tau=1}^{k-1}\mathbf{1}\{(x,a) = (x_h^\tau,a_h^\tau)\}r_h(x_h^\tau,a_h^\tau), \;\text{ for all } (x,a)\in\mathcal{S}\times\mathcal{A}.$$

$$\widehat{g}_h^k(x,a) \;=\; \frac{1}{n_h^k(x,a)+\lambda}\sum_{\tau=1}^{k-1}\mathbf{1}\{(x,a) = (x_h^\tau,a_h^\tau)\}g_h(x_h^\tau,a_h^\tau), \;\text{ for all } (x,a)\in\mathcal{S}\times\mathcal{A}.$$

where $\lambda > 0$ is the regularization parameter. Moreover, we introduce the bonus term $\Gamma_h^k\colon \mathcal{S}\times\mathcal{A}\to\mathbb{R}$,

$$\Gamma_h^k(x,a) \;=\; \beta\left(n_h^k(x,a)+\lambda\right)^{-1/2}$$

which adapts the counter-based bonus terms in the literature [9, 36], where $\beta > 0$ is to be determined later. Using the estimated transition kernels $\{\widehat{\mathbb{P}}_h^k\}_{h=1}^H$, the estimated reward/utility functions $\{\widehat{r}_h^k, \widehat{g}_h^k\}_{h=1}^H$, and the bonus terms $\{\Gamma_h^k\}_{h=1}^H$, we now can estimate the action-value function via

$$Q_{\diamond,h}^k(x,a) \;=\; \min\left(\widehat{\diamond}_h^k(x,a) + \sum_{x'\in\mathcal{S}}\widehat{\mathbb{P}}_h(x'\,|\,x,a)V_{\diamond,h+1}^k(x') + 2\Gamma_h^k(x,a),\, H-h+1\right)^+$$

for any $(x,a)\in\mathcal{S}\times\mathcal{A}$, where $\diamond = r$ or g. Thus, $V_{\diamond,h}^k(x) = \langle Q_{\diamond,h}^k(x,\cdot),\pi_h^k(\cdot\,|\,x)\rangle_{\mathcal{A}}$. We summarize the above procedure in Algorithm 3. Using already estimated $\{Q_{r,h}^k(\cdot,\cdot), Q_{g,h}^k(\cdot,\cdot)\}_{h=1}^H$, we can execute the policy improvement and the dual update in Algorithm 1.

Similar to Theorem 1, we prove the following regret and constraint violation bounds.

Theorem 3 (Regret and Constraint Violation). *For the tabular CMDP with feature maps* (39)*, let Assumption 1 hold. Fix* $p\in(0,1)$*. In Algorithm 1, we set* $\alpha = \sqrt{\log|\mathcal{A}|/(H^2K)}$*,* $\beta = C_1H\sqrt{|\mathcal{S}|\log(|\mathcal{S}||\mathcal{A}|T/p)}$*,* $\eta = 1/\sqrt{K}$*,* $\theta = 1/K$*, and* $\lambda = 1$ *where* C_1 *is an absolute constant. Then, the regret and the constraint violation in* (3) *satisfy*

$$\text{Regret}(K) \leq C|\mathcal{S}|\sqrt{|\mathcal{A}|H^5T}\log\left(\frac{|\mathcal{S}||\mathcal{A}|T}{p}\right) \quad\textit{and}\quad [\text{Violation}(K)]_+ \leq C'|\mathcal{S}|\sqrt{|\mathcal{A}|H^5T}\log\left(\frac{|\mathcal{S}||\mathcal{A}|T}{p}\right)$$

with probability $1-p$ *where* C *and* C' *are absolute constants.*

Proof. The proof is similar to Theorem 1. Since we only change the policy evaluation, all previous policy improvement results still hold. By Lemma 4, we have

$$\text{Regret}(K) \;=\; C_3 H^{2.5}\sqrt{T\log|\mathcal{A}|} \,+\, \sum_{k=1}^{K}\sum_{h=1}^{H}\left(\mathbb{E}_{\pi^\star}\left[\iota_{r,h}^k(x_h,a_h)\right] - \iota_{r,h}^k(x_h^k,a_h^k)\right) \,+\, M_{r,H,2}^K$$

where $\iota_{r,h}^k$ is the model prediction error given by (11) and $\{M_{r,h,m}^k\}_{(k,h,m)\in[K]\times[H]\times[2]}$ is a martingale adapted to the filtration $\{\mathcal{F}_{h,m}^k\}_{(k,h,m)\in[K]\times[H]\times[2]}$ in terms of time index t defined in (16). By Lemma 6, it holds with probability $1-p/3$ that $|M_{r,H,2}^K| \le 4\sqrt{H^2T\log(4/p)}$. The rest is to bound the double sum term. As shown in Section D.4, with probability $1-p/2$ it holds that for any $(k,h)\in[K]\times[H]$ and $(x,a)\in\mathcal{S}\times\mathcal{A}$,

$$-\,4\Gamma_h^k(x,a) \;\le\; \iota_{r,h}^k(x,a) \;\le\; 0. \tag{40}$$

Together with the choice of Γ_h^k, we have

$$\begin{aligned}
\sum_{k=1}^{K}\sum_{h=1}^{H}\left(\mathbb{E}_{\pi^\star}\left[\iota_{r,h}^k(x_h,a_h)\,|\,x_1\right] - \iota_{r,h}^k(x_h^k,a_h^k)\right) \;&\le\; 4\sum_{k=1}^{K}\sum_{h=1}^{H}\Gamma_h^k(x_h^k,a_h^k)\\
&=\; 4\beta\sum_{k=1}^{K}\sum_{h=1}^{H}\left(n_h^k(x_h^k,a_h^k)+\lambda\right)^{-1/2}.
\end{aligned}$$

Define mapping $\bar\phi$: $\mathcal{S}\times\mathcal{A}\to\mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$ as $\bar\phi(x,a)=\mathbf{e}_{(x,a)}$, we can utilize Lemma 13. For any $(k,h)\in[K]\times[H]$, we have

$$\begin{aligned}
\bar\Lambda_h^k \;&=\; \sum_{\tau=1}^{k-1}\bar\phi(x_h^\tau,a_h^\tau)\bar\phi(x_h^\tau,a_h^\tau)^\top + \lambda I \;\in\; \mathbb{R}^{|\mathcal{S}||\mathcal{A}|\times|\mathcal{S}||\mathcal{A}|}\\
\Gamma_h^k(x,a) \;&=\; \beta\left(n_h^k(x,a)+\lambda\right)^{-1/2} \;=\; \beta\sqrt{\bar\phi(x,a)(\bar\Lambda_h^k)^{-1}\bar\phi(x,a)^\top}
\end{aligned}$$

where $\bar\Lambda_h^k$ is a diagonal matrix whose the (x,a)th diagonal entry is $n_h^k(x,a)+\lambda$. Therefore, we have

$$\begin{aligned}
\sum_{k=1}^{K}\sum_{h=1}^{H}\left(\mathbb{E}_{\pi^\star}\left[\iota_{r,h}^k(x_h,a_h)\right] - \iota_{r,h}^k(x_h^k,a_h^k)\right) \;&\le\; 4\beta\sum_{k=1}^{K}\sum_{h=1}^{H}\left(\bar\phi(x_h^k,a_h^k)(\bar\Lambda_h^k)^{-1}\bar\phi(x_h^k,a_h^k)^\top\right)^{1/2}\\
&\le\; 4\beta\sum_{h=1}^{H}\left(K\sum_{k=1}^{K}\bar\phi(x_h^k,a_h^k)(\bar\Lambda_h^k)^{-1}\bar\phi(x_h^k,a_h^k)^\top\right)^{1/2}\\
&\le\; 4\beta\sqrt{2K}\sum_{h=1}^{H}\log^{1/2}\left(\frac{\det\left(\bar\Lambda_h^{K+1}\right)}{\det\bar\Lambda_h^1}\right)
\end{aligned}$$

where we apply the Cauchy-Schwartz inequality for the second inequality and Lemma 13 for the third inequality. Notice that $(K+\lambda)I\succeq\bar\Lambda_h^K$ and $\bar\Lambda_h^1=\lambda I$. Hence,

$$\text{Regret}(K) \;=\; C_3H^{2.5}\sqrt{T\log|\mathcal{A}|} \,+\, 4\beta\sqrt{2|\mathcal{S}||\mathcal{A}|HT}\sqrt{\log\left(\frac{K+\lambda}{\lambda}\right)} \,+\, 4\sqrt{H^2T\log\left(\frac{6}{p}\right)}.$$

Notice that $\log|\mathcal{A}| \le O\big(|\mathcal{S}|^2|\mathcal{A}|\log^2(|\mathcal{S}||\mathcal{A}|T/p)\big)$. By setting $\lambda=1$ and $\beta = C_1H\sqrt{|\mathcal{S}|\log(|\mathcal{S}||\mathcal{A}|T/p)}$, we conclude the desired regret bound.

