Optimistic Actor-Critic with Parametric Policies for Linear Markov Decision Processes

Max Qiushi Lin
Simon Fraser University

Reza Asad
Simon Fraser University

Kevin Tan
University of Pennsylvania

Haque Ishfaq
Mila, McGill University

Csaba Szepesvári
Google DeepMind, University of Alberta

Sharan Vaswani
Simon Fraser University

Abstract

Although actor-critic methods have been successful in practice, their theoretical analyses have several limitations. Specifically, existing theoretical work either sidesteps the exploration problem by making strong assumptions or analyzes impractical methods with complicated algorithmic modifications. Moreover, the actor-critic methods analyzed for linear MDPs often employ natural policy gradient and construct "implicit" policies without explicit parameterization. Such policies are computationally expensive to sample from, making the environment interactions inefficient. To that end, we focus on the finite-horizon linear MDPs and propose an optimistic actor-critic framework that uses parametric log-linear policies. In particular, we introduce a tractable *logit-matching* regression objective for the actor. For the critic, we use approximate Thompson sampling via Langevin Monte Carlo to obtain optimistic value estimates. We prove that the resulting algorithm achieves $\widetilde{\mathcal{O}}(\epsilon^{-4})$ and $\widetilde{\mathcal{O}}(\epsilon^{-2})$ sample complexity in the on-policy and off-policy setting, respectively. Our results match prior theoretical work in achieving the state-of-the-art sample complexity, while our algorithm is more aligned with practice.

1 INTRODUCTION

Reinforcement learning (RL) is a general framework for sequential decision making under uncertainty and has been successful in various real-world applications, such as robotics (Kober et al., 2013) and large language models (Uc-Cetina et al., 2023). Policy Gradient (PG) methods (Williams, 1992; Sutton et al., 1999; Kakade, 2001; Schulman et al., 2017a) are an important class of algorithms that assume a differentiable parameterization of the policy, and directly optimize the policy parameters using the return from interacting with the environment. PG methods are widely used in practice as they can easily handle function approximation or structured state-action spaces. However, since the environment is typically stochastic in practice, the estimated returns usually have high variance, resulting in poor sample efficiency (Dulac-Arnold et al., 2019).

Actor-critic (AC) methods (Konda and Tsitsiklis, 1999; Peters et al., 2005; Bhatnagar et al., 2009) alleviate this issue by using value-based approaches in conjunction with PG methods. In particular, they utilize a critic that estimates the policy's value and an actor that performs PG to improve the policy towards obtaining higher returns. These AC methods have been proven to be empirically successful in both on-policy (Schulman et al., 2015, 2017b) and off-policy (Lillicrap et al., 2015; Fujimoto et al., 2018; Haarnoja et al., 2018) settings.

Subsequently, there have been many attempts to provide a theoretical understanding of actor-critic methods, especially in the presence of function approximation (Cai et al., 2020; Zhong and Zhang, 2023; Liu et al., 2023). However, there are two prevalent issues that result in mismatches between theory and practice: the studied methods either (i) do not consider strategic exploration in a systematic manner or (ii) analyze complicated and impractical variants of the algorithm. In particular, much of the literature makes unrealistic assumptions to avoid dealing with exploration, a central challenge in RL. For instance, existing works on PG methods (Agarwal et al., 2021a; Yuan et al., 2023; Alfano and Rebeschini, 2022; Asad et al., 2025) obtain convergence rates that involve a mismatch ratio between the optimal policy and the initial state

Proceedings of the 29th International Conference on Artificial Intelligence and Statistics (AISTATS) 2026, Tangier, Morocco. PMLR: Volume 300.

Algorithm	Sample Complexity (On-Policy)	Sample Complexity (Off-Policy)	Policy Param.	Clipping Q-Function	Computational Cost for Policy Inference
Liu et al. (2023)	$\widetilde{\mathcal{O}}(\frac{1}{\epsilon^4})$	✗	implicit	yes	$\mathcal{O}(d_{\boldsymbol{c}}^2\, H\, \lvert\mathcal{A}\rvert\, \epsilon^{-2})$
Sherman et al. (2023) Cassel and Rosenberg (2024)	✗	$\widetilde{\mathcal{O}}(\frac{1}{\epsilon^2})$	implicit	no	$\mathcal{O}(d_{\boldsymbol{c}}^2\, H\, \lvert\mathcal{A}\rvert\, \epsilon^{-2})$
Ours	$\widetilde{\mathcal{O}}(\frac{1}{\epsilon^4})$	$\widetilde{\mathcal{O}}(\frac{1}{\epsilon^2})$	explicit	yes	$\mathcal{O}(d_{\boldsymbol{a}}^2\, H\, \lvert\mathcal{A}\rvert)$

Table 1: Comparison with the state-of-the-art algorithms for episodic finite-horizon linear MDPs (with feature dimension $d_{\boldsymbol{c}}$ and horizon H). The sample complexity refers to the number of interactions required for outputting an ϵ-optimal policy, whereas the cost of policy inference refers to the per-episode cost for interacting with the environment for H steps. Our proposed algorithm matches the optimal sample efficiency in both the on- and off-policy settings. Furthermore, in contrast to existing works, our algorithm employs an explicit policy parameterization (log-linear policies with dimension $d_{\boldsymbol{a}}$), resulting in lower computational cost for policy inference.

distribution. These results are only meaningful if the mismatch ratio is bounded. However, a bounded mismatch ratio indicates that the initial state distribution already provides a good coverage over the state space, thereby sidestepping the exploration problem. Within actor-critic methods, some early analyses make assumptions on the reachability of the state-action space or the coverage of collected data (Abbasi-Yadkori et al., 2019; Neu et al., 2017; Bhandari and Russo, 2024; Agarwal et al., 2021a; Cen et al., 2022; Gaur et al., 2024), which again imply that the state-action space is already relatively easy to explore. Follow-up works (Hong et al., 2023; Fu et al., 2021; Xu et al., 2020; Cayci et al., 2024) assume a bounded mismatch ratio, while others (Khodadadian et al., 2022; Gaur et al., 2023) require mixing assumptions on the induced Markov chain.

On the other hand, recent works (Cai et al., 2020; Jin et al., 2021; Zanette et al., 2021; Zhong and Zhang, 2023; Agarwal et al., 2023; He et al., 2023; Liu et al., 2023; Sherman et al., 2023; Cassel and Rosenberg, 2024; Tan et al., 2025) tackle the exploration issue directly. However, the algorithms analyzed are significantly different from those implemented in practice. Much of this body of work studies AC methods that use the natural policy gradient (NPG) update for policy optimization. However, the canonical implementation of the NPG update does not consider an explicit policy parameterization. Instead, the update involves constructing "implicit" policies *on the fly* using all previously stored Q-functions. This makes it computationally expensive to sample from these policies and use them to interact with the environment. In contrast, the algorithms used in practice typically employ explicitly parameterized complex models as learnable policies and optimize them with gradient descent-based methods. Therefore, we aim to address the following question:

Can we design a provably efficient actor–critic algorithm with parametric policies for linear MDPs in both on- and off-policy settings?

Contributions. We answer the above question affirmatively and make the following contributions.

1. General framework with an explicitly parameterized actor. In Section 3, we propose a general optimistic actor-critic framework that utilizes an explicitly parameterized policy. We apply this framework in the setting of linear function approximation for both the environment (i.e., linear MDP (Jin et al., 2020)) and the policy (i.e., log-linear policy class). In Section 4, we propose an actor algorithm that learns a log-linear policy by solving a specific regression problem at each iteration. This allows us to directly control the error between the explicitly parametrized policy and the implicit policy induced by NPG. Using this error bound in conjunction with the well-established theoretical results of NPG (Hazan et al., 2016; Szepesvári, 2022) enables us to analyze the performance of the parameterized actor. We show that the proposed algorithm benefits from a substantially improved memory complexity, while retaining similar theoretical guarantees.

2. LMC critic for practical strategic exploration. In Section 5, instead of constructing UCB bonuses (Jin et al., 2020), which are ubiquitous within prior works (Zhong and Zhang, 2023; Liu et al., 2023; Sherman et al., 2023; Cassel and Rosenberg, 2024), we adopt a more practical approach. We employ Langevin Monte Carlo (LMC) (Welling and Teh, 2011) to update the critic parameters at each episode. Unlike UCB-based approaches that require computing confidence sets at every episode, LMC simply perturbs (by Gaussian noise) the gradient descent update on the critic loss. This gradient descent-based approach is both easier to implement (Ishfaq et al., 2025) and to extend to general function approximation (Ishfaq et al., 2024b). Furthermore, the LMC algorithm directly leads to an optimistic estimate of the Q-function that has similar guarantees as UCB bonuses. Nevertheless, previous works have only successfully designed provably efficient algorithms for solving multi-armed bandits (Mazumdar et al., 2020),

contextual bandits (Xu et al., 2022), and linear MDPs via value-based methods (Ishfaq et al., 2024a). Our paper is the first to analyze an LMC-based approach in the context of policy optimization.

3. End-to-end theoretical guarantees for actor-critic. In Section 6, we analyze the proposed actor-critic framework in both the on-policy and off-policy settings without making any assumptions on the mismatch ratio or data coverage. In particular, in the on-policy setting, we prove that our method requires $\widetilde{\mathcal{O}}(\epsilon^{-4})$ samples to learn an ϵ-optimal policy. This matches the result in (Liu et al., 2023) that uses an implicit NPG policy in conjunction with UCB bonuses. On the other hand, we also prove that our framework can attain a sample complexity of $\widetilde{\mathcal{O}}(\epsilon^{-2})$ in the off-policy setting. This matches the results of Sherman et al. (2023); Cassel and Rosenberg (2024), but with a far less complicated algorithm design.

We thus demonstrate that our proposed algorithm is both sample-efficient and aligned with practice.

2 PRELIMINARIES

In this section, we introduce the finite-horizon linear MDP setting and the log-linear policy class.

Finite-Horizon Linear MDP. A finite-horizon MDP is a tuple $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathbb{P}, r, H)$ where $\mathcal{S}$ denotes the state space, $\mathcal{A}$ is the action set, and $H \in \mathbb{Z}_+$ is the length of the horizon. $\mathbb{P} = \{\mathbb{P}_h\}_{h\in[H]}$ is a set of time-dependent transition kernels, and $r = \{r_h\}_{h\in[H]}$ denotes a sequence of reward functions. We assume that the state space $\mathcal{S}$ is a (possibly infinite) measurable space, whereas $\mathcal{A}$ is a finite set with cardinality $|\mathcal{A}|$. $\mathbb{P}_h(\cdot \mid s, a) \in \Delta(\mathcal{S})$ is the distribution over states when taking action $a \in \mathcal{A}$ in state $s \in \mathcal{S}$ at step $h \in [H]$, and $r_h(s, a) \in [0, 1]$ is the corresponding reward. Additionally, for any given function $V : \mathcal{S} \to \mathbb{R}$, we define that $[\mathbb{P}_h V_{h+1}](s, a) \coloneqq \mathbb{E}_{s'\sim\mathbb{P}_h(\cdot|s,a)} V_{h+1}(s')$.

The agent interacts with the environment by starting at an initial state (w.l.o.g., fixed to be $s_1 \in \mathcal{S}$). At step h, the agent first observes the current state $s_h \in \mathcal{S}$, then takes an action $a_h \in \mathcal{A}$ and receives the reward $r_h(s_h, a_h)$. After that, the agent transitions to $s_{h+1} \sim \mathbb{P}_h(\cdot \mid s_h, a_h)$. The agent follows a given policy $\pi : [H] \times \mathcal{S} \mapsto \Delta(\mathcal{A})$ in which $\pi_h(\cdot \mid s) \in \Delta(\mathcal{A})$ is the probability distribution over $\mathcal{A}$ in state s at step h. To quantify the performance of any policy π, we define the value function as $V_h^\pi(s) \coloneqq \mathbb{E}_{\pi,\mathbb{P}}[\sum_{\tau=h}^{H} r_\tau(s_\tau, a_\tau) \mid s_h = s]$, and the corresponding state-action value function is defined as $Q_h^\pi(s, a) \coloneqq \mathbb{E}_{\pi,\mathbb{P}}[\sum_{\tau=h}^{H} r_\tau(s_\tau, a_\tau) \mid s_h = s, a_h = a]$, where the expectation is with respect to the randomness in the stochastic policy and the transition dynamics. The value function (resp. Q-function) corresponds to the expected cumulative rewards when starting in state s (resp. state-action (s, a)) at step h, and subsequently following the policy π until reaching step H.

We assume that both $\mathbb{P}$ and r are unknown to the agent. In order to efficiently learn these quantities, we consider the linear MDP assumption (Jin et al., 2020) where both the transition kernel and the reward function are assumed to be linear functions of given features.

Definition 2.1 (Linear MDP)**.** *A finite-horizon MDP $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathbb{P}, r, H)$ is a linear MDP with a feature map $\phi : \mathcal{S} \times \mathcal{A} \mapsto \mathbb{R}^{d_\mathbf{c}}$ if the following holds. There exist H signed measures $\psi_h : \mathcal{S} \to \mathbb{R}^{d_\mathbf{c}}$ and $\upsilon_h : \mathbb{R}^{d_\mathbf{c}}$ such that $\mathbb{P}_h(s' \mid s, a) = \langle \phi(s, a), \psi_h(s') \rangle$ and $r_h(s, a) = \langle \phi(s, a), \upsilon_h \rangle$ for all h, s, and a. It should also satisfy the following constraints: $\|\phi(s, a)\|_2 \le 1$ and $\|\upsilon_h\|_2 \le \sqrt{d_\mathbf{c}}$ for all h, s, and a. Additionally, for any measurable function $V : \mathcal{S} \to [0, 1]$, $\left\|\int_{s\in\mathcal{S}} V(s)\, \psi_h(s)\, \mathrm{d}s\right\|_2 \le \sqrt{d_\mathbf{c}}$.*

According to Jin et al. (2020, Proposition 2.3), for a linear MDP and any policy π, Q_h^π is a linear function of the features: for all (h, s, a), there exists a $w_h \in \mathbb{R}^{d_\mathbf{c}}$ such that $Q_h^\pi(s, a) = \langle \phi(s, a), w_h \rangle$.

Learning Objective. For this linear MDP setting, we assume that only ϕ is given to the learner whereas ψ and υ are not. The agent sequentially interacts with the environment for T episodes and aims to minimize the *cumulative regret* defined as $\mathtt{Reg}(T) \coloneqq \sum_{t=1}^{T}[V_1^\star(s_1) - V_1^{\pi_t}(s_1)]$, where $V_1^\star \coloneqq V_1^{\pi^\star} \coloneqq \sup_\pi V_1^\pi$ is the value function of the optimal policy $\pi^\star \coloneqq \arg\sup_\pi V_1^\pi(s_1)$. Equivalently, if $\overline{\pi}^T$ denotes the mixture policy that picks a policy among $\{\pi^1, \ldots, \pi^T\}$ uniformly randomly, we aim to learn an ϵ-optimal $\overline{\pi}^T$, i.e., its *optimality gap* (OG) is bounded such that

$$\mathtt{OG}(T) \coloneqq \mathbb{E}\left[V_1^\star(s_1) - V_1^{\overline{\pi}^T}(s_1)\right] = \frac{\mathtt{Reg}(T)}{T} \le \widetilde{\mathcal{O}}(\epsilon)\,,$$

where the expectation is taken with respect to the randomness of the mixture policy.

Log-Linear Policy. We consider a restricted policy class Π_{lin} consisting of *log-linear policies*. Log-linear policies are represented using the softmax function with linear function approximation. In particular, a log-linear policy is defined as follows: for all $h \in [H]$,

$$\pi_h(a \mid s, \theta) = \frac{\exp(z_h(s, a \mid \theta_h))}{\sum_{a'\in\mathcal{A}} \exp(z_h(s, a' \mid \theta_h))}\,, \tag{1}$$

where $z_h(s, a \mid \theta_h) = \langle \varphi(s, a), \theta_h \rangle$ represents the logits parameterized by θ_h, and $\varphi : \mathcal{S} \times \mathcal{A} \mapsto \mathbb{R}^{d_\mathbf{a}}$ are policy features given to the learner. W.l.o.g, we assume that $\|\varphi(s, a)\| \le 1$ for all s and a. For convenience, we use the shorthand $\pi(\theta) : [H] \times \mathcal{S} \to \Delta(\mathcal{A})$ to refer to the log-linear policy corresponding to the parameters θ.

Algorithm 1 Actor-Critic with Parametric Policies

1: **Input**: number of update steps T, data collection batch size N (only for on-policy)
2: set $\mathcal{D}^0 \leftarrow \emptyset$, $w_h^1 \leftarrow \mathbf{0}$, $\pi_h^1(\cdot \mid s) \leftarrow \text{Unif}(\mathcal{A}) \quad \forall (h,s)$
3: **for** $t = 1, \dots, T-1$ **do**
4: $\quad \mathcal{D}^t \leftarrow \begin{cases} \text{On-Policy:} & \{N \text{ fresh traj.} \overset{\text{i.i.d.}}{\sim} \pi^t\} \\ \text{Off-Policy:} & \mathcal{D}^{t-1} \cup \{1 \text{ traj.} \sim \pi^t\} \end{cases}$
5: $\quad w^{t+1} \leftarrow \texttt{CRITIC\_UPDATE}(\mathcal{D}^t, \pi^t, w^t)$
6: $\quad \theta^{t+1} \leftarrow \texttt{ACTOR\_UPDATE}(w^{t+1}, \theta^t)$
7: $\quad \pi^{t+1} = \pi(\theta^{t+1})$
8: **Return**: mixture policy $\overline{\pi}^T$

3 A GENERAL ACTOR-CRITIC FRAMEWORK WITH PARAMETRIC POLICIES

In this section, we start by introducing our general optimistic actor-critic framework as shown in Algorithm 1. Starting with a uniform policy π^1, at the beginning of every learning episode $t \in [T]$, the agent interacts with the environment using policy π^t (Line 4). Our framework allows for collecting data from the environment in either an on-policy or off-policy fashion. In the on-policy setting, at episode t, the agent collects N fresh trajectories $\mathcal{D}^t$ by interacting with the environment using the current policy π^t. On the other hand, in the off-policy setting, at episode t, the agent collects only 1 trajectory from the environment using π^t. However, the agent stores all the historical data collected by the previous policies, and hence, $\mathcal{D}^t$ consists of t trajectories, each collected by $\pi^1, \dots, \pi^t$ respectively.

The *critic* uses the collected data and estimates an (optimistic) Q-function via learning the critic parameters $w^{t+1} \in [H] \times \mathbb{R}^{d_c}$ (Line 5). The *actor* then uses the estimated Q-function, and updates the parameters of $\theta^t \in [H] \times \mathbb{R}^{d_a}$ of the log-linear policy (Line 6). The updated log-linear policy is denoted by π^{t+1} (Line 7), and is used to collect data in the next episode.

Given this general framework, we will next instantiate the actor in Section 4 and the critic in Section 5.

4 INSTANTIATING THE ACTOR: PROJECTED NATURAL POLICY GRADIENT

In this section, we instantiate the actor using natural policy gradient (NPG) with parametric policies and analyze its behavior. In particular, in Section 4.1, we devise an algorithm that projects the standard NPG update onto the class of realizable policies. In Section 4.2, we analyze and control the errors induced by the projection step. Finally, in Section 4.3, we put everything together and instantiate the complete actor algorithm for the log-linear policy class.

4.1 Projected Natural Policy Gradient

At episode $t \in [T]$, given $\widehat{Q}_h^t$, the estimated Q-function, NPG updates the policy as: for every $(h,s) \in [H] \times \mathcal{S}$,

$$\pi_h^{t+1}(\cdot|s) \propto \pi_h^t(\cdot|s) \exp(\eta\, \widehat{Q}_h^t(s,\cdot)) \tag{2}$$

with the corresponding normalization across $\mathcal{A}$. Existing work on policy optimization in linear MDPs (Liu et al., 2023; Sherman et al., 2023; Cassel and Rosenberg, 2024) uses NPG to update the actor because of its favorable theoretical properties. Importantly, *these results do not consider any explicit parameterization for the actor*. Directly implementing the update in Eq. (2) requires $\mathcal{O}(|\mathcal{S}||\mathcal{A}|)$ memory, and is therefore impractical with large state-action spaces.

Consequently, existing work uses the following equivalent form of the NPG update: for every $(h,s) \in [H] \times \mathcal{S}$,

$$\pi_h^{t+1}(\cdot|s) \propto \pi_h^1(\cdot|s) \exp\Big(\eta \textstyle\sum_{i=1}^t \widehat{Q}_h^i(s,\cdot)\Big),$$

which characterizes the policy implicitly. In particular, at episode t, for any (h,s,a), we can compute the policy *on the fly* if we have access to the sum of all the parameterized Q-functions up to episode t. However, in existing works (Liu et al., 2023; Sherman et al., 2023; Cassel and Rosenberg, 2024), the sum of parameterized Q functions cannot be stored in a succinct manner, and they require storing *all* the parameterized Q functions. Consequently, when interacting with the environment using such an implicit policy, these works suffer from an extensive per-episode computational cost of $\mathcal{O}(d_{\mathbf{c}}^2\, H\, |\mathcal{A}|\, T)$. Therefore, the resulting algorithm is far from practice that typically uses an explicit (and often sophisticated) actor parameterization.

