Provably Efficient Exploration in Policy Optimization

Qi Cai [1] **Zhuoran Yang** [2] **Chi Jin** [3] **Zhaoran Wang** [1]

Abstract

While policy-based reinforcement learning (RL) achieves tremendous successes in practice, it is significantly less understood in theory, especially compared with value-based RL. In particular, it remains elusive how to design a provably efficient policy optimization algorithm that incorporates exploration. To bridge such a gap, this paper proposes an Optimistic variant of the Proximal Policy Optimization algorithm (OPPO), which follows an "optimistic version" of the policy gradient direction. This paper proves that, in the problem of episodic Markov decision process with linear function approximation, unknown transition, and adversarial reward with full-information feedback, OPPO achieves $\widetilde{O}(\sqrt{d^2 H^3 T})$ regret. Here d is the feature dimension, H is the episode horizon, and T is the total number of steps. To the best of our knowledge, OPPO is the first provably efficient policy optimization algorithm that explores.[1]

1. Introduction

Coupled with powerful function approximators such as neural networks, policy optimization plays a key role in the tremendous empirical successes of deep reinforcement learning (Silver et al., 2016; 2017; Duan et al., 2016; OpenAI, 2019; Wang et al., 2018). In sharp contrast, the theoretical understandings of policy optimization remain rather limited from both computational and statistical perspectives. More specifically, from the computational perspective, it remains unclear until recently whether policy optimization converges to the globally optimal policy in a finite number of iterations, even given infinite data. Meanwhile, from the statistical perspective, it still remains unclear how to attain the globally optimal policy with a finite regret or sample complexity.

A line of recent work (Fazel et al., 2018; Yang et al., 2019a; Abbasi-Yadkori et al., 2019a;b; Bhandari & Russo, 2019; Liu et al., 2019; Agarwal et al., 2019; Wang et al., 2019) answers the computational question affirmatively by proving that a wide variety of policy optimization algorithms, such as policy gradient (PG) (Williams, 1992; Baxter & Bartlett, 2000; Sutton et al., 2000), natural policy gradient (NPG) (Kakade, 2002), trust-region policy optimization (TRPO) (Schulman et al., 2015), proximal policy optimization (PPO) (Schulman et al., 2017), and actor-critic (AC) (Konda & Tsitsiklis, 2000), converge to the globally optimal policy at sublinear rates of convergence, even when they are coupled with neural networks (Liu et al., 2019; Wang et al., 2019). However, such computational efficiency guarantees rely on the regularity condition that the state space is already well explored. Such a condition is often implied by assuming either the access to a "simulator" (also known as the generative model) (Koenig & Simmons, 1993; Azar et al., 2011; 2012a;b; Sidford et al., 2018a;b; Wainwright, 2019) or finite concentratability coefficients (Munos & Szepesvári, 2008; Antos et al., 2008; Farahmand et al., 2010; Tosatto et al., 2017; Yang et al., 2019b; Chen & Jiang, 2019), both of which are often unavailable in practice.

In a more practical setting, the agent sequentially explores the state space, and meanwhile, exploits the information at hand by taking the actions that lead to higher expected total rewards. Such an exploration-exploitation tradeoff is better captured by the aforementioned statistical question regarding the regret or sample complexity, which remains even more challenging to answer than the computational question. As a result, such a lack of statistical understanding hinders the development of more sample-efficient policy optimization algorithms beyond heuristics. In fact, empirically, vanilla policy gradient is known to exhibit a possibly worse sample complexity than random search (Mania et al., 2018), even in basic settings such as linear-quadratic regulators. Meanwhile, theoretically, vanilla policy gradient can be shown to suffer from exponentially large variance in the well-known "combination lock" setting (Kakade, 2003;

[1]Department of Industrial Engineering and Management Sciences, Northwestern University [2]Department of Operations Research and Financial Engineering, Princeton University [3]Department of Electrical Engineering, Princeton University. Correspondence to: Qi Cai <qicai2022@u.northwestern.edu>, Zhuoran Yang <zy6@princeton.edu>, Chi Jin <chij@princeton.edu>, Zhaoran Wang <zhaoranwang@gmail.com>.

Proceedings of the 37^{th} International Conference on Machine Learning, Online, PMLR 119, 2020.

[1]Full version of this paper is available on `arXiv` via https://arxiv.org/pdf/1912.05830.pdf.

Leffler et al., 2007; Azar et al., 2012a), which only has a finite state space.

In this paper, we aim to answer the following fundamental question:

Can we design a policy optimization algorithm that incorporates exploration and is provably sample-efficient?

To answer this question, we propose the first policy optimization algorithm that incorporates exploration in a principled manner. In detail, we develop an Optimistic variant of the PPO algorithm, namely OPPO. Our algorithm is also closely related to NPG and TRPO. At each update, OPPO solves a Kullback-Leibler (KL)-regularized policy optimization subproblem, where the linear component of the objective function is defined using the action-value function. As is shown subsequently, solving such a subproblem corresponds to one iteration of infinite-dimensional mirror descent (Nemirovsky & Yudin, 1983) or dual averaging (Xiao, 2010), where the action-value function plays the role of the gradient. To encourage exploration, we explicitly incorporate a bonus function into the action-value function, which quantifies the uncertainty that arises from only observing finite historical data. Through uncertainty quantification, such a bonus function ensures the (conservative) optimism of the updated policy. Based on NPG, TRPO, and PPO, OPPO only augments the action-value function with the bonus function in an additive manner, which makes it easily implementable in practice.

Theoretically, we establish the sample efficiency of OPPO in an episodic setting of Markov decision processes (MDPs) with full-information feedback, where the transition dynamics are linear in features (Yang & Wang, 2019b;a; Jin et al., 2019; Ayoub et al., 2020; Zhou et al., 2020). In particular, we allow the transition dynamics to be nonstationary within each episode. See also the work of (Du et al., 2019a; Van Roy & Dong, 2019; Lattimore & Szepesvari, 2019) for a related discussion on the necessity of the linear representation. In detail, we prove that OPPO attains a $\sqrt{d^2H^3T}$-regret up to logarithmic factors, where d is the feature dimension, H is the episode horizon, and T is the total number of steps taken by the agent. Note that such a regret does not depend on the numbers of states and actions, and therefore, allows them to be even infinite. In particular, OPPO attains such a regret without knowing the transition dynamics or accessing a "simulator". Moreover, we prove that, even when the reward functions are adversarially chosen across the episodes, OPPO attains the same regret in terms of competing with the globally optimal policy in hindsight (Cesa-Bianchi & Lugosi, 2006; Bubeck & Cesa-Bianchi, 2012). In comparison, existing algorithms based on value iteration, e.g., optimistic least-squares value iteration (LSVI) (Jin et al., 2019), do not allow adversarially chosen reward functions. Such a notion of robustness partially justifies the empirical advantages of KL-regularized policy optimization (Neu et al., 2017; Geist et al., 2019). To the best of our knowledge, OPPO is the first provably sample-efficient policy optimization algorithm that incorporates exploration.

1.1. Related Work

Our work is based on the aforementioned line of recent work (Fazel et al., 2018; Yang et al., 2019a; Abbasi-Yadkori et al., 2019a;b; Bhandari & Russo, 2019; Liu et al., 2019; Agarwal et al., 2019; Wang et al., 2019) on the computational efficiency of policy optimization, which covers PG, NPG, TRPO, PPO, and AC. In particular, OPPO is based on PPO (and similarly, NPG and TRPO), which is shown to converge to the globally optimal policy at sublinear rates in tabular and linear settings, as well as nonlinear settings involving neural networks (Liu et al., 2019; Wang et al., 2019). However, without assuming the access to a "simulator" or finite concentratability coefficients, both of which imply that the state space is already well explored, it remains unclear whether any of such algorithms is sample-efficient, that is, attains a finite regret or sample complexity. In comparison, by incorporating uncertainty quantification into the action-value function at each update, which explicitly encourages exploration, OPPO not only attains the same computational efficiency as NPG, TRPO, and PPO, but is also shown to be sample-efficient with a $\sqrt{d^2H^3T}$-regret up to logarithmic factors.

Our work is closely related to another line of work (Even-Dar et al., 2009; Yu et al., 2009; Neu et al., 2010a;b; Zimin & Neu, 2013; Neu et al., 2012; Rosenberg & Mansour, 2019b;a) on online MDPs with adversarially chosen reward functions, which mostly focuses on the tabular setting.

- Assuming the transition dynamics are known and the full information of the reward functions is available, the work of (Even-Dar et al., 2009) establishes a $\sqrt{\tau^2 T \cdot \log|\mathcal{A}|}$-regret, where $\mathcal{A}$ is the action space, $|\mathcal{A}|$ is its cardinality, and τ upper bounds the mixing time of the MDP. See also the work of (Yu et al., 2009), which establishes a $T^{2/3}$-regret in a similar setting.
- Assuming the transition dynamics are known but only the bandit feedback of the received rewards is available, the work of (Neu et al., 2010a;b; Zimin & Neu, 2013) establishes an $H^2\sqrt{|\mathcal{A}|T}/\beta$-regret (Neu et al., 2010b), a $T^{2/3}$-regret (Neu et al., 2010a), and a $\sqrt{H|\mathcal{S}||\mathcal{A}|T}$-regret (Zimin & Neu, 2013), respectively, all up to logarithmic factors. Here $\mathcal{S}$ is the state space and $|\mathcal{S}|$ is its cardinality. In particular, it is assumed by (Neu et al., 2010b) that, with probability at least β, any state is reachable under any policy.
- Assuming the full information of the reward functions

is available but the transition dynamics are unknown, the work of (Neu et al., 2012; Rosenberg & Mansour, 2019b) establishes an $H|\mathcal{S}||\mathcal{A}|\sqrt{T}$-regret (Neu et al., 2012) and an $H|\mathcal{S}|\sqrt{|\mathcal{A}|T}$-regret (Rosenberg & Mansour, 2019b), respectively, both up to logarithmic factors.

- Assuming the transition dynamics are unknown and only the bandit feedback of the received rewards is available, the recent work of (Rosenberg & Mansour, 2019a) establishes an $H|\mathcal{S}|\sqrt{|\mathcal{A}|T}/\beta$-regret up to logarithmic factors. In particular, it is assumed by (Rosenberg & Mansour, 2019a) that, with probability at least β, any state is reachable under any policy. Without such an assumption, an $H^{3/2}|\mathcal{S}||\mathcal{A}|^{1/4}T^{3/4}$-regret is established.

In the latter two settings with unknown transition dynamics, all the existing algorithms (Neu et al., 2012; Rosenberg & Mansour, 2019b;a) follow the gradient direction with respect to the visitation measure, and thus, differ from most practical policy optimization algorithms. In comparison, OPPO is not restricted to the tabular setting and indeed follows the gradient direction with respect to the policy. OPPO is simply an optimistic variant of NPG, TRPO, and PPO, which makes it also a practical policy optimization algorithm. In particular, when specialized to the tabular setting, our setting corresponds to the third setting with $d = |\mathcal{S}|^2|\mathcal{A}|$, where OPPO attains an $H^{3/2}|\mathcal{S}|^2|\mathcal{A}|\sqrt{T}$-regret up to logarithmic factors.

Broadly speaking, our work is related to a vast body of work on value-based reinforcement learning in tabular (Jaksch et al., 2010; Osband et al., 2014; Osband & Van Roy, 2016; Azar et al., 2017; Dann et al., 2017; Strehl et al., 2006; Jin et al., 2018) and linear settings (Yang & Wang, 2019b;a; Jin et al., 2019; Ayoub et al., 2020; Zhou et al., 2020), as well as nonlinear settings involving general function approximators (Wen & Van Roy, 2017; Jiang et al., 2017; Du et al., 2019b; Dong et al., 2019). In particular, our setting is the same as the linear setting studied by (Ayoub et al., 2020; Zhou et al., 2020), which generalizes the one proposed by (Yang & Wang, 2019a). We remark that our setting differs from the linear setting studied by (Yang & Wang, 2019b; Jin et al., 2019). It can be shown that the two settings are incomparable in the sense that one does not imply the other (Zhou et al., 2020). Also, our setting is related to the low-Bellman-rank setting studied by (Jiang et al., 2017; Dong et al., 2019). In comparison, we focus on policy-based reinforcement learning, which is significantly less studied in theory. In particular, compared with the work of (Yang & Wang, 2019b;a; Jin et al., 2019; Ayoub et al., 2020; Zhou et al., 2020), which focuses on value-based reinforcement learning, OPPO attains the same $\sqrt{T}$-regret even in the presence of adversarially chosen reward functions. Compared with optimism-led iterative value-function elimination (OLIVE) (Jiang et al., 2017; Dong et al., 2019), which handles the more general low-Bellman-rank setting but is only sample-efficient, OPPO simultaneously attains computational efficiency and sample efficiency in the linear setting. Despite the differences between policy-based and value-based reinforcement learning, our work shows that the general principle of "optimism in the face of uncertainty" (Auer et al., 2002; Bubeck & Cesa-Bianchi, 2012) can be carried over from existing algorithms based on value iteration, e.g., optimistic LSVI, into policy optimization algorithms, e.g., NPG, TRPO, and PPO, to make them sample-efficient, which further leads to a new general principle of "conservative optimism in the face of uncertainty and adversary" that additionally allows adversarially chosen reward functions.

1.2. Notation

We denote by $\|\cdot\|_2$ the ℓ_2-norm of a vector or the spectral norm of a matrix. We denote by $\Delta(\mathcal{A})$ the set of probability distributions on a set $\mathcal{A}$ and correspondingly define

$$\Delta(\mathcal{A}\,|\,\mathcal{S},H) = \big\{\{\pi_h(\cdot\,|\,\cdot)\}_{h=1}^H : \pi_h(\cdot\,|\,x)\in\Delta(\mathcal{A}) \text{ for any } x\in\mathcal{S} \text{ and } h\in[H]\big\}$$

for any set $\mathcal{S}$ and $H\in\mathbb{Z}_+$. For $p_1, p_2 \in \Delta(\mathcal{A})$, we denote by $D_{\mathrm{KL}}(p_1\,\|\,p_2)$ the KL-divergence,

$$D_{\mathrm{KL}}(p_1\,\|\,p_2) = \sum_{a\in\mathcal{A}} p_1(a)\log\frac{p_1(a)}{p_2(a)}.$$

Throughout this paper, we denote by $C, C', C'', \ldots$ absolute constants whose values can vary from line by line.

2. Preliminaries

2.1. MDPs with Adversarial Rewards

In this paper, we consider an episodic MDP $(\mathcal{S},\mathcal{A},H,\mathcal{P},r)$, where $\mathcal{S}$ and $\mathcal{A}$ are the state and action spaces, respectively, H is the length of each episode, $\mathcal{P}_h(\cdot\,|\,\cdot,\cdot)$ is the transition kernel from a state-action pair to the next state at the h-th step of each episode, and $r_h^k : \mathcal{S}\times\mathcal{A}\to[0,1]$ is the reward function at the h-th step of the k-th episode. We assume that the reward function is deterministic, which is without loss of generality, as our subsequent regret analysis readily generalizes to the setting where the reward function is stochastic.