For the constraint violation analysis, Lemmas 7 still holds. Similar to (37), we have

$$\begin{aligned}
&V_{r,1}^{\pi^\star}(x_1) - V_{r,1}^{\pi'}(x_1) \,+\, \chi\left[b - V_{g,1}^{\pi'}(x_1)\right]_+\\
&\le\; \frac{C_4H^{2.5}\sqrt{T\log|\mathcal{A}|}}{K} \,+\, \frac{4}{K}\sum_{k=1}^{K}\sum_{h=1}^{H}\Gamma_h^k(x_h^k,a_h^k) \,+\, \frac{4\chi}{K}\sum_{k=1}^{K}\sum_{h=1}^{H}\Gamma_h^k(x_h^k,a_h^k) \,+\, \frac{1}{K}M_{r,H,2}^K \,+\, \frac{\chi}{K}\left|M_{g,H,2}^K\right|
\end{aligned}$$

where $V_{r,1}^{\pi'}(x_1) \;=\; \frac{1}{K}\sum_{k=1}^{K}V_{r,1}^{\pi^k}(x_1)$ and $V_{g,1}^{\pi'}(x_1) \;=\; \frac{1}{K}\sum_{k=1}^{K}V_{g,1}^{\pi^k}(x_1)$. Similar to Lemma 6, it holds with probability $1-p/3$ that $|M_{g,H,2}^K|\le 4\sqrt{H^2T\log(6/p)}$ for $\diamond = r$ or g. As shown in Section D.4, with probability

$1-p/3$ it holds that $-4\Gamma_h^k(x,a) \le \iota_{\diamond,h}^k(x,a) \le 0$ for any $(k,h) \in [K]\times[H]$ and $(x,a) \in \mathcal{S}\times\mathcal{A}$. Therefore, we have

$$
\begin{aligned}
& V_{r,1}^{\pi^\star}(x_1) - V_{r,1}^{\pi'}(x_1) \;+\; \chi\left[b - V_{g,1}^{\pi'}(x_1)\right]_+ \\
& \le\; \frac{C_4 H^{2.5}\sqrt{T\log(|\mathcal{A}|)}}{K} \;+\; \frac{4(1+\chi)\beta\sqrt{2|\mathcal{S}||\mathcal{A}|HT}}{K}\sqrt{\log\left(\frac{K+\lambda}{\lambda}\right)} \;+\; \frac{4(1+\chi)}{K}\sqrt{H^2T\log\left(\frac{6}{p}\right)}
\end{aligned}
$$

which leads to the desired constraint violation bound due to Lemma 10 and we set λ and β as previously. ☐

D Other Verifications

In this section, we collect some verifications for readers' convenience.

D.1 Proof of Formulas (12) and (14)

For any $(k,h) \in [K]\times[H]$, we recall the definitions of $V_{r,h}^{\pi^\star}$ in the Bellman equations (1) and $V_{r,h}^k$ from line 12 in Algorithm 2,

$$
V_{r,h}^{\pi^\star}(x) \;=\; \left\langle Q_h^{\pi^\star}(x,\cdot), \pi_h^\star(\cdot\,|\,x)\right\rangle \;\text{ and }\; V_{r,h}^k(x) \;=\; \left\langle Q_h^k(x,\cdot), \pi_h^k(\cdot\,|\,x)\right\rangle.
$$

We can expand the difference $V_{r,h}^{\pi^\star}(x) - V_{r,h}^k(x)$ into

$$
\begin{aligned}
V_{r,h}^{\pi^\star}(x) \;-\; V_{r,h}^k(x) \;&=\; \left\langle Q_h^{\pi^\star}(x,\cdot), \pi_h^\star(\cdot\,|\,x)\right\rangle \;-\; \left\langle Q_h^k(x,\cdot), \pi_h^k(\cdot\,|\,x)\right\rangle \\
&=\; \left\langle Q_h^{\pi^\star}(x,\cdot) - Q_h^k(x,\cdot), \pi_h^\star(\cdot\,|\,x)\right\rangle \;+\; \left\langle Q_h^k(x,\cdot), \pi_h^\star(\cdot\,|\,x) - \pi_h^k(\cdot\,|\,x)\right\rangle \\
&=\; \left\langle Q_h^{\pi^\star}(x,\cdot) - Q_h^k(x,\cdot), \pi_h^\star(\cdot\,|\,x)\right\rangle \;+\; \xi_h^k(x),
\end{aligned} \tag{41}
$$

where $\xi_h^k(x) := \langle Q_h^k(x,\cdot), \pi_h^\star(\cdot\,|\,x) - \pi_h^k(\cdot\,|\,x)\rangle$.

Recall the equality in the Bellman equations (1) and the model prediction error,

$$
Q_{r,h}^{\pi^\star} \;=\; r_h^k \;+\; \mathbb{P}_h V_{r,h+1}^{\pi^\star} \;\text{ and }\; \iota_{r,h}^k \;=\; r_h \;+\; \mathbb{P}_h V_{r,h+1}^k \;-\; Q_{r,h}^k.
$$

As a result of the above two, it is easy to see that

$$
Q_{r,h}^{\pi^\star} \;-\; Q_{r,h}^k \;=\; \mathbb{P}_h\big(V_{r,h+1}^{\pi^\star} - V_{r,h+1}^k\big) \;+\; \iota_{r,h}^k.
$$

Substituting the above difference into the right-hand side of (41) yields,

$$
V_{r,h}^{\pi^\star}(x) \;-\; V_{r,h}^k(x) \;=\; \left\langle \mathbb{P}_h\big(V_{r,h+1}^{\pi^\star} - V_{r,h+1}^k\big)(x,\cdot), \pi_h^\star(\cdot\,|x)\right\rangle \;+\; \left\langle \iota_{r,h}^k(x,\cdot), \pi_h^\star(\cdot\,|\,x)\right\rangle \;+\; \xi_h^k(x).
$$

which displays a recursive formula over h. Thus, we expand $V_{r,1}^{\pi^\star}(x_1) - V_{r,1}^k(x_1)$ recursively with $x = x_1$ as

$$
\begin{aligned}
V_{r,1}^{\pi^\star}(x_1) \;-\; V_{r,1}^k(x_1) \;&=\; \left\langle \mathbb{P}_1\big(V_{r,2}^{\pi^\star} - V_{r,2}^k\big)(x_1,\cdot), \pi_1^\star(\cdot|x_1)\right\rangle \;+\; \left\langle \iota_{r,1}^k(x_1,\cdot), \pi_1^\star(\cdot|x_1)\right\rangle \;+\; \xi_1^k(x_1) \\
&=\; \left\langle \mathbb{P}_1\left\langle \mathbb{P}_2\big(V_{r,3}^{\pi^\star} - V_{r,3}^k\big)(x_2,\cdot), \pi_2^\star(\cdot|x_2)\right\rangle(x_1,\cdot), \pi_1^\star(\cdot|x_1)\right\rangle \\
&\quad+\; \left\langle \mathbb{P}_1\left\langle \iota_{r,2}^k(x_2,\cdot), \pi_2^\star(\cdot|x_2)\right\rangle(x_1,\cdot), \pi_1^\star(\cdot|x_1)\right\rangle \;+\; \left\langle \iota_{r,1}^k(x_1,\cdot), \pi_1^\star(\cdot|x_1)\right\rangle \\
&\quad+ \left\langle \mathbb{P}_1\xi_2^k(x_1,\cdot), \pi_1^\star(\cdot|x_1)\right\rangle \;+\; \xi_1^k(x_1).
\end{aligned} \tag{42}
$$

