# A Policy Gradient Primal-Dual Algorithm for Constrained MDPs with Uniform PAC Guarantees

**Toshinori Kitamura** [*1], **Tadashi Kozuno**[2,6], **Masahiro Kato**[3,1], **Yuki Ichihara**[4],

**Soichiro Nishimori**[1], **Akiyoshi Sannai**[7,1], **Sho Sonoda**[5,1], **Wataru Kumagai**[5], **Yutaka Matsuo**[1]

## Abstract

We study a primal-dual (PD) reinforcement learning (RL) algorithm for online constrained Markov decision processes (CMDPs). Despite its widespread practical use, the existing theoretical literature on PD-RL algorithms for this problem only provides sublinear regret guarantees and fails to ensure convergence to optimal policies. In this paper, we introduce a novel policy gradient PD algorithm with uniform probably approximate correctness (Uniform-PAC) guarantees, simultaneously ensuring convergence to optimal policies, sublinear regret, and polynomial sample complexity for any target accuracy. Notably, this represents the first Uniform-PAC algorithm for the online CMDP problem. In addition to the theoretical guarantees, we empirically demonstrate in a simple CMDP that our algorithm converges to optimal policies, while baseline algorithms exhibit oscillatory performance and constraint violation.

## 1 Introduction

This paper studies a primal-dual (PD) reinforcement learning (RL) algorithm for the online constrained Markov decision processes (CMDP) problem (Efroni et al., 2020), where the agent explores the environment with the aim of identifying an optimal policy that maximizes the return while satisfying certain constraints. The CMDP framework is particularly promising for designing policies in safety-critical decision-making applications, such as autonomous driving with collision avoidance (He et al., 2023b; Gu et al., 2023) and controlling thermal power plants with temperature satisfaction (Zhan et al., 2022). Please refer to Gu et al. (2022) for more examples.

Two primary approaches for the online CMDP problem are the linear programming (LP) approach and the PD approach. While the LP approach is common in theoretical literature (Efroni et al., 2020; Liu et al., 2021a; Bura et al., 2022; HasanzadeZonuzy et al., 2021; Zheng & Ratliff, 2020), the PD approach is more popular in practice due to its adaptability to high-dimensional problem settings. The PD approach typically involves iterative policy gradient ascent over the Lagrange function, making it amenable to recent deep policy gradient RL algorithms (Achiam et al., 2017; Tessler et al., 2018; Wang et al., 2022; Le et al., 2019; Russel et al., 2020).

Despite its practical importance, theoretical results of PD RL algorithms are currently scarce. Existing results on PD RL for online CMDPs are limited to sublinear regret guarantees (Efroni et al., 2020; Liu et al., 2021a; Wei et al., 2021; Ding et al., 2021; Ghosh et al., 2023). However, sublinear regret guarantees only bound the integral of the magnitude of mistakes during the training, and they cannot ensure the performance of the last-iterate policy up to arbitrary accuracy (Dann et al., 2017). An alternative performance measure is $(\varepsilon, \delta)$-PAC, which ensures that the last-iterate policy's performance is sufficiently close to an optimal policy. However, $(\varepsilon, \delta)$-PAC has only been established

*Correspondence to: <toshinori-k@weblab.t.u-tokyo.ac.jp>.
[1] The University of Tokyo, Japan, [2] OMRON SINIC X, Japan, [3] Mizuho–DL Financial Technology, Japan,
[4] Nara Institute of Science and Technology, Japan, [5] RIKEN Center for Advanced Intelligence Project, Japan.
[6] Osaka University, Japan. [7] Kyoto University, Japan.

Table 1: Regret bound and $(\varepsilon, \delta)$-PAC bound comparison of online CMDP algorithms. The "LP" and "PD" rows correspond to linear programming and primal-dual algorithms, respectively. The "Optimality" and "VIO" columns correspond to the bounds for optimality gap and constraint violation, respectively. In the "VIO" column, "same" means that the bound for constraint violation is the same as that for the optimality gap, and "const" means that the bound does not depend on $K$. In the "Regret" column, the subscript "$_+$" means that the bound is concerning to strong regret measures rather than weak measures (see Appendix B). If the "LIC?" column is "✓", the algorithm is guaranteed the last-iterate convergence (LIC) to optimal policies. This table is presented under single constraint settings (i.e., $N = 1$) for a fair comparison. The algorithms are: OptCMDP, OptPrimalDual-CMDP (Efroni et al., 2020), OptPress-LP, OptPress-PrimalDual (Liu et al., 2021a), DOPE (Bura et al., 2022), OPDOP (Ding et al., 2021), Triple-Q (Wei et al., 2021), Online-CRL (HasanzadeZonuzy et al., 2021), and Regularized Primal-Dual Algorithm (Müller et al., 2024).

| | Algorithm | Regret | | $(\varepsilon, \delta)$-PAC | | LIC? |
|---|---|---|---|---|---|---|
| | | Optimality | VIO | Optimality | VIO | |
| LP | OptCMDP | $\widetilde{\mathcal{O}}\left(XA^{\frac{1}{2}}H^2K^{\frac{1}{2}}\right)_+$ | same$_+$ | - | - | ✗ |
| | OptPress-LP | $\widetilde{\mathcal{O}}\left(b_{\text{gap}}^{-1}X^{\frac{3}{2}}A^{\frac{1}{2}}H^3K^{\frac{1}{2}}\right)$ | $0_+$ | - | - | ✗ |
| | DOPE | $\widetilde{\mathcal{O}}\left(b_{\text{gap}}^{-1}XA^{\frac{1}{2}}H^3K^{\frac{1}{2}}\right)$ | $0_+$ | - | - | ✗ |
| | Online-CRL | - | - | $\widetilde{\mathcal{O}}\left(X^2AH^2\varepsilon^{-2}\right)$ | same | ✗ |
| PD | OPDOP | $\widetilde{\mathcal{O}}\left(XA^{\frac{1}{2}}H^{\frac{5}{2}}K^{\frac{1}{2}}\right)$ | same | - | - | ✗ |
| | OptPD-CMDP | $\widetilde{\mathcal{O}}\left(b_{\text{gap}}^{-1}X^{\frac{3}{2}}A^{\frac{1}{2}}H^2K^{\frac{1}{2}}\right)$ | same | - | - | ✗ |
| | OptPress-PD | $\widetilde{\mathcal{O}}\left(b_{\text{gap}}^{-1}X^{\frac{3}{2}}A^{\frac{1}{2}}H^3K^{\frac{1}{2}}\right)$ | const | - | - | ✗ |
| | Triple-Q | $\widetilde{\mathcal{O}}\left(b_{\text{gap}}^{-1}X^{\frac{1}{2}}A^{\frac{1}{2}}H^4K^{\frac{4}{5}}\right)$ | 0 | - | - | ✗ |
| | Regularized PD | $\widetilde{\mathcal{O}}\left(b_{\text{gap}}^{-2}X^{\frac{1}{2}}A^{\frac{1}{4}}H^{\frac{9}{2}}K^{0.93}\right)_+$ | same$_+$ | - | - | ✗ |
| | **UOpt-RPGPD** | $\widetilde{\mathcal{O}}\left(b_{\text{gap}}^{-1}X^{\frac{1}{2}}A^{\frac{2}{7}}H^4K^{\frac{6}{7}}\right)_+$ | same$_+$ | $\widetilde{\mathcal{O}}\left(b_{\text{gap}}^{-7}X^{\frac{7}{2}}A^2H^{25}\varepsilon^{-7}\right)$ | same | ✓ |

for LP algorithms in the online CMDP problem (HasanzadeZonuzy et al., 2021). Furthermore, it is known that both sublinear regret and $(\varepsilon, \delta)$-PAC are insufficient to ensure convergence to optimal policies (Dann et al., 2017). CMDP algorithms lacking convergence guarantees may yield policies exhibiting oscillatory performance and constraint violation, those are undesirable in practical applications due to their potential impact on system stability and safety (Moskovitz et al., 2023).

Numerous studies have tackled the convergence problem of PD algorithms for CMDPs, but most of them are limited to cases without exploration. Even in the absence of exploration, these studies possess unfavorable limitations for application, such as optimization over occupancy measures [1] rather than policies (Moskovitz et al., 2023), providing convergence guarantees only through an average of past returns or a mixture of past policies (Li et al., 2021; Chen et al., 2021b; Ding et al., 2020; Liu et al., 2021b), and converging to a biased solution with fixed $\varepsilon > 0$ (Ying et al., 2022; Ding et al., 2023). More related works can be found in Appendix A. In light of these limitations, a natural question then arises:

*Is it possible to design a policy gradient PD algorithm for online CMDPs that ensures the triplet of sublinear regret, $(\varepsilon, \delta)$-PAC, and convergence to optimal policies?*

We provide an affirmative response by proposing a novel policy gradient PD algorithm for online CMDPs with a uniform probably approximate correctness (Uniform-PAC) guarantee (Dann et al., 2017), called **U**niform-PAC **Opt**imistic **R**egularized **P**olicy **G**radient **P**rimal-**D**ual (`UOpt-RPGPD`). Uniform-PAC, a stronger performance metric than sublinear regret and $(\varepsilon, \delta)$-PAC, ensures not only convergence to optimal policies but also sublinear regret and polynomial sample complexity for any target accuracy. Notably, `UOpt-RPGPD` is the first-ever online CMDP algorithm with Uniform-PAC guarantees for both LP and PD approaches. Furthermore, `UOpt-RPGPD` ensures strong regret guarantees whereas most of the existing PD algorithms (Efroni et al., 2020; Ding et al., 2021; Liu et al., 2021a; Wei et al., 2021) provide weaker guarantees as relying on the *error cancellation*

[1] An occupancy measure of a policy denotes the set of distributions over the state-action space generated by executing the policy in the environment. See Equation (12) for the definition.

technique (see Remark B.3 in Appendix B for details). The very recent Müller et al. (2024) provides a PD algorithm with a strong regret guarantee but they cannot be Uniform-PAC due to their bonus design (see Appendix C). Table 1 compares the theoretical guarantees of online CMDP algorithms.

Finally, we empirically demonstrate the effectiveness of the three techniques through an ablation study on a simple CMDP. Our results illustrate that `UOpt-RPGPD` converges to optimal policies, while other baseline algorithms fail to converge or even exhibit oscillatory behaviors.

## 2 Preliminary

We use the shorthand $\mathbb{R}_+ \coloneqq [0, \infty)$. The set of probability distributions over a set $\mathbf{S}$ is denoted by $\Delta(\mathbf{S})$. For a positive integer $N \in \mathbb{N}$, we define $[N] \coloneqq \{1, \dots, N\}$. All scalar operators and inequalities should be understood point-wise when used for vectors and functions. For example, for functions $f, g, z : \mathbf{X} \to \mathbb{R}$, we express "$f(x) \geq g(x)$ for all $x$" as $f \geq g$ and "$z(x) = f(x) + g(x)$ for all $x$" as $z = f + g$. For $p_1, p_2 \in \Delta(\mathbf{A})$ with $p_1, p_2 > \mathbf{0}$, we define the KL divergence between $p_1$ and $p_2$ as $\mathrm{KL}[p_1, p_2] \coloneqq \sum_{a \in \mathbf{A}} p_1(a) \ln \frac{p_1(a)}{p_2(a)}$. For $x, a, b \in \mathbb{R}$, we define a clipping function such that $\mathrm{clip}(x, a, b) = \min\{\max\{x, a\}, b\}$. For two positive sequences $\{a_n\}$ and $\{b_n\}$ with $n = 1, 2, \dots$, we write $a_n = O(b_n)$ if there exists an absolute constant $C > 0$ such that $a_n \leq C b_n$ holds for all $n \geq 1$. We use $\widetilde{O}(\cdot)$ to further hide the polylogarithmic factors.

**Constrained Markov Decision Processes.** Let $N \in \{0, 1, \dots\}$ be the number of constraints. A finite-horizon and episodic CMDP is defined as a tuple $(\mathbf{X}, \mathbf{A}, H, P, \mathbf{r}, b, x_1)$, where $\mathbf{X}$ denotes the finite state space with size $X$, $\mathbf{A}$ denotes the finite action space with size $A$, $H \in \mathbb{N}$ denotes the horizon of an episode, $b \in [0, H]^N$ denotes the constrained threshold vector, where $b^n$ is the threshold scalar for the $n$-th constraint, $x_1$ denotes the fixed initial state, and $\mathbf{r} \coloneqq \{r^n\}_{n \in \{0\} \cup [N]}$ denotes the set of reward functions, where $r^n_\cdot(\cdot, \cdot) : [H] \times \mathbf{X} \times \mathbf{A} \to [0, 1]$ denotes the $n$ th reward function, and $r^n_h(x, a)$ is the $n$-th reward when taking an action $a$ at a state $x$ in step $h$. The reward function $r^0$ is for the objective to optimize and the reward functions $\{r^1, \dots, r^N\}$ are for constraints. $P_\cdot(\cdot \mid \cdot, \cdot) : [H] \times \mathbf{X} \times \mathbf{A} \to \Delta(\mathbf{X})$ denotes the transition probability kernel, where $P_h(y \mid x, a)$ is the state transition probability to a new state $y$ from a state $x$ when taking an action $a$ in step $h$. With an abuse of notation, for a function $V : \mathbf{X} \to \mathbb{R}$, let $P_h$ be an operator such that $(P_h V)(x, a) = \sum_{y \in \mathbf{X}} V(y) P_h(y \mid x, a)$.

**Policy and Regularized Value Functions.** A policy is defined as $\pi_\cdot(\cdot \mid \cdot) : [H] \times \mathbf{X} \to \Delta(\mathbf{A})$, where $\pi_h(a \mid x)$ denotes the probability of taking an action $a$ at state $x$ in step $h$. For a policy $\pi$ with $\pi > \mathbf{0}$, we denote $\ln \pi_\cdot(\cdot|\cdot) : [H] \times \mathbf{X} \times \mathbf{A} \to \mathbb{R}$ as the function such that $\ln \pi_h(x|a) = \ln(\pi_h(a \mid x))$. The set of all the policies is denoted as $\Pi$. With an abuse of notation, for any policy $\pi$ and $Q : \mathbf{X} \times \mathbf{A} \to \mathbb{R}$, let $\pi_h$ be an operator such that $(\pi_h Q)(x) = \sum_{a \in \mathbf{A}} \pi_h(a \mid x) Q(x, a)$.

For a policy $\pi$, transition kernel $P$, reward function $r_\cdot(\cdot, \cdot) : [H] \times \mathbf{X} \times \mathbf{A} \to \mathbb{R}$, and an entropy coefficient $\tau \geq 0$, let $V^\pi_\cdot[P, r, \tau](\cdot) : [H] \times \mathbf{X} \to \mathbb{R}$ be the regularized value function such that

$$V^\pi_h[P, r, \tau](x) = \mathbb{E}\left[\sum_{i=h}^{H} r_h(x_i, a_i) - \tau \ln \pi_h(a_i \mid x_i) \,\middle|\, x_h = x, \pi, P\right],$$

where the expectation is taken over all possible trajectories, in which $a_h \sim \pi_h(\cdot \,|\, x_h)$ and $x_{h+1} \sim P_h(\cdot \,|\, x_h, a_h)$. We set $V^\pi_{H+1}[P, r, \tau](x) = 0$ for all $x \in \mathbf{X}$. We define the regularized action-value function $Q^\pi_\cdot[P, r, \tau](\cdot, \cdot) : [H] \times \mathbf{X} \times \mathbf{A} \to \mathbb{R}$ such that

$$Q^\pi_h[P, r, \tau](x, a) = r_h(x, a) + \left(P_h V^\pi_{h+1}[P, r, \tau]\right)(x, a).$$

We set $Q^\pi_{H+1}[P, r, \tau](x, a) = 0$ for all $(x, a) \in \mathbf{X} \times \mathbf{A}$. When $\tau = 0$, we omit $\tau$ from notations. For example, we write $Q^\pi_h[P, r] \coloneqq Q^\pi_h[P, r, 0]$.

### 2.1 Learning Problem Setup

We consider an algorithm operating an agent that repeatedly interacts with a CMDP $(\mathbf{X}, \mathbf{A}, H, P, \mathbf{r}, b, x_1)$ by playing a sequence of policies $\pi^1, \pi^2, \dots$, where $\pi^k \in \Pi$ denotes the policy that the agent follows in the $k$-th episode. Each episode $k$ starts from the fixed initial state $x_1$. At the beginning of each time-step $h \in [H]$ in an episode $k$, the agent observes a state $x_h^k$ and chooses an action $a_h^k$, which is sampled from $\pi_h^k(\cdot \mid x_h^k)$. The next state is sampled as $x_{h+1}^k \sim P_h(\cdot \mid x_h^k, a_h^k)$. The learning agent lacks prior knowledge of the transition kernel $P$. We assume that the set of reward functions $\mathbf{r}$ is known for simplicity, but extending our algorithm to unknown stochastic rewards poses no real difficulty (Azar et al., 2017; Ayoub et al., 2020).

Let $\Pi_{\text{safe}} \coloneqq \left\{\pi \,\middle|\, \min_{n\in[N]}(V_1^\pi[P, r^n](x_1) - b^n) \geq 0\right\}$ be a set of policies that do not violate the constraints. An optimal policy $\pi^\star$, which maximizes the non-regularized return while satisfying all the constraints, is defined as

$$\pi^\star \in \arg\max_{\pi\in\Pi_{\text{safe}}} V_1^\pi[P, r^0](x_1) \,.$$

Finally, we assume the following slater condition. This assumption is mild as it holds when there exists some strictly feasible policy (Efroni et al., 2020; Liu et al., 2021a; Paternain et al., 2019).

**Assumption 2.1** (Slater Point)**.** There exists an unknown policy $\pi_{\text{safe}} \in \Pi_{\text{safe}}$ such that $V_1^{\pi_{\text{safe}}}[P, r^n](x_1) = b_{\text{safe}}^n$, where $b_{\text{safe}}^n > b^n$ for all $n \in [N]$. Let $b_{\text{gap}} \coloneqq \min_{n\in[N]}(b_{\text{safe}}^n - b^n)$.

### 2.2 Performance Measure

Let $\Delta_{\text{opt}}^k \coloneqq V_1^{\pi^\star}[P, r^0](x_1) - V_1^{\pi^k}[P, r^0](x_1)$ and $\Delta_{\text{vio}}^k \coloneqq \max_{n\in[N]} b^n - V_1^{\pi^k}[P, r^n](x_1)$ be the temporal optimality gap and constraint violation, respectively. Let $\Delta_{\text{opt}+}^k \coloneqq \max\{\Delta_{\text{opt}}^k, 0\}$ and $\Delta_{\text{vio}+}^k \coloneqq \max\{\Delta_{\text{vio}}^k, 0\}$ be their positively clipped values. For any $K \in \mathbb{N}$ and $\varepsilon > 0$, the following notations are useful to introduce the performance measures:

$$\mathfrak{M}_{\text{opt}}^\varepsilon \coloneqq \sum_{k=1}^{\infty} \mathbb{1}\{\Delta_{\text{opt}}^k > \varepsilon\}, \quad \text{and} \quad \mathfrak{M}_{\text{vio}}^\varepsilon \coloneqq \sum_{k=1}^{\infty} \mathbb{1}\{\Delta_{\text{vio}}^k > \varepsilon\} \,. \tag{1}$$

$\mathfrak{M}_{\text{opt}}^\varepsilon$ and $\mathfrak{M}_{\text{vio}}^\varepsilon$ measure the count of mistakes that exceed $\varepsilon > 0$ related to optimality gap and constraint violation, respectively.

The performance of an online CMDP algorithm is typically measured by either the high-probability regret (Efroni et al., 2020; Liu et al., 2021a; Bura et al., 2022) or $(\varepsilon, \delta)$-PAC (HasanzadeZonuzy et al., 2021; Zeng et al., 2022; Vaswani et al., 2022; Bennett et al., 2023). Their formal definitions are deferred to Appendix B. Since neither sublinear regret nor $(\varepsilon, \delta)$-PAC guarantees convergence to optimal policies (Dann et al., 2017), we consider the following Uniform-PAC measure to evaluate the algorithm's performance.

