Rate-Optimal Policy Optimization for Linear Markov Decision Processes

Uri Sherman
Blavatnik School of Computer Science
Tel Aviv University
urisherman@mail.tau.ac.il

Alon Cohen
School of Electrical Engineering
Tel Aviv University and Google Research
alonco@tauex.tau.ac.il

Tomer Koren
Blavatnik School of Computer Science
Tel Aviv University and Google Research
tkoren@tauex.tau.ac.il

Yishay Mansour
Blavatnik School of Computer Science
Tel Aviv University and Google Research
mansour.yishay@gmail.com

Abstract

We study regret minimization in online episodic linear Markov Decision Processes, and propose a policy optimization algorithm that is computationally efficient, and obtains rate optimal $\widetilde{O}(\sqrt{K})$ regret where K denotes the number of episodes. Our work is the first to establish the optimal rate (in terms of K) of convergence in the stochastic setting with bandit feedback using a policy optimization based approach, and the first to establish the optimal rate in the adversarial setup with full information feedback, for which no algorithm with an optimal rate guarantee was previously known.

1 Introduction

Policy Optimization (PO) algorithms are a class of methods in Reinforcement Learning (RL; Sutton and Barto, 2018, Mannor et al., 2022) where the agent's policy is iteratively updated according to the (possibly preconditioned) gradient of the value function w.r.t. policy parameters. From a theoretical perspective, framing the optimization process as one that follows Mirror Descent Nemirovskij and Yudin [1983], Beck and Teboulle [2003] updates leads to strong online guarantees that go beyond stationary or stochastic rewards, and apply more generally for *any* (possibly adversarial) reward sequence Shani et al. [2020], Luo et al. [2021]. Furthermore, PO methods are easy to implement in practice and (perhaps, one could say, somewhat in line with theory) exhibit favorable robustness properties when applied to real world problems ranging from robotics Levine and Koltun [2013], Schulman et al. [2015], Haarnoja et al. [2018], computer games Schulman et al. [2017], and more recently training of large language models Ouyang et al. [2022].

Notwithstanding their popularity and theoretical appeal, current results Agarwal et al. [2020], Zanette et al. [2021], Liu et al. [2023b], Zhong and Zhang [2023] in the function approximation setting with linear MDP Jin et al. [2020] assumptions fall short of establishing the optimal dependence on the number of episodes K; arguably, the most important problem parameter.

In this work, we establish that an optimistic variant of the classic natural policy gradient [1] (NPG; Kakade, 2001) obtains the optimal (up to logarithmic factors) $\widetilde{O}(\sqrt{K})$ regret when combined with a short reward free warmup period and a suitable bonus update schedule. Our results hold for

[1] To be precise, our algorithm is the classic NPG with softmax parametrization equipped with an optimistic linear function approximation routine for action-value estimates.

17th European Workshop on Reinforcement Learning (EWRL 2024).

adversarial losses when the learner is given full information feedback, and for stochastic losses when given bandit feedback. Thus our algorithm is also the first (and currently, the only) method to obtain rate optimal regret (be it by PO or any other approach) for adversarial losses with full feedback in the linear MDP setup.

1.1 Summary of contributions

We consider online learning in a finite horizon episodic linear MDP, where an agent interacts with the environment over the course of K episodes. In each episode $k \in [K]$, the agent interacts with the MDP $\mathcal{M}_k = (\mathcal{S}, \mathcal{A}, H, \{\mathbb{P}_h\}, \{\ell_h^k\}, s_1)$, that shares all elements with MDPs of other episodes except for the loss functions. Our central structural assumption is that the dynamics and losses are *linear*; that there exist feature embeddings $\phi, \psi_1, \ldots, \psi_H$ such that $\mathbb{P}_h(s'|s,a) = \phi(s,a)^\top \psi_h(s')$, and $\mathbb{E}[\ell_h^k(s,a)|s,a] = \phi(s,a)^\top g_{h,k}$, for some $g_{h,k} \in \mathbb{R}^d$. The objective of the agent is to minimize her *regret*, defined by the sum of value functions (namely, the expected cumulative loss) of the agent minus the sum of values of the best policy in hindsight.

Our main contribution in this paper is a computationally efficient policy optimization algorithm (see Algorithm 1), that guarantees an $\widetilde{O}(\sqrt{K})$ regret bound under either of the following two conditions:

- For any (possibly adversarial) loss sequence $\{g_{h,k}\}$, when given *full feedback*, meaning the agent observes $g_{1,k}, \ldots, g_{H,k}$ after each episode k.
- For stationary losses, namely $g_{h,k} = g_h \forall k$, when given noisy *bandit feedback*, meaning the agent observes only $l_h^k := \ell_h^k(s_h^k, a_h^k)$, and it holds that $l_h^k \in [-1, 1]$ and that the expected value of l_h^k conditioned on past interactions is $\phi(s_h^k, a_h^k)^\top g_h$.

1.2 Overview of techniques

The difficulty encountered in recent attempts (Liu et al., 2023b, Zhong and Zhang, 2023, and to an extent also in Sherman et al., 2023) towards establishing the rate optimal $\sqrt{K}$ stems from the need to control the capacity of the policy class explored by the optimization process. Since the dynamics in linear MDPs cannot be estimated pointwise, the estimation procedure of the action-value function involves a linear regression sub-routine where the dependent variable is given by the value function estimate from the previous timestep, which depends on past rollouts in a way that breaks the martingale structure. Thus, to establish concentration, an additional uniform convergence argument is required in which the capacity of the policy class plays a central role.

To illustrate, let us consider a simplified, non-optimistic estimation routine with non-zero immediate losses only at step H, and let $\left\{(s_h^i, a_h^i, s_{h+1}^i)\right\}_{i=1}^{k-1}$ denote a dataset of past agent transitions, and $\widehat{V}_{h+1}^k$ the value function estimated in step $h+1$. Then the estimation step on time h is given by:

$$\widehat{v}_h^k = \operatorname*{arg\,min}_{v \in \mathbb{R}^d} \left\{ \sum_{i=1}^{k-1} \left(\phi(s_h^i, a_h^i)^\top v - \widehat{V}_{h+1}^k(s_{h+1}^i) \right)^2 \right\},$$
$$\widehat{Q}_h^k(s,a) = \text{truncate}\left[\widehat{\mathbb{P}}_h^k \widehat{V}_{h+1}^k(s,a) := \phi(s,a)^\top \widehat{v}_h^k \right],$$

where $\text{truncate}[\cdot]$ denotes some form of clipping used to keep the estimated action-values in reasonable range (e.g., $[-H, H]$). Notably, $\widehat{V}_{h+1}^k$ was itself estimated using the same procedure in the previous backward induction step, combined with an expectation given by the agent's policy:

$$\widehat{V}_{h+1}^k(s) = \left\langle \pi_{h+1}^k(\cdot|s), \widehat{Q}_{h+1}^k(s,\cdot) \right\rangle,$$

which means the estimated quantity is a random variable that depends on *all* past trajectories through the agent's policy. Hence, to establish a least squares concentration bound, the common technique (originally proposed in this context in the work of Jin et al., 2020) dictates arguing uniform convergence over the class of all possible value functions $\widehat{V}_{h+1}^k$ explored by the learner. Further, the capacity of the class of learner value functions is inevitably tied to the capacity of the learner's

policies, and when employing mirror descent updates, these are parameterized by the sum of past action-value functions:

$$\pi^k_{h+1}(a|s) \propto \exp\left(-\eta \sum_{i=1}^{k-1} \widehat{Q}^i_{h+1}(s,a)\right).$$

Now, the problem is that the truncation of the Q-functions implies the above expression does not admit a low dimensional (independent of k) representation, and thus leads to the agent's policy and value classes having prohibitively large covering number.

The main component of our approach is to employ a reward free warmup period, that eventually allows to forgo the truncation of the action value function, thereby reducing the policy class capacity. Indeed, if the action-value functions were not truncated, the policy parameterization could be made effectively independent (up to log factors) of k, as the sum of Q-functions will "collapse" into a single d dimensional parameter of larger norm:

$$\pi^k_{h+1}(a|s) \propto \exp\left(\phi(s,a)^\top \theta^k_{h+1}\right),$$

where $\theta^k_{h+1} = -\eta \sum_{i=1}^{k-1} \widehat{v}^i_{h+1}$. In order to remove the truncations, we observe they are actively involved only in those regions of the state space that are poorly explored; indeed, assume the least squares errors are boudned as:

$$\left|\widehat{\mathbb{P}}^k_h \widehat{V}^k_{h+1}(s,a) - \mathbb{P}_h \widehat{V}^k_{h+1}(s,a)\right| \le \beta \left\|\phi(s,a)\right\|_{\Lambda^{-1}_{k,h}},$$

where $\Lambda_{k,h} \coloneqq I + \sum_i \phi(s^i_h, a^i_h)\phi(s^i_h, a^i_h)^\top$ for some β that depends (among other quantities) on $\max_{s'} \widehat{V}^k_{h+1}(s')$, and assume we have already shown that $\widehat{V}^k_{h+1}(s') \lesssim H$ for all s'. Then as long as $\phi(s,a)$ points in a well explored direction in the state-action space — concretely one where $\|\phi(s,a)\|_{\Lambda^{-1}_{k,h}} \le 1/(\beta H)$ — we will get that:

$$\begin{aligned}
&\widehat{Q}^k_h(s,a) = \mathbb{P}_h \widehat{V}^k_{h+1}(s,a) \pm \frac{1}{H} \\
&\quad\implies \left|\widehat{Q}^k_h(s,a)\right| \le \left|\mathbb{P}_h \widehat{V}^k_{h+1}(s,a)\right| + \frac{1}{H} \lesssim H + \frac{1}{H}.
\end{aligned}$$

Thus, forgoing truncations and if all directions were well explored, we would get $\left\|\widehat{V}^k_h\right\|_\infty \le \left\|\widehat{V}^k_{h+1}\right\|_\infty + \frac{1}{H}$, and continuing inductively we accumulate errors across the horizon in an additive manner; $\left\|\widehat{V}^k_h\right\|_\infty \lesssim H + (H-h)/H$. Now, while we cannot ensure sufficient exploration in *all* directions, we can in fact ensure it in "most" directions (those which are reachable w.p. $\gtrsim 1/\sqrt{K}$) using a properly tuned reward free warmup phase, which is based on the algorithm developed in Wagenmaker et al. [2022b]. The technical argument roughly follows the above intuition, carefully controlling the least squares errors through an inductive argument. This way, we establish the estimated value functions remain in the low capacity function class, for which we have a suitable uniform concentration bound.

1.3 Additional related work

Linear MDPs with adversarial costs. Most relevant to our paper is the recent work of Zhong and Zhang [2023], who consider the same adversarial setup as ours and establish a $\widetilde{O}(K^{3/4})$ regret bound, using an optimistic policy optimization framework similar to ours, but with an additional batching mechanism. Several recent papers consider the more general setting consisting of *adversarial costs and bandit-feedback*. Neu and Olkhovskaya [2021] obtain a rate optimal regret bound assuming *known dynamics* and a certain exploratory condition. In the general setting without additional assumptions, Luo et al. [2021] was the first to establish a sublinear regret bound. The followup works of Dai et al. [2023], Sherman et al. [2023] obtain respectively, $\widetilde{O}(K^{8/9}), \widetilde{O}(K^{6/7})$ regret, and Kong et al. [2023] obtain $\widetilde{O}(K^{4/5} + 1/\lambda^\star_{\min})$ (here, $\lambda^\star_{\min}$ denotes the minimum eigenvalue of the best exploratory policy's 2nd moment matrix) albeit with a computationally inefficient algorithm. Finally, a very recent preprint [Liu et al., 2023a] establishes the current state-of-the-art results for this setting; $\widetilde{O}(K^{3/4})$ with a computationally efficient algorithm, and $\widetilde{O}(\sqrt{K})$ with a computationally inefficient one.

Policy optimization in tabular and linear MDPs. Most of the currently published works that consider policy optimization algorithms in the learning setup that necessitates exploration were mentioned in the introduction. In particular, the work of Liu et al. [2023b] considers the same stochastic setup as ours and obtains a $\widetilde{O}(1/\epsilon^3)$ sample complexity for a different variant of the optimistic NPG algorithm. Many recent works [e.g., Bhandari and Russo, 2019, Liu et al., 2019, Agarwal et al., 2021, Lan, 2022, Xiao, 2022, Yuan et al., 2022] study convergence properties of policy optimization methods from a pure optimization perspective or subject to exploratory assumptions; in this setup, exploration need not be handled algorithmically, and rates much faster than $O(\sqrt{K})$ regret are achievable when access to exact value function gradients is granted.

RL with function approximation The study of MDPs with linear structure in the form we adopt here was initiated with the works of Yang and Wang [2019, 2020], Jin et al. [2020], and has lead to an abundance of papers considering algorithmic approaches to various problem setups [e.g., Zanette et al., 2020, Wei et al., 2021, Wagenmaker et al., 2022b]. The linear mixture MDP Modi et al. [2020], Ayoub et al. [2020], Zhou et al. [2021a,b] is a different model that in general is incomparable with the linear MDP Zhou et al. [2021b]. There is also a rich line of works studying statistical properties of RL with more general function approximation [e.g., Jiang et al., 2017, Jin et al., 2021, Du et al., 2021], although these usually do not provide computationally efficient algorithms.

2 Preliminaries

Episodic MDPs. A finite horizon episodic MDP is defined by the tuple $\mathcal{M} = (\mathcal{S}, \mathcal{A}, H, \mathbb{P}, \ell, s_1)$, where $\mathcal{S}$ denotes the state space, $\mathcal{A}$ the action set, $H \in \mathbb{Z}_+$ the length of the horizon, $\mathbb{P} = \{\mathbb{P}_h\}_{h\in[H]}$ the time dependent transition function, $\ell = \{\ell_h\}_{h\in[H]}$ a sequence of loss functions, and $s_1 \in \mathcal{S}$ the initial state that we assume to be fixed w.l.o.g. The transition density given the agent is at state $s \in \mathcal{S}$ at time h and takes action a is given by $\mathbb{P}_h(\cdot|s,a) \in \Delta(\mathcal{S})$. After the agent takes an action on the last time step H, she transitions to a fixed terminal state $s_{H+1} \in \mathcal{S}$ and the episode terminates immediately. We assume the state space $\mathcal{S}$ is a (possibly infinite) measurable space, and that the action set $\mathcal{A}$ is finite with $A \coloneqq |\mathcal{A}|$. A policy is defined by a mapping $\pi\colon \mathcal{S}\times[H] \to \Delta(\mathcal{A})$, where $\Delta(\mathcal{A})$ denotes the probability simplex over the action set $\mathcal{A}$. We let $\pi_h(\cdot|s) \in \Delta(\mathcal{A})$ denote the distribution over actions given by π at s, h. Finally, we use the convention that for any function $V\colon \mathcal{S}\to\mathbb{R}$, we interpret $\mathbb{P}_h V\colon \mathcal{S}\times\mathcal{A}\to\mathbb{R}$ as the result of applying the conditional expectation operator $\mathbb{P}_h$; $\mathbb{P}_h V(s,a) \coloneqq \mathbb{E}_{s'\sim\mathbb{P}_h(\cdot|s,a)} V(s')$.

Episodic Linear MDPs. Our central structural assumption is that the learner interacts with a *linear MDP* Jin et al. [2020], defined next.