For notational simplicity, for any $(k,h) \in [K]\times[H]$, we define an operator $\mathcal{I}_h$ for function $f : \mathcal{S}\times\mathcal{A} \to \mathbb{R}$,

$$
(\mathcal{I}_h f)(x) \;=\; \left\langle f(x,\cdot), \pi_h^\star(\cdot\,|x)\right\rangle.
$$

With this notation, repeating the above recursion (42) over $h \in [H]$ yields

$$
\begin{aligned}
& V_{r,1}^{\pi^\star}(x_1) - V_{r,1}^k(x_1) \\
&=\; \mathcal{I}_1\mathbb{P}_1\mathcal{I}_2\mathbb{P}_2\big(V_{r,3}^{\pi^\star} - V_{r,3}^k\big) + \mathcal{I}_1\mathbb{P}_1\mathcal{I}_2\iota_{r,2}^k + \mathcal{I}_1\iota_{r,1}^k + \mathcal{I}_1\mathbb{P}_1\xi_2^k + \xi_1^k \\
&=\; \mathcal{I}_1\mathbb{P}_1\mathcal{I}_2\mathbb{P}_2\mathcal{I}_3\mathbb{P}_3\big(V_{r,4}^{\pi^\star} - V_{r,4}^k\big) + \mathcal{I}_1\mathbb{P}_1\mathcal{I}_2\mathbb{P}_2\mathcal{I}_3\iota_{r,3}^k + \mathcal{I}_1\mathbb{P}_1\mathcal{I}_2\iota_{r,2}^k + \mathcal{I}_1\iota_{r,1}^k + \mathcal{I}_1\mathbb{P}_1\mathcal{I}_2\mathbb{P}_2\xi_3^k + \mathcal{I}_1\mathbb{P}_1\xi_2^k + \xi_1^k \\
&\;\;\vdots \\
&=\; \left(\prod_{h=1}^{H}\mathcal{I}_h\mathbb{P}_h\right)\big(V_{r,H+1}^{\pi^\star} - V_{r,H+1}^k\big) + \sum_{h=1}^{H}\left(\prod_{i=1}^{h-1}\mathcal{I}_i\mathbb{P}_i\right)\mathcal{I}_h\iota_{r,h}^k + \sum_{h=1}^{H}\left(\prod_{h=1}^{h-1}\mathcal{I}_i\mathbb{P}_i\right)\xi_h^k.
\end{aligned}
$$

Finally, notice that $V_{r,H+1}^{\pi^\star} = V_{r,H+1}^k = 0$, we use the definitions of $\mathbb{P}_h$ and $\mathcal{I}_h$ to conclude (12). Similarly, we can also use the above argument to verify (14).

D.2 Proof of Formulas (17) and (18)

We recall the definition of $V_{r,h}^{\pi^k}$ and define an operator $\mathcal{I}_h^k$ for function $f: \mathcal{S}\times\mathcal{A}\to\mathbb{R}$,

$$V_{r,h}^{\pi^k}(x) \;=\; \left\langle Q_h^{\pi^k}(x,\,\cdot\,), \pi_h^k(\,\cdot\,|\,x)\right\rangle \;\text{ and }\; \left(\mathcal{I}_h^k f\right)(x) \;=\; \left\langle f(x,\,\cdot\,), \pi_h^k(\,\cdot\,|\,x)\right\rangle.$$

We expand the model prediction error $\iota_{r,h}^k$ into,

$$\begin{aligned}
\iota_{r,h}^k(x_h^k,a_h^k) \;&=\; r_h(x_h^k,a_h^k) \,+\, (\mathbb{P}_h V_{r,h+1}^k)(x_h^k,a_h^k) - Q_{r,h}^k(x_h^k,a_h^k)\\
&=\; \left(r_h(x_h^k,a_h^k) + (\mathbb{P}_h V_{r,h+1}^k)(x_h^k,a_h^k) - Q_{r,h}^{\pi^k}(x_h^k,a_h^k)\right) \,+\, \left(Q_{r,h}^{\pi^k}(x_h^k,a_h^k) - Q_{r,h}^k(x_h^k,a_h^k)\right)\\
&=\; \left(\mathbb{P}_h V_{r,h+1}^k - \mathbb{P}_h V_{r,h+1}^{\pi^k}\right)(x_h^k,a_h^k) \,+\, \left(Q_{r,h}^{\pi^k}(x_h^k,a_h^k) - Q_{r,h}^k(x_h^k,a_h^k)\right),
\end{aligned}$$

where we use the Bellman equation $Q_{r,h}^{\pi^k}(x_h^k,a_h^k) = r_h(x_h^k,a_h^k) + (\mathbb{P}_h V_{r,h+1}^{\pi^k})(x_h^k,a_h^k)$ in the last equality. With the above formula, we expand the difference $V_{r,1}^k(x_1) - V_{r,1}^{\pi^k}(x_1)$ into

$$\begin{aligned}
V_{r,h}^k(x_h^k) - V_{r,h}^{\pi^k}(x_h^k) \;&=\; \left(\mathcal{I}_h^k(Q_{r,h}^k - Q_{r,h}^{\pi^k})\right)(x_h^k) \,-\, \iota_{r,h}^k(x_h^k,a_h^k)\\
&\quad +\, \left(\mathbb{P}_h V_{r,h+1}^k - \mathbb{P}_h V_{r,h+1}^{\pi^k}\right)(x_h^k,a_h^k) \,+\, \left(Q_{r,h}^{\pi^k} - Q_{r,h}^k\right)(x_h^k,a_h^k).
\end{aligned}$$

Let

$$\begin{aligned}
D_{r,h,1}^k \;&:=\; \left(\mathcal{I}_h^k(Q_{r,h}^k - Q_{r,h}^{\pi^k})\right)(x_h^k) \,-\, \left(Q_{r,h}^k - Q_{r,h}^{\pi^k}\right)(x_h^k,a_h^k),\\
D_{r,h,2}^k \;&:=\; \left(\mathbb{P}_h V_{r,h+1}^k - \mathbb{P}_h V_{r,h+1}^{\pi^k}\right)(x_h^k,a_h^k) \,-\, \left(V_{r,h+1}^k - V_{r,h+1}^{\pi^k}\right)(x_{h+1}^k).
\end{aligned}$$

Therefore, we have the following recursive formula over h,

$$V_{r,h}^k(x_h^k) - V_{r,h}^{\pi^k}(x_h^k) \;=\; D_{r,h,1}^k \,+\, D_{r,h,2}^k \,+\, \left(V_{r,h+1}^k - V_{r,h+1}^{\pi^k}\right)(x_{h+1}^k) \,-\, \iota_{r,h}^k(x_h^k,a_h^k).$$

Notice that $V_{r,H+1}^{\pi^k} = V_{r,H+1}^k = 0$. Summing the above equality over $h\in[H]$ yields

$$V_{r,1}^k(x_1) \,-\, V_{r,1}^{\pi^k}(x_1) \;=\; \sum_{h=1}^H \left(D_{r,h,1}^k + D_{r,h,2}^k\right) \,-\, \sum_{h=1}^H \iota_{r,h}^k(x_h^k,a_h^k). \tag{43}$$

Following the definitions of $\mathcal{F}_{h,1}^k$ and $\mathcal{F}_{h,2}^k$, we know $D_{r,h,1}^k \in \mathcal{F}_{h,1}^k$ and $D_{r,h,2}^k \in \mathcal{F}_{h,2}^k$. Thus, for any $(k,h)\in[K]\times[H]$,

$$\mathbb{E}\left[D_{r,h,1}^k \,|\, \mathcal{F}_{h-1,2}^k\right] \;=\; 0 \;\text{ and }\; \mathbb{E}\left[D_{r,h,2}^k \,|\, \mathcal{F}_{h,1}^k\right] \;=\; 0.$$

Notice that $t(k,0,2) = t(k-1,H,2) = 2H(k-1)$. Clearly, $\mathcal{F}_{0,2}^k = \mathcal{F}_{H,2}^{k-1}$ for any $k\geq 2$. Let $\mathcal{F}_{0,2}^1$ be empty. We define a martingale sequence,

$$\begin{aligned}
M_{r,h,m}^k \;&=\; \sum_{\tau=1}^{k-1}\sum_{i=1}^H \left(D_{r,i,1}^\tau + D_{r,i,2}^\tau\right) \,+\, \sum_{i=1}^{h-1}\left(D_{r,i,1}^k + D_{r,i,2}^k\right) \,+\, \sum_{\ell=1}^m D_{r,h,\ell}^k\\
&=\; \sum_{(\tau,i,\ell)\,\in\,[K]\times[H]\times[2],\, t(\tau,i,\ell)\,\leq\, t(k,h,m)} D_{r,i,\ell}^\tau,
\end{aligned}$$

where $t(k,h,m) := 2(k-1)H + 2(h-1) + m$ is the time index. Clearly, this martingale is adapted to the filtration $\{\mathcal{F}_{h,m}^k\}_{(k,h,m)\in[K]\times[H]\times[2]}$, and particularly,

$$\sum_{k=1}^K\sum_{h=1}^H (D_{r,h,1}^k + D_{r,h,2}^k) \;=\; M_{r,H,2}^K.$$

Finally, we combine the above martingale with (43) to obtain (17). Similarly, we can show (18).