At the beginning of the k-th episode, the agent determines a policy $\pi^k = \{\pi_h^k\}_{h=1}^H \in \Delta(\mathcal{A}\,|\,\mathcal{S},H)$. We assume that the initial state x_1^k is fixed to $x_1\in\mathcal{S}$ across all the episodes, which is without loss of generality, as our subsequent regret analysis readily generalizes to the setting where x_1^k is sampled from a fixed distribution across all the episodes. Then the agent iteratively interacts with the environment as follows. At the h-th step, the agent receives a state x_h^k and takes an action following $a_h^k \sim \pi_h^k(\cdot \,|\, x_h^k)$. Subsequently, the agent receives a reward $r_h^k(x_h^k, a_h^k)$ and the next state following $x_{h+1}^k \sim \mathcal{P}_h(\cdot \,|\, x_h^k, a_h^k)$. The k-th episode ends after the agent receives the last reward $r_H^k(x_H^k, a_H^k)$.

We allow the reward function $r^k = \{r_h^k\}_{h=1}^H$ to be adversarially chosen by the environment at the beginning of the k-th episode, which can depend on the $(k-1)$ historical trajectories. The reward function r_h^k is revealed to the agent after it takes the action a_h^k at the state x_h^k, which together determine the received reward $r_h^k(x_h^k, a_h^k)$. We define the regret in terms of competing with the globally optimal policy in hindsight (Cesa-Bianchi & Lugosi, 2006; Bubeck & Cesa-Bianchi, 2012) as

$$\text{Regret}(T) = \max_{\pi \in \Delta(\mathcal{A} \,|\, \mathcal{S}, H)} \sum_{k=1}^K \big(V_1^{\pi,k}(x_1^k) - V_1^{\pi^k,k}(x_1^k)\big), \tag{2.1}$$

where $T = HK$ is the total number of steps taken by the agent in all the K episodes. For any policy $\pi = \{\pi_h\}_{h=1}^H \in \Delta(\mathcal{A} \,|\, \mathcal{S}, H)$, the value function $V_h^{\pi,k}: \mathcal{S} \to \mathbb{R}$ associated with the reward function $r^k = \{r_h^k\}_{h=1}^H$ is defined by

$$V_h^{\pi,k}(x) = \mathbb{E}_\pi\Big[\sum_{i=h}^H r_i^k(x_i, a_i) \,\Big|\, x_h = x\Big]. \tag{2.2}$$

Here we denote by $\mathbb{E}_\pi[\cdot]$ the expectation with respect to the randomness of the state-action sequence $\{(x_h, a_h)\}_{h=1}^H$, where the action a_h follows the policy $\pi_h(\cdot \,|\, x_h)$ at the state x_h and the next state x_{h+1} follows the transition dynamics $\mathcal{P}_h(\cdot \,|\, x_h, a_h)$. Correspondingly, we define the action-value function (also known as the Q-function) $Q_h^{\pi,k}: \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ as

$$Q_h^{\pi,k}(x, a) = \mathbb{E}_\pi\Big[\sum_{i=h}^H r_i^k(x_i, a_i) \,\Big|\, x_h = x,\, a_h = a\Big]. \tag{2.3}$$

By the definitions in (2.2) and (2.3), we have the following Bellman equation,

$$V_h^{\pi,k} = \langle Q_h^{\pi,k}, \pi_h \rangle_{\mathcal{A}}, \quad Q_h^{\pi,k} = r_h^k + \mathbb{P}_h V_{h+1}^{\pi,k}. \tag{2.4}$$

Here $\langle \cdot, \cdot \rangle_{\mathcal{A}}$ denotes the inner product over $\mathcal{A}$, where the subscript is omitted subsequently if it is clear from the context. Also, $\mathbb{P}_h$ is the operator form of the transition kernel $\mathcal{P}_h(\cdot \,|\, \cdot, \cdot)$, which is defined by

$$(\mathbb{P}_h f)(x, a) = \mathbb{E}[f(x') \,|\, x' \sim \mathcal{P}_h(\cdot \,|\, x, a)] \tag{2.5}$$

for any function $f: \mathcal{S} \to \mathbb{R}$. By allowing the reward function to be adversarially chosen in each episode, our setting generalizes the stationary setting commonly adopted by the existing work on value-based reinforcement learning (Jaksch et al., 2010; Osband et al., 2014; Osband & Van Roy, 2016; Azar et al., 2017; Dann et al., 2017; Strehl et al., 2006; Jin et al., 2018; 2019; Yang & Wang, 2019b;a), where the reward function is fixed across all the episodes.

2.2. Linear Function Approximations

We consider the linear setting where the transition dynamics are linear in a feature map, which is formalized in the following assumption.

Assumption 2.1 (Linear MDP)**.** We assume that the MDP $(\mathcal{S}, \mathcal{A}, H, \mathcal{P}, r)$ is a linear MDP with the known feature map $\psi: \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}^d$, that is, for any $h \in [H]$, there exists $\theta_h \in \mathbb{R}^d$ with $\|\theta_h\|_2 \le \sqrt{d}$ such that

$$\mathcal{P}_h(x' \,|\, x, a) = \psi(x, a, x')^\top \theta_h$$

for any $(x, a, x') \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}$. Also, we assume that

$$\Big\| \int_{\mathcal{S}} \psi(x, a, x') \cdot V(x') \mathrm{d}x' \Big\|_2 \le \sqrt{d} H$$

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

See (Ayoub et al., 2020; Zhou et al., 2020) for various examples of linear MDPs, including the one proposed by (Yang & Wang, 2019a). In particular, a tabular MDP corresponds to the linear MDP with $d = |\mathcal{S}|^2|\mathcal{A}|$ and the feature vector $\psi(x, a, x')$ being the canonical basis $e_{(x,a,x')}$ of $\mathbb{R}^{|\mathcal{S}|^2|\mathcal{A}|}$. See also (Du et al., 2019a; Van Roy & Dong, 2019; Lattimore & Szepesvari, 2019) for a related discussion on the necessity of the linear representation.

We remark that (Yang & Wang, 2019b; Jin et al., 2019) study another variant of linear MDPs, where the transition kernel can be written as $\mathcal{P}_h(x' \,|\, x, a) = \varphi(x, a)^\top \mu_h(x')$ for any $h \in [H]$ and $(x, a, x') \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}$. Here $\varphi: \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$ is a known feature map and $\mu_h: \mathcal{S} \to \mathbb{R}^d$ is an unknown function on $\mathcal{S}$ for any $h \in [H]$. Although the variant of linear MDPs defined in Assumption 2.1 and the one studied by (Yang & Wang, 2019b; Jin et al., 2019) both cover the tabular setting and the one proposed by (Yang & Wang, 2019a) as special cases, they are two different definitions of linear MDPs as their feature maps $\psi(\cdot, \cdot, \cdot)$ and $\varphi(\cdot, \cdot)$ are defined on different domains. It can be shown that the two variants are incomparable in the sense that one does not imply the other (Zhou et al., 2020).

3. Algorithm and Theory

3.1. Optimistic PPO (OPPO)

We present Optimistic PPO (OPPO) in Algorithm 1, which involves a policy improvement step and a policy evaluation step.

Policy Improvement Step. In the k-th episode, OPPO updates π^k based on π^{k-1} (Lines 4-9 of Algorithm 1). In detail, we define the following linear function of the policy $\pi \in \Delta(\mathcal{A} \,|\, \mathcal{S}, H)$,

$$
\begin{aligned}
&L_{k-1}(\pi) \\
&\quad = V_1^{\pi^{k-1},k-1}(x_1^k) + \mathbb{E}_{\pi^{k-1}}\Big[\sum_{h=1}^{H} \langle Q_h^{\pi^{k-1},k-1}(x_h, \cdot), \\
&\qquad\qquad \pi_h(\cdot \,|\, x_h) - \pi_h^{k-1}(\cdot \,|\, x_h)\rangle \,\Big|\, x_1 = x_1^k\Big], \quad (3.1)
\end{aligned}
$$

which is a local linear approximation of $V_1^{\pi,k-1}(x_1^k)$ at π^{k-1} (Schulman et al., 2015; 2017). In particular, we have that $L_{k-1}(\pi^{k-1}) = V_1^{\pi^{k-1},k-1}(x_1^k)$. The policy improvement step is defined by

$$
\begin{aligned}
\pi^k \leftarrow \mathop{\mathrm{argmax}}_{\pi \in \Delta(\mathcal{A} \,|\, \mathcal{S}, H)} \ & L_{k-1}(\pi) \qquad (3.2)\\
& - \alpha^{-1} \cdot \mathbb{E}_{\pi^{k-1}}\big[\widetilde{D}_{\mathrm{KL}}(\pi \,\|\, \pi^{k-1}) \,|\, x_1 = x_1^k\big],
\end{aligned}
$$

where $\widetilde{D}_{\mathrm{KL}}(\pi \,\|\, \pi^{k-1}) = \sum_{h=1}^{H} D_{\mathrm{KL}}\big(\pi_h(\cdot \,|\, x_h) \,\big\|\, \pi_h^{k-1}(\cdot \,|\, x_h)\big)$.

Here the KL-divergence regularizes π to be close to π^{k-1} so that $L_{k-1}(\pi)$ well approximates $V_1^{\pi,k-1}(x_1^k)$, which further ensures that the updated policy π^k improves the expected total reward (associated with the reward function r^{k-1}) upon π^{k-1}. Also, $\alpha > 0$ is the stepsize, which is specified in Theorem 3.1. By executing the updated policy π^k, the agent receives the state-action sequence $\{(x_h^k, a_h^k)\}_{h=1}^H$ and observes the reward function r^k, which together determine the received rewards $\{r_h^k(x_h^k, a_h^k)\}_{h=1}^H$.

The policy improvement step defined in (3.2) corresponds to one iteration of NPG (Kakade, 2002), TRPO (Schulman et al., 2015), and PPO (Schulman et al., 2017). In particular, PPO solves the same KL-regularized policy optimization subproblem as in (3.2) at each iteration, while TRPO solves an equivalent KL-constrained subproblem. In the special case where the reward function r_h^{k-1} is linear in the feature map ϕ_h^{k-1} defined subsequently, which implies that the Q-function $Q_h^{\pi^{k-1},k-1}$ is also linear in ϕ_h^{k-1}, the updated policy π^k can be equivalently obtained by one iteration of NPG when the policy is parameterized by an energy-based distribution (Agarwal et al., 2019; Wang et al., 2019). Such a policy improvement step can also be cast as one iteration of infinite-dimensional mirror descent (Nemirovsky & Yudin, 1983) or dual averaging (Xiao, 2010), where the Q-function plays the role of the gradient (Liu et al., 2019; Wang et al., 2019).

The updated policy π^k obtained in (3.2) takes the following closed form,

$$
\pi_h^k(\cdot \,|\, x) \propto \pi_h^{k-1}(\cdot \,|\, x) \cdot \exp\big(\alpha \cdot Q_h^{\pi^{k-1},k-1}(x, \cdot)\big) \quad (3.3)
$$

for any $h \in [H]$ and $x \in \mathcal{S}$. However, the Q-function $Q_h^{\pi^{k-1},k-1}$ remains to be estimated through the subsequent policy evaluation step. We denote by Q_h^{k-1} the estimated Q-function, which replaces the Q-function $Q_h^{\pi^{k-1},k-1}$ in (3.1)-(3.3) and is correspondingly used in Line 6 of Algorithm 1.

Algorithm 1 Optimistic PPO (OPPO)

1: Initialize $\{\pi_h^0(\cdot \,|\, \cdot)\}_{h=1}^H$ as uniform distributions on $\mathcal{A}$ and $\{Q_h^0(\cdot, \cdot)\}_{h=1}^H$ as zero functions.
2: **For** episode $k = 1, 2, \ldots, K$ **do**
3: Receive the initial state x_1^k.
4: **For** step $h = 1, 2, \ldots, H$ **do**
5: Update the policy by
6: $\pi_h^k(\cdot \,|\, \cdot) \propto \pi_h^{k-1}(\cdot \,|\, \cdot) \cdot \exp\{\alpha \cdot Q_h^{k-1}(\cdot, \cdot)\}$.
7: Take the action following $a_h^k \sim \pi_h^k(\cdot \,|\, x_h^k)$.
8: Observe the reward function $r_h^k(\cdot, \cdot)$.
9: Receive the next state x_{h+1}^k.
10: Initialize $V_{H+1}^k(\cdot)$ as a zero function.
11: **For** step $h = H, H-1, \ldots, 1$ **do**
12: $\Lambda_h^k \leftarrow \sum_{\tau=1}^{k-1} \phi_h^\tau(x_h^\tau, a_h^\tau)\phi_h^\tau(x_h^\tau, a_h^\tau)^\top + \lambda \cdot I$.
13: $w_h^k \leftarrow (\Lambda_h^k)^{-1} \sum_{\tau=1}^{k-1} \phi_h^\tau(x_h^\tau, a_h^\tau) \cdot V_{h+1}^\tau(x_{h+1}^\tau)$.
14: $\phi_h^k(\cdot, \cdot) \leftarrow \int_{\mathcal{S}} \psi(\cdot, \cdot, x') \cdot V_{h+1}^k(x')\mathrm{d}x'$.
15: $\Gamma_h^k(\cdot, \cdot) \leftarrow \beta \cdot [\phi_h^k(\cdot, \cdot)^\top (\Lambda_h^k)^{-1} \phi_h^k(\cdot, \cdot)]^{1/2}$.
16: $\overline{Q}_h^k(\cdot, \cdot) \leftarrow r_h^k(\cdot, \cdot) + \phi_h^k(\cdot, \cdot)^\top w_h^k + \Gamma_h^k(\cdot, \cdot)$.
17: $Q_h^k(\cdot, \cdot) \leftarrow \min\{\overline{Q}_h^k(\cdot, \cdot), H - h + 1\}^+$.
18: $V_h^k(\cdot) \leftarrow \langle Q_h^k(\cdot, \cdot), \pi_h^k(\cdot \,|\, \cdot)\rangle_{\mathcal{A}}$.