Definition 1 (Linear MDP)**.** An MDP $\mathcal{M} = (\mathcal{S}, \mathcal{A}, H, \mathbb{P}, \ell, s_1)$ is a linear MDP if the following holds. There is a feature mapping $\phi\colon \mathcal{S}\times\mathcal{A}\to\mathbb{R}^d$ that is **known** to the learner, and H signed vector-valued measures $\psi_h\colon \mathcal{S}\to\mathbb{R}^d$ that are **unknown**, such that for all $h, s, a, s' \in [H-1]\times\mathcal{S}\times\mathcal{A}\times\mathcal{S}$:

$$\mathbb{P}_h(s'|s,a) = \phi(s,a)^\top \psi_h(s'). \tag{1}$$

W.l.o.g., we assume $\|\phi(s,a)\| \le 1$ for all s, a, and that for any measurable function $f\colon \mathcal{S}\to\mathbb{R}$ with $\|f\|_\infty \le 1$, it holds that $\left\|\int \psi_h(s')f(s')\mathrm{d}s'\right\| \le \sqrt{d}$ for all $h \in [H]$. In addition, for all s, a, h:

$$\ell_h(s,a) = \phi(s,a)^\top g_h, \tag{2}$$

where $\{g_h\} \subset \mathbb{R}^d$. W.l.o.g., we assume $\left|\phi(s,a)^\top g_h\right| \le 1$ for all s, a, h, and $\|g_h\| \le \sqrt{d}$ for all h.

Problem setup. We consider linear MDPs in two setups; *adversarial* and *stochastic*. In the adversarial setup defined formally next, we assume the agent interacts with a sequence of $K \ge 1$ MDPs over the course of K episodes that share all elements other than the loss functions, which may change adversarially.

Assumption 1 (Adversarial Linear MDP with full-feedback)**.** The learner interacts with a sequence of MDPs $\left\{\mathcal{M}^k\right\}_{k=1}^K$, $\mathcal{M}^k = (\mathcal{S}, \mathcal{A}, H, \mathbb{P}, \ell^k, s_1)$ that share all elements other than the loss functions. Each MDP $\mathcal{M}^k$ is a linear MDP as per Definition 1. The feedback provided to the learner on episode k time step h is the low dimensional cost vector $g_{k,h} \in \mathbb{R}^d$, where $g_k = (g_{k,1}, \ldots, g_{k,H}) \in \mathbb{R}^{dH}$ is the d dimensional representation of $\ell^k = \left(\ell_1^k, \ldots, \ell_H^k\right)$.

In the stochastic setup, we assume the agent interacts with a single linear MDP over the course of $K \geq 1$ episodes, and receives only noisy *bandit*-feedback.

Assumption 2 (Stochastic Linear MDP with bandit-feedback)**.** In each episode, the learner interacts with the same linear MDP $\mathcal{M} = (\mathcal{S}, \mathcal{A}, H, \mathbb{P}, \ell, s_1)$. The feedback provided to the learner on episode k time step h is the random instantaneous loss $l_h^k := \ell_h^k(s_h^k, a_h^k)$, where s_h^k, a_h^k denote the state and action visited by the agent on episode k time step h. It holds that $\mathbb{E}\left[l_h^k \mid s_h^k, a_h^k, \left(l_h^{k'}, s_h^{k'}, a_h^{k'}\right)_{k'<k}\right] = \ell_h(s_h^k, a_h^k)$, and $\left|\ell_h^k(s_h^k, a_h^k)\right| \leq 1$ almost surely.

The pseudocode for learner environment interaction, encompassing both assumptions is provided below in Protocol 1. We make the following final notes with regards to the model we consider: (1) for any $s, a \in \mathcal{S} \times \mathcal{A}$, the agent may evaluate $\phi(s, a)$ in $O(1)$ time; (2) In the adversarial setup, we assume an oblivious and deterministic adversary. Specifically, that the sequence of loss functions is chosen in advance, before interaction begins.

Protocol 1 Learner-Environment Interaction

parameters: $(\mathcal{S}, \mathcal{A}, H, \mathbb{P}, \phi, s_1; K)$

Nature chooses $\begin{cases} \textit{Adv.:} & \{g_k\}_{k=1}^K \in \mathbb{R}^{dH}; \\ \textit{Stoch.:} & g \in \mathbb{R}^{dH}, \text{ and sets } g_k \equiv g \ \forall k \end{cases}$

for $k = 1, \ldots, K$ **do**
  agent decides on a policy π^k
  environment resets to $s_1^k = s_1$
  for $h = 1, \ldots, H$ **do**
    agent observes $s_h^k \in \mathcal{S}$
    agent chooses $a_h^k \sim \pi_h^k(\cdot|s_h^k)$
    agent incurs loss $\phi(s_h^k, a_h^k)^\top g_{k,h}$
    agent observes $\begin{cases} \textit{Full-feedback:} & g_{k,h} \\ \textit{Bandit-feedback:} & \ell_h^k(s_h^k, a_h^k) \end{cases}$
    environment transitions to $s_{h+1}^k \sim \mathbb{P}_h(\cdot|s, a)$
  end for
end for

Learning objective. The expected loss of a policy π when starting from state $s \in \mathcal{S}$ at time step $h \in [H]$ is given by the value function;

$$V_h^\pi(s; \ell) := \mathbb{E}\left[\sum_{t=h}^{H} \ell_t(s_t, a_t) \mid s_h = s, \pi, \ell\right], \tag{3}$$

where we use the extra $(; \ell)$ notation to emphasize the specific loss function considered. The expected loss conditioned on the agent taking action $a \in \mathcal{A}$ on time step h at s and then continuing with π is given by the action-value function;

$$Q_h^\pi(s, a; \ell) := \mathbb{E}\left[\sum_{t=h}^{H} \ell_t(s_t, a_t) \mid s_h = s, a_h = a, \pi, \ell\right]. \tag{4}$$

The value and action-value functions of a policy π in the MDP $\left(\mathcal{S}, \mathcal{A}, H, \mathbb{P}, \ell^k, s_1\right)$ associated with episode $k \in [K]$ are denoted by, respectively;

$$V_h^{k,\pi}(s) := V_h^\pi(s; \ell^k); \ Q_h^{k,\pi}(s, a) := Q_h^\pi(s, a; \ell^k), \tag{5}$$

where $V_h^\pi(s; \ell^k)$ and $Q_h^\pi(s, a; \ell^k)$ have been defined in Eqs. (3) and (4). For the sake of conciseness, we further define

$$V^{k,\pi} := V_1^{k,\pi}(s_1)$$

We let $\pi^\star$ denote the best policy in hindsight;

$$\pi^\star \coloneqq \arg\min_{\pi} \left\{ \sum_{k=1}^{K} V_1^{k,\pi}(s_1) \right\},$$

and seek to minimize the *pseudo regret* of the agent policy sequence $\pi^1, \ldots, \pi^K$;

$$\text{Regret} \coloneqq \sum_{k=1}^{K} V^{k,\pi^k} - V^{k,\pi^\star}. \tag{6}$$

Occupancy measures. We denote the occupancy measure of a policy π by

$$\mu_h^\pi(s,a) \coloneqq \Pr\left(s_h = s, a_h = a \mid \pi\right), \tag{7}$$

and additionally denote $\mu_h^k \coloneqq \mu_h^{\pi^k}$, and $\mu_h^\star \coloneqq \mu_h^{\pi^\star}$.

Additional notation. We let $\|\cdot\| = \|\cdot\|_2$ denote the standard Euclidean norm, and for a positive definite matrix $\Lambda \in \mathbb{R}^{d\times d}$, we let $\|v\|_\Lambda = \sqrt{v^\top \Lambda v}$ denote the weighted norm induced by Λ. Further, we let $\|\Lambda\| = \|\Lambda\|_{\text{op}} = \max_{v,\|v\|=1} v^\top \Lambda v$ denote the operator norm of Λ.

3 Algorithm and Main Result

In this section, we present Algorithm 1 and our main theorem providing its regret guarantees. At a high level, Algorithm 1 follows an optimistic policy optimization paradigm similar to Shani et al. [2020] in the tabular case and more recently Liu et al. [2023b], Zhong and Zhang [2023] in the linear MDP case. The important difference is the utilization of a pure exploration warmup period provided by Algorithm 2 (which we describe in more detail in Section 3.1), and the usage of *restricted* value functions. The restricted value functions, in contrast to truncated ones, take zero value outside the confidence state set.

The core property required from the warmup period is that the data it collects is sufficient to ensure a small error when using it in the least squares regression step of Algorithm 1. The degree to which the error should be small is determined by the multiplicative factor in the confidence bound for a single regression step (determined by the bonus parameter β along with other problem parameters), and the number of times we perform this step (H; the length of the horizon). The analysis leads to the following definition for the "known" states set of step h:

$$\mathcal{Z}_h \coloneqq \left\{ s \in \mathcal{S} \mid \forall a, \|\phi(s,a)\|_{\Lambda_{0,h}^{-1}} \le 1/(2\beta H) \right\}, \tag{8}$$

where $\Lambda_{0,h}$ denotes the warmup covariate matrix returned by Algorithm 2 for step h. The set $\mathcal{Z}_h$ contains the states for which we collected enough data, so that the least squares regression error when estimating their value can be well controlled without employing truncation.

On episode k, the standard optimistic estimates value function estimates are denoted $\widetilde{Q}_h^k, \widetilde{V}_h^k$, while their restricted counterparts are defined by:

$$\widetilde{Q}_h^{k;\circ}(s,a) = \mathbf{I}\left\{s \in \mathcal{Z}_h\right\} \widetilde{Q}_h^k(s,a),$$
$$\widetilde{V}_h^{k;\circ}(s) = \left\langle \widetilde{Q}_h^{k;\circ}(s,\cdot), \pi_h^k(\cdot|s) \right\rangle.$$

During the backward dynamic programming step, the estimate of the non-restricted action-value function $\widetilde{Q}_{h-1}^k$ then makes use of the least squares solution w.r.t. the restricted $\widetilde{V}_h^{k;\circ}$, which has a well bounded $\|\cdot\|_\infty$. Further, the warmup ensures the known state set $\mathcal{Z}_h$ is large enough so that we do not lose much by this restriction; concretely, that no policy has total occupancy larger than $O(\epsilon_{\text{cov}})$ outside the known states set.

The other important ingredient of Algorithm 1 is the epoch schedule in the updates of bonus functions $\hat{b}_h^k$, determined by the determinant of the covariate matrices $\Lambda_{k,h}$. This ensures we update the bonus functions at most $O(\log K)$ times, which, when combined with the truncation-less least squares routine, allows keeping the number of variables in the policy parameterization $O(d^2 \log K)$. We conclude this section with our main theorem, providing the regret guarantees of Algorithm 1.

Theorem 1. *Let $\delta > 0$, assume $K \geq H^5 d^4 \log^8(dHK/\delta)$, $H \geq 3$, $\log A \leq K$, and consider setting $\beta = 2c_\beta d^{3/2} H \log(dHK/\delta)$ where c_β is specified by Lemma 3, $\epsilon_{\mathrm{cov}} = H^{3/2} d^2 \log^4(dHK/\delta)/\sqrt{K}$ and $\eta = \sqrt{\log A}/(H\sqrt{K})$. Suppose we run Algorithm 1 with these parameters for either the adversarial case with full-feedback (Assumption 1), or the stochastic case with bandit-feedback (Assumption 2). Then we obtain the following bound w.p. $1 - 4\delta$:*

$$\sum_{k=1}^{K} V^{k,\pi^k} - V^{k,\pi^\star} = O\left(d^2 H^{7/2} \log^4 \frac{dHK}{\delta} \sqrt{K \log A}\right),$$

where big-O hides only constant factors independent of problem parameters.

Algorithm 1 Optimistic PO for Linear MDPs

input: $(\eta, \delta, \beta, \epsilon_{\mathrm{cov}})$.
$\left\{\left(\mathcal{D}_h^0, \Lambda_{0,h}\right)\right\}_{h\in[H]} \leftarrow$ Algorithm 2 $(\delta, \beta, \epsilon_{\mathrm{cov}})$
Let $K_0 - 1$ be the number of rounds Algorithm 2 played
Init $\forall s:\ \pi_h^1(\cdot|s) = \mathrm{Unif}(\mathcal{A}), \forall h \in [H]:\ \widehat{\Lambda}_{K_0,h} = 0.$
for $k = K_0, \ldots, K$ **do**
  Rollout π^k to generate $\left\{(s_h^k, a_h^k, \ell_h^k)\right\}_{h=1}^H$.
  $\widetilde{V}_{H+1}^k(\cdot) \equiv 0.$
  for $h = H, \ldots, 1$ **do**
    $\mathcal{D}_h^k \leftarrow \mathcal{D}_h^0 \cup \left\{(s_h^i, a_h^i, s_{h+1}^i)\right\}_{i=K_0}^{k-1}$
    $\Lambda_{k,h} \leftarrow I + \sum_{i\in\mathcal{D}_h^k} \phi(s_h^i, a_h^i)\phi(s_h^i, a_h^i)^\top$
    if $\det \Lambda_{k,h} \geq 2 \det \widehat{\Lambda}_{k,h}$ **then**
      $\widehat{\Lambda}_{k,h} \leftarrow \Lambda_{k,h}$
      $\hat{b}_h^k(s,a) = \beta\sqrt{\phi(s,a)^\top \widehat{\Lambda}_{k,h}^{-1} \phi(s,a)}$
    end if
    $\widehat{v}_h^k \leftarrow \Lambda_{k,h}^{-1} \sum_{i\in\mathcal{D}_h^k} \phi(s_h^i, a_h^i) \widetilde{V}_{h+1}^{k;\circ}(s_{h+1}^i)$
    $\widehat{\mathbb{P}}_h^k \widetilde{V}_{h+1}^{k;\circ}(s,a) = \phi(s,a)^\top \widehat{v}_h^k$
    $\widehat{g}_{k,h} \leftarrow \begin{cases} \textit{Adv.:} & g_{k,h} \\ \textit{Stoch.:} & \Lambda_{k,h}^{-1} \sum_{i\in\mathcal{D}_h^k} \phi(s_h^i, a_h^i)\ell_h^i(s_h^i, a_h^i) \end{cases}$
    $\widehat{\ell}_h^k(s,a) = \phi(s,a)^\top \widehat{g}_{k,h}$
    Set

$$\begin{cases} \widetilde{Q}_h^k(s,a) & = \widehat{\ell}_h^k(s,a) + \widehat{\mathbb{P}}_h^k \widetilde{V}_{h+1}^{k;\circ}(s,a) - \hat{b}_h^k(s,a) \\ \widetilde{Q}_h^{k;\circ}(s,a) & = \mathbf{I}\{s \in \mathcal{Z}_h\}\, \widetilde{Q}_h^k(s,a) \\ \widetilde{V}_h^k(s) & = \left\langle \widetilde{Q}_h^k(s,\cdot), \pi_h^k(\cdot|s) \right\rangle \\ \widetilde{V}_h^{k;\circ}(s) & = \left\langle \widetilde{Q}_h^{k;\circ}(s,\cdot), \pi_h^k(\cdot|s) \right\rangle \end{cases}$$

  end for
  # Policy improvement:

$$\pi_h^{k+1}(a|s) \propto \pi_h^k(a|s) e^{-\eta \widetilde{Q}_h^k(s,a)}$$

end for

3.1 Reward-free warmup

In this section we present Algorithm 2, which we employ for a pure exploration warmup period. The algorithm invokes the CovTraj algorithm Wagenmaker et al. [2022b] for each step of the horizon, and thus follows the same high level design of reward free exploration outlined in Algorithm 1 of Wagenmaker et al. [2022b].