D.3 Proof of Formula (19)

We recall the definition of the feature map $\phi_{r,h}^k$,

$$\phi_{r,h}^k(x,a) \;=\; \int_{\mathcal{S}} \psi(x,a,x')V_{r,h+1}^k(x')dx'$$

for any $(k,h) \in [K]\times[H]$ and $(x,a)\in\mathcal{S}\times\mathcal{A}$. By Assumption 2, we have

$$\begin{aligned}
\left(\mathbb{P}_h V_{r,h+1}^k\right)(x,a) \;&=\; \int_{\mathcal{S}} \psi\left(x,a,x'\right)^\top \theta_h \cdot V_{r,h+1}^k(x')dx' \\
&=\; \phi_{r,h}^k(x,a)^\top \theta_h \\
&=\; \phi_{r,h}^k(x,a)^\top (\Lambda_{r,h}^k)^{-1}\Lambda_{r,h}^k\theta_h \\
&=\; \phi_{r,h}^k(x,a)^\top (\Lambda_{r,h}^k)^{-1}\left(\sum_{\tau=1}^{k-1}\phi_{r,h}^\tau(x_h^\tau,a_h^\tau)\phi_{r,h}^\tau(x_h^\tau,a_h^\tau)^\top\theta_h + \lambda\theta_h\right) \\
&=\; \phi_{r,h}^k(x,a)^\top (\Lambda_{r,h}^k)^{-1}\left(\sum_{\tau=1}^{k-1}\phi_{r,h}^\tau(x_h^\tau,a_h^\tau)\cdot\left(\mathbb{P}_h V_{r,h+1}^\tau\right)(x_h^\tau,a_h^\tau) + \lambda\theta_h\right)
\end{aligned}$$

where the second equality is due to the definition of $\phi_{r,h}^k$, we exploit $\Lambda_{r,h}^k = \sum_{\tau=1}^{k-1}\phi_{r,h}^\tau(x_h^\tau,a_h^\tau)\phi_{r,h}^\tau(x_h^\tau,a_h^\tau)^\top + \lambda I$ from line 4 of Algorithm 2 in the fourth equality, and we recursively replace $\phi_{r,h}^\tau(x_h^\tau,a_h^\tau)^\top\theta_h$ by $\left(\mathbb{P}_h V_{r,h+1}^\tau\right)(x_h^\tau,a_h^\tau)$ for all $\tau\in[k-1]$ in the last equality.

We recall the update $w_{r,h}^k = (\Lambda_{r,h}^k)^{-1}\sum_{\tau=1}^{k-1}\phi_{r,h}^\tau(x_h^\tau,a_h^\tau)V_{r,h+1}^\tau(x_{h+1}^\tau)$ from line 5 of Algorithm 2. Therefore,

$$\begin{aligned}
&\left|\phi_{r,h}^k(x,a)^\top w_{r,h}^k - \left(\mathbb{P}_h V_{r,h+1}^k\right)(x,a)\right| \\
&= \left|\phi_{r,h}^k(x,a)^\top(\Lambda_{r,h}^k)^{-1}\sum_{\tau=1}^{k-1}\phi_{r,h}^\tau(x_h^\tau,a_h^\tau)\cdot\left(V_{r,h+1}^\tau(x_{h+1}^\tau) - \left(\mathbb{P}_h V_{r,h+1}^\tau\right)(x_h^\tau,a_h^\tau)\right)\right| \\
&\quad + \left|\lambda\cdot\phi_{r,h}^k(x,a)^\top(\Lambda_{r,h}^k)^{-1}\theta_h\right| \\
&\leq \left(\phi_{r,h}^k(x,a)^\top(\Lambda_{r,h}^k)^{-1}\phi_{r,h}^k(x,a)\right)^{1/2}\left\|\sum_{\tau=1}^{k-1}\phi_{r,h}^\tau(x_h^\tau,a_h^\tau)\cdot\left(V_{r,h+1}^\tau(x_{h+1}^\tau) - \left(\mathbb{P}_h V_{r,h+1}^\tau\right)(x_h^\tau,a_h^\tau)\right)\right\|_{(\Lambda_{r,h}^k)^{-1}} \\
&\quad + \lambda\left(\phi_{r,h}^k(x,a)^\top(\Lambda_{r,h}^k)^{-1}\phi_{r,h}^k(x,a)\right)^{1/2}\|\theta_h\|_{(\Lambda_{r,h}^k)^{-1}}
\end{aligned}$$

for any $(k,h)\in[K]\times[H]$ and $(x,a)\in\mathcal{S}\times\mathcal{A}$, where we apply the Cauchy-Schwarz inequality twice in the inequality. By Lemma 12, set $\lambda = 1$, with probability $1-p/2$ it holds that

$$\left\|\sum_{\tau=1}^{k-1}\phi_{r,h}^\tau(x_h^\tau,a_h^\tau)\cdot\left(V_{r,h+1}^\tau(x_{h+1}^\tau) - \left(\mathbb{P}_h V_{r,h+1}^\tau\right)(x_h^\tau,a_h^\tau)\right)\right\|_{(\Lambda_{r,h}^k)^{-1}} \;\leq\; C\sqrt{dH^2\log\left(\frac{dT}{p}\right)}.$$

Also notice that $\Lambda_{r,h}^k \succeq \lambda I$ and $\|\theta_h\|\leq\sqrt{d}$, thus $\|\theta_h\|_{(\Lambda_{r,h}^k)^{-1}}\leq\sqrt{\lambda d}$. Thus, by taking an appropriate absolute constant C, we obtain that

$$\left|\phi_{r,h}^k(x,a)^\top w_{r,h}^k - \left(\mathbb{P}_h V_{r,h+1}^k\right)(x,a)\right| \;\leq\; C\left(\phi_{r,h}^k(x,a)^\top(\Lambda_{r,h}^k)^{-1}\phi_{r,h}^k(x,a)\right)^{1/2}\sqrt{dH^2\log\left(\frac{dT}{p}\right)}$$

for any $(k,h)\in[K]\times[H]$ and $(x,a)\in\mathcal{S}\times\mathcal{A}$ under the event of Lemma 12.

We now set $C>1$ and $\beta = C\sqrt{dH^2\log\left(dT/p\right)}$. By the exploration bonus $\Gamma_{r,h}^k$ in line 7 of Algorithm 2, with probability $1-p/2$ it holds that

$$\left|\phi_{r,h}^k(x,a)^\top w_{r,h}^k - \left(\mathbb{P}_h V_{r,h+1}^k\right)(x,a)\right| \;\leq\; \Gamma_{r,h}^k(x,a) \tag{44}$$

for any $(k,h) \in [K] \times [H]$ and $(x,a) \in \mathcal{S} \times \mathcal{A}$.