**Definition 2.2** (Uniform-PAC)**.** Given $\varepsilon > 0$ and $\delta \in (0, 1]$, let $F_{\text{UPAC}}(\cdots)$ be shorthand for $F_{\text{UPAC}}(X, A, H, N, 1/b_{\text{gap}}, 1/\varepsilon, \ln(1/\delta))$, where $F_{\text{UPAC}}$ is a real-valued function polynomial in all its arguments. An algorithm achieves Uniform-PAC for $\delta > 0$ if there exists $F_{\text{UPAC}}(\cdots)$ such that

$$\mathbb{P}\big(\exists \varepsilon > 0 \text{ such that } \mathfrak{M}_{\text{opt}}^\varepsilon > F_{\text{UPAC}}(\cdots) \,\vee\, \mathfrak{M}_{\text{vio}}^\varepsilon > F_{\text{UPAC}}(\cdots)\big) \leq \delta \,.$$

**Theorem 2.3.** *Suppose an algorithm is Uniform-PAC for $\delta$ with $F_{\text{UPAC}}(\cdots) = \widetilde{\mathcal{O}}(C\varepsilon^{-\alpha})$, where $C, \alpha > 0$ are constants independent of $\varepsilon$. Then, the algorithm*

1. *converges with high probability: $\mathbb{P}\big(\lim_{k\to\infty} \Delta_{\text{opt}}^k = 0 \wedge \lim_{k\to\infty} \Delta_{\text{vio}}^k = 0\big) \geq 1 - \delta$.*
2. *is $(\varepsilon, \delta)$-PAC with sample complexity $F_{\text{UPAC}}(\cdots)$ for all $\varepsilon > 0$.*
3. *achieves $\widetilde{\mathcal{O}}\Big(C^{\frac{1}{\alpha}} K^{1-\frac{1}{\alpha}}\Big)$ regret with probability at least $1 - \delta$, where $K \in \mathbb{N}$.*

The first and the second claims are direct applications of **Theorem 3** from Dann et al. (2017). The third claim follows from Lemma F.2 in Appendix F.

**Algorithm 1** UOpt-RPGPD

**Input:** $\delta \in (0,1]$, $\alpha_\eta \in (0,1]$, $\alpha_\tau \in (0,1]$
Set $\lambda^1 \coloneqq \mathbf{0}$. Set $\pi_h^1(a \mid x) \coloneqq \frac{1}{A} \quad \forall (x,a,h) \in [H] \times \mathbf{X} \times \mathbf{A}$
Set $\eta_k \coloneqq (k+3)^{-\alpha_\eta}$ and $\tau_k \coloneqq (k+3)^{-\alpha_\tau}$
**for** $k = 1, 2, 3, \ldots$ **do**
  Compute bonus $\beta^{k,\delta}$ by Equation (3)
  $\widetilde{Q}^{k,0}, \widetilde{V}^{k,0} \coloneqq \texttt{RegularizedPolicyEvaluation}(r^0, \beta^{k,\delta}, \overline{P}^k, \pi^k, \tau_k)$
  $\widetilde{Q}^{k,n}, \widetilde{V}^{k,n} \coloneqq \texttt{RegularizedPolicyEvaluation}(r^n, \beta^{k,\delta}, \overline{P}^k, \pi^k, 0)$ for all $n \in [N]$
  $\widetilde{Q}^k \coloneqq \widetilde{Q}^{k,0} + \sum_{n=1}^N \lambda^{k,n} \widetilde{Q}^{k,n}$
  Compute $\pi^{k+1}$ by Equation (4) and compute $\lambda^{k+1}$ by Equation (5)
  Rollout $\pi^{k+1}$ and then update $n^k$ and $\overline{P}^k$
**end for**

## 3 The UOpt-RPGPD Algorithm

This section provides our **U**niform-PAC **Opt**imistic **R**egularized **P**olicy **G**radient **P**rimal-**D**ual (`UOpt-RPGPD`) algorithm. Our `UOpt-RPGPD` relies on the combination of three key techniques: (i) the Lagrange function regularized by policy entropy and Lagrange multipliers, (ii) Uniform-PAC exploration bonus, and (iii) careful adjustment of the regularization coefficient and learning rate. We present the pseudo-code of our algorithm in Algorithm 1. It is important to remark that existing online CMDP algorithms are designed only for a fixed iteration length $K \in [N]$ (Efroni et al., 2020; Liu et al., 2021a; Bura et al., 2022; HasanzadeZonuzy et al., 2021; Zheng & Ratliff, 2020; Wei et al., 2021; 2022; Ding et al., 2021; Amani et al., 2021; Ghosh et al., 2023). In contrast, our algorithm works with an infinite episode length.

**Regularized Lagrange function.** `UOpt-RPGPD` is designed to solve the following regularized Lagrange function in Equation (2) while exploring the environment. For a policy $\pi \in \Pi$, Lagrange multipliers $\lambda \in \mathbb{R}_+^N$, and an entropy coefficient $\tau \geq 0$, we define the regularized Lagrange function as

$$L_\tau(\pi, \lambda) \coloneqq V_1^\pi\big[P, r^0, \tau\big](x_1) + \sum_{n=1}^N \lambda^n (V_1^\pi[P, r^n](x_1) - b^n) + \frac{\tau}{2}\|\lambda\|_2^2 \,. \tag{2}$$

Let $\pi_\tau^\star \coloneqq \arg\max_{\pi \in \Pi} \min_{\lambda \in \mathbb{R}_+^N} L_\tau(\pi, \lambda)$ and $\lambda_\tau^\star \coloneqq \arg\min_{\lambda \in \mathbb{R}_+^N} \max_{\pi \in \Pi} L_\tau(\pi, \lambda)$. Note that $(\pi_\tau^\star, \lambda_\tau^\star)$ is the unique saddle point of $L_\tau$ as the following lemma shows (the proof is provided in Appendix G).

**Lemma 3.1.** *For any $\tau \in (0, \infty)$, there exists a unique saddle point $(\pi_\tau^\star, \lambda_\tau^\star) \in \Pi \times \mathbb{R}_+^N$ such that*

$$L_\tau(\pi_\tau^\star, \lambda) \geq L_\tau(\pi_\tau^\star, \lambda_\tau^\star) \geq L_\tau(\pi, \lambda_\tau^\star) \;\; \forall (\pi, \lambda) \in \Pi \times \mathbb{R}_+^N \,.$$

The regularized Lagrange technique is derived from the work by Ding et al. (2023), wherein the consideration of exploration is absent. The introduced regularization affords us to upper bound the value of $\sum_{h=1}^H \sum_{x \in \mathbf{X}} w_h^{\pi_\tau^\star}[P](x) \mathrm{KL}[\pi_{\tau_k,h}^\star(\cdot \mid x), \pi_h^k(\cdot \mid x)]$ by a decreasing function on $k$. Intuitively, it implies that the pair $(\pi^k, \lambda^k)$ converges to $(\pi_{\tau_k}^\star, \lambda_{\tau_k}^\star)$, leading to the decreasing optimality gap and constraint violation of `UOpt-RPGPD`. The detailed upper bound is provided in Appendix H.4.

**Uniform-PAC Exploration Bonus.** The second technique in our algorithm is the use of the Uniform-PAC bonus function. Let $\mathrm{llnp}(x) \coloneqq \ln(\ln(\max\{x, e\}))$. Let $n_h^k(x,a) \coloneqq \sum_{i=1}^k \mathbb{1}\big(x_h^i = x, a_h^i = a\big)$ be the number of times a pair $(x,a)$ was observed at step $h$ before episode $k+1$. We define the empirical estimation of the transition model as

$$\overline{P}_h^k(y \mid x, a) \coloneqq \frac{\sum_{i=1}^k \mathbb{1}\left(x_h^i = x, a_h^i = a, x_{h+1}^i = y\right)}{n_h^k(x,a) \vee 1} \,.$$

**Algorithm 2** RegularizedPolicyEvaluation

**Input:** $r$, $\beta$, $\overline{P}$, $\pi \in \Pi$, $\tau \in \mathbb{R}_+$
Set $\widetilde{V}_{H+1} \coloneqq \mathbf{0}$
**for** $h = H, \cdots 1$ **do**
  $\widetilde{Q}_h \coloneqq \min\Big\{r_h + (1+\tau\ln A)H\beta_h + \overline{P}_h\widetilde{V}_{h+1},\ (1+\tau\ln A)(H-h+1)\mathbf{1}\Big\}$
  $\widetilde{V}_h \coloneqq \pi_h\Big(\widetilde{Q}_h - \tau\ln\pi_h\Big)$
**end for**
**Return:** $\widetilde{Q}$, $\widetilde{V}$

Given a failure probability $\delta$, we define the bonus function $\beta^{k,\delta}(\cdot,\cdot) : [H]\times\mathbf{X}\times\mathbf{A}\to\mathbb{R}$ such that

$$\beta_h^{k,\delta}(x,a) = \sum_{y\in\mathbf{X}}\sqrt{4\overline{P}_h^k(y\mid x,a)}\phi + 5\phi^2 \quad \text{where} \quad \phi \coloneqq \sqrt{\frac{2\,\mathrm{llnp}\big(2n_h^k(x,a)\big) + 2\ln(48X^2AH\delta^{-1})}{n_h^k(x,a)\vee 1}}\,. \tag{3}$$

Using the bonus $\beta^{k,\delta}$, UOpt-RPGPD computes the regularized optimistic value functions $\widetilde{Q}^{k,0}, \widetilde{V}^{k,0}$, and the non-regularized optimistic value functions $\Big\{\widetilde{Q}^{k,n}, \widetilde{V}^{k,n}\Big\}_{n\in[N]}$ by a regularized policy evaluation (Algorithm 2).

Our bonus design is inspired by the work of Dann et al. (2017). While naive bonus functions (e.g., Efroni et al. (2020)) scale to $\sqrt{\ln(K)/n_h^k(x,a)}$ with a fixed iteration length $K$, our bonus scales to $\sqrt{\ln\ln(n_h^k(x,a))/n_h^k(x,a)}$. This allows the bonus to diminish sufficiently quickly even when $K\to\infty$, in contrast to existing bonuses that can increase when $K$ becomes large.

**Adjust Regularization Coefficients and Learning Rate.** The combination of the above two techniques may not be sufficient for Uniform-PAC because it can introduce a biased solution due to the regularization in Equation (2). To overcome this issue, we decrease the regularization coefficient and the policy learning rate as $\eta_k \coloneqq k^{-\alpha_\eta}$ and $\tau_k \coloneqq k^{-\alpha_\tau}$ with $0 < \alpha_\tau < 0.5 < \alpha_\eta < 1$. We remark that employing naive learning rates such as $\eta_k = \tau_k \propto 1/k$ or $\eta_k = \tau_k \propto 1/\sqrt{k}$, as seen in prior works like Efroni et al. (2020), fails to guarantee Uniform-PAC. To attain Uniform-PAC, we applied careful sequential analysis techniques akin to those utilized in bandit studies (e.g., Cai et al. (2023)) to our primal-dual CMDP algorithm.

Coupled with this adjustment technique, UOpt-RPGPD updates the policy and the Lagrange multipliers based on the regularized Lagrange objective (Equation (2)) with the Uniform-PAC bonus (Equation (3)). Specifically, it updates the policy through an entropy-regularized natural policy gradient ascent (Cen et al., 2022) as

$$\pi_h^{k+1}(\cdot\mid x) \propto \big(\pi_h^k(\cdot\mid x)\big)^{(1-\eta_k\tau_k)}\exp\Big(\eta_k\widetilde{Q}_h^k(x,\cdot)\Big) \tag{4}$$

for all $(x,h)\in\mathbf{X}\times[H]$, where $\widetilde{Q}^k$ is defined as $\widetilde{Q}^k \coloneqq \widetilde{Q}^{k,0} + \sum_{n=1}^N \lambda^{k,n}\widetilde{Q}^{k,n}$.

UOpt-RPGPD then updates the Lagrange multipliers through a projected regularized gradient descent, given by

$$\lambda^{k+1,n} \coloneqq \mathrm{clip}\left[\Lambda,\ 0,\ \frac{H(1+\tau_k\ln A)}{b_{\mathrm{gap}}}\right] \quad \forall n\in[N] \tag{5}$$

where $\Lambda \coloneqq \lambda^{k,n} + \eta_k\Big(b^n - \widetilde{V}_1^{k,n}(x_1) - \tau_k\lambda^{k,n}\Big)$ and $\lambda^{k,n}$ denotes the $n$-th value of $\lambda^k$.

## 4 Uniform-PAC Analysis

Our UOpt-RPGPD achieves the following Uniform-PAC guarantee.

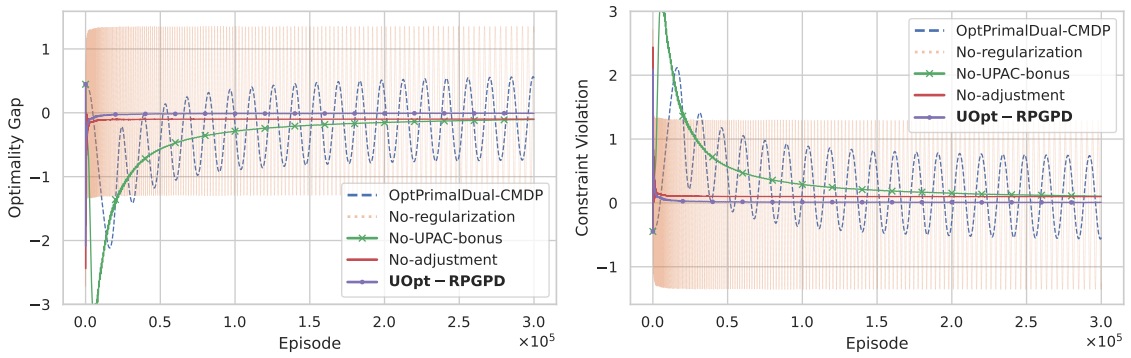


Figure 1: Comparison of the algorithms described in Section 5. **Left**: optimality gap ($\Delta^k_{\text{opt}}$) and **Right**: constraint violation ($\Delta^k_{\text{vio}}$).

**Theorem 4.1.** *Suppose that Assumption 2.1 holds. Set the regularization coefficient $\alpha_\tau$ and the learning rate $\alpha_\eta$ such that $0 < \alpha_\tau < 0.5 < \alpha_\eta < 1$ and $\alpha_\eta + \alpha_\tau < 1$. Then, for $\delta > 0$, UOpt-RPGPD achieves Uniform-PAC for $F_{UPAC}(\cdots)$ such that*

$$F_{UPAC}(\cdots) = \underbrace{\widetilde{\mathcal{O}}\left(\left(b_{\text{gap}}^{-2}(1+N)XA^{\frac{1}{2}}H^7\varepsilon^{-2}\right)^{\frac{1}{\alpha_\eta-0.5}}\right)}_{\text{(i)}} + \underbrace{\widetilde{\mathcal{O}}\left(\left(b_{\text{gap}}^{-1}H\varepsilon^{-1}\right)^{\frac{1}{\alpha_\tau}}\right)}_{\text{(ii)}} + \underbrace{\left(\frac{24}{1-(\alpha_\eta+\alpha_\tau)}\ln\frac{12}{1-(\alpha_\eta+\alpha_\tau)}\right)^{\frac{1}{1-(\alpha_\eta+\alpha_\tau)}}}_{\text{(iii)}}. \tag{6}$$

The proof is provided in Appendix H. Note that this bound depends on the values of $\alpha_\tau$ and $\alpha_\eta$. Below, we establish a concrete bound by setting the values of $\alpha_\tau$ and $\alpha_\eta$, focusing on the order of $\varepsilon$. The first term (i) decreases as $\alpha_\eta$ approaches 1. However, the conditions $0 < \alpha_\tau < 0.5 < \alpha_\eta < 1$ and $\alpha_\eta + \alpha_\tau < 1$ restrict $\alpha_\eta$ from nearing 1 due to the second term (ii) and the third term (iii). This tradeoff makes achieving $\widetilde{\mathcal{O}}(\varepsilon^{-6})$ in Equation (6) unattainable.

Hence, we select values for $\alpha_\eta$ and $\alpha_\tau$ to make the order of $\varepsilon$ in Equation (6) to be $\widetilde{\mathcal{O}}(\varepsilon^{-7})$. The ensuing corollary is presented as follows.

**Corollary 4.2.** *Suppose that Assumption 2.1 holds. With $\alpha_\eta = \frac{11}{14}$ and $\alpha_\tau = \frac{1}{7}$, UOpt-RPGPD achieves Uniform-PAC for $\delta > 0$ for $F_{UPAC}$ such that*

$$F_{UPAC}(\cdots) = \widetilde{\mathcal{O}}\left(b_{\text{gap}}^{-7}(1+N)^{\frac{7}{2}}X^{\frac{7}{2}}A^2H^{25}\varepsilon^{-7}\right).$$

*Therefore, with probability at least $1-\delta$, UOpt-RPGPD converges to optimal policies and has regret*

$$\widetilde{\mathcal{O}}\left(b_{\text{gap}}^{-1}(1+N)^{\frac{1}{2}}X^{\frac{1}{2}}A^{\frac{2}{7}}H^4K^{\frac{6}{7}}\right).$$

In Corollary 4.2, the convergence and regret bound follows immediately from Theorem 2.3. To our knowledge, UOpt-RPGPD is the first RL algorithm for online CMDPs that achieves the triplet of sublinear regret, $(\varepsilon, \delta)$-PAC, and convergence to optimal policies.

## 5 Experiments

This section describes the experimental behavior of our UOpt-RPGPD on a simple CMDP. We randomly instantiate a CMDP with $X = 30$, $A = 3$, $H = 10$, and $N = 1$. The construction of a

CMDP is based on a tabular (C)MDP experiment conducted by Dann et al. (2017) and Moskovitz et al. (2023). A detailed description of the experimental setup can be found in Appendix D.

In this experiment, we compare `UOpt-RPGPD` with `OptPrimalDual-CMDP` from Efroni et al. (2020) as it adheres to the naive primal-dual framework. The detail of the algorithm is provided in Appendix C. Furthermore, to empirically validate the efficacy of the three techniques introduced in `UOpt-RPGPD` as expounded in Section 3, we compare: (i) `UOpt-RPGPD` without regularization technique (i.e., $\tau_k = 0$), (ii) `UOpt-RPGPD` with the naive bonus function by Equation (10), and (iii) `UOpt-RPGPD` with fixed $\eta_k$ and $\tau_k$. We call the three algorithms `No-regularization`, `No-UPAC-bonus`, and `No-adjustment`, respectively.

Figure 1 compares algorithms and presents their optimality gap (Left) and constraint violation (Right). The results are from a single run of the same randomly generated CMDP, yet it is illustrative. We reran the experiment with different random seeds, consistently obtaining qualitatively similar results.

Compared to other algorithms, our `UOpt-RPGPD` quickly converges to optimal policies. Algorithms without regularization, such as `No-regularization` and `OptPrimalDual-CMDP`, fail to converge and even display oscillatory behaviors. `No-UPAC-bonus` results in slow learning and `No-adjustment` converges to a biased solution. These results highlight the importance of the three techniques introduced in `UOpt-RPGPD`.

## 6 Conclusion

We introduced `UOpt-RPGPD`, the first primal-dual RL algorithm for online CMDPs with Uniform-PAC guarantees. `UOpt-RPGPD` ensures simultaneous convergence to optimal policies, sublinear regret, and polynomial sample complexity for any target accuracy. In addition to the theoretical analysis, we empirically demonstrated that our algorithm successfully converges to optimal policies in a simple CMDP, whereas the existing primal-dual algorithm exhibits oscillatory behavior.

**Limitation and Future Work.** Although `UOpt-RPGPD` achieves Uniform-PAC, it may violate constraints during exploration. The development of a zero-violation algorithm is currently a significant topic in the study on online CMDP (Liu et al., 2021a; Bura et al., 2022; Wei et al., 2021). We defer the extension of our results to a zero-violation algorithm as part of our future work.

Another future research involves the extension to function approximation. A theoretical study of the primal-dual approach with function approximation could reveal opportunities for improvement in existing practical primal-dual deep RL algorithms. While there are CMDP algorithms with linear function approximation (Ding et al., 2021; Amani et al., 2021; Ghosh et al., 2023), none establish Uniform-PAC guarantees when using function approximation. The extension to function approximation, such as linear MDPs (Zhou et al., 2021; Hu et al., 2022; He et al., 2023a) or general function approximation (Jiang et al., 2017; Du et al., 2021; Jin et al., 2021), represents another promising direction for future work.

Finally, our Uniform-PAC bound may not be tight. Compared to the $\widetilde{\mathcal{O}}(\sqrt{K})$ regret bound by the LP algorithm Efroni et al. (2020), our Corollary 4.2 provides $\widetilde{\mathcal{O}}(K^{\frac{6}{7}})$ regret bound. It remains unclear whether this is an artifact of our analysis or a genuine limitation of Uniform-PAC primal-dual algorithms. We leave this topic as a future work.

### Broader Impact Statement

This paper presents work whose goal is to advance the field of RL theory. There are many potential societal consequences of our work, none of which we feel must be specifically highlighted here.