The basic guarantee provided by the warmup period is given by the next lemma.

Lemma 1. *Assume we execute Algorithm 2 with the setting of $\beta = \widetilde{O}(d^{3/2}H)$ and $\epsilon_{\mathrm{cov}} \geq 1/K$. Then it will terminate after $O\left(\frac{d^4H^5}{\epsilon_{\mathrm{cov}}}\log^7\frac{dHK}{\delta}\right)$ episodes, and with probability $\geq 1-\delta$, outputs $\Lambda_{0,1},\ldots,\Lambda_{0,H}$ such that:*

$$\forall h, \forall \pi, \Pr_{s_h\sim\mu_h^\pi}(s_h \notin \mathcal{Z}_h) \leq \epsilon_{\mathrm{cov}}.$$

The proof of Lemma 1 is provided in Appendix B, and mostly follows from the basic guarantees of the CovTraj algorithm.

Algorithm 2 Reward Free Warmup

input: $\delta, \beta, \epsilon_{\mathrm{cov}}$
Set $m = \lceil \log \frac{1}{\epsilon_{\mathrm{cov}}} \rceil$
Set $\forall i \in [m],\ \gamma_i = 1/(2\beta H)$
$\mathcal{Z}_{H+1} \coloneqq \{s_{H+1}\}$
for $h = H, \ldots, 1$ **do**
  $\left\{\left(\mathcal{X}_{h,i}, \widetilde{\mathcal{D}}_{h,i}, \widetilde{\Lambda}_{h,i}\right)\right\}_{i=1}^{m} \leftarrow \text{CovTraj}(h, \delta/H, m, \{\gamma_i\})$
  $\mathcal{D}_h^0 \leftarrow \bigcup_i \widetilde{\mathcal{D}}_{h,i}$
  $\Lambda_{0,h} \leftarrow I + \sum_{t\in\mathcal{D}_h^0} \phi(s_h^t, a_h^t)\phi(s_h^t, a_h^t)^\top$
end for
return $\left\{\left(\mathcal{D}_h^0, \Lambda_{0,h}\right)\right\}_{h\in[H]}$

4 Analysis overview

In this section, we outline the technical arguments leading up to the proof of Theorem 1. At the core of most of the analyses of linear MDP algorithms that involve a value estimation step, is a uniform convergence argument that ensures the regression errors concentrate uniformly over the class of value functions explored by the algorithm. The need for uniform convergence stems from the fact that we estimate the value function using past rollouts and a previously estimated value function, which in itself depends on past rollouts through the current agent policy. The lemma below is used to establish this part of the argument, and is stated in a generic manner — this is essentially the same argument used in Jin et al. [2020].

Lemma 2. *Let $\mathcal{V} \subseteq \mathcal{S} \to \mathbb{R}$ be a class of functions where $\forall f \in \mathcal{V}, \|f\|_\infty \leq C$, fix $h \in [H]$, and consider a transitions dataset $\mathcal{D}_h = \left\{(s_h^i, a_h^i, s_{h+1}^i)\right\}_{i\in[k]}$ collected by agent rollouts in the environment. Let $\widehat{\mathbb{P}}_h f(s,a) = \phi(s,a)^\top \hat{v}_h^f$ be the approximation of $\mathbb{P}_h f(s,a) = \mathbb{E}_{s'\sim\mathbb{P}_h(\cdot|s,a)} f(s')$ given by the least squares estimate $\hat{v}_h^f = \Lambda_h^{-1}\sum_{i=1}^k \phi(s_h^i, a_h^i) f(s_{h+1}^i)$, $\Lambda_h \coloneqq I + \sum_{i=1}^k \phi(s_h^i, a_h^i)\phi(s_h^i, a_h^i)^\top$. Then, w.p. $1-\delta$ over the generation of $\mathcal{D}_h$, we have $\forall s, a \in \mathcal{S}\times\mathcal{A}, \forall f \in \mathcal{V}$;*

$$\left|\left(\mathbb{P}_h - \widehat{\mathbb{P}}_h\right) f(s,a)\right| \leq \left(4C\sqrt{2d\log k + \log\frac{\mathcal{N}_{1/k}(\mathcal{V})}{\delta}}\right)\|\phi(s,a)\|_{\Lambda_h^{-1}},$$

where $\mathcal{N}_\nu(\mathcal{V})$ denotes the $\|\cdot\|_\infty$ covering number of $\mathcal{V}$.

Compared to other works, our analysis differs most significantly in how the above lemma is applied in order to control the estimation errors of the algorithm; the important parameter being the bound C on the magnitude of the target function. Usually, the regression step for $\widetilde{V}_h^k$ uses a truncated version of $\widetilde{V}_{h+1}^k$ from the previous horizon step, which simplifies the analysis and unfortunately leads to a too large value function class. The core of our argument lies in the next lemma; specifically, in the proof of $\mathcal{E}^{\mathrm{vbu}}$.

Lemma 3 (The good event)**.** *There exists a universal constant c_β, such that for any $\delta > 0$, executing Algorithm 1 with $\beta \geq c_\beta d^{3/2}H\log(dHK/\delta)$, we have that the following hold w.p. $> 1-4\delta$.*

$\forall\pi, \forall h$:

$$\Pr_{s_h\sim\mu_h^\pi}(s_h \notin \mathcal{Z}_h) \leq \epsilon_{\mathrm{cov}}, \tag{$\mathcal{E}^{\mathrm{rfw}}$}$$

$\forall k \geq K_0, h, s, a$:

$$\left|\widetilde{Q}_h^{k;\circ}(s,a)\right| \leq 2H, \tag{$\mathcal{E}^{\mathrm{qbd}}$}$$

$$|(\mathbb{P}_h - \widehat{\mathbb{P}}_h^k)\widetilde{V}_{h+1}^{k;\circ}(s,a)| \leq \frac{\beta}{2}\left\|\phi(s,a)\right\|_{\Lambda_{k,h}^{-1}}, \tag{$\mathcal{E}^{\mathrm{vbu}}$}$$

$$|\widehat{\ell}_h^k(s,a) - \ell_h^k(s,a)| \leq \frac{\beta}{2}\left\|\phi(s,a)\right\|_{\Lambda_{k,h}^{-1}}, \tag{$\mathcal{E}^{\mathrm{sle}}$}$$

and $\forall h \in [H]$:

$$\sum_{k=K_0}^{K} \mathbb{E}_{s_h,a_h\sim\mu_h^k}\left[\left\|\phi(s_h,a_h)\right\|_{\Lambda_{k,h}^{-1}}\right] \leq 2\sum_{k=K_0}^{K}\left\|\phi(s_h^k,a_h^k)\right\|_{\Lambda_{k,h}^{-1}} + 4\log\frac{4KH}{\delta}. \tag{$\mathcal{E}^{\mathrm{bon}}$}$$

The proof of Lemma 3 is provided fully in Appendix A.3; below we provide an overview. The success of $\mathcal{E}^{\mathrm{rfw}}$ is given by Lemma 1, while the proofs for $\mathcal{E}^{\mathrm{sle}}$ and $\mathcal{E}^{\mathrm{bon}}$ follow from standard arguments; we provide their proofs in Lemmas 6 and 8, respectively.

Proof sketch ($\mathcal{E}^{\mathrm{rfw}} \cup \mathcal{E}^{\mathrm{sle}} \implies \mathcal{E}^{\mathrm{qbd}} \cup \mathcal{E}^{\mathrm{vbu}}$). Establishing the bound in $\mathcal{E}^{\mathrm{vbu}}$ involves showing (i) $\mathcal{E}^{\mathrm{qbd}}$ holds for $\widetilde{Q}_{h+1}^{k;\circ}$, and (ii) that the policy π_{h+1}^k belongs to a "small" policy class. Given (i) and (ii), it immediately follows that $\widetilde{V}_{h+1}^{k;\circ}$ belongs to a small and bounded value function class, which leads to $\mathcal{E}^{\mathrm{vbu}}$ through an application of Lemma 2.

We proceed by an inductive argument as follows. Let k, h, and assume we have already proved (i), (ii) and $\mathcal{E}^{\mathrm{vbu}}$ for $(k', h'), k' < k$ and $(k, h'), h' > h$. Our Q estimate on step h decomposes as:

$$\left|\widetilde{Q}_h^{k;\circ}(s,a)\right| = \left|\widehat{\ell}_h^k(s,a) + \widehat{\mathbb{P}}_h^k\widetilde{V}_{h+1}^{k;\circ}(s,a) - \hat{b}_h^k(s,a)\right|.$$

Now we may apply $\mathcal{E}^{\mathrm{sle}}$ for the loss term, and the inductive hypothesis combined with $\mathcal{E}^{\mathrm{rfw}}$ for the regression term; which gives us that it is close to the true and well bounded value — this gives (i). For (ii), we employ the inductive hypothesis for $k' < k$ to show that the policy π_h^k has compact parametric form. As mentioned above, (i) + (ii) now lead to $\mathcal{E}^{\mathrm{vbu}}$, which completes the inductive step and thus the proof. □

Proceeding, we consider the regret decomposition given by our next lemma, and continue with a proof sketch of Theorem 1 immediately after. Formal proofs are deferred to the relevant subsections in Appendix A.

Lemma 4 (Regret decomposition)**.** *Upon execution of Algorithm 1, conditioned on the good event Lemma 3, it holds that:*

$$\begin{aligned}
&\sum_{k=K_0}^{K} V^{k,\pi^k} - V^{k,\pi^\star} \leq 4\epsilon_{\mathrm{cov}}H^2K \\
&+ \underbrace{\sum_{h=1}^{H}\sum_{k=K_0}^{K}\mathbb{E}_{s_h,a_h\sim\mu_h^k}\left[-\Delta_h^k(s_h,a_h) + \hat{b}_h^k(s_h,a_h) \mid s_h \in \mathcal{Z}_h\right]}_{\textit{Bias}} \\
&+ \underbrace{\sum_{h=1}^{H}\sum_{k=K_0}^{K}\mathbb{E}_{s_h\sim\mu_h^\star}\left[\left\langle \widetilde{Q}_h^k(s_h,\cdot), \pi_h^k(\cdot|s_h) - \pi_h^\star(\cdot|s_h)\right\rangle \mid s_h \in \mathcal{Z}_h\right]}_{\textit{OMD}} \\
&+ \underbrace{\sum_{h=1}^{H}\sum_{k=K_0}^{K}\mathbb{E}_{s_h,a_h\sim\mu_h^\star}\left[\Delta_h^k(s_h,a_h) - \hat{b}_h^k(s_h,a_h) \mid s_h \in \mathcal{Z}_h\right]}_{\textit{Optimism}},
\end{aligned}$$

where

$$\Delta_h^k(s_h,a_h) \coloneqq \widehat{\ell}_h^k(s_h,a_h) - \ell_h^k(s_h,a_h) + \left(\widehat{\mathbb{P}}_h^k - \mathbb{P}_h\right)\widetilde{V}_{h+1}^{k;\circ}(s_h,a_h).$$

Proof sketch of Theorem 1 . Given our choice of parameters, by Lemma 1, we have that the number of warmup episodes satisfies

$$K_0 = O\left(\frac{d^4H^5}{\epsilon_{\mathrm{cov}}}\log^7\frac{dHK}{\delta}\right).$$

For the remainder of the proof, we assume the good event defined in Lemma 3 holds, which indeed occurs w.p. $1-4\delta$ by that lemma. Proceeding, owed to $\mathcal{E}^{\mathrm{vbu}}$, it is not hard to establish that

$$\text{Bias} \le 3\beta\sum_{h=1}^{H}\sum_{k=K_0}^{K}\mathbb{E}_{\mu_h^k}\left[\left\|\phi(s_h^k,a_h^k)\right\|_{\Lambda_{k,h}^{-1}}\right] \le 6\beta\sum_{h=1}^{H}\sum_{k=K_0}^{K}\left\|\phi(s_h^k,a_h^k)\right\|_{\Lambda_{k,h}^{-1}} + 12\beta H\log\frac{4KH}{\delta},$$

where the second inequality follows from $\mathcal{E}^{\mathrm{bon}}$. Using the elliptical potential lemma (Lemma 20), we can then obtain

$$\text{Bias} \le 12\beta H\left(\sqrt{Kd\log K}+\log\frac{4KH}{\delta}\right).$$

By standard arguments, the OMD term is bounded as

$$\text{OMD} \le \frac{H\log A}{\eta} + 4\eta H^3K,$$

and the optimism term is non positive, again owed to $\mathcal{E}^{\mathrm{vbu}}$. To conclude the proof, we combine the bound on the number of warmup episodes K_0 with Lemma 4, and the bounds argued above on all three terms, to obtain:

$$\begin{aligned}\sum_{k=1}^{K}V^{\pi^k}-V^\star \lesssim{}& \frac{d^4H^5}{\epsilon_{\mathrm{cov}}}\log^7\frac{dHK}{\delta} + \epsilon_{\mathrm{cov}}H^2K + \frac{H\log A}{\eta}\\ &+\eta H^3K + \beta H\left(\sqrt{Kd\log K}+\log\frac{KH}{\delta}\right),\end{aligned}$$

where $\lesssim$ hides only constant factors. Setting $\epsilon_{\mathrm{cov}},\beta$, and η as specified in the theorem statement yields:

$$\sum_{k=1}^{K}V^{\pi^k}-V^\star \lesssim d^2H^{7/2}\log^4\frac{dHK}{\delta}\sqrt{K\log A},$$

which completes the proof. □

References

Y. Abbasi-Yadkori, D. Pál, and C. Szepesvári. Improved algorithms for linear stochastic bandits. *Advances in neural information processing systems*, 24, 2011.

A. Agarwal, M. Henaff, S. Kakade, and W. Sun. Pc-pg: Policy cover directed exploration for provable policy gradient learning. *Advances in neural information processing systems*, 33:13399–13412, 2020.

A. Agarwal, S. M. Kakade, J. D. Lee, and G. Mahajan. On the theory of policy gradient methods: Optimality, approximation, and distribution shift. *J. Mach. Learn. Res.*, 22(98):1–76, 2021.

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

A. Beck and M. Teboulle. Mirror descent and nonlinear projected subgradient methods for convex optimization. *Operations Research Letters*, 31(3):167–175, 2003.

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

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

A. Cohen, T. Koren, and Y. Mansour. Learning linear-quadratic regulators efficiently with only $\sqrt{T}$ regret. In *International Conference on Machine Learning*, pages 1300–1309. PMLR, 2019.

Y. Dai, H. Luo, C.-Y. Wei, and J. Zimmert. Refined regret for adversarial mdps with linear function approximation. *arXiv preprint arXiv:2301.12942*, 2023.

S. Du, S. Kakade, J. Lee, S. Lovett, G. Mahajan, W. Sun, and R. Wang. Bilinear classes: A structural framework for provable generalization in rl. In *International Conference on Machine Learning*, pages 2826–2836. PMLR, 2021.

T. Haarnoja, A. Zhou, K. Hartikainen, G. Tucker, S. Ha, J. Tan, V. Kumar, H. Zhu, A. Gupta, P. Abbeel, et al. Soft actor-critic algorithms and applications. *arXiv preprint arXiv:1812.05905*, 2018.

E. Hazan et al. Introduction to online convex optimization. *Foundations and Trends® in Optimization*, 2(3-4):157–325, 2016.