We note that reward/utility functions are fixed over episodes, $r_h(x_h^\tau, a_h^\tau) := \varphi(x_h^\tau, a_h^\tau)^\top \theta_{r,h}$ For the difference $\varphi(x,a)^\top u_{r,h}^k - r_h(x,a)$, we have

$$
\begin{aligned}
& \left| \varphi(x,a)^\top u_{r,h}^k - r_h(x,a) \right| \\
= \;& \left| \varphi(x,a)^\top u_{r,h}^k - \varphi(x,a)^\top \theta_{r,h} \right| \\
= \;& \left| \varphi(x,a)^\top (\Lambda_h^k)^{-1} \left(\sum_{\tau=1}^{k-1} \varphi(x_h^\tau, a_h^\tau) r_h(x_h^\tau, a_h^\tau) - \Lambda_h^k \, \theta_{r,h} \right) \right| \\
= \;& \left| \varphi(x,a)^\top (\Lambda_h^k)^{-1} \left(\sum_{\tau=1}^{k-1} \varphi(x_h^\tau, a_h^\tau) \left(r_h(x_h^\tau, a_h^\tau) - \varphi(x_h^\tau, a_h^\tau)^\top \theta_{r,h} \right) + \lambda \theta_{r,h} \right) \right| \\
= \;& \lambda \left| \varphi(x,a)^\top (\Lambda_h^k)^{-1} \theta_{r,h} \right| \\
\leq \;& \lambda \left(\varphi(x,a)^\top (\Lambda_h^k)^{-1} \varphi(x,a) \right)^{1/2} \|\theta_{r,h}\|_{(\Lambda_h^k)^{-1}}
\end{aligned}
$$

where we apply the Cauchy-Schwartz inequality in the inequality. Notice that $\Lambda_h^k \succeq \lambda I$ and $\|\theta_{r,h}\| \leq \sqrt{d}$, thus $\|\theta_{r,h}\|_{(\Lambda_h^k)^{-1}} \leq \sqrt{\lambda d}$. Hence, if we set $\lambda = 1$ and $\beta = C\sqrt{dH^2 \log(dT/p)}$, then any $(k,h) \in [K] \times [H]$ and $(x,a) \in \mathcal{S} \times \mathcal{A}$,

$$
\left| \varphi(x,a)^\top u_{r,h}^k \, - \, r_h(x,a) \right| \; \leq \; \Gamma_h^k(x,a). \tag{45}
$$

We recall the model prediction error $\iota_{r,h}^k := r_h + \mathbb{P}_h V_{r,h+1}^k - Q_{r,h}^k$ and the estimated state-action value function $Q_{r,h}^k$ in line 11 of Algorithm 2,

$$
Q_{r,h}^k(x,a) \; = \; \min \left(\varphi(x,a)^\top u_{r,h}^k + \phi_{r,h}^k(x,a)^\top w_{r,h}^k + (\Gamma_h^k + \Gamma_{r,h}^k)(x,a), \, H - h + 1 \right)^+
$$

for any $(k,h) \in [K] \times [H]$ and $(x,a) \in \mathcal{S} \times \mathcal{A}$. By (44) and (45), we first have

$$
\phi_{r,h}^k(x,a)^\top w_{r,h}^k \, + \, \Gamma_{r,h}^k(x,a) \; \geq \; 0 \quad \text{and} \quad \varphi(x,a)^\top u_{r,h}^k \, + \, \Gamma_h^k(x,a) \; \geq \; 0.
$$

Then, we can show that

$$
\begin{aligned}
& -\iota_{r,h}^k(x,a) \\
= \;& Q_{r,h}^k(x,a) \, - \, (r_h + \mathbb{P}_h V_{r,h+1}^k)(x,a) \\
\leq \;& \varphi(x,a)^\top u_{r,h}^k \, + \, \phi_{r,h}^k(x,a)^\top w_{r,h}^k \, + \, (\Gamma_h^k + \Gamma_{r,h}^k)(x,a) \, - \, (r_h^k + \mathbb{P}_h V_{r,h+1}^k)(x,a) \\
\leq \;& (\varphi(x,a)^\top u_{r,h}^k - r_h(x,a)) + \Gamma_h^k(x,a) + 2\Gamma_{r,h}^k(x,a)
\end{aligned} \tag{46}
$$

for any $(k,h) \in [K] \times [H]$ and $(x,a) \in \mathcal{S} \times \mathcal{A}$.

Therefore, (46) reduces to

$$
-\iota_{r,h}^k(x,a) \; \leq \; 2\Gamma_h^k(x,a) + 2\Gamma_{r,h}^k(x,a) \; = \; 2(\Gamma_h^k + \Gamma_{r,h}^k)(x,a).
$$

On the other hand, notice that $(r_h^k + \mathbb{P}_h V_{r,h+1}^k)(x,a) \leq H - h + 1$, thus

$$
\begin{aligned}
& \iota_{r,h}^k(x,a) \\
= \;& (r_h + \mathbb{P}_h V_{r,h+1}^k)(x,a) \, - \, Q_{r,h}^k(x,a) \\
\leq \;& (r_h + \mathbb{P}_h V_{r,h+1}^k)(x,a) \, - \, \min \left(\varphi(x,a)^\top u_{r,h}^k + \phi_{r,h}^k(x,a)^\top w_{r,h}^k + (\Gamma_h^k + \Gamma_{r,h}^k)(x,a), \, H - h + 1 \right)^+ \\
\leq \;& \max \left(r_h(x,a) - \varphi(x,a)^\top u_{r,h}^k - \Gamma_h^k(x,a) + (\mathbb{P}_h V_{r,h+1}^k)(x,a) - \phi_{r,h}^k(x,a)^\top w_{r,h}^k - \Gamma_{r,h}^k(x,a), \, 0 \right)^+ \\
\leq \;& 0
\end{aligned}
$$

for any $(k,h) \in [K] \times [H]$ and $(x,a) \in \mathcal{S} \times \mathcal{A}$.

Therefore, we have proved that with probability $1 - p/2$ it holds that

$$-2(\Gamma_h^k + \Gamma_{r,h}^k)(x,a) \;\leq\; \iota_{r,h}^k(x,a) \;\leq\; 0$$

for any $(k,h) \in [K] \times [H]$ and $(x,a) \in \mathcal{S} \times \mathcal{A}$.

Similarly, we can show another inequality $-2(\Gamma_h^k + \Gamma_{g,h}^k)(x,a) \leq \iota_{g,h}^k(x,a) \leq 0$.

D.4 Proof of Formula (40)

Let $\mathcal{V} = \{V : \mathcal{S} \to [0,H]\}$ be a set of bounded function on $\mathcal{S}$. Fo any $V \in \mathcal{V}$, we consider the difference between $\sum_{x' \in \mathcal{S}} \widehat{\mathbb{P}}_h^k(x' \,|\, \cdot, \cdot) V(x')$ and $\sum_{x' \in \mathcal{S}} \mathbb{P}_h(x' \,|\, \cdot, \cdot) V(x')$ as follows,

$$\begin{aligned}
& \left(n_h^k(x,a) + \lambda\right)^{1/2} \left| \sum_{x' \in \mathcal{S}} \left(\widehat{\mathbb{P}}_h^k(x' \,|\, x,a) V(x') - \mathbb{P}_h(x' \,|\, x,a) V(x') \right) \right| \\
&= \left(n_h^k(x,a) + \lambda\right)^{-1/2} \left| \sum_{x' \in \mathcal{S}} n_h^k(x,a,x') V(x') - (n_h^k(x,a) + \lambda)(\mathbb{P}_h V)(x,a) \right| \\
&\leq \left(n_h^k(x,a) + \lambda\right)^{-1/2} \left| \sum_{x' \in \mathcal{S}} n_h^k(x,a,x') V(x') - n_h^k(x,a)(\mathbb{P}_h V)(x,a) \right| \\
&\quad + \left(n_h^k(x,a) + \lambda\right)^{-1/2} \left| \lambda (\mathbb{P}_h V)(x,a) \right| \\
&= \left(n_h^k(x,a) + \lambda\right)^{-1/2} \left| \sum_{\tau=1}^{k-1} \mathbf{1}\{(x,a) = (x_h^\tau, a_h^\tau)\} \left(V(x_{h+1}^\tau) - (\mathbb{P}_h V)(x,a) \right) \right| \\
&\quad + \left(n_h^k(x,a) + \lambda\right)^{-1/2} \left| \lambda (\mathbb{P}_h V)(x,a) \right|
\end{aligned} \tag{47}$$

for any $(k,h) \in [K] \times [H]$ and $(x,a) \in \mathcal{S} \times \mathcal{A}$, where we apply the triangle inequality for the inequality.