### Acknowledgments

SS was supported by JST PRESTO JPMJPR2125. WK was partially supported by the JSPS KAKENHI Grant Numbers 19H04071 and 23H04974. AS was supported by JSPS KAKENHI Grant Number JP20K03743, JP23H04484 and JST PRESTO JPMJPR2123.

## References


Joshua Achiam, David Held, Aviv Tamar, and Pieter Abbeel. Constrained Policy Optimization. In *International conference on machine learning*, 2017.

Eitan Altman. *Constrained Markov Decision Processes*, volume 7. CRC Press, 1999.

Sanae Amani, Christos Thrampoulidis, and Lin Yang. Safe Reinforcement Learning with Linear Function Approximation. In *International Conference on Machine Learning*, 2021.

Alex Ayoub, Zeyu Jia, Csaba Szepesvari, Mengdi Wang, and Lin Yang. Model-Based Reinforcement Learning with Value-Targeted Regression. In *International Conference on Machine Learning*, 2020.

Mohammad Gheshlaghi Azar, Ian Osband, and Rémi Munos. Minimax Regret Bounds for Reinforcement Learning. In *International Conference on Machine Learning*, 2017.

Andrew Bennett, Dipendra Misra, and Nathan Kallus. Provable Safe Reinforcement Learning with Binary Feedback. In *International Conference on Artificial Intelligence and Statistics*, 2023.

Vivek S Borkar. A convex analytic approach to markov decision processes. *Probability Theory and Related Fields*, 78(4):583–602, 1988.

Archana Bura, Aria HasanzadeZonuzy, Dileep Kalathil, Srinivas Shakkottai, and Jean-Francois Chamberland. DOPE: Doubly Optimistic and Pessimistic Exploration for Safe Reinforcement Learning. In *Advances in Neural Information Processing Systems*, 2022.

Yang Cai, Haipeng Luo, Chen-Yu Wei, and Weiqiang Zheng. Uncoupled and Convergent Learning in Two-Player Zero-Sum Markov Games. *arXiv preprint arXiv:2303.02738*, 2023.

Shicong Cen, Chen Cheng, Yuxin Chen, Yuting Wei, and Yuejie Chi. Fast Global Convergence of Natural Policy Gradient Methods with Entropy Regularization. *Operations Research*, 70(4): 2563–2578, 2022.

Liyu Chen, Haipeng Luo, and Chen-Yu Wei. Impossible Tuning Made Possible: A New Expert Algorithm and Its Applications. In *Conference on Learning Theory*, 2021a.

Yi Chen, Jing Dong, and Zhaoran Wang. A Primal-Dual Approach to Constrained Markov Decision Processes. *arXiv preprint arXiv:2101.10895*, 2021b.

Christoph Dann, Tor Lattimore, and Emma Brunskill. Unifying PAC and Regret: Uniform PAC Bounds for Episodic Reinforcement Learning. In *Advances in Neural Information Processing Systems*, 2017.

Christoph Dann, Lihong Li, Wei Wei, and Emma Brunskill. Policy Certificates: Towards Accountable Reinforcement Learning. In *International Conference on Machine Learning*, 2019.

Dongsheng Ding, Kaiqing Zhang, Tamer Basar, and Mihailo Jovanovic. Natural Policy Gradient Primal-Dual Method for Constrained Markov Decision Processes. In *Advances in Neural Information Processing Systems*, 2020.

Dongsheng Ding, Xiaohan Wei, Zhuoran Yang, Zhaoran Wang, and Mihailo Jovanovic. Provably Efficient Safe Exploration via Primal-Dual Policy Optimization. In *International Conference on Artificial Intelligence and Statistics*, 2021.

Dongsheng Ding, Chen-Yu Wei, Kaiqing Zhang, and Alejandro Ribeiro. Last-Iterate Convergent Policy Gradient Primal-Dual Methods for Constrained MDPs. *arXiv preprint arXiv:2306.11700*, 2023.

Simon Du, Sham Kakade, Jason Lee, Shachar Lovett, Gaurav Mahajan, Wen Sun, and Ruosong Wang. Bilinear Classes: A Structural Framework for Provable Generalization in RL. In *International Conference on Machine Learning*, 2021.

Yonathan Efroni, Shie Mannor, and Matteo Pirotta. Exploration-Exploitation in Constrained MDPs. *arXiv preprint arXiv:2003.02189*, 2020.

Arnob Ghosh, Xingyu Zhou, and Ness Shroff. Achieving Sub-linear Regret in Infinite Horizon Average Reward Constrained MDP with Linear Function Approximation. In *International Conference on Learning Representations*, 2023.

Shangding Gu, Long Yang, Yali Du, Guang Chen, Florian Walter, Jun Wang, Yaodong Yang, and Alois Knoll. A Review of Safe Reinforcement Learning: Methods, Theory and Applications. *arXiv preprint arXiv:2205.10330*, 2022.

Ziqing Gu, Lingping Gao, Haitong Ma, Shengbo Eben Li, Sifa Zheng, Wei Jing, and Junbo Chen. Safe-State Enhancement Method for Autonomous Driving via Direct Hierarchical Reinforcement Learning. *IEEE Transactions on Intelligent Transportation Systems*, 24(9):9966–9983, 2023.

Aria HasanzadeZonuzy, Archana Bura, Dileep Kalathil, and Srinivas Shakkottai. Learning with Safety Constraints: Sample Complexity of Reinforcement Learning for Constrained MDPs. In *AAAI Conference on Artificial Intelligence*, 2021.

Jiafan He, Heyang Zhao, Dongruo Zhou, and Quanquan Gu. Nearly Minimax Optimal Reinforcement Learning for Linear Markov Decision Processes. In *International Conference on Machine Learning*, 2023a.

Xiangkun He, Haohan Yang, Zhongxu Hu, and Chen Lv. Robust Lane Change Decision Making for Autonomous Vehicles: An Observation Adversarial Reinforcement Learning Approach. *IEEE Transactions on Intelligent Vehicles*, 8(1):184–193, 2023b.

Pihe Hu, Yu Chen, and Longbo Huang. Nearly Minimax Optimal Reinforcement Learning with Linear Function Approximation. In *International Conference on Machine Learning*, 2022.

Nan Jiang, Akshay Krishnamurthy, Alekh Agarwal, John Langford, and Robert E Schapire. Contextual Decision Processes with Low Bellman Rank are PAC-Learnable. In *International Conference on Machine Learning*, 2017.

Chi Jin, Qinghua Liu, and Sobhan Miryoosefi. Bellman Eluder Dimension: New Rich Classes of RL Problems, and Sample-Efficient Algorithms. In *Advances in Neural Information Processing Systems*, 2021.

Tadashi Kozuno, Eiji Uchibe, and Kenji Doya. Theoretical Analysis of Efficiency and Robustness of Softmax and Gap-Increasing Operators in Reinforcement Learning. In *International Conference on Artificial Intelligence and Statistics*, 2019.

Hoang Le, Cameron Voloshin, and Yisong Yue. Batch Policy Learning Under Constraints. In *International Conference on Machine Learning*, 2019.

Tianjiao Li, Ziwei Guan, Shaofeng Zou, Tengyu Xu, Yingbin Liang, and Guanghui Lan. Faster Algorithm and Sharper Analysis for Constrained Markov Decision Process. *arXiv preprint arXiv:2110.10351*, 2021.

Tao Liu, Ruida Zhou, Dileep Kalathil, Panganamala Kumar, and Chao Tian. Learning Policies with Zero or Bounded Constraint Violation for Constrained MDPs. In *Advances in Neural Information Processing Systems*, 2021a.

Tao Liu, Ruida Zhou, Dileep Kalathil, PR Kumar, and Chao Tian. Policy Optimization for Constrained MDPs with Provable Fast Global Convergence. *arXiv preprint arXiv:2111.00552*, 2021b.

Sobhan Miryoosefi, Kianté Brantley, Hal Daume III, Miro Dudik, and Robert E Schapire. Reinforcement Learning with Convex Constraints. In *Advances in Neural Information Processing Systems*, 2019.

Ted Moskovitz, Brendan O'Donoghue, Vivek Veeriah, Sebastian Flennerhag, Satinder Singh, and Tom Zahavy. ReLOAD: Reinforcement Learning with Optimistic Ascent-Descent for Last-Iterate Convergence in Constrained MDPs. In *International Conference on Machine Learning*, 2023.

Adrian Müller, Pragnya Alatur, Volkan Cevher, Giorgia Ramponi, and Niao He. Truly no-regret learning in constrained mdps. *arXiv preprint arXiv:2402.15776*, 2024.

Santiago Paternain, Luiz Chamon, Miguel Calvo-Fullana, and Alejandro Ribeiro. Constrained Reinforcement Learning Has Zero Duality Gap. In *Advances in Neural Information Processing Systems*, 2019.

Reazul Hasan Russel, Mouhacine Benosman, and Jeroen Van Baar. Robust constrained-MDPs: Soft-constrained robust policy optimization under model uncertainty. *arXiv preprint arXiv:2010.04870*, 2020.

Chen Tessler, Daniel J Mankowitz, and Shie Mannor. Reward Constrained Policy Optimization. In *International Conference on Learning Representations*, 2018.

Sharan Vaswani, Lin Yang, and Csaba Szepesvári. Near-Optimal Sample Complexity Bounds for Constrained MDPs. *Advances in Neural Information Processing Systems*, 2022.

Yue Wang, Fei Miao, and Shaofeng Zou. Robust Constrained Reinforcement Learning. *arXiv preprint arXiv:2209.06866*, 2022.

Honghao Wei, Xin Liu, and Lei Ying. A Provably-Efficient Model-Free Algorithm for Constrained Markov Decision Processes. *arXiv preprint arXiv:2106.01577*, 2021.

Honghao Wei, Xin Liu, and Lei Ying. A Provably-Efficient Model-Free Algorithm for Infinite-Horizon Average-Reward Constrained Markov Decision Processes. In *AAAI Conference on Artificial Intelligence*, 2022.

Donghao Ying, Yuhao Ding, and Javad Lavaei. A Dual Approach to Constrained Markov Decision Processes with Entropy Regularization. In *International Conference on Artificial Intelligence and Statistics*, 2022.

Tom Zahavy, Brendan O'Donoghue, Guillaume Desjardins, and Satinder Singh. Reward is Enough for Convex MDPs. In *Advances in Neural Information Processing Systems*, 2021.

Sihan Zeng, Thinh T Doan, and Justin Romberg. Finite-Time Complexity of Online Primal-Dual Natural Actor-Critic Algorithm for Constrained Markov Decision Processes. In *2022 IEEE 61st Conference on Decision and Control (CDC)*, 2022.

Xianyuan Zhan, Haoran Xu, Yue Zhang, Xiangyu Zhu, Honglei Yin, and Yu Zheng. DeepThermal: Combustion Optimization for Thermal Power Generating Units Using Offline Reinforcement Learning. In *AAAI Conference on Artificial Intelligence*, 2022.

Liyuan Zheng and Lillian Ratliff. Constrained Upper Confidence Reinforcement Learning. In *Learning for Dynamics and Control*, 2020.

Dongruo Zhou, Quanquan Gu, and Csaba Szepesvari. Nearly Minimax Optimal Reinforcement Learning for Linear Mixture Markov Decision Processes. In *Conference on Learning Theory*, 2021.

## A Related Work

**Online CMDP Algorithms.** The seminal work by Efroni et al. (2020) provides both LP and primal-dual algorithms with sublinear regret. Liu et al. (2021a) and Bura et al. (2022) extend this work, achieving sublinear performance regret with a zero constraint violation guarantee during learning. Wei et al. (2021) propose a model-free primal-dual algorithm with sublinear regret, and Wei et al. (2022)

extend it to the average-reward setting. Ding et al. (2021), Amani et al. (2021), and Ghosh et al. (2023) propose CMDP algorithms with linear function approximation and sublinear regret guarantees. While the sublinear regret guarantee is common in online CMDPs, even an optimal regret algorithm may still make infinitely many mistakes of arbitrary quality (e.g., **Theorem 2** in Dann et al. (2017)).

While numerous online CMDP algorithms guarantee sublinear regret, those equipped with $(\varepsilon, \delta)$-PAC guarantees remain scarce. HasanzadeZonuzy et al. (2021) present an LP algorithm with an $(\varepsilon, \delta)$-PAC guarantee for online CMDPs but do not include a primal-dual algorithm. Zeng et al. (2022) and Vaswani et al. (2022) offer primal-dual algorithms with $(\varepsilon, \delta)$-PAC guarantees for infinite horizon CMDPs. However, Zeng et al. (2022) assume that the MDP is ergodic, and Vaswani et al. (2022) assume access to a simulator, both potentially obscuring the challenges associated with exploration. Bennett et al. (2023) provide an $(\varepsilon, \delta)$-PAC algorithm for the problem wherein the state-action-wise constraint signal is given by binary feedback. Although these $(\varepsilon, \delta)$-PAC algorithms ensure the performance of the last-iterate policy, they share a common limitation: they halt learning once an $\varepsilon$-optimal policy is found. This implies that existing $(\varepsilon, \delta)$-PAC CMDP algorithms never converge to optimal policies since they may make infinitely many mistakes with accuracy $\varepsilon/2$ (Dann et al., 2017).

**Convergence Guarantees of Primal-dual Algorithms without Exploration.** Numerous studies have struggled to design primal-dual algorithms that provide convergence guarantees even in non-exploration settings.

One fundamental approach involves utilizing the average of model parameters. By exploiting the convexity of CMDPs concerning the occupancy measure (Altman, 1999), a straightforward primal-dual method over the occupancy measure ensures that the average of the updated occupancy measures throughout training converges to an optimal solution (Zahavy et al., 2021). However, optimization over the occupancy measure becomes impractical when the state space is large. To address this challenge, Miryoosefi et al. (2019), Chen et al. (2021b), Li et al. (2021), and Liu et al. (2021b) propose policy optimization algorithms. While they offer certain performance guarantees for the average of past policies, there is no assurance for the last-iterate policy. This poses a challenge in cases where policy averaging is impractical, as in deep RL applications. Ding et al. (2020) provide a policy gradient algorithm independent of policy averaging, but their performance is guaranteed only for the average of past performances, not for the last-iterate policy.

Rather than the average-based approach, certain studies try to ensure the performance of the last-iterate policy. Moskovitz et al. (2023) propose a primal-dual algorithm with a last-iterate convergence guarantee. However, their optimization is over the occupancy measure, and furthermore, they do not provide non-asymptotic performance guarantees. Ying et al. (2022) achieve a non-asymptotic performance guarantee by employing a policy-entropy regularized Lagrange function, and Ding et al. (2023) present an extended algorithm regularized by both policy-entropy and Lagrange multipliers. However, their algorithms converge to a biased solution rather than the optimal solution due to the regularization.

## B Other Performance Measures

We use the following notations to define the high-probability regret.

$$\mathfrak{R}^K_{\mathrm{opt}+} \coloneqq \sum_{k=1}^{K} \Delta^k_{\mathrm{opt}+}, \ \mathfrak{R}^K_{\mathrm{vio}+} \coloneqq \sum_{k=1}^{K} \Delta^k_{\mathrm{vio}+}, \tag{7}$$

Intuitively, $\mathfrak{R}^K_{\mathrm{opt}+}$ and $\mathfrak{R}^K_{\mathrm{vio}+}$ quantify the cumulative optimality gap and cumulative constraint violation up to episode $K \in \mathbb{N}$, respectively.

**Definition B.1** (Regret)**.** For an episode length $K \in \mathbb{N}$ and a failure probability $\delta \in (0, 1]$, let $F_{\mathrm{HPR}}(\cdots)$ be shorthand for $F_{\mathrm{HPR}}(X, A, H, K, N, 1/b_{\mathrm{gap}}, \ln(1/\delta))$, where $F_{\mathrm{HPR}}$ is a real-valued function polynomial in all its arguments. An algorithm achieves $F_{\mathrm{HPR}}(\cdots)$ regret for $\delta$ if there

exists $F_{\mathrm{HPR}}(\cdots)$ such that

$$\mathbb{P}\big(\mathfrak{R}^K_{\mathrm{opt}+} > F_{\mathrm{HPR}}(\cdots) \ \vee \ \mathfrak{R}^K_{\mathrm{vio}+} > F_{\mathrm{HPR}}(\cdots)\big) \le \delta \ ,$$

where $\mathbb{P}$ is for the randomness of $\pi^k$ due to the stochastic interaction to the CMDP. We say that the regret is sublinear if $F_{\mathrm{HPR}}(\cdots)$ is sublinear with respect to $K$.

**Definition B.2** ($(\varepsilon,\delta)$-PAC)**.** For an admissible accuracy $\varepsilon > 0$, let $F_{\mathrm{PAC}}(\cdots)$ be shorthand for $F_{\mathrm{PAC}}(X, A, H, N, 1/b_{\mathrm{gap}}, 1/\varepsilon, \ln(1/\delta))$, where $F_{\mathrm{PAC}}$ is a real-valued function polynomial in all its arguments. For $\varepsilon > 0$ and $\delta > 0$, an algorithm achieves $(\varepsilon,\delta)$-PAC if there exists $F_{\mathrm{PAC}}(\cdots)$ such that

$$\mathbb{P}\big(\mathfrak{M}^{\varepsilon}_{\mathrm{opt}} > F_{\mathrm{PAC}}(\cdots) \ \vee \ \mathfrak{M}^{\varepsilon}_{\mathrm{vio}} > F_{\mathrm{PAC}}(\cdots)\big) \le \delta \ .$$

*Remark* B.3 (Weak Regret Measures). Rather than $\mathfrak{R}^K_{\mathrm{opt}+}$ and $\mathfrak{R}^K_{\mathrm{vio}+}$, the regret of an algorithm is often measured by the following $\mathfrak{R}^K_{\mathrm{opt}}$ and $\mathfrak{R}^K_{\mathrm{vio}}$ (Efroni et al., 2020; Liu et al., 2021a; Bura et al., 2022; Wei et al., 2021):

$$\mathfrak{R}^K_{\mathrm{opt}} \coloneqq \sum_{k=1}^{K} \Delta^k_{\mathrm{opt}} \ , \ \mathfrak{R}^K_{\mathrm{vio}} \coloneqq \sum_{k=1}^{K} \Delta^k_{\mathrm{vio}} \ . \tag{8}$$

Note that $\Delta^k_{\mathrm{opt}}$ and $\Delta^k_{\mathrm{vio}}$ might be negative since policy $\pi^k$ might violate the constraints. The negative temporal value allows an algorithm to cancel the past positive regret with the future negative regret. On the other hand, such *error cancellations* are not permitted in $\mathfrak{R}^K_{\mathrm{opt}+}$ and $\mathfrak{R}^K_{\mathrm{vio}+}$. Hence, regret guarantees for $\mathfrak{R}^K_{\mathrm{opt}+}$ and $\mathfrak{R}^K_{\mathrm{vio}+}$ are stronger than those for $\mathfrak{R}^K_{\mathrm{opt}}$ and $\mathfrak{R}^K_{\mathrm{vio}}$ in the sense that guarantees on the former imply guarantees on the latter, but not vice versa.

## C Naive Primal-Dual RL for Online CMDPs

**Algorithm 3** Naive Primal-Dual RL for online CMDPs

**Input:** $\delta \in (0,1]$ and iteration length $K \in \mathbb{N}$
Set $\lambda^1 \coloneqq \mathbf{0}$. Set $\pi^1_h(a \mid x) \coloneqq \frac{1}{A} \quad \forall (x,a,h) \in [H] \times \mathbf{X} \times \mathbf{A}$
**for** $k = 1, \dots K$ **do**
  Compute bonus $\beta^{k,\delta}$ by Equation (10)
  Compute $\widetilde{L}^k(\pi, \lambda)$ by Equation (11)
  Compute $\pi^{k+1}$ by a policy optimization over $\widetilde{L}^k(\cdot, \lambda^k)$
  Compute $\lambda^{k+1}$ by a gradient descent over $\widetilde{L}^k(\pi^k, \cdot)$
  Rollout $\pi^{k+1}$ and then update $n^k$ and $\overline{P}^k$
**end for**

For a better understanding of our proposed algorithm in Section 3, this section introduces the naive PD-RL algorithm for online CMDPs under Assumption 2.1 (e.g., Efroni et al. (2020); Liu et al. (2021a); Wei et al. (2021)). We provide the pseudocode of the algorithm as Algorithm 3.