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

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

C. Jin, Q. Liu, and S. Miryoosefi. Bellman eluder dimension: New rich classes of rl problems, and sample-efficient algorithms. *Advances in neural information processing systems*, 34:13406–13418, 2021.

S. M. Kakade. A natural policy gradient. *Advances in neural information processing systems*, 14, 2001.

F. Kong, X. Zhang, B. Wang, and S. Li. Improved regret bounds for linear adversarial mdps via linear optimization. *arXiv preprint arXiv:2302.06834*, 2023.

G. Lan. Policy mirror descent for reinforcement learning: Linear convergence, new sampling complexity, and generalized problem classes. *Mathematical programming*, pages 1–48, 2022.

T. Lattimore and C. Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020.

S. Levine and V. Koltun. Guided policy search. In *International conference on machine learning*, pages 1–9. PMLR, 2013.

B. Liu, Q. Cai, Z. Yang, and Z. Wang. Neural trust region/proximal policy optimization attains globally optimal policy. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

H. Liu, C.-Y. Wei, and J. Zimmert. Towards optimal regret in adversarial linear mdps with bandit feedback. *arXiv preprint arXiv:2310.11550*, 2023a.

Q. Liu, G. Weisz, A. György, C. Jin, and C. Szepesvári. Optimistic natural policy gradient: a simple efficient policy optimization framework for online rl. *arXiv preprint arXiv:2305.11032*, 2023b.

H. Luo, C.-Y. Wei, and C.-W. Lee. Policy optimization in adversarial mdps: Improved exploration via dilated bonuses. *Advances in Neural Information Processing Systems*, 34:22931–22942, 2021.

S. Mannor, Y. Mansour, and A. Tamar. *Reinforcement Learning: Foundations*. -, 2022. URL `https://sites.google.com/view/rlfoundations/home`.

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

A. S. Nemirovskij and D. B. Yudin. Problem complexity and method efficiency in optimization, 1983.

G. Neu and J. Olkhovskaya. Online learning in mdps with linear function approximation and bandit feedback. *Advances in Neural Information Processing Systems*, 34:10407–10417, 2021.

L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744, 2022.

A. Rosenberg, A. Cohen, Y. Mansour, and H. Kaplan. Near-optimal regret bounds for stochastic shortest path. In *International Conference on Machine Learning*, pages 8210–8219. PMLR, 2020.

J. Schulman, P. Moritz, S. Levine, M. Jordan, and P. Abbeel. High-dimensional continuous control using generalized advantage estimation. *arXiv preprint arXiv:1506.02438*, 2015.

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

L. Shani, Y. Efroni, A. Rosenberg, and S. Mannor. Optimistic policy optimization with bandit feedback. In *International Conference on Machine Learning*, pages 8604–8613. PMLR, 2020.

U. Sherman, T. Koren, and Y. Mansour. Improved regret for efficient online reinforcement learning with linear function approximation. *arXiv preprint arXiv:2301.13087*, 2023.

R. S. Sutton and A. G. Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

A. Wagenmaker and K. G. Jamieson. Instance-dependent near-optimal policy identification in linear mdps via online experiment design. *Advances in Neural Information Processing Systems*, 35: 5968–5981, 2022.

A. J. Wagenmaker, Y. Chen, M. Simchowitz, S. Du, and K. Jamieson. First-order regret in reinforcement learning with linear function approximation: A robust estimation approach. In *International Conference on Machine Learning*, pages 22384–22429. PMLR, 2022a.

A. J. Wagenmaker, Y. Chen, M. Simchowitz, S. Du, and K. Jamieson. Reward-free rl is no harder than reward-aware rl in linear markov decision processes. In *International Conference on Machine Learning*, pages 22430–22456. PMLR, 2022b.

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

L. Xiao. On the convergence rates of policy gradient methods. *The Journal of Machine Learning Research*, 23(1):12887–12922, 2022.

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

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

R. Yuan, S. S. Du, R. M. Gower, A. Lazaric, and L. Xiao. Linear convergence of natural policy gradient methods with log-linear policies. *arXiv preprint arXiv:2210.01400*, 2022.

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

A. Zanette, C.-A. Cheng, and A. Agarwal. Cautiously optimistic policy optimization and exploration with linear function approximation. In *Conference on Learning Theory*, pages 4473–4525. PMLR, 2021.

H. Zhong and T. Zhang. A theoretical analysis of optimistic proximal policy optimization in linear markov decision processes. *arXiv preprint arXiv:2305.08841*, 2023.

D. Zhou, Q. Gu, and C. Szepesvari. Nearly minimax optimal reinforcement learning for linear mixture markov decision processes. In *Conference on Learning Theory*, pages 4532–4576. PMLR, 2021a.

D. Zhou, J. He, and Q. Gu. Provably efficient reinforcement learning for discounted mdps with feature mapping. In *International Conference on Machine Learning*, pages 12793–12802. PMLR, 2021b.

A Deferred Proofs

In this section, we provide details of the analysis that weren't fully included in the main text. The central component is the proof of Lemma 3, which is given in Appendix A.3. In Appendix A.2 we define the value and policy classes which will be shown in Appendix A.3 to contain (w.h.p.) the values and policies explored by the algorithm. Appendix A.4 includes the technical details for the covering number bounds of the value classes defined in Appendix A.2.

Additional notation. We will make use of the following filtration;

$$\mathcal{F}_h^k \coloneqq \sigma\left((s_{h'}^1, a_{h'}^1, l_{h'}^1)_{h'=1}^H, \ldots, (s_{h'}^{k-1}, a_{h'}^{k-1}, l_{h'}^{k-1})_{h'=1}^H, (s_{h'}^k, a_{h'}^k, l_{h'}^k)_{h'=1}^h\right), \ \mathcal{F}^k \coloneqq \mathcal{F}_H^k, \quad (9)$$

where (s_h^i, a_h^i, l_h^i) are the (state, action, loss) random variables generated during policy rollouts. In addition, for any function class $E \subseteq \mathcal{X} \to \mathbb{R}$, where $\mathcal{X}$ is an arbitrary set, we let $\mathcal{N}_\nu(E)$ denote the $\|\cdot\|_\infty$ covering number of E; that is, the cardinality of the smallest set $\widetilde{E} \subset E$ such that for all $f \in E$, there exists $\tilde{f} \in \widetilde{E}$ such that $\max_{x\in\mathcal{X}} \left|f(x) - \tilde{f}(x)\right| \le \nu$.

A.1 Proof of Lemma 2

Proof of Lemma 2. Denote

$$v_h \coloneqq \int \psi_h(s') f(s') \mathrm{d}s'; \quad \hat{v}_h \coloneqq \Lambda_h^{-1} \sum_{i=1}^k \phi(s_h^i, a_h^i) f(s_{h+1}^i).$$

Then, we have

$$\begin{aligned}
\hat{v}_h - v_h &= \Lambda_h^{-1}\left(\sum_{i=1}^k \phi(s_h^i, a_h^i) f(s_{h+1}^i) - \left(I + \sum_{i=1}^k \phi(s_h^i, a_h^i)\phi(s_h^i, a_h^i)^\top\right) v_h\right) \\
&= \Lambda_h^{-1} \sum_{i=1}^k \phi(s_h^i, a_h^i)\left(f(s_{h+1}^i) - \phi(s_h^i, a_h^i)^\top v_h\right) - \Lambda_h^{-1} v_h \\
&= \Lambda_h^{-1} \sum_{i=1}^k \phi(s_h^i, a_h^i)\left(f(s_{h+1}^i) - \mathbb{E}_{s'}\left[f(s') \mid s_h^i, a_h^i\right]\right) - \Lambda_h^{-1} v_h \qquad (10)
\end{aligned}$$

Now, note that

$$\left\|\Lambda_h^{-1} v_h\right\|_{\Lambda_h}^2 = \|v_h\|_{\Lambda_h^{-1}}^2 \le \|v_h\|^2 \le dC^2.$$

In addition, for the first term in Eq. (10), we consider the filtration defined in Eq. (9), and note that $\phi(s_h^i, a_h^i)$ is $\mathcal{F}_h^i$-measurable while s_{h+1}^i is $\mathcal{F}_{h+1}^i$-measurable. Hence we may apply Lemma 23 to obtain that for any $\epsilon, p > 0$, with probability $\ge 1 - p$ we have

$$\begin{aligned}
&\left\|\Lambda_h^{-1} \sum_{i=1}^k \phi(s_h^i, a_h^i)\left(f(s_{h+1}^i) - \mathbb{E}_{s'}\left[f(s') \mid s_h^i, a_h^i\right]\right)\right\|_{\Lambda_h}^2 \\
&\quad \le 4C^2\left(\frac{d}{2}\log(k+1) + \log\frac{\mathcal{N}_\epsilon(\mathcal{V})}{p}\right) + 8k^2\epsilon^2 \\
&\quad \le 4C^2\left(d\log(k) + \log\frac{\mathcal{N}_{1/k}(\mathcal{V})}{p}\right) + 8 \qquad \text{(setting } \epsilon = 1/k\text{)}
\end{aligned}$$

Combining Eq. (10) with the inequalities from the last two displays gives;

$$\begin{aligned}
&\|\hat{v}_h - v_h\|_{\Lambda_h}^2 \le 16C^2\left(2d\log(k) + \log\frac{\mathcal{N}_{1/k}(\mathcal{V})}{p}\right) \\
\implies &\|\hat{v}_h - v_h\|_{\Lambda_h} \le 4C\sqrt{2d\log(k) + \log\frac{\mathcal{N}_{1/k}(\mathcal{V})}{p}}.
\end{aligned}$$

Finally,

$$\left|\left(\mathbb{P}_h - \overline{\mathbb{P}}_h\right)V(s,a)\right| = \left|\phi(s,a)^\top(\hat{v}_h - v_h)\right| \le \|\phi(s,a)\|_{\Lambda_h^{-1}} \|\hat{v}_h - v_h\|_{\Lambda_h},$$

which completes the proof after plugging in the bound from the previous display. □

A.2 Value and policy classes

Given an input parameter β given to Algorithm 1, we consider the softmax policy class as defined below:

$$\begin{aligned}\mathcal{Y}(D_w, \lambda_-, \lambda_+, J_{\max}) &\coloneqq \{y(\cdot; w, W_{1:J}) \mid \|w\| \le D_w, \lambda_- I \preceq W_j \preceq \lambda_+ I, J \le J_{\max}\}, \\ &\text{where } y(x; w, W_{1:J}) \coloneqq x^\top w + \sum_{j=1}^{J} \|x\|_{W_j}; \\ \Pi &\coloneqq \left\{\pi(\cdot|\cdot; y) \mid y \in \mathcal{Y}(3dHK^2, K^{-2}, \beta^2 K^2, 2d\log K)\right\}, \\ &\text{where } \pi(a|s; y) \coloneqq \frac{e^{y(\phi(s,a))}}{\sum_b e^{y(\phi(s,b))}}. \end{aligned} \tag{11}$$

We further consider the following class of empirical restricted (Q-)functions:

$$\begin{aligned}\widetilde{Q}^\circ(s, a; w, W, \mathcal{Z}) &\coloneqq \mathbf{I}\{s \in \mathcal{Z}\}\left(\phi(s,a)^\top w - \sqrt{\phi(s,a)^\top W \phi(s,a)}\right), \\ \widetilde{\mathcal{Q}}^\circ(\mathcal{Z}, C) &\coloneqq \left\{\widetilde{Q}^\circ(\cdot,\cdot; w, W, \mathcal{Z}) \mid \|w\|_2 \le 2dHK, \|W\|_2 \le \beta^2, \left\|\widetilde{Q}^\circ(\cdot,\cdot; w, W, \mathcal{Z})\right\|_\infty \le C\right\},\end{aligned} \tag{12}$$

and their corresponding value functions:

$$\widetilde{V}(s; \pi_h, \widetilde{Q}^\circ) \coloneqq \left\langle \pi_h(\cdot|s), \widetilde{Q}^\circ(s,\cdot)\right\rangle.$$

Now define the following empirical restricted value function class:

$$\widetilde{\mathcal{V}}^\circ(\mathcal{Z}, C) = \left\{\widetilde{V}(\cdot; \pi_h, \widetilde{Q}^\circ) \colon \mathcal{S} \to \mathbb{R} \mid \widetilde{Q}^\circ \in \mathcal{Q}^\circ(\mathcal{Z}, C), \pi_h \in \Pi\right\}. \tag{13}$$

The following lemma (of which the proof is deferred to Appendix A.4) provides the bound on the covering number of the function class defined in Eq. (13) above.

Lemma 5. *There exists a universal constant $c_{\mathcal{N}}$, such that for any $\nu > 0$, $\mathcal{Z} \subseteq \mathcal{S}$,*

$$\log \mathcal{N}_\nu(\widetilde{\mathcal{V}}^\circ(\mathcal{Z}, C)) \le c_{\mathcal{N}} d^3 \log(\beta CK/\nu).$$

A.3 Proof of Lemma 3

In this section we provide the full technical details for the analysis of the good event Lemma 3. The core part of the argument establishes the confidence bounds for the regression step in spite of the absence of the truncation. To begin, we first define an additional success event; the concentration of least squares errors uniformly over the class of empirical value functions (recall the function class $\widetilde{\mathcal{V}}^\circ$ is defined in Eq. (13)).

$$\forall k \ge K_0, h; \forall V_{h+1} \in \widetilde{\mathcal{V}}^\circ(\mathcal{Z}_{h+1}, 2H); \forall s, a: \left|\left(\mathbb{P}_h - \widehat{\mathbb{P}}_h^k\right) V_{h+1}(s,a)\right| \le (\beta/2)\|\phi(s,a)\|_{\Lambda_{k,h}^{-1}}, \tag{$\mathcal{E}^{\mathrm{uls}}$}$$

The core argument pertaining to the regression errors proceeds as follows.

1. Lemma 1, establishes the success probability of $\mathcal{E}^{\mathrm{rfw}}$. For the most part this follows from the guarantees of the CovTraj algorithm developed in the prior work of Wagenmaker et al. [2022b].
2. Lemma 7 establishes the success probability of $\mathcal{E}^{\mathrm{uls}}$; ensuring concentration of the regression errors w.r.t. the value function classes $\widetilde{\mathcal{V}}^\circ(\mathcal{Z}_{h+1}, 2H)$.
3. Given that $\mathcal{E}^{\mathrm{rfw}}$ and $\mathcal{E}^{\mathrm{uls}}$ both hold, Lemma 11 provides, using a careful inductive argument, that the value functions estimated in Algorithm 1 are contained in the function class $\widetilde{\mathcal{V}}^\circ(\mathcal{Z}_{h+1}, 2H)$. Thus, $\mathcal{E}^{\mathrm{qbd}}$ and $\mathcal{E}^{\mathrm{vbu}}$ hold.