To alleviate these issues, we aim to compute a policy that is (i) realizable by the explicit actor parameterization and (ii) provably approximates the policy induced by the NPG update in Eq. (2) (referred to as the *implicit policy*). To this end, we use a projected NPG update:

$$\pi_h^{t+1}(\cdot \mid s) = \text{Proj}_{\Pi}\left[\frac{\pi_h^t(\cdot \mid s) \exp(\eta\, \widehat{Q}_h^t(s,\cdot))}{\sum_{a'} \pi_h^t(a' \mid s) \exp(\eta\, \widehat{Q}_h^t(s,a'))}\right],$$

where Proj is the projection operator, which will be instantiated subsequently in Section 4.2.

When theoretically analyzing policy optimization methods, an important intermediate result is the bound on the regret for a specific online linear optimization problem. For the standard NPG update in Eq. (2), this regret can be bounded by $\widetilde{\mathcal{O}}(\sqrt{T})$ (Hazan et al., 2016; Szepesvári, 2022). In the following lemma, we analyze the effect of the projection operator and bound the regret for the projected NPG.

Theorem 4.1. *Given a sequence of linear functions $\{\langle p^t, g^t\rangle\}_{t\in[T]}$ for a sequence of vectors $\{g^t\}_{t\in[T]}$ where for any $t \in [T]$, $p^t \in \Delta(\mathcal{A})$, $g^t \in \mathbb{R}^{|\mathcal{A}|}$, and $\|g^t\|_\infty \le H$. Consider $p^{t\in[T]}$ where p^1 is the uniform distribution, and for all $t \in [T]$,*

$$p^{t+1/2} = \arg\min_{p\in\Delta_A}\left\{\langle p, -\eta\, g^t\rangle + \mathrm{KL}(p \,\|\, p^t)\right\}, \quad (3)$$

$$p^{t+1} = \mathrm{Proj}_\Pi(p^{t+1/2}). \quad (4)$$

Let $\epsilon^t := \mathrm{KL}(u \,\|\, p^{t+1}) - \mathrm{KL}(u \,\|\, p^{t+1/2})$ be the projection error induced by Eq. (4). Then, for any comparator $u \in \Delta(\mathcal{A})$, it holds that

$$\sum_{t=1}^T \langle u - p^t, g^t\rangle \le \frac{\log|\mathcal{A}| + \sum_{t=1}^T \epsilon^t}{\eta} + \frac{\eta\, H^2\, T}{2}.$$

The update in Eq. (3) with $p^t = \pi_h^t(\cdot|s)$ is equivalent to the standard NPG update in Eq. (2) (Xiao, 2022). Using this lemma for each state s and step h, with $g^t = \widehat{Q}_h^t(s,\cdot)$ and an appropriate choice of η gives the following regret bound[1] for the projected NPG:

$$\sum_{t=1}^T \sum_{h=1}^H \max_{s\in\mathcal{S}} \left\langle \pi_h^\star(\cdot|s) - \pi_h^t(\cdot|s), \widehat{Q}_h^t(s,\cdot)\right\rangle \le \mathcal{O}\left(H^2\sqrt{\log|\mathcal{A}|}\,\sqrt{T} + H^2\sqrt{\bar\epsilon}\,T\right), \quad (5)$$

where $\bar\epsilon := \max_{t,s,h} \epsilon_h^t(s)$ is the largest error across all t, s, and h. For the NPG in Eq. (2) without projection, $\bar\epsilon = 0$ and the above result recovers the standard regret bound for NPG. The above lemma suggests that by choosing the projection operator carefully and controlling the projection errors, we can bound the regret.

4.2 Controlling the Projection Error for Log-Linear Policies

To bound the projection error in Theorem 4.1, one could choose that $\mathrm{Proj}_\Pi(p) = \arg\min_p \mathrm{KL}(u \,\|\, p) - \mathrm{KL}(u \,\|\, p^{t+1/2})$, and hence directly control ϵ_h^t. However, this results in a non-convex optimization problem. Consequently, we instead choose Proj to minimize the following regression loss in the logit space: $\frac{1}{2}\|z - (z^t + \eta\, g^t)\|$ where z^t is the logit corresponding to p^t such that $p^t \propto \exp(z^t)$. For the projected NPG with log-linear policies, we aim to minimize the sum of such regression losses (across all $(s,a) \in \mathcal{S}\times\mathcal{A}$) at episode t and step h, and obtain the loss function:

$$\ell_h^t(\theta) = \frac{1}{2}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \left[\left\langle \varphi(s,a), \theta - \widehat{\theta}_h^t\right\rangle - \eta\,\widehat{Q}_h^t(s,a)\right]^2,$$

where $\widehat{Q}_h^t(s,a)$ is the estimated Q-function from the critic. As a regression problem, this actor loss can be easily optimized via gradient descent-based methods.

However, note that the above actor loss requires a minimization over the entire state-action space, which may be impractical. Therefore, we propose to construct a good and preferably small subset $\mathcal{D}_{\text{exp}} \subset \mathcal{S}\times\mathcal{A}$ along with a corresponding distribution $\rho_{\text{exp}} \in \Delta(\mathcal{D}_{\text{exp}})$ that offers good coverage of the feature space. Given $\mathcal{D}_{\text{exp}}$ and ρ_{exp} (the construction of which will be detailed subsequently), we instantiate the actor loss, $\widetilde{\ell}_h^t(\theta)$, in terms of a *logit-matching* regression:

$$\widetilde{\ell}_h^t(\theta) = \frac{1}{2}\sum_{(s,a)\in\mathcal{D}_{\text{exp}}} \rho_{\text{exp}}(s,a)\left[\langle\varphi(s,a),\theta\rangle - \widehat{Z}_h^t(s,a)\right]^2,$$
$$\text{where } \widehat{Z}_h^t(s,a) := \left\langle \varphi(s,a), \widehat{\theta}_h^t\right\rangle + \eta\,\widehat{Q}_h^t(s,a). \quad (6)$$

In order to show that optimizing the above actor loss can indeed bound the projection error, we require the following assumptions. We assume that the given policy features φ are expressive enough to control the bias when minimizing $\ell_h^t(\theta)$.

Assumption 4.1 (Bias)**.** $\min_\theta \sup_{t,h} \widetilde{\ell}_h^t(\theta) \le \epsilon_{\text{bias}}$

In practice, ϵ_{bias} can be controlled by choosing high-dimensional features (e.g., $d_{\boldsymbol{a}} \gg d_{\boldsymbol{c}}$) or a sufficiently expressive policy class (e.g., neural network).

Next, we assume the loss $\widetilde{\ell}_h^t(\theta)$ is sufficiently minimized.

Assumption 4.2 (Optimization Error)**.** *Suppose θ_h^t is obtained by minimizing $\widetilde{\ell}_h^t(\theta)$ in the critic update.* $\sup_{t,h}\left|\widetilde{\ell}_h^t(\theta_h^t) - \min_\theta \widetilde{\ell}_h^t(\theta)\right| \le \epsilon_{\text{opt}}$.

In practice, minimizing $\widetilde{\ell}_h^t(\theta)$ by K_t steps of gradient descent ensures that $\epsilon_{\text{opt}} \le \mathcal{O}(\exp(-K_t))$. Given these two mild assumptions, we can then proceed to bound the projection error. Using Theorem 4.1 for the projected NPG update at state s, step h, and setting $u = \pi^\star(\cdot \mid s)$, the projection error $\epsilon_h^t(s)$ can be bounded as follows.

Lemma 4.1. *Let $\overline{\varphi}_G := \sup_{(s,a)\in\mathcal{S}\times\mathcal{A}}\|\varphi(s,a)\|_{G^{-1}}$ where $G := \sum_{(s,a)\in\mathcal{D}_{\text{exp}}} \rho_{\text{exp}}(s,a)\,\varphi(s,a)\,\varphi(s,a)^\top$. Under Assumptions 4.1 and 4.2, it holds that*

$$\left|\epsilon_h^t(s)\right| \le \bar\epsilon := 2\,(\overline{\varphi}_G + 1)\sqrt{\epsilon_{\text{bias}}} + 2\sqrt{\epsilon_{\text{opt}}} \quad \forall (t,h,s).$$

The above lemma is true for any choice of $\mathcal{D}_{\text{exp}}$ and ρ_{exp}, and suggests that if we can control $\overline{\varphi}_G$, the projection error can be bounded.

Constructing $\mathcal{D}_{\text{exp}}$ and ρ_{exp}. Therefore, we would like to construct a suitable $\mathcal{D}_{\text{exp}}$ and ρ_{exp} to bound $\|\varphi(s,a)\|_{G^{-1}}$ for any $(s,a) \in \mathcal{S}\times\mathcal{A}$ and solve the fol-

[1] This generalized regret bound holds for any other mirror descent-based policy optimization method (e.g., SPMA (Asad et al., 2025) in Appendix B.3), but we discuss NPG within the main text for the ease of exposition.

lowing optimization problem:

$$\inf_{\substack{\mathcal{D}_{\text{exp}}\in\mathcal{S}\times\mathcal{A}\\ \rho_{\text{exp}}\in\Delta(\mathcal{D}_{\text{exp}})}} \sup_{(s,a)\in\mathcal{S}\times\mathcal{A}} \|\varphi(s,a)\|_{G^{-1}} \quad \text{s.t.} \quad G = \sum_{(s,a)\in\mathcal{D}_{\text{exp}}} \rho_{\text{exp}}(s,a)\,\varphi(s,a)\,\varphi(s,a)^\top ,$$

which fits the form of experimental design. Ideally, we would also like $|\mathcal{D}_{\text{exp}}|$ to be relatively small so that the actor parameters can be updated efficiently.

There are standard techniques to solve this problem. The most common approach is the G-optimal design, which involves constructing a *coreset* (i.e., $\mathcal{D}_{\text{exp}}$ and ρ_{exp}) and bounds $\|\varphi(s,a)\|_{G^{-1}}$. In particular, the Kiefer–Wolfowitz theorem (Kiefer and Wolfowitz, 1960) guarantees that there exists a coreset such that $\|\varphi(s,a)\|_{G^{-1}} \le \mathcal{O}(d_{\boldsymbol{a}})$ and $|\mathcal{D}_{\text{exp}}| \le \widetilde{\mathcal{O}}(d_{\boldsymbol{a}})$. Constructing such a coreset can be achieved using various methods, such as the Frank-Wolfe algorithm (Frank et al., 1956; Szepesvári, 2022). We remark that this method only uses the given policy features φ, and does not involve the linear MDP features ϕ. Furthermore, the required coreset can be constructed offline, even before the learning procedure or without any knowledge of the environment (see Appendix C.1 for details). Giving access to such a coreset guarantees that $\bar{\epsilon} \le \mathcal{O}(d_{\boldsymbol{a}}\sqrt{\epsilon_{\text{bias}}} + \sqrt{\epsilon_{\text{opt}}})$, and optimizing the actor loss upon that only requires $\mathcal{O}(d_{\boldsymbol{a}})$ computation.

Rather than forming a coreset, alternative approaches assume $\varphi = \phi$, and use some limited interaction with the environment to construct $\mathcal{D}_{\text{exp}}$. In particular, under some standard assumptions (e.g., Wagenmaker and Jamieson, 2022, Assumption 1), we can apply methods such as `CoverTraj` (Wagenmaker et al., 2022) and `OptCov` (Wagenmaker et al., 2022) that construct $\mathcal{D}_{\text{exp}}$ and ρ_{exp} and can subsequently ensure that $\|\varphi(s,a)\|_{G^{-1}}$ is bounded. The sample complexity for such procedures is $\widetilde{\mathcal{O}}(d_{\boldsymbol{c}}^4 H^3 \epsilon_{\mathcal{M}}^{-1})$ where $\epsilon_{\mathcal{M}}$ is a problem-dependent constant. We defer all the details to Appendix C.

Remark 4.1. *Having access to $\mathcal{D}_{\text{exp}}$ and ρ_{exp} does not obviate the necessity of exploration. Specifically, we do not assume random access, i.e., the agent cannot visit all the state-actions in $\mathcal{D}_{\text{exp}}$. Hence, we cannot effectively calculate $Q^\pi(s,a)$ for any $(s,a) \in \mathcal{D}_{\text{exp}}$.*

4.3 Putting Everything Together: Projected NPG with Log-Linear Policies

In Algorithm 2, we instantiate the complete actor algorithm, which uses the projected `NPG` update for log-linear policies. Unlike the standard `NPG` update, Algorithm 2 alleviates the necessity of storing past Q-functions, improving the memory complexity to $\mathcal{O}(d_{\boldsymbol{a}})$, while enjoying similar theoretical guarantees. Furthermore, the actor parameters are updated by using gradient descent on a properly defined surrogate loss, rendering it closer to the practical implementation of common algorithms (e.g., PPO (Schulman et al., 2017b)).

Algorithm 2 Actor: Projected NPG

Input: critic parameters w^t, policy optimization learning rate η, number of actor updates K_t, actor learning rate $\alpha_{\boldsymbol{a}}^t$, subset and distribution of the state-action space $\mathcal{D}_{\text{exp}}$ and ρ_{exp}
for $h = 1, 2, \dots, H$ **do**
  $\widehat{Q}_h^t(\cdot,\cdot) = \min\{\langle \phi(\cdot,\cdot), w_h^t\rangle, H-h+1\}^+$ [2]
  Define the actor loss $\widetilde{\ell}_h^t(\theta)$ using Eq. (6)
  for $k = 1, \dots, K_t$ **do**
    $\theta_h^{t,k} \leftarrow \theta_h^{t,k-1} - \alpha_{\boldsymbol{a}}^t \nabla_\theta \widetilde{\ell}_h^t(\theta_h^{t,k-1})$
Return: actor parameters for the policy θ^t

We remark that although we focused on the log-linear policies, our theoretical guarantees readily extend to general function approximation when Assumptions 4.1 and 4.2 are satisfied, and one has access to an exploratory policy (Hao et al., 2021, Definition 1). In the next section, we instantiate the critic in Algorithm 1.

5 INSTANTIATING THE CRITIC: LANGEVIN MONTE CARLO

In this section, we use Langevin Monte Carlo (`LMC`) to instantiate the critic. We describe the resulting algorithm in Section 5.1, and analyze it in Section 5.2

The `LMC` approaches allow for sampling from a posterior distribution and have recently been used in sequential decision-making problems. For example, Mazumdar et al. (2020) achieves optimal instance-dependent regret bounds for multi-armed bandits using Langevin dynamics for approximate Thompson sampling. On the other hand, Xu et al. (2022) uses `LMC` for contextual bandits, achieving comparable theoretical results to Thompson sampling. More recently, Ishfaq et al. (2024a) leverages `LMC` for linear MDPs by using it to sample the Q-function from its posterior distribution, achieving the optimal $\widetilde{\mathcal{O}}(\sqrt{T})$ regret.

Nevertheless, all existing `LMC`-based approaches for MDPs, including those for general function approximation (Ishfaq et al., 2024b; Jorge et al., 2024) use value-based algorithms. To the best of our knowledge, such approaches have never been theoretically analyzed in the context of policy optimization. Next, we incorporate the `LMC` algorithm into our actor-critic framework and provide the first provable result.

5.1 LMC for Linear MDPs

At episode t, the critic uses the collected dataset $\mathcal{D}^t$ to obtain an optimistic estimate of the Q function. In order to instantiate the critic loss, we consider the dataset $\mathcal{D}^t$ as split into H disjoint subsets $\{\mathcal{D}_h^t\}_{h\in[H]}$,

[2] $\min\{x,a\}^+ := \max\{\min\{x,a\},0\}$ is the clipping function that bounds the given value x to the range $[0,a]$.

where $\mathcal{D}_h^t$ consists of (s_h, a_h, s_{h+1}) tuples indexed as $\{(s_h^i, a_h^i, s_{h+1}^i)\}_{i=1}^{|\mathcal{D}^t|}$ [3]. The critic loss at episode t and step h uses the estimated value function at step $h+1$, and forms the following ridge regression problem:

$$\mathcal{L}_h^t(w) = \tfrac{1}{2}\sum_{i=1}^{|\mathcal{D}^t|}\left[\chi_i^{t,h} - \langle \phi(s_h^i, a_h^i), w\rangle\right]^2 + \tfrac{\lambda}{2}\|w\|^2,$$
$$\text{where } \chi_i^{t,h} = r_h(s_h^i, a_h^i) + \widehat{V}_{h+1}^t(s_{h+1}^i)\,. \tag{7}$$

For each step h, the LMC algorithm iteratively adds Gaussian noise to the gradient descent updates on $\mathcal{L}_h^t(w)$, and aims to produce approximate samples of the critic parameters from its underlying posterior distribution (Line 6-8). In particular, for an arbitrary loss ℓ, the LMC update can be written as:

$$w^{t+1} = w^t - \alpha^t \nabla_w \ell(w^t) + \sqrt{\alpha^t/\zeta}\,\nu^t\,,$$

where α_t is the learning rate, ζ is the inverse temperature parameter, and ν_t is sampled from an isotropic Gaussian distribution. After J_t steps of the LMC update on the critic loss (Lines 6-8 in Algorithm 3), the resulting critic parameters are used to produce an optimistic sample of the Q-function (Line 9). From a theoretical perspective, we note that it is important to clip Q_h^t appropriately. In order to improve the optimism guarantees of the LMC algorithm, we follow the idea in Ishfaq et al. (2021), and repeat the LMC update M times, taking the maximum over these samples (Line 9). Iterating this procedure backwards from $h = H$ to 1, we can obtain the desired critic parameters.

Note that compared to UCB-based approaches, LMC does not require computing confidence sets at every episode. Instead, it simply perturbs gradient descent by injecting Gaussian noise, allowing for a natural extension beyond the linear function approximation setting and rendering it easier to implement in practice.

5.2 Optimism Guarantee and Error Bound

In order to theoretically analyze Algorithm 3, we first define the following model prediction error.

Definition 5.1. *Given an estimated Q-function $\widehat{Q}^t$ and the corresponding estimated value function $\widehat{V}^t$, for all (t, h, s, a), the model prediction error is defined as: $\iota_h^t(s,a) \coloneqq r_h(s,a) + \mathbb{P}_h\,\widehat{V}_{h+1}^t(s,a) - \widehat{Q}_h^t(s,a)$.*

The theoretical analyses in existing work (Jin et al., 2020; Zhong and Zhang, 2023; Liu et al., 2023) that use UCB bonuses typically proceed by proving an upper bound of 0 on ι_h^t (*optimism*) and a lower bound of $\widetilde{\mathcal{O}}(\sqrt{T})$. The following lemma shows that LMC can offer similar guarantees.

Lemma 5.1. *Let $\Lambda_h^t \coloneqq \sum_{(s,a,s')\in\mathcal{D}_h^t}\phi(s,a)\,\phi(s,a)^\top + \lambda I$. With appropriate choices of λ, ζ, J_t, $\alpha_{\boldsymbol{c}}^t$, M and for any $\delta \in (0,1)$, Algorithm 1 with the LMC critic in Algorithm 3 ensures that in both the on-policy and off-policy settings, for all t, h, s, a and some constant $\Gamma_{\texttt{LMC}} = \widetilde{\mathcal{O}}(H\,d_{\boldsymbol{c}})$, with probability at least $1-\delta$,*

$$-\Gamma_{\texttt{LMC}} \times \|\phi(s,a)\|_{(\Lambda_h^t)^{-1}} \le \iota_h^t(s,a) \le 0.$$

[3] $|\mathcal{D}^t|$ represents the number of trajectories in $\mathcal{D}^t$ or the number of (s_h, a_h, s_{h+1}) tuples in $\mathcal{D}_h^t$

Algorithm 3 Critic: LMC

1: **Input**: collected data $\mathcal{D}^t$, policy π^{t-1}, number of critic updates J_t, critic learning rate $\alpha_{\boldsymbol{c}}^{h,t}$, inverse temperature ζ, number of critic samples M
2: $\widehat{V}_{H+1}^t(\cdot) \leftarrow 0$
3: **for** $h = H, H-1, \ldots, 1$ **do**
4: Define the critic loss $\mathcal{L}_h^t(w)$ using Eq. (7)
5: $w_h^{t,m,0} \leftarrow w_h^{t-1,m,J_{t-1}} \quad \forall m \in [M]$
6: **for** $j = 1, \ldots, J_t$ **do**
7: $\nu_h^{t,m,j} \leftarrow \mathbf{N}(0, I) \quad \forall m \in [M]$
8: $w_h^{t,m,j} \leftarrow w_h^{t,m,j-1} - \alpha_{\boldsymbol{c}}^{h,t}\nabla_w \mathcal{L}_h^t(w_h^{t,m,j-1}) + \sqrt{\alpha_{\boldsymbol{c}}^{h,t}/\zeta}\,\nu_h^{t,m,j} \quad \forall m \in [M]$
9: $\widehat{Q}_h^t(\cdot,\cdot) = \min\left\{\max_{m\in[M]}\left\langle \phi(\cdot,\cdot), w_h^{t,m,J_t}\right\rangle, H\right\}^+$
10: $\widehat{V}_h^t(\cdot) = \mathbb{E}_{a\sim\pi^{t-1}(\cdot|s)}\widehat{Q}_h^t(\cdot,a)$
11: **Return**: critic parameters for the estimated Q-function $\{w_h^{t,m,J_t}\}_{(m,h)\in[M]\times[H]}$

The exact definition of $\Gamma_{\texttt{LMC}}$ varies between the on-policy and off-policy settings, although they are both bounded by $\widetilde{O}(H\,d_{\boldsymbol{c}})$ (see Appendix D for the full version of this lemma). In order to prove this result in the on-policy setting, we use the fact that all the data points in $\mathcal{D}_h^t$ are collected via independent trajectories from the same policy π^t, and are therefore independent and identically distributed. Hence, we can use the self-normalized bounds in Abbasi-Yadkori et al. (2011) to analyze the dependence in h, and prove the corresponding result. In the off-policy setting, since the data points in $\mathcal{D}_h^t$ are collected by different data-dependent policies, these samples are correlated in a complicated manner. Hence, we use the value-aware uniform concentration result from Jin et al. (2020). We remark that this result requires control over the covering number of the value function class, which is deferred to Section 6.