Policy Evaluation Step. At the end of the k-th episode, OPPO evaluates the policy π^k based on the $(k-1)$ historical trajectories (Lines 11-18 of Algorithm 1). In detail, for any $h \in [H]$, we define the empirical mean-squared Bellman error (MSBE) (Sutton & Barto, 2018) as

$$
M_h^k(w) = \sum_{\tau=1}^{k-1} \big(V_{h+1}^\tau(x_{h+1}^\tau) - \phi_h^\tau(x_h^\tau, a_h^\tau)^\top w\big)^2, \quad (3.4)
$$

$$
\begin{aligned}
\text{where } \ \phi_h^\tau(\cdot, \cdot) &= \int_{\mathcal{S}} \psi(\cdot, \cdot, x') \cdot V_{h+1}^\tau(x')\mathrm{d}x', \\
V_{h+1}^\tau(\cdot) &= \langle Q_{h+1}^\tau(\cdot, \cdot), \pi_{h+1}^\tau(\cdot \,|\, \cdot)\rangle_{\mathcal{A}},
\end{aligned}
$$

while we initialize V_{H+1}^τ as a zero function on $\mathcal{S}$. The policy evaluation step is defined by iteratively updating the estimated Q-function $Q^k = \{Q_h^k\}_{h=1}^H$ associated with the reward function $r^k = \{r_h^k\}_{h=1}^H$ by

$$
\begin{aligned}
& w_h^k \leftarrow \mathop{\mathrm{argmin}}_{w \in \mathbb{R}^d} M_h^k(w) + \lambda \cdot \|w\|_2^2, \\
& Q_h^k(\cdot, \cdot) \leftarrow \min\{r_h^k(\cdot, \cdot) + \phi_h^k(\cdot, \cdot)^\top w_h^k + \Gamma_h^k(\cdot, \cdot), H - h + 1\}^+ \qquad (3.5)
\end{aligned}
$$

in the order of $h = H, H-1, \ldots, 1$. Here $\lambda > 0$ is the regularization parameter, which is specified in Theorem 3.1. Also, $\Gamma_h^k : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^+$ is a bonus function, which quantifies the uncertainty in estimating the Q-function $Q_h^{\pi^k,k}$

based on only finite historical data. In particular, the weight vector w_h^k obtained in (3.5) and the bonus function Γ_h^k take the following closed forms,

$$w_h^k = (\Lambda_h^k)^{-1}\Big(\sum_{\tau=1}^{k-1} \phi_h^\tau(x_h^\tau, a_h^\tau) \cdot V_{h+1}^\tau(x_{h+1}^\tau)\Big),$$
$$\Gamma_h^k(\cdot,\cdot) = \beta \cdot \big(\phi_h^k(\cdot,\cdot)^\top (\Lambda_h^k)^{-1} \phi_h^k(\cdot,\cdot)\big)^{1/2}, \qquad (3.6)$$
$$\text{where } \Lambda_h^k = \sum_{\tau=1}^{k-1} \phi_h^\tau(x_h^\tau, a_h^\tau)\phi_h^\tau(x_h^\tau, a_h^\tau)^\top + \lambda \cdot I.$$

Here $\beta > 0$ scales with d, H, and K, which is specified in Theorem 3.1.

The policy evaluation step defined in (3.5) corresponds to one iteration of least-squares temporal difference (LSTD) (Bradtke & Barto, 1996; Boyan, 2002). In particular, as we have

$$\mathbb{E}[V_{h+1}^\tau(x') \,|\, x' \sim \mathcal{P}_h(\cdot \,|\, x, a)] = (\mathbb{P}_h V_{h+1}^\tau)(x, a)$$

for any $\tau \in [k-1]$ and $(x, a) \in \mathcal{S} \times \mathcal{A}$ in the empirical MSBE defined in (3.4), $\phi_h^{k\top} w_h^k$ in (3.5) is an estimator of $\mathbb{P}_h V_{h+1}^k$ in the Bellman equation defined in (2.4) (with $V_{h+1}^{\pi^k,k}$ replaced by V_{h+1}^k). Meanwhile, we construct the bonus function Γ_h^k according to (3.6) so that $\phi_h^{k\top} w_h^k + \Gamma_h^k$ is an upper confidence bound (UCB), that is, it holds that

$$\phi_h^k(\cdot,\cdot)^\top w_h^k + \Gamma_h^k(\cdot,\cdot) \geq (\mathbb{P}_h V_{h+1}^k)(\cdot,\cdot)$$

with high probability, which is subsequently characterized in Lemma 4.3. Here the inequality holds uniformly for any $(h, k) \in [H] \times [K]$ and $(x, a) \in \mathcal{S} \times \mathcal{A}$. As the fact that $r_h^k \in [0, 1]$ for any $h \in [H]$ implies that $Q_h^{\pi^k,k} \in [0, H - h + 1]$, we truncate Q_h^k to the range $[0, H - h + 1]$ in (3.5), which is correspondingly used in Line 17 of Algorithm 1.

3.2. Regret Analysis

We establish an upper bound of the regret of OPPO (Algorithm 1) in the following theorem. Recall that the regret is defined in (2.1) and $T = HK$ is the total number of steps taken by the agent, where H is the length of each episode and K is the total number of episodes. Also, $|\mathcal{A}|$ is the cardinality of $\mathcal{A}$ and d is the dimension of the feature map ψ.

Theorem 3.1 (Total Regret)**.** Let $\alpha = \sqrt{2 \log |\mathcal{A}|/(HT)}$ in (3.2) and Line 6 of Algorithm 1, $\lambda = 1$ in (3.5) and Line 12 of Algorithm 1, and $\beta = C\sqrt{dH^2 \cdot \log(dT/\zeta)}$ in (3.6) and Line 15 of Algorithm 1, where $C > 1$ is an absolute constant and $\zeta \in (0, 1]$. Under Assumption 2.1 and the assumption that $\log |\mathcal{A}| = O(d^2 \cdot [\log(dT/\zeta)]^2)$, the regret of OPPO satisfies

$$\text{Regret}(T) \leq C'\sqrt{d^2 H^3 T} \cdot \log(dT/\zeta)$$

with probability at least $1 - \zeta$, where $C' > 0$ is an absolute constant.

Proof. See Section 4 for a proof sketch and Appendix C for a detailed proof. □

Theorem 3.1 proves that OPPO attains a $\sqrt{d^2H^3T}$-regret up to logarithmic factors, where the dependency on the total number of steps T is optimal. In the stationary setting where the reward function and initial state are fixed across all the episodes, such a regret translates to a d^2H^4/ε^2-sample complexity (up to logarithmic factors) following the argument of (Jin et al., 2018) (Section 3.1). Here $\varepsilon > 0$ measures the suboptimality of the obtained policy π^k in the following sense,

$$\max_{\pi \in \Delta(\mathcal{A} \,|\, \mathcal{S}, H)} V_1^\pi(x_1) - V_1^{\pi^k}(x_1) \leq \varepsilon,$$

where k is sampled from $[K]$ uniformly at random. Here we denote the value function by $V_1^\pi = V_1^{\pi,k}$ and the initial state by $x_1 = x_1^k$ for any $k \in [K]$, as the reward function and initial state are fixed across all the episodes. Moreover, compared with the work of (Yang & Wang, 2019b;a; Jin et al., 2019; Ayoub et al., 2020; Zhou et al., 2020), OPPO additionally allows adversarially chosen reward functions without exacerbating the regret, which leads to a notion of robustness. Also, as a tabular MDP satisfies Assumption 2.1 with $d = |\mathcal{S}|^2|\mathcal{A}|$ and ψ being the canonical basis of $\mathbb{R}^d$, Theorem 3.1 yields an $|\mathcal{S}|^2|\mathcal{A}|\sqrt{H^3T}$-regret in the tabular setting. Our subsequent discussion intuitively explains how OPPO achieves such a notion of robustness while attaining the $\sqrt{d^2H^3T}$-regret (up to logarithmic factors).

Discussion of Mechanisms. In the sequel, we consider the ideal setting where the transition dynamics are known, which, by the Bellman equation defined in (2.4), allows us to access the Q-function $Q_h^{\pi,k}$ for any policy π and $(h, k) \in [H] \times [K]$ once given the reward function r^k. The following lemma connects the difference between two policies to the difference between their expected total rewards through the Q-function.

Lemma 3.2 (Performance Difference)**.** For any policies $\pi, \pi' \in \Delta(\mathcal{A} \,|\, \mathcal{S}, H)$ and $k \in [K]$, it holds that

$$V_1^{\pi',k}(x_1^k) - V_1^{\pi,k}(x_1^k) \qquad (3.7)$$
$$= \mathbb{E}_{\pi'}\Big[\sum_{h=1}^H \langle Q_h^{\pi,k}(x_h, \cdot), \pi'_h(\cdot \,|\, x_h) - \pi_h(\cdot \,|\, x_h)\rangle \,\Big|\, x_1 = x_1^k\Big].$$

Proof. See Appendix A.1 for a detailed proof. □

For notational simplicity, we omit the conditioning on $x_1 = x_1^k$, e.g., in (3.7) of Lemma 3.2, subsequently. The following lemma characterizes the policy improvement step defined in (3.2), where the updated policy π^k takes the closed form in (3.3).

Lemma 3.3 (One-Step Descent)**.** For any distributions $p^*, p \in \Delta(\mathcal{A})$, state $x \in \mathcal{S}$, and function $Q : \mathcal{S} \times \mathcal{A} \to [0, H]$, it holds for $p' \in \Delta(\mathcal{A})$ with $p'(\cdot) \propto p(\cdot) \cdot \exp\{\alpha \cdot Q(x, \cdot)\}$ that

$$\langle Q(x,\cdot), p^*(\cdot) - p(\cdot)\rangle \le \alpha H^2/2 + \alpha^{-1} \cdot \Big(D_{\mathrm{KL}}\big(p^*(\cdot) \,\|\, p(\cdot)\big) - D_{\mathrm{KL}}\big(p^*(\cdot) \,\|\, p'(\cdot)\big)\Big).$$

Proof. See Appendix A.2 for a detailed proof. □

Corresponding to the definition of the regret in (2.1), we define the globally optimal policy in hindsight (Cesa-Bianchi & Lugosi, 2006; Bubeck & Cesa-Bianchi, 2012) as

$$\pi^* = \operatorname*{argmax}_{\pi \in \Delta(\mathcal{A} \,|\, \mathcal{S}, H)} \sum_{k=1}^{K} V_1^{\pi,k}(x_1^k), \tag{3.8}$$

which attains a zero-regret. In the ideal setting where the Q-function $Q_h^{\pi^k,k}$ associated with the reward function r^k is known and the updated policy π_h^{k+1} takes the closed form in (3.3), Lemma 3.3 implies

$$\begin{aligned} &\langle Q_h^{\pi^k,k}(x,\cdot), \pi_h^*(\cdot \,|\, x) - \pi_h^k(\cdot \,|\, x)\rangle \\ &\quad \le \alpha H^2/2 + \alpha^{-1} \cdot \Big(D_{\mathrm{KL}}\big(\pi_h^*(\cdot \,|\, x) \,\|\, \pi_h^k(\cdot \,|\, x)\big) \\ &\qquad - D_{\mathrm{KL}}\big(\pi_h^*(\cdot \,|\, x) \,\|\, \pi_h^{k+1}(\cdot \,|\, x)\big)\Big) \end{aligned} \tag{3.9}$$

for any $(h, k) \in [H] \times [K]$ and $x \in \mathcal{S}$. Combining (3.9) with Lemma 3.2, we obtain

$$\begin{aligned} &\mathrm{Regret}(T) \\ &= \sum_{k=1}^{K} \big(V_1^{\pi^*,k}(x_1^k) - V_1^{\pi^k,k}(x_1^k)\big) \\ &= \mathbb{E}_{\pi^*}\Big[\sum_{k=1}^{K}\sum_{h=1}^{H} \langle Q_h^{\pi^k,k}(x_h,\cdot), \pi_h^*(\cdot \,|\, x_h) - \pi_h^k(\cdot \,|\, x_h)\rangle\Big] \\ &\le \alpha H^3 K/2 + \alpha^{-1} \cdot \sum_{h=1}^{H} \mathbb{E}_{\pi^*}\big[D_{\mathrm{KL}}\big(\pi_h^*(\cdot \,|\, x_h) \,\|\, \pi_h^1(\cdot \,|\, x_h)\big)\big] \\ &\le \alpha H^3 K/2 + \alpha^{-1} H \cdot \log|\mathcal{A}|. \end{aligned} \tag{3.10}$$

Here the first inequality follows from telescoping the right-hand side of (3.9) across all the episodes and the fact that the KL-divergence is nonnegative. Also, the second inequality follows from the initialization of the policy and Q-function in Line 1 of Algorithm 1. Setting $\alpha = \sqrt{2\log|\mathcal{A}|/(HT)}$ in (3.10), we establish a $\sqrt{H^3 T \cdot \log|\mathcal{A}|}$-regret in the ideal setting.

Such an ideal setting demonstrates the key role of the KL-divergence in the policy improvement step defined in (3.2), where $\alpha > 0$ is the stepsize. Intuitively, without the KL-divergence, that is, setting $\alpha \to \infty$, the upper bound of the regret on the right-hand side of (3.10) tends to infinity. In fact, for any $\alpha < \infty$, the updated policy π_h^k in (3.3) is "conservatively" greedy with respect to the Q-function $Q_h^{\pi^{k-1},k-1}$ associated with the reward function r^{k-1}. In particular, the regularization effect of both π_h^{k-1} and α in (3.3) ensures that π_h^k is not "fully" committed to perform well only with respect to r^{k-1}, just in case the subsequent adversarially chosen reward function r^k significantly differs from r^{k-1}. In comparison, the "fully" greedy policy improvement step, which is commonly adopted by the existing work on value-based reinforcement learning (Jaksch et al., 2010; Osband et al., 2014; Osband & Van Roy, 2016; Azar et al., 2017; Dann et al., 2017; Strehl et al., 2006; Jin et al., 2018; 2019; Yang & Wang, 2019b;a), lacks such a notion of robustness. On the other hand, an intriguing question is whether being "conservatively" greedy is less sample-efficient than being "fully" greedy in the stationary setting, where the reward function is fixed across all the episodes. In fact, in the ideal setting where the Q-function $Q_h^{\pi^{k-1},k-1}$ associated with the reward function r^{k-1} in (3.3) is known, the "fully" greedy policy improvement step with $\alpha \to \infty$ corresponds to one step of policy iteration (Sutton & Barto, 2018), which converges to the globally optimal policy π^* within $K = H$ episodes and hence equivalently induces an H^2-regret. However, in the realistic setting, the Q-function $Q_h^{\pi^{k-1},k-1}$ in (3.1)-(3.3) is replaced by the estimated Q-function Q_h^{k-1} in Line 6 of Algorithm 1, which is obtained by the policy evaluation step defined in (3.5). As a result of the estimation uncertainty that arises from only observing finite historical data, it is indeed impossible to do better than the $\sqrt{T}$-regret even in the tabular setting (Jin et al., 2018), which is shown to be an information-theoretic lower bound. In the linear setting, OPPO attains such a lower bound in terms of the total number of steps $T = HK$. In other words, in the stationary setting, being "conservatively" greedy suffices to achieve sample-efficiency, which complements its advantages in terms of robustness in the more challenging setting with adversarially chosen reward functions.

4. Proof Sketch

4.1. Regret Decomposition

For the simplicity of discussion, we define the model prediction error as

$$\iota_h^k = r_h^k + \mathbb{P}_h V_{h+1}^k - Q_h^k, \tag{4.1}$$

which arises from estimating $\mathbb{P}_h V_{h+1}^k$ in the Bellman equation defined in (2.4) (with $V_{h+1}^{\pi^k,k}$ replaced by V_{h+1}^k) based on only finite historical data. Also, we define the following filtration generated by the state-action sequence and reward functions.