Lemma 6 (success of $\mathcal{E}^{\mathrm{bon}}$)**.** *For any $\delta > 0$, we have that with probability $\ge 1 - \delta$, for all h:*

$$\sum_{k=1}^{K} \mathbb{E}_{\mu_h^k}\left[\left\|\phi(s_h, a_h)\right\|_{\Lambda_{k,h}^{-1}}\right] \le 2\sum_{k=1}^{K}\left\|\phi(s_h^k, a_h^k)\right\|_{\Lambda_{k,h}^{-1}} + 4\log\frac{4KH}{\delta}.$$

Proof. Denote $X_k = \left\|\phi(s_h^k, a_h^k)\right\|_{\Lambda_{k,h}^{-1}}$, and recall the definition of $\mathcal{F}^k$ in Eq. (9). Then X_k is $\mathcal{F}^k$ measurable, and

$$\mathbb{E}_{\mu_h^k}\left[\left\|\phi(s_h, a_h)\right\|_{\Lambda_{k,h}^{-1}}\right] = \mathbb{E}\left[X_k \mid \mathcal{F}^{k-1}\right].$$

In addition, by the definition of $\Lambda_{k,h} = I + \sum_{i\in\mathcal{D}_h^k} \phi(s_h^i, a_h^i)\phi(s_h^i, a_h^i)^\top$ in Algorithm 1, and by the assumption that $\|\phi(s,a)\| \le 1$ (Definition 1), we have that $0 \le X_k \le 1$. Thus by Lemma 19 and the union bound, we have that w.p. $1-\delta$, for all $h \in [H]$:

$$\sum_{k=1}^K \mathbb{E}\left[X_k \mid \mathcal{F}^{k-1}\right] \le 2\sum_{k=1}^K X_k + \log\frac{2KH}{(\delta/KH)} \le 2\sum_{k=1}^K \left\|\phi(s_h^k, a_h^k)\right\|_{\Lambda_{k,h}^{-1}} + 4\log\frac{2KH}{\delta},$$

which completes the proof. □

Lemma 7 (success of $\mathcal{E}^{\mathrm{uls}}$)**.** *There exists a constant $c_\beta > 0$, such that when running Algorithm 1 with $\beta \ge 2c_\beta dH\log(dHK/\delta)$, we have that the event $\mathcal{E}^{\mathrm{uls}}$ holds with probability $\ge 1-\delta$.*

Proof. Let $h, k \in [H] \times \{K_0, \ldots, K\}$, and note that since we define the regression solution using

$$\mathcal{D}_h^k = \mathcal{D}_h^0 \cup \left\{(s_h^i, a_h^i, s_{h+1}^i)\right\}_{i=K_0}^{k-1},$$

this dataset is independent of the known states set $\mathcal{Z}_{h+1}$ which is defined using $\mathcal{D}_{h+1}^0$. Indeed, this is because $\left\{\mathcal{D}_h^0\right\}_{h\in[H]}$ were generated by *independent runs* of CovTraj in Algorithm 2. Hence, the value class $\widetilde{\mathcal{V}}^\circ(\mathcal{Z}_{h+1}, 2H)$ is independent of $\mathcal{D}_h^k$, and we may apply Lemma 2 to obtain that w.p. $\ge 1-\delta'$,

$$\forall V_{h+1} \in \widetilde{\mathcal{V}}^\circ(\mathcal{Z}_{h+1}, 2H), \forall s\in\mathcal{S}, a\in\mathcal{A}: \ \left|\left(\mathbb{P}_h - \widehat{\mathbb{P}}_h^k\right)V_{h+1}(s,a)\right| \le \hat{\beta}\left\|\phi(s,a)\right\|_{\Lambda_{k,h}^{-1}},$$

where

$$\hat{\beta} = \left(8H\sqrt{2d\log k + \log\frac{\mathcal{N}_{1/k}(\widetilde{\mathcal{V}}^\circ(\mathcal{Z}_{h+1}, 2H))}{\delta'}}\right).$$

By Lemma 5, we have

$$\log\mathcal{N}_{1/K}(\widetilde{\mathcal{V}}^\circ(\mathcal{Z}_{h+1}, 2H)) \le cd^3\log(\beta HK),$$

for some universal constant c, which implies that

$$\hat{\beta} \le c'd^{3/2}H\log(\beta HK/\delta'),$$

for a suitable constant $c' > 0$. Setting $\delta' = \delta/KH$, we now have by the union bound that,

$$\begin{aligned}&\forall k,h; \forall V_{h+1}\in\widetilde{\mathcal{V}}^\circ(\mathcal{Z}_{h+1}, 2H); \forall s,a: \\ &\qquad\left|\left(\mathbb{P}_h - \widehat{\mathbb{P}}_h^k\right)V_{h+1}(s,a)\right| \le 2c'd^{3/2}H\log(\beta HK/\delta)\left\|\phi(s,a)\right\|_{\Lambda_{k,h}^{-1}}.\end{aligned}$$

Finally, by Lemma 18, for a suitable $c_\beta > 0$ we have

$$\beta/2 \ge c_\beta d^{3/2}H\log(dHK/\delta) \ge 2c'd^{3/2}H\log(\beta HK/\delta),$$

which completes the proof.

□

Lemma 8 (success of $\mathcal{E}^{\mathrm{sle}}$)**.** *Consider running Algorithm 1 in the stochastic case with bandit feedback, with $\beta \ge 2c_\beta dH\log(dHK/\delta)$ as specified by Lemma 7. Then, we have that the event $\mathcal{E}^{\mathrm{sle}}$ holds w.p. $1-\delta$.*

Proof. For a given k, h, we have

$$\forall s,a: \ \left|\widehat{\ell}_h^k(s,a) - \ell_h^k(s,a)\right| = \left|\phi(s,a)^\top\left(\widehat{g}_{k,h} - g_{k,h}\right)\right| \le \left\|\phi(s,a)\right\|_{\Lambda_{k,h}^{-1}}\left\|\widehat{g}_{k,h} - g_{k,h}\right\|_{\Lambda_{k,h}}. \quad (14)$$

Following the same algebraic argument as that given in Lemma 2, we have

$$\begin{aligned}\|\widehat{g}_{k,h} - g_{k,h}\|_{\Lambda_{k,h}} &= \left\|\Lambda_{k,h}^{-1}\sum_{i=1}^{k-1}\phi(s_h^i,a_h^i)\left(\ell_h^i(s_h^i,a_h^i)-\ell_h(s_h^i,a_h^i)\right)-\Lambda_{k,h}^{-1}g_{k,h}\right\|_{\Lambda_{k,h}}\\ &\leq \left\|\sum_{i=1}^{k-1}\phi(s_h^i,a_h^i)\left(\ell_h^i(s_h^i,a_h^i)-\ell_h(s_h^i,a_h^i)\right)\right\|_{\Lambda_{k,h}^{-1}}+\|g_{k,h}\|_{\Lambda_{k,h}^{-1}}.\end{aligned}$$

By Lemma 22 (the application of which is legitimate due to Assumption 2) and the union bound, for any $\delta>0$ the first term above is bounded by $\sqrt{4d\log(HK/\delta)}$ for all k,h, while the second term is bounded a.s. by $\sqrt{d}$ owed to the assumption in Definition 1 and that $\Lambda_{k,h}^{-1}\preceq I$. Concluding, we have w.p. $\geq 1-\delta$, for all k,h:

$$\|\widehat{g}_{k,h}-g_{k,h}\|_{\Lambda_{k,h}}\leq\sqrt{4d\log(HK/\delta)}+\sqrt{d}\leq\beta/2.$$

The proof is complete after plugging the above inequality into Eq. (14). □

Lemma 9. *Let* $\mathcal{D}_h=\left\{(s_h^i,a_h^i)\right\}_{i\in[k]}$*, and* $\Lambda_h:=I+\sum_{i\in\mathcal{D}_h}\phi(s_h^i,a_h^i)\phi(s_h^i,a_h^i)^\top$*. Then,*

$$\left\|\Lambda_h^{-1}\sum_{i\in\mathcal{D}_h}\phi(s_h^i,a_h^i)\right\|_2\leq\sqrt{dk}.$$

Proof. Follows from the exact same argument as in Jin et al. [2020], Lemma B.2. □

Lemma 10. *Let* $K_0\leq\tau<K$*,* $h\in[H]$*, and assume* $\widetilde{V}_{h+1}^{k;\circ}\in\widetilde{\mathcal{V}}^\circ(\mathcal{Z}_h,2H)$ *for all* $k\in\{K_0,\ldots,\tau-1\}$*. Then* $\pi_h^{\tau+1}\in\Pi$*, where* Π *is defined in Eq.* (11)*, and* $\pi_h^{\tau+1}$ *is a mirror descent step from* π_h^τ *as defined in Algorithm 1.*

Proof. By the definition of the OMD update step in Algorithm 1, we have for any a,s;

$$\pi_h^{\tau+1}(a|s)=\frac{e^{-\eta\sum_{k=K_0}^{\tau}\widetilde{Q}_h^k(s,a)}}{\sum_{a'}e^{-\eta\sum_{k=K_0}^{\tau}\widetilde{Q}_h^k(s,a')}}.$$

In addition, by the definition of the estimated Q-functions $\widetilde{Q}_h^k$ in Algorithm 1, we have;

$$\begin{aligned}-\eta\sum_{k=K_0}^{\tau}\widetilde{Q}_h^k(s,a)&=-\eta\sum_{k=K_0}^{\tau}\left(\widehat{\ell}_h^k(s,a)+\widehat{\mathbb{P}}_h^k\widetilde{V}_{h+1}^{k;\circ}(s,a)\right)-\hat{b}_h^k(s,a)\\&=-\eta\sum_{k=K_0}^{\tau}\phi(s,a)^\top\left(\widehat{g}_{k,h}+\widehat{v}_h^k\right)+\eta\sum_{k=K_0}^{\tau}\hat{b}_h^k(s,a)\\&=-\eta\sum_{k=K_0}^{\tau}\phi(s,a)^\top\left(\widehat{g}_{k,h}+\widehat{v}_h^k\right)+\eta\beta\sum_{k=K_0}^{\tau}\|\phi(s,a)\|_{\widehat{\Lambda}_{k,h}^{-1}}\\&=\phi(s,a)^\top\left(-\eta\sum_{k=K_0}^{\tau}\widehat{g}_{k,h}+\widehat{v}_h^k\right)+\eta\beta\sum_{j=1}^{J}(k_{j+1}-k_j)\|\phi(s,a)\|_{\Lambda_{k_j,h}^{-1}},\end{aligned}$$

where k_j are the episodes on which we update the bonus matrices $\widehat{\Lambda}_{k,h}$ in Algorithm 1. Now, since for all $K_0\leq k\leq K$ we have

$$\|\Lambda_{k,h}\|=\left\|I+\sum_{i\in\mathcal{D}_h^k}\phi(s_h^i,a_h^i)\phi(s_h^i,a_h^i)^\top\right\|\leq\|I\|+\sum_{i\in\mathcal{D}_h^k}\left\|\phi(s_h^i,a_h^i)\phi(s_h^i,a_h^i)^\top\right\|\leq 1+K,$$

and $I\preceq\Lambda_{k,h}$, it follows that

$$2^J\det\Lambda_{K_0,h}\leq\det\Lambda_{K,h}\leq\|\Lambda_{k,h}\|^d\leq(K+1)^d,$$

and $1 \leq \det \Lambda_{K_0,h}$. Thus it is implied that $J \leq d\log(K+1) \leq 2d\log K$. In addition, $\eta\beta(k_{j+1}-k_j)\left\|\phi(s,a)\right\|_{\widehat{\Lambda}^{-1}_{k_j,h}} = \left\|\phi(s,a)\right\|_{W_j}$ when we define

$$W_j = \eta^2\beta^2(k_{j+1}-k_j)^2\Lambda^{-1}_{k_j,h}, \text{ and thus } \frac{1}{K^2}I \preceq W_j \preceq \beta^2K^2\Lambda^{-1}_{k_j,h} \preceq \beta^2K^2I.$$

Furthermore, in the adversarial case $\|\widehat{g}_{k,h}\| = \|g_{k,h}\| \leq \sqrt{d}$ by assumption (see Definition 1), and in the stochastic case,

$$\|\widehat{g}_{k,h}\| = \left\|\Lambda^{-1}_{k,h}\sum_{i\in\mathcal{D}^k_h}\phi(s^i_h,a^i_h)\ell^i_h(s^i_h,a^i_h)\right\| \leq \sqrt{dK},$$

where the inequality follows from Lemma 9 and our assumption that $\left|\ell^k_h(s^i_h,a^i_h)\right| \leq 1$. In addition, in both the stochastic and adversarial cases, we have

$$\|\widehat{v}^k_h\| = \left\|\Lambda^{-1}_{k,h}\sum_{i\in\mathcal{D}^k_h}\phi(s^i_h,a^i_h)\widetilde{V}^{k;\circ}_{h+1}(s^i_{h+1})\right\| \leq 2H\sqrt{dK},$$

which follows again by Lemma 9, and our assumption that $\widetilde{V}^{k;\circ}_{h+1} \in \widetilde{\mathcal{V}}^\circ(\mathcal{Z}_h, 2H) \implies \left\|\widetilde{V}^{k;\circ}_{h+1}\right\|_\infty \leq 2H$ for all $k \leq \tau$. Thus, $\left\|\widehat{g}_{k,h}+\widehat{v}^k_h\right\| \leq 3H\sqrt{dK}$ for all $k \leq \tau$. Concluding, we have shown that

$$\pi^{\tau+1}_h(a|s) \propto \exp\left(\phi(s,a)^\top w^\tau_h + \sum_{j=1}^J \left\|\phi(s,a)\right\|_{W_j}\right),$$

where $\|w^\tau_h\| \leq 3dHK^2$ and $K^{-2}I \preceq W_j \preceq \beta^2K^2I$, therefore $\pi^{\tau+1}_h \in \Pi$, as required. □

Lemma 11 (success of $\mathcal{E}^{\mathrm{qbd}} \cup \mathcal{E}^{\mathrm{vbu}}$)**.** *Assume that the event $\mathcal{E}^{\mathrm{rfw}} \cup \mathcal{E}^{\mathrm{uls}} \cup \mathcal{E}^{\mathrm{sle}}$ holds. Then, we have that,*

$$\forall k \geq K_0, h \in [H] : \widetilde{Q}^{k;\circ}_h \in \widetilde{\mathcal{Q}}^\circ(\mathcal{Z}_h, C_h), \widetilde{V}^{k;\circ}_h \in \widetilde{\mathcal{V}}^\circ(\mathcal{Z}_h, C_h),$$

where $C_h \coloneqq (H-h+1)(1+2/H)$. Furthermore, we have that the event $\mathcal{E}^{\mathrm{vbu}}$ holds, that is,

$$\forall k \geq K_0, h \in [H]; \forall s,a : \left|\left(\mathbb{P}_h - \widehat{\mathbb{P}}^k_h\right)\widetilde{V}^{k;\circ}_{h+1}(s,a)\right| \leq (\beta/2)\left\|\phi(s,a)\right\|_{\Lambda^{-1}_{k,h}}. \tag{15}$$

Proof. We begin first by establishing simple bounds on the instantaneous loss estimates. For any $k \geq K_0$, we have in the adversarial case $\widehat{\ell}^k_h(s,a) = \ell^k_h(s,a)$ for all s,a,h,k, so $\left|\widehat{\ell}^k_h(s,a)\right| \leq 1$ by the assumption in Definition 1. In the stochastic case on the other hand, for any $s,a \in \mathcal{Z}_h \times \mathcal{A}$, owed to our assumption that $\mathcal{E}^{\mathrm{sle}}$ holds;

$$\left|\widehat{\ell}^k_h(s,a)\right| \leq \left|\ell^k_h(s,a)\right| + \left|\widehat{\ell}^k_h(s,a) - \ell^k_h(s,a)\right| \leq 1 + \beta\left\|\phi(s,a)\right\|_{\Lambda^{-1}_{k,h}}.$$

Furthermore, we have,

$$\forall h \in [H], s,a \in \mathcal{Z}_h \times \mathcal{A} : \left\|\phi(s,a)\right\|_{\Lambda^{-1}_{k,h}} \leq \left\|\phi(s,a)\right\|_{\widehat{\Lambda}^{-1}_{k,h}} \leq \left\|\phi(s,a)\right\|_{\Lambda^{-1}_{0,h}} \leq 1/(2\beta H), \tag{16}$$

where the last inequality follows from the definition of $\mathcal{Z}_h$, thus we obtain $\beta\left\|\phi(s,a)\right\|_{\Lambda^{-1}_{k,h}} \leq 1/(2H)$. To conclude, in both the stochastic and adversarial cases we have:

$$\forall k \geq K_0, h \in [H], s \in \mathcal{Z}_h, a \in \mathcal{A};\ \left|\widehat{\ell}^k_h(s,a)\right| \leq 1 + 1/(2H). \tag{17}$$

The rest of the proof proceeds by an inductive argument as follows. Fix $K_0 \leq k \leq K$, and assume we have already proved the claim for all $k', h \in \{K_0, \ldots, k-1\} \times [H]$. We will now establish the claim for episode k by induction on $h = H, \ldots, 1$.