Let $\eta_h^\tau := V(x_{h+1}^\tau) - (\mathbb{P}_h V)(x_h^\tau, a_h^\tau)$. Conditioning on the filtration $\mathcal{F}_{h,1}^k$, η_h^τ is a zero-mean and $H/2$-subGaussian random variable. By Lemma 11, we use $Y = \lambda I$ and $X_\tau = \mathbf{1}\{(x,a) = (x_h^\tau, a_h^\tau)\}$ and thus with probability at least $1 - \delta$ it holds that

$$\begin{aligned}
& \left(n_h^k(x,a) + \lambda\right)^{-1/2} \left| \sum_{\tau=1}^{k-1} \mathbf{1}\{(x,a) = (x_h^\tau, a_h^\tau)\} \left(V(x_{h+1}^\tau) - (\mathbb{P}_h V)(x,a) \right) \right| \\
&\leq \sqrt{\frac{H^2}{2} \log\left(\frac{\left(n_h^k(x,a) + \lambda\right)^{1/2} \lambda^{-1/2}}{\delta/H} \right)} \\
&\leq \sqrt{\frac{H^2}{2} \log\left(\frac{T}{\delta} \right)}
\end{aligned}$$

for any $(k,h) \in [K] \times [H]$. Also, since $0 \leq V \leq H$, we have

$$\left(n_h^k(x,a) + \lambda\right)^{-1/2} \left| \lambda (\mathbb{P}_h V)(x,a) \right| \;\leq\; \sqrt{\lambda} H.$$

By returning to (47) and setting $\lambda = 1$, with probability at least $1 - \delta$ it holds that

$$\left(n_h^k(x,a) + \lambda\right) \left| \sum_{x' \in \mathcal{S}} \left(\widehat{\mathbb{P}}_h^k(x' \,|\, x,a) V(x') - \mathbb{P}_h(x' \,|\, x,a) V(x') \right) \right|^2 \;\leq\; H^2 \left(\log\left(\frac{T}{\delta} \right) + 2 \right) \tag{48}$$

for any $k \geq 1$.

Let $d(V,V') = \max_{x \in \mathcal{S}} |V(x) - V'(x)|$ be a distance on $\mathcal{V}$. For any ϵ, an ϵ-covering $\mathcal{V}_\epsilon$ of $\mathcal{V}$ with respect to distance $d(\cdot,\cdot)$ satisfies

$$|\mathcal{V}_\epsilon| \;\leq\; \left(1 + \frac{2\sqrt{|\mathcal{S}|} H}{\epsilon} \right)^{|\mathcal{S}|}.$$

Thus, for any $V \in \mathcal{V}$, there exists $V' \in \mathcal{V}_\epsilon$ such that $\max_{x \in \mathcal{S}} |V(x) - V'(x)| \leq \epsilon$. By the triangle inequality, we have

$$
\begin{aligned}
&\left(n_h^k(x,a)+\lambda\right)^{1/2}\left|\sum_{x' \in \mathcal{S}}\left(\widehat{\mathbb{P}}_h^k(x' \,|\, x,a)V(x') - \mathbb{P}_h(x' \,|\, x,a)V(x')\right)\right| \\
=\;& \left(n_h^k(x,a)+\lambda\right)^{1/2}\left|\sum_{x' \in \mathcal{S}}\left(\widehat{\mathbb{P}}_h^k(x' \,|\, x,a)V'(x') - \mathbb{P}_h(x' \,|\, x,a)V'(x')\right)\right| \\
&+ \left(n_h^k(x,a)+\lambda\right)^{1/2}\left|\sum_{x' \in \mathcal{S}}\left(\widehat{\mathbb{P}}_h^k(x' \,|\, x,a)(V(x')-V'(x')) - \mathbb{P}_h(x' \,|\, x,a)(V(x')-V'(x'))\right)\right| \\
\leq\;& \left(n_h^k(x,a)+\lambda\right)^{1/2}\left|\sum_{x' \in \mathcal{S}}\left(\widehat{\mathbb{P}}_h^k(x' \,|\, x,a)V'(x') - \mathbb{P}_h(x' \,|\, x,a)V'(x')\right)\right| \\
&+ 2\left(n_h^k(x,a)+\lambda\right)^{-1/2}\epsilon.
\end{aligned}
$$

Furthermore, we choose $\delta = (p/3) \,/\, (|\mathcal{V}_\epsilon||\mathcal{S}||\mathcal{A}|)$ and take a union bound over $V \in \mathcal{V}_\epsilon$ and $(x,a) \in \mathcal{S}\times\mathcal{A}$. By (48), with probability at least $1-p/2$ it holds that

$$
\begin{aligned}
&\sup_{V \in \mathcal{V}}\left\{\left(n_h^k(x,a)+\lambda\right)^{1/2}\left|\sum_{x' \in \mathcal{S}}\left(\widehat{\mathbb{P}}_h^k(x' \,|\, x,a)V(x') - \mathbb{P}_h(x' \,|\, x,a)V(x')\right)\right|\right\} \\
\leq\;& \sqrt{H^2\left(\log\left(\frac{T}{\delta}\right)+2\right)} \;+\; 2\left(n_h^k(x,a)+\lambda\right)^{-1/2}\frac{H}{K} \\
\leq\;& \sqrt{2H^2\left(\log|\mathcal{V}_\epsilon| + \log\left(\frac{2|\mathcal{S}||\mathcal{A}|T}{p}\right)+2\right)} \;+\; 2\left(n_h^k(x,a)+\lambda\right)^{-1/2}\frac{H}{K} \\
\leq\;& C_1 H\sqrt{|\mathcal{S}|\log\left(\frac{|\mathcal{S}||\mathcal{A}|T}{p}\right)} \;:=\; \beta
\end{aligned}
$$

for all (k,h) and (x,a), where C_1 is an absolute constant. We recall our choice of Γ_h^k and β. Hence, with probability at least $1-p/2$ it holds that

$$
\left|\sum_{x' \in \mathcal{S}}\left(\widehat{\mathbb{P}}_h^k(x' \,|\, x,a)V(x') - \mathbb{P}_h(x' \,|\, x,a)V(x')\right)\right| \;\leq\; \beta\left(n_h^k(x,a)+\lambda\right)^{-1/2} \;:=\; \Gamma_h^k(x,a)
$$

for any $(k,h) \in [K]\times[H]$ and $(x,a) \in |\mathcal{S}|\times|\mathcal{A}|$, where $\beta := C_1 H\sqrt{|\mathcal{S}|\log(|\mathcal{S}||\mathcal{A}|T/p)}$.

We recall the definition $r_h(x,a) = \mathbf{e}_{(x,a)}^\top \theta_{r,h}$. By our estimation $\widehat{r}_h^k(x,a)$ in Algorithm 3, we have

$$
\widehat{r}_h^k(x,a) \;=\; \frac{1}{n_h^k(x,a)+\lambda}\sum_{\tau=1}^{k-1}\mathbf{1}\{(x,a) = (x_h^\tau, a_h^\tau)\}[\theta_{r,h}]_{(x_h^\tau, a_h^\tau)}
$$

and thus

$$
\begin{aligned}
&\left|\widehat{r}_h^k(x,a) - r_h(x,a)\right| \\
=\;& \left|\widehat{r}_h^k(x,a) - [\theta_{r,h}]_{(x,a)}\right| \\
=\;& \left(n_h^k(x,a)+\lambda\right)^{-1}\left|\sum_{\tau=1}^{k-1}\mathbf{1}\{(x,a) = (x_h^\tau, a_h^\tau)\}\left([\theta_{r,h}]_{(x_h^\tau,a_h^\tau)} - [\theta_{r,h}]_{(x,a)}\right) - \lambda[\theta_{r,h}]_{(x,a)}\right| \\
=\;& \left(n_h^k(x,a)+\lambda\right)^{-1}\left|\lambda[\theta_{r,h}]_{(x,a)}\right| \\
\leq\;& \left(n_h^k(x,a)+\lambda\right)^{-1}\lambda \\
\leq\;& \left(n_h^k(x,a)+\lambda\right)^{-1/2}\lambda \\
\leq\;& \Gamma_h^k(x,a)
\end{aligned}
$$

where we utilize $\lambda = 1$ and $\beta \geq 1$ in the inequalities.