Definition 4.1 (Filtration). For any $(k,h) \in [K] \times [H]$, we define $\mathcal{F}_{k,h,1}$ as the σ-algebra generated by the following state-action sequence and reward functions,

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

For any $(k,h) \in [K] \times [H-1]$, we define $\mathcal{F}_{k,h,2}$ as the σ-algebra generated by

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

while for any $k \in [K]$, we define $\mathcal{F}_{k,H,2}$ as the σ-algebra generated by

$$\{(x_i^\tau, a_i^\tau)\}_{(\tau,i)\in[k]\times[H]} \cup \{r^\tau\}_{\tau\in[k+1]}.$$

The σ-algebra sequence $\{\mathcal{F}_{k,h,m}\}_{(k,h,m)\in[K]\times[H]\times[2]}$ is a filtration with respect to the timestep index

$$t(k,h,m) = (k-1)\cdot 2H + (h-1)\cdot 2 + m. \tag{4.2}$$

In other words, for any $t(k,h,m) \le t(k',h',m')$, it holds that $\mathcal{F}_{k,h,m} \subseteq \mathcal{F}_{k',h',m'}$.

By the definition of the σ-algebra $\mathcal{F}_{k,h,m}$, for $(k,h,m) \in [K] \times [H] \times [2]$, the estimated value function V_h^k and Q-function Q_h^k are measurable to $\mathcal{F}_{k-1,H,2}$, as they are obtained based on the $(k-1)$ historical trajectories and the reward function r^k adversarially chosen by the environment at the beginning of the k-th episode, both of which are measurable to $\mathcal{F}_{k-1,H,2}$.

In the following lemma, we decompose the regret defined in (2.1) into three terms. Recall that the globally optimal policy in hindsight π^* is defined in (3.8) and the model prediction error ι_h^k is defined in (4.1).

Lemma 4.2 (Regret Decomposition). It holds that

$$\begin{aligned}
&\text{Regret}(T) \qquad\qquad (4.3)\\
&= \sum_{k=1}^K \big(V_1^{\pi^*,k}(x_1^k) - V_1^{\pi^k,k}(x_1^k)\big)\\
&= \underbrace{\sum_{k=1}^K \sum_{h=1}^H \mathbb{E}_{\pi^*}\big[\langle Q_h^k(x_h,\cdot), \pi_h^*(\cdot\,|\,x_h) - \pi_h^k(\cdot\,|\,x_h)\rangle\big]}_{\text{(i)}}\\
&\quad + \underbrace{\mathcal{M}_{K,H,2}}_{\text{(ii)}} + \underbrace{\sum_{k=1}^K \sum_{h=1}^H \big(\mathbb{E}_{\pi^*}[\iota_h^k(x_h,a_h)] - \iota_h^k(x_h^k,a_h^k)\big)}_{\text{(iii)}},
\end{aligned}$$

which is independent of the linear setting in Assumption 2.1. Here $\{\mathcal{M}_{k,h,m}\}_{(k,h,m)\in[K]\times[H]\times[2]}$ is a martingale adapted to the filtration $\{\mathcal{F}_{k,h,m}\}_{(k,h,m)\in[K]\times[H]\times[2]}$, both with respect to the timestep index $t(k,h,m)$ defined in (4.2) of Definition 4.1.

Proof. See Appendix B.1 for a detailed proof. □

Lemma 4.2 allows us to characterize the regret by upper bounding terms (i), (ii), and (iii) in (4.3), respectively. In detail, term (i) corresponds to the right-hand side of (3.2) in Lemma 3.2 with the Q-function $Q_h^{\pi^k,k}$ replaced by the estimated Q-function Q_h^k, which is obtained by the policy evaluation step defined in (3.5). In particular, as the updated policy π_h^{k+1} is obtained by the policy improvement step in Line 6 of Algorithm 1 using π_h^k and Q_h^k, term (i) can be upper bounded following a similar analysis to the discussion in Section 3.2, which is based on Lemmas 3.2 and 3.3 as well as (3.10). Also, by the Azuma-Hoeffding inequality, term (ii) is a martingale that scales as $O(B_{\mathcal{M}}\sqrt{T_{\mathcal{M}}})$ with high probability, where $T_{\mathcal{M}}$ is the total number of timesteps and $B_{\mathcal{M}}$ is an upper bound of the martingale differences. More specifically, we prove that $T_{\mathcal{M}} = 2HK = 2T$ and $B_{\mathcal{M}} = 2H$ in Appendix C, which implies that term (ii) is $O(\sqrt{H^2T})$ with high probability. Meanwhile, term (iii) corresponds to the model prediction error, which is characterized subsequently in Section 4.2. Note that the regret decomposition in (4.3) of Lemma 4.2 is independent of the linear setting in Assumption 2.1, and therefore, applies to any forms of estimated Q-functions Q_h^k in more general settings. In particular, as long as we can upper bound term (iii) in (4.3), our regret analysis can be carried over even beyond the linear setting.

4.2. Model Prediction Error

To upper bound term (iii) in (4.3) of Lemma 4.2, we characterize the model prediction error ι_h^k defined in (4.1) in the following lemma. Recall that the bonus function Γ_h^k is defined in (3.6).

Lemma 4.3 (Upper Confidence Bound). Let $\lambda = 1$ in (3.5) and Line 12 of Algorithm 1, and $\beta = C\sqrt{dH^2 \cdot \log(dT/\zeta)}$ in (3.6) and Line 15 of Algorithm 1, where $C > 1$ is an absolute constant and $\zeta \in (0,1]$. Under Assumption 2.1, it holds with probability at least $1-\zeta/2$ that

$$-2\Gamma_h^k(x,a) \le \iota_h^k(x,a) \le 0$$

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

Proof. See Appendix B.2 for a detailed proof. □

Lemma 4.3 demonstrates the key role of uncertainty quantification in achieving sample-efficiency. More specifically, due to the uncertainty that arises from only observing finite historical data, the model prediction error $\iota_h^k(x,a)$ can be possibly large for the state-action pairs (x,a) that are less visited or even unseen. However, as is shown in Lemma 4.3, explicitly incorporating the bonus function Γ_h^k into the estimated Q-function Q_h^k ensures that $\iota_h^k(x,a) \le 0$ with high probability for any $(k,h) \in [K] \times [H]$ and $(x,a) \in \mathcal{S} \times \mathcal{A}$. In other words, the estimated Q-function Q_h^k is "optimistic in the face of uncertainty", as $\iota_h^k(x,a) \le 0$ or equivalently

$$Q_h^k(x,a) \ge r_h^k(x,a) + (\mathbb{P}_h V_{h+1}^k)(x,a) \tag{4.4}$$

implies that $\mathbb{E}_{\pi^*}[\iota_h^k(x_h, a_h)]$ in term (iii) of (4.3) is upper bounded by zero. Also, Lemma 4.3 implies that $-\iota_h^k(x_h^k, a_h^k) \le 2\Gamma_h^k(x_h^k, a_h^k)$ with high probability for any $(k,h) \in [K] \times [H]$. As a result, it only remains to upper bound the cumulative sum $\sum_{k=1}^K \sum_{h=1}^H 2\Gamma_h^k(x_h^k, a_h^k)$ corresponding to term (iii) in (4.3), which can be characterized by the elliptical potential lemma (Dani et al., 2008; Rusmevichientong & Tsitsiklis, 2010; Chu et al., 2011; Abbasi-Yadkori et al., 2011; Jin et al., 2019). See Appendix C for a detailed proof.

To illustrate the intuition behind the model prediction error ι_h^k defined in (4.1), we define the implicitly estimated transition dynamics as

$$\begin{aligned} &\widehat{\mathcal{P}}_{k,h}(x' \,|\, x,a) \\ &\quad = \psi(x,a,x')^\top (\Lambda_h^k)^{-1} \sum_{\tau=1}^{k-1} \phi_h^\tau(x_h^\tau, a_h^\tau) \cdot V_{h+1}^\tau(x_{h+1}^\tau), \end{aligned}$$

where Λ_h^k is defined in (3.6). Correspondingly, the policy evaluation step defined in (3.5) takes the following equivalent form (ignoring the truncation step for the simplicity of discussion),

$$Q_h^k \leftarrow r_h^k + \widehat{\mathbb{P}}_{k,h} V_{h+1}^k + \Gamma_h^k. \tag{4.5}$$

Here $\widehat{\mathbb{P}}_{k,h}$ is the operator form of the implicitly estimated transition kernel $\widehat{\mathcal{P}}_{k,h}(\cdot \,|\, \cdot, \cdot)$, which is defined by

$$(\widehat{\mathbb{P}}_{k,h} f)(x,a) = \int_{\mathcal{S}} \widehat{\mathcal{P}}_{k,h}(x' \,|\, x,a) \cdot f(x') \mathrm{d}x'$$

for any function $f : \mathcal{S} \to \mathbb{R}$. Correspondingly, by (4.1) and (4.5) we have

$$\begin{aligned} \iota_h^k &= r_h^k + \mathbb{P}_h V_{h+1}^k - Q_h^k \\ &= (\mathbb{P}_h - \widehat{\mathbb{P}}_{k,h}) V_{h+1}^k - \Gamma_h^k, \end{aligned} \tag{4.6}$$

where $\mathbb{P}_h - \widehat{\mathbb{P}}_{k,h}$ is the error that arises from implicitly estimating the transition dynamics based on only finite historical data. Such a model estimation error enters the regret in (4.3) of Lemma 4.2 only through the model prediction error $(\mathbb{P}_h - \widehat{\mathbb{P}}_{k,h}) V_{h+1}^k$, which allows us to bypass explicitly estimating the transition dynamics, and instead, employ the estimated Q-function Q_h^k obtained by the policy evaluation step defined in (4.5). As is shown in Appendix B.2, the bonus function Γ_h^k upper bounds $(\mathbb{P}_h - \widehat{\mathbb{P}}_{k,h}) V_{h+1}^k$ in (4.6) with high probability for any $(k,h) \in [K] \times [H]$ and $(x,a) \in \mathcal{S} \times \mathcal{A}$, which then ensures the optimism of the estimated Q-function Q_h^k in the sense of (4.4).

5. Conclusion

We study the sample efficiency of policy-based reinforcement learning in the episodic setting of linear MDPs with full-information feedback. We proposed an optimistic variant of the proximal policy optimization algorithm, dubbed as OPPO, which incorporates the principle of "optimism in the face of uncertainty" into policy optimization. When applied to the episodic MDP with unknown transition and adversarial reward, OPPO provably achieves a near-optimal $\sqrt{d^2 H^3 T}$-regret. To the best of our knowledge, OPPO is the first provably efficient policy optimization algorithm that explicitly incorporates exploration.

Acknowledgements

The authors would like to thank Lingxiao Wang, Wen Sun, and Sham Kakade for pointing out a technical issue in the initial version regarding the covering number of value functions in the linear setting. Such a technical issue is fixed in the camera-ready version with a definition of the linear MDP different from the one in the initial version. The authors would like to thank Csaba Szepesvári, Lin F. Yang, Yining Wang, and Simon S. Du for helpful discussions. The authors would also like to thank the anonymous reviewers for the valuable comments and the program chairs for helping with preparing for the camera-ready version.

References

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

Abbasi-Yadkori, Y., Bartlett, P., Bhatia, K., Lazic, N., Szepesvári, C., and Weisz, G. POLITEX: Regret bounds for policy iteration using expert prediction. In *International Conference on Machine Learning*, volume 97, pp. 3692–3702, 2019a.

Abbasi-Yadkori, Y., Lazic, N., Szepesvari, C., and Weisz, G. Exploration-enhanced POLITEX. *arXiv preprint arXiv:1908.10479*, 2019b.

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

Antos, A., Szepesvári, C., and Munos, R. Fitted Q-iteration in continuous action-space mdps. In *Advances in Neural Information Processing Systems*, pp. 9–16, 2008.

Auer, P., Cesa-Bianchi, N., and Fischer, P. Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 47(2-3):235–256, 2002.

Ayoub, A., Jia, Z., Szepesvari, C., Wang, M., and Yang, L. Model-based reinforcement learning with value-targeted regression. *arXiv preprint arXiv:2006.01107*, 2020.

Azar, M. G., Munos, R., Ghavamzadaeh, M., and Kappen, H. J. Speedy Q-learning. In *Advances in Neural Information Processing Systems*, 2011.

Azar, M. G., Gómez, V., and Kappen, H. J. Dynamic policy programming. *Journal of Machine Learning Research*, 13(Nov):3207–3245, 2012a.

Azar, M. G., Munos, R., and Kappen, B. On the sample complexity of reinforcement learning with a generative model. *arXiv preprint arXiv:1206.6461*, 2012b.

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

Baxter, J. and Bartlett, P. L. Direct gradient-based reinforcement learning. In *International Symposium on Circuits and Systems*, pp. 271–274, 2000.

Bhandari, J. and Russo, D. Global optimality guarantees for policy gradient methods. *arXiv preprint arXiv:1906.01786*, 2019.

Boyan, J. A. Least-squares temporal difference learning. *Machine Learning*, 49(2-3):233–246, 2002.

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

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

Cesa-Bianchi, N. and Lugosi, G. *Prediction, Learning, and Games*. Cambridge, 2006.

Chen, J. and Jiang, N. Information-theoretic considerations in batch reinforcement learning. *arXiv preprint arXiv:1905.00360*, 2019.

Chu, W., Li, L., Reyzin, L., and Schapire, R. Contextual bandits with linear payoff functions. In *International Conference on Artificial Intelligence and Statistics*, pp. 208–214, 2011.

Dani, V., Hayes, T. P., and Kakade, S. M. Stochastic linear optimization under bandit feedback. *Conference on Learning Theory*, 2008.

Dann, C., Lattimore, T., and Brunskill, E. Unifying PAC and regret: Uniform PAC bounds for episodic reinforcement learning. In *Advances in Neural Information Processing Systems*, pp. 5713–5723, 2017.

Dong, K., Peng, J., Wang, Y., and Zhou, Y. $\sqrt{n}$-regret for learning in Markov decision processes with function approximation and low Bellman rank. *arXiv preprint arXiv:1909.02506*, 2019.

Du, S. S., Kakade, S. M., Wang, R., and Yang, L. Is a good representation sufficient for sample efficient reinforcement learning? *arXiv preprint arXiv:1910.03016*, 2019a.

Du, S. S., Luo, Y., Wang, R., and Zhang, H. Provably efficient Q-learning with function approximation via distribution shift error checking oracle. *arXiv preprint arXiv:1906.06321*, 2019b.

Duan, Y., Chen, X., Houthooft, R., Schulman, J., and Abbeel, P. Benchmarking deep reinforcement learning for continuous control. In *International Conference on Machine Learning*, pp. 1329–1338, 2016.