Base case $h = H$: Here, we have

$$\left|\widetilde{Q}^k_H(s,a)\right| = \left|\widehat{\ell}^k_H(s,a) - \hat{b}^k_h(s,a)\right| \leq 1 + 1/(2H) + \beta\left\|\phi(s,a)\right\|_{\widehat{\Lambda}^{-1}_{k,h}} \leq 1 + 1/H,$$

where the first inequality follows from Eq. (17), and the last inequality from Eq. (16). Thus, we obtain $\widetilde{Q}^{k;\circ}_H \in \widetilde{\mathcal{Q}}^\circ(\mathcal{Z}_H, 1+1/H) \subset \widetilde{\mathcal{Q}}^\circ(\mathcal{Z}_H, C_H)$. Further, since $\widetilde{V}^{k';\circ}_{H+1} \equiv 0$ for any $k' \in [K]$, we may apply Lemma 10 which ensures $\pi^k_H \in \Pi$. Thus, it also follows that $\widetilde{V}^{k;\circ}_H(s) \in \widetilde{V}^\circ(\mathcal{Z}_H, C_H)$.

Inductive step: Let $h < H$ and assume $\widetilde{V}^{k;\circ}_{h+1} \in \widetilde{\mathcal{V}}^{\circ}(\mathcal{Z}_{h+1}, C_{h+1})$. For $s \in \mathcal{Z}_h, a \in \mathcal{A}$, we have;

$$\begin{aligned}\left|\widetilde{Q}^{k;\circ}_h(s,a)\right| &= \left|\widehat{\ell}^k_h(s,a) + \widehat{\mathbb{P}}^k_h \widetilde{V}^{k;\circ}_{h+1}(s,a) - \hat{b}^k_h(s,a)\right| \\ &= \left|\widehat{\ell}^k_h(s,a) + \mathbb{P}_h \widetilde{V}^{k;\circ}_{h+1}(s,a) + (\widehat{\mathbb{P}}^k_h - \mathbb{P}_h)\widetilde{V}^{k;\circ}_{h+1}(s,a) - \hat{b}^k_h(s,a)\right| \\ &\le 1 + 1/H + C_{h+1} + \beta \left\|\phi(s,a)\right\|_{\Lambda^{-1}_{k,h}} + \beta \left\|\phi(s,a)\right\|_{\widehat{\Lambda}^{-1}_{k,h}},\end{aligned}$$

where the last inequality follows from Eq. (17), the inductive hypothesis, and by the assumption that $\mathcal{E}^{\mathrm{uls}}$ holds. Applying Eq. (16) again, this implies that the empirical Q is well bounded on the known states;

$$\left|\widetilde{Q}^{k;\circ}_h(s,a)\right| \le 1 + 1/H + C_{h+1} + 1/H = C_h.$$

In addition, for any $s, a \in \mathcal{S} \times \mathcal{A}$;

$$\widetilde{Q}^k_h(s,a) = \phi(s,a)^\top \widehat{g}_{k,h} + \widehat{\mathbb{P}}^k_h \widetilde{V}^{k;\circ}_{h+1}(s,a) - \hat{b}^k_h(s,a) = \phi(s,a)^\top \left(\widehat{g}_{k,h} + \widehat{v}^k_h\right) - \beta \left\|\phi(s,a)\right\|_{\widehat{\Lambda}^{-1}_{k,h}},$$

Further, as argued in the proof of Lemma 10, by Lemma 9 and our assumption that $\left\|\widetilde{V}^{k;\circ}_{h+1}\right\|_\infty \le 2H$, we have that

$$\left\|\widehat{v}^k_h\right\| \le 2H\sqrt{dK}, \quad \text{and } \left\|\widehat{g}_{k,h}\right\| \le \sqrt{dK}.$$

In addition, $\beta \left\|\phi(s,a)\right\|_{\widehat{\Lambda}^{-1}_{k,h}} = \left\|\phi(s,a)\right\|_W$ for

$$W = \beta^2 \widehat{\Lambda}^{-1}_{k,h}, \text{ and thus } \|W\| = \beta^2 \left\|\widehat{\Lambda}^{-1}_{k,h}\right\| \le \beta^2.$$

Therefore, we establish that $\widetilde{Q}^{k;\circ}_h \in \widetilde{\mathcal{Q}}^{\circ}(\mathcal{Z}_h, C_h)$. Now, by our (first) inductive assumption that $\widetilde{V}^{k';\circ}_{h+1} \in \widetilde{\mathcal{V}}^{\circ}(\mathcal{Z}_{h+1}, C_{h+1})$ for all $k' < k$, we may apply Lemma 10 to obtain that $\pi^k_h \in \Pi$. This immediately implies that $\widetilde{V}^{k;\circ}_h \in \widetilde{\mathcal{V}}(\mathcal{Z}_h, C_h)$, and completes the inductive argument. Finally, combined with our assumption that $\mathcal{E}^{\mathrm{uls}}$ holds, this implies $\mathcal{E}^{\mathrm{vbu}}$ holds, which completes the proof. □

We conclude this section with the proof of the good event Lemma 3, which now follows easily by combining the above lemmas.

Proof of Lemma 3. By Lemmas 1, 6 and 7, and the union bound, we have that $\mathcal{E}^{\mathrm{rfw}} \cup \mathcal{E}^{\mathrm{uls}} \cup \mathcal{E}^{\mathrm{sle}} \cup \mathcal{E}^{\mathrm{bon}}$ holds w.p.$\ge 1 - 4\delta$. By Lemma 11, this now implies that $\mathcal{E}^{\mathrm{qbd}} \cup \mathcal{E}^{\mathrm{vbu}}$ holds as well, which completes the proof. □

A.4 Covering of empirical value functions

Lemma 12 (Policy class is Lipschitz)**.** *For any $\pi_h, \tilde{\pi}_h \in \Pi$, $\pi_h(\cdot|\cdot) = \pi(\cdot|\cdot; y_h)$, $\tilde{\pi}_h(\cdot|\cdot) = \tilde{\pi}(\cdot|\cdot; \tilde{y}_h)$, parameterized by $y_h(\cdot) = y_h(\cdot; w, W_{1:J})$, $\tilde{y}_h(\cdot) = y_h(\cdot; \tilde{w}, \tilde{W}_{1:J})$, we have for any $s \in \mathcal{S}$:*

$$\left\|\pi_h(\cdot|s) - \tilde{\pi}_h(\cdot|s)\right\|_1 \le 6K \sqrt{\|w - \tilde{w}\|^2 + \sum_{j=1}^{J} \left\|W_j - \tilde{W}_j\right\|^2}.$$

Proof. We have, for any $x \in \mathbb{R}^d$,

$$\begin{aligned}\nabla_w y(x; w, W_{1:J}) &= x \\ \nabla_{W_j} y(x; w, W_{1:J}) &= \nabla_{W_j} \left(\sqrt{x^\top W_j x}\right) = \frac{1}{2\sqrt{x^\top W_j x}} x x^\top.\end{aligned}$$

Thus, considering $y(x; w, W_{1:J}) \in \mathcal{Y}$ implies $K^{-2} I \preceq W_j$;

$$\left\|\nabla_{W_j} y(x; w, W_{1:J})\right\|_F = \frac{1}{2\sqrt{x^\top W_j x}} \left\|x x^\top\right\|_F = \frac{1}{2\sqrt{x^\top W_j x}} \left\|x\right\|^2 \le \frac{1}{2\sqrt{\lambda_{\min}(W_j)\left\|x\right\|}} \left\|x\right\|^2 \le K \left\|x\right\|,$$

which implies that when $\|x\| \leq 1$,

$$\|\nabla_\theta y(x;\theta)\| = \sqrt{\|\nabla_w y(x;\theta)\|^2 + \sum_{j=1}^{J} \left\|\nabla_{W_j} y(x;\theta)\right\|_F^2} \leq \sqrt{\|x\|^2 + K\sum_{j=1}^{J}\|x\|^2} \leq 3\|x\| K \leq 3K.$$

Hence, the parameterization $\theta \mapsto y(\cdot;\theta)$ is $(3K)$-Lipschitz, and the result follows from Lemma 13. □

The next lemma follows from similar arguments to those given in Wagenmaker and Jamieson [2022, Lemma A.12].

Lemma 13. *Let $f_\theta \colon \mathbb{R}^d \to \mathbb{R}$ be any function parameterized by $\theta \in \mathbb{R}^p$, and assume the mapping $\theta \mapsto f_\theta(\phi(s,a)) \in \mathbb{R}$ is L-Lipschitz for any s, a. Consider softmax policies $\pi_h^\theta(\cdot|\cdot) = \pi_h(\cdot|\cdot; f_\theta)$, $\pi_h^{\tilde\theta}(\cdot|\cdot) = \pi_h(\cdot|\cdot; f_{\tilde\theta}) \colon \mathcal{S} \to \Delta_{\mathcal{A}}$ as defined in Eq.* (11)*. Then, for any $\theta, \tilde\theta \in \mathbb{R}^p$, it holds that for any $s \in \mathcal{S}$:*

$$\left\|\pi_h^\theta(\cdot|s) - \pi_h^{\tilde\theta}(\cdot|s)\right\|_1 \leq 2L\left\|\theta - \tilde\theta\right\|_2.$$

Proof. Let $v^s(\theta) \coloneqq f_\theta(\phi(s,\cdot)) \in \mathbb{R}^A$, and let

$$\mathbf{J}v^s(\theta) \coloneqq \begin{pmatrix} \nabla_\theta f_\theta(\phi(s,a_1))^\top \\ \vdots \\ \nabla_\theta f_\theta(\phi(s,a_A))^\top \end{pmatrix} \in \mathbb{R}^{A\times p}$$

denote the Jacobian of v^s at $\theta \in \mathbb{R}^p$. Then, we have by the chain rule:

$$\nabla_\theta \pi_h^\theta(a|s) = \mathbf{J}v^s(\theta)^\top \nabla_u\left(\sigma(u)_a\right),$$

where $u \coloneqq v^s(\theta)$ and $\sigma(u)_i = e^{u_i}/(\sum_j e^{u_j})$ denotes the softmax function. Combining with the softmax gradient $\nabla_u\left(\sigma(u)_a\right) = \sigma(u)_a\left(e_a - \sigma(u)\right)$, we get

$$\left\|\nabla_\theta \pi_h^\theta(a|s)\right\| = \left(\sigma(u)_a\right)\left\|\mathbf{J}v^s(\theta)^\top\left(e_a - \sigma(u)\right)\right\| \leq 2\sigma(u)_a \max_a \|\nabla_\theta f_\theta(\phi(s,a))\| \leq 2L\pi_h^\theta(a|s),$$

where the last inequality uses our Lipschitz assumption and that $\sigma(u)_a = \pi_h^\theta(a|s)$. Now, by the mean-value theorem, we get that for some $\theta' \in [\theta, \tilde\theta]$,

$$\left|\pi_h^\theta(a|s) - \pi_h^{\tilde\theta}(a|s)\right| = \left|\nabla\pi_h^{\theta'}(a|s)\left(\theta - \tilde\theta\right)\right| \leq 2L\pi_h^{\theta'}(a|s)\left\|\theta - \tilde\theta\right\|_2,$$

which implies

$$\left\|\pi_h^\theta(\cdot|s) - \pi_h^{\tilde\theta}(\cdot|s)\right\|_1 \leq 2L\left\|\theta - \tilde\theta\right\|_2,$$

and completes the proof. □

Proof of Lemma 5. Let $\pi_h, \pi_h' \in \Pi$ be parameterized by $\pi_h(\cdot|\cdot) = \pi(\cdot|\cdot; y_h), \pi_h'(\cdot|\cdot) = \pi(\cdot|\cdot; y_h')$, where $y_h(\cdot) = y(\cdot; w, W_{1:J})$, $y_h'(\cdot) = y(\cdot; w', W_{1:J}')$, and consider $q, q' \in \widetilde{\mathcal{Q}}^\circ(\mathcal{Z}, C)$. For any $s \in \mathcal{Z}$, we have;

$$\left|\widetilde{V}(s;\pi_h,q) - \widetilde{V}(s;\pi_h',q')\right| \leq \left|\widetilde{V}(s;\pi_h,q) - \widetilde{V}(s;\pi_h,q')\right| + \left|\widetilde{V}(s;\pi_h,q') - \widetilde{V}(s;\pi_h',q')\right|.$$

For the first term,

$$\begin{aligned}\left|\widetilde{V}(s;\pi_h,q) - \widetilde{V}(s;\pi_h,q')\right| &\leq \max_a\left\{\left|\phi(s,a)^\top(w - w')\right| + \sqrt{\left|\phi(s,a)^\top(W - W')\,\phi(s,a)\right|}\right\} \\ &\leq \|w - w'\| + \sqrt{\|W - W'\|}. \end{aligned} \tag{18}$$

For the second term,

$$\begin{aligned}\left|\widetilde{V}(s;\pi_h,q') - \widetilde{V}(s;\pi_h',q')\right| &= \left|\langle \pi_h(\cdot|s) - \pi_h'(\cdot|s), q'(s,\cdot)\rangle\right| \\ &\leq C\left\|\pi_h(\cdot|s) - \tilde\pi_h(\cdot|s)\right\|_1 \\ &\leq 6CK\sqrt{\|w - w'\|^2 + \sum_{j=1}^{J}\left\|W_j - W_j'\right\|^2} \\ &\leq 6CK\left(\|w - w'\| + \sum_{j=1}^{J}\left\|W_j - W_j'\right\|\right),\end{aligned}$$

where the last inequality follows from Lemma 12. As per Eq. (11), we have that $w, w' \in \mathcal{B}^d(3dHK^2)$, $J \leq 2d \log K$, and $W_j, W_j' \in \mathcal{B}^{d^2}(\sqrt{d}\beta^2 K^2)$ for all $j \leq J$, where this last claim follows since the Frobenius norm of any matrix is larger than its spectral norm by a factor of at most $\sqrt{d}$. Thus, for simplicity, we consider covering the larger set given by $E \coloneqq \mathcal{B}^p(4dH\beta^2K^2)$ and $p = 4d^3 \log K$. By Lemma 17, given any ν, we have a $(\nu_1 = \nu/(12CK))$-covering with cardinality $\leq (1+(4dH\beta^2K^2)* 12CK/\nu)^p = (1 + 48dCH\beta^2K^3/\nu)^p$.