Therefore, we conclude that, compared to UCB bonuses, LMC offers significant practical advantages while still providing similar theoretical guarantees.

6 SAMPLE COMPLEXITY ANALYSIS

In this section, we analyze the sample complexity of Algorithm 1 with the projected NPG actor from Algorithm 2 and the LMC critic from Algorithm 3. Section 6.1 focuses on the on-policy setting, while Section 6.2 addresses the off-policy setting.

6.1 On-Policy Setting

We now present the following theorem that shows that our proposed algorithm achieves a sample complexity of $\widetilde{\mathcal{O}}(\epsilon^{-4})$ in the on-policy setting.

Theorem 6.1. *Under Assumptions 4.1 and 4.2, consider Algorithm 1 in the on-policy setting with the projected* NPG *actor (Algorithm 2) and the* LMC *critic (Algorithm 3). Suppose $\overline{\epsilon}$ is the projection error in the actor. For an appropriate choice of the actor and critic parameters, including $N = H^4/\epsilon^2$, and any $\delta \in (0,1)$, it holds that with probability at least $1-\delta$,*

$$\texttt{OG}(T) \leq \widetilde{\mathcal{O}}\left(\frac{H^2\sqrt{d_{\boldsymbol{c}}^3\,\log|\mathcal{A}|}}{\sqrt{T}} + H^2\sqrt{\overline{\epsilon}}\right).$$

Hence, for any $\epsilon > 0$, Algorithm 1 with $T = d_{\boldsymbol{c}}^3\,H^4\,\log|\mathcal{A}|\,\epsilon^{-2}$ yields a $(\epsilon + H^2\sqrt{\overline{\epsilon}})$-optimal mixture policy, and therefore requires $T \times N = \widetilde{\mathcal{O}}(\epsilon^{-4})$ samples.

Proof Sketch. We decompose the difference between $V_1^{\overline{\pi}^T}(s_1)$ and $V_1^{\star}(s_1)$ into two terms that only depend on either the actor or the critic.

$$\mathbb{E}\Big[V_1^{\pi^\star} - V_1^{\overline{\pi}^T}(s_1)\Big] = \underbrace{\frac{1}{T}\sum_{t=1}^{T}\sum_{h=1}^{H}\mathbb{E}_{\pi^\star}\Big[\Big\langle \pi_h^\star(\cdot\mid s_h) - \pi_h^t(\cdot\mid s_h), \widehat{Q}_h^t(s_h,\cdot)\Big\rangle\Big]}_{\text{policy optimization (actor) error}} + \underbrace{\frac{1}{T}\sum_{t=1}^{T}\sum_{h=1}^{H}\big(\mathbb{E}_{\pi^\star}[\iota_h^t(s_h,a_h)] - \mathbb{E}_{\pi^t}[\iota_h^t(s_h,a_h)]\big)}_{\text{policy evaluation (critic) error}}.$$

The policy optimization (actor) error can be bounded using Eq. (5), and the policy evaluation (critic) error is bounded using Lemma D.2. In particular, in the on-policy setting, the lower-bound in Lemma D.2 can be instantiated as:

$$\sum_{t=1}^{T}\sum_{h=1}^{H}\mathbb{E}_{\pi^t}\big[-\iota_h^t(s,a)\big] \leq \mathcal{O}\left(\sqrt{d_{\boldsymbol{c}}^3\,H^4\,T\,\log^2(N/\delta)/N}\right).$$

Putting everything together with the chosen value of N leads to the stated sample complexity. □

Comparison to Liu et al. (2023). The on-policy sample-complexity in Theorem 6.1 matches the bound in Liu et al. (2023, Theorem 1). However, unlike the proposed algorithm, Liu et al. (2023) uses NPG with implicit policies for the actor. Consequently, sampling from the current policy (to interact with the environment) requires calculating $\sum_{\tau=1}^{t-1}\widehat{Q}_h^\tau(\cdot,\cdot)$ for each encountered state-action pair. Since they use clipped Q functions with UCB bonuses for the critic, the above sum of Q functions cannot be stored succinctly. Hence, sampling from the policy requires instantiating each previous Q function for each step h and episode t, resulting in a computational cost of $\mathcal{O}(d_{\boldsymbol{c}}^2\,H\,\epsilon^{-2})$.

6.2 Off-Policy Setting

Next, we show that, in the off-policy setting, Algorithm 1 can achieve $\widetilde{\mathcal{O}}(\epsilon^{-2})$ sample complexity.

Theorem 6.2. *Under Assumptions 4.1 and 4.2, consider Algorithm 1 in the off-policy setting with the projected* NPG *actor (Algorithm 2) and the* LMC *critic (Algorithm 3). Suppose $\overline{\epsilon}$ is the projection error in the actor. For an appropriate choice of the actor and critic parameters and any $\delta \in (0,1)$, it holds that with probability at least $1-\delta$,*

$$\texttt{OG}(T) \leq \widetilde{\mathcal{O}}\left(\frac{H^2\sqrt{d_{\boldsymbol{c}}^3\,\max\{d_{\boldsymbol{a}},d_{\boldsymbol{c}}\}\,\log|\mathcal{A}|}}{\sqrt{T}} + H^2\sqrt{\overline{\epsilon}}\right).$$

Hence, for any $\epsilon > 0$, Algorithm 1 with $T = d_{\boldsymbol{c}}^3\,\max\{d_{\boldsymbol{a}},d_{\boldsymbol{c}}\}\,H^4\,\log|\mathcal{A}|\,\epsilon^{-2}$ yields a $(\epsilon + H^2\sqrt{\overline{\epsilon}})$-optimal mixture policy, and therefore requires $T \times 1 = \widetilde{\mathcal{O}}(\epsilon^{-2})$ samples.

Proof Sketch. The proof uses a similar regret decomposition to Theorem 6.1. Compared to the on-policy setting, the most significant difference is the bound on the policy evaluation (critic) errors. First, using Lemma 5.1, we have that

$$\sum_{t=1}^{T}\sum_{h=1}^{H}\mathbb{E}_{\pi^t}\big[-\iota_h^t(s,a)\big] \leq \Gamma\sum_{t=1}^{T}\sum_{h=1}^{H}\mathbb{E}_{\pi^t}\|\phi(s,a)\|_{(\Lambda_h^t)^{-1}}.$$

The term, $\sum_{t=1}^{T}\sum_{h=1}^{H}\mathbb{E}_{\pi^t}\|\phi(s,a)\|_{(\Lambda_h^t)^{-1}}$, can be bounded using the standard elliptical potential lemma. However, since we are in the off-policy setting, bounding $\Gamma > 0$ requires the uniform concentration argument from Jin et al. (2020), which yields that

$$\Gamma \leq \mathcal{O}\left(H\sqrt{d_{\boldsymbol{c}}\log T + \log\left(\frac{\mathcal{N}_\Delta(\mathcal{V})}{\delta}\right)} + T\,\Delta\right).$$

This involves obtaining a bound on $\log(\mathcal{N}_\Delta(\mathcal{V}))$, the logarithm of the covering number of the value function class. The covering number is a measure of the complexity of the space of value functions. In particular, we show that for an actor with log-linear policies, we can bound the logarithm of the covering number using the following lemma.

Lemma 6.1. *Let Π_{lin} be the policy class induced by Eq. (1) such that $\sup_{\theta,h,s,a}\|z_h(s,a\mid\theta)\| \leq \overline{Z}$. Let $\mathcal{Q} = \left\{\min\{\langle\phi(\cdot,\cdot),w\rangle), H\}^+ \mid \|w\| \leq \overline{W}\right\}$ be the Q-function class and $\mathcal{V} = \{\langle Q(\cdot,\cdot),\pi(\cdot\mid\cdot,\theta)\rangle_{\mathcal{A}} \mid Q \in \mathcal{Q},\ \pi \in \Pi_{lin}\}$ be the corresponding value function class. Then, it holds that $\log\mathcal{N}_\Delta(\mathcal{V}) \leq \mathsf{V}$ where $\mathsf{V} := d_{\boldsymbol{c}}\log\left(1 + \frac{4\overline{W}+4H\sqrt{2\overline{Z}}}{\Delta}\right) + d_{\boldsymbol{a}}\log\left(1 + \frac{4H\sqrt{2\overline{Z}}}{\Delta}\right)$.*

In particular, we can show that $\overline{W} \leq \mathcal{O}(\sqrt{T})$ (Lemma D.7), and $\overline{Z} \leq \mathcal{O}(\bar{\epsilon}\, T)$ (Lemma F.1). Putting everything together and setting $\Delta = 1/T$ yields that

$$\sum_{t=1}^{T}\sum_{h=1}^{H} \mathbb{E}_{\pi^t}\left[-\iota_h^t(s,a)\right] \leq \widetilde{\mathcal{O}}\left(\sqrt{d_{\boldsymbol{c}}^3\, \max\{d_{\boldsymbol{a}}, d_{\boldsymbol{c}}\}\, H^4\, T}\right).$$

Following a proof similar to Theorem 6.1 leads to the desired sample complexity. □

Remark 6.1. *In order to effectively bound the logarithm of the covering number, previous work (Sherman et al., 2023; Cassel and Rosenberg, 2024; Tan et al., 2025) has incorporated various algorithmic tweaks, including reward-free warm-ups, feature contractions, and rare-switching. On the contrary, since our algorithm projects the implicit policy onto the log-linear policy class at each iteration, the logarithm of the covering number can be bounded in a more direct manner.*

Comparison to Sherman et al. (2023); Cassel and Rosenberg (2024). The off-policy sample-complexity in Theorem 6.2 matches the bound in Sherman et al. (2023); Cassel and Rosenberg (2024). Similar to Liu et al. (2023), both works use NPG with implicit policies for the actor and `UCB` bonuses for the critic. Consequently, the resulting methods suffer from a high cost of policy inference. Finally, we note that similar to the proposed algorithm (see Section 4.2), Sherman et al. (2023) also uses a reward-free warm-up phase (Wagenmaker et al., 2022). However, while we require the warm-up phase to identify a coreset to efficiently minimize the actor loss (a computational reason), Sherman et al. (2023) requires the warm-up procedure to restrict the subsequent regret minimization procedure to high occupancy regions of the state-action space and effectively control the capacity of the policy class (a statistical reason).

7 EXPERIMENTS

In this section, we first evaluate our proposed algorithm in linear MDPs, consistent with our theoretical analyses. To further demonstrate its versatility, we extend the proposed algorithm to large-scale deep RL applications, evaluating its performance across several Atari games.

7.1 Experiments in Linear MDPs

To demonstrate the practical value of our proposed algorithm, we evaluate it over various benchmarks. We first test our proposed algorithm in the randomly generated linear MDPs (Random MDP). We compare our proposed algorithm, `LMC-NPG-EXP`, with the memory-intensive variant with implicit policy parameterization, `LMC-NPG-IMP`, and the value-based baseline, `LMC` (Ishfaq et al., 2024a). As shown in Figure 1, our proposed algorithm can achieve comparable performance with the memory-intensive variant and better performance than `LMC`. In Appendix G.1, we also do the same experiment in the linear MDP version of the Deep Sea environment (Osband et al., 2019). In Appendix G.2, we further conduct some ablation studies in these two environments of linear MDPs.

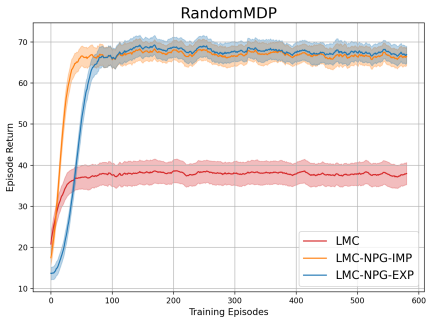

Figure 1: Comparison of `LMC-NPG-EXP` (our proposed algorithm), `LMC-NPG-IMP` (memory-intensive variant), and `LMC` (value-based baseline) in the Random MDP.

7.2 Experiments Beyond Linear MDPs: Atari

In Appendix G.3, we extend our proposed algorithm to large-scale deep RL applications. We then conduct experiments in Atari (Mnih et al., 2013) using Stable Baselines3 (Raffin et al., 2021) and compare our proposed algorithm to PPO (Schulman et al., 2017b) in the on-policy setting and DQN (Mnih et al., 2015) in the off-policy setting. We demonstrate that our algorithm can achieve similar or even better performance.

8 DISCUSSION

We proposed an optimistic actor–critic algorithm with explicitly parameterized policies and a systematic exploration mechanism. In particular, for the actor, we demonstrated that using projected `NPG` with parametric policies is not only practical, but also equipped with theoretical guarantees. For the critic, we demonstrated that `LMC` is a principled and easy-to-implement exploration scheme for policy optimization methods. We derived theoretical guarantees in both the on-policy and off-policy settings, showcasing that the proposed actor-critic algorithm can simultaneously achieve sample efficiency and practicality.

For future work, eliminating the dependence on $\mathcal{D}_{\text{exp}}$ and ρ_{exp} will result in a more practical algorithm. We also aim to investigate the actor-critic algorithms in more realistic setups (e.g., infinite-horizon discounted MDPs) with more general function approximation schemes beyond linear models for both the environment and the policy. It would also be fruitful to further evaluate their empirical performance in more challenging large-scale deep RL applications.

ACKNOWLEDGMENTS

We would like to thank Xingtu Liu, Yunxiang Li, and the anonymous reviewers for their helpful feedback. This work was partially supported by the Natural Sciences and Engineering Research Council of Canada (NSERC) Discovery Grant RGPIN-2022-04816, and enabled in part by support provided by the Digital Research Alliance of Canada (alliancecan.ca).

References

Abbasi-Yadkori, Y., Bartlett, P., Bhatia, K., Lazic, N., Szepesvari, C., and Weisz, G. (2019). Politex: Regret bounds for policy iteration using expert prediction. In *International Conference on Machine Learning*, pages 3692–3702. PMLR.

Abbasi-Yadkori, Y., Pál, D., and Szepesvári, C. (2011). Improved algorithms for linear stochastic bandits. *Advances in Neural Information Processing Systems*, 24.

Abramowitz, M. and Stegun, I. A. (1948). *Handbook of mathematical functions with formulas, graphs, and mathematical tables*, volume 55. US Government Printing Office.

Agarwal, A., Jin, Y., and Zhang, T. (2023). VOQL: Towards optimal regret in model-free RL with nonlinear function approximation. In *The Thirty Sixth Annual Conference on Learning Theory*, pages 987–1063. PMLR.

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

Agarwal, N., Chaudhuri, S., Jain, P., Nagaraj, D., and Netrapalli, P. (2021b). Online target q-learning with reverse experience replay: Efficiently finding the optimal policy for linear mdps. *arXiv preprint arXiv:2110.08440*.

Alfano, C. and Rebeschini, P. (2022). Linear convergence for natural policy gradient with log-linear policy parametrization. *arXiv preprint arXiv:2209.15382*.

Asad, R., Harikandeh, R. B., Laradji, I. H., Le Roux, N., and Vaswani, S. (2025). Fast convergence of softmax policy mirror ascent. In *International Conference on Artificial Intelligence and Statistics*, pages 3943–3951. PMLR.

Bhandari, J. and Russo, D. (2024). Global optimality guarantees for policy gradient methods. *Operations Research*, 72(5):1906–1927.

Bhatnagar, S., Sutton, R. S., Ghavamzadeh, M., and Lee, M. (2009). Natural actor–critic algorithms. *Automatica*, 45(11):2471–2482.

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

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

Cayci, S., He, N., and Srikant, R. (2024). Finite-time analysis of entropy-regularized neural natural actor-critic algorithm. *Transactions on Machine Learning Research*.

Cen, S., Cheng, C., Chen, Y., Wei, Y., and Chi, Y. (2022). Fast global convergence of natural policy gradient methods with entropy regularization. *Operations Research*, 70(4):2563–2578.

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

Frank, M., Wolfe, P., et al. (1956). An algorithm for quadratic programming. *Naval research logistics quarterly*, 3(1-2):95–110.

Fu, Z., Yang, Z., and Wang, Z. (2021). Single-timescale actor-critic provably finds globally optimal policy. In *International Conference on Learning Representations*.

Fujimoto, S., Hoof, H., and Meger, D. (2018). Addressing function approximation error in actor-critic methods. In *International conference on machine learning*, pages 1587–1596. PMLR.

Gaur, M., Bedi, A., Wang, D., and Aggarwal, V. (2024). Closing the gap: Achieving global convergence (last iterate) of actor-critic under Markovian sampling with neural network parametrization. In *International Conference on Machine Learning*, pages 15153–15179. PMLR.

Gaur, M., Bedi, A. S., Wang, D., and Aggarwal, V. (2023). On the global convergence of natural actor-critic with two-layer neural network parametrization. *arXiv preprint arXiv:2306.10486*.

Haarnoja, T., Zhou, A., Abbeel, P., and Levine, S. (2018). Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *International conference on machine learning*, pages 1861–1870. Pmlr.

Hao, B., Lattimore, T., Szepesvári, C., and Wang, M. (2021). Online sparse reinforcement learning. In *International Conference on Artificial Intelligence and Statistics*, pages 316–324. PMLR.

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

He, J., Zhao, H., Zhou, D., and Gu, Q. (2023). Nearly minimax optimal reinforcement learning for linear markov decision processes. In *International Conference on Machine Learning*, pages 12790–12822. PMLR.

Hong, M., Wai, H.-T., Wang, Z., and Yang, Z. (2023). A two-timescale stochastic algorithm framework for bilevel optimization: Complexity analysis and application to actor-critic. *SIAM Journal on Optimization*, 33(1):147–180.

Ishfaq, H., Cui, Q., Nguyen, V., Ayoub, A., Yang, Z., Wang, Z., Precup, D., and Yang, L. (2021). Randomized exploration in reinforcement learning with general value function approximation. In *International Conference on Machine Learning*, pages 4607–4616. PMLR.

Ishfaq, H., Lan, Q., Xu, P., Mahmood, A. R., Precup, D., Anandkumar, A., and Azizzadenesheli, K. (2024a). Provable and practical: Efficient exploration in reinforcement learning via langevin monte carlo. In *The Twelfth International Conference on Learning Representations*.

Ishfaq, H., Tan, Y., Yang, Y., Lan, Q., Lu, J., Mahmood, A. R., Precup, D., and Xu, P. (2024b). More efficient randomized exploration for reinforcement learning via approximate sampling. In *Reinforcement Learning Conference*.

Ishfaq, H., Wang, G., Islam, S. N., and Precup, D. (2025). Langevin soft actor-critic: Efficient exploration through uncertainty-driven critic learning. In *The Thirteenth International Conference on Learning Representations*.

Jin, C., Liu, Q., and Miryoosefi, S. (2021). Bellman eluder dimension: New rich classes of rl problems, and sample-efficient algorithms. *Advances in Neural Information Processing Systems*, 34:13406–13418.

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

Jorge, E., Dimitrakakis, C., and Basu, D. (2024). Isoperimetry is all we need: Langevin posterior sampling for rl with sublinear regret. *arXiv preprint arXiv:2412.20824*.

Kakade, S. M. (2001). A natural policy gradient. *Advances in Neural Information Processing Systems*, 14.

Khodadadian, S., Doan, T. T., Romberg, J., and Maguluri, S. T. (2022). Finite-sample analysis of two-time-scale natural actor–critic algorithm. *IEEE Transactions on Automatic Control*, 68(6):3273–3284.