Even-Dar, E., Kakade, S. M., and Mansour, Y. Online Markov decision processes. *Mathematics of Operations Research*, 34(3):726–736, 2009.

Farahmand, A.-m., Szepesvári, C., and Munos, R. Error propagation for approximate policy and value iteration. In *Advances in Neural Information Processing Systems*, pp. 568–576, 2010.

Fazel, M., Ge, R., Kakade, S. M., and Mesbahi, M. Global convergence of policy gradient methods for the linear quadratic regulator. *arXiv preprint arXiv:1801.05039*, 2018.

Geist, M., Scherrer, B., and Pietquin, O. A theory of regularized Markov decision processes. *arXiv preprint arXiv:1901.11275*, 2019.

Jaksch, T., Ortner, R., and Auer, P. Near-optimal regret bounds for reinforcement learning. *Journal of Machine Learning Research*, 11(4):1563–1600, 2010.

Jiang, N., Krishnamurthy, A., Agarwal, A., Langford, J., and Schapire, R. E. Contextual decision processes with low Bellman rank are PAC-learnable. In *International Conference on Machine Learning*, pp. 1704–1713, 2017.

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

Jin, C., Yang, Z., Wang, Z., and Jordan, M. I. Provably efficient reinforcement learning with linear function approximation. *arXiv preprint arXiv:1907.05388*, 2019.

Kakade, S. M. A natural policy gradient. In *Advances in Neural Information Processing Systems*, 2002.

Kakade, S. M. *On the Sample Complexity of Reinforcement Learning*. PhD thesis, University of London, 2003.

Koenig, S. and Simmons, R. G. Complexity analysis of real-time reinforcement learning. In *Association for the Advancement of Artificial Intelligence*, pp. 99–107, 1993.

Konda, V. R. and Tsitsiklis, J. N. Actor-critic algorithms. In *Advances in Neural Information Processing Systems*, 2000.

Lattimore, T. and Szepesvari, C. Learning with good feature representations in bandits and in RL with a generative model. *arXiv preprint arXiv:1911.07676*, 2019.

Leffler, B. R., Littman, M. L., and Edmunds, T. Efficient reinforcement learning with relocatable action models. In *Association for the Advancement of Artificial Intelligence*, pp. 572–577, 2007.

Liu, B., Cai, Q., Yang, Z., and Wang, Z. Neural proximal/trust region policy optimization attains globally optimal policy. *arXiv preprint arXiv:1906.10306*, 2019.

Mania, H., Guy, A., and Recht, B. Simple random search provides a competitive approach to reinforcement learning. *arXiv preprint arXiv:1803.07055*, 2018.

Munos, R. and Szepesvári, C. Finite-time bounds for fitted value iteration. *Journal of Machine Learning Research*, 9 (May):815–857, 2008.

Nemirovsky, A. S. and Yudin, D. B. *Problem Complexity and Method Efficiency in Optimization.* Wiley, 1983.

Neu, G., Antos, A., György, A., and Szepesvári, C. Online Markov decision processes under bandit feedback. In *Advances in Neural Information Processing Systems*, pp. 1804–1812, 2010a.

Neu, G., György, A., and Szepesvári, C. The online loop-free stochastic shortest-path problem. In *Conference on Learning Theory*, volume 2010, pp. 231–243, 2010b.

Neu, G., György, A., and Szepesvári, C. The adversarial stochastic shortest path problem with unknown transition probabilities. In *International Conference on Artificial Intelligence and Statistics*, pp. 805–813, 2012.

Neu, G., Jonsson, A., and Gómez, V. A unified view of entropy-regularized Markov decision processes. *arXiv preprint arXiv:1705.07798*, 2017.

OpenAI. OpenAI Five. https://openai.com/five/, 2019.

Osband, I. and Van Roy, B. On lower bounds for regret in reinforcement learning. *arXiv preprint arXiv:1608.02732*, 2016.

Osband, I., Van Roy, B., and Wen, Z. Generalization and exploration via randomized value functions. *arXiv preprint arXiv:1402.0635*, 2014.

Rosenberg, A. and Mansour, Y. Online stochastic shortest path with bandit feedback and unknown transition function. In *Advances in Neural Information Processing Systems*, pp. 2209–2218, 2019a.

Rosenberg, A. and Mansour, Y. Online convex optimization in adversarial Markov decision processes. *arXiv preprint arXiv:1905.07773*, 2019b.

Rusmevichientong, P. and Tsitsiklis, J. N. Linearly parameterized bandits. *Mathematics of Operations Research*, 35 (2):395–411, 2010.

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

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

Sidford, A., Wang, M., Wu, X., Yang, L., and Ye, Y. Near-optimal time and sample complexities for solving Markov decision processes with a generative model. In *Advances in Neural Information Processing Systems*, pp. 5186–5196, 2018a.

Sidford, A., Wang, M., Wu, X., and Ye, Y. Variance reduced value iteration and faster algorithms for solving Markov decision processes. In *Symposium on Discrete Algorithms*, pp. 770–787, 2018b.

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Van Den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., et al. Mastering the game of Go with deep neural networks and tree search. *Nature*, 529(7587):484, 2016.

Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., Hubert, T., Baker, L., Lai, M., Bolton, A., et al. Mastering the game of Go without human knowledge. *Nature*, 550(7676):354, 2017.

Strehl, A. L., Li, L., Wiewiora, E., Langford, J., and Littman, M. L. PAC model-free reinforcement learning. In *International Conference on Machine Learning*, pp. 881–888, 2006.

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

Sutton, R. S., McAllester, D. A., Singh, S. P., and Mansour, Y. Policy gradient methods for reinforcement learning with function approximation. In *Advances in Neural Information Processing Systems*, 2000.

Tosatto, S., Pirotta, M., D'Eramo, C., and Restelli, M. Boosted fitted Q-iteration. In *International Conference on Machine Learning*, pp. 3434–3443, 2017.

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

Wainwright, M. J. Variance-reduced Q-learning is minimax optimal. *arXiv preprint arXiv:1906.04697*, 2019.

Wang, L., Cai, Q., Yang, Z., and Wang, Z. Neural policy gradient methods: Global optimality and rates of convergence. *arXiv preprint arXiv:1909.01150*, 2019.

Wang, W. Y., Li, J., and He, X. Deep reinforcement learning for NLP. In *Association for Computational Linguistics*, pp. 19–21, 2018.

Wen, Z. and Van Roy, B. Efficient reinforcement learning in deterministic systems with value function generalization. *Mathematics of Operations Research*, 42(3): 762–782, 2017.

Williams, R. J. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, 8(3-4):229–256, 1992.

Xiao, L. Dual averaging methods for regularized stochastic learning and online optimization. *Journal of Machine Learning Research*, 11(Oct):2543–2596, 2010.

Yang, L. and Wang, M. Reinforcement leaning in feature space: Matrix bandit, kernels, and regret bound. *arXiv preprint arXiv:1905.10389*, 2019a.

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

Yang, Z., Chen, Y., Hong, M., and Wang, Z. On the global convergence of actor-critic: A case for linear quadratic regulator with ergodic cost. *arXiv preprint arXiv:1907.06246*, 2019a.

Yang, Z., Xie, Y., and Wang, Z. A theoretical analysis of deep Q-learning. *arXiv preprint arXiv:1901.00137*, 2019b.

Yu, J. Y., Mannor, S., and Shimkin, N. Markov decision processes with arbitrary reward processes. *Mathematics of Operations Research*, 34(3):737–757, 2009.

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

Zimin, A. and Neu, G. Online learning in episodic Markovian decision processes by relative entropy policy search. In *Advances in Neural Information Processing Systems*, pp. 1583–1591, 2013.

A. Proofs of Lemmas in Section 3

A.1. Proof of Lemma 3.2

Proof. In this section, we focus on the k-th episode and omit the episode index k for notational simplicity. For any $h \in [H]$ and policy $\pi \in \Delta(\mathcal{A} \,|\, \mathcal{S}, H)$, we define the Bellman evaluation operator $\mathbb{T}_{h,\pi}$ by

$$\begin{aligned}(\mathbb{T}_{h,\pi}V)(x) &= \mathbb{E}[r_h(x,a) + V(x') \,|\, a \sim \pi_h(\cdot \,|\, x),\, x' \sim \mathcal{P}_h(\cdot \,|\, x,a)] \\ &= \langle (r_h + \mathbb{P}_h V)(x,\cdot), \pi_h(\cdot \,|\, x)\rangle \end{aligned} \tag{A.1}$$

for any function $V : \mathcal{S} \to \mathbb{R}$. By the definition of the value function V_h^π in (2.2), we have

$$V_h^\pi = \prod_{i=h}^H \mathbb{T}_{i,\pi}\mathbf{0} \tag{A.2}$$

for any $h \in [H]$, where $\mathbf{0}$ is a zero function on $\mathcal{S}$. Here $\prod_{i=h}^H \mathbb{T}_{i,\pi}$ denotes the sequential composition of the Bellman evaluation operators $\mathbb{T}_{i,\pi}$. Thus, for any policies $\pi', \pi \in \Delta(\mathcal{A} \,|\, \mathcal{S}, H)$, it holds that

$$\begin{aligned} V_1^{\pi'} - V_1^\pi &= \prod_{h=1}^H \mathbb{T}_{h,\pi'}\mathbf{0} - \prod_{h=1}^H \mathbb{T}_{h,\pi}\mathbf{0} \\ &= \prod_{h=1}^H \mathbb{T}_{h,\pi'}\mathbf{0} - \sum_{h=1}^{H-1}\Big(\prod_{i=1}^h \mathbb{T}_{i,\pi'} \prod_{i=h+1}^H \mathbb{T}_{i,\pi}\mathbf{0} - \prod_{i=1}^h \mathbb{T}_{i,\pi'} \prod_{i=h+1}^H \mathbb{T}_{i,\pi}\mathbf{0}\Big) - \prod_{h=1}^H \mathbb{T}_{h,\pi}\mathbf{0} \\ &= \sum_{h=H}^1 \Big(\prod_{i=1}^h \mathbb{T}_{i,\pi'} \prod_{i=h+1}^H \mathbb{T}_{i,\pi}\mathbf{0} - \prod_{i=1}^{h-1} \mathbb{T}_{i,\pi'} \prod_{i=h}^H \mathbb{T}_{i,\pi}\mathbf{0}\Big). \end{aligned} \tag{A.3}$$

Meanwhile, by (A.2) we have that, on the right-hand side of (A.3),

$$\begin{aligned} &\prod_{i=1}^h \mathbb{T}_{i,\pi'} \prod_{i=h+1}^H \mathbb{T}_{i,\pi}\mathbf{0} - \prod_{i=1}^{h-1} \mathbb{T}_{i,\pi'} \prod_{i=h}^H \mathbb{T}_{i,\pi}\mathbf{0} \\ &\quad = \prod_{i=1}^{h-1} \mathbb{T}_{i,\pi'}(\mathbb{T}_{h,\pi'} - \mathbb{T}_{h,\pi}) \prod_{i=h+1}^H \mathbb{T}_{i,\pi}\mathbf{0} = \prod_{i=1}^{h-1} \mathbb{T}_{i,\pi'}(\mathbb{T}_{h,\pi'} - \mathbb{T}_{h,\pi}) V_{h+1}^\pi. \end{aligned} \tag{A.4}$$

By the definition of the Bellman evaluation operator $\mathbb{T}_{h,\pi}$ in (A.1), we have

$$(\mathbb{T}_{h,\pi'} - \mathbb{T}_{h,\pi})V_{h+1}^\pi = \langle r_h + \mathbb{P}_h V_{h+1}^\pi, \pi_h' - \pi_h\rangle_{\mathcal{A}} = \langle Q_h^\pi, \pi_h' - \pi_h\rangle_{\mathcal{A}}, \tag{A.5}$$

where the last equality follows from (2.4). Combining (A.3), (A.4), (A.5), and the linearity of the Bellman evaluation operator defined in (A.1), we obtain

$$\begin{aligned} V_1^{\pi'}(x_1) - V_1^\pi(x_1) &= \sum_{h=1}^H \Big(\prod_{i=1}^{h-1} \mathbb{T}_{i,\pi'} \langle Q_h^\pi, \pi_h' - \pi_h\rangle_{\mathcal{A}}\Big)(x_1) \\ &= \mathbb{E}_{\pi'}\Big[\sum_{h=1}^H \langle Q_h^\pi(x_h,\cdot), \pi_h'(\cdot \,|\, x_h) - \pi_h(\cdot \,|\, x_h)\rangle \,\Big|\, x_1\Big], \end{aligned}$$

which concludes the proof of Lemma 3.2. □

A.2. Proof of Lemma 3.3

Proof. For any function $g : \mathcal{A} \to \mathbb{R}$ and distributions $p, p', p^* \in \Delta(\mathcal{A})$ that satisfy

$$p'(\cdot) \propto p(\cdot) \cdot \exp\big(\alpha \cdot g(\cdot)\big),$$

we have

$$\begin{aligned}\alpha \cdot \langle g, p^* - p' \rangle &= \langle z + \log(p'/p), p^* - p' \rangle \\ &= \langle z, p^* - p' \rangle + \langle \log(p^*/p), p^* \rangle + \langle \log(p'/p^*), p^* \rangle + \langle \log(p'/p), -p' \rangle \\ &= D_{\mathrm{KL}}(p^* \,\|\, p) - D_{\mathrm{KL}}(p^* \,\|\, p') - D_{\mathrm{KL}}(p' \,\|\, p). \end{aligned} \tag{A.6}$$