Similarly, we $\nu/4$ construct a cover corresponding to each of the terms in Eq. (18) with sets of cardinality $(1 + 64\beta^2/\nu^2)^{d^2}$ and $(1 + 16dHK^2/\nu)^d$. This gives,

$$\begin{aligned}\log \mathcal{N}_\nu(\widetilde{\mathcal{V}}^\circ(\mathcal{Z}, C)) &\leq p \log\left(1 + 48dCH\beta^2K^3/\nu\right) + 2d^2 \log\left(1 + 64\beta/\nu\right) + d\log(1 + 16dHK^2/\nu) \\ &\leq c_{\mathcal{N}} d^3 \log\left(\beta CK/\nu\right),\end{aligned}$$

for an appropriate constant $c_{\mathcal{N}}$, which completes the proof. □

A.5 Proof of Lemma 4

Proof of Lemma 4. For any k, we have by Lemma 16;

$$\begin{aligned}V_1^{k,\pi^k} - V_1^{k,\pi^\star} &= V_1^{k,\pi^k} - \widetilde{V}_1^{k;\circ} + \widetilde{V}_1^{k;\circ} - V_1^{k,\pi^\star} \\ &= \sum_{h=1}^H \mathbb{E}_{\mu_h^k}\left[\ell_h^k(s_h, a_h) + \mathbb{P}_h \widetilde{V}_{h+1}^{k;\circ}(s_h, a_h) - \widetilde{Q}_h^{k;\circ}(s_h, a_h)\right] \\ &\quad + \sum_{h=1}^H \mathbb{E}_{\mu_h^\star}\left[\left\langle \widetilde{Q}_h^{k;\circ}(s_h, \cdot), \pi_h^k(\cdot|s_h) - \pi_h^\star(\cdot|s_h)\right\rangle\right] \\ &\quad + \sum_{h=1}^H \mathbb{E}_{\mu_h^\star}\left[\widetilde{Q}_h^{k;\circ}(s_h, a_h) - \ell_h^k(s_h, a_h) - \mathbb{P}_h \widetilde{V}_{h+1}^{k;\circ}(s_h, a_h)\right].\end{aligned}$$

Now, note that for any $s \in \mathcal{Z}_h, a \in \mathcal{A}$;

$$\widetilde{Q}_h^{k;\circ}(s, a) = \phi(s, a)^\top \widehat{g}_{k,h} + \widehat{\mathbb{P}}_h^k \widetilde{V}_h^{k;\circ}(s, a) - \hat{b}_h^k(s, a),$$

thus

$$\ell_h^k(s, a) + \mathbb{P}_h \widetilde{V}_{h+1}^{k;\circ}(s, a) - \widetilde{Q}_h^{k;\circ}(s, a) = \phi(s, a)^\top (g_{k,h} - \widehat{g}_{k,h}) + \left(\mathbb{P}_h - \widehat{\mathbb{P}}_h^k\right) \widetilde{V}_{h+1}^{k;\circ}(s, a) + \hat{b}_h^k(s, a).$$

In addition, by the good event, specifically $\mathcal{E}^{\mathrm{qbd}}$, and the assumption that the instantaneous loss is $\in [-1, 1]$, we have for any $s \notin \mathcal{Z}_h, a \in \mathcal{A}$:

$$\left|\ell_h^k(s, a) + \mathbb{P}_h \widetilde{V}_{h+1}^{k;\circ}(s, a) - \widetilde{Q}_h^{k;\circ}(s, a)\right| = \left|\ell_h^k(s, a) + \mathbb{P}_h \widetilde{V}_{h+1}^{k;\circ}(s, a)\right| \leq 1 + H + 2 \leq 2H.$$

Thus by the law of total expectation,

$$\begin{aligned}&\mathbb{E}_{\mu_h^k}\left[\ell_h^k(s_h, a_h) + \mathbb{P}_h \widetilde{V}_{h+1}^{k;\circ}(s_h, a_h) - \widetilde{Q}_h^{k;\circ}(s_h, a_h)\right] \\ &\quad \leq \mathbb{E}_{\mu_h^k}\left[\ell_h^k(s_h, a_h) - \widehat{\ell}_h^k(s_h, a_h) + \left(\mathbb{P}_h - \widehat{\mathbb{P}}_h^k\right) \widetilde{V}_{h+1}^{k;\circ}(s_h, a_h) + \hat{b}_h^k(s, a) \mid s_h \in \mathcal{Z}_h\right] + 2\epsilon_{\mathrm{cov}} H,\end{aligned}$$

where the inequality follows since the good event $\mathcal{E}^{\mathrm{rfw}}$ implies $\mu_h^k(\mathcal{S} \setminus \mathcal{Z}_h) \leq \epsilon_{\mathrm{cov}}$, and for similar reasons;

$$\begin{aligned}&\mathbb{E}_{\mu_h^\star}\left[\widetilde{Q}_h^{k;\circ}(s_h, a_h) - \ell_h^k(s_h, a_h) - \mathbb{P}_h \widetilde{V}_{h+1}^{k;\circ}(s_h, a_h)\right] \\ &\quad \leq \mathbb{E}_{\mu_h^\star}\left[\widehat{\ell}_h^k(s_h, a_h) - \ell_h^k(s_h, a_h) + \left(\widehat{\mathbb{P}}_h^k - \mathbb{P}_h\right) \widetilde{V}_{h+1}^{k;\circ}(s_h, a_h) - \hat{b}_h^k(s, a) \mid s_h \in \mathcal{Z}_h\right] + 2\epsilon_{\mathrm{cov}} H.\end{aligned}$$

Finally, again by the law of total expectation and definition of the restricted Q-function;

$$\mathbb{E}_{s_h \sim \mu_h^\star}\left[\left\langle \widetilde{Q}_h^{k;\circ}(s_h, \cdot), \pi_h^k(\cdot|s_h) - \pi_h^\star(\cdot|s_h)\right\rangle\right] \leq \mathbb{E}_{s_h \sim \mu_h^\star}\left[\left\langle \widetilde{Q}_h^k(s_h, \cdot), \pi_h^k(\cdot|s_h) - \pi_h^\star(\cdot|s_h)\right\rangle \mid s_h \in \mathcal{Z}_h\right].$$

Combining the last three displays with the first equation and summing over $k = K_0, \ldots, K$ and $h \in [H]$ completes the proof. □

Lemma 14. *Upon execution of Algorithm 1, for all k, h it holds that*

$$\forall u \in \mathbb{R}^d; \quad \|u\|_{\Lambda_{k,h}^{-1}} \leq \|u\|_{\widehat{\Lambda}_{k,h}^{-1}} \leq \sqrt{2}\,\|u\|_{\Lambda_{k,h}^{-1}}\,.$$

Proof. By definition, we have at all times $\widehat{\Lambda}_{k,h} \preceq \Lambda_{k,h}$ and $\det \Lambda_{k,h} \leq 2 \det \widehat{\Lambda}_{k,h}$. Therefore, $\Lambda_{k,h}^{-1} \preceq \widehat{\Lambda}_{k,h}^{-1}$ and $\frac{\det \widehat{\Lambda}_{k,h}^{-1}}{\det \Lambda_{k,h}^{-1}} \leq 2$. Now, by Lemma 21, we have

$$\Lambda_{k,h}^{-1} \preceq \widehat{\Lambda}_{k,h}^{-1} \preceq 2\Lambda_{k,h}^{-1},$$

which completes the proof. □

A.6 Proof of Theorem 1

Proof of Theorem 1. Given our choice of parameters, by Lemma 1, we have that the number of warmup episodes satisfies

$$K_0 = O\left(\frac{d^4 H^5}{\epsilon_{\mathrm{cov}}} \log^7 \frac{dHK}{\delta}\right). \tag{19}$$

For the remainder of the proof, we assume the good event defined in Lemma 3 holds, which indeed occurs w.p. $1 - 4\delta$ by that lemma. Proceeding, we will bound the regret for the remaining rounds using the decomposition given by Lemma 4.

Bias term. By Eq. ($\mathcal{E}^{\mathrm{vbu}}$), we have for all s, a:

$$\left(\mathbb{P}_h - \widehat{\mathbb{P}}_h^k\right)\widetilde{V}_{h+1}^{k;\circ}(s,a) + \frac{1}{2}\hat{b}_h^k(s,a) \leq \frac{\beta}{2}\,\|\phi(s,a)\|_{\Lambda_{k,h}^{-1}} + \frac{1}{2}\hat{b}_h^k(s,a).$$

In addition, in the stochastic case, owed to Eq. ($\mathcal{E}^{\mathrm{sle}}$), for all s, a:

$$\begin{aligned}\ell_h(s,a) - \widehat{\ell}_h^k(s,a) &= \phi(s_h,a_h)^\top\left(g_{k,h} - \widehat{g}_{k,h}\right) + \frac{1}{2}\hat{b}_h^k(s,a) \\ &\leq \frac{\beta}{2}\,\|\phi(s,a)\|_{\Lambda_{k,h}^{-1}} + \frac{1}{2}\hat{b}_h^k(s,a).\end{aligned}$$

In the adversarial case, the above bound holds trivially since $\ell_h(s,a) = \widehat{\ell}_h^k(s,a)$. By a simple algebraic argument given in Lemma 14, we additionally have $\hat{b}_h^k(s,a) = \beta\,\|\phi(s,a)\|_{\widehat{\Lambda}_{k,h}^{-1}} \leq 2\beta\,\|\phi(s,a)\|_{\Lambda_{k,h}^{-1}}$, thus the sum of the last two displays is bounded by $3\beta\,\|\phi(s,a)\|_{\Lambda_{k,h}^{-1}}$, therefore

$$\begin{aligned}\text{Bias} &\leq 3\beta \sum_{h=1}^{H}\sum_{k=K_0}^{K} \mathbb{E}_{\mu_h^k}\left[\left\|\phi(s_h^k,a_h^k)\right\|_{\Lambda_{k,h}^{-1}}\right] \\ &\leq 6\beta \sum_{h=1}^{H}\sum_{k=K_0}^{K} \left\|\phi(s_h^k,a_h^k)\right\|_{\Lambda_{k,h}^{-1}} + 12\beta H \log\frac{4KH}{\delta},\end{aligned}$$

where the second inequality follows from Eq. ($\mathcal{E}^{\mathrm{bon}}$). Further, by Lemma 20, for any $h \in [H]$,

$$\sum_{k=K_0}^{K} \left\|\phi(s_h^k,a_h^k)\right\|_{\Lambda_{k,h}^{-1}} \leq \sqrt{K\sum_{k=K_0}^{K}\left\|\phi(s_h^k,a_h^k)\right\|_{\Lambda_{k,h}^{-1}}^2} \leq 2\sqrt{Kd\log K},$$

hence,

$$\text{Bias} \leq 12\beta H\left(\sqrt{Kd\log K} + \log\frac{4KH}{\delta}\right).$$

OMD Term. By Eq. ($\mathcal{E}^{\mathrm{qbd}}$) we have that for all $k \geq K_0, h, s \in \mathcal{Z}_h, a \in \mathcal{A}$; $\left|\widetilde{Q}_h^k(s,a)\right| = \left|\widetilde{Q}_h^{k;\circ}(s,a)\right| \leq 2H$. Thus, applying the OMD regret bound Lemma 24 for any $s \in \mathcal{Z}_h, h \in [H]$ we have;

$$\begin{aligned}
&\sum_{k=K_0}^{K} \left\langle \widetilde{Q}_h^k(s,\cdot), \pi_h^k(\cdot|s) - \pi_h^\star(\cdot|s) \right\rangle \\
&\qquad \leq \frac{\log A}{\eta} + \eta \sum_{k=K_0}^{K} \sum_{a\in\mathcal{A}} \pi_h^k(a|s) \widetilde{Q}_h^k(s,a)^2 \\
&\qquad \leq \frac{\log A}{\eta} + 4\eta H^2 K.
\end{aligned}$$

Therefore, we may bound the OMD term as follows:

$$\begin{aligned}
&\sum_{h=1}^{H} \sum_{k=K_0}^{K} \mathbb{E}_{s_h\sim\mu_h^\star} \left[\left\langle \widetilde{Q}_h^k(s_h,\cdot), \pi_h^k(\cdot|s_h) - \pi_h^\star(\cdot|s_h) \right\rangle \mid \mathcal{K}_h \right] \\
&= \sum_{h=1}^{H} \mathbb{E}_{s_h\sim\mu_h^\star} \left[\sum_{k=K_0}^{K} \left\langle \widetilde{Q}_h^k(s_h,\cdot), \pi_h^k(\cdot|s_h) - \pi_h^\star(\cdot|s_h) \right\rangle \mid \mathcal{K}_h \right] \\
&\leq \sum_{h=1}^{H} \mathbb{E}_{s_h\sim\mu_h^\star} \left[\frac{\log A}{\eta} + 4\eta H^2 K \mid \mathcal{K}_h \right] \\
&= \frac{H \log A}{\eta} + 4\eta H^3 K.
\end{aligned}$$

Optimism term. By Eq. ($\mathcal{E}^{\mathrm{vbu}}$), for $s, a \in \mathcal{Z}_h \times \mathcal{A}$:

$$\left(\widehat{\mathbb{P}}_h^k - \mathbb{P}_h\right)\widetilde{V}_{h+1}^{k;\circ}(s,a) - \frac{1}{2}\hat{b}_h^k(s,a) \leq \frac{\beta}{2}\left\|\phi(s,a)\right\|_{\Lambda_{k,h}^{-1}} - \frac{\beta}{2}\left\|\phi(s,a)\right\|_{\widehat{\Lambda}_{k,h}^{-1}} \leq 0,$$

since $\widehat{\Lambda}_{k,h}^{-1} \succeq \Lambda_{k,h}^{-1}$ by construction. Similarly, owed to Eq. ($\mathcal{E}^{\mathrm{sle}}$):

$$\phi(s,a)^\top\left(\widehat{g}_{k,h} - g_{k,h}\right) - \frac{1}{2}\hat{b}_h^k(s,a) \leq \frac{\beta}{2}\left\|\phi(s,a)\right\|_{\Lambda_{k,h}^{-1}} - \frac{\beta}{2}\left\|\phi(s,a)\right\|_{\widehat{\Lambda}_{k,h}^{-1}} \leq 0.$$

Thus, we immediately obtain the optimism term is non positive.

Concluding the proof. Combining the bound on the number of warmup episodes Eq. (19), with Lemma 4 and the bounds on all three terms, we have:

$$\begin{aligned}
\sum_{k=1}^{K} V^{\pi^k} - V^\star \lesssim \frac{d^4H^5}{\epsilon_{\mathrm{cov}}}\log^7\frac{dHK}{\delta} + \epsilon_{\mathrm{cov}}H^2K + \frac{H\log A}{\eta} \\
+ \eta H^3 K + \beta H\left(\sqrt{Kd\log K} + \log\frac{KH}{\delta}\right),
\end{aligned}$$

where $\lesssim$ hides only constant factors. Finally, setting $\epsilon_{\mathrm{cov}} = \frac{H^{3/2}d^2\log^4(dHK/\delta)}{\sqrt{K}}, \beta = 2c_\beta d^{3/2}H\log(dHK/\delta)$ and $\eta = \frac{\sqrt{\log A}}{H\sqrt{K}}$, we obtain:

$$\sum_{k=1}^{K} V^{\pi^k} - V^\star \lesssim d^2H^{7/2}\log^4\frac{dHK}{\delta}\sqrt{K\log A},$$

which completes the proof. □

B Proof of Lemma 1

In this section, we provide the technical details of the reward free algorithm guarantees. As mentioned, the algorithm is based on the work of Wagenmaker et al. [2022b] — the basic guarantee we build upon is formally stated below and follows immediately from Theorem 2 and Corollary 2 in their work. The bound on the number of episodes T follows from plugging the guarantees of FORCE [Wagenmaker et al., 2022a, Algorithm 1] into the precise setting of K_i given in the beginning of Appendix B of Wagenmaker et al. [2022b].