Kiefer, J. and Wolfowitz, J. (1960). The equivalence of two extremum problems. *Canadian Journal of Mathematics*, 12:363–366.

Kober, J., Bagnell, J. A., and Peters, J. (2013). Reinforcement learning in robotics: A survey. *The International Journal of Robotics Research*, 32(11):1238–1274.

Konda, V. and Tsitsiklis, J. (1999). Actor-critic algorithms. *Advances in Neural Information Processing Systems*, 12.

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

Lillicrap, T. P., Hunt, J. J., Pritzel, A., Heess, N., Erez, T., Tassa, Y., Silver, D., and Wierstra, D. (2015). Continuous control with deep reinforcement learning. *arXiv preprint arXiv:1509.02971*.

Liu, Q., Weisz, G., György, A., Jin, C., and Szepesvári, C. (2023). Optimistic natural policy gradient: a simple efficient policy optimization framework for online rl. *Advances in Neural Information Processing Systems*, 36:3560–3577.

Mazumdar, E., Pacchiano, A., Ma, Y., Jordan, M., and Bartlett, P. (2020). On approximate thompson sampling with langevin algorithms. In *international conference on machine learning*, pages 6797–6807. PMLR.

Mnih, V., Kavukcuoglu, K., Silver, D., Graves, A., Antonoglou, I., Wierstra, D., and Riedmiller, M. (2013). Playing atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602*.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., et al. (2015). Human-level control through deep reinforcement learning. *nature*, 518(7540):529–533.

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

Osband, I., Doron, Y., Hessel, M., Aslanides, J., Sezener, E., Saraiva, A., McKinney, K., Lattimore, T., Szepesvari, C., Singh, S., et al. (2019). Behaviour suite for reinforcement learning. *arXiv preprint arXiv:1908.03568*.

Peters, J., Vijayakumar, S., and Schaal, S. (2005). Natural actor-critic. In *European Conference on Machine Learning*, pages 280–291. Springer.

Raffin, A. (2020). Rl baselines3 zoo. https://github.com/DLR-RM/rl-baselines3-zoo.

Raffin, A., Hill, A., Gleave, A., Kanervisto, A., Ernestus, M., and Dormann, N. (2021). Stable-baselines3: Reliable reinforcement learning implementations.

Journal of Machine Learning Research, 22(268):1–8.

Schulman, J., Chen, X., and Abbeel, P. (2017a). Equivalence between policy gradients and soft q-learning. *arXiv preprint arXiv:1704.06440*.

Schulman, J., Levine, S., Abbeel, P., Jordan, M., and Moritz, P. (2015). Trust region policy optimization. In *International conference on machine learning*, pages 1889–1897. PMLR.

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

Sherman, U., Cohen, A., Koren, T., and Mansour, Y. (2023). Rate-optimal policy optimization for linear markov decision processes. *arXiv preprint arXiv:2308.14642*.

Sutton, R. S., McAllester, D., Singh, S., and Mansour, Y. (1999). Policy gradient methods for reinforcement learning with function approximation. *Advances in Neural Information Processing Systems*, 12.

Szepesvári, C. (2022). *Algorithms for reinforcement learning*. Springer nature.

Tan, K., Fan, W., and Wei, Y. (2025). Actor-critics can achieve optimal sample efficiency. *arXiv preprint arXiv:2505.03710*.

Todd, M. J. (2016). *Minimum-volume ellipsoids: Theory and algorithms*. SIAM.

Tomar, M., Shani, L., Efroni, Y., and Ghavamzadeh, M. (2020). Mirror descent policy optimization. *arXiv preprint arXiv:2005.09814*.

Uc-Cetina, V., Navarro-Guerrero, N., Martin-Gonzalez, A., Weber, C., and Wermter, S. (2023). Survey on reinforcement learning for language processing. *Artificial Intelligence Review*, 56(2):1543–1575.

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

Wagenmaker, A. J., Chen, Y., Simchowitz, M., Du, S., and Jamieson, K. (2022). 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.

Welling, M. and Teh, Y. W. (2011). Bayesian learning via stochastic gradient langevin dynamics. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pages 681–688.

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

Xiao, L. (2022). On the convergence rates of policy gradient methods. *Journal of Machine Learning Research*, 23(282):1–36.

Xu, P., Zheng, H., Mazumdar, E. V., Azizzadenesheli, K., and Anandkumar, A. (2022). Langevin monte carlo for contextual bandits. In *International Conference on Machine Learning*, pages 24830–24850. PMLR.

Xu, T., Wang, Z., and Liang, Y. (2020). Improving sample complexity bounds for (natural) actor-critic algorithms. *Advances in Neural Information Processing Systems*, 33:4358–4369.

Yuan, R., Du, S. S., Gower, R. M., Lazaric, A., and Xiao, L. (2023). Linear convergence of natural policy gradient methods with log-linear policies. In *International Conference on Learning Representations*.

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

Zhong, H. and Zhang, T. (2023). A theoretical analysis of optimistic proximal policy optimization in linear markov decision processes. *Advances in Neural Information Processing Systems*, 36:73666–73690.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Yes]
2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes]
   (b) Complete proofs of all theoretical results. [Yes]
   (c) Clear explanations of any assumptions. [Yes]
3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Yes]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Not Applicable]
   (b) The license information of the assets, if applicable. [Not Applicable]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]
   (d) Information about consent from data providers/curators. [Not Applicable]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]
5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Not Applicable]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

Appendix

Contents

A NOTATIONS

Notation	Meaning
Problem Definition	
$\mathcal{S}$, $\mathcal{A}$	state space and action space
H, h	horizon length (total number of steps), current index of step
$r \in \mathbb{R}^{\lvert\mathcal{S}\rvert\times\lvert\mathcal{A}\rvert}$	reward function
$\mathbb{P} \in \mathbb{R}^{\lvert\mathcal{S}\rvert\times\lvert\mathcal{A}\rvert\times\lvert\mathcal{S}\rvert}$	transition kernel
$\phi : \mathcal{S}\times\mathcal{A} \mapsto \mathbb{R}^{d_c}$	features for the linear MDP environment
$\varphi : \mathcal{S}\times\mathcal{A} \mapsto \mathbb{R}^{d_a}$	features for the learnable policy
Algorithm Design	
T, t	total number of learning episodes, index of current episode
$\mathcal{D}^t$, $\mathcal{D}^t_h$	collected data of at episode t, split data at h-th step (subset of $\mathcal{D}^t$)
N	number of samples collected for on-policy learning
$w \in \mathbb{R}^{d_c}$	learnable critic parameters
J	number of critic updates
α_c	critic learning rate
ν	noise vector for LMC sampled from the standard normal distribution
ζ	inverse temperature for the LMC critic loss
M	number of samples for the critic parameters
$\mathcal{D}_{\text{exp}}$, ρ_{exp}	subset of $\mathcal{S}\times\mathcal{A}$, distribution over the subset
$\theta \in \mathbb{R}^{d_a}$	learnable actor parameters
K	number of actor updates
α_a	actor learning rate
η	policy optimization learning rate

Table 2: Notations for Problem Definition and Algorithm Design

Additional Notations. Throughout this paper, we use subscripts to represent the index of the step within the horizon of the episodic MDP and superscripts to denote the index of the episode for learning. For example, V^t_h means the value function for the h-th step derived at the learning episode t. In some cases, where the subscripts are omitted, it represents a set of H functions for all steps $h \in [H]$ (e.g., $V^t := \{V^t_h\}_{h\in[H]}$). $|\mathcal{D}^t|$ represents the number of trajectories in $\mathcal{D}^t$ or the number of (s_h, a_h, s_{h+1}) tuples in $\mathcal{D}^t_h$. Additionally, for any vector $v \in \mathbb{R}^d$ and any matrix $M \in \mathbb{R}^{d\times d}$, we denote $\|v\|_M = \sqrt{v^\top M\, v}$.

B ANALYSIS FOR THE ACTOR

B.1 Generalized OMD Regret (Proof of Theorem 4.1)

Proof of Theorem 4.1. Given the update of $p^{t+1/2}$ and the fact that $\Delta(\mathcal{A})$ is a convex set, we have the following optimality condition:

$$\left\langle u - p^{t+1/2}, -\eta\, g^t + \log(p^{t+1/2}) - \log(p^t) \right\rangle \geq 0\,. \tag{8}$$

Then, for each $t \in [T]$, we have that

$$\begin{aligned}
\left\langle u - p^t, \eta\, g^t \right\rangle &= \left\langle u - p^{t+1/2}, \eta\, g^t \right\rangle + \left\langle p^{t+1/2} - p^t, \eta\, g^t \right\rangle \\
&= \left\langle u - p^{t+1/2}, \eta\, g^t - \log(p^{t+1/2}) + \log(p^t) \right\rangle + \left\langle u - p^{t+1/2}, \log(p^{t+1/2}) - \log(p^t) \right\rangle \\
&\quad + \left\langle p^{t+1/2} - p^t, \eta\, g^t \right\rangle \\
&\overset{\text{(i)}}{\leq} \left\langle u - p^{t+1/2}, \log(p^{t+1/2}) - \log(p^t) \right\rangle + \left\langle p^{t+1/2} - p^t, \eta\, g^t \right\rangle \\
&\overset{\text{(ii)}}{=} \mathrm{KL}(u \,\|\, p^t) - \mathrm{KL}(u \,\|\, p^{t+1/2}) - \mathrm{KL}(p^{t+1/2} \,\|\, p^t) + \left\langle p^{t+1/2} - p^t, \eta\, g^t \right\rangle
\end{aligned}$$

$$
\begin{aligned}
&\overset{\text{(iii)}}{\leq} \mathrm{KL}(u \,\|\, p^t) - \mathrm{KL}(u \,\|\, p^{t+1/2}) - \mathrm{KL}(p^{t+1/2} \,\|\, p^t) + \frac{1}{2}\left\|p^{t+1/2} - p^t\right\|_1^2 + \frac{1}{2}\left\|\eta\, g^t\right\|_\infty^2 \\
&\overset{\text{(iv)}}{\leq} \mathrm{KL}(u \,\|\, p^t) - \mathrm{KL}(u \,\|\, p^{t+1/2}) - \mathrm{KL}(p^{t+1/2} \,\|\, p^t) + \mathrm{KL}(p^{t+1/2} \,\|\, p^t) + \frac{\eta^2\, H^2}{2} \\
&\leq \mathrm{KL}(u \,\|\, p^t) - \mathrm{KL}(u \,\|\, p^{t+1/2}) + \frac{\eta^2\, H^2}{2} \\
&\leq \mathrm{KL}(u \,\|\, p^t) - \mathrm{KL}(u \,\|\, p^{t+1}) + \mathrm{KL}(u \,\|\, p^{t+1}) - \mathrm{KL}(u \,\|\, p^{t+1/2}) + \frac{\eta^2\, H^2}{2} \\
&= \mathrm{KL}(u \,\|\, p^t) - \mathrm{KL}(u \,\|\, p^{t+1}) + \epsilon^t + \frac{\eta^2\, H^2}{2}\,.
\end{aligned}
$$

(i) drops the first term due to the optimality condition from Eq. (8). (ii) applies the three-point property of Bregman divergence (Lemma B.2) by setting $x = u$, $y = p^{t+1/2}$, and $z = p^t$. (iii) follows from the Hölder's inequality and then the Young's inequality (i.e., $\langle u, v\rangle \leq \|u\|_1 \|v\|_\infty \leq \|u\|_1^2/2 + \|v\|_\infty^2/2$), and (iv) applies the Pinkster's inequality and $|g^t| \leq H$. Summing up the above inequality from $t = 1$ to T yields that

$$
\begin{aligned}
\sum_{t=1}^{T} \left\langle u - p^t, \eta\, g^t \right\rangle &= \sum_{t=1}^{T} \mathrm{KL}(u \,\|\, p^t) - \mathrm{KL}(u \,\|\, p^{t+1}) + \sum_{t=1}^{T} \epsilon^t + \frac{\eta^2\, H^2\, T}{2} \\
&= \mathrm{KL}(u \,\|\, p^1) - \mathrm{KL}(u \,\|\, p^{T+1}) + \sum_{t=1}^{T} \epsilon^t + \frac{\eta^2\, H^2\, T}{2} \\
&\overset{\text{(v)}}{\leq} \mathrm{KL}(u \,\|\, p^1) + \sum_{t=1}^{T} \epsilon^t + \frac{\eta^2\, H^2\, T}{2} \\
&\leq \sum_{a \in \mathcal{A}} u(a) \log(u(a)) - \sum_{a \in \mathcal{A}} u(a)\, \log(p^1(a)) + \sum_{t=1}^{T} \epsilon^t + \frac{\eta^2\, H^2\, T}{2} \\
&\overset{\text{(vi)}}{\leq} \log|\mathcal{A}| + \sum_{t=1}^{T} \epsilon^t + \frac{\eta^2\, H^2\, T}{2}\,.
\end{aligned}
$$

(v) follows from the fact that KL-divergence is non-negative, and (vi) stands because the first term is negative, and for the second term, p^1 is a uniform distribution. Dividing both side by η, we have that

$$
\sum_{t=1}^{T} \left\langle u - p^t, g^t \right\rangle \leq \frac{\log|\mathcal{A}| + \sum_{t=1}^{T} \epsilon^t}{\eta} + \frac{\eta\, H^2\, T}{2}\,.
$$

This concludes the proof. □

B.2 Projection Error (Proof of Lemma 4.1)

Proof of Lemma 4.1. First, we define Φ as the log-sum-exp mirror map and $\Phi^\star$ as negative entropy, its Fenchel conjugate. Based on this, for any softmax policy π, we can also define its logit as $z := \nabla\Phi^\star(\pi) = (\nabla\Phi)^{-1}(\pi)$. Consequently, $\pi = \nabla\Phi(z)$. Additionally, for any two softmax policies π, π' and their corresponding logits z, z', it holds that, $D_\Phi(z, z') = \mathrm{KL}(\pi', \pi)$.

Since we are using the log-linear policy class, we have $z_h^t(s,a) = \left\langle \varphi(s,a), \widehat{\theta}_h^t \right\rangle$ for any $(s,a) \in \mathcal{S} \times \mathcal{A}$ where $\widehat{\theta}_h^t$ represents the parameters we attain at episode t. Therefore, for any $s \in \mathcal{S}$,

$$
\begin{aligned}
\epsilon_h^t(s) &= \mathrm{KL}(\pi^\star(\cdot \mid s) \,\|\, \pi_h^{t+1}(\cdot \mid s)) - \mathrm{KL}(\pi_h^\star(\cdot \mid s) \,\|\, \pi_h^{t+1/2}(\cdot \mid s)) \\
&= D_\Phi(z_h^{t+1}(\cdot \mid s), z_h^\star(s,\cdot)) - D_\Phi(z_h^{t+1/2}(\cdot \mid s), z_h^\star(\cdot \mid s)) \\
&\overset{\text{(i)}}{=} \left\langle \nabla\Phi(z_h^{t+1}(s,\cdot))) - \nabla\Phi(z_h^\star(s,\cdot))), z_h^{t+1}(s,\cdot)) - z_h^{t+1/2}(\cdot \mid s)) \right\rangle - D_\Phi(z_h^{t+1/2}(\cdot \mid s), z_h^{t+1}(\cdot \mid s)) \\
&= \left\langle \pi_h^{t+1}(s,\cdot) - \pi_h^\star(s,\cdot), z_h^{t+1}(s,\cdot)) - z_h^{t+1/2}(s,\cdot)) \right\rangle - \mathrm{KL}(\pi_h^{t+1}(\cdot \mid s)) \,\|\, \pi_h^{t+1/2}(\cdot \mid s))) \\
&\overset{\text{(ii)}}{\leq} \left\langle \pi_h^{t+1}(\cdot \mid s) - \pi_h^\star(\cdot \mid s), z_h^{t+1}(s,\cdot)) - z_h^{t+1/2}(s,\cdot)) \right\rangle
\end{aligned}
$$

$$
\begin{aligned}
&\overset{\text{(iii)}}{\leq} \left\|\pi_h^{t+1}(\cdot \mid s) - \pi^\star(\cdot \mid s)\right\|_2 \left\|z_h^{t+1}(s,\cdot)) - z_h^{t+1/2}(s,\cdot))\right\|_2 \\
&\leq \sqrt{2}\left\|z_h^{t+1}(s,\cdot) - z_h^{t+1/2}(s,\cdot)\right\|_2 \\
&\overset{\text{(iv)}}{=} \sqrt{2}\left\|\left\langle \varphi(s,\cdot), \widehat{\theta}_h^{t+1} - \widehat{\theta}_h^t \right\rangle - \eta\, \widehat{Q}_h^t(s,\cdot)\right\|_2 \\
&\overset{\text{(v)}}{\leq} \sqrt{2}\left|\left\langle \varphi(s,\cdot), \widehat{\theta}_h^{t+1} - \widehat{\theta}_h^t \right\rangle - \eta\, \widehat{Q}_h^t(s,\cdot)\right|.
\end{aligned}
$$

(i) follows from the three-point property of Bregman divergence (Lemma B.2) by setting $x = z_h^{t+1/2}(\cdot \mid s)$, $y = z_h^{t+1}(\cdot \mid s)$, and $z = z_h^\star(\cdot \mid s)$ where $z^\star$ is the logit of $\pi^\star$. (ii) is based on the fact that KL-divergence is non-negative. (iii) uses the Cauchy-Schwarz inequality. (iv) uses the NPG update. (v) holds because $\|\cdot\|_2 \leq \|\cdot\|_1$. Since the actor is designed to minimize the ridge regression in Algorithm 2, the minimizer can be written as

$$
\widehat{\theta}_h^{t,\star} = \arg\min_{\theta_h} \frac{1}{2} \sum_{(s,a)\in\mathcal{D}_{\text{exp}}} \rho(s,a)\left[\langle \varphi(s,a), \theta_h\rangle - \widehat{Z}_h^t(s,a)\right]^2,
$$

where $\widehat{Z}_h^t(s,a) \coloneqq \left\langle \varphi(s,\cdot), \widehat{\theta}_h^t(s,\cdot)\right\rangle + \eta\, \widehat{Q}_h^t(s,a)$ for all $t \in [T]$. We define $\widehat{\theta}_h^{t,\star}$ as the minimizer, and it has the following explicit solution:

$$
\widehat{\theta}_h^{t,\star} = G^{-1}\left[\sum_{(s',a')\in\mathcal{D}_{\text{exp}}} \rho(s',a')\, \widehat{Z}_h^t(s',a')\, \varphi(s',a')\right],
$$

where $G \coloneqq \sum_{(s,a)\in\mathcal{D}_{\text{exp}}} \rho(s,a)\, \varphi(s,a)\, \varphi(s,a)^\top \in \mathbb{R}^{d_{\boldsymbol{a}}\times d_{\boldsymbol{a}}}$.