We now are ready to check the model prediction error $\iota_{r,h}^k$ defined by (11),

$$
\begin{aligned}
& -\iota_{r,h}^k(x,a) \\
= \;& Q_{r,h}^k(x,a) - (r_h + \mathbb{P}_h V_{r,h+1}^k)(x,a) \\
\leq \;& \widehat{r}_h^k(x,a) + \textstyle\sum_{x' \in \mathcal{S}} \widehat{\mathbb{P}}_h^k(x' \,|\, x,a) V_{r,h+1}^k(x') + 2\Gamma_h^k(x,a) - (r_h + \mathbb{P}_h V_{r,h+1}^k)(x,a) \\
\leq \;& 4\Gamma_h^k(x,a)
\end{aligned}
$$

for any $(x,a) \in \mathcal{S} \times \mathcal{A}$. On the other hand, notice that $(r_h + \mathbb{P}_h V_{r,h+1}^k)(x,a) \leq H - h + 1$, thus

$$
\begin{aligned}
& \iota_{r,h}^k(x,a) \\
= \;& (r_h + \mathbb{P}_h V_{r,h+1}^k)(x,a) - Q_{r,h}^k(x,a) \\
\leq \;& (r_h + \mathbb{P}_h V_{r,h+1}^k)(x,a) - \min\big(\widehat{r}_h^k(x,a) + \textstyle\sum_{x' \in \mathcal{S}} \widehat{\mathbb{P}}_h^k(x' \,|\, x,a) V_{r,h+1}^k(x') + 2\Gamma_h^k(x,a),\, H - h + 1\big)^+ \\
\leq \;& \max\big((r_h - \widehat{r}_h)(x,a) - \Gamma_h^k(x,a) + (\mathbb{P}_h V_{r,h+1}^k)(x,a) - \textstyle\sum_{x' \in \mathcal{S}} \widehat{\mathbb{P}}_h^k(x' \,|\, x,a) V_{r,h+1}^k(x') - \Gamma_h^k(x,a),\, 0\big)^+ \\
\leq \;& 0
\end{aligned}
$$

for any $(k,h) \in [K] \times [H]$ and $(x,a) \in \mathcal{S} \times \mathcal{A}$. Hence, we complete the proof of (40).

E Supporting Lemmas from Optimization

We collect some standard results from the literature for readers' convenience. We rephrase them for our constrained problem (2),

$$
\underset{\pi \in \Delta(\mathcal{A} \,|\, \mathcal{S}, H)}{\text{maximize}} \;\; V_{r,1}^{\pi}(x_1) \;\; \text{subject to} \;\; V_{g,1}^{\pi}(x_1) \;\geq\; b
$$

in which we maximize over all policies and $b \in (0, H]$. Let the optimal solution be $\pi^\star$ such that

$$
V_{r,1}^{\pi^\star}(x_1) \;=\; \underset{\pi \in \Delta(\mathcal{A} \,|\, \mathcal{S}, H)}{\text{maximize}} \;\{ V_{r,1}^{\pi}(x_1) \,|\, V_{g,1}^{\pi}(x_1) \;\geq\; b \}.
$$

Let the Lagrangian be $\mathcal{L}(\pi, Y) := V_{r,1}^{\pi}(x_1) + Y(V_{g,1}^{\pi}(x_1) - b)$, where $Y \geq 0$ is the Lagrange multiplier or dual variable. The associated dual function is defined as

$$
\mathcal{D}(Y) \;:=\; \underset{\pi \in \Delta(\mathcal{A} \,|\, \mathcal{S}, H)}{\text{maximize}} \;\mathcal{L}(\pi, Y) := V_{r,1}^{\pi}(x_1) \,+\, Y(V_{g,1}^{\pi}(x_1) - b)
$$

and the optimal dual is $Y^\star := \text{argmin}_{Y \geq 0} \mathcal{D}(Y)$,

$$
\mathcal{D}(Y^\star) \;:=\; \underset{\lambda \geq 0}{\text{minimize}} \;\mathcal{D}(Y)
$$

We recall that the problem (2) enjoys strong duality under the strict feasibility condition (also called Slater condition). The proof follows [52, Proposition 1] in finite-horizon.

Assumption 3 (Slater Condition)**.** *There exists* $\gamma > 0$ *and* $\bar{\pi}$ *such that* $V_{g,1}^{\bar{\pi}}(x_1) - b \geq \gamma$.

Lemma 8 (Strong Duality)**.** *[52, Proposition 1] If the Slater condition holds, then the strong duality holds,*

$$
V_{r,1}^{\pi^\star}(x_1) \;=\; \mathcal{D}(Y^\star).
$$

It is implied by the strong duality that the optimal solution to the dual problem: $\text{minimize}_{Y \geq 0} \;\mathcal{D}(Y)$ is obtained at $Y^\star$. Denote the set of all optimal dual variables as $\Lambda^\star$.

Under the Slater condition, a useful property of the dual variable is that the sublevel sets are bounded [12, Section 8.5].

Lemma 9 (Boundedness of Sublevel Sets of the Dual Function)**.** *Let the Slater condition hold. Fix $C \in \mathbb{R}$. For any $Y \in \{Y \geq 0 \,|\, \mathcal{D}(Y) \leq C\}$, it holds that*

$$Y \;\leq\; \frac{1}{\gamma}\left(C - V_{r,1}^{\bar{\pi}}(x_1)\right).$$

Proof. By $Y \in \{Y \geq 0 \,|\, \mathcal{D}(Y) \leq C\}$,

$$C \;\geq\; \mathcal{D}(Y) \;\geq\; V_{r,1}^{\bar{\pi}}(x_1) + Y\,(V_{g,1}^{\bar{\pi}}(x_1) - b) \;\geq\; V_r^{\bar{\pi}}(\rho) + Y\,\gamma$$

where we utilize the Slater point $\bar{\pi}$ in the last inequality. We complete the proof by noting $\gamma > 0$. □

Corollary 2 (Boundedness of $Y^\star$)**.** *If we take $C = V_{r,1}^{\pi^\star}(x_1) = \mathcal{D}(Y^\star)$, then $\Lambda^\star = \{Y \geq 0 \,|\, \mathcal{D}(Y) \leq C\}$. Thus, for any $Y \in \Lambda^\star$,*

$$Y \;\leq\; \frac{1}{\gamma}\left(V_{r,1}^{\pi^\star}(x_1) - V_{r,1}^{\bar{\pi}}(x_1)\right).$$

Another useful theorem from the optimization [12, Section 3.5] is given as follows. It describes that the constraint violation $b - V_{g,1}^{\pi}(x_1)$ can be bounded similarly even if we have some weak bound. We next state and prove it for our problem, which is used in our constraint violation analysis in Section B.

Lemma 10 (Constraint Violation)**.** *Let the Slater condition hold and $Y^\star \in \Lambda^\star$. Let $C^\star \geq 2Y^\star$. Assume that $\pi \in \Delta(\mathcal{A} \,|\, \mathcal{S}, H)$ satisfies*

$$V_{r,1}^{\pi^\star}(x_1) \;-\; V_{r,1}^{\pi}(x_1) \;+\; C^\star\left[b - V_{g,1}^{\pi}(x_1)\right]_+ \;\leq\; \delta.$$

Then,

$$\left[b - V_{g,1}^{\pi}(x_1)\right]_+ \;\leq\; \frac{2\delta}{C^\star}$$

where $[x]_+ = \max(x, 0)$.