Theorem 2 (Wagenmaker et al., 2022b)**.** *The* COVERTRAJ *algorithm [Wagenmaker et al., 2022b, Algorithm 4] when instantiated with* FORCE *[Wagenmaker et al., 2022a, Algorithm 1] enjoys the following guarantee. Given a sequence of tolerance parameters $\gamma_1, \dots, \gamma_m > 0$ and $h \in [H]$, the algorithm interacts with the environment for T steps, where*

$$T \leq C \sum_{i=1}^{m} 2^i \max\left\{ \frac{d}{\gamma_i^2} \log \frac{2^i}{\gamma_i^2}, d^4 H^3 m^3 \log^{7/2} \frac{1}{\delta} \right\}, \quad C > 0 \text{ a constant,}$$

and outputs $\left\{ \left(\mathcal{X}_{h,i}, \widetilde{\mathcal{D}}_{h,i}, \widetilde{\Lambda}_{h,i} \right) \right\}_{i=1}^{m}$ such that $\bigcup_{i=1}^{m+1} \mathcal{X}_{h,i} = B_0^d(1)$ partitions the euclidean unit ball, $\widetilde{\Lambda}_{h,i} = I + \sum_{\tau \in \widetilde{\mathcal{D}}_{h,i}} \phi(s_h^\tau, a_h^\tau) \phi(s_h^\tau, a_h^\tau)^\top$, and with probability $1-\delta$, it holds that:

$$\forall i \in [m],\ \phi^\top \widetilde{\Lambda}_{h,i}^{-1} \phi \leq \gamma_i^2,\ \forall \phi \in \mathcal{X}_{h,i};$$

$$\text{and } \forall i \in [m+1], \sup_{\pi} \left\{ \int_{\mathcal{S} \times \mathcal{A}} \mathbf{I}\left\{ \phi(s,a) \in \mathcal{X}_{h,i} \right\} \mu_h^\pi(s,a) \right\} \leq 2^{-i+1}.$$

Lemma 15. *Assume $h \in [H], \epsilon, \delta > 0, \gamma_m \geq \dots \geq \gamma_1 > 0$, and let $\{\Lambda_{h,i}\}_{i \in [m]}$ be the covariate matrices returned from CovTraj$(h, \delta, m = \log(1/\epsilon), \{\gamma_i\})$. Then under the assumption that the event from Theorem 2 holds, we have for any policy π and $i \in [m]$:*

$$\Pr_{s_h \sim \mu_h^\pi} \left(\exists a \text{ s.t. } \|\phi(s_h, a)\|_{\Lambda_{h,i}^{-1}} > \gamma_m \right) \leq \epsilon.$$

Proof. By Theorem 2, we have that the total probability density induced by any policy $\pi \in [H] \times \mathcal{S} \to \Delta(\mathcal{A})$ on the last partition set $\mathcal{X}_{h,m+1}$ is at most $2^{-m} = \epsilon$. In addition, since on each of the remaining partition sets $\{\mathcal{X}_{h,i}\}_{i \in [m]}$ we have the guarantee that $\phi \in \mathcal{X}_{h,i} \implies \|\phi\|_{\Lambda_{h,i}^{-1}} \leq \gamma_i \leq \gamma_m$, it follows that,

$$\forall \pi; \quad \Pr_{s_h, a_h \sim \mu_h^\pi} \left(\|\phi(s_h, a_h)\|_{\Lambda_{h,i}^{-1}} > \gamma_m \right) = \Pr_{s_h, a_h \sim \mu_h^\pi} \left(\phi(s_h, a_h) \in \mathcal{X}_{h,m+1} \right) \leq \epsilon. \tag{2}$$

Assume by contradiction that π is a policy for which the statement of the theorem does not hold. Then

$$\Pr_{s_h \sim \mu_h^\pi} \left(\exists a, \|\phi(s_h, a)\|_{\Lambda_{h,i}^{-1}} > \gamma_m \right) > \epsilon.$$

But, if this happens, we can consider a transformed policy $\tilde{\pi}$; that rolls into timestep h with π, then takes (with probability 1) the action a that maximizes $\|\phi(s_h, a)\|_{\Lambda_{h,i}^{-1}}$. Formally, $\tilde{\pi}_{h'} = \pi_{h'}$ for all $h' \neq h$, and $\tilde{\pi}_h(a|s) = \mathbf{I}\left\{ a \in \arg\max_{a'} \|\phi(s, a')\|_{\Lambda_{h,i}^{-1}} \right\}$. This implies,

$$\Pr_{s_h, a_h \sim \mu_h^{\tilde{\pi}}} \left(\|\phi(s_h, a_h)\|_{\Lambda_h^{-1}} > \gamma_m \right) > \epsilon,$$

thus reaching a contradiction which completes the proof. □

Proof of Lemma 1. For the episode count, in order to apply Theorem 2, first note that given $\beta = O(d^{3/2} H \log(dHK/\delta)), \epsilon_{\text{cov}} \geq 1/K$, we have:

$$\forall i : \frac{d}{\gamma_i^2} \log \frac{2^i}{\gamma_i^2} = O(d\beta^2 H^2 \log(2^i \beta H)) = O(d\beta^2 H^2 \log^2(\beta HK)) = O(d^4 H^4 \log^4(dHK/\delta)).$$

In addition,

$$d^4 H^3 m^3 \log^{7/2} \frac{1}{\delta} = O\left(d^4 H^3 \log^3 K \log^{7/2} \frac{1}{\delta} \right) = O\left(d^4 H^3 \log^7 \frac{K}{\delta} \right)$$

Hence, we have that for every h, with T_h denoting the number of episodes run by CovTraj, by Theorem 2;

$$T_h = O\left(d^4H^4\log^7(dHK/\delta)\sum_{i=1}^{m}2^i\right) = O\left(2^{m+1}d^4H^4\log^7(dHK/\delta)\right) = O\left(\frac{d^4H^4}{\epsilon_{\mathrm{cov}}}\log^7(dHK/\delta)\right).$$

Given that Algorithm 2 executes CovTraj H times, the claim follows. For the claim on the unreachability of $\mathcal{S}\setminus\mathcal{Z}_h$, fix $h\in[H]$, and observe that by Lemma 15, w.p. $1-\delta/H$, for any π;

$$\Pr_{s_h\sim\mu_h^\pi}(s_h\notin\mathcal{Z}_h) = \Pr_{s_h\sim\mu_h^\pi}\left(\exists a \text{ s.t. } \|\phi(s_h,a)\|_{\Lambda_{0,h}^{-1}} > \gamma_m\right) \le \epsilon_{\mathrm{cov}},$$

where in the inequality we use that $\widetilde{\Lambda}_{h,i}\preceq\Lambda_{0,h}$. The proof is complete by a union bound over h. □

C Auxiliary Lemmas

Lemma 16 (Extended value difference; Shani et al., 2020, Lemma 1; Cai et al., 2020)**.** *Let $M=(H,\mathcal{S},\mathcal{A},\mathbb{P},\ell)$ be any MDP and $\pi,\pi'\in\mathcal{S}\to\Delta_{\mathcal{A}}$ be any two policies. Then, for any sequence of functions $\widehat{Q}_h^\pi\colon\mathcal{S}\times\mathcal{A}\to\mathbb{R}, V_h^\pi\colon\mathcal{S}\to\mathbb{R}$, where $\widehat{V}_h^\pi(s)\coloneqq\left\langle\pi(\cdot|s),\widehat{Q}_h(s,\cdot)\right\rangle$, $h=1,\ldots,H$, we have*

$$\begin{aligned}\widehat{V}_1^\pi - V_1^{\pi'} &= \sum_{h=1}^{H}\mathbb{E}_{s_h,a_h\sim d_h^{\pi'}}\left[\left\langle\widehat{Q}_h^\pi(s_h,\cdot),\pi_h(\cdot|s_h)-\pi'_h(\cdot|s_h)\right\rangle\right]\\ &\quad+\sum_{h=1}^{H}\mathbb{E}_{s_h,a_h\sim d_h^{\pi'}}\left[\widehat{Q}_h^\pi(s_h,a_h)-\ell_h(s_h,a_h)-\mathbb{P}\widehat{V}_{h+1}^\pi(s_h,a_h)\right].\end{aligned}$$

Lemma 17 (Covering number of Euclidean Ball)**.** *For any $\epsilon>0$, the ϵ-covering of the Euclidean ball in $\mathbb{R}^d$ with radius $R>0$ is upper bounded by $(1+2R/\epsilon)^d$.*

Lemma 18. *Let $R,z\ge1$, and $x\ge2z\log(Rz)$. Then $z\log(Rx)\le x$.*

Proof. If $x=2z\log(Rz)$;

$$\begin{aligned}z\log(Rx) &= z\log R + z\log(2z\log(Rz))\\ &= z\log R + z\log(2z) + z\log\log(Rz)\\ &\le z\log R + z\log z + z\log(Rz)\\ &= 2z\log R + 2z\log z\\ &= x.\end{aligned}$$

For larger values, the result follows by noting $x-z\sqrt{\log(Rx)}$ is monotonically increasing in x for all $x\ge z$. □

Lemma 19 (Lemma D.4 in Rosenberg et al., 2020)**.** *Let $(\mathcal{F}_k)_{k=1}^\infty$ be a filtration, and let $(X_k)_{k=1}^\infty$ be a sequence of random variables that are $\mathcal{F}_k$-measurable, and supported on $[0,B]$. Then with probability $\ge1-\delta$, we have that for any $K\ge1$;*

$$\sum_{k=1}^{K}\mathbb{E}\left[X_k\mid\mathcal{F}_{k-1}\right]\le2\sum_{k=1}^{K}X_k+4B\log\frac{2K}{\delta}.$$

Lemma 20 (Elliptical potential lemma, Lattimore and Szepesvári, 2020, Lemma 19.4)**.** *Let $(\phi_i)_{k=1}^K\subset\mathbb{R}^d$ with $\|\phi_k\|\le1$, and set $\Lambda_k\coloneqq\lambda I+\sum_{i=1}^{k-1}\phi_i\phi_i^\top$ where $\lambda\ge1$. Then,*

$$\sum_{k=1}^{K}\|\phi_i\|^2_{\Lambda_k^{-1}}\le2d\log\left(1+\frac{K}{d\lambda}\right)$$

Proof. Note that $\lambda\ge1$ implies $\|\phi_i\|^2_{\Lambda_k^{-1}}\le\lambda_{\max}(\Lambda_k^{-1})\|\phi_i\|^2\le\lambda^{-1}\le1$. Thus

$$\sum_{k=1}^{K}\|\phi_i\|^2_{\Lambda_k^{-1}}=\sum_{k=1}^{K}\min\left\{1,\|\phi_i\|^2_{\Lambda_k^{-1}}\right\}.$$

The rest of the proof is identical to Lattimore and Szepesvári [2020], with $L=1$ and $V_0=\lambda I$. □

Lemma 21 (Cohen et al., 2019, Lemma 27)**.** *For any two matrices $A, B \in \mathbb{R}^{d\times d}$ which satisfy $0 \preceq A \preceq B$, we have $B \preceq \frac{\det B}{\det A} A$.*

The following lemma is a direct consequence of the concentration of Self-Normalized Processes due to Abbasi-Yadkori et al. [2011].

Lemma 22. *Let $k \in \mathbb{N}$ and let $\ell\colon \mathbb{R}^d \to \mathbb{R}$ denote a linear function $\ell(\phi) = \phi^\top g^\star$, $g^\star \in \mathbb{R}^d$. Assume $\{\mathcal{F}_i\}_{i=1}^k$ is a filtration, and that $\phi_i \in \mathcal{F}_{i-1}$ is an $\mathbb{R}^d$ valued stochastic process with $\|\phi_i\| \le 1$. Further, assume $\ell^i = \ell(\phi_i) + \xi_i$ where ξ_i is a random variable such that $\mathbb{E}[\xi_i \mid \mathcal{F}_{i-1}] = 0$, and $\left|\ell^i\right| \le D$ almost surely. Then for any $\delta > 0$, w.p. $1-\delta$, we have*

$$\left\|\sum_{\tau=1}^{k} \phi_\tau \Big(\ell^\tau - \ell(\phi_\tau)\Big)\right\|^2_{\Lambda_k^{-1}} \le 2D^2 d \log\left(\frac{k+\lambda}{\lambda}\right),$$

where $\Lambda_k = \lambda I + \sum_{i=1}^k \phi_i \phi_i^\top$

The next lemma establishes the *uniform* concentration of least squares solutions over a *class* of functions, and follows from a standard covering argument combined with the concentration of Self-Normalized Processes Abbasi-Yadkori et al. [2011].

Lemma 23 (**OLS uniform concentration**; Jin et al., 2020, Lemma D.4)**.** *Let $\{\mathcal{F}_\tau\}_{\tau=1}^\infty$ be a filtration. Let $\{x_\tau\}$ be a stochastic process on state space $\mathcal{S}$ that is $\mathcal{F}_\tau$-measurable, and $\{\phi_\tau\}$ be an $\mathbb{R}^d$-valued stochastic process that is $\mathcal{F}_{\tau-1}$-measurable and satisfies $\|\phi_\tau\| \le 1$. Further, let $\Lambda_n = \lambda I + \sum_{\tau=1}^n \phi_\tau \phi_\tau^\top$. Then for any $\delta > 0$, with probability at least $1-\delta$, for all $n \ge 1$ and any $V \in \mathcal{V}$ so that $\|V\|_\infty \le D$, we have;*

$$\left\|\sum_{\tau=1}^{n} \phi_\tau \Big(V(x_\tau) - \mathbb{E}\left[V(x_\tau)|\mathcal{F}_{\tau-1}\right]\Big)\right\|^2_{\Lambda_n^{-1}} \le 4D^2\left(\frac{d}{2}\log\left(\frac{n+\lambda}{\lambda}\right) + \log\frac{\mathcal{N}_\epsilon(\mathcal{V})}{\delta}\right) + \frac{8n^2\epsilon^2}{\lambda},$$

where $\mathcal{N}_\epsilon(\mathcal{V})$ is the $\|\cdot\|_\infty$ ϵ-covering number of $\mathcal{V}$.

The next lemma is standard, for proof see e.g., Hazan et al. [2016], Lattimore and Szepesvári [2020].

Lemma 24 (**Entropy regularized OMD**)**.** *Let $y_1, \ldots, y_T \in \mathbb{R}^A$ be any sequence of vectors, and $\eta > 0$ such that $\eta y_t(a) \ge -1$ for all $t \in [T], a \in [A]$. Then if $\{x_t\} \subset \Delta_A$ is given by $x_1(a) = 1/n \forall a$, and for $t \ge 1$:*

$$x_{t+1}(a) = \frac{x_t(a) e^{-\eta y_t(a)}}{\sum_{a' \in [A]} x_t(a') e^{-\eta y_t(a')}},$$

then,

$$\max_{x \in \Delta_A}\left\{\sum_{t=1}^{T} \langle y_t, x_t - x\rangle\right\} \le \frac{\log A}{\eta} + \eta \sum_{k=1}^{K}\sum_{a=1}^{A} x_t(a) y_t(a)^2.$$