Suppose $\theta_h^{t,\star}$ is the minimizer of the regression loss over the entire state-action space, $\widehat{\theta}_h^{t,\star}$ is the minimizer over the coreset, and $\widehat{\theta}_h^t$ is the parameters produced by the actor after K_t rounds of gradient descent as shown in Algorithm 2. Under Assumptions 4.1 and 4.2, we know that for any t and h,

$$
\begin{aligned}
\left|\left\langle \varphi(s,a), \widehat{\theta}_h^{t,\star}\right\rangle - \widehat{Z}_h^t(s,a))\right| &\leq \sqrt{2\,\epsilon_{\text{bias}}}\,, \\
\left|\left\langle \varphi(s,a), \widehat{\theta}_h^t\right\rangle - \left\langle \varphi(s,a), \widehat{\theta}_h^{t,\star}\right\rangle\right| &\leq \sqrt{2\,\epsilon_{\text{opt}}}\,.
\end{aligned}
$$

Then, for any arbitrary $(s,a) \in \mathcal{S}\times\mathcal{A}$, using the triangular inequality, we have that

$$
\begin{aligned}
&\left|\left\langle \varphi(s,a), \widehat{\theta}_h^t\right\rangle - \widehat{Z}_h^t(s,a))\right| \\
&\quad\leq \left|\left\langle \varphi(s,a), \theta_h^{t,\star}\right\rangle - \widehat{Z}_h^t(s,a))\right| + \left|\left\langle \varphi(s,a), \widehat{\theta}_h^t\right\rangle - \left\langle \varphi(s,a), \theta_h^{t,\star}\right\rangle\right| \\
&\quad= \sqrt{2\,\epsilon_{\text{bias}}} + \left|\left\langle \varphi(s,a), \widehat{\theta}_h^t\right\rangle - \left\langle \varphi(s,a), \theta_h^{t,\star}\right\rangle\right| \\
&\quad\leq \sqrt{2\,\epsilon_{\text{bias}}} + \left|\left\langle \varphi(s,a), \widehat{\theta}_h^t\right\rangle - \left\langle \varphi(s,a), \widehat{\theta}_h^{t,\star}\right\rangle\right| + \left|\left\langle \varphi(s,a), \widehat{\theta}_h^{t,\star}\right\rangle - \left\langle \varphi(s,a), \theta_h^{t,\star}\right\rangle\right| \\
&\quad= \sqrt{2\,\epsilon_{\text{bias}}} + \sqrt{2\,\epsilon_{\text{opt}}} + \left|\left\langle \varphi(s,a), \widehat{\theta}_h^{t,\star} - \theta_h^{t,\star}\right\rangle\right|.
\end{aligned}
$$

Therefore, it suffices to bound $\left|\left\langle \varphi(s,a), \widehat{\theta}_h^{t,\star}(s,a) - \theta_h^{t,\star}(s,a)\right\rangle\right|$. To do that, we first define $\Upsilon(s',a') \coloneqq \widehat{Z}_h^t(s',a') - \left\langle \varphi(s',a'), \theta_h^{t,\star}\right\rangle$ for any $(s',a') \in \mathcal{D}_{\text{exp}}$. Then, we have that

$$
\begin{aligned}
\widehat{\theta}_h^{t,\star} &= G^{-1}\left[\sum_{(s',a')\in\mathcal{D}_{\text{exp}}} \rho(s',a')\left[\Upsilon(s',a') + \left\langle \varphi(s',a'), \theta_h^{t,\star}\right\rangle\right]\varphi(s',a')\right] \\
&= G^{-1}\left[\sum_{(s',a')\in\mathcal{D}_{\text{exp}}} \rho(s',a')\, \varphi(s',a')\, \varphi(s',a')^\top\right]\theta_h^{t,\star} + G^{-1}\left[\sum_{(s',a')\in\mathcal{D}_{\text{exp}}} \rho(s',a')\, \Upsilon(s',a')\, \varphi(s',a')\right]
\end{aligned}
$$

$$= \theta_h^{t,\star} + G^{-1} \left[\sum_{(s',a') \in \mathcal{D}_{\text{exp}}} \rho(s',a')\, \Upsilon(s',a')\, \varphi(s',a') \right].$$

This implies that

$$\widehat{\theta}_h^{t,\star} - \theta_h^{t,\star} = G^{-1} \left[\sum_{(s',a') \in \mathcal{D}_{\text{exp}}} \rho(s',a')\, \Upsilon(s',a')\, \varphi(s',a') \right].$$

Hence, for any arbitrary $(s,a) \in \mathcal{S} \times \mathcal{A}$,

$$\begin{aligned}
\left| \left\langle \varphi(s,a), \widehat{\theta}_h^{t,\star} - \theta_h^{t,\star} \right\rangle \right| &= \left| \sum_{(s',a') \in \mathcal{D}_{\text{exp}}} \rho(s',a')\, \Upsilon(s',a') \varphi(s,a)^\top G^{-1}\, \varphi(s',a') \right| \\
&\overset{\text{(vi)}}{\leq} \sum_{(s',a') \in \mathcal{D}_{\text{exp}}} |\Upsilon(s',a')|\, \rho(s',a') \left| \varphi(s,a)^\top G^{-1}\, \varphi(s',a') \right| \\
&\leq \left(\max_{(s',a') \in \mathcal{D}_{\text{exp}}} |\Upsilon(s',a')| \right) \sum_{(s',a') \in \mathcal{D}_{\text{exp}}} \rho(s',a') \left| \varphi(s,a)^\top G^{-1}\, \varphi(s',a') \right| \\
&\leq \sqrt{2\,\epsilon_{\text{bias}}} \sum_{(s',a') \in \mathcal{D}_{\text{exp}}} \rho(s',a') \left| \varphi(s,a)^\top G^{-1}\, \varphi(s',a') \right| \\
&= \sqrt{2\,\epsilon_{\text{bias}}} \sqrt{\left(\mathbb{E}_{(s',a') \sim \rho} \left| \varphi(s,a)^\top G^{-1}\, \varphi(s',a') \right| \right)^2} \\
&\overset{\text{(vii)}}{\leq} \sqrt{2\,\epsilon_{\text{bias}}} \sqrt{\mathbb{E}_{(s',a') \sim \rho} \left| \varphi(s,a)^\top G^{-1}\, \varphi(s',a') \right|^2} \\
&= \sqrt{2\,\epsilon_{\text{bias}}} \sqrt{\varphi(s,a)^\top G^{-1} \left[\sum_{(s',a') \in \mathcal{D}_{\text{exp}}} \rho(s',a')\, \varphi(s',a')\, \varphi(s',a')^\top \right] G^{-1} \varphi(s,a)} \\
&= \sqrt{2\,\epsilon_{\text{bias}}}\, \|\varphi(s,a)\|_{G^{-1}}\,.
\end{aligned}$$

(vi) applies the Cauchy-Schwarz inequality, and (vii) follows from Jensen's inequality.

Putting everything together, we have that

$$\begin{aligned}
\left| \left\langle \varphi(s,a), \widehat{\theta}_h^t \right\rangle - \widehat{Z}_h^t(s,a)) \right| &\leq \sqrt{2} \left[(\|\varphi(s,a)\|_{G^{-1}} + 1) \sqrt{\epsilon_{\text{bias}}} + \sqrt{\epsilon_{\text{opt}}} \right] \\
&\leq \sqrt{2} \left[(\overline{\varphi}_G + 1) \sqrt{\epsilon_{\text{bias}}} + \sqrt{\epsilon_{\text{opt}}} \right].
\end{aligned}$$

Recall that $\epsilon_h^t(s) \leq \sqrt{2} \left\| \left\langle \varphi(s,\cdot), \widehat{\theta}_h^{t+1}(s,\cdot) \right\rangle - \widehat{Z}_h^t(s,\cdot) \right\|_2$. Therefore, for any $s \in \mathcal{S}$,

$$\epsilon_h^t(s) \leq \sqrt{2} \left| \left\langle \varphi(s,\cdot), \widehat{\theta}_h^t \right\rangle - \widehat{Z}_h^t(s,a)) \right| \leq 2\,(\overline{\varphi}_G + 1) \sqrt{\epsilon_{\text{bias}}} + 2 \sqrt{\epsilon_{\text{opt}}}\,.$$

This concludes the proof. □

B.3 Instantiating the Actor with SPMA

Lemma 4.1 can not only be applied to NPG but also other mirror descent-based policy optimization methods such as MDPO (Tomar et al., 2020) and SPMA (Asad et al., 2025). In this section, as an example, we show that the projected variant of SPMA (projected SPMA) is also compatible with our framework and can enjoy similar sample complexity guarantees as projected NPG.

We can instantiate the actor in Algorithm 1 with the projected SPMA by setting the actor loss in Algorithm 2 as

$$\tilde{\ell}_h^t(\theta) = \frac{1}{2} \sum_{(s,a) \in \mathcal{D}_{\text{exp}}} \rho_{\text{exp}}(s,a) \left[\langle \varphi(s,a), \theta \rangle - \widehat{Z}_h^t(s,a) \right]^2,$$

$$\text{where } \widehat{Z}_h^t(s,a) \coloneqq \left\langle \varphi(s,a), \widehat{\theta}_h^t \right\rangle + \log(1+\eta\, A^{\pi^t}(s,\cdot))\,.$$

Equivalently, the projected SPMA update can be expressed as follows. For any $s \in \mathcal{S}$, $\pi^1(\cdot \mid s)$ is a uniform distribution, and

$$\pi^{t+1/2}(\cdot \mid s) = \arg\min_{p \in \Delta_{\mathcal{A}}} \left\{ \left\langle \pi^t(\cdot \mid s), -\log\left(1+\eta\, A^{\pi^t}(s,\cdot)\right) \right\rangle + \mathrm{KL}\big(p \,\|\, \pi^t(\cdot \mid s)\big) \right\},$$
$$\pi^{t+1}(\cdot \mid s) = \mathrm{Proj}_{\Pi}(\pi^{t+1/2}(\cdot \mid s))\,.$$

Hence, we introduce the following alternative lemma to show that Theorem 4.1 also holds for the projected SPMA.

Lemma B.1. *Given a sequence of linear functions $\{\langle p^t, g^t\rangle\}_{t\in[T]}$ for a sequence of vectors $\{g^t\}_{t\in[T]}$ where for any $t \in [T]$, $p^t \in \Delta(\mathcal{A})$, $g^t \in \mathbb{R}^{|\mathcal{A}|}$, and $g^t(a) \in [0,H]$ for all $a \in \mathcal{A}$. Consider $p^{t\in[T]}$ where p^1 is the uniform distribution, and for all $t \in [T]$,*

$$p^{t+1/2} = \arg\min_{p \in \Delta_{\mathcal{A}}} \left\{ \left\langle p, -\log\left(1+\eta\left(g^t - \left\langle p^t, g^t\right\rangle \mathbf{1}\right)\right) \right\rangle + \mathrm{KL}(p \,\|\, p^t) \right\},$$
$$p^{t+1} = \mathrm{Proj}_{\Pi}(p^{t+1/2})\,,$$

where $\mathbf{1} \in \mathbb{R}^{|\mathcal{A}|}$ is an all-one vector. Let $\epsilon^t \coloneqq \mathrm{KL}(u \,\|\, p^{t+1}) - \mathrm{KL}(u \,\|\, p^{t+1/2})$ be the projection error induced by Eq. (4). If $\eta \le \frac{1}{2H}$, then for any comparator $u \in \Delta(\mathcal{A})$, it holds that

$$\sum_{t=1}^{T} \left\langle u - p^t, g^t \right\rangle \le \frac{\log|\mathcal{A}| + \sum_{t=1}^T \epsilon^t}{\eta} + \frac{3\,\eta\, H^2\, T}{2}\,.$$

Proof of Lemma B.1. We first denote that $d^t = \log(1+\eta\,(g^t - \langle p^t, g^t\rangle\,\mathbf{1}))$ for all $t \in [T]$. Then, for all $a \in \mathcal{A}$, since $\eta \le \frac{1}{2H}$ and $g^t(a) - \langle p^t, g^t\rangle \in [-H, H]$, we have $\eta\,(g^t(a) - \langle p^t, g^t\rangle) > -\frac{1}{2}$ and therefore

$$d^t(a) \overset{\text{(i)}}{\le} \eta\left(g^t(a) - \left\langle p^t, g^t\right\rangle\right) \le \eta\, H\,,$$
$$d^t(a) \overset{\text{(ii)}}{\ge} \eta\left(g^t(a) - \left\langle p^t, g^t\right\rangle\right) - \eta^2\left(g^t(a) - \left\langle p^t, g^t\right\rangle\right)^2 \ge \eta\left(g^t(a) - \left\langle p^t, g^t\right\rangle\right) - \eta\, H^2\,,$$

where (i) follows from $\log(1+x) \le x$ for all $x > -1$, and (ii) holds because $\log(1+x) \ge x - x^2$ for all $x > -\frac{1}{2}$. Given the update of $p^{t+1/2}$ and the fact that $\Delta(\mathcal{A})$ is a convex set, we have the following optimality condition:

$$\left\langle u - p^{t+1/2}, -d^t + \log(p^{t+1/2}) - \log(p^t) \right\rangle \ge 0\,. \tag{9}$$

Then, for all $t \in [T]$, we have that

$$\begin{aligned}
\left\langle u - p^t, d^t \right\rangle &= \left\langle u - p^{t+1/2}, d^t \right\rangle + \left\langle p^{t+1/2} - p^t, d^t \right\rangle \\
&= \left\langle u - p^{t+1/2}, d^t - \log(p^{t+1/2}) + \log(p^t) \right\rangle + \left\langle u - p^{t+1/2}, \log(p^{t+1/2}) - \log(p^t) \right\rangle \\
&\quad + \left\langle p^{t+1/2} - p^t, d^t \right\rangle \\
&\overset{\text{(iii)}}{\le} \left\langle u - p^{t+1/2}, \log(p^{t+1/2}) - \log(p^t) \right\rangle + \left\langle p^{t+1/2} - p^t, d^t \right\rangle \\
&\overset{\text{(iv)}}{=} \mathrm{KL}(u \,\|\, p^t) - \mathrm{KL}(u \,\|\, p^{t+1/2}) - \mathrm{KL}(p^{t+1/2} \,\|\, p^t) + \left\langle p^{t+1/2} - p^t, d^t \right\rangle \\
&\overset{\text{(v)}}{\le} \mathrm{KL}(u \,\|\, p^t) - \mathrm{KL}(u \,\|\, p^{t+1/2}) - \mathrm{KL}(p^{t+1/2} \,\|\, p^t) + \frac{1}{2}\left\|p^{t+1/2} - p^t\right\|_1^2 + \frac{1}{2}\left\|d^t\right\|_\infty^2 \\
&\overset{\text{(vi)}}{\le} \mathrm{KL}(u \,\|\, p^t) - \mathrm{KL}(u \,\|\, p^{t+1/2}) - \mathrm{KL}(p^{t+1/2} \,\|\, p^t) + \mathrm{KL}(p^{t+1/2} \,\|\, p^t) + \frac{\eta^2\, H^2}{2} \\
&\le \mathrm{KL}(u \,\|\, p^t) - \mathrm{KL}(u \,\|\, p^{t+1/2}) + \frac{\eta^2\, H^2}{2}
\end{aligned}$$

$$\leq \mathrm{KL}(u \,\|\, p^t) - \mathrm{KL}(u \,\|\, p^{t+1}) + \mathrm{KL}(u \,\|\, \pi^{t+1}) - \mathrm{KL}(u \,\|\, p^{t+1/2}) + \frac{\eta^2 H^2}{2}$$
$$= \mathrm{KL}(u \,\|\, p^t) - \mathrm{KL}(u \,\|\, p^{t+1}) + \epsilon^t + \frac{\eta^2 H^2}{2}\,.$$

(iii) drops the first term due to the optimality condition from Eq. (9). (iv) applies the three-point property of Bregman divergence (Lemma B.2) by setting $x = u$, $y = p^{t+1/2}$, and $z = p^t$. (v) follows from the Hölder's inequality and then the Young's inequality (i.e., $\langle u, v\rangle \leq \|u\|_1 \|v\|_\infty \leq \|u\|_1^2/2 + \|v\|_\infty^2/2$), and (vi) applies the Pinkster's inequality and $\|d^t\|_\infty \leq H$.

Moreover, we have that

$$\left\langle u - p^t, d^t \right\rangle \overset{\text{(vii)}}{\geq} \left\langle u - p^t, \eta \left(g^t - \left\langle p^t, g^t \right\rangle \mathbf{1}\right)\right\rangle - \eta H^2 \overset{\text{(viii)}}{\geq} \left\langle u - p^t, \eta\, g^t \right\rangle - \eta H^2\,,$$

where (vii) comes from the fact that $d^t(a) \geq \eta\,(g^t(a) - \langle p^t, g^t\rangle) - \eta H^2$ for all $a \in \mathcal{A}$, and (viii) follows from the fact that $\langle p^t, g^t\rangle \geq 0$ since $p^t \in \Delta(\mathcal{A})$ and $g^t(a) \in [0, H]$ for all $a \in \mathcal{A}$. This implies that

$$\left\langle u - p^t, \eta\, g^t \right\rangle \leq \left\langle u - p^t, d^t \right\rangle + \eta H^2$$
$$\leq \mathrm{KL}(u \,\|\, p^t) - \mathrm{KL}(u \,\|\, p^{t+1}) + \epsilon^t + \frac{3\,\eta^2 H^2}{2}\,.$$

Summing up the above inequality from $t = 1$ to T yields that

$$\sum_{t=1}^{T} \left\langle u - p^t, \eta\, g^t \right\rangle = \sum_{t=1}^{T} \mathrm{KL}(u \,\|\, p^t) - \mathrm{KL}(u \,\|\, p^{t+1}) + \sum_{t=1}^{T} \epsilon^t + \frac{3\,\eta^2 H^2 T}{2}$$
$$= \mathrm{KL}(u \,\|\, p^1) - \mathrm{KL}(u \,\|\, p^{T+1}) + \sum_{t=1}^{T} \epsilon^t + \frac{3\,\eta^2 H^2 T}{2}$$
$$\overset{\text{(ix)}}{\leq} \mathrm{KL}(u \,\|\, p^1) + \sum_{t=1}^{T} \epsilon^t + \frac{3\,\eta^2 H^2 T}{2}$$
$$\leq \sum_{a \in \mathcal{A}} u(a) \log(u(a)) - \sum_{a \in \mathcal{A}} u(a)\, \log(p^1(a)) + \sum_{t=1}^{T} \epsilon^t + \frac{3\,\eta^2 H^2 T}{2}$$
$$\overset{\text{(x)}}{\leq} \log|\mathcal{A}| + \sum_{t=1}^{T} \epsilon^t + \frac{3\,\eta^2 H^2 T}{2}\,.$$

(ix) follows from the fact that KL-divergence is non-negative, and (x) stands because the first term is negative, and for the second term, p^1 is a uniform distribution. Dividing both side by η, we have that

$$\sum_{t=1}^{T} \left\langle u - p^t, g^t \right\rangle \leq \frac{\log|\mathcal{A}| + \sum_{t=1}^{T} \epsilon^t}{\eta} + \frac{3\,\eta\, H^2 T}{2}\,.$$

This concludes the proof. □

In order to obtain a meaningful regret bound, we should set $\eta = \min\left\{\frac{1}{2\,H}, \sqrt{\frac{2\,(\log|\mathcal{A}| + \bar{\epsilon}\,T)}{3\,H^2\,T}}\right\}$.

Therefore, under Assumptions 4.1 and 4.2, we can easily prove that Lemma 4.1 also holds for the projected SPMA, and consequently, all the sample complexity guarantees for the projected NPG should also hold.

B.4 Technical Tools

Lemma B.2 (Three-Point Property of Bregman Divergence)**.** *Suppose $X \subseteq \mathbb{R}^d$ is closed and convex. Consider a strictly convex function $\Phi : X \to \mathbb{R}$. For all $x \in X$ and $y, z \in intX$,*

$$D_\Phi(x, y) + D_\Phi(y, z) - D_\Phi(x, z) = \langle \nabla\Phi(z) - \nabla\Phi(y), x - y\rangle\,.$$

C CONSTRUCTING $\mathcal{D}_{\text{exp}}$ and ρ_{exp} VIA EXPERIMENTAL DESIGN

In this section, we introduce various methods of experimental design to bound $\overline{\varphi}_G$ defined in Lemma 4.1. The experimental design problem can be written as

$$\inf_{\substack{\mathcal{D}_{\text{exp}} \in \mathcal{S}\times\mathcal{A} \\ \rho_{\text{exp}} \in \Delta(\mathcal{D}_{\text{exp}})}} \sup_{(s,a)\in\mathcal{S}\times\mathcal{A}} \|\varphi(s,a)\|_{G^{-1}}$$
$$\text{s.t.} \quad G = \sum_{(s,a)\in\mathcal{D}_{\text{exp}}} \rho_{\text{exp}}(s,a)\,\varphi(s,a)\,\varphi(s,a)^\top .$$

In Appendix C.1, we consider constructing a coreset for the policy features. The Kiefer–Wolfowitz theorem guarantees that there exists a coreset that can ensure that $\overline{\varphi}_G$ is bounded, and that such a coreset has a small $O(d)$ size. Such a coreset can be formed using G-experimental design. In Appendix C.2, we consider using the linear MDP features as the policy features and constructing $\mathcal{D}_{\text{exp}}$ through limited interaction with the environment.

C.1 Kiefer–Wolfowitz Theorem and G-Experimental Design

We first introduce the Kiefer–Wolfowitz theorem (Kiefer and Wolfowitz, 1960) which guarantees that there exists a coreset $\mathcal{D}_{\text{exp}}$ and its corresponding distribution ρ_{exp} that can be used to bound $\overline{\varphi}_G$.

Proposition C.1 (Kiefer–Wolfowitz). *Let $G \coloneqq \sum_{(s,a)\in\mathcal{D}_{\text{exp}}} \rho_{\text{exp}}(s,a)\,\varphi(s,a)\,\varphi(s,a)^\top$ be the covariance matrix for any $\mathcal{D}_{\text{exp}} \subset \mathcal{S}\times\mathcal{A}$ and $\rho \in \Delta(\mathcal{D}_{\text{exp}})$. There exists a coreset $\mathcal{D}_{\text{exp}}$ and a distribution ρ_{exp} such that*

$$\sup_{(s,a)\in\mathcal{S}\times\mathcal{A}} \|\varphi(s,a)\|_{G^{-1}} \le 2\,d_{\boldsymbol{a}} \quad \textit{and} \quad |\mathcal{D}_{\text{exp}}| \le 4 d_{\boldsymbol{a}} \log\log(d_{\boldsymbol{a}}+4) + 28\,.$$

Note that the size of $\mathcal{D}_{\text{exp}}$ is also bounded by $\widetilde{\mathcal{O}}(d_{\boldsymbol{a}})$, suggesting that the computation cost of calculating the actor loss over $\mathcal{D}_{\text{exp}}$ is inexpensive. The problem of constructing such a coreset is often framed as G-experimental design, and it can typically be solved using numerous efficient approximation algorithms such as the Franke-Wolfe algorithm (Frank et al., 1956) as mentioned in Todd (2016); Lattimore and Szepesvári (2020). Using $\mathcal{D}_{\text{exp}}$ and ρ_{exp} produced by such methods to construct the actor loss in Algorithm 2 offers the guarantees that $\overline{\varphi}_G \le \mathcal{O}(d_{\boldsymbol{a}})$, which is consequently used to bound the projection error in Lemma 4.1 as $\overline{\epsilon} \le \mathcal{O}(d_{\boldsymbol{a}}\,\epsilon_{\text{bias}} + \epsilon_{\text{opt}})$.