Here $z : \mathcal{A} \to \mathbb{R}$ is a constant function defined by

$$z(a) = \log\Big(\sum_{a' \in \mathcal{A}} p(a') \cdot \exp\big(\alpha \cdot g(a')\big)\Big),$$

which implies that $\langle z, p^* - p' \rangle = 0$ in (A.6) as $p', p^* \in \Delta(\mathcal{A})$. Moreover, by (A.6) we have

$$\begin{aligned}\alpha \cdot \langle Q(x, \cdot), p^*(\cdot) - p(\cdot) \rangle &= \alpha \cdot \langle Q(x, \cdot), p^*(\cdot) - p'(\cdot) \rangle - \alpha \cdot \langle Q(x, \cdot), p(\cdot) - p'(\cdot) \rangle \\ &\le D_{\mathrm{KL}}\big(p^*(\cdot) \,\big\|\, p(\cdot)\big) - D_{\mathrm{KL}}\big(p^*(\cdot) \,\big\|\, p'(\cdot)\big) - D_{\mathrm{KL}}\big(p'(\cdot) \,\big\|\, p(\cdot)\big) \\ &\quad + \alpha \cdot \|Q(x, \cdot)\|_\infty \cdot \|p(\cdot) - p'(\cdot)\|_1 \end{aligned} \tag{A.7}$$

for any state $x \in \mathcal{S}$. Meanwhile, by Pinsker's inequality, it holds that

$$D_{\mathrm{KL}}(p' \,\|\, p) \ge \|p - p'\|_1^2/2. \tag{A.8}$$

Combining (A.7), (A.8), and the fact that $\|Q(x, \cdot)\|_\infty \le H$ for any state $x \in \mathcal{S}$, we obtain

$$\begin{aligned}&\alpha \cdot \langle Q(x, \cdot), p^*(\cdot) - p(\cdot) \rangle \\ &\quad \le D_{\mathrm{KL}}\big(p^*(\cdot) \,\big\|\, p(\cdot)\big) - D_{\mathrm{KL}}\big(p^*(\cdot) \,\big\|\, p'(\cdot)\big) - \|p(\cdot) - p'(\cdot)\|_1^2/2 + \alpha H \cdot \|p(\cdot) - p'(\cdot)\|_1 \\ &\quad \le D_{\mathrm{KL}}\big(p^*(\cdot) \,\big\|\, p(\cdot)\big) - D_{\mathrm{KL}}\big(p^*(\cdot) \,\big\|\, p'(\cdot)\big) + \alpha^2 H^2/2, \end{aligned}$$

which concludes the proof of Lemma 3.3. □

B. Proofs of Lemmas in Section 4

For notational simplicity, we define the operators $\mathbb{J}_h$ and $\mathbb{J}_{k,h}$ respectively by

$$(\mathbb{J}_h f)(x) = \langle f(x, \cdot), \pi_h^*(\cdot \,|\, x) \rangle, \quad (\mathbb{J}_{k,h} f)(x) = \langle f(x, \cdot), \pi_h^k(\cdot \,|\, x) \rangle \tag{B.1}$$

for any $(k, h) \in [K] \times [H]$ and function $f : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$. Also, we define

$$\xi_h^k(x) = (\mathbb{J}_h Q_h^k)(x) - (\mathbb{J}_{k,h} Q_h^k)(x) = \langle Q_h^k(x, \cdot), \pi_h^*(\cdot \,|\, x) - \pi_h^k(\cdot \,|\, x) \rangle \tag{B.2}$$

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

B.1. Proof of Lemma 4.2

Proof. We decompose the instantaneous regret at the k-th episode into the following two terms,

$$V_1^{\pi^*, k}(x_1^k) - V_1^{\pi^k, k}(x_1^k) = \underbrace{V_1^{\pi^*, k}(x_1^k) - V_1^k(x_1^k)}_{\text{(i)}} + \underbrace{V_1^k(x_1^k) - V_1^{\pi^k, k}(x_1^k)}_{\text{(ii)}}. \tag{B.3}$$

Term (i). By the definitions of the value function $V_h^{\pi^*, k}$ in (2.4), the estimated value function V_h^k in (3.4), the operators $\mathbb{J}_h$ and $\mathbb{J}_{k,h}$ in (B.1), and ξ_h^k in (B.2), we have

$$\begin{aligned}V_h^{\pi^*, k} - V_h^k &= \mathbb{J}_h Q_h^{\pi^*, k} - \mathbb{J}_{k,h} Q_h^k \\ &= \mathbb{J}_h(Q_h^{\pi^*, k} - Q_h^k) + (\mathbb{J}_h - \mathbb{J}_{k,h}) Q_h^k = \mathbb{J}_h(Q_h^{\pi^*, k} - Q_h^k) + \xi_h^k \end{aligned} \tag{B.4}$$

for any $(k, h) \in [K] \times [H]$. Meanwhile, by the definition of the model prediction error, that is, $\iota_h^k = r_h^k + \mathbb{P}_h V_{h+1}^k - Q_h^k$, we have that, on the right-hand side of (B.4),

$$Q_h^{\pi^*,k} = r_h^k + \mathbb{P}_h V_{h+1}^{\pi^*,k}, \quad Q_h^k = r_h^k + \mathbb{P}_h V_{h+1}^k - \iota_h^k,$$

which implies

$$Q_h^{\pi^*,k} - Q_h^k = \mathbb{P}_h (V_{h+1}^{\pi^*,k} - V_{h+1}^k) + \iota_h^k. \tag{B.5}$$

Combining (B.4) and (B.5), we obtain

$$V_h^{\pi^*,k} - V_h^k = \mathbb{J}_h \mathbb{P}_h (V_{h+1}^{\pi^*,k} - V_{h+1}^k) + \mathbb{J}_h \iota_h^k + \xi_h^k \tag{B.6}$$

for any $(k, h) \in [K] \times [H]$. For any $k \in [K]$, recursively expanding (B.6) across $h \in [H]$ yields

$$V_1^{\pi^*,k} - V_1^k = \Big(\prod_{h=1}^{H} \mathbb{J}_h \mathbb{P}_h\Big)(V_{H+1}^{\pi^*,k} - V_{H+1}^k) + \sum_{h=1}^{H}\Big(\prod_{i=1}^{h-1} \mathbb{J}_i \mathbb{P}_i\Big)\mathbb{J}_h \iota_h^k + \sum_{h=1}^{H}\Big(\prod_{i=1}^{h-1} \mathbb{J}_i \mathbb{P}_i\Big)\xi_h^k,$$

where $V_{H+1}^{\pi^*,k} = V_{H+1}^k = \mathbf{0}$. Therefore, we obtain

$$V_1^{\pi^*,k} - V_1^k = \sum_{h=1}^{H}\Big(\prod_{i=1}^{h-1} \mathbb{J}_i \mathbb{P}_i\Big)\mathbb{J}_h \iota_h^k + \sum_{h=1}^{H}\Big(\prod_{i=1}^{h-1} \mathbb{J}_i \mathbb{P}_i\Big)\xi_h^k.$$

By the definitions of $\mathbb{P}_h$ in (2.5), $\mathbb{J}_h$ in (B.1), and ξ_h^k in (B.2), we further obtain

$$\begin{aligned} & V_1^{\pi^*,k}(x_1^k) - V_1^k(x_1^k) \\ & = \sum_{h=1}^{H} \mathbb{E}_{\pi^*}[\iota_h^k(x_h, a_h) \,|\, x_1 = x_1^k] + \sum_{h=1}^{H} \mathbb{E}_{\pi^*}\big[\langle Q_h^k(x_h, \cdot), \pi_h^*(\cdot \,|\, x_h) - \pi_h^k(\cdot \,|\, x_h)\rangle \,\big|\, x_1 = x_1^k\big] \end{aligned} \tag{B.7}$$

for any $k \in [K]$.

Term (ii). By the definitions of the value function $V_h^{\pi^k,k}$ in (2.4), the estimated value function V_h^k in (3.4), and the operator $\mathbb{J}_{k,h}$ in (B.1), we have

$$V_h^k(x_h^k) - V_h^{\pi^k,k}(x_h^k) = \big(\mathbb{J}_{k,h}(Q_h^k - Q_h^{\pi^k,k})\big)(x_h^k) + \iota_h^k(x_h^k, a_h^k) - \iota_h^k(x_h^k, a_h^k) \tag{B.8}$$

for any $(k, h) \in [K] \times [H]$. By the definition of the model prediction error ι_h^k in (4.1), we have

$$\begin{aligned} \iota_h^k(x_h^k, a_h^k) &= r_h^k(x_h^k, a_h^k) + (\mathbb{P}_h V_{h+1}^k)(x_h^k, a_h^k) - Q_h^k(x_h^k, a_h^k) \\ &= \big(r_h^k(x_h^k, a_h^k) + (\mathbb{P}_h V_{h+1}^k)(x_h^k, a_h^k) - Q_h^{\pi^k,k}(x_h^k, a_h^k)\big) + \big(Q_h^{\pi^k,k}(x_h^k, a_h^k) - Q_h^k(x_h^k, a_h^k)\big) \\ &= \big(\mathbb{P}_h(V_{h+1}^k - V_{h+1}^{\pi^k,k})\big)(x_h^k, a_h^k) + (Q_h^{\pi^k,k} - Q_h^k)(x_h^k, a_h^k), \end{aligned} \tag{B.9}$$

where the last equality follows from (2.4). Plugging (B.9) into (B.8), we obtain

$$\begin{aligned} V_h^k(x_h^k) - V_h^{\pi^k,k}(x_h^k) = {} & \big(\mathbb{J}_{k,h}(Q_h^k - Q_h^{\pi^k,k})\big)(x_h^k) + (Q_h^{\pi^k,k} - Q_h^k)(x_h^k, a_h^k) \\ & + \big(\mathbb{P}_h(V_{h+1}^k - V_{h+1}^{\pi^k,k})\big)(x_h^k, a_h^k) - \iota_h^k(x_h^k, a_h^k), \end{aligned} \tag{B.10}$$

which implies

$$\begin{aligned} V_h^k(x_h^k) - V_h^{\pi^k,k}(x_h^k) = {} & \underbrace{\big(\mathbb{J}_{k,h}(Q_h^k - Q_h^{\pi^k,k})\big)(x_h^k) - (Q_h^k - Q_h^{\pi^k,k})(x_h^k, a_h^k)}_{D_{k,h,1}} \\ & + \underbrace{\big(\mathbb{P}_h(V_{h+1}^k - V_{h+1}^{\pi^k,k})\big)(x_h^k, a_h^k) - (V_{h+1}^k - V_{h+1}^{\pi^k,k})(x_{h+1}^k)}_{D_{k,h,2}} \\ & + (V_{h+1}^k - V_{h+1}^{\pi^k,k})(x_{h+1}^k) - \iota_h^k(x_h^k, a_h^k) \end{aligned} \tag{B.11}$$

for any $(k, h) \in [K] \times [H]$. For any $k \in [K]$, recursively expanding (B.11) across $h \in [H]$ yields

$$\begin{aligned}
&V_1^k(x_1^k) - V_1^{\pi^k,k}(x_1^k) \\
&\quad = V_{H+1}^k(x_{H+1}^k) - V_{H+1}^{\pi^k,k}(x_{H+1}^k) - \sum_{h=1}^H \iota_h^k(x_h^k, a_h^k) + \sum_{h=1}^H (D_{k,h,1} + D_{k,h,2}),
\end{aligned}$$

where $V_{H+1}^k(x_{H+1}^k) = V_{H+1}^{\pi^k,k}(x_{H+1}^k) = 0$. Therefore, we obtain

$$V_1^k(x_1^k) - V_1^{\pi^k,k}(x_1^k) = -\sum_{h=1}^H \iota_h^k(x_h^k, a_h^k) + \sum_{h=1}^H (D_{k,h,1} + D_{k,h,2}). \tag{B.12}$$

By Definition 4.1 and the definitions of $D_{k,h,1}$ and $D_{k,h,2}$ in (B.11), we have

$$D_{k,h,1} \in \mathcal{F}_{k,h,1}, \quad D_{k,h,2} \in \mathcal{F}_{k,h,2}, \quad \mathbb{E}[D_{k,h,1} \,|\, \mathcal{F}_{k,h-1,2}] = 0, \quad \mathbb{E}[D_{k,h,2} \,|\, \mathcal{F}_{k,h,1}] = 0 \tag{B.13}$$

for any $(k, h) \in [K] \times [H]$. Here we have that $\mathcal{F}_{k,0,2} = \mathcal{F}_{k-1,H,2}$ for any $k \geq 2$, as (4.2) of Definition 4.1 implies

$$t(k, 0, 2) = t(k-1, H, 2) = (k-1) \cdot 2H.$$

Also, we define $\mathcal{F}_{1,0,2}$ to be empty. Thus, (B.13) allows us to define the martingale

$$\begin{aligned}
\mathcal{M}_{k,h,m} &= \sum_{\tau=1}^{k-1} \sum_{i=1}^H (D_{\tau,i,1} + D_{\tau,i,2}) + \sum_{i=1}^{h-1} (D_{k,i,1} + D_{k,i,2}) + \sum_{\ell=1}^m D_{k,h,\ell} \\
&= \sum_{\substack{(\tau,i,\ell) \in [K] \times [H] \times [2], \\ t(\tau,i,\ell) \leq t(k,h,m)}} D_{\tau,i,\ell}
\end{aligned} \tag{B.14}$$

with respect to the timestep index $t(k, h, m)$ defined in (4.2) of Definition 4.1. Such a martingale is adapted to the filtration $\{\mathcal{F}_{k,h,m}\}_{(k,h,m) \in [K] \times [H] \times [2]}$. In particular, we have that, on the right-hand side of (B.12),

$$\sum_{k=1}^K \sum_{h=1}^H (D_{k,h,1} + D_{k,h,2}) = \mathcal{M}_{K,H,2}. \tag{B.15}$$

Combining (B.3), (B.7), (B.12), and (B.15), we obtain

$$\begin{aligned}
\sum_{k=1}^K \bigl(V_1^{\pi^*,k}(x_1^k) - V_1^{\pi^k,k}(x_1^k)\bigr) &= \sum_{k=1}^K \sum_{h=1}^H \mathbb{E}_{\pi^*}[\iota_h^k(x_h, a_h) \,|\, x_1 = x_1^k] \\
&\quad + \sum_{k=1}^K \sum_{h=1}^H \mathbb{E}_{\pi^*}\bigl[\langle Q_h^k(x_h, \cdot), \pi_h^*(\cdot \,|\, x_h) - \pi_h^k(\cdot \,|\, x_h) \rangle \,\big|\, x_1 = x_1^k\bigr] \\
&\quad - \sum_{k=1}^K \sum_{h=1}^H \iota_h^k(x_h^k, a_h^k) + \mathcal{M}_{K,H,2},
\end{aligned}$$

which concludes the proof of Lemma 4.2. □

B.2. Proof of Lemma 4.3

Proof. Recall that ϕ_h^k defined in (3.4) takes the following form,

$$\phi_h^k(x, a) = \int_{\mathcal{S}} \psi(x, a, x') \cdot V_{h+1}^k(x') \mathrm{d}x'$$

for any $(k, h) \in [K] \times [H]$ and $(x, a) \in \mathcal{S} \times \mathcal{A}$. Also, recall that the estimated Q-function Q_h^k obtained by the policy evaluation step defined in (3.5) takes the following form,

$$Q_h^k(x, a) = \min\{r_h^k(x, a) + \phi_h^k(x, a)^\top w_h^k + \Gamma_h^k(x, a), H - h + 1\}^+, \quad \text{where } w_h^k = (\Lambda_h^k)^{-1}\Big(\sum_{\tau=1}^{k-1} \phi_h^\tau(x_h^\tau, a_h^\tau) \cdot V_{h+1}^\tau(x_{h+1}^\tau)\Big) \tag{B.16}$$

for any $(k, h) \in [K] \times [H]$ and $(x, a) \in \mathcal{S} \times \mathcal{A}$. Here Γ_h^k and Λ_h^k are defined in (3.6). Meanwhile, by Assumption 2.1 we have

$$\begin{aligned}(\mathbb{P}_h V_{h+1}^k)(x, a) &= \int_{\mathcal{S}} \psi(x, a, x')^\top \theta_h \cdot V_{h+1}^k(x') \mathrm{d}x' \\ &= \phi_h^k(x, a)^\top \theta_h = \phi_h^k(x, a)^\top (\Lambda_h^k)^{-1} \Lambda_h^k \theta_h\end{aligned} \tag{B.17}$$