Let $L : \Pi \times \mathbb{R}^N_+ \to \mathbb{R}$ be the Lagrange function such that for a policy $\pi \in \Pi$ and its multipliers $\lambda \in \mathbb{R}^N_+$,

$$L(\pi, \lambda) \coloneqq V^{\pi}_1\big[P, r^0\big](x_1) + \sum_{n=1}^{N} \lambda^n (V^{\pi}_1[P, r^n](x_1) - b^n) \ . \tag{9}$$

Let $\lambda^\star \in \arg\min_{\lambda \in \mathbb{R}^N_+} \max_{\pi \in \Pi} L(\pi, \lambda)$. The central idea of the naive algorithm involves exploring the environment to identify a pair $(\pi^\star, \lambda^\star)$, which is a saddle point of $L$ (Altman, 1999). In other words, the pair satisfies: $L(\pi^\star, \lambda) \ge L(\pi^\star, \lambda^\star) \ge L(\pi, \lambda^\star)$ for any $(\pi, \lambda) \in \Pi \times \mathbb{R}^N_+$.

The key to encouraging exploration is adhering to the *optimism-in-the-face-of-uncertainty* principle, which propels the agent with an optimistic policy. Given a failure probability $\delta \in (0,1]$ and

a fixed iteration length $K \in \mathbb{N}$, the principle is typically realized by introducing a bonus term $\beta_{\cdot}^{k,\delta}(\cdot,\cdot) : [H] \times \mathbf{X} \times \mathbf{A} \to \mathbb{R}$ into the reward function, where $\beta^{k,\delta}$ takes the form of

$$\beta_h^{k,\delta}(x,a) \approx \frac{\ln\left(K\delta^{-1}\right)}{n_h^k(x,a) \vee 1} + \sqrt{\frac{\ln(K\delta^{-1})}{n_h^k(x,a) \vee 1}} \,. \tag{10}$$

Here, we highlight the terms dependent on $n_h^k(x,a)$, $K$, and $\delta$ and conceal all other relevant terms for simplicity. Using $\overline{P}^k$ and $\beta^{k,\delta}$, the algorithm computes an optimistic Lagrange function $\widetilde{L}^k : \Pi \times \mathbb{R}_+^N \to \mathbb{R}$ such that

$$\widetilde{L}^k(\pi,\lambda) = V_1^\pi\left[\overline{P}^k, r^0 + \beta^{k,\delta}\right](x_1) + \sum_{n=1}^{N} \lambda^n \left(V_1^\pi\left[\overline{P}^k, r^n + \beta^{k,\delta}\right](x_1) - b^n\right) . \tag{11}$$

Let $\pi^k \in \Pi$ and $\lambda^k \in \mathbb{R}_+^N$ be the policy and the Lagrange multipliers at the episode $k$, respectively. The algorithm computes $\pi^{k+1}$ by a policy optimization method, such as policy gradient, with the aim of maximizing $\widetilde{L}^k(\cdot,\lambda^k)$. Subsequently, it computes $\lambda^{k+1}$ by a projected gradient descent to minimize $\widetilde{L}^k(\pi^k,\cdot)$. The naive algorithm iterates this update scheme until the episode reaches $K$.

**Challenges towards a Uniform-PAC Algorithm.** While the naive algorithm attains sublinear regret (Efroni et al., 2020; Liu et al., 2021a; Wei et al., 2021), it may fall short in delivering Uniform-PAC guarantees, encountering two primary challenges.

Firstly, even in the absence of exploration, i.e., $\overline{P}^k = P$ for all $k$, finding a saddle point $(\pi^\star,\lambda^\star)$ of the Lagrange function in Equation (9) is non-trivial. Even if we have $\lambda^k = \lambda^\star$ and an associated maximum policy $\pi^{k+1} \in \arg\max_{\pi\in\Pi} L(\pi,\lambda^\star)$, it is not guaranteed that $\pi^{k+1}$ represents an optimal policy. In some CMDPs where feasible policies in $\Pi_{\text{safe}}$ must be stochastic (Altman, 1999), the maximization can provide a deterministic $\pi^{k+1}$ that cannot be feasible. Hence, ensuring that $(\pi^k,\lambda^k)$ will be close to $(\pi^\star,\lambda^\star)$ poses a potential challenge.

Secondly, the naive bonus function in Equation (10) might not be adequate for achieving a uniform PAC guarantee. The inclusion of $\ln(K)$ in the bonus function implies that the algorithm attempts each action in each state infinitely often, potentially leading to an infinite number of mistakes (Dann et al., 2017).

## D Experiment Details

**Environment Construction.** We instantiated a CMDP with $X = 30$, $A = 3$, $H = 10$, and $N = 1$, employing a construction strategy akin to that of Dann et al. (2017). For all $x, a, h$, the transition probabilities $P_h(\cdot \mid x,a)$ were independently sampled from Dirichlet$(0.1,\ldots,0.1)$. This transition probability kernel is concentrated yet encompasses non-deterministic transition probabilities.

The reward values for the objective $r_h^0(x,a)$ are set to 0 with probability 0.5 and uniformly chosen at random from $[0,1]$ otherwise. The reward values for the constraint $r_h^1(x,a)$ are set to $1 - r_h^0(x,a)$. Thus, the constraint and objective are in conflict in this CMDP. This aligns with the CMDP construction strategy proposed by Moskovitz et al. (2023) to generate a straightforward CMDP where a naive primal-dual algorithm might struggle to converge.

The initial state $x_0$ is randomly chosen from $\mathbf{X}$ and fixed during the training. The constraint threshold is set as $b^1 = \frac{1}{2}\max_{\pi\in\Pi} V_1^\pi[P,r^1](x_1)$.

**Algorithm Implementations.** We modify the `OptPrimalDual-CMDP` algorithm from Efroni et al. (2020) for the setting where the reward functions $\mathbf{r}$ are known.

All the algorithms use $\delta = 0.1$. For `OptPrimalDual-CMDP` and `No-UPAC-bonus` that use the naive bonus function (Equation (10)), we set $K = 10^5$. For `UOpt-RPGPD` and `No-UPAC-bonus`, we set

$\alpha_\eta = 0.53$ and $\alpha_\tau = 0.4$, that do not contradict to Theorem 4.1. For `No-regularization`, we set $\alpha_\eta = 0.53$ and $\tau_k = 0$. For `No-adjustment`, we set $\tau_k = \eta_k = 0.1$.

Finally, we scale the bonus functions of all the algorithms by a factor of $10^{-3}$ to observe the algorithms' behavior in relatively smaller episodes.

# E Notation for Proofs

For any $h \in [H]$, let $w_\cdot^\pi[P] : [H] \to \Delta(\mathbf{X} \times \mathbf{A})$ be the occupancy measure of $\pi$ in $P$. In other words, for any $(h, x, a) \in [H] \times \mathbf{X} \times \mathbf{A}$, it satisfies

$$w_h^\pi[P](x, a) = \mathbb{P}(x_h = x, a_h = a \mid \pi, P) \,. \tag{12}$$

With an abuse of notation, we write $w_h^\pi[P](x) = \sum_{a \in \mathbf{A}} w_h^\pi[P](x, a)$.

Let $\widetilde{V}_\cdot^k(\cdot) : [H] \times \mathbf{X} \to \mathbb{R}$ be the regularized optimistic value function such that:

$$\widetilde{V}_h^k(x) = \widetilde{V}_h^{k,0}(x) + \sum_{n=1}^{N} \lambda^{k,n} \widetilde{V}_h^{k,n}(x) \quad \forall (h, x) \in [H] \times \mathbf{X} \,, \tag{13}$$

where $\widetilde{V}^{k;n}$ for $n \in \{0\} \cup [N]$ are defined in Algorithm 1.

We use the shorthand $H_{\text{ent}} \coloneqq H(1 + \ln A)$. Finally, for a set of positive values $\{a_n\}_{n=1}^N$, we write $x = \text{polylog}\,(a_1, \ldots, a_N)$ if there exists an absolute constant $C > 0$ and $\alpha > 0$ such that $x \le C\Big(\sum_{n=1}^N \ln a_n\Big)^\alpha$ holds.

# F Useful Lemmas

## F.1 RL Useful Lemmas

**Lemma F.1** (**Lemma 34** in Efroni et al. (2020))**.** *Let $\pi, \pi'$ be two policies, $P$ be a transition model, and $g$ be a reward function. Let $\widetilde{V}_\cdot^\pi(\cdot) : [H] \times \mathbf{X} \to \mathbb{R}$ be a function such that $\widetilde{V}_h^\pi = \pi_h \widetilde{Q}_h$ for all $h \in [H]$, where we defined $\widetilde{Q}_\cdot(\cdot, \cdot) : [H] \times \mathbf{X} \times \mathbf{A} \to \mathbb{R}$. Then,*

$$\begin{aligned}
&\widetilde{V}_1^\pi(x_1) - V_1^{\pi'}[P; g](x_1) \\
&= \sum_{h=1}^{H} \sum_{x \in \mathbf{X}} w_h^{\pi'}[P](x) \sum_{a \in \mathbf{A}} (\pi'_h(a \mid x) - \pi_h(a \mid x)) \widetilde{Q}_h(x, a) \\
&\quad + \sum_{h=1}^{H} \sum_{x,a \in \mathbf{X} \times \mathbf{A}} w_h^{\pi'}[P](x, a) \Big( \widetilde{Q}_h(x, a) - g_h(x, a) - \Big( P'_h \widetilde{V}_{h+1}^\pi \Big)(x, a) \Big)
\end{aligned}$$

**Lemma F.2** (Error to regret)**.** *Consider two sequences of real values $x_1, x_2, \ldots$ and $y_1, y_2, \ldots$. Assume that $0 \le x_i, y_i \le u$ for any $i \in \mathbb{N}$ with $u > 0$. For any $\varepsilon \in (0, u]$, assume that*

$$x_k - y_k \le \varepsilon$$

*on all $k \in \mathbb{N}$ except at most*

$$\left\lceil \frac{Z_1}{\varepsilon^\alpha} \left( \ln\left( \frac{Z_2}{\varepsilon} \right) \right)^\beta \right\rceil$$

*times, where $\alpha \ge 1$, $\beta \ge 0$, $Z_1 > 0$, and $Z_2 \ge \max\{u, e\}$ are constants that do not depend on $\varepsilon$. We also assume that $Z_2 \ge e Z_1^{\frac{1}{\alpha}}$. Then, for any $K \in \mathbb{N}$, it holds that*

$$\sum_{k=1}^{K} \max\{x_k - y_k, 0\} \le K^{1 - \frac{1}{\alpha}} Z_1^{\frac{1}{\alpha}} \,\text{polylog}(K, Z_1, Z_2, u) \,.$$

*Proof.* This proof follows the strategy of **Theorem A.1** in Dann et al. (2017).

Let $z \coloneqq \left(\frac{K}{Z_1}\right)^{\frac{1}{\alpha}}$ with $K \in \mathbb{N}$. Due to the assumption that $Z_2 \geq e Z_1^{\frac{1}{\alpha}}$, we have

$$Z_2 z = \frac{Z_2}{Z_1^{\frac{1}{\alpha}}} K^{\frac{1}{\alpha}} \geq \frac{Z_2}{Z_1^{\frac{1}{\alpha}}} \geq e \,. \tag{14}$$

Also, let $g(\varepsilon) \coloneqq \frac{Z_1}{\varepsilon^\alpha}\left(\ln\left(\frac{Z_2}{\varepsilon}\right)\right)^\beta$ and $\varepsilon_{\min} \coloneqq \frac{(\ln(Z_2 z))^{\frac{\beta}{\alpha}}}{z}$. Note that $g(\varepsilon)$ is well-defined since $\ln Z_2 - \ln \varepsilon \geq 0$ due to $Z_2 \geq \max\{u, e\}$. Then, it holds that

$$\begin{aligned} g(\varepsilon_{\min}) &= \frac{Z_1}{\varepsilon_{\min}^\alpha}\left(\ln \frac{Z_2}{\varepsilon_{\min}}\right)^\beta = Z_1 \frac{z^\alpha}{(\ln(Z_2 z))^\beta}\left(\ln(Z_2 z) - \frac{\beta}{\alpha} \underbrace{\ln \ln(Z_2 z)}_{\geq 0 \text{ due to Equation (14)}}\right)^\beta \\ &\leq Z_1 \frac{z^\alpha}{(\ln(Z_2 z))^\beta} (\ln(Z_2 z))^\beta = Z_1 z^\alpha = K \,. \end{aligned}$$

Since $g(\varepsilon)$ is monotonically decreasing for $\varepsilon > 0$, due to $g(\varepsilon_{\min}) \leq K$, we have $g(\varepsilon) \leq K$ for any $\varepsilon \in [\varepsilon_{\min}, u]$. Using the above results with $x_k - y_k \leq u$ for any $k$, it holds that

$$\sum_{k=1}^K \max\{x_k - y_k, 0\} \leq \int_0^u g(\varepsilon) \leq K \varepsilon_{\min} + \int_{\varepsilon_{\min}}^u g(\varepsilon) d\varepsilon \,.$$

For the intuition of the above inequality, see **Figure 3** in Dann et al. (2017). We are going to bound both terms in the right-hand side separately.

For the first term, we have

$$K \varepsilon_{\min} = K \frac{\left(\ln Z_2 + \frac{1}{\alpha} \ln K - \frac{1}{\alpha} \ln Z_1\right)^{\frac{\beta}{\alpha}}}{\left(\frac{K}{Z_1}\right)^{\frac{1}{\alpha}}} = K^{1-\frac{1}{\alpha}} Z_1^{\frac{1}{\alpha}} \operatorname{polylog}(K, Z_1, Z_2) \,.$$

For the second term, we have

$$\int_{\varepsilon_{\min}}^u g(\varepsilon) d\varepsilon = \int_{\varepsilon_{\min}}^u \frac{Z_1}{\varepsilon^\alpha}\left(\ln\left(\frac{Z_2}{\varepsilon}\right)\right)^\beta d\varepsilon \leq Z_1 \left(\ln\left(\frac{Z_2}{\varepsilon}\right)\right)^\beta \int_{\varepsilon_{\min}}^u \varepsilon^{-\alpha} d\varepsilon \,.$$

When $\alpha = 1$, we have

$$Z_1 \left(\ln\left(\frac{Z_2}{\varepsilon}\right)\right)^\beta \int_{\varepsilon_{\min}}^u \varepsilon^{-1} d\varepsilon = Z_1 \left(\ln\left(\frac{Z_2}{\varepsilon}\right)\right)^\beta (\ln u - \ln \varepsilon_{\min}) = Z_1 \operatorname{polylog}(K, Z_1, Z_2, u) \,.$$

When $\alpha > 1$, we have

$$Z_1 \left(\ln\left(\frac{Z_2}{\varepsilon}\right)\right)^\beta \int_{\varepsilon_{\min}}^u \varepsilon^{-\alpha} d\varepsilon \leq \frac{Z_1}{\alpha - 1}\left(\ln\left(\frac{Z_2}{\varepsilon}\right)\right)^\beta \varepsilon_{\min}^{1-\alpha} = K^{1-\frac{1}{\alpha}} Z_1^{\frac{1}{\alpha}} \operatorname{polylog}(K, Z_1, Z_2) \,,$$

where we used $\varepsilon_{\min}^{1-\alpha} = K^{1-\frac{1}{\alpha}} Z_1^{\frac{1}{\alpha}-1} \operatorname{polylog}(K, Z_1, Z_2)$.

Therefore, we conclude that

$$\sum_{k=1}^K \max\{x_k - y_k, 0\} \leq K^{1-\frac{1}{\alpha}} Z_1^{\frac{1}{\alpha}} \operatorname{polylog}(K, Z_1, Z_2, u) \,.$$

□

**Lemma F.3.** *Suppose that Assumption 2.1 holds. For any $\tau \in (0, \infty)$ and $(\pi, \lambda) \in \Pi \times \mathbb{R}_+^N$, it holds that*

$$V_1^{\pi}[P, r^0](x_1) + \sum_{n=1}^{N} \lambda_\tau^{\star,n}(V_1^{\pi}[P, r^n](x_1) - b^n) \le V_1^{\pi_\tau^\star}[P, r^0](x_1) + \sum_{n=1}^{N} \lambda_\tau^{\star,n}\left(V_1^{\pi_\tau^\star}[P, r^n](x_1) - b^n\right) + \tau H_{\text{ent}} ,$$

*and* $$\sum_{n=1}^{N} \lambda_\tau^{\star,n}\left(V_1^{\pi_\tau^\star}[P, r^n](x_1) - b^n\right) \le \sum_{n=1}^{N} \lambda^n\left(V_1^{\pi_\tau^\star}[P, r^n](x_1) - b^n\right) + \frac{\tau}{2}\|\lambda\|_2^2 .$$

*Proof.* $L_\tau(\pi, \lambda_\tau^\star) \le L_\tau(\pi_\tau^\star, \lambda_\tau^\star)$ due to Lemma 3.1 indicates that

$$\begin{aligned}
&V_1^{\pi}\left[P, r^0\right](x_1) + \sum_{n=1}^{N} \lambda_\tau^{\star,n}(V_1^{\pi}[P, r^n](x_1) - b^n) + \frac{\tau}{2}\|\lambda_\tau^\star\|_2^2 \\
\le &V_1^{\pi}\left[P, r^0, \tau\right](x_1) + \sum_{n=1}^{N} \lambda_\tau^{\star,n}(V_1^{\pi}[P, r^n](x_1) - b^n) + \frac{\tau}{2}\|\lambda_\tau^\star\|_2^2 \\
\le &V_1^{\pi_\tau^\star}\left[P, r^0\right](x_1) + \tau H_{\text{ent}} + \sum_{n=1}^{N} \lambda_\tau^{\star,n}\left(V_1^{\pi_\tau^\star}[P, r^n](x_1) - b^n\right) + \frac{\tau}{2}\|\lambda_\tau^\star\|_2^2 .
\end{aligned}$$

The first claim holds by rearranging the above inequality.

Also, $L_\tau(\pi_\tau^\star, \lambda_\tau^\star) \le L_\tau(\pi_\tau^\star, \lambda)$ due to Lemma 3.1 indicates that

$$\begin{aligned}
&V_1^{\pi_\tau^\star}\left[P, r^0, \tau\right](x_1) + \sum_{n=1}^{N} \lambda_\tau^{\star,n}\left(V_1^{\pi_\tau^\star}[P, r^n](x_1) - b^n\right) \\
\le &V_1^{\pi_\tau^\star}\left[P, r^0, \tau\right](x_1) + \sum_{n=1}^{N} \lambda_\tau^{\star,n}\left(V_1^{\pi_\tau^\star}[P, r^n](x_1) - b^n\right) + \frac{\tau}{2}\|\lambda_\tau^\star\|_2^2 \\
\le &V_1^{\pi_\tau^\star}\left[P, r^0, \tau\right](x_1) + \sum_{n=1}^{N} \lambda^n\left(V_1^{\pi_\tau^\star}[P, r^n](x_1) - b^n\right) + \frac{\tau}{2}\|\lambda\|_2^2 .
\end{aligned}$$

The second claim holds by rearranging the above inequality. □

**Lemma F.4** (Properties of $\lambda_\tau^\star$ and $\pi_\tau^\star$)**.** *Suppose that $0 \le \tau \le 1$. Under Assumption 2.1, we have*

$$\sum_{n=1}^{N} \lambda_\tau^{\star,n} \le \frac{H_{\text{ent}}}{b_{\text{gap}}} \quad \text{and} \quad V_1^{\pi_\tau^\star}[P, r^n](x_1) - b^n \ge -\frac{\tau H_{\text{ent}}}{b_{\text{gap}}} \quad \forall n \in [N] .$$

*Proof.* We first assume $\tau > 0$. Since $\lambda_\tau^{\star,n} = \arg\min_{\lambda^n \in \mathbb{R}_+} \lambda^n\left(V_1^{\pi_\tau^\star}[P, r^n](x_1) - b^n\right) + \frac{\tau}{2}(\lambda^n)^2$, $\lambda_\tau^{\star,n}$ can be analytically computed as $\frac{1}{\tau}\max\left\{b^n - V_1^{\pi_\tau^\star}[P, r^n](x_1), 0\right\}$ for every $n \in [N]$. Thus, from Lemma 3.1 and Assumption 2.1,

$$\begin{aligned}
L_\tau(\pi_{\text{safe}}, \lambda_\tau^\star) &= V_1^{\pi_{\text{safe}}}[P, r^0, \tau](x_1) + \sum_{n=1}^{N} \lambda_\tau^{\star,n} \underbrace{(V_1^{\pi_{\text{safe}}}[P, r^n](x_1) - b^n)}_{\ge b_{\text{gap}} \text{ by Assumption 2.1}} \\
&\le V_1^{\pi_\tau^\star}[P, r^0, \tau](x_1) + \underbrace{\sum_{n=1}^{N} \lambda_\tau^{\star,n}\left(V_1^{\pi_\tau^\star}[P, r^n](x_1) - b^n\right)}_{\le 0 \text{ from the analytical expression of } \lambda_\tau^{\star,n}} \le V_1^{\pi_\tau^\star}[P, r^0, \tau](x_1) .
\end{aligned} \tag{15}$$

By rearrangement,

$$b_{\text{gap}} \sum_{n=1}^{N} \lambda_\tau^{\star,n} \le V_1^{\pi_\tau^\star}[P, r^0, \tau](x_1) - V_1^{\pi_{\text{safe}}}[P, r^0, \tau](x_1) \le H(1 + \tau \ln A) \le H_{\text{ent}} ,$$

which concludes the proof of the first claim for $\tau > 0$. Furthermore, as $\lambda_\tau^{\star,n} \leq \sum_{m=1}^N \lambda_\tau^{\star;m}$ for any $n \in [N]$,

$$b^n - V_1^{\pi_\tau^\star}[P, r^n](x_1) \leq \max\Big\{b^n - V_1^{\pi_\tau^\star}[P, r^n](x_1), 0\Big\} = \tau\lambda_\tau^{\star,n} \leq \frac{\tau H_{\mathrm{ent}}}{b_{\mathrm{gap}}} \ ,$$

which concludes the proof of the second claim for $\tau > 0$.