We remark that the coreset construction can be done before the learning process in the actor-critic algorithm since it is independent of the linear MDP environment. However, these algorithms typically require traversing through all the policy features in $\mathcal{S}\times\mathcal{A}$, which is not ideal for large state-action spaces.

C.2 Exploratory Policy and Minimum Eigenvalue

Alternatively, we can choose to use the linear MDP features as the policy features (i.e., $\varphi = \phi$) and construct $\mathcal{D}_{\text{exp}}$ via interacting with the environment. Note that bounding $\overline{\varphi}_G$ is equivalent to controlling $\|\phi(s,a)\|_{G^{-1}}$ for all $(s,a) \in \mathcal{S}\times\mathcal{A}$. Consequently, given that $\|\phi(s,a)\|_2 \le 1$ by the linear MDP assumption and since

$$\|\phi(s,a)\|_{G^{-1}} \le \frac{\|\phi(s,a)\|_2}{\lambda_{\min}(G)} = \frac{1}{\lambda_{\min}(G)}\,,$$

we only need a well-conditioned covariance matrix G that has a positive minimum eigenvalue.

Several existing works (Hao et al., 2021; Agarwal et al., 2021b) assume access to an exploratory (not necessarily optimal) policy π_{exp} that is able to collect such covariance matrices with minimum eigenvalue bounded away from 0. Given that, we can directly apply π_{exp} to roll-out trajectories and collect observations, which can be used to construct $\mathcal{D}_{\text{exp}}$ and the corresponding covariance G.

However, in practice, we rarely have access to such an oracle policy. Consequently, Wagenmaker et al. (2022) proposed a reward-free approach, `CoverTraj`, that can effectively collect such observations without assuming access to an exploratory policy. In particular, the `CoverTraj` algorithm offers the following theoretical guarantee.

Proposition C.2 (Wagenmaker et al. 2022, Theorem 4). *Fix $h \in [H]$ and $\gamma \in [0,1]$. Suppose there exists a problem-dependent constant $\epsilon_{\mathcal{M}} > 0$ such that $\sup_{\pi\in\Pi} \lambda_{\min}\big(\mathbb{E}_\pi\big[\phi(s,a)\,\phi(s,a)^\top\big]\big) \ge \epsilon_{\mathcal{M}}$. Running K rounds of* `CoverTraj` *to collect $\mathcal{D}_{\text{exp}} = \{(s_h^\tau, a_h^\tau)\}_{\tau=1}^K$ where*

$$K = \widetilde{\mathcal{O}}\bigg(\frac{1}{\epsilon_{\mathcal{M}}} \cdot \max\bigg\{\frac{d_{\boldsymbol{c}}}{\gamma^2}, d_{\boldsymbol{c}}^4\,H^3\,\log^3\frac{1}{\delta}\bigg\}\bigg)$$

ensures that for any $\delta \in (0,1)$*, with probability of at least* $1-\delta$*,*

$$\lambda_{\min}(G) \geq \frac{\epsilon_{\mathcal{M}}}{\gamma^2}\,,$$

where $G = \sum_{(s,a)\in\mathcal{D}_{\text{exp}}} \phi(s,a)\,\phi(s,a)^\top$.

Note that `CoverTraj` does not utilize the reward function of the MDP and merely use the transition kernel when interacting with the environment. Alternatively, Wagenmaker and Jamieson (2022) provides another approach, `OptCov`, that utilizes regret minimization algorithms to construct the desired covariance matrix. According to Wagenmaker and Jamieson (2022, Theorem 9), `OptCov` can also offer a similar guarantee of the minimum eigenvalue ensuring that

$$\lambda_{\min}(G) \geq \max\left\{ d_{\boldsymbol{c}}\,\log\left(\frac{1}{\delta}\right), \epsilon_{\mathcal{M}} \right\}.$$

To conclude, the Frank-Wolfe algorithm can be used to form a coreset and subsequently bound $\overline{\varphi}_G$ for any given policy features. If we use the linear MDP features as the policy features, we can construct $\mathcal{D}_{\text{exp}}$ by interacting with the environment. Either having access to an exploratory policy or running `CoverTraj` or `OptCOv` can offer guarantees on the minimum eigenvalues of the covariance matrix, which will consequently control $\overline{\varphi}_G$.

D ANALYSIS FOR THE CRITIC

D.1 Proof of Lemma 5.1

In order to prove Lemma 5.1, we introduce the following "good" event for the estimated value function.

Lemma D.1 (Good Event)**.** *There exists some* $C_\delta > 0$ *such that for any fixed* $\delta \in (0,1)$*, the following event,*

$$\mathcal{E}_\delta \coloneqq \left\{ \forall (t,h) \in [T]\times[H] : \left\| \textstyle\sum_{(s,a,s')\in\mathcal{D}_h^t} \phi(s,a)\left[\widehat{V}_{h+1}^t(s') - \mathbb{P}_h \widehat{V}_{h+1}^t(s,a)\right] \right\|_{(\Lambda_h^t)^{-1}} \leq C_\delta\, H\, \sqrt{d_{\boldsymbol{c}}} \right\},$$

holds with probability at least $1-\delta$ *(i.e.,* $\Pr(\mathcal{E}_\delta) \geq 1-\delta$*).*

The exact definition of C_δ varies between the on-policy and the off-policy settings. We will prove that $\Pr(\mathcal{E}_\delta) \geq 1-\delta$ for the on-policy and off-policy setting in Appendix E and Appendix F respectively.

Next, conditioned on the above event, we present a formal version of Lemma 5.1, which provides an upper and a lower bound for the model prediction error induced by the `LMC` critic.

Lemma D.2 (Formal version of Lemma 5.1)**.** *Consider Algorithm 1 with the* `LMC` *critic from Algorithm 3. Conditioned on* $\mathcal{E}_\delta$ *defined in Lemma D.1, if we choose that* $\lambda = 1$, $\zeta = \left(2\,H\,\sqrt{d_{\boldsymbol{c}}}\,C_\delta + 8/3\right)^{-2}$, $\alpha_{\boldsymbol{c}}^{h,t} = 1/(2\,\lambda_{\max}(\Lambda_h^t)$, $J_t \geq 2\,\kappa_t\,\log(1/\sigma)$*, and* $M = \log(H\,T/\delta)/\log(1/(1-c))$ *where* $\kappa_t = \max_{h\in[H]} \lambda_{\max}(\Lambda_h^t)/\lambda_{\min}(\Lambda_h^t)$, $\sigma = 1/\left(4\,H\,(|\mathcal{D}^t|+1)\,\sqrt{d_{\boldsymbol{c}}}\right)$*, and* $c = 1/(2\sqrt{2e\pi})$*, then, for all* $(t,h,s,a) \in [T]\times[H]\times\mathcal{S}\times\mathcal{A}$ *and for any* $\delta\in(0,1)$ *, with probability at least* $1-\delta$*,*

$$-\Gamma_{\texttt{LMC}} \times \|\phi(s,a)\|_{(\Lambda_h^t)^{-1}} \leq \iota_h^t(s,a) \leq 0\,, \tag{10}$$

where $\Gamma_{\texttt{LMC}} = C_\delta\,H\,\sqrt{d_{\boldsymbol{c}}} + \frac{4}{3}\sqrt{\frac{2\,d_{\boldsymbol{c}}\log(1/\delta)}{3\,\zeta}} + \frac{4}{3} \leq \mathcal{O}\left(C_\delta\,H\,d_{\boldsymbol{c}}\sqrt{\log(1/\delta)}\right)$.

D.1.1 Preliminary Properties

In this section, we introduce some useful properties of `LMC` and state the supporting lemmas that will be helpful in proving the above result.

First, we obtain the derivative of the critic loss defined in Algorithm 3.

$$\nabla L_h^t(w_h) = \Lambda_h^t w_h - b_h^t, \tag{11}$$

where $\Lambda_h^t \coloneqq \sum_{(s,a)\in\mathcal{D}_h^t} \phi(s,a)\,\phi(s,a)^\top + \lambda\, I$ and $b_h^t \coloneqq \sum_{(s,a,s')\in\mathcal{D}_h^t}\left[r_h(s,a) + \widehat{V}_{h+1}^t(s')\right]\phi(s,a)$. Consequently, by setting $\nabla L_h^t(w_h) = 0$, we get the minimizer of $L_h^t(w_h)$ as

$$\widehat{w}_h^t \coloneqq (\Lambda_h^t)^{-1}\, b_h^t\,. \tag{12}$$

We now introduce the following lemma, showing that the noisy gradient descent performed by the `LMC` critic ensures that the sampled critic parameter w follows a Gaussian distribution.

Lemma D.3 (Ishfaq et al. 2024a, Proposition B.1)**.** *Consider Algorithm 1 with the* LMC *critic from Algorithm 3. For any $(t,h,m) \in [T]\times[H]\times[M]$, the sampled parameters w_h^{t,m,J_t} follows a Gaussian distribution $\mathbf{N}\Big(\mu_h^{t,m,J_t}, \Sigma_h^{t,m,J_t}\Big)$. The mean and the covariance are defined as*

$$\mu_h^{t,J_t} = A_t^{J_t} \dots A_1^{J_1}\, w^{1,0} + \sum_{i=1}^{t} A_t^{J_t} \dots A_{i+1}^{J_{i+1}} (I - A_i^{J_i})\, \widehat{w}_h^i\,, \tag{13}$$

$$\Sigma_h^{t,J_t} = \frac{1}{\zeta} \sum_{i=1}^{t} A_t^{J_t} \dots A_{i+1}^{J_{i+1}}\, (I - A_i^{2\,J_i})\, (\Lambda_h^i)^{-1}\, (I + A_i)^{-1}\, A_{i+1}^{J_{i+1}} \dots A_t^{J_t}\,, \tag{14}$$

where $A_t \coloneqq I - \alpha_{\mathbf{c}}^t\, \Lambda_h^t$ for all $t \in [T]$.

Since w_h^{t,m,J_t} follows the Gaussian distribution of $\mathbf{N}\Big(\mu_h^{t,m,J_t}, \Sigma_h^{t,m,J_t}\Big)$, $\Big\langle \phi_h(s,a), w_h^{t,m,J_t} \Big\rangle$ also follows the Gaussian distribution of $\mathbf{N}\Big(\phi_h(s,a)^\top\, \mu_h^{t,m,J_t}, \phi_h(s,a)^\top\, \Sigma_h^{t,m,J_t}\, \phi_h(s,a)\Big)$. Therefore, we introduce the following lemmas to bound the terms related to the mean and variance.

Lemma D.4. *Consider Algorithm 1 with the* LMC *critic from Algorithm 3. If we follow the hyperparameter choices of Lemma D.2, then for any $(s,a) \in \mathcal{S}\times\mathcal{A}$,*

$$\left|\left\langle \phi(s,a), \left(\mu_h^{t,J_t} - \widehat{w}_h^t\right)\right\rangle\right| \le \frac{4}{3}\, \|\phi(s,a)\|_{(\Lambda_h^t)^{-1}}\,.$$

Lemma D.5. *Consider Algorithm 1 with the* LMC *critic from Algorithm 3. If we follow the hyperparameter choices of Lemma D.2, then for any $(s,a) \in \mathcal{S}\times\mathcal{A}$,*

$$\frac{1}{2\sqrt{6\,\zeta}}\, \|\phi(s,a)\|_{(\Lambda_h^t)^{-1}} \le \|\phi(s,a)\|_{\Sigma_h^{t,m,J_t}} \le \frac{4}{3}\sqrt{\frac{2}{3\,\zeta}}\, \|\phi(s,a)\|_{(\Lambda_h^t)^{-1}}\,.$$

Additionally, we outline the necessary supporting lemmas that are useful for bounding the model prediction error. Recall that $|\mathcal{D}^t| \coloneqq \sup_{h\in[H]} |\mathcal{D}_h^t|$ represents the number of trajectories in $\mathcal{D}^t$ or the number of (s_h, a_h, s_{h+1}) tuples in $\mathcal{D}_h^t$, where $|\mathcal{D}^t| = N$ in the on-policy setting, and $|\mathcal{D}^t| = t$ in the off-policy setting.

Lemma D.6. *Consider Algorithm 1 with the* LMC *critic from Algorithm 3. For any $(t,h) \in [T]\times[H]$, it holds that*

$$\left\|\widehat{w}_h^t\right\|_2 \le 2\, H\, \sqrt{d_{\mathbf{c}}\, |\mathcal{D}^t|/\lambda}\,.$$

Lemma D.7. *Consider Algorithm 1 with the* LMC *critic from Algorithm 3. If we follow the hyperparameter choices of Lemma D.2, then for any $(t,m,h) \in [T]\times[M]\times[H]$ and for any $\delta \in (0,1)$, with probability at least $1-\delta$,*

$$\left\|w_h^{t,m,J_t}\right\|_2 \le \overline{W}_\delta^t \coloneqq \frac{16}{3}\, H\, \sqrt{d_{\mathbf{c}}\, |\mathcal{D}^t|} + \sqrt{\frac{2\, d_{\mathbf{c}}^3\, t}{3\,\zeta\,\delta}}\,.$$

Lemma D.8. *Consider Algorithm 1 with the* LMC *critic from Algorithm 3. If we follow the hyperparameter choices of Lemma D.2, then for any $(t,m,h,s,a) \in [T]\times[M]\times[H]\times\mathcal{S}\times\mathcal{A}$ and for any $\delta \in (0,1)$, with probability at least $1-\delta$,*

$$\left|\left\langle \phi(s,a), \widehat{w}_h^t - w_h^{t,m,J_t}\right\rangle\right| \le \left(\frac{8}{3}\sqrt{\frac{2\, d_{\mathbf{c}}\, \log(1/\delta)}{3\,\zeta}} + \frac{4}{3}\right) \|\phi(s,a)\|_{(\Lambda_h^t)^{-1}}\,.$$

Lemma D.9. *Consider Algorithm 1 with the* LMC *critic from Algorithm 3. Conditioned on $\mathcal{E}_\delta$ defined in Lemma D.1, if we follow the hyperparameter choices of Lemma D.2, then for any $(t,h,s,a) \in [T]\times[H]\times\mathcal{S}\times\mathcal{A}$ and for any $\delta \in (0,1)$, it holds that*

$$\left|\left\langle \phi(s,a), \widehat{w}_h^t\right\rangle - r_h(s,a) - \mathbb{P}_h \widehat{V}_{h+1}^t(s,a)\right| \le 3\, C_\delta\, H\, \sqrt{d_{\mathbf{c}}}\, \|\phi(s,a)\|_{(\Lambda_h^t)^{-1}}\,.$$

D.1.2 Main Analysis

We will use the above lemmas to complete the main proof in this section.

Proof of Lemma D.2.

Optimism (RHS of Eq. (10)) Using the definition of the model prediction error, we need to show that with high probability, $\widehat{Q}_h^t(s,a) \geq r_h(s,a) + \mathbb{P}_h \widehat{V}_{h+1}^t(s,a)$. Recall that $\widehat{Q}_h^t(s,a) = \min\Big\{\Big\langle \phi(s,a), w_h^{t,m,J_t}\Big\rangle, H-h+1\Big\}$. Since $r_h(s,a)+\mathbb{P}_h\widehat{V}_{h+1}^t(s,a) \leq H-h+1$, when $\Big\langle \phi(s,a), w_h^{t,m,J_t}\Big\rangle > H-h+1$, the statement is trivially true. Thus, we only need to consider the case when $\Big\langle \phi(s,a), w_h^{t,m,J_t}\Big\rangle \leq H-h+1$ and thus $\widehat{Q}_h^t(s,a) = \Big\langle \phi(s,a), w_h^{t,m,J_t}\Big\rangle$.

Based on the mean and covariance matrix defined in Lemma D.3, we have that $\Big\langle \phi(s,a), w_h^{t,m,J_t}\Big\rangle$ follows the distribution $\mathbf{N}\Big(\phi(s,a)^\top \mu_h^{t,J_t}, \phi(s,a)^\top \Sigma_h^{t,J_t}\, \phi(s,a)\Big)$.

In order to prove that $\widehat{Q}_h^t(s,a) \geq r_h(s,a) + \mathbb{P}_h\widehat{V}_{h+1}^t(s,a)$, we consider the following variable $X_t \coloneqq \frac{r_h(s,a)+\mathbb{P}_h\widehat{V}_{h+1}^t(s,a) - \left\langle \phi(s,a), \mu_h^{t,J_t}\right\rangle}{\sqrt{\phi(s,a)^\top \Sigma_h^{t,J_t}\, \phi(s,a)}}$ and will next show that $|X_t| \leq 1$. First, we have that

$$
\begin{aligned}
&\left| r_h(s,a) + \mathbb{P}_h \widehat{V}_{h+1}^t(s,a) - \Big\langle \phi(s,a), \mu_h^{t,J_t}\Big\rangle \right| \\
&\quad \overset{\text{(i)}}{\leq} \left| r_h(s,a) + \mathbb{P}_h \widehat{V}_{h+1}^t(s,a) - \left\langle \phi(s,a), \widehat{w}_h^t\right\rangle \right| + \left| \Big\langle \phi(s,a), \widehat{w}_h^t - \mu_h^{t,J_t}\Big\rangle \right| \\
&\quad \overset{\text{(ii)}}{\leq} C_\delta\, H\, \sqrt{d_{\boldsymbol{c}}}\, \|\phi(s,a)\|_{(\Lambda_h^t)^{-1}} + \frac{4}{3}\, \|\phi(s,a)\|_{(\Lambda_h^t)^{-1}} \\
&\quad = \left(C_\delta\, H\, \sqrt{d_{\boldsymbol{c}}} + \frac{4}{3}\right) \|\phi(s,a)\|_{(\Lambda_h^t)^{-1}}\,,
\end{aligned}
$$

where (i) uses the triangular inequality, and (ii) is implied by Lemmas D.4 and D.9. Therefore,

$$
\begin{aligned}
|X_t| &= \left| \frac{r_h(s,a) + \mathbb{P}_h \widehat{V}_{h+1}^t(s,a) - \Big\langle \phi(s,a), \mu_h^{t,J_t}\Big\rangle}{\sqrt{\phi(s,a)^\top \Sigma_h^{t,J_t}\, \phi(s,a)}} \right| \\
&\leq \sqrt{\zeta}\Big(2\, H\, \sqrt{d_{\boldsymbol{c}}} C_\delta + 8/3\Big)\,.
\end{aligned}
$$

Since we choose $\zeta = \big(2\, H\, \sqrt{d_{\boldsymbol{c}}} C_\delta + 8/3\big)^{-2}$, we have that $|X_t| \leq 1$. Then, using Lemma D.12, we can get that

$$
\begin{aligned}
&\Pr\Big(\Big\langle \phi(s,a), w_h^{t,m,J_t}\Big\rangle \geq r_h(s,a) + \mathbb{P}_h \widehat{V}_{h+1}^t(s,a)\Big) \\
&\quad = \Pr\left(\frac{\Big\langle \phi(s,a), w_h^{t,m,J_t}\Big\rangle - \Big\langle \phi(s,a), \mu_h^{t,J_t}\Big\rangle}{\sqrt{\phi(s,a)^\top \Sigma_h^{t,J_t}\, \phi(s,a)}} \geq \frac{r_h(s,a) + \mathbb{P}_h \widehat{V}_{h+1}^t(s,a) - \Big\langle \phi(s,a), \mu_h^{t,J_t}\Big\rangle}{\sqrt{\phi(s,a)^\top \Sigma_h^{t,J_t}\, \phi(s,a)}} \right) \\
&\quad = \Pr\left(\frac{\Big\langle \phi(s,a), w_h^{t,m,J_t}\Big\rangle - \Big\langle \phi(s,a), \mu_h^{t,J_t}\Big\rangle}{\sqrt{\phi(s,a)^\top \Sigma_h^{t,J_t}\, \phi(s,a)}} \geq X_t \right) \\
&\quad \geq \frac{1}{2\sqrt{2\pi}} \exp\left(-X_t^2/2\right) \\
&\quad \geq \frac{1}{2\sqrt{2e\pi}}\,.
\end{aligned}
$$