for any $(k, h) \in [K] \times [H]$ and $(x, a) \in \mathcal{S} \times \mathcal{A}$. Plugging the definition of Λ_h^k in (3.6) into (B.17), we obtain

$$\begin{aligned}(\mathbb{P}_h V_{h+1}^k)(x, a) &= \phi_h^k(x, a)^\top (\Lambda_h^k)^{-1}\Big(\sum_{\tau=1}^{k-1} \phi_h^\tau(x_h^\tau, a_h^\tau)\phi_h^\tau(x_h^\tau, a_h^\tau)^\top \theta_h + \lambda \cdot \theta_h\Big) \\ &= \phi_h^k(x, a)^\top (\Lambda_h^k)^{-1}\Big(\sum_{\tau=1}^{k-1} \phi_h^\tau(x_h^\tau, a_h^\tau) \cdot (\mathbb{P}_h V_{h+1}^\tau)(x_h^\tau, a_h^\tau) + \lambda \cdot \theta_h\Big)\end{aligned} \tag{B.18}$$

for any $(k, h) \in [K] \times [H]$ and $(x, a) \in \mathcal{S} \times \mathcal{A}$. Here the second equality follows from (B.17) with V_{h+1}^k replaced by V_{h+1}^τ for any $\tau \in [k - 1]$. Combining (B.16) and (B.18), we obtain

$$\begin{aligned}&\phi_h^k(x, a)^\top w_h^k - (\mathbb{P}_h V_{h+1}^k)(x, a) \\ &= \underbrace{\phi_h^k(x, a)^\top (\Lambda_h^k)^{-1}\Big(\sum_{\tau=1}^{k-1} \phi_h^\tau(x_h^\tau, a_h^\tau) \cdot \big(V_{h+1}^\tau(x_{h+1}^\tau) - (\mathbb{P}_h V_{h+1}^\tau)(x_h^\tau, a_h^\tau)\big)\Big)}_{\text{(i)}} - \underbrace{\lambda \cdot \phi_h^k(x, a)^\top (\Lambda_h^k)^{-1} \theta_h}_{\text{(ii)}}\end{aligned} \tag{B.19}$$

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

Term (i). As is defined in (3.6), $(\Lambda_h^k)^{-1}$ is a positive-definite matrix. By the Cauchy-Schwarz inequality, the absolute value of term (i) is upper bounded as

$$|\text{(i)}| \le \sqrt{\phi_h^k(x, a)^\top (\Lambda_h^k)^{-1} \phi_h^k(x, a)} \cdot \Big\|\sum_{\tau=1}^{k-1} \phi_h^\tau(x_h^\tau, a_h^\tau) \cdot \big(V_{h+1}^\tau(x_{h+1}^\tau) - (\mathbb{P}_h V_{h+1}^\tau)(x_h^\tau, a_h^\tau)\big)\Big\|_{(\Lambda_h^k)^{-1}} \tag{B.20}$$

for any $(k, h) \in [K] \times [H]$ and $(x, a) \in \mathcal{S} \times \mathcal{A}$. Under the event $\mathcal{E}$ defined in (D.1) of Lemma D.1, which happens with probability at least $1 - \zeta/2$, it holds that

$$|\text{(i)}| \le C'' \sqrt{dH^2 \cdot \log(dT/\zeta)} \cdot \sqrt{\phi_h^k(x, a)^\top (\Lambda_h^k)^{-1} \phi_h^k(x, a)} \tag{B.21}$$

for any $(k, h) \in [K] \times [H]$ and $(x, a) \in \mathcal{S} \times \mathcal{A}$. Here $C'' > 0$ is an absolute constant and $\zeta \in (0, 1]$.

Term (ii). Similar to (B.20), the absolute value of term (ii) is upper bounded as

$$\begin{aligned}|\text{(ii)}| &\le \lambda \cdot \sqrt{\phi_h^k(x, a)^\top (\Lambda_h^k)^{-1} \phi_h^k(x, a)} \cdot \|\theta_h\|_{(\Lambda_h^k)^{-1}} \\ &\le \sqrt{\lambda} \cdot \sqrt{\phi_h^k(x, a)^\top (\Lambda_h^k)^{-1} \phi_h^k(x, a)} \cdot \|\theta_h\|_2 \le \sqrt{\lambda d} \cdot \sqrt{\phi_h^k(x, a)^\top (\Lambda_h^k)^{-1} \phi_h^k(x, a)}\end{aligned} \tag{B.22}$$

for any $(k, h) \in [K] \times [H]$ and $(x, a) \in \mathcal{S} \times \mathcal{A}$. Here the first inequality follows from the Cauchy-Schwarz inequality, the second inequality follows from the fact that $\Lambda_h^k \succeq \lambda \cdot I$, and the last inequality follows from Assumption 2.1, which assumes that $\|\theta_h\|_2 \le \sqrt{d}$.

Combining (B.19), (B.21), (B.22), and the fact that $\lambda = 1$, we obtain

$$\begin{aligned} &|\phi_h^k(x,a)^\top w_h^k - (\mathbb{P}_h V_{h+1}^k)(x,a)| \\ &\quad \le C\sqrt{dH^2 \cdot \log(dT/\zeta)} \cdot \sqrt{\phi_h^k(x,a)^\top (\Lambda_h^k)^{-1} \phi_h^k(x,a)} \end{aligned} \tag{B.23}$$

for any $(k,h) \in [K] \times [H]$ and $(x,a) \in \mathcal{S} \times \mathcal{A}$ under the event $\mathcal{E}$ defined in (D.1) of Lemma D.1. Here $C > 1$ is an absolute constant. Setting

$$\beta = C\sqrt{dH^2 \cdot \log(dT/\zeta)}$$

in the bonus function Γ_h^k defined in (3.6), by (B.23) we obtain

$$|\phi_h^k(x,a)^\top w_h^k - (\mathbb{P}_h V_{h+1}^k)(x,a)| \le \Gamma_h^k(x,a) \tag{B.24}$$

for any $(k,h) \in [K] \times [H]$ and $(x,a) \in \mathcal{S} \times \mathcal{A}$ under $\mathcal{E}$. As (3.5) implies that $(\mathbb{P}_h V_{h+1}^k)(x,a) \ge 0$, by (B.24) we have

$$\phi_h^k(x,a)^\top w_h^k + \Gamma_h^k(x,a) \ge 0 \tag{B.25}$$

for any $(k,h) \in [K] \times [H]$ and $(x,a) \in \mathcal{S} \times \mathcal{A}$ under $\mathcal{E}$. Hence, for the model prediction error ι_h^k defined in (4.1), by (B.16), (B.24), and (B.25) we have

$$\begin{aligned} -\iota_h^k(x,a) &= Q_h^k(x,a) - (r_h^k + \mathbb{P}_h V_{h+1}^k)(x,a) \\ &\le r_h^k(x,a) + \phi_h^k(x,a)^\top w_h^k + \Gamma_h^k(x,a) - (r_h^k + \mathbb{P}_h V_{h+1}^k)(x,a) \le 2\Gamma_h^k(x,a) \end{aligned} \tag{B.26}$$

for any $(k,h) \in [K] \times [H]$ and $(x,a) \in \mathcal{S} \times \mathcal{A}$ under $\mathcal{E}$. Meanwhile, as (3.5) implies that $(\mathbb{P}_h V_{h+1}^k)(x,a) \le H - h$ and hence

$$(r_h^k + \mathbb{P}_h V_{h+1}^k)(x,a) \le H - h + 1,$$

by (4.1), (B.16), and (B.24) we have

$$\begin{aligned} \iota_h^k(x,a) &= (r_h^k + \mathbb{P}_h V_{h+1}^k)(x,a) - Q_h^k(x,a) \\ &\le (r_h^k + \mathbb{P}_h V_{h+1}^k)(x,a) - \min\{r_h^k(x,a) + \phi_h^k(x,a)^\top w_h^k + \Gamma_h^k(x,a), H - h + 1\} \\ &= \max\{(\mathbb{P}_h V_{h+1}^k)(x,a) - \phi_h^k(x,a)^\top w_h^k - \Gamma_h^k(x,a), (r_h^k + \mathbb{P}_h V_{h+1}^k)(x,a) - (H - h + 1)\} \\ &\le 0 \end{aligned} \tag{B.27}$$

for any $(k,h) \in [K] \times [H]$ and $(x,a) \in \mathcal{S} \times \mathcal{A}$ under $\mathcal{E}$. Thus, combining (B.26), (B.27), and Lemma D.1, which ensures that $\mathcal{E}$ happens with probability at least $1 - \zeta/2$, we conclude the proof of Lemma 4.3. □

C. Proof of Theorem 3.1

Proof. We upper bound terms (i)-(iii) in (4.3) of Lemma 4.2 respectively, that is,

$$\begin{aligned} \text{Regret}(T) &= \sum_{k=1}^{K} \big(V_1^{\pi^*,k}(x_1^k) - V_1^{\pi^k,k}(x_1^k)\big) \\ &= \underbrace{\sum_{k=1}^{K}\sum_{h=1}^{H} \mathbb{E}_{\pi^*}\big[\langle Q_h^k(x_h,\cdot), \pi_h^*(\cdot\,|\,x_h) - \pi_h^k(\cdot\,|\,x_h)\rangle \,\big|\, x_1 = x_1^k\big]}_{\text{(i)}} + \underbrace{\mathcal{M}_{K,H,2}}_{\text{(ii)}} \\ &\quad + \underbrace{\sum_{k=1}^{K}\sum_{h=1}^{H} \big(\mathbb{E}_{\pi^*}[\iota_h^k(x_h,a_h)\,|\,x_1 = x_1^k] - \iota_h^k(x_h^k,a_h^k)\big)}_{\text{(iii)}}. \end{aligned}$$

Term (i). By Lemma 3.3 and the policy improvement step in Line 6 of Algorithm 1, we have

$$
\begin{aligned}
&\sum_{k=1}^{K}\sum_{h=1}^{H}\mathbb{E}_{\pi^*}\big[\langle Q_h^k(x_h,\cdot),\pi_h^*(\cdot\,|\,x_h)-\pi_h^k(\cdot\,|\,x_h)\rangle\,\big|\,x_1=x_1^k\big]\\
&\le\sum_{k=1}^{K}\sum_{h=1}^{H}\Big(\alpha H^2/2+\alpha^{-1}\cdot\mathbb{E}_{\pi^*}\big[D_{\mathrm{KL}}\big(\pi_h^*(\cdot\,|\,x_h)\,\big\|\,\pi_h^k(\cdot\,|\,x_h)\big)-D_{\mathrm{KL}}\big(\pi_h^*(\cdot\,|\,x_h)\,\big\|\,\pi_h^{k+1}(\cdot\,|\,x_h)\big)\,\big|\,x_1=x_1^k\big]\Big)\\
&\le\alpha H^3K/2+\alpha^{-1}\cdot\sum_{h=1}^{H}\mathbb{E}_{\pi^*}\big[D_{\mathrm{KL}}\big(\pi_h^*(\cdot\,|\,x_h)\,\big\|\,\pi_h^1(\cdot\,|\,x_h)\big)\,\big|\,x_1=x_1^k\big]\\
&\le\alpha H^3K/2+\alpha^{-1}H\cdot\log|\mathcal{A}|.
\end{aligned}
\tag{C.1}
$$

Here the second last inequality follows from the fact that the KL-divergence is nonnegative. Also, the last inequality follows from the initialization of the policy and Q-function in Line 1 of Algorithm 1, which implies that $\pi_h^1(\cdot\,|\,x_h)$ is a uniform distribution on $\mathcal{A}$ and hence

$$
\begin{aligned}
D_{\mathrm{KL}}\big(\pi_h^*(\cdot\,|\,x_h)\,\big\|\,\pi_h^1(\cdot\,|\,x_h)\big)&=\sum_{a\in\mathcal{A}}\pi_h^*(a\,|\,x_h)\cdot\log\big(|\mathcal{A}|\cdot\pi_h^*(a\,|\,x_h)\big)\\
&=\log|\mathcal{A}|+\sum_{a\in\mathcal{A}}\pi_h^*(a\,|\,x_h)\cdot\log\big(\pi_h^*(a\,|\,x_h)\big)\le\log|\mathcal{A}|.
\end{aligned}
$$

Here the inequality follows from the fact that the entropy of $\pi_h^*(\cdot\,|\,x_h)$ is nonnegative. Thus, setting $\alpha=\sqrt{2\log|\mathcal{A}|/(HT)}$ in Line 6 of Algorithm 1, by (C.1) we obtain

$$
\sum_{k=1}^{K}\sum_{h=1}^{H}\mathbb{E}_{\pi^*}\big[\langle Q_h^k(x_h,\cdot),\pi_h^*(\cdot\,|\,x_h)-\pi_h^k(\cdot\,|\,x_h)\rangle\,\big|\,x_1=x_1^k\big]\le\sqrt{2H^3T\cdot\log|\mathcal{A}|},
\tag{C.2}
$$

where $T=HK$.

Term (ii). Recall that the martingale differences $D_{k,h,1}$ and $D_{k,h,2}$ defined in (B.11) take the following forms,

$$
\begin{aligned}
D_{k,h,1}&=\big(\mathbb{J}_{k,h}(Q_h^k-Q_h^{\pi^k,k})\big)(x_h^k)-(Q_h^k-Q_h^{\pi^k,k})(x_h^k,a_h^k),\\
D_{k,h,2}&=\big(\mathbb{P}_h(V_{h+1}^k-V_{h+1}^{\pi^k,k})\big)(x_h^k,a_h^k)-(V_{h+1}^k-V_{h+1}^{\pi^k,k})(x_{h+1}^k).
\end{aligned}
$$

By the truncation of Q_h^k to the range $[0,H-h+1]$ in (3.5), we have

$$
Q_h^k,Q_h^{\pi^k,k},V_{h+1}^k,V_{h+1}^{\pi^k,k}\in[0,H],
$$

which implies that $|D_{k,h,1}|\le 2H$ and $|D_{k,h,2}|\le 2H$ for any $(k,h)\in[K]\times[H]$. Therefore, applying the Azuma-Hoeffding inequality to the martingale defined in (B.14), we obtain

$$
P\big(|\mathcal{M}_{K,H,2}|>t\big)\le 2\exp\bigg(\frac{-t^2}{16H^2T}\bigg)
$$

for any $t>0$. Setting $t=\sqrt{16H^2T\cdot\log(4/\zeta)}$ with $\zeta\in(0,1]$, we obtain

$$
|\mathcal{M}_{K,H,2}|\le\sqrt{16H^2T\cdot\log(4/\zeta)}
\tag{C.3}
$$

with probability at least $1-\zeta/2$, where $T=HK$.