Proof. Let

$$v(\tau) \;=\; \underset{\pi \,\in\, \Delta(\mathcal{A}\,|\,\mathcal{S},H)}{\text{maximize}} \;\{\, V_{r,1}^{\pi}(x_1) \,|\, V_{g,1}^{\pi}(x_1) \;\geq\; b + \tau \,\}.$$

By definition, $v(0) = V_{r,1}^{\pi^\star}(x_1)$. It has been shown as a special case of [52, Proposition 1] that $v(\tau)$ is concave. First, we show that $-Y^\star \in \partial v(0)$. By the Lagrangian and the strong duality,

$$\mathcal{L}(\pi, Y^\star) \;\leq\; \underset{\pi \,\in\, \Delta(\mathcal{A}\,|\,\mathcal{S},H)}{\text{maximize}} \;\mathcal{L}(\pi, Y^\star) \;=\; \mathcal{D}(Y^\star) \;=\; V_{r,1}^{\pi^\star}(x_1) \;=\; v(0), \;\;\text{for all } \pi \,\in\, \Delta(\mathcal{A} \,|\, \mathcal{S}, H).$$

For any $\pi \in \{\pi \in \Delta(\mathcal{A} \,|\, \mathcal{S}, H) \,|\, V_{g,1}^{\pi}(x_1) \geq b + \tau\}$,

$$\begin{aligned}
v(0) - \tau Y^\star \;\;&\geq\;\; \mathcal{L}(\pi, Y^\star) - \tau Y^\star \\
&=\;\; V_{r,1}^{\pi}(x_1) + Y^\star(V_{g,1}^{\pi}(x_1) - b) - \tau Y^\star \\
&=\;\; V_{r,1}^{\pi}(x_1) + Y^\star(V_{g,1}^{\pi}(x_1) - b - \tau) \\
&\geq\;\; V_{r,1}^{\pi}(x_1).
\end{aligned}$$

If we maximize the right-hand side of above inequality over $\pi \in \{\pi \in \Delta(\mathcal{A} \,|\, \mathcal{S}, H) \,|\, V_{g,1}^{\pi}(x_1) \geq b + \tau\}$, then

$$v(0) - \tau Y^\star \;\geq\; v(\tau)$$

which show that $-Y^\star \in \partial v(0)$. On the other hand, if we take $\tau = \bar{\tau} := -(b - V_{g,1}^{\bar{\pi}}(x_1))_+$, then

$$V_{r,1}^{\bar{\pi}}(x_1) \;\leq\; V_{r,1}^{\pi^\star}(x_1) \;=\; v(0) \;\leq\; v(\bar{\tau}).$$

Combing the above two yields

$$V_{r,1}^{\bar{\pi}}(x_1) - V_{r,1}^{\star}(x_1) \;\leq\; -\bar{\tau} Y^\star.$$

Thus,

$$
\begin{aligned}
(C^\star - Y^\star)\,|\bar\tau| &= -Y^\star\,|\bar\tau| + C^\star\,|\bar\tau| \\
&= \bar\tau Y^\star + C^\star\,|\bar\tau| \\
&\leq V_{r,1}^{\pi^\star}(x_1) - V_{r,1}^{\bar\pi}(x_1) + C^\star\,|\bar\tau|\,.
\end{aligned}
$$

By our assumption and $\bar\tau = \left[b - V_g^{\bar\pi}(\rho)\right]_+$,

$$
\left[b - V_{g,1}^{\bar\pi}(x_1)\right]_+ \;\leq\; \frac{\delta}{C^\star - Y^\star} \;\leq\; \frac{2\delta}{C^\star}.
$$

□

F Other Supporting Lemmas

First, we state a useful concentration inequality for the standard self-normalized processes.

Lemma 11 (Concentration of Self-normalized Processes)**.** *Let $\{\mathcal{F}_t\}_{t=0}^\infty$ be a filtration and $\{\eta_t\}_{t=1}^\infty$ be a $\mathbb{R}$-valued stochastic process such that η_t is $\mathcal{F}_t$-measurable for any $t \geq 0$. Assume that for any $t \geq 0$, conditioning on $\mathcal{F}_t$, η_t is a zero-mean and σ-subGaussian random variable with the variance proxy $\sigma^2 > 0$, i.e., $\mathbb{E}\left[e^{\lambda\eta_t}\,|\,\mathcal{F}_t\right] \leq e^{\lambda^2\sigma^2/2}$ for any $\lambda \in \mathbb{R}$. Let $\{X_t\}_{t=1}^\infty$ be an $\mathbb{R}^d$-valued stochastic process such that X_t is $\mathcal{F}_t$-measurable for any $t \geq 0$. Let $Y \in \mathbb{R}^{d\times d}$ be a deterministic and positive-definite matrix. For any $t \geq 0$, we define*

$$
\bar Y_t \;:=\; Y \,+\, \sum_{\tau=1}^{t} X_\tau X_\tau^\top \;\;\text{and}\;\; S_t \;=\; \sum_{\tau=1}^{t} \eta_\tau X_\tau.
$$

Then, for any fixed $\delta \in (0,1)$, it holds with probability at least $1-\delta$ that

$$
\|S_t\|^2_{(\bar Y_t)^{-1}} \;\leq\; 2\sigma^2 \log\left(\frac{\det\left(\bar Y_t\right)^{1/2}\det\left(Y\right)^{-1/2}}{\delta}\right)
$$

for any $t \geq 0$.

Proof. See the proof of Theorem 1 in [1]. □

The above concentration inequality can be customized to our setting in the following form without using covering number arguments as in [37].

Lemma 12. *Let $\lambda = 1$ in Algorithm 2. Fix $\delta \in (0,1)$. Then, for any $(k,h) \in [K]\times[H]$ it holds for $\diamond = r$ or g that*

$$
\left\|\sum_{\tau=1}^{k-1} \phi_{\diamond,h}^\tau(x_h^\tau,a_h^\tau)^\top\left(V_{\diamond,h+1}^k(x_{h+1}^\tau) - (\mathbb{P}_h V_{\diamond,h+1}^k)(x_h^\tau,a_h^\tau)\right)\right\|_{(\Lambda_{\diamond,h}^k)^{-1}} \;\leq\; C\sqrt{dH^2\log\left(\frac{dT}{\delta}\right)}
$$

with probability at least $1-\delta/2$ where $C > 0$ is an absolute constant.

Proof. See the proof of Lemma D.1 in [20]. □

Lemma 13 (Elliptical Potential Lemma)**.** *Let $\{\phi_t\}_{t=1}^\infty$ be a sequence of functions in $\mathbb{R}^d$ and $\Lambda_0 \in \mathbb{R}^{d\times d}$ be a positive definite matrix. Let $\Lambda_t = \Lambda_0 + \sum_{i=1}^{t-1}\phi_i\phi_i^\top$. Assume $\|\phi_t\|_2 \leq 1$ and $\lambda_{\min}(\Lambda_0) \geq 1$. Then for any $t \geq 1$ it holds that*

$$
\log\left(\frac{\det\left(\Lambda_{t+1}\right)}{\det\left(\Lambda_1\right)}\right) \;\leq\; \sum_{i=1}^{t} \phi_i^\top\Lambda_i^{-1}\phi_i \;\leq\; 2\log\left(\frac{\det\left(\Lambda_{t+1}\right)}{\det\left(\Lambda_1\right)}\right).
$$

Proof. See the proof of Lemma D.2 in [37] or [20]. □

Lemma 14 (Pushback Property of KL-divergence)**.** *Let $f:\Delta\to\mathbb{R}$ be a concave function where Δ is a probability simplex in $\mathbb{R}^d$. Let Δ^o be the interior of Δ. Let $x^\star = \operatorname{argmax}_{x\,\in\,\Delta} f(x) - \alpha^{-1}D(x,y)$ for a fixed $y\in\Delta^o$ and $\alpha>0$. Then, for any $z\in\Delta$,*

$$f(x^\star) \,-\, \frac{1}{\alpha}D(x^\star,y) \;\geq\; f(z) \,-\, \frac{1}{\alpha}D(z,y) \,+\, \frac{1}{\alpha}D(z,x^\star).$$

Proof. See the proof of Lemma 14 in [68]. □

Lemma 15 (Bounded KL-divergence Difference)**.** *Let π_1,π_2 be two probability distributions in $\Delta(\mathcal{A})$. Let $\widetilde{\pi}_2 = (1-\theta)\pi_2 + \mathbf{1}\theta/|\mathcal{A}|$ where $\theta\in(0,1]$. Then,*

$$D\left(\pi_1\,|\,\widetilde{\pi}_2\right) \,-\, D\left(\pi_1\,|\,\pi_2\right) \;\leq\; \theta\log|\mathcal{A}|.$$

Moreover, we have a uniform bound, $D\left(\pi_1\,|\,\widetilde{\pi}_2\right) \leq \log(|\mathcal{A}|/\theta)$.

Proof. See the proof of Lemma 31 in [68]. □