The above result holds for any $m \in [M]$. Since we have M parallel critic parameters, it holds that

$$
\begin{aligned}
&\Pr\Big(\exists (s,a) \in \mathcal{S}\times\mathcal{A} : \widehat{Q}_h^t(s,a) \leq r_h(s,a) + \mathbb{P}_h \widehat{V}_{h+1}^t(s,a)\Big) \\
&\quad = \Pr\Big(\exists (s,a) \in \mathcal{S}\times\mathcal{A} : \max_{m\in[M]} \widehat{Q}_h^{t,m}(s,a) \leq r_h(s,a) + \mathbb{P}_h \widehat{V}_{h+1}^t(s,a)\Big)
\end{aligned}
$$

$$
\begin{aligned}
&= \Pr\Big(\exists(s,a) \in \mathcal{S}\times\mathcal{A} : \forall m \in [M],\ \ \widehat{Q}_h^{t,m}(s,a) \le r_h(s,a) + \mathbb{P}_h \widehat{V}_{h+1}^t(s,a)\Big) \\
&\le \Pr\Big(\forall m \in [M],\ \ \exists(s^m,a^m) \in \mathcal{S}\times\mathcal{A} : \widehat{Q}_h^{t,m}(s^m,a^m) \le r_h(s^m,a^m) + \mathbb{P}_h \widehat{V}_{h+1}^t(s^m,a^m)\Big) \\
&= \prod_{m=1}^{M} \Pr\Big(\exists(s,a) \in \mathcal{S}\times\mathcal{A} : \widehat{Q}_h^{t,m}(s,a) \le r_h(s,a) + \mathbb{P}_h \widehat{V}_{h+1}^t(s,a)\Big) \\
&= \prod_{m=1}^{M} \Big(1 - \Pr\Big(\forall(s,a) \in \mathcal{S}\times\mathcal{A} : \widehat{Q}_h^{t,m}(s,a) \ge r_h(s,a) + \mathbb{P}_h \widehat{V}_{h+1}^t(s,a)\Big)\Big) \\
&= \prod_{m=1}^{M} \Big(1 - \Pr\Big(\forall(s,a) \in \mathcal{S}\times\mathcal{A} : \Big\langle \phi(s,a), w_h^{t,m,J_t}\Big\rangle \ge r_h(s,a) + \mathbb{P}_h \widehat{V}_{h+1}^t(s,a)\Big)\Big) \\
&\le \left(1 - \frac{1}{2\sqrt{2e\pi}}\right)^M .
\end{aligned}
$$

This further implies that

$$
\begin{aligned}
&\Pr\big(\forall(s,a) \in \mathcal{S}\times\mathcal{A} : \iota_h^t(s,a) \le 0\big) \\
&\quad = \Pr\Big(\forall(s,a) \in \mathcal{S}\times\mathcal{A} : \widehat{Q}_h^t(s,a) \ge r_h(s,a) + \mathbb{P}_h \widehat{V}_{h+1}^t(s,a)\Big) \\
&\quad = 1 - \Pr\Big(\exists(s,a) \in \mathcal{S}\times\mathcal{A} : \widehat{Q}_h^t(s,a) \le r_h(s,a) + \mathbb{P}_h \widehat{V}_{h+1}^t(s,a)\Big) \\
&\quad = 1 - \left(1 - \frac{1}{2\sqrt{2e\pi}}\right)^M .
\end{aligned}
$$

Let $1 - \left(1 - \frac{1}{2\sqrt{2e\pi}}\right)^M \ge 1 - \delta/(H\,T)$, which yields that $M = \log(H\,T/\delta)/\log(1/(1-c))$ where $c = 1/(2\sqrt{2e\pi})$. Therefore, we have that

$$
\Pr\big(\iota_h^t(s,a) \le 0,\ \ \forall(s,a) \in \mathcal{S}\times\mathcal{A}\big) \ge 1 - \frac{\delta}{H\,T}\,.
$$

Applying union bound over $[H]$ and $[T]$, we have that $\iota_h^t(s,a) \le 0$ with probability $1-\delta$.

Error Bound (LHS of Eq. (10)) We can lower bound ι_h^t as follows.

$$
\begin{aligned}
-\iota_h^t(s,a) &= \widehat{Q}_h^t(s,a) - r_h(s,a) - \mathbb{P}_h \widehat{V}_{h+1}^t(s,a) \\
&= \min\left\{ \max_{m\in[M]} \Big\langle \phi(s,a), w_h^{t,m,J_t}\Big\rangle, H-h+1 \right\}^+ - r_h(s,a) - \mathbb{P}_h \widehat{V}_{h+1}^t(s,a) \\
&\le \max_{m\in[M]} \Big\langle \phi(s,a), w_h^{t,m,J_t}\Big\rangle - r_h(s,a) - \mathbb{P}_h \widehat{V}_{h+1}^t(s,a) \\
&= \max_{m\in[M]} \Big\langle \phi(s,a), w_h^{t,m,J_t}\Big\rangle - \big\langle \phi(s,a), \widehat{w}_h^t\big\rangle + \big\langle \phi(s,a), \widehat{w}_h^t\big\rangle - r_h(s,a) - \mathbb{P}_h \widehat{V}_{h+1}^t(s,a) \\
&\le \left| \max_{m\in[M]} \Big\langle \phi(s,a), w_h^{t,m,J_t}\Big\rangle - \big\langle \phi(s,a), \widehat{w}_h^t\big\rangle \right| + \left| \big\langle \phi(s,a), \widehat{w}_h^t\big\rangle - r_h(s,a) - \mathbb{P}_h \widehat{V}_{h+1}^t(s,a) \right| \\
&\overset{\text{(iii)}}{\le} \left(C_\delta\, H\, \sqrt{d_{\boldsymbol{c}}} + \frac{4}{3}\sqrt{\frac{2\,d_{\boldsymbol{c}} \log(1/\delta)}{3\,\zeta}} + \frac{4}{3} \right) \|\phi(s,a)\|_{(\Lambda_h^t)^{-1}}\,,
\end{aligned}
$$

where (iii) is derived from Lemmas D.8 and D.9. This concludes the proof. □

D.2 Proofs of Preliminary Properties

D.2.1 Proof of Lemma D.3

Proof. For any $(t,m) \in [T]\times[M]$, the critic update rule at j-th round can be written as

$$
w_h^{t,m,j} = w_h^{t,m,j-1} - \alpha_{\boldsymbol{c}}^{h,t}\, \nabla L_h^t(w_h^{t,m,j-1}) + \sqrt{\alpha_{\boldsymbol{c}}^{h,t}\, \zeta^{-1}}\, \nu_h^{t,m,j}\,.
$$

Considering $j = J_t$ and plugging in Eq. (11), we have that

$$
\begin{aligned}
w_h^{t,m,J_t} &= w_h^{t,m,J_t-1} - \alpha_{\boldsymbol{c}}^{h,t}\Big(\Lambda_h^t\, w_h^{t,m,J_t-1} - b_h^t\Big) + \sqrt{\alpha_{\boldsymbol{c}}^{h,t}\,\zeta^{-1}}\,\nu_h^{t,m,J_t} \\
&= \big(I - \alpha_{\boldsymbol{c}}^{h,t}\,\Lambda_h^t\big)\, w_h^{t,m,J_t-1} + \alpha_{\boldsymbol{c}}^{h,t}\, b_h^t + \sqrt{\alpha_{\boldsymbol{c}}^{h,t}\,\zeta^{-1}}\,\nu_h^{t,m,J_t} \\
&\overset{\text{(i)}}{=} \big(I - \alpha_{\boldsymbol{c}}^{h,t}\,\Lambda_h^t\big)^{J_t}\, w_h^{t,m,0} + \sum_{l=0}^{J_t-1}\big(I - \alpha_{\boldsymbol{c}}^{h,t}\,\Lambda_h^t\big)^l\Big(\alpha_{\boldsymbol{c}}^{h,t} b_h^t + \sqrt{\alpha_{\boldsymbol{c}}^{h,t}\,\zeta^{-1}}\,\nu_h^{t,m,J_t-l}\Big) \\
&= \big(I - \alpha_{\boldsymbol{c}}^{h,t}\,\Lambda_h^t\big)^{J_t}\, w_h^{t,m,0} + \alpha_{\boldsymbol{c}}^{h,t}\sum_{l=0}^{J_t-1}\big(I - \alpha_{\boldsymbol{c}}^{h,t}\,\Lambda_h^t\big)^l\, b_h^t + \sqrt{\alpha_{\boldsymbol{c}}^{h,t}\,\zeta^{-1}}\sum_{l=0}^{J_t-1}\big(I - \alpha_{\boldsymbol{c}}^{h,t}\,\Lambda_h^t\big)^l\,\nu_h^{t,m,J_t-l} \\
&\overset{\text{(ii)}}{=} A_t^{J_t}\, w_h^{t,m,0} + \alpha_{\boldsymbol{c}}^{h,t}\sum_{l=0}^{J_t-1} A_t^l\,\Lambda_h^t\,\widehat{w}_h^t + \sqrt{\alpha_{\boldsymbol{c}}^{h,t}\,\zeta^{-1}}\sum_{l=0}^{J_t-1} A_t^l\,\nu_h^{t,m,J_t-l} \\
&\overset{\text{(iii)}}{=} A_t^{J_t}\, w_h^{t,m,0} + (I - A_t)\Big(A_t^0 + A_t^1 + \ldots + A_t^{J_t-1}\Big)\,\widehat{w}_h^t + \sqrt{\alpha_{\boldsymbol{c}}^{h,t}\,\zeta^{-1}}\sum_{l=0}^{J_t-1} A_t^l\nu_h^{t,m,J_t-l} \\
&\overset{\text{(iv)}}{=} A_t^{J_t}\, w_h^{t,m,0} + \Big(I - A_t^{J_t}\Big)\,\widehat{w}_h^t + \sqrt{\alpha_{\boldsymbol{c}}^{h,t}\,\zeta^{-1}}\sum_{l=0}^{J_t-1} A_t^l\,\nu_h^{t,m,J_t-l}\,.
\end{aligned}
$$

(i) comes from telescoping the previous equation from $l = 0$ to $J_t - 1$. (ii) uses the definition that $A_t = I - \alpha_{\boldsymbol{c}}^{h,t}\,\Lambda_h^t$ and $b_h^t = \Lambda_h^t\widehat{w}_h^t$. (iii) uses the definition of A_t. (iv) follows from $I + A + \ldots + A^{n-1} = (I - A^n)(I - A)^{-1}$. Since we set $\alpha_{\boldsymbol{c}}^{h,t} = 1/(2\,\lambda_{\max}(\Lambda_h^t)$, A_t satisfies $I \succ A_t \succ 0$ for all $t \in [T]$. Note that we warm-start the parameters from the previous episode and set $w_h^{t,m,0} = w_h^{t-1,m,J_{t-1}}$. Therefore, by telescoping the above equation from $i = 0$ to t, we further have that

$$
\begin{aligned}
w_h^{t,m,J_t} &= A_t^{J_t}\, w_h^{t-1,m,J_{t-1}} + \Big(I - A_t^{J_t}\Big)\,\widehat{w}_h^t + \sqrt{\alpha_{\boldsymbol{c}}^{h,t}\,\zeta^{-1}}\sum_{l=0}^{J_t-1} A_t^l\,\nu_h^{t,m,J_t-l} \\
&= A_t^{J_t}\ldots A_1^{J_1} w_h^{1,m,0} + \sum_{i=1}^{t} A_t^{J_t}\ldots A_{i+1}^{J_{i+1}}\Big(I - A_i^{J_i}\Big)\,\widehat{w}_h^i + \sum_{i=1}^{t}\sqrt{\alpha_{\boldsymbol{c}}^i\,\zeta^{-1}}\,A_t^{J_t}\ldots A_{i+1}^{J_{i+1}}\sum_{l=0}^{J_i-1} A_i^l\nu_h^{i,J_i-l}\,.
\end{aligned}
$$

Note that if $\xi \sim \mathbf{N}(0, I_{d\times d})$, then we have that $A\,\xi + \mu \sim \mathbf{N}(\mu, A\,A^\top)$ for any $A \in \mathbb{R}^{d\times d}$ and $\mu \in \mathbb{R}^d$. This implies that w_h^{t,m,J_t} follows the Gaussian distribution $N(\mu_h^{t,m,J_t}, \Sigma_h^{t,m,J_t})$, where

$$
\mu_h^{t,m,J_t} = A_t^{J_t}\ldots A_1^{J_1} w_h^{1,m,0} + \sum_{i=1}^{t} A_t^{J_t}\ldots A_{i+1}^{J_{i+1}}\Big(I - A_i^{J_i}\Big)\,\widehat{w}_h^i\,.
$$

We then derive the covariance matrix Σ_h^{t,m,J_t}. For any $i \in [t]$, we denote that $\mathscr{A}_{i+1} = A_t^{J_t}\ldots A_{i+1}^{J_{i+1}}$. Therefore,

$$
\begin{aligned}
\sqrt{\alpha_{\boldsymbol{c}}^i\,\zeta^{-1}}\,\mathscr{A}_{i+1}\sum_{l=0}^{J_i-1} A_i^l\,\nu_h^{i,J_i-l} &= \sum_{l=0}^{J_i-1}\sqrt{\alpha_{\boldsymbol{c}}^i\,\zeta^{-1}}\,\mathscr{A}_{i+1}\,A_i^l\,\nu_h^{i,J_i-l} \\
&\sim \mathbf{N}\Bigg(0, \sum_{l=0}^{J_i-1}\alpha_{\boldsymbol{c}}^i\,\zeta^{-1}\,\mathscr{A}_{i+1}\,A_i^l\,(\mathscr{A}_{i+1}\,A_i^l)^\top\Bigg) \sim \mathbf{N}\Bigg(0, \alpha_{\boldsymbol{c}}^i\,\zeta^{-1}\,\mathscr{A}_{i+1}\Big(\sum_{l=0}^{J_i-1} A_i^{2l}\Big)\,\mathscr{A}_{i+1}^\top\Bigg)\,.
\end{aligned}
$$

This further implies that

$$
\begin{aligned}
\Sigma_h^{t,m,J_t} &= \sum_{i=1}^{t}\alpha_{\boldsymbol{c}}^i\,\zeta^{-1}\,\mathscr{A}_{i+1}\Big(\sum_{l=0}^{J_i-1} A_i^{2l}\Big)\,\mathscr{A}_{i+1}^\top \\
&= \sum_{i=1}^{t}\alpha_{\boldsymbol{c}}^i\,\zeta^{-1}\,A_t^{J_t}\ldots A_{i+1}^{J_{i+1}}\Big(\sum_{l=0}^{J_i-1} A_i^{2l}\Big)\,A_{i+1}^{J_{i+1}}\ldots A_t^{J_t}
\end{aligned}
$$

$$
\begin{aligned}
&\overset{\text{(v)}}{=} \sum_{i=1}^{t} \alpha_{\boldsymbol{c}}^{i}\, \zeta^{-1}\, A_t^{J_t} \dots A_{i+1}^{J_{i+1}} \left(I - A_i^{2J_i}\right) \left(I - A_i^2\right)^{-1} A_{i+1}^{J_{i+1}} \dots A_t^{J_t} \\
&= \sum_{i=1}^{t} \alpha_{\boldsymbol{c}}^{i}\, \zeta^{-1}\, A_t^{J_t} \dots A_{i+1}^{J_{i+1}} \left(I - A_i^{2J_i}\right) \left(\Lambda_h^i\right) \left(\Lambda_h^i\right)^{-1} (I - A_i)^{-1} (I + A_i)^{-1} A_{i+1}^{J_{i+1}} \dots A_t^{J_t} \\
&\overset{\text{(vi)}}{=} \sum_{i=1}^{t} \zeta^{-1}\, A_t^{J_t} \dots A_{i+1}^{J_{i+1}} \left(I - A_i^{2J_i}\right) \left(\Lambda_h^i\right)^{-1} (I + A_i)^{-1} A_{i+1}^{J_{i+1}} \dots A_t^{J_t}\,.
\end{aligned}
$$

(v) uses the fact that $I + A + \ldots + A^{n-1} = (I - A^n)(I - A)^{-1}$, and (vi) uses the fact that $\alpha_{\boldsymbol{c}}^{h,t}\, \Lambda_h^t = I - A_t$. This concludes the proof. □

D.2.2 Proof of Lemma D.4

Proof. Using Lemma D.3, we first have that

$$
\begin{aligned}
\mu_h^{t,J_t} &= A_t^{J_t} \dots A_1^{J_1} w_h^{1,m,0} + \sum_{i=1}^{t} A_t^{J_t} \dots A_{i+1}^{J_{i+1}} \left(I - A_i^{J_i}\right) \widehat{w}_h^i \\
&= A_t^{J_t} \dots A_1^{J_1} w_h^{1,m,0} + \sum_{i=1}^{t} A_t^{J_t} \dots A_{i+1}^{J_{i+1}}\, \widehat{w}_h^i - \sum_{i=1}^{t} A_t^{J_t} \dots A_i^{J_i}\, \widehat{w}_h^i \\
&= A_t^{J_t} \dots A_1^{J_1} w_h^{1,m,0} + \sum_{i=1}^{t-1} A_t^{J_t} \dots A_{i+1}^{J_{i+1}} \left(\widehat{w}_h^i - \widehat{w}_h^{i+1}\right) - A_t^{J_t} \dots A_1^{J_1}\, \widehat{w}_h^1 + \widehat{w}_h^t \\
&= A_t^{J_t} \dots A_1^{J_1} \left(w_h^{1,m,0} - \widehat{w}_h^1\right) + \sum_{i=1}^{t-1} A_t^{J_t} \dots A_{i+1}^{J_{i+1}} \left(\widehat{w}_h^i - \widehat{w}_h^{i+1}\right) + \widehat{w}_h^t\,.
\end{aligned}
$$

This implies that

$$
\begin{aligned}
\left|\left\langle \phi(s,a), \left(\mu_h^{t,J_t} - \widehat{w}_h^t\right)\right\rangle\right| &= \phi(s,a)^\top A_t^{J_t} \dots A_1^{J_1} \left(w_h^{1,m,0} - \widehat{w}_h^1\right) + \phi(s,a)^\top \sum_{i=1}^{t-1} A_t^{J_t} \dots A_{i+1}^{J_{i+1}} \left(\widehat{w}_h^i - \widehat{w}_h^{i+1}\right) \\
&\overset{\text{(i)}}{=} \left|\phi(s,a)^\top \sum_{i=0}^{t-1} A_t^{J_t} \dots A_{i+1}^{J_{i+1}} \left(\widehat{w}_h^i - \widehat{w}_h^{i+1}\right)\right| \\
&= \left|\sum_{i=0}^{t-1} \phi(s,a)^\top A_t^{J_t} \dots A_{i+1}^{J_{i+1}} \left(\widehat{w}_h^i - \widehat{w}_h^{i+1}\right)\right| \\
&\overset{\text{(ii)}}{\leq} \sum_{i=0}^{t-1} \prod_{j=i+1}^{t} \left(1 - \alpha_{\boldsymbol{c}}^{h,j}\, \lambda_{\min}\left(\Lambda_h^j\right)\right)^{J_j} \|\phi(s,a)\|_2 \|\widehat{w}_h^i - \widehat{w}_h^{i+1}\|_2 \\
&\overset{\text{(iii)}}{\leq} \sum_{i=0}^{t-1} \prod_{j=i+1}^{t} \left(1 - \alpha_{\boldsymbol{c}}^{h,j}\, \lambda_{\min}\left(\Lambda_h^j\right)\right)^{J_j} \|\phi(s,a)\|_2 \left(\|\widehat{w}_h^i\|_2 + \|\widehat{w}_h^{i+1}\|_2\right) \\
&\overset{\text{(iv)}}{\leq} 4\, H \sqrt{d_{\boldsymbol{c}}\, |\mathcal{D}^t| / \lambda} \sum_{i=0}^{t-1} \prod_{j=i+1}^{t} \left(1 - \alpha_{\boldsymbol{c}}^{h,j}\, \lambda_{\min}\left(\Lambda_h^j\right)\right)^{J_j} \|\phi(s,a)\|_2 \\
&\overset{\text{(v)}}{\leq} 4\, H \left(|\mathcal{D}^t| + 1\right) \sqrt{d_{\boldsymbol{c}} / \lambda} \sum_{i=0}^{t-1} \prod_{j=i+1}^{t} \left(1 - \alpha_{\boldsymbol{c}}^{h,j}\, \lambda_{\min}\left(\Lambda_h^j\right)\right)^{J_j} \|\phi(s,a)\|_{\left(\Lambda_h^i\right)^{-1}} \\
&\overset{\text{(vi)}}{\leq} 4\, H \left(|\mathcal{D}^t| + 1\right) \sqrt{d_{\boldsymbol{c}}} \sum_{i=0}^{t-1} \sigma^{t-i} \|\phi(s,a)\|_{\left(\Lambda_h^i\right)^{-1}} \\
&\overset{\text{(vii)}}{\leq} 4\, H \left(|\mathcal{D}^t| + 1\right) \sqrt{d_{\boldsymbol{c}}} \left(\sum_{i=0}^{t-1} \sigma^{t-i}\right) \|\phi(s,a)\|_{\left(\Lambda_h^t\right)^{-1}}
\end{aligned}
$$

$$
\begin{aligned}
&\leq 4\,H\left(\left|\mathcal{D}^t\right|+1\right)\sqrt{d_{\boldsymbol{c}}}\left(\sum_{i=0}^{t-1}\sigma^{t-i}\right)\left\|\phi(s,a)\right\|_{\left(\Lambda_h^t\right)^{-1}}\\
&= 4\,H\left(\left|\mathcal{D}^t\right|+1\right)\sqrt{d_{\boldsymbol{c}}}\left(\sum_{i=1}^{t-1}\sigma^{i}\right)\left\|\phi(s,a)\right\|_{\left(\Lambda_h^t\right)^{-1}}\\
&\overset{\text{(viii)}}{\leq} 4\,H\left(\left|\mathcal{D}^t\right|+1\right)\sqrt{d_{\boldsymbol{c}}}\left(\frac{\sigma}{1-\sigma}\right)\left\|\phi(s,a)\right\|_{\left(\Lambda_h^t\right)^{-1}}\\
&= \left(\frac{1}{1-\sigma}\right)\left\|\phi(s,a)\right\|_{\left(\Lambda_h^t\right)^{-1}}\\
&\leq \frac{4}{3}\left\|\phi(s,a)\right\|_{\left(\Lambda_h^t\right)^{-1}}.
\end{aligned}
$$