Term (iii). As is shown in Lemma 4.3, it holds with probability at least $1-\zeta/2$ that

$$
-2\Gamma_h^k(x,a)\le\iota_h^k(x,a)\le 0
\tag{C.4}
$$

for any $(k,h)\in[K]\times[H]$ and $(x,a)\in\mathcal{S}\times\mathcal{A}$. Meanwhile, by the definitions of ι_h^k and Q_h^k in (4.1) and (3.5), respectively, we have that $|\iota_h^k(x,a)|\le 2H$, which together with (C.4) implies

$$
-\iota_h^k(x,a)\le 2\min\{H,\Gamma_h^k(x,a)\}
$$

for any $(k, h) \in [K] \times [H]$ and $(x, a) \in \mathcal{S} \times \mathcal{A}$ with probability at least $1 - \zeta/2$. Hence, we obtain

$$\sum_{k=1}^{K} \sum_{h=1}^{H} \bigl(\mathbb{E}_{\pi^*}[\iota_h^k(x_h, a_h) \,|\, x_1 = x_1^k] - \iota_h^k(x_h^k, a_h^k)\bigr) \le 2 \sum_{k=1}^{K} \sum_{h=1}^{H} \min\{H, \Gamma_h^k(x_h^k, a_h^k)\} \tag{C.5}$$

with probability at least $1 - \zeta/2$. By the definition of Γ_h^k in (3.6), we have

$$\sum_{k=1}^{K} \sum_{h=1}^{H} \min\{H, \Gamma_h^k(x_h^k, a_h^k)\} = \beta \cdot \sum_{h=1}^{H} \sum_{k=1}^{K} \min\Bigl\{H/\beta, \sqrt{\phi_h^k(x_h^k, a_h^k)^\top (\Lambda_h^k)^{-1} \phi_h^k(x_h^k, a_h^k)}\Bigr\}. \tag{C.6}$$

Recall that we set

$$\beta = C\sqrt{dH^2 \cdot \log(dT/\zeta)} \tag{C.7}$$

with $C > 1$ being an absolute constant, which implies that $H \le \beta$. Thus, (C.6) implies

$$\sum_{k=1}^{K} \sum_{h=1}^{H} \min\{H, \Gamma_h^k(x_h^k, a_h^k)\} \le \beta \cdot \sum_{h=1}^{H} \sum_{k=1}^{K} \min\Bigl\{1, \sqrt{\phi_h^k(x_h^k, a_h^k)^\top (\Lambda_h^k)^{-1} \phi_h^k(x_h^k, a_h^k)}\Bigr\}. \tag{C.8}$$

By Lemma D.3 and the definition of Λ_h^k in (3.6), we obtain

$$\sum_{k=1}^{K} \min\{1, \phi_h^k(x_h^k, a_h^k)^\top (\Lambda_h^k)^{-1} \phi_h^k(x_h^k, a_h^k)\} \le 2 \log\biggl(\frac{\det(\Lambda_h^{K+1})}{\det(\Lambda_h^1)}\biggr) \tag{C.9}$$

for any $h \in [H]$, where $\Lambda_h^1 = \lambda \cdot I$ and $\Lambda_h^{K+1} \in \mathcal{F}_{K,H,2}$ by Definition 4.1. Moreover, Assumption 2.1 implies

$$\|\phi_h^k(x, a)\|_2 \le \sqrt{d}H$$

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

$$\Lambda_h^{K+1} = \sum_{k=1}^{K} \phi_h^k(x_h^k, a_h^k)\phi_h^k(x_h^k, a_h^k)^\top + \lambda \cdot I \preceq (dH^2K + \lambda) \cdot I$$

for any $h \in [H]$. As we set $\lambda = 1$, it holds for any $h \in [H]$ that

$$2 \log\biggl(\frac{\det(\Lambda_h^{K+1})}{\det(\Lambda_h^1)}\biggr) \le 2d \cdot \log\bigl((dH^2K + \lambda)/\lambda\bigr) \le 4d \cdot \log(dHT). \tag{C.10}$$

Combining (C.8)-(C.10) and the Cauchy-Schwarz inequality, we obtain

$$\begin{aligned}
\sum_{k=1}^{K} \sum_{h=1}^{H} \min\{H, \Gamma_h^k(x_h^k, a_h^k)\} &\le \beta \cdot \sum_{h=1}^{H} \Bigl(K \cdot \sum_{k=1}^{K} \min\{1, \phi_h^k(x_h^k, a_h^k)^\top (\Lambda_h^k)^{-1} \phi_h^k(x_h^k, a_h^k)\}\Bigr)^{1/2} \\
&\le \beta \cdot \sum_{h=1}^{H} \sqrt{K} \cdot \biggl(2 \log\biggl(\frac{\det(\Lambda_h^{K+1})}{\det(\Lambda_h^1)}\biggr)\biggr)^{1/2} \\
&\le 2\beta\sqrt{dH^2K \cdot \log(dHT)}.
\end{aligned} \tag{C.11}$$

By (C.5), (C.7), and (C.11), it holds with probability at least $1 - \zeta/2$ that

$$\begin{aligned}
&\sum_{k=1}^{K} \sum_{h=1}^{H} \bigl(\mathbb{E}_{\pi^*}[\iota_h^k(x_h, a_h) \,|\, x_1 = x_1^k] - \iota_h^k(x_h^k, a_h^k)\bigr) \\
&\qquad \le 4\beta\sqrt{dH^2K \cdot \log(dHT)} \le 8C\sqrt{d^2H^3T} \cdot \log(dT/\zeta),
\end{aligned} \tag{C.12}$$

where $C > 1$ is an absolute constant, $\zeta \in (0, 1]$, and $T = HK$.

Plugging the upper bounds of terms (i)-(iii) in (C.2), (C.3), and (C.12), respectively, into (4.3) of Lemma 4.2, we obtain

$$\text{Regret}(T) \le C' \sqrt{d^2 H^3 T} \cdot \log(dT/\zeta)$$

with probability at least $1-\zeta$, where $C' > 0$ is an absolute constant. Here we use the fact that $\log|\mathcal{A}| = O(d^2 \cdot [\log(dT/\zeta)]^2)$ in (C.2) and (C.12). Therefore, we conclude the proof of Theorem 3.1. □

D. Supporting Lemmas

In this section, we present the supporting lemmas.

Lemma D.1. Let $\lambda = 1$ in (3.5) and Line 12 of Algorithm 1. For any $\zeta \in (0, 1]$, the event $\mathcal{E}$ that, for any $(k, h) \in [K] \times [H]$,

$$\Big\| \sum_{\tau=1}^{k-1} \phi_h^\tau(x_h^\tau, a_h^\tau) \cdot \big(V_{h+1}^\tau(x_{h+1}^\tau) - (\mathbb{P}_h V_{h+1}^\tau)(x_h^\tau, a_h^\tau)\big) \Big\|_{(\Lambda_h^k)^{-1}} \le C'' \sqrt{dH^2 \cdot \log(dT/\zeta)} \tag{D.1}$$

happens with probability at least $1 - \zeta/2$, where $C'' > 0$ is an absolute constant.

Proof. By the definition of the filtration $\{\mathcal{F}_{k,h,m}\}_{(k,h,m)\in[K]\times[H]\times[2]}$ in Definition 4.1 and the Markov property, we have

$$\mathbb{E}[V_{h+1}^\tau(x_{h+1}^\tau) \,|\, \mathcal{F}_{\tau,h,1}] = (\mathbb{P}_h V_{h+1}^\tau)(x_h^\tau, a_h^\tau). \tag{D.2}$$

Conditioning on $\mathcal{F}_{\tau,h,1}$, the only randomness comes from x_{h+1}^τ, while V_{h+1}^τ is a deterministic function. To see this, note that V_{h+1}^τ is determined by Q_{h+1}^τ and π_{h+1}^τ, which are further determined by the historical data in $\mathcal{F}_{\tau,h,1}$. We define

$$\eta_{\tau,h} = V_{h+1}^\tau(x_{h+1}^\tau) - (\mathbb{P}_h V_{h+1}^\tau)(x_h^\tau, a_h^\tau).$$

By (D.2), conditioning on $\mathcal{F}_{\tau,h,1}$, $\eta_{\tau,h}$ is a zero-mean random variable. Moreover, as $V_{h+1}^\tau \in [0, H]$, conditioning on $\mathcal{F}_{\tau,h,1}$, $\eta_{\tau,h}$ is an $H/2$-sub-Gaussian random variable, which is defined in (D.5) of Lemma D.2. Also, $\eta_{\tau,h}$ is $\mathcal{F}_{k,h,2}$-measurable, as $x_{h+1}^\tau \in \mathcal{F}_{\tau,h,2}$ for any $\tau \in [k-1]$. Hence, for any fixed $h \in [H]$, by Lemma D.2, it holds with probability at least $1 - \zeta/(2H)$ that

$$\begin{aligned}
&\Big\| \sum_{\tau=1}^{k-1} \phi_h^\tau(x_h^\tau, a_h^\tau) \cdot \big(V_{h+1}^\tau(x_{h+1}^\tau) - (\mathbb{P}_h V_{h+1}^\tau)(x_h^\tau, a_h^\tau)\big) \Big\|_{(\Lambda_h^k)^{-1}}^2 \\
&\quad \le H^2/2 \cdot \Big(\log\big(\det(\Lambda_h^k)^{1/2} \det(\lambda \cdot I)^{-1/2}\big) + \log(2H/\zeta) \Big)
\end{aligned} \tag{D.3}$$

for any $k \in [K]$. To upper bound $\det(\Lambda_h^k)$ in (D.3), recall that Λ_h^k is defined by

$$\Lambda_h^k = \sum_{\tau=1}^{k-1} \phi_h^\tau(x_h^\tau, a_h^\tau) \phi_h^\tau(x_h^\tau, a_h^\tau)^\top + \lambda \cdot I.$$

By the triangle inequality, the spectral norm of Λ_h^k is upper bounded as

$$\|\Lambda_h^k\|_2 \le \lambda + \sum_{\tau=1}^{k-1} \|\phi_h^\tau(x_h^\tau, a_h^\tau)\|_2^2 \le \lambda + dH^2 K.$$

Here the last inequality follows from Assumption 2.1, which implies

$$\sup_{(x,a)\in\mathcal{S}\times\mathcal{A}} \Big\| \int_{\mathcal{S}} \psi(x, a, x') \cdot V(x') \mathrm{d}x' \Big\|_2 \le \sqrt{d} H$$

for any $V : \mathcal{S} \to [0, H]$. Hence, $\det(\Lambda_h^k)$ in (D.3) is upper bounded as

$$\det(\Lambda_h^k) \le \|\Lambda_h^k\|_2^d \le (\lambda + dH^2 K)^d. \tag{D.4}$$

Moreover, setting $\lambda = 1$, combining (D.3) and (D.4), and applying the union bound for any $h \in [H]$, we obtain that, with probability at least $1 - \zeta/2$,

$$\begin{aligned}&\Big\|\sum_{\tau=1}^{k-1} \phi_h^\tau(x_h^\tau, a_h^\tau) \cdot \big(V_{h+1}^\tau(x_{h+1}^\tau) - (\mathbb{P}_h V_{h+1}^\tau)(x_h^\tau, a_h^\tau)\big)\Big\|^2_{(\Lambda_h^k)^{-1}}\\ &\quad\le H^2/2 \cdot \Big(d/2 \cdot \log\big((\lambda + dH^2K)/\lambda\big) + \log(2H/\zeta)\Big) \le C''^2 dH^2 \cdot \log(dT/\zeta)\end{aligned}$$

for any $(k, h) \in [K] \times [H]$, where $C'' > 0$ is an absolute constant. Thus, we conclude the proof of Lemma D.1. □

Lemma D.2 (Concentration of Self-Normalized Process (Abbasi-Yadkori et al., 2011))**.** Let $\{\widetilde{\mathcal{F}}_t\}_{t=0}^\infty$ be a filtration and $\{\eta_t\}_{t=1}^\infty$ be an $\mathbb{R}$-valued stochastic process such that η_t is $\widetilde{\mathcal{F}}_t$-measurable for any $t \ge 0$. Moreover, we assume that, for any $t \ge 0$, conditioning on $\widetilde{\mathcal{F}}_t$, η_t is a zero-mean and σ-sub-Gaussian random variable with the variance proxy $\sigma^2 > 0$, that is,

$$\mathbb{E}[e^{\lambda\eta_t} \,|\, \widetilde{\mathcal{F}}_t] \le e^{\lambda^2\sigma^2/2} \tag{D.5}$$

for any $\lambda \in \mathbb{R}$. Let $\{X_t\}_{t=1}^\infty$ be an $\mathbb{R}^d$-valued stochastic process such that X_t is $\widetilde{\mathcal{F}}_t$-measurable for any $t \ge 0$. Also, let $Y \in \mathbb{R}^{d\times d}$ be a deterministic and positive-definite matrix. For any $t \ge 0$, we define

$$\overline{Y}_t = Y + \sum_{s=1}^t X_s X_s^\top, \quad S_t = \sum_{s=1}^t \eta_s \cdot X_s.$$

For any $\delta > 0$, it holds with probability at least $1 - \delta$ that

$$\|S_t\|^2_{\overline{Y}_t^{-1}} \le 2\sigma^2 \cdot \log\bigg(\frac{\det(\overline{Y}_t)^{1/2}\det(Y)^{-1/2}}{\delta}\bigg)$$

for any $t \ge 0$.

Proof. See Theorem 1 of (Abbasi-Yadkori et al., 2011) for a detailed proof. □

Lemma D.3 (Elliptical Potential Lemma (Dani et al., 2008; Rusmevichientong & Tsitsiklis, 2010; Chu et al., 2011; Abbasi-Yadkori et al., 2011; Jin et al., 2019))**.** Let $\{\phi_t\}_{t=1}^\infty$ be an $\mathbb{R}^d$-valued sequence. Meanwhile, let $\Lambda_0 \in \mathbb{R}^{d\times d}$ be a positive-definite matrix and $\Lambda_t = \Lambda_0 + \sum_{j=1}^{t-1} \phi_j\phi_j^\top$. It holds for any $t \in \mathbb{Z}_+$ that

$$\sum_{j=1}^t \min\{1, \phi_j^\top \Lambda_j^{-1}\phi_j\} \le 2\log\bigg(\frac{\det(\Lambda_{t+1})}{\det(\Lambda_1)}\bigg).$$

Moreover, assuming that $\|\phi_j\|_2 \le 1$ for any $j \in \mathbb{Z}_+$ and $\lambda_{\min}(\Lambda_0) \ge 1$, it holds for any $t \in \mathbb{Z}_+$ that

$$\log\bigg(\frac{\det(\Lambda_{t+1})}{\det(\Lambda_1)}\bigg) \le \sum_{j=1}^t \phi_j^\top \Lambda_j^{-1}\phi_j \le 2\log\bigg(\frac{\det(\Lambda_{t+1})}{\det(\Lambda_1)}\bigg).$$

Proof. See Lemma 11 of (Abbasi-Yadkori et al., 2011) for a detailed proof. □