The first and second claims when $\tau = 0$ obviously hold. Indeed, $V_1^{\pi_0^\star}[P, r^n](x_1) - b^n \geq 0$ for every $n \in [N]$ by definition, and thus, $\lambda_\tau^{\star,n}\Big(V_1^{\pi_\tau^\star}[P, r^n](x_1) - b^n\Big) = 0$ when $\tau = 0$, which provides Equation (15). □

**Lemma F.5** (KL to optimality gap)**.** *Let $\tau > 0$ and $\pi \in \Pi$ with $\pi > \mathbf{0}$. Assume that Assumption 2.1 holds and*

$$\sum_{h=1}^H \sum_{x \in \mathbf{X}} w_h^{\pi_\tau^\star}[P](x) \mathrm{KL}[\pi_{\tau,h}^\star(\cdot \mid x), \pi_h(\cdot \mid x)] \leq \varepsilon \ .$$

*Then, $\pi$ satisfies*

$$V_1^{\pi^\star}[P, r](x_1) - V_1^{\pi}[P, r](x_1) \leq \tau H_{\mathrm{ent}} + H\sqrt{2H\varepsilon}$$
$$\text{and} \quad b^n - V_1^{\pi}[P, r^n](x_1) \leq \frac{\tau H_{\mathrm{ent}}}{b_{\mathrm{gap}}} + H\sqrt{2H\varepsilon} \quad \forall n \in [N] \ .$$

*Proof.* Note that

$$V_1^{\pi^\star}[P, r](x_1) - V_1^{\pi}[P, r](x_1) = \underbrace{V_1^{\pi^\star}[P, r](x_1) - V_1^{\pi_\tau^\star}[P, r](x_1)}_{\text{(i)}} + \underbrace{V_1^{\pi_\tau^\star}[P, r](x_1) - V_1^{\pi}[P, r](x_1)}_{\text{(ii)}} .$$

For the term (ii), we have

$$\begin{aligned}
\text{(ii)} &= V_1^{\pi_\tau^\star}[P, r](x_1) - V_1^{\pi}[P, r](x_1) \\
&\overset{(a)}{\leq} H \sum_{h=1}^H \sum_{x \in \mathbf{X}} w_h^{\pi_\tau^\star}[P](x) \big\|\pi_{\tau,h}^\star(\cdot \mid x) - \pi_h(\cdot \mid x)\big\|_1 \\
&\overset{(b)}{\leq} H \sum_{h=1}^H \sqrt{\sum_{x \in \mathbf{X}} w_h^{\pi_\tau^\star}[P](x) \Big\|\pi_{\tau,h}^\star(\cdot \mid x) - \pi_h(\cdot \mid x)\Big\|_1^2} \\
&\overset{(c)}{\leq} H \sum_{h=1}^H \sqrt{2 \sum_{x \in \mathbf{X}} w_h^{\pi_\tau^\star}[P](x) \mathrm{KL}[\pi_{\tau,h}^\star(\cdot \mid x), \pi_h(\cdot \mid x)]} \\
&\overset{(d)}{\leq} H \sqrt{2H \sum_{h=1}^H \sum_{x \in \mathbf{X}} w_h^{\pi_\tau^\star}[P](x) \mathrm{KL}[\pi_{\tau,h}^\star(\cdot \mid x), \pi_h(\cdot \mid x)]} \leq H\sqrt{2H\varepsilon} \ ,
\end{aligned}$$

where (a) is due to Lemma F.1 (b) uses the fact that $(\mathbb{E}[x])^2 \leq \mathbb{E}[x^2]$, (c) uses the Pinsker's inequality, and (d) uses Lemma F.10.

For the term (i), the second claim of Lemma F.3 with $\lambda = \mathbf{0}$ indicates that $\sum_{n=1}^N \lambda_\tau^{\star,n}\Big(V_1^{\pi_\tau^\star}[P, r^n](x_1) - b^n\Big) \leq 0$ . Then, the first claim of Lemma F.3 with $\pi = \pi^\star$ indicates that

$$\text{(i)} = V_1^{\pi^\star}[P, r^0](x_1) - V_1^{\pi_\tau^\star}[P, r^0](x_1) \leq \tau H_{\mathrm{ent}} \ .$$

Therefore, the optimality gap is bounded as

$$V_1^{\pi^\star}[P, r^0](x_1) - V_1^{\pi^k}[P, r^0](x_1) \leq \tau H_{\mathrm{ent}} + H\sqrt{2H\varepsilon} \ .$$

This concludes the proof of the first claim.

For the second claim, for any $n \in [N]$, we have

$$b^n - V_1^{\pi^k}[P, r^n](x_1) = \underbrace{b^n - V_1^{\pi^\star_{\tau_k}}[P, r^n](x_1)}_{\text{(i)}} + \underbrace{V_1^{\pi^\star_{\tau_k}}[P, r^n](x_1) - V_1^{\pi^k}[P, r^n](x_1)}_{\text{(ii)}} .$$

By taking a similar transformation of the first claim, the term (ii) can be bounded by (ii) $\le H\sqrt{2H\varepsilon}$. Furthermore, Lemma F.4 indicates that (i) $\le \frac{\tau H_{\text{ent}}}{b_{\text{gap}}}$. Therefore, we have

$$b^n - V_1^{\pi^k}[P, r^n](x_1) \le \frac{\tau H_{\text{ent}}}{b_{\text{gap}}} + H\sqrt{2H\varepsilon} \quad \forall n \in [N] .$$

□

### F.2 Other Useful Lemmas

**Lemma F.6.** *For $x \in \Delta(\mathbf{A})$ with $|\mathbf{A}| = A$ and $0 < c \le 1$, it holds that*

$$\sum_{a \in \mathbf{A}} (x(a))^c (\ln x(a))^2 \le A^{1-c} \left( \frac{2}{c} - 1 + \ln A \right)^2 .$$

*Proof.* Note that the second derivative of $f(x) \coloneqq x^c \big(1 - \frac{2}{c} + \ln x\big)^2$ is

$$\frac{\partial^2}{\partial x^2} f(x) = x^{c-2} \left( \ln(x) \left( \underbrace{(c-1)c\ln(x)}_{\ge 0} + \underbrace{2(c-1)c + 2}_{\ge 0} \right) + (c-1)c \right) \le 0$$

Therefore, $f(x)$ is a concave function.

Accordingly,

$$\sum_{a \in \mathbf{A}} (x(a))^c (\ln x(a))^2 \le \sum_{a \in \mathbf{A}} (x(a))^c \left( 1 - \frac{2}{c} + \ln x(a) \right)^2 \le A^{1-c} \left( \frac{2}{c} - 1 + \ln A \right)^2 ,$$

where the second inequality is due to Jensen's inequality.

□

**Lemma F.7** (**Lemma 12** in Cai et al. (2023))**.** *For $x \in (0, 1)$ and $y > 0$, we have $x^{1-y} - x \le -yx^{1-y} \ln x$.*

**Lemma F.8** (Inequality for Mirror Descent)**.** *For $\ell : \mathbf{A} \to \mathbb{R}$, $x \in \Delta(\mathbf{A})$, $1 \ge \eta > 0$, and $1 \ge \kappa \ge 0$, let*

$$x' = \arg\min_{\tilde{x} \in \Delta(\mathbf{A})} \left\{ \sum_{a \in \mathcal{A}} \tilde{x}(a)(\ell(a) + \kappa \ln x(a)) + \frac{1}{\eta} \mathrm{KL}[\tilde{x}, x] \right\}$$

*Then, for any $u \in \Delta(\mathbf{A})$, it holds that*

$$\sum_{a \in \mathbf{A}} (x(a) - u(a))(\ell(a) + \kappa \ln x(a)) \le \frac{\mathrm{KL}[u, x] - \mathrm{KL}[u, x']}{\eta} + 2\eta\kappa^2 (1 + \ln A)^2 + 2\eta \sum_a x(a)(\ell(a))^2 .$$

*Proof.* By the standard analysis of online mirror descent (e.g., **Lemma 14** from Chen et al. (2021a)), we have

$$\sum_{a \in \mathbf{A}} (x_a - u_a)(\ell_a + \tau \ln x_a) \le \frac{\mathrm{KL}[u, x] - \mathrm{KL}[u, x']}{\eta} + \underbrace{\sum_{a \in \mathbf{A}} (x_a - x'_a)(\ell_a + \tau \ln x_a) - \frac{1}{\eta} \mathrm{KL}[x', x]}_{\clubsuit} . \tag{16}$$

We will bound ♣. For two positive vectors $x, y \in \mathbb{R}^A$ such that $x(a), y(a) > 0$, define a mapping $\phi$ such that,

$$\phi(x, y) = \sum_a x(a)(\ln x(a) - \ln y(a)) - x(a) + y(a) \ .$$

Note that when $x, y \in \Delta(\mathbf{A})$, $\phi(x, y)$ is equivalent to $\mathrm{KL}[x, y]$. Then, for any $g \in \mathbb{R}^A$ and $x' \in \Delta(\mathbf{A})$,

$$\begin{aligned}
&\sum_{a\in\mathbf{A}} -x'(a)g(a) - \frac{1}{\eta}\mathrm{KL}[x', x] \\
=&\sum_{a\in\mathbf{A}} -x'(a)g(a) - \frac{1}{\eta}\phi(x', x) \\
\le&\max_{y\in\mathbb{R}^A} \sum_{a\in\mathbf{A}} -y(a)g(a) - \frac{1}{\eta}\phi(y, x) \\
=&\max_{y\in\mathbb{R}^A} \sum_{a\in\mathbf{A}} -y(a)g(a) - \frac{1}{\eta}\sum_a y(a)(\ln y(a) - \ln x(a) - 1) + x(a) \\
=&\max_{y\in\mathbb{R}^A} f(y) - \frac{1}{\eta}\sum_{a\in\mathbf{A}} x(a) \ ,
\end{aligned}$$

where we defined a function $f : y \in \mathbb{R}^A \mapsto \sum_{a\in\mathbf{A}} -y(a)g(a) - \frac{1}{\eta}\sum_a y(a)(\ln y(a) - \ln x(a) - 1)$. Note that $f(y)$ is a strongly concave function and has a unique maxima. Let $y^\star \coloneqq \arg\max_{y\in\mathbb{R}^A} f(y)$. It is easy to verify that $y^\star$ satisfies

$$y^\star(a) = x(a)\exp(-\eta g(a)) \ .$$

As $\ln y^\star(a) = \ln x(a) - \eta g(a)$, we have $\sum_{a\in\mathbf{A}} -y^\star(a)g(a) = \frac{1}{\eta}\sum_{a\in\mathbf{A}} y^\star(a)(\ln y^\star(a) - \ln x(a))$ and $\sum_{a\in\mathbf{A}} x(a)g(a) = \frac{1}{\eta}\sum_{a\in\mathbf{A}} x(a)(\ln y^\star(a) - \ln x(a))$. Therefore,

$$\begin{aligned}
\sum_{a\in\mathbf{A}}(x(a) - y^\star(a))g(a) - \frac{1}{\eta}\phi(y^\star, x) &= \frac{1}{\eta}\phi(x, y^\star) \\
&= \frac{1}{\eta}\sum_{a\in\mathbf{A}} x(a)(\ln x(a) - \ln y^\star(a)) - x(a) + y^\star(a) \\
&= \frac{1}{\eta}\sum_{a\in\mathbf{A}} x(a)(\eta g(a) - 1 + \exp(-\eta g(a))) \\
&\overset{(a)}{\le} \frac{1}{\eta}\sum_{a\in\mathbf{A}} x(a)(\eta g(a))^2 = \eta\sum_{a\in\mathbf{A}} x(a)(g(a))^2 \ ,
\end{aligned} \tag{17}$$

where (a) uses $\exp(-z) - 1 + z \le z^2$ for $z \ge -1$.

Then, by setting $g(a) = \ell(a) + \tau\ln x(a)$ in Equation (17), ♣ can be bounded as

$$\begin{aligned}
♣ =& \sum_{a\in\mathbf{A}}(x(a) - x'(a))(\ell(a) + \tau\ln x(a)) - \frac{1}{\eta}\mathrm{KL}[x', x] \\
=& \sum_{a\in\mathbf{A}}(x(a) - x'(a))(\ell(a) + \tau\ln x(a)) - \frac{1}{\eta}\phi(x', x) \\
\le& \max_{y\in\mathbb{R}^A}\sum_{a\in\mathbf{A}}(x(a) - y(a))(\ell(a) + \tau\ln x(a)) - \frac{1}{\eta}\phi(y, x) \\
\le& \eta\sum_{a\in\mathbf{A}} x(a)(\ell(a) + \tau\ln x(a))^2 \\
\overset{(a)}{\le}& 2\eta\sum_{a\in\mathbf{A}} x(a)(\ell(a))^2 + 2\eta\tau^2\sum_{a\in\mathbf{A}} x(a)(\ln x(a))^2 \\
\overset{(b)}{\le}& 2\eta\sum_{a\in\mathbf{A}} x(a)(\ell(a))^2 + 2\eta\tau^2(1 + \ln A)^2
\end{aligned}$$

where (a) uses $(a+b)^2 \le 2a^2 + 2b^2$ and (b) uses Lemma F.6.

The claim holds by plugging this result to Equation (16). □

**Lemma F.9** (Inequality for Gradient Descent)**.** *For some $u, \lambda \in \mathbb{R}_+$, $\eta > 0$, and $\ell \in \mathbb{R}$, let $\lambda' \coloneqq \mathrm{clip}[\lambda + \eta\ell, 0, u]$ . Then, for any $\lambda^\star \in [0, u]$,*

$$\ell(\lambda - \lambda^\star) \le \frac{1}{2\eta}\big((\lambda - \lambda^\star)^2 - (\lambda' - \lambda^\star)^2\big) + \frac{\eta}{2}\ell^2 \,.$$

*Proof.* Let $\bar{\lambda}' \coloneqq \lambda + \eta\ell$. Since $\lambda^\star \in [0, u]$, we have

$$(\lambda' - \lambda^\star)^2 = (\mathrm{clip}\big[\bar{\lambda}', 0, u\big] - \lambda^\star)^2 \le (\bar{\lambda}' - \lambda^\star)^2 \,.$$

Therefore,

$$(\lambda' - \lambda^\star)^2 \le (\bar{\lambda}' - \lambda^\star)^2 = (\lambda + \eta\ell - \lambda^\star)^2 = (\lambda - \lambda^\star)^2 - 2\eta\ell(\lambda - \lambda^\star) + \eta^2\ell^2 \,.$$

The claim holds by rearranging the above inequality. □

**Lemma F.10.** *For any positive real numbers $x_1, x_2, \ldots, x_n$, $\sum_{i=1}^n \sqrt{x_i} \le \sqrt{n}\sqrt{\sum_{i=1}^n x_i}$.*

*Proof.* Due to the Cauchy-Schwarz inequality, we have $\left(\frac{\sum_{i=1}^n \sqrt{x_i}}{n}\right)^2 \le \frac{\sum_{i=1}^n x_i}{n}$. Taking the square root of the inequality proves the claim. □

**Lemma F.11.** *Let $g : \mathbb{N} \to \mathbb{R}$ be a function such that*

$$g(k) = Z_1 k^{-\alpha}(\ln(Z_2 k))^\beta$$

*where $\alpha, \beta > 0$, $Z_1 > 0$, and $Z_2 \ge 1$ are constants that do not depend on $k$. Then, for any $\varepsilon \in (0, \infty)$, there exists a constant $k^\star = \widetilde{\mathcal{O}}\left(Z_1^{\frac{1}{\alpha}} \varepsilon^{-\frac{1}{\alpha}}\right)$ such that $g(k) \le \varepsilon$ for all $k \ge k^\star$.*

*Proof.* Consider a function $\kappa \in [1/Z_2, \infty) \mapsto Z_1\kappa^{-\alpha}(\ln(Z_2\kappa))^\beta$. Note that

$$Z_1\kappa^{-\alpha}(\ln(Z_2\kappa))^\beta \le \varepsilon \iff \frac{1}{(Z_2\kappa)^{\alpha/\beta}} \ln(Z_2\kappa) \le \left(\frac{\varepsilon}{Z_1}\right)^{1/\beta} \frac{1}{Z_2^{\alpha/\beta}} \iff \frac{\ln x}{x^\eta} \le c \,,$$

where $x \coloneqq Z_2\kappa$, $\eta \coloneqq \alpha/\beta$, and $c \coloneqq \varepsilon^{1/\beta} Z_1^{-1/\beta} Z_2^{-\alpha/\beta}$. Let $f(x) \coloneqq x^{-\eta} \ln x$. Its derivative is given by

$$f'(x) = x^{-1-\eta}(1 - \eta \ln x) \,.$$

Therefore $f$ is increasing when $1 \le x < e^{1/\eta}$, takes its maximum $1/(\eta e)$ at $x = e^{1/\eta}$, and decreasing towards 0. Hence there exists some $x^\star$ such that $f(x) \le c$ for all $x \ge x^\star$. Desired $k^\star$ can be obtained by $\lceil x^\star / Z_2 \rceil$, where $\lceil \cdot \rceil$ is the ceiling function.

As $f(x) \le 1/(\eta e)$, we assume $1/(\eta e) > c$ otherwise the claim trivially holds. Following the same discussion as above, it can be shown that a function $x \mapsto x^{-\eta\lambda} \ln x$ for $\lambda \in (0, 1)$ takes its maximum $1/(e\eta\lambda)$ at $x = e^{1/(\eta\lambda)}$. Then, since

$$f(x) = \frac{1}{x^{(1-\lambda)\eta}} \frac{1}{x^{\lambda\eta}} \ln x \le \frac{1}{e\eta\lambda} \frac{1}{x^{(1-\lambda)\eta}} \,,$$

it suffices to find $x^\star$ such that $(ce\eta\lambda)^{-1} \le (x^\star)^{\eta(1-\lambda)}$, that is,

$$x^\star \ge \frac{1}{\eta(1-\lambda)} \ln \frac{1}{ce\eta\lambda} \,.$$

Finally we set $\lambda$ to $1-(ce\eta)^{1/\eta}$. Note that $1-(ce\eta)^{1/\eta} \in (0,1)$ due to the assumption that $1/(\eta e) > c$. Then,

$$\frac{1}{\eta(1-\lambda)} \ln \frac{1}{ce\eta\lambda} = \frac{1}{\eta(ce\eta)^{1/\eta}} \ln \frac{1}{ce\eta(1-(ce\eta)^{1/\eta})} = \frac{Z_2}{\eta(e\eta)^{1/\eta}} \left(\frac{Z_1}{\varepsilon}\right)^{1/\alpha} \ln \frac{1}{ce\eta(1-(ce\eta)^{1/\eta})}$$
$$= \frac{Z_2\beta}{\alpha} \left(\frac{\beta}{e\alpha}\right)^{\beta/\alpha} \left(\frac{Z_1}{\varepsilon}\right)^{1/\alpha} \ln \frac{\beta}{ce\alpha(1-(ce\alpha/\beta)^{\beta/\alpha})} .$$

Therefore, $k^\star$ is obtained as

$$k^\star = \left\lceil \frac{\beta}{\alpha} \left(\frac{\beta}{e\alpha}\right)^{\beta/\alpha} \left(\frac{Z_1}{\varepsilon}\right)^{1/\alpha} \ln \frac{\beta}{ce\alpha(1-(ce\alpha/\beta)^{\beta/\alpha})} \right\rceil = \widetilde{\mathcal{O}}\left(Z_1^{\frac{1}{\alpha}} \varepsilon^{-\frac{1}{\alpha}}\right) .$$

□

**Lemma F.12** (**Lemma E.6** in Dann et al. (2017)). $\operatorname{llnp}(xy) \le \operatorname{llnp}(x) + \operatorname{llnp}(y) + 1$ *for all* $x, y \ge 0$.

**Lemma F.13** (Cai et al. (2023) **Lemma 3**). *Let* $0 < \alpha < 1$ *and* $k \ge \left(\frac{24}{1-\alpha} \ln \frac{12}{1-\alpha}\right)^{\frac{1}{1-\alpha}}$. *Then,* $k^{1-\alpha} \ge 12 \ln k$.

**Lemma F.14.** *Let* $0 < \alpha < 1$, $0 \le \beta \le 2$, $c \in \{0\} \cup \mathbb{N}$, *and let* $k \ge \left(\frac{24}{1-\alpha} \ln \frac{12}{1-\alpha}\right)^{\frac{1}{1-\alpha}}$. *Then,*

$$\sum_{i=1}^{k} \left( (i+c)^{-\beta} \prod_{j=i+1}^{k} \left(1-(j+c)^{-\alpha}\right) \right) \le 9 \ln(k+c)(k+c)^{-\beta+\alpha}$$

*Proof.* The case when $c = 0$ is equivalent to **Lemma 1** in Cai et al. (2023). For $c \ge 1$, we have

$$\sum_{i=1}^{k} \left( (i+c)^{-\beta} \prod_{j=i+1}^{k} \left(1-(j+c)^{-\alpha}\right) \right) = \sum_{i=1+c}^{k+c} \left( i^{-\beta} \prod_{j=i+1}^{k+c} \left(1-j^{-\alpha}\right) \right)$$
$$\le \sum_{i=1}^{k+c} \left( i^{-\beta} \prod_{j=i+1}^{k+c} \left(1-j^{-\alpha}\right) \right) \le 9 \ln(k+c)(k+c)^{-\beta+\alpha} ,$$

where the last inequality uses the result when $c = 0$ with replacing $k$ by $k + c$. □

**Lemma F.15.** *Let* $0 < \alpha < 1$, $0 \le \beta \le 2$, $c \in \{0\} \cup \mathbb{N}$, *and let* $k \ge \left(\frac{24}{1-\alpha} \ln \frac{12}{1-\alpha}\right)^{\frac{1}{1-\alpha}}$. *Then,*

$$\max_{1 \le i \le k} \left( (i+c)^{-\beta} \prod_{j=i+1}^{k} \left(1-(j+c)^{-\alpha}\right) \right) \le 4(k+c)^{-\beta}$$

*Proof.* The case when $c = 0$ is equivalent to **Lemma 2** in Cai et al. (2023). For $c \ge 1$, we have

$$\max_{1 \le i \le k} \left( (i+c)^{-\beta} \prod_{j=i+1}^{k} \left(1-(j+c)^{-\alpha}\right) \right) \le \max_{1+c \le i \le k+c} \left( (i+c)^{-\beta} \prod_{j=i+1}^{k+c} \left(1-(j+c)^{-\alpha}\right) \right)$$
$$\max_{1 \le i \le k+c} \left( (i+c)^{-\beta} \prod_{j=i+1}^{k+c} \left(1-(j+c)^{-\alpha}\right) \right) \le 4(k+c)^{-\beta} .$$

where the last inequality uses the result when $c = 0$ with replacing $k$ by $k + c$. □

## G Proof of Lemma 3.1

This section provides the proof of Lemma 3.1. While the proof is a direct modification of **Lemma 6** in Ding et al. (2023) to the finite-horizon setting, we include the proof here for completeness. The following lemma is the restatement of Lemma 3.1.

**Lemma G.1.** *For any $\tau \in (0, \infty)$, there exists a unique saddle point $(\pi_\tau^\star, \lambda_\tau^\star) \in \Pi \times \mathbb{R}_+^N$ such that*

$$L_\tau(\pi_\tau^\star, \lambda) \geq L_\tau(\pi_\tau^\star, \lambda_\tau^\star) \geq L_\tau(\pi, \lambda_\tau^\star) \;\; \forall (\pi, \lambda) \in \Pi \times \mathbb{R}_+^N \, .$$

*Proof.* Let $\mathbf{W} \coloneqq \{w^\pi[P] \mid \pi \in \Pi\}$ be the set of all the occupancy measures on $P$. Let $\bar{L}_\tau : \mathbf{W} \times \mathbb{R}_+^N \to \mathbb{R}$ be the regularized Lagrange function in terms of occupancy measure such that

$$\begin{aligned} \bar{L}_\tau(w, \lambda) &= \sum_{h,x,a \in [H] \times \mathbf{X} \times \mathbf{A}} w_h(x,a) \left( r_h^0(x,a) + \sum_{n=1}^N \lambda^n r_h^n(x,a) - \frac{b^n}{H} \right) + \tau \mathcal{H}(w) + \frac{\tau}{2} \|\lambda\|_2^2 \\ \text{where } \mathcal{H}(w) &\coloneqq - \sum_{h,x,a \in [H] \times \mathbf{X} \times \mathbf{A}} w_h(x,a) \ln \frac{w_h(x,a)}{\sum_{a' \in \mathbf{A}} w_h(x,a')} \end{aligned} \tag{18}$$

For this Lagrange function, $\bar{L}_\tau(w^\pi[P], \lambda) = L_\tau(\pi, \lambda)$ holds for any $\pi$ and $\lambda$. From the one-to-one correspondence between policies and occupancy measures, it is sufficient to prove that there exists a unique saddle point $(w_\tau^\star, \lambda_\tau^\star) \in \mathbf{W} \times \mathbb{R}_+^N$ such that

$$\bar{L}_\tau(w_\tau^\star, \lambda) \geq \bar{L}_\tau(w_\tau^\star, \lambda_\tau^\star) \geq \bar{L}_\tau(w, \lambda_\tau^\star) \;\; \forall (w, \lambda) \in \mathbf{W} \times \mathbb{R}_+^N \, .$$

Note that $\mathbf{W}$ is convex and compact (Borkar, 1988). Furthermore, for any $w \in \mathbf{W}$, there exists some finite $\lambda_w \in \mathbb{R}_+^N$ such that $\bar{L}_\tau(w, \lambda_w) = \min_{\lambda \in \mathbb{R}_+^N} \bar{L}_\tau(w, \lambda)$ due to the regularization $\frac{\tau}{2}\|\lambda\|_2^2$. Thus, according to Sion's minimax theorem, the claim immediately holds by showing that $\bar{L}_\tau(w, \lambda)$ is strictly concave in $w$ and strictly convex in $\lambda$.

It is obvious that $\bar{L}_\tau(w, \lambda)$ is strictly convex in $\lambda$. We then show that $\bar{L}_\tau(w, \lambda)$ is strictly concave in $w$. According to Equation (18), it is sufficient to show that $\mathcal{H}(w)$ is strictly convex in $w$. For any $w^1, w^2 \in \mathbf{W}$ and $\alpha \in [0, 1]$, we have

$$\begin{aligned} &\mathcal{H}(\alpha w^1 + (1-\alpha) w^2) \\ =& - \sum_{h,x,a \in [H] \times \mathbf{X} \times \mathbf{A}} \left( \alpha w_h^1(x,a) + (1-\alpha) w_h^2(x,a) \right) \ln \frac{\alpha w_h^1(x,a) + (1-\alpha) w_h^2(x,a)}{\sum_{a' \in \mathbf{A}} \alpha w_h^1(x,a') + (1-\alpha) w_h^2(x,a')} \\ \overset{(a)}{\geq}& - \sum_{h,x,a \in [H] \times \mathbf{X} \times \mathbf{A}} \alpha w_h^1(x,a) \ln \frac{\alpha w_h^1(x,a)}{\sum_{a' \in \mathbf{A}} \alpha w_h^1(x,a')} - \sum_{h,x,a \in [H] \times \mathbf{X} \times \mathbf{A}} (1-\alpha) w_h^2(x,a) \ln \frac{(1-\alpha) w_h^2(x,a)}{\sum_{a' \in \mathbf{A}} (1-\alpha) w_h^2(x,a')} \\ =& \alpha \mathcal{H}(w^1) + (1-\alpha) \mathcal{H}(w^2) \, , \end{aligned}$$

where (a) is due to the log-sum inequality $(\sum_i a_i) \ln \frac{\sum_i a_i}{\sum_i b_i} \leq \sum_i a_i \ln \frac{a_i}{b_i}$ for non-negative $a_i$ and $b_i$. Note that the equality of the log-sum inequality holds if and only if $\frac{a_i}{b_i}$ are equal for all $i$. Therefore, when $w^1 \neq w^2$, we have

$$\mathcal{H}(\alpha w^1 + (1-\alpha) w^2) > \alpha \mathcal{H}(w^1) + (1-\alpha) \mathcal{H}(w^2) \, .$$

This concludes the proof. □

# H Proofs for Theorem 4.1

## H.1 Failure Events and Their Probabilities

In this section, we use a refined notation of the bonus function: $\beta_h^{k,\delta}(x,a) = \sum_{y\in\mathbf{X}} \beta_h^{k,\delta}(x,a,y)$, where

$$\beta_h^{k,\delta}(x,a,y) = 2\sqrt{\overline{P}_h^k(y \mid x,a)}\phi\big(n_h^k(x,a)\big) + 5\phi\big(n_h^k(x,a)\big)^2 ,$$

and

$$\phi(n) = 1 \wedge \sqrt{\frac{2}{n\vee 1}\left(\mathrm{llnp}(2n) + \ln\frac{48X^2AH}{\delta}\right)} .$$

For any $\delta > 0$, we define the following failure events.

$$\mathcal{F}_\delta^P \coloneqq \left\{\exists x,y,a,h,k : \left|\overline{P}_h^k(y \mid x,a) - P_h(y \mid x,a)\right| \geq \beta_h^{k,\delta}(x,a,y)\right\} ,$$

$$\mathcal{F}_\delta^N \coloneqq \left\{\exists x,a,h,k : n_h^k(x,a) < \frac{1}{2}\sum_{j<k} w_h^{\pi^j}[P](x,a) - \ln\frac{4XAH}{\delta}\right\} ,$$

$$\mathcal{F}_\delta^{L1} \coloneqq \left\{\exists x,a,h,k : \left\|\overline{P}_h^k(\cdot \mid x,a) - P_h(\cdot \mid x,a)\right\|_1 \geq \sqrt{\frac{4}{n_h^k(x,a)\vee 1}\left(2\,\mathrm{llnp}\left(n_h^k(x,a)\right) + \ln\frac{12XAH(2^X-2)}{\delta}\right)}\right\} ,$$

and $\mathcal{F}_\delta \coloneqq \mathcal{F}_\delta^P \cup \mathcal{F}_\delta^N \cup \mathcal{F}_\delta^{L1}$ , for which the following results hold.

**Lemma H.1.** *For any $\delta$, the failure probabilities are bounded as follows:*

$$\mathbb{P}\big(\mathcal{F}_\delta^P\big) \leq \frac{\delta}{2} , \ \mathbb{P}\big(\mathcal{F}_\delta^N\big) \leq \frac{\delta}{4X} , \ \mathbb{P}\big(\mathcal{F}_\delta^{L1}\big) \leq \frac{\delta}{4X} , \ \textit{and } \mathbb{P}(\mathcal{F}_\delta) \leq \delta ,$$

*Proof.* The bound for $\mathcal{F}_\delta^P$ holds by a direct application of **Lemma 6** from Dann et al. (2019). The bounds for $\mathcal{F}_\delta^N$ and $\mathcal{F}_\delta^{L_1}$ hold by **Corollary E.4** and **Corollary E.3** from Dann et al. (2017), respectively.

Accordingly,

$$\mathbb{P}(\mathcal{F}_\delta) \leq \mathbb{P}\big(\mathcal{F}_\delta^P\big) + \mathbb{P}\big(\mathcal{F}_\delta^N\big) + \mathbb{P}\big(\mathcal{F}_\delta^{L1}\big) \leq \delta .$$

□

## H.2 Bounds for Policy Estimation

**Lemma H.2** (Policy Estimation Optimism)**.** *Assume that $\mathcal{F}_\delta^c$ holds. For any $(h,x,a) \in [H]\times\mathbf{X}\times\mathbf{A}$ and for any episode $k$, the following bound holds*

$$\widetilde{Q}_h^{k,n}(x,a) - r_h^n(x,a) - P_h\widetilde{V}_{h+1}^{k,n}(x,a) \geq 0 \quad \forall n \in \{0\}\cup[N]$$

$$\textit{and } \widetilde{Q}_h^k(x,a) - r_h^0(x,a) - \sum_{n=1}^N \lambda^{k,n} r_h^n(x,a) - P_h\widetilde{V}_{h+1}^k(x,a) \geq 0 .$$

*Proof.* For $n = 0$, we have

$$
\begin{aligned}
&\widetilde{Q}_h^{k,0}(x,a) - r_h^0(x,a) - P_h\widetilde{V}_{h+1}^{k,0}(x,a) \\
&= \min\Big\{r_h^0(x,a) + (1+\tau_k \ln A)H\beta_h^k(x,a) + \overline{P}_h\widetilde{V}_{h+1}^{k,0}(x,a),\ (1+\tau_k \ln A)(H-h+1)\Big\} - r_h^0(x,a) - P_h\widetilde{V}_{h+1}^{k,0}(x,a) \\
&= \min\Big\{(1+\tau_k \ln A)H\beta_h^k(x,a) + \big(\overline{P}_h - P_h\big)\widetilde{V}_{h+1}^{k,0}(x,a),\ (1+\tau_k \ln A)(H-h+1) - r_h^0(x,a) - P_h\widetilde{V}_{h+1}^{k,0}(x,a)\Big\} \\
&\overset{(a)}{\geq} \min\Big\{(1+\tau_k \ln A)H\beta_h^k(x,a) + \big(\overline{P}_h - P_h\big)\widetilde{V}_{h+1}^{k,0}(x,a),\ 0\Big\} \\
&\geq \min\left\{\sum_{y\in\mathbf{X}}(1+\tau_k \ln A)H\beta_h^k(x,a,y) - \big|\overline{P}_h(y \mid x,a) - P_h(y \mid x,a)\big|\big|\widetilde{V}_{h+1}^{k,0}(y)\big|,\ 0\right\} \\
&\overset{(b)}{\geq} \min\left\{(1+\tau_k \ln A)H\sum_{y\in\mathbf{X}}\big(\beta_h^k(x,a,y) - \big|\overline{P}_h(y \mid x,a) - P_h(y \mid x,a)\big|\big),\ 0\right\} \overset{(c)}{\geq} 0\,,
\end{aligned}
\tag{19}
$$

where (a) is due to $r_h^0(x,a)+P_h\widetilde{V}_{h+1}^{k,0}(x,a) \leq 1+(1+\tau_k \ln A)(H-(h+1)+1) = (1+\tau_k \ln A)(H-h+1)$, (b) is due to $\left|\widetilde{V}_{h+1}^{k,0}(y)\right| \leq (1+\tau_k \ln A)H$, and (c) is due to the good event $\mathcal{F}^c$. Similarly, for $n \in [N]$, it is easy to verify that

$$
\begin{aligned}
&\widetilde{Q}_h^{k,n}(x,a) - r_h^n(x,a) - P_h\widetilde{V}_h^{k,n}(x,a) \\
&\geq \min\left\{H\sum_{y\in\mathbf{X}}\big(\beta_h^k(x,a,y) - \big|\overline{P}_h(y \mid x,a) - P_h(y \mid x,a)\big|\big),\ 0\right\} \geq 0\,.
\end{aligned}
\tag{20}
$$

The first claim holds by Equation (19) and Equation (20).

According to the definition of $\widetilde{Q}^k$ in Algorithm 1, we have

$$
\begin{aligned}
&\widetilde{Q}_h^k(x,a) - r_h^0(x,a) - \sum_{n=1}^N \lambda^{k,n} r_h^n(x,a) - P_h\widetilde{V}_{h+1}^k(x,a) \\
=&\widetilde{Q}_h^{k,0}(x,a) - r_h^0(x,a) - P_h\widetilde{V}_{h+1}^{k,0}(x,a) + \sum_{n=1}^N \lambda^{k,n}\Big(\widetilde{Q}_h^{k,n}(x,a) - r_h^n(x,a) - P_h\widetilde{V}_{h+1}^{k,n}(x,a)\Big)\,.
\end{aligned}
\tag{21}
$$

The second claim holds by inserting Equation (19) and Equation (20) into Equation (21). □

The following *nice-episode* technique from Dann et al. (2017) is useful to derive the estimation error bound (Lemma H.6).

**Definition H.3** ($\varepsilon$-Nice Episode)**.** For $\varepsilon > 0$, let $w_{\min}(\varepsilon) \coloneqq \frac{\varepsilon}{HH_{\text{ent}}XA}$. An episode $k$ is $\varepsilon$-nice if and only if for all $h, x, a \in [H] \times \mathbf{X} \times \mathbf{A}$, the following two conditions hold:

$$
w_h^k[P](x,a) \leq w_{\min}(\varepsilon) \quad \vee \quad n_h^k(x,a) \geq \frac{1}{4}\sum_{i<k} w_h^i(x,a)\,.
$$

We also define a set

$$
\mathbf{U}_h^k(\varepsilon) \coloneqq \big\{(x,a) \in \mathbf{X}\times\mathbf{A} \;\big|\; w_h^k[P](x,a) \geq w_{\min}(\varepsilon)\big\}\,.
$$

**Lemma H.4** (**Lemma E.2** in Dann et al. (2017))**.** *On the good event $\mathcal{F}_\delta^c$, the number of episodes that are not $\varepsilon$-nice is at most*

$$
\frac{6X^2AH^3}{\varepsilon}\ln\frac{4HXA}{\delta}\,.
$$

**Lemma H.5** (**Lemma E.3** in Dann et al. (2017)). *Fix $r \geq 1$, $\varepsilon > 0$, $C > 0$ and $D \geq 1$. $C$ may depend polynomially on $\varepsilon^{-1}$ and relevant quantities of $X, A, H, \delta^{-1}$. $D$ may depend poly-logarithmically on relevant quantities. Then,*

$$\sum_{h\in[H]} \sum_{x,a\in\mathbf{U}_h^k(\varepsilon)} w_h^k[P](x,a) \left( \frac{C\left(\mathrm{llnp}\left(n_h^k(x,a)\right) + D\right)}{n_h^k(x,a)} \right)^{1/r} \leq \varepsilon$$

*on all but at most*

$$\frac{CXAH^r}{\varepsilon^r} \operatorname{polylog}\left(X, A, H, \delta^{-1}, \varepsilon^{-1}\right)$$

*$\varepsilon$-nice episodes.*

**Lemma H.6** (Estimation Error Bound). *Assume that the good event $\mathcal{F}_\delta^c$ holds. For any $k \in \mathbb{N}$ and $\varepsilon > 0$, it holds that*

$$\widetilde{V}_1^{k,0}(x_1) - V_1^{\pi^k}[P, r^0, \tau_k](x_1) \leq \varepsilon$$
$$\text{and } \widetilde{V}_1^{k,n}(x_1) - V_1^{\pi^k}[P, r^n](x_1) \leq \varepsilon \quad \forall n \in [N]$$

*on all episodes $k \in \mathbb{N}$ except at most*

$$\frac{X^2AH^4}{\varepsilon^2} \operatorname{polylog}(X, A, H, \delta^{-1}, \varepsilon^{-1}) + \frac{XAH^3}{\varepsilon} \operatorname{polylog}(X, A, H, \delta^{-1}, \varepsilon^{-1})$$

*episodes.*

*Proof.* Consider $n = 0$. Note that $\widetilde{V}_h^{k,0} = \pi_h^k\left(\widetilde{Q}_h^{k,0} - \tau_k \ln \pi_h^k\right)$. Using Lemma F.1, we have

$$\begin{aligned}
&\widetilde{V}_1^{k,0}(x_1) - V_1^{\pi^k}\left[P, r^0, \tau_k\right](x_1) \\
=&\sum_{h=1}^{H} \sum_{x,a\in\mathbf{X}\times\mathbf{A}} w_h^{\pi^\star_{\tau_k}}[P](x)\left(\left(\widetilde{Q}_h^k(x,a) - \tau_k \ln \pi_h^k(x,a)\right) - \left(r_h^0(x,a) - \tau_k \ln \pi_h^k(x,a)\right) - \left(P_h\widetilde{V}_h^k\right)(x,a)\right) \\
=&\sum_{h=1}^{H} \sum_{x,a\in\mathbf{X}\times\mathbf{A}} w_h^{\pi^\star_{\tau_k}}[P](x)\left(\widetilde{Q}_h^k(x,a) - r_h^0(x,a) - \left(P_h\widetilde{V}_h^k\right)(x,a)\right) \geq 0\,,
\end{aligned}$$

where the last inequality is due to Lemma H.2.

Accordingly, we have

$$\begin{aligned}
0 \leq&\widetilde{V}_1^{k,0}(x_1) - V_1^{\pi^k}\left[P, r^0, \tau_k\right](x_1) \\
=&\sum_{h=1}^{H} \sum_{x,a\in\mathbf{X}\times\mathbf{A}} w_h^{\pi_k}[P](x,a)\left(\widetilde{Q}_h^{k,0}(x,a) - r_h^n(x,a) - P_h\widetilde{V}_{h+1}^{k,0}(x,a)\right) \\
=&\sum_{h=1}^{H} \sum_{x,a\notin\mathbf{U}_h^k(\frac{\varepsilon}{3})} \underbrace{w_h^{\pi_k}[P](x,a)}_{\leq w_{\min}(\frac{\varepsilon}{3})} \underbrace{\left(\widetilde{Q}_h^{k,0}(x,a) - r_h^n(x,a) - P_h\widetilde{V}_{h+1}^{k,0}(x,a)\right)}_{\leq H_{\mathrm{ent}}} \\
&+\sum_{h=1}^{H} \sum_{x,a\in\mathbf{U}_h^k(\frac{\varepsilon}{3})} w_h^{\pi_k}[P](x,a)\left(\widetilde{Q}_h^{k,0}(x,a) - r_h^n(x,a) - P_h\widetilde{V}_{h+1}^{k,0}(x,a)\right) \\
\leq&\underbrace{HH_{\mathrm{ent}}XAw_{\min}\left(\frac{\varepsilon}{3}\right)}_{=\frac{\varepsilon}{3} \text{ due to } w_{\min}(\frac{\varepsilon}{3})} + \underbrace{\sum_{h=1}^{H} \sum_{x,a\in\mathbf{U}_h^k(\frac{\varepsilon}{3})} w_h^{\pi_k}[P](x,a)\left(\widetilde{Q}_h^{k,0}(x,a) - r_h^n(x,a) - P_h\widetilde{V}_{h+1}^{k,0}(x,a)\right)}_{=:\diamondsuit^k}\,.
\end{aligned}$$

where $\mathbf{U}_h^k(\frac{\varepsilon}{3})$ is defined in Definition H.3.

According to the definition of $\widetilde{Q}^{k,n}$, $\diamondsuit^k$ is bounded as

$$\diamondsuit^k \leq \sum_{h=1}^{H} \sum_{x,a \in \mathbf{U}_h^k(\frac{\varepsilon}{3})} w_h^{\pi_k}[P](x,a)\left(H_{\mathrm{ent}}\beta_h^k(x,a) + \left|\left(\overline{P}_h^k - P_h\right)\widetilde{V}_{h+1}^{k,0}(x,a)\right|\right).$$

On the good event $\mathcal{F}_k^{L1}$, Hölder's inequality indicates that

$$\begin{aligned}
\left|\left(\overline{P}_h^k - P_h\right)\widetilde{V}_{h+1}^{k,0}(x,a)\right| &\leq \left\|\left(\overline{P}_h^k(\cdot \mid x,a) - P_h(\cdot \mid x,a)\right)\right\|_1 \left\|\widetilde{V}_{h+1}^{k,0}\right\|_\infty \\
&\leq H_{\mathrm{ent}}\sqrt{\frac{4}{n_h^k(x,a)\vee 1}\left(2\,\mathrm{llnp}\left(n_h^k(x,a)\right) + \ln\frac{12XAH(2^X-2)}{\delta}\right)} \\
&\leq \sqrt{\frac{8H_{\mathrm{ent}}^2 X}{n_h^k(x,a)\vee 1}\left(\mathrm{llnp}\left(n_h^k(x,a)\right) + \frac{1}{2}\ln\frac{24XAH}{\delta}\right)}.
\end{aligned}$$

Also, the bonus term is bounded as

$$\begin{aligned}
\beta_h^k(x,a) \leq& \frac{5X}{n_h^k(x,a)\vee 1}\left(2\,\mathrm{llnp}\left(2n_h^k(x,a)\right) + 2\ln\frac{48X^2AH}{\delta}\right) \\
&+ \sum_{y\in X}\sqrt{\frac{4\overline{P}_h\left(y \mid x,a\right)}{n_h^k(x,a)\vee 1}\left(2\,\mathrm{llnp}\left(2n_h^k(x,a)\right) + 2\ln\frac{48X^2AH^2}{\delta}\right)} \\
\leq& \frac{10X}{n_h^k(x,a)\vee 1}\left(\mathrm{llnp}\left(n_h^k(x,a)\right) + 1 + \ln\frac{48X^2AH}{\delta}\right) \\
&+ \sqrt{\frac{8X}{n_h^k(x,a)\vee 1}\left(\mathrm{llnp}\left(n_h^k(x,a)\right) + 1 + \ln\frac{48X^2AH}{\delta}\right)} \\
\leq& \frac{10X}{n_h^k(x,a)\vee 1}\left(\mathrm{llnp}\left(n_h^k(x,a)\right) + 2\ln\frac{48X^2AH}{\delta}\right) + \sqrt{\frac{8X}{n_h^k(x,a)\vee 1}\left(\mathrm{llnp}\left(n_h^k(x,a)\right) + 2\ln\frac{48X^2AH}{\delta}\right)},
\end{aligned}$$

where the second inequality is due to Lemma F.10 and Lemma F.12.

Accordingly, we have

$$\begin{aligned}
\diamondsuit^k \leq& \sum_{h=1}^{H}\sum_{x,a\in\mathbf{U}_h^k(\frac{\varepsilon}{3})} w_h^{\pi_k}[P](x,a)\left(\frac{10X}{n_h^k(x,a)\vee 1}\left(\mathrm{llnp}\left(n_h^k(x,a)\right) + 2\ln\frac{48X^2AH}{\delta}\right)\right) \\
&+ \sum_{h=1}^{H}\sum_{x,a\in\mathbf{U}_h^k(\frac{\varepsilon}{3})} w_h^{\pi_k}[P](x,a)\sqrt{\frac{35H_{\mathrm{ent}}^2X}{n_h^k(x,a)\vee 1}\left(\mathrm{llnp}\left(n_h^k(x,a)\right) + 2\ln\frac{48X^2AH}{\delta}\right)},
\end{aligned}$$

where we used $\sqrt{8} + \sqrt{8H_{\mathrm{ent}}^2} \leq \sqrt{16 + 16H_{\mathrm{ent}}^2} \leq \sqrt{35H_{\mathrm{ent}}^2}$ due to $(a+b)^2 \leq 2a^2 + 2b^2$ for $a, b \in \mathbb{R}$.

For the first term, we apply Lemma H.5 with $r = 1$, $C = 10X$, and $D = 2\ln\frac{48X^2AH}{\delta}$ to bound this term by $\frac{\varepsilon}{3}$ on all but at most

$$\frac{X^2AH}{\varepsilon}\,\mathrm{polylog}(X, A, H, \delta^{-1}, \varepsilon^{-1})$$

nice episodes.

For the second term, we apply Lemma H.5 with $r = 2$, $C = 35H_{\mathrm{ent}}^2X$, and $D = 2\ln\frac{48X^2AH}{\delta}$ to bound this term by $\frac{\varepsilon}{3}$ on all but at most

$$\frac{X^2AH_{\mathrm{ent}}^2H^2}{\varepsilon^2}\,\mathrm{polylog}(X, A, H, \delta^{-1}, \varepsilon^{-1}) = \frac{X^2AH^4}{\varepsilon^2}\,\mathrm{polylog}(X, A, H, \delta^{-1}, \varepsilon^{-1})$$

nice episodes.

By combining the above results, it holds that $\widetilde{V}_1^{k,0}(x_1) - V_1^{\pi^k}\big[P, r^0, \tau_k\big](x_1) \leq \varepsilon$ on all $\varepsilon$-nice episodes $k \in \mathbb{N}$ except at most

$$\frac{X^2AH^4}{\varepsilon^2}\operatorname{polylog}(X, A, H, \delta^{-1}, \varepsilon^{-1})$$

nice episodes. Due to Lemma H.4, the number of not $\varepsilon$-nice episodes is at most $\frac{XAH^3}{\varepsilon}\operatorname{polylog}(X, A, H, \delta^{-1}, \varepsilon^{-1})$. Therefore, $\widetilde{V}_1^{k,0}(x_1) - V_1^{\pi^k}\big[P, r^0, \tau_k\big](x_1) \leq \varepsilon$ holds on all episodes $k \in \mathbb{N}$ except at most

$$\frac{X^2AH^4}{\varepsilon^2}\operatorname{polylog}(X, A, H, \delta^{-1}, \varepsilon^{-1}) + \frac{XAH^3}{\varepsilon}\operatorname{polylog}(X, A, H, \delta^{-1}, \varepsilon^{-1})$$

episodes. This concludes the proof of the first claim. It is easy to verify that the second claim (for $n \in [N]$) holds using the same proof strategy. □

### H.3 Duality Gap Analysis

Recall the regularized optimistic value function $\widetilde{V}^k$ defined in Equation (13). We decompose the duality gap at episode $k$ as

$$\begin{aligned}
0 &\leq L_{\tau_k}(\pi^\star_{\tau_k}, \lambda^k) - L_{\tau_k}(\pi^k, \lambda^\star_{\tau_k}) \\
&= L_{\tau_k}(\pi^\star_{\tau_k}, \lambda^k) - \Big(\widetilde{V}_1^k(x_1) + \frac{\tau_k}{2}\|\lambda^k\|_2^2\Big) + \Big(\widetilde{V}_1^k(x_1) + \frac{\tau_k}{2}\|\lambda^k\|_2^2\Big) - L_{\tau_k}(\pi^k, \lambda^\star_{\tau_k}) \\
&= \underbrace{V_1^{\pi^\star_{\tau_k}}\big[P, r^0, \tau_k\big](x_1) + \sum_{n=1}^N \lambda^{k,n}\Big(V_1^{\pi^\star_{\tau_k}}[P, r^n](x_1) - b^n\Big) - \widetilde{V}_1^k(x_1) + \sum_{n=1}^N \lambda^{k,n} b^n}_{\clubsuit^k} \\
&\quad \underbrace{- \sum_{n=1}^N \lambda^{k,n} b^n + \widetilde{V}_1^k(x_1) - V_1^{\pi^k}\big[P, r^0, \tau_k\big](x_1) - \sum_{n=1}^N \lambda^{\star,n}_{\tau_k}\Big(V_1^{\pi^k}[P, r^n](x_1) - b^n\Big) + \frac{\tau_k}{2}\|\lambda^k\|_2^2 - \frac{\tau_k}{2}\|\lambda^\star_{\tau_k}\|_2^2}_{\heartsuit^k} .
\end{aligned}$$

#### H.3.1 $\clubsuit^k$ Bound

Let $\gamma_h^k(x) \coloneqq \mathrm{KL}[\pi^\star_{\tau_k,h}(\cdot \mid x), \pi_h^k(\cdot \mid x)]$. It is easy to see that

$$V_1^{\pi^\star_{\tau_k}}\big[P, -\ln \pi^k\big](x_1) - V_1^{\pi^\star_{\tau_k}}\big[P, -\ln \pi^\star_{\tau_k}\big](x_1) = \sum_{h=1}^H \sum_{x \in \mathbf{X}} w_h^{\pi^\star_{\tau_k}}[P](x)\gamma_h^k(x) .$$

**Lemma H.7.** *Let* $C_1 \coloneqq 2H_{\mathrm{ent}}^2\Big(1 + \frac{H}{b_{\mathrm{gap}}}\Big)^2$ *and* $C_{2,k} \coloneqq 2(1 + \ln A)^2$. *Assume that the good event* $\mathcal{F}_\delta^c$ *holds. Then,*

$$\clubsuit^k \leq \sum_{h=1}^H \sum_{x \in \mathbf{X}} w_h^{\pi^\star_{\tau_k}}[P](x)\frac{1}{\eta_k}\big((1 - \eta_k\tau_k)\gamma_h^k(x) - \gamma_h^{k+1}(x)\big) + \eta_k C_1 + \eta_k\tau_k^2 C_{2,k} .$$

*Proof.* Let $g^k \coloneqq r^0 - \tau_k \ln \pi^k + \sum_{n=1}^N \lambda^{k,n} r^n$. Note that $V_1^\pi\left[P, r^0, \tau\right] = V_1^\pi\left[P, r^0\right] + \tau V_1^\pi[P, -\ln \pi]$ for any $\pi \in \Pi$ and $\tau \geq 0$. Accordingly,

$$\begin{aligned}
\clubsuit^k &= V_1^{\pi^\star_{\tau_k}}\left[P, r^0, \tau_k\right](x_1) + \sum_{n=1}^N \lambda^{k,n}\left(V_1^{\pi^\star_{\tau_k}}[P, r^n](x_1) - b^n\right) - \widetilde{V}_1^k(x_1) + \sum_{n=1}^N \lambda^{k,n} b^n \\
&= V_1^{\pi^\star_{\tau_k}}\left[P, r^0 + \sum_{n=1}^N \lambda^{k,n} r^n\right](x_1) - \widetilde{V}_1^k(x_1) + \tau_k V_1^{\pi^\star_{\tau_k}}\left[P, -\ln \pi^\star_{\tau_k}\right](x_1) \\
&= V_1^{\pi^\star_{\tau_k}}\left[P, g^k\right](x_1) - \widetilde{V}_1^k(x_1) + \tau_k V_1^{\pi^\star_{\tau_k}}\left[P, -\ln \pi^\star_{\tau_k}\right](x_1) - \tau_k V_1^{\pi^\star_{\tau_k}}\left[P, -\ln \pi^k\right](x_1) \\
&= \underbrace{V_1^{\pi^\star_{\tau_k}}\left[P, g^k\right](x_1) - \widetilde{V}_1^k(x_1)}_{\diamondsuit} - \tau_k \sum_{h=1}^H \sum_{x \in \mathbf{X}} w_h^{\pi^\star_{\tau_k}}[P](x) \gamma_h^k(x) \, .
\end{aligned}$$

Using the definition of $\widetilde{Q}^k$ in Algorithm 1, we have

$$\widetilde{V}_h^k = \widetilde{V}_h^{k,0} + \sum_{n=1}^N \lambda^{k,n} \widetilde{V}_h^{k,n} = \pi_h^k\left(\widetilde{Q}_h^{k,0} - \tau_k \ln \pi_h^k + \sum_{n=1}^N \lambda^{k,n} \widetilde{Q}_h^{k,n}\right) = \pi_h^k\left(\widetilde{Q}_h^k - \tau_k \ln \pi_h^k\right) .$$

Using Lemma F.1, we have

$$\begin{aligned}
-\diamondsuit &= \widetilde{V}_1^k(x_1) - V_1^{\pi^\star_{\tau_k}}\left[P, g^k\right](x_1) \\
&= \sum_{h=1}^H \sum_{x,a \in \mathbf{X} \times \mathbf{A}} w_h^{\pi^\star_{\tau_k}}[P](x)\left(\pi^\star_{\tau_k, h}(a \mid x) - \pi_h^k(a \mid x)\right)\left(\widetilde{Q}_h^k(x,a) - \tau_k \ln \pi_h^k(x,a)\right) \\
&\quad + \underbrace{\sum_{h=1}^H \sum_{x,a \in \mathbf{X} \times \mathbf{A}} w_h^{\pi^\star_{\tau_k}}[P](x,a)\left(\widetilde{Q}_h^k(x,a) - \tau_k \ln \pi_h^k(x,a) - g_h^k(x,a) - \left(P_h \widetilde{V}_h^k\right)(x,a)\right)}_{\geq 0 \text{ due to Lemma H.2}} \\
&\geq \sum_{h=1}^H \sum_{x,a \in \mathbf{X} \times \mathbf{A}} w_h^{\pi^\star_{\tau_k}}[P](x)\left(\pi^\star_{\tau_k, h}(a \mid x) - \pi_h^k(a \mid x)\right)\left(\widetilde{Q}_h^k(x,a) - \tau_k \ln \pi_h^k(x,a)\right) .
\end{aligned} \tag{22}$$

Note that $\pi^{k+1}$ is the closed-form solution of the KL-regularized greedy policy (e.g., **Equation (5)** of Kozuno et al. (2019)):

$$\begin{aligned}
\pi_h^{k+1}(\cdot \mid x) &\propto \pi_h^k(\cdot \mid x) \exp\left(\eta_k\left(\widetilde{Q}_h^k - \tau_k \ln \pi_h^k\right)(x, \cdot)\right) \\
&= \arg\min_{\widetilde{\pi} \in \Delta_{\mathbf{A}}}\left\{\sum_{a \in \mathbf{A}} \widetilde{\pi}(a)\left(\left(-\widetilde{Q}_h^k + \tau_k \ln \pi_h^k\right)(x,a)\right) + \frac{1}{\eta_k} \mathrm{KL}[\widetilde{\pi}, \pi_h^k(\cdot \mid x)]\right\} .
\end{aligned} \tag{23}$$

Also, due to the definition of $\lambda^k$,

$$\left\|\widetilde{Q}^k\right\|_\infty \leq \underbrace{\left\|\widetilde{Q}^{k,0}\right\|_\infty}_{\leq H_{\mathrm{ent}}} + \underbrace{\sum_{n=1}^N \lambda^{k,n}}_{\leq H_{\mathrm{ent}}/b_{\mathrm{gap}}} \underbrace{\left\|\widetilde{Q}^{k,n}\right\|_\infty}_{\leq H} \leq H_{\mathrm{ent}}\left(1 + \frac{H}{b_{\mathrm{gap}}}\right) . \tag{24}$$

Using Lemma F.8 with Equation (23), we have

$$\begin{aligned}
&\big(\pi^\star_{\tau_k,h}(a\mid x)-\pi^k_h(a\mid x)\big)\Big(\widetilde{Q}^k_h-\tau_k\ln\pi^k_h\Big)(x,a)\\
\le&\frac{1}{\eta_k}\big(\gamma^k_h(x)-\gamma^{k+1}_h(x)\big)+2\eta_k\sum_{a\in\mathbf{A}}\pi^k_h(a\mid x)\Big(\widetilde{Q}^k_h(x,a)\Big)^2+2\eta_k\tau_k^2(1+\ln A)^2\\
\overset{(a)}{\le}&\frac{1}{\eta_k}\big(\gamma^k_h(x)-\gamma^{k+1}_h(x)\big)+\eta_kC_1+\eta_k\tau_k^2C_{2,k}\ ,
\end{aligned}\tag{25}$$

where (a) is due to Equation (24). By substituting Equation (22) and Equation (25) to $\diamondsuit$, we have

$$\clubsuit^k=\diamondsuit-\sum_{h=1}^H\sum_{x\in\mathbf{X}}w_h^{\pi^\star_{\tau_k}}[P](x)\gamma^k_h(x)\le\sum_{h=1}^H\sum_{x\in\mathbf{X}}w_h^{\pi^\star_{\tau_k}}[P](x)\frac{1}{\eta_k}\big((1-\eta_k\tau_k)\gamma^k_h(x)-\gamma^{k+1}_h(x)\big)+\eta_kC_1+\eta_k\tau_k^2C_{2,k}\ .$$

□

### H.3.2 $\heartsuit^k$ Bound

**Lemma H.8.** *Let* $\rho^k\coloneqq\widetilde{V}_1^{k,0}(x_1)-V_1^{\pi^k}\big[P,r^0,\tau_k\big](x_1)+\sum_{n=1}^N\lambda^{\star,n}_{\tau_k}\Big(\widetilde{V}_1^{k,n}(x_1)-V_1^{\pi^k}[P,r^n](x_1)\Big)$. *Let* $C_3\coloneqq\frac{N}{2}\Big(H+\frac{H_{\mathrm{ent}}}{b_{\mathrm{gap}}}\Big)^2$. *Then,*

$$\heartsuit^k\le\frac{1}{2\eta_k}\big((1-\eta_k\tau_k)\|\lambda^\star_{\tau_k}-\lambda^k\|_2^2-\|\lambda^\star_{\tau_k}-\lambda^k\|_2^2\big)+\rho^k+\frac{1}{2}\eta_kC_3\ .$$

*Proof.* Recall that

$$\heartsuit^k=\widetilde{V}_1^k(x_1)-\sum_{n=1}^N\lambda^{k,n}b^n-V_1^{\pi^k}\big[P,r^0,\tau_k\big](x_1)-\sum_{n=1}^N\lambda^{\star,n}_{\tau_k}\Big(V_1^{\pi^k}[P,r^n](x_1)-b^n\Big)+\frac{\tau_k}{2}\|\lambda^k\|_2^2-\frac{\tau_k}{2}\|\lambda^\star_{\tau_k}\|_2^2\ .$$

Note that

$$\begin{aligned}
&\widetilde{V}_1^k(x_1)-\sum_{n=1}^N\lambda^{k,n}b^n-V_1^{\pi^k}\big[P,r^0,\tau_k\big](x_1)-\sum_{n=1}^N\lambda^{\star,n}_{\tau_k}\Big(V_1^{\pi^k}[P,r^n](x_1)-b^n\Big)\\
=&\widetilde{V}_1^{k,0}(x_1)+\sum_{n=1}^N\lambda^{k,n}\Big(\widetilde{V}_1^{k,n}(x_1)-b^n\Big)-V_1^{\pi^k}\big[P,r^0,\tau_k\big](x_1)-\sum_{n=1}^N\lambda^{\star,n}_{\tau_k}\Big(V_1^{\pi^k}[P,r^n](x_1)-b^n\Big)\\
=&\underbrace{\widetilde{V}_1^{k,0}(x_1)-V_1^{\pi^k}\big[P,r^0,\tau_k\big](x_1)+\sum_{n=1}^N\lambda^{\star,n}_{\tau_k}\Big(\widetilde{V}_1^{k,n}(x_1)-V_1^{\pi^k}[P,r^n](x_1)\Big)}_{=\rho^k}+\sum_{n=1}^N\big(\lambda^{k,n}-\lambda^{\star,n}_{\tau_k}\big)\Big(\widetilde{V}_1^{k,n}(x_1)-b^n\Big)
\end{aligned}\tag{26}$$

Also, note that

$$\|\lambda^k\|_2^2-\|\lambda^\star_{\tau_k}\|_2^2=\sum_{n=1}^N(\lambda^{k,n})^2-(\lambda^{\star,n}_{\tau_k})^2=\sum_{n=1}^N2\lambda^{k,n}(\lambda^{k,n}-\lambda^{\star,n}_{\tau_k})-(\lambda^{k,n}-\lambda^{\star,n}_{\tau_k})^2\tag{27}$$

and

$$\left(\underbrace{\widetilde{V}_1^{k,n}(x_1)-b^n}_{\in[-H,H]}+\underbrace{\tau_k\lambda^{k,n}}_{\le H_{\mathrm{ent}}/b_{\mathrm{gap}}}\right)^2\le\left(H+\frac{H_{\mathrm{ent}}}{b_{\mathrm{gap}}}\right)^2\ .\tag{28}$$

Combining Equation (26), Equation (27), and Equation (28), we have

$$\begin{aligned}
\heartsuit^k \overset{(a)}{\leq} & \rho^k + \sum_{n=1}^{N} \left(\lambda^{k,n} - \lambda_{\tau_k}^{\star,n}\right)\left(\widetilde{V}_1^{k,n}(x_1) - b^n\right) + \sum_{n=1}^{N} \tau_k \lambda^{k,n}(\lambda^{k,n} - \lambda_{\tau_k}^{\star,n}) - \frac{\tau_k}{2}(\lambda^{k,n} - \lambda_{\tau_k}^{\star,n})^2 \\
= & \rho^k + \sum_{n=1}^{N} \left(\lambda^{k,n} - \lambda_{\tau_k}^{\star,n}\right)\left(\widetilde{V}_1^{k,n}(x_1) - b^n + \tau_k \lambda^{k,n}\right) - \frac{\tau_k}{2}(\lambda^{k,n} - \lambda_{\tau_k}^{\star,n})^2 \\
\overset{(b)}{\leq} & \rho^k + \sum_{n=1}^{N} \frac{1}{2\eta_k}\left((\lambda^{k,n} - \lambda_{\tau_k}^{\star,n})^2 - (\lambda^{k+1,n} - \lambda_{\tau_k}^{\star,n})^2\right) - \frac{\tau_k}{2}(\lambda^{k,n} - \lambda_{\tau_k}^{\star,n})^2 + \frac{\eta_k}{2}\left(H + \frac{H_{\text{ent}}}{b_{\text{gap}}}\right)^2 \\
\leq & \frac{1}{2\eta_k}\left((1-\eta_k\tau_k)\|\lambda_{\tau_k}^{\star} - \lambda^k\|_2^2 - \|\lambda_{\tau_k}^{\star} - \lambda^{k+1}\|_2^2\right) + \rho^k + \eta_k C_3
\end{aligned}$$

where (a) uses Equation (27) and (b) uses Lemma F.9 with the definition of $\lambda^{k+1}$ and Equation (28). □

By combining the bounds of $\clubsuit^k$ and $\heartsuit^k$, under the good event $\mathcal{F}_\delta^c$, we have

$$\begin{aligned}
0 \leq \clubsuit^k + \heartsuit^k \leq & \sum_{h=1}^{H}\sum_{x\in\mathbf{X}} w_h^{\pi_{\tau_k}^\star}[P](x)\left((1-\eta_k\tau_k)\gamma_h^k(x) - \gamma_h^{k+1}(x)\right) + \eta_k^2 C_1 + \eta_k^2\tau_k^2 C_{2,k} \\
& + \frac{1}{2}\left((1-\eta_k\tau_k)\|\lambda_{\tau_k}^{\star} - \lambda^k\|_2^2 - \|\lambda_{\tau_k}^{\star} - \lambda^{k+1}\|_2^2\right) + \eta_k\rho^k + \eta_k^2 C_3 \,.
\end{aligned} \tag{29}$$

### H.4 Optimaility Gap and Constraint Violation Analysis

Let $\Phi^k \coloneqq \sum_{h=1}^{H}\sum_{x\in\mathbf{X}} w_h^{\pi_{\tau_k}^\star}[P](x)\gamma_h^k(x) + \frac{1}{2}\|\lambda_{\tau_k}^{\star} - \lambda^k\|_2^2$ and $C \coloneqq \max_{k\in[N]}\{C_1 + \tau_k^2 C_{2,k} + C_3\}$. By rearranging Equation (29), we get

$$\begin{aligned}
\Phi^{k+1} & \leq (1-\eta_k\tau_k)\Phi^k + \eta_k^2 C + \eta_k\rho^k \\
& \leq (1-\eta_k\tau_k)(1-\eta_{k-1}\tau_{k-1})\Phi^{k-1} + \left((1-\eta_k\tau_k)\eta_{k-1}^2 + \eta_k^2\right)C + \left((1-\eta_k\tau_k)\eta_{k-1}\rho^{k-1} + \eta_k\rho^k\right) \\
& \leq \cdots \\
& \leq A_1^k\Phi^1 + B_k C + E_k \,,
\end{aligned}$$

where $A_t^k = \prod_{i=t}^{k}(1-\eta_i\tau_i)$, $B_k = \sum_{i=1}^{k} A_{i+1}^k\eta_i^2$, and $E_k = \sum_{i=1}^{k} A_{i+1}^k\eta_i\rho^i$. For $\Phi^{k+1}$, the following lemma holds.

**Lemma H.9.** *Set the learning rate and the regularization coefficient as $\eta_k = (k+3)^{-\alpha_\eta}$ and $\tau_k = (k+3)^{-\alpha_\tau}$. Set $\alpha_\tau$ and $\alpha_\eta$ such that $0 < \alpha_\tau < 0.5 < \alpha_\eta < 1$ and $\alpha_\eta + \alpha_\tau < 1$. Let $k^\star \coloneqq \left(\frac{24}{1-(\alpha_\eta+\alpha_\tau)}\ln\frac{12}{1-(\alpha_\eta+\alpha_\tau)}\right)^{\frac{1}{1-(\alpha_\eta+\alpha_\tau)}}$.*

*Assume that Assumption 2.1 and the good event $\mathcal{F}_\delta^c$ hold. Then, for any $\varepsilon > 0$, $\Phi^{k+1} \leq \varepsilon$ is satisfied for any $k \in \mathbb{N}$ except at most*

$$\widetilde{\mathcal{O}}\left(\left(b_{\text{gap}}^{-1}(1+N)X\sqrt{A}H^3\varepsilon^{-1}\right)^{\frac{1}{0.5-\alpha_\tau}}\right) + \widetilde{\mathcal{O}}\left(\left(b_{\text{gap}}^{-2}(1+N)H^4\varepsilon^{-1}\right)^{\frac{1}{\alpha_\eta-\alpha_\tau}}\right) + k^\star$$

*episodes.*

*Proof.* Using Lemma F.14, for $k \geq k^\star$, we have

$$B_k = \sum_{i=1}^{k} A_{i+1}^k\eta_i^2 = \sum_{i=1}^{k}\eta_i^2\prod_{j=i+1}^{k}(1-\eta_j\tau_j) = \sum_{i=1}^{k}(i+3)^{-2\alpha_\eta}\prod_{j=i+1}^{k}(1-(j+3)^{-\alpha_\eta-\alpha_\tau}) \leq 9\ln(k+3)(k+3)^{\alpha_\tau-\alpha_\eta} \,.$$

Note that $\frac{1}{\prod_{j=2}^{3}\left(1-j^{-\alpha_\eta-\alpha_\tau}\right)} \leq \frac{1}{(1-2^{-0.5})(1-3^{-0.5})} \leq 9$. For $A_1^k$, when $k \geq k^\star$, we have

$$A_1^k = \prod_{i=1}^{k}\left(1-(i+3)^{-\alpha_\eta-\alpha_\tau}\right) = \prod_{i=4}^{k+3}\left(1-i^{-\alpha_\eta-\alpha_\tau}\right) \leq 9\prod_{i=2}^{k+3}\left(1-i^{-\alpha_\eta-\alpha_\tau}\right) \leq 9\left(1-(k+3)^{-\alpha_\eta-\alpha_\tau}\right)^{k+3}$$
$$\overset{(a)}{\leq} 9\left(\exp\left(-(k+3)^{-(\alpha_\eta+\alpha_\tau)}\right)\right)^{k+3} = 9\exp\left(-(k+3)^{1-(\alpha_\eta+\alpha_\tau)}\right) \overset{(b)}{\leq} 9\exp(-12\ln(k+3)) = 9(k+3)^{-12}\,.$$

where (a) uses $1-x \leq \exp(-x)$ and (b) uses Lemma F.13.

Using Lemma F.15, for $k \geq k^\star$, we have

$$\max_{1\leq i\leq k}\eta_i A_{i+1}^k = \max_{1\leq i\leq k}\eta_i \prod_{j=i+1}^{k}(1-\eta_j\tau_j) = \max_{1\leq i\leq k}(i+3)^{-\alpha_\eta}\prod_{j=i+1}^{k}\left(1-(j+3)^{-\alpha_\eta-\alpha_\tau}\right) \leq 4(k+3)^{-\alpha_\eta}\,.$$

This indicates that

$$E_k = \sum_{i=1}^{k} A_{i+1}^k\eta_i\rho^i \leq 4(k+3)^{-\alpha_\eta}\sum_{i=1}^{k}\rho^i\,.$$

Therefore, $\Phi^{k+1}$ is bounded as

$$\Phi^{k+1} \leq A_1^k\Phi^1 + B_kC + E_k \leq \underbrace{9\Phi^1(k+3)^{-12}}_{\text{(i)}} + \underbrace{(9C\ln(k+3))(k+3)^{\alpha_\tau-\alpha_\eta}}_{\text{(ii)}} + \underbrace{4(k+3)^{-\alpha_\eta}\sum_{i=1}^{k}\rho^i}_{\text{(iii)}}\,.$$

(i) **bound.** Since $\pi^1$ is a uniform policy and $\lambda^1 = \mathbf{0}$, Lemma F.4 indicates that

$$\Phi^1 = \sum_{h=1}^{H}\sum_{x\in\mathbf{X}} w_h^{\pi_{\tau_k}^\star}[P](x)\underbrace{\mathrm{KL}[\pi_{\tau_1;h}^\star(\cdot\mid x),\pi_h^1(\cdot\mid x)]}_{\leq 2\ln A} + \frac{1}{2}\underbrace{\|\lambda_{\tau_1}^\star-\lambda^1\|_2^2}_{=\|\lambda_{\tau_1}^\star\|_2^2\leq NH_{\mathrm{ent}}^2/b_{\mathrm{gap}}^2} \leq 2H\ln A + \frac{NH_{\mathrm{ent}}^2}{2b_{\mathrm{gap}}^2}\,.$$

Therefore, for any $\varepsilon > 0$, (i) $= 9\Phi^1(k+3)^{-12} \leq \varepsilon$ is satisfied for any $k \in \mathbb{N}$ except at most

$$\widetilde{\mathcal{O}}\left(\left(\varepsilon^{-1}\left(H + b_{\mathrm{gap}}^{-2}NH^2\right)\right)^{\frac{1}{12}}\right) + k^\star$$

episodes.

(ii) **bound.** Recall that

$$C = (1+\ln A)^2 + H_{\mathrm{ent}}^2\left(1+\frac{H}{b_{\mathrm{gap}}}\right)^2 + \frac{N}{2}\left(H+\frac{H_{\mathrm{ent}}}{b_{\mathrm{gap}}}\right)^2\,.$$

Accordingly, we have

$$9C\ln(k+3) = \left(\frac{H^4}{b_{\mathrm{gap}}^2} + \frac{NH^2}{b_{\mathrm{gap}}^2}\right)\mathrm{polylog}(k)\,.$$

Lemma F.11 indicates that, for any $\varepsilon > 0$, (ii) $= (9C\ln(k+3))(k+3)^{\alpha_\tau-\alpha_\eta} \leq \varepsilon$ is satisfied for any $k \in \mathbb{N}$ except at most

$$\widetilde{\mathcal{O}}\left(\left(\varepsilon^{-1}\left(b_{\mathrm{gap}}^{-2}H^4 + b_{\mathrm{gap}}^{-2}NH^2\right)\right)^{\frac{1}{\alpha_\eta-\alpha_\tau}}\right) + k^\star$$

episodes.

(iii) **bound.** Lemma H.2 indicates that

$$0 \leq \sum_{i=1}^{k} \rho^i \leq \sum_{i=1}^{k} \left( \widetilde{V}_1^{i,0}(x_1) - V_{\tau_i;1}^{\pi^i}[P, r^0](x_1) \right) + \frac{H_{\mathrm{ent}}}{b_{\mathrm{gap}}} \sum_{n=1}^{N} \sum_{i=1}^{k} \left( \widetilde{V}_1^{i,n}(x_1) - V_1^{\pi^i}[P, r^n](x_1) \right) .$$

By applying Lemma F.2 to Lemma H.6, we have

$$\sum_{i=1}^{k} \left( \widetilde{V}_1^{i,0}(x_1) - V_{\tau_i;1}^{\pi^i}[P, r^0](x_1) \right) \leq \sqrt{kX^2AH^4}\, \mathrm{polylog}\left(k, X, A, H, \delta^{-1}\right)$$

$$\text{and} \quad \sum_{i=1}^{k} \left( \widetilde{V}_1^{i,n}(x_1) - V_1^{\pi^i}[P, r^n](x_1) \right) \leq \sqrt{kX^2AH^4}\, \mathrm{polylog}\left(k, X, A, H, \delta^{-1}\right) \quad \forall n \in [N] .$$

This indicates that

$$\sum_{i=1}^{k} \rho^i \leq \left( 1 + \frac{NH_{\mathrm{ent}}}{b_{\mathrm{gap}}} \right) \sqrt{kX^2AH^4}\, \mathrm{polylog}\left(k, X, A, H, \delta^{-1}\right) .$$

Lemma F.11 indicates that, for any $\varepsilon > 0$, (iii) $= 4(k+3)^{-\alpha_\eta} \sum_{i=1}^{k} \rho^i \leq \varepsilon$ is satisfied for any $k \in \mathbb{N}$ except at most

$$\widetilde{\mathcal{O}}\left( \left( \varepsilon^{-1} X \sqrt{A} \left( H^2 + b_{\mathrm{gap}}^{-1} NH^3 \right) \right)^{\frac{1}{\alpha_\eta - 0.5}} \right) + k^\star$$

episodes. By combining the above results, we have $\Phi^{k+1} \leq \varepsilon$ for all $k \in \mathbb{N}$ except at most

$$\begin{aligned}
&\widetilde{\mathcal{O}}\left( \left( \varepsilon^{-1} \left( H + b_{\mathrm{gap}}^{-2} NH^2 \right) \right)^{\frac{1}{12}} \right) + \widetilde{\mathcal{O}}\left( \left( \varepsilon^{-1} \left( b_{\mathrm{gap}}^{-2} H^4 + b_{\mathrm{gap}}^{-2} NH^2 \right) \right)^{\frac{1}{\alpha_\eta - \alpha_\tau}} \right) \\
&+ \widetilde{\mathcal{O}}\left( \left( \varepsilon^{-1} X \sqrt{A} \left( H^2 + b_{\mathrm{gap}}^{-1} NH^3 \right) \right)^{\frac{1}{\alpha_\eta - 0.5}} \right) + k^\star \\
=&\widetilde{\mathcal{O}}\left( \left( \varepsilon^{-1} X \sqrt{A} H^4 b_{\mathrm{gap}}^{-2} (1+N) \right)^{\frac{1}{\alpha_\eta - 0.5}} \right) + k^\star
\end{aligned}$$

episodes, where we used $\alpha_\tau < 0.5$ and $b_{\mathrm{gap}}^{-2} H^4 \geq H^2$ due to $b_{\mathrm{gap}} \leq H$. □

### H.5 Proof of Theorem 4.1

We are now ready to prove the main claim. Consider $\alpha_\tau$ and $\alpha_\eta$ satisfy conditions specified in Theorem 4.1. Suppose that the good event $\mathcal{F}_\delta^c$ holds.

Using $k^\star$ defined in Lemma H.9, for any $\varepsilon > 0$, we have $\Phi^k \leq \frac{\varepsilon^2}{H^3}$ for any $k \in \mathbb{N}$ except at most

$$\widetilde{\mathcal{O}}\left( \left( \left( \frac{\varepsilon^2}{H^3} \right)^{-1} X \sqrt{A} H^4 b_{\mathrm{gap}}^{-2} (1+N) \right)^{\frac{1}{\alpha_\eta - 0.5}} \right) + k^\star = \widetilde{\mathcal{O}}\left( \left( b_{\mathrm{gap}}^{-2} (1+N) X \sqrt{A} H^7 \varepsilon^{-2} \right)^{\frac{1}{\alpha_\eta - 0.5}} \right) + k^\star$$

episodes.

Also, when $k \geq \left( \frac{H_{\mathrm{ent}}}{\varepsilon \min\{b_{\mathrm{gap}}, 1\}} \right)^{\frac{1}{\alpha_\tau}}$, we have $\tau_k H_{\mathrm{ent}} \leq \varepsilon$ and $\frac{\tau_k H_{\mathrm{ent}}}{b_{\mathrm{gap}}} \leq \varepsilon$. Furthermore, it is easy to see that $\Phi^k \leq \frac{\varepsilon^2}{H^3}$ indicates $\sum_{h=1}^{H} \sum_{x \in \mathbf{X}} w_h^{\pi_{\tau_k}^\star}[P](x) \gamma_h^k(x) \leq \frac{\varepsilon^2}{H^3}$. Then, Lemma F.5 indicates that

$$V_1^{\pi^\star}[P, r](x_1) - V_1^{\pi^k}[P, r](x_1) \leq \varepsilon \quad \text{and} \quad b^n - V_1^{\pi^k}[P, r^n](x_1) \leq \varepsilon \quad \forall n \in [N]$$

hold for any $\varepsilon > 0$ and for any $k \in \mathbb{N}$ except at most

$$\widetilde{\mathcal{O}}\left( \left( b_{\mathrm{gap}}^{-2} (1+N) X \sqrt{A} H^7 \varepsilon^{-2} \right)^{\frac{1}{\alpha_\eta - 0.5}} \right) + \widetilde{\mathcal{O}}\left( \left( \varepsilon^{-1} b_{\mathrm{gap}}^{-1} H \right)^{\frac{1}{\alpha_\tau}} \right) + \left( \frac{24}{1 - (\alpha_\eta + \alpha_\tau)} \ln \frac{12}{1 - (\alpha_\eta + \alpha_\tau)} \right)^{\frac{1}{1 - (\alpha_\eta + \alpha_\tau)}}$$

episodes.

Finally, Lemma H.1 shows that the good event $\mathcal{F}_\delta^c$ holds with probability at least $1 - \delta$. This concludes the proof of Theorem 4.1.