For (i), we choose $w_h^{1,m,0} = \mathbf{0}$ and denote that $\widehat{w}_h^0 = \mathbf{0}$. (ii) comes from $A_i \prec \left(1-\alpha_{\boldsymbol{c}}^{h,j}\,\lambda_{\min}\left(\Lambda_h^j\right)\right) I$ and the Hölder's inequality. (iii) uses the triangular inequality. (iv) uses Lemma D.6. (v) uses the fact that $\|\phi(s,a)\| \leq \sqrt{|\mathcal{D}^t|+1}\|\phi(s,a)\|_{(\Lambda_h^i)^{-1}}$. (vi) hold because we set $\lambda = 1$ and uses Lemma D.16 by setting $J_j \geq \kappa_j \log(1/\sigma)$ where $\sigma = 1/\left(4\,H\left(|\mathcal{D}^t|+1\right)\sqrt{d_{\boldsymbol{c}}}\right)$. (vii) follows from $\|\phi(s,a)\|_{\left(\Lambda_h^i\right)^{-1}} \leq \|\phi(s,a)\|_2 \leq \sqrt{|\mathcal{D}^t|+1}\,\|\phi(s,a)\|_{\left(\Lambda_h^t\right)^{-1}}$. (viii) follows from $\sum_{i=1}^t \sigma^t \leq \sum_{i=1}^\infty \sigma^i \leq \sigma/(1-\sigma)$. This concludes the proof. □

D.2.3 Proof of Lemma D.5

Proof. We first bound the RHS. Using Lemma D.3, we have that

$$
\begin{aligned}
&\phi(s,a)^\top \Sigma_h^{t,J_t}\,\phi(s,a)\\
&\quad= \frac{1}{\zeta}\sum_{i=1}^{t}\phi(s,a)^\top A_t^{J_t}\ldots A_{i+1}^{J_{i+1}}\left(I-A^{2J_i}\right)\left(\Lambda_h^i\right)^{-1}\left(I+A_i\right)^{-1}A_{i+1}^{J_{i+1}}\ldots A_t^{J_t}\,\phi(s,a)\\
&\quad\overset{\text{(i)}}{=} \frac{1}{\zeta}\sum_{i=1}^{t}\phi(s,a)^\top \mathscr{A}_{i+1}\left(I-A^{2J_i}\right)\left(\Lambda_h^i\right)^{-1}\left(I+A_i\right)^{-1}\mathscr{A}_{i+1}^\top\,\phi(s,a)\\
&\quad\overset{\text{(ii)}}{\leq} \frac{2}{3\,\zeta}\sum_{i=1}^{t}\phi(s,a)^\top \mathscr{A}_{i+1}\left(\left(\Lambda_h^i\right)^{-1}-A_i^{J_i}\left(\Lambda_h^i\right)^{-1}A_i^{J_i}\right)\mathscr{A}_{i+1}^\top\,\phi(s,a)\\
&\quad= \frac{2}{3\,\zeta}\left(\sum_{i=1}^{t}\phi(s,a)^\top \mathscr{A}_{i+1}\left(\Lambda_h^i\right)^{-1}\mathscr{A}_{i+1}^\top\,\phi(s,a)-\sum_{i=1}^{t}\phi(s,a)^\top \mathscr{A}_{i}\left(\Lambda_h^i\right)^{-1}\mathscr{A}_{i}^\top\,\phi(s,a)\right)\\
&\quad\overset{\text{(iii)}}{\leq} \frac{2}{3\,\zeta}\sum_{i=1}^{t}\phi(s,a)^\top \mathscr{A}_{i+1}\left(\Lambda_h^i\right)^{-1}\mathscr{A}_{i+1}^\top\,\phi(s,a)\\
&\quad= \frac{2}{3\,\zeta}\left(\left\|\phi(s,a)\right\|^2_{\left(\Lambda_h^i\right)^{-1}}+\sum_{i=1}^{t-1}\left\|\mathscr{A}_{i+1}^\top\,\phi(s,a)\right\|^2_{\left(\Lambda_h^i\right)^{-1}}\right)\\
&\quad\leq \frac{2}{3\,\zeta}\left\|\phi(s,a)\right\|^2_{\left(\Lambda_h^t\right)^{-1}}+\frac{2}{3\,\zeta}\sum_{i=1}^{t-1}\prod_{j=i+1}^{t}\left(1-\alpha_{\boldsymbol{c}}\,\lambda_{\min}(\Lambda_h^j)\right)^{2\,J_j}\left\|\phi(s,a)\right\|^2_{\left(\Lambda_h^i\right)^{-1}}.
\end{aligned}
$$

For (i), we denote $\mathscr{A}_{i+1} = A_t^{J_t}\ldots A_{i+1}^{J_{i+1}}$. (ii) follows from $I+A_i \succeq \frac{3}{2}I$ since we set $\alpha_{\boldsymbol{c}}^{h,j} = 1/\left(2\,\lambda_{\max}(\Lambda_h^j)\right)$. In particular, it is easy to prove that A and $(\Lambda_h^t)^{-1}$ are commuting matrices and thus

$$
\begin{aligned}
A^{2J_i}\left(\Lambda_h^i\right)^{-1} &= A^{2J_i-1}\left(I-\alpha_{\boldsymbol{c}}^{h,t}\,\Lambda_h^t\right)\left(\Lambda_h^t\right)^{-1}\\
&= A^{2J_i-1}\left(\Lambda_h^t\right)^{-1}\left(I-\alpha_{\boldsymbol{c}}^{h,t}\,\Lambda_h^t\right)\\
&= A^{2J_i-1}\left(\Lambda_h^t\right)^{-1}A\\
&\;\;\vdots
\end{aligned}
$$

$$= A^{J_i} \left(\Lambda_h^t\right)^{-1} A^{J_i} .$$

(iii) follows from the fact that $\sum_{i=1}^t \phi(s,a)^\top \mathscr{A}_i \left(\Lambda_h^i\right)^{-1} \mathscr{A}_i^\top \phi(s,a) > 0$. Therefore,

$$\begin{aligned}
\left\|\phi(s,a)^\top \left(\Sigma_h^{t,J_t}\right)^{1/2}\right\|_2 &= \sqrt{\phi(s,a)^\top \Sigma_h^{t,J_t} \phi(s,a)} \\
&\overset{\text{(iv)}}{\leq} \sqrt{\frac{2}{3\,\zeta}} \|\phi(s,a)\|_{\left(\Lambda_h^t\right)^{-1}} + \sqrt{\frac{2}{3\,\zeta}} \sum_{i=1}^{t-1} \prod_{j=i+1}^{t} \left(1 - \alpha_{\boldsymbol{c}}\, \lambda_{\min}(\Lambda_h^j)\right)^{J_j} \|\phi(s,a)\|_{\left(\Lambda_h^i\right)^{-1}} \\
&\overset{\text{(v)}}{\leq} \sqrt{\frac{2}{3\,\zeta}} \|\phi(s,a)\|_{\left(\Lambda_h^t\right)^{-1}} + \sqrt{\frac{2}{3\,\zeta}} \sum_{i=1}^{t-1} \sigma^{t-i} \|\phi(s,a)\|_{\left(\Lambda_h^i\right)^{-1}} \\
&\overset{\text{(vi)}}{\leq} \sqrt{\frac{2}{3\,\zeta}} \|\phi(s,a)\|_{\left(\Lambda_h^t\right)^{-1}} + \sqrt{\frac{2\left(|\mathcal{D}^t|+1\right)}{3\,\zeta}} \left(\sum_{i=1}^{t-1} \sigma^{t-i}\right) \|\phi(s,a)\|_{\left(\Lambda_h^t\right)^{-1}} \\
&\overset{\text{(vii)}}{\leq} \sqrt{\frac{2}{3\,\zeta}} \|\phi(s,a)\|_{\left(\Lambda_h^t\right)^{-1}} + \sqrt{\frac{2\left(|\mathcal{D}^t|+1\right)}{3\,\zeta}} \left(\frac{\sigma}{1-\sigma}\right) \|\phi(s,a)\|_{\left(\Lambda_h^t\right)^{-1}} \\
&\leq \sqrt{\frac{2}{3\,\zeta}} \|\phi(s,a)\|_{\left(\Lambda_h^t\right)^{-1}} + \frac{1}{4} \sqrt{\frac{2}{3\,\zeta}} \left(\frac{1}{1-\sigma}\right) \|\phi(s,a)\|_{\left(\Lambda_h^t\right)^{-1}} \\
&\leq \left(\sqrt{\frac{2}{3\,\zeta}} + \frac{1}{3}\sqrt{\frac{2}{3\,\zeta}}\right) \|\phi(s,a)\|_{\left(\Lambda_h^t\right)^{-1}} \\
&\leq \frac{4}{3}\sqrt{\frac{2}{3\,\zeta}} \|\phi(s,a)\|_{\left(\Lambda_h^t\right)^{-1}} .
\end{aligned}$$

(iv) follows from the fact that $\sqrt{a+b} \leq a+b$ for all $a, b > 0$. (v) uses Lemma D.16 by setting $J_j \geq 2\,\kappa_j \log\left(1/\sigma\right)$ where $\sigma = 1/\left(4\,H\left(|\mathcal{D}^t|+1\right)\sqrt{d_{\boldsymbol{c}}}\right)$. (vi) follows from $\|\phi(s,a)\|_{\left(\Lambda_h^i\right)^{-1}} \leq \|\phi(s,a)\|_2 \leq \sqrt{|\mathcal{D}^t|+1}\, \|\phi(s,a)\|_{\left(\Lambda_h^t\right)^{-1}}$. (vii) follows from $\sum_{i=1}^t \sigma^{t-i} \leq \sum_{i=1}^\infty \sigma^i \leq \sigma/(1-\sigma)$.

We then proceed to bound the LHS. Using the definition of Σ_h^{t,J_t} from Eq. (14), we have

$$\begin{aligned}
&\phi(s,a)^\top \Sigma_h^{t,J_t} \phi(s,a) \\
&\quad = \sum_{i=1}^{t} \frac{1}{\zeta} \phi(s,a)^\top A_t^{J_t} \dots A_{i+1}^{J_{i+1}} \left(I - A^{2J_i}\right) \left(\Lambda_h^i\right)^{-1} \left(I + A_i\right)^{-1} A_{i+1}^{J_{i+1}} \dots A_t^{J_t} \phi(s,a) \\
&\quad \overset{\text{(iii)}}{\geq} \frac{1}{2\,\zeta} \sum_{i=1}^{t} \phi(s,a)^\top \mathscr{A}_{i+1} \left(I - A^{2J_i}\right) \left(\Lambda_h^i\right)^{-1} \mathscr{A}_{i+1}^\top \phi(s,a) \\
&\quad = \frac{1}{2\,\zeta} \sum_{i=1}^{t} \frac{1}{2\,\zeta} \phi(s,a)^\top \mathscr{A}_{i+1} \left(\left(\Lambda_h^i\right)^{-1} - A_t^{J_t} \left(\Lambda_h^i\right)^{-1} A_t^{J_t}\right) \mathscr{A}_{i+1}^\top \phi(s,a) \\
&\quad = \frac{1}{2\,\zeta} \sum_{i=1}^{t-1} \phi(s,a)^\top \mathscr{A}_{i+1} \left((\Lambda_h^i)^{-1} - (\Lambda_h^{i+1})^{-1}\right) \mathscr{A}_{i+1}^\top \phi(s,a) \\
&\qquad - \frac{1}{2\,\zeta} \phi(s,a)^\top A_t^{J_t} \dots A_1^{J_1} (\Lambda_h^1)^{-1} A_1^{J_1} \dots A_t^{J_t} \phi(s,a) + \frac{1}{2\,\zeta} \phi(s,a)^\top (\Lambda_h^t)^{-1} \phi(s,a) ,
\end{aligned}$$

where (iii) follows from $(I + A_t)^{-1} \succeq \frac{1}{2}\, I$ for all $t \in [T]$.

$$\begin{aligned}
&\left|\phi(s,a)^\top \mathscr{A}_{i+1} \left((\Lambda_h^i)^{-1} - (\Lambda_h^{i+1})^{-1}\right) \mathscr{A}_{i+1}^\top \phi(s,a)\right| \\
&\qquad \leq \left|\phi(s,a)^\top \mathscr{A}_{i+1} \left(\Lambda_h^i\right)^{-1} \mathscr{A}_{i+1}^\top \phi(s,a)\right| \\
&\qquad\quad + \left|\left\langle \phi(s,a), \mathscr{A}_{i+1} \left(\Lambda_h^{i+1}\right)^{-1} \mathscr{A}_{i+1}^\top \phi(s,a)\right\rangle\right| \\
&\qquad \leq \left\|\phi(s,a)^\top \mathscr{A}_{i+1} \left(\Lambda_h^i\right)^{-1/2}\right\|^2 + \left\|\phi(s,a)^\top \mathscr{A}_{i+1} \left(\Lambda_h^{i+1}\right)^{-1/2}\right\|^2
\end{aligned}$$

$$= \prod_{j=i+1}^{t} \left(1 - \alpha_{\boldsymbol{c}}^{h,j}\, \lambda_{\min}(\Lambda_h^j)\right)^{2\,J_j} \left(\|\phi(s,a)\|^2_{(\Lambda_h^i)^{-1}} + \|\phi(s,a)\|^2_{(\Lambda_h^{i+1})^{-1}}\right)$$
$$\leq 2 \prod_{j=i+1}^{t} \left(1 - \alpha_{\boldsymbol{c}}^{h,j}\, \lambda_{\min}(\Lambda_h^j)\right)^{2\,J_j} \|\phi(s,a)\|_2^2\,,$$

where we used $0 < \|\phi(s,a)\|_{(\Lambda_h^i)^{-1}} \leq \|\phi(s,a)\|_2$. Therefore, we have that

$$\begin{aligned}
&\phi(s,a)^\top \Sigma_h^{t,J_t}\, \phi(s,a) \\
&\quad \geq \frac{1}{2\,\zeta} \|\phi(s,a)\|^2_{(\Lambda_h^t)^{-1}} - \frac{1}{2\,\zeta} \prod_{i=1}^{t} \left(1 - \alpha_{\boldsymbol{c}}^{h,j} \lambda_{\min}(\Lambda_h^i)\right)^{2J_i} \|\phi(s,a)\|_2^2 \\
&\qquad - \frac{1}{\zeta} \sum_{i=1}^{t-1} \prod_{j=i+1}^{t} \left(1 - \alpha_{\boldsymbol{c}}^{h,j}\, \lambda_{\min}(\Lambda_h^j)\right)^{2\,J_j} \|\phi(s,a)\|_2^2 \\
&\quad \overset{\text{(iv)}}{\geq} \frac{1}{2\,\zeta} \left(\|\phi(s,a)\|^2_{(\Lambda_h^t)^{-1}} - \sigma^t \|\phi(s,a)\|_2^2 - \sum_{i=1}^{t-1} 2\,\sigma^i\, \|\phi(s,a)\|_2^2\right) \\
&\quad \overset{\text{(v)}}{\geq} \frac{1}{2\,\zeta} \|\phi(s,a)\|^2_{(\Lambda_h^t)^{-1}} \left(1 - \left(\left|\mathcal{D}^t\right| + 1\right) \sigma^t - 2\left(\left|\mathcal{D}^t\right| + 1\right) \sum_{i=1}^{t-1} \sigma^i\right) \\
&\quad \geq \frac{1}{2\,\zeta} \|\phi(s,a)\|^2_{(\Lambda_h^t)^{-1}} \left(1 - \sigma^{t-1} - \frac{1}{2\,(1-\sigma)}\right) \\
&\quad \geq \frac{1}{2\,\zeta} \|\phi(s,a)\|^2_{(\Lambda_h^t)^{-1}} \left(1 - \frac{1}{4} - \frac{2}{3}\right) \\
&\quad = \frac{1}{24\,\zeta} \|\phi(s,a)\|^2_{(\Lambda_h^t)^{-1}}\,,
\end{aligned}$$

where (iv) uses Lemma D.16 by setting $J_j \geq 2\,\kappa_j \log(1/\sigma)$ where $\sigma = 1/\left(4\,H\left(\left|\mathcal{D}^t\right| + 1\right)\sqrt{d_{\boldsymbol{c}}}\right)$, and (v) use $\|\phi(s,a)\|_2 \leq \sqrt{\left|\mathcal{D}^t\right| + 1}\, \|\phi(s,a)\|_{(\Lambda_h^t)^{-1}}$. This concludes the proof. □

D.2.4 Proof of Lemma D.6

Proof. Given the definition of $\widehat{w}_h^t$ in Eq. (12), we have that

$$\begin{aligned}
\left\|\widehat{w}_h^t\right\| &= \left\| \left(\Lambda_h^t\right)^{-1} \sum_{(s,a)\in\mathcal{D}_h^t} \left[r_h(s,a) + \widehat{V}_{h+1}^t(s)\right] \cdot \phi(s_h,a) \right\| \\
&\leq \sqrt{\frac{\left|\mathcal{D}^t\right|}{\lambda}} \left(\sum_{(s,a)\in\mathcal{D}_h^t} \left\| \left[r_h(s,a) + \widehat{V}_{h+1}^t(s)\right] \cdot \phi(s,a) \right\|^2_{(\Lambda_h^t)^{-1}} \right)^{1/2} \\
&\leq 2\,H \sqrt{\frac{\left|\mathcal{D}^t\right|}{\lambda}} \left(\sum_{(s,a)\in\mathcal{D}_h^t} \|\phi(s,a)\|^2_{(\Lambda_h^t)^{-1}} \right)^{1/2} \\
&\leq 2\,H \sqrt{d_{\boldsymbol{c}}\, \left|\mathcal{D}^t\right| / \lambda}\,,
\end{aligned}$$

where the first inequality follows from Lemma D.15, the second inequality is due to the fact that $V_h^t \in [0, H]$ and the reward function is bounded by 1, and the last inequality follows from Lemma D.10. □

D.2.5 Proof of Lemma D.7

Proof. From Lemma D.3, we know w_h^{t,m,J_t} follows Gaussian distribution $\mathbf{N}(\mu_h^{t,J_t}, \Sigma_h^{t,J_t})$. Therefore, we have that

$$\left\|w_h^{t,m,J_t}\right\|_2 = \left\|\mu_h^{t,J_t} + \xi_h^{t,J_t}\right\|_2 \leq \underbrace{\left\|\mu_h^{t,J_t}\right\|_2}_{\text{(I)}} + \underbrace{\left\|\xi_h^{t,J_t}\right\|_2}_{\text{(II)}},$$

where $\xi_h^{t,J_t} \sim \mathbf{N}(0, \Sigma_h^{t,J_t})$. We first start by bounding Term (I). Given Lemma D.3, by setting $w_h^{1,m,0} = \mathbf{0}$, we can obtain that

$$
\begin{aligned}
\left\|\mu_h^{t,J_t}\right\|_2 &= \left\|A_t^{J_t} \dots A_1^{J_1} w_h^{1,m,0} + \sum_{i=1}^{t} A_t^{J_t} \dots A_{i+1}^{J_{i+1}} \left(I - A_i^{J_i}\right) \widehat{w}_h^i\right\|_2 \\
&\leq \sum_{i=1}^{t} \left\|A_t^{J_t} \dots A_{i+1}^{J_{i+1}} \left(I - A_i^{J_i}\right) \widehat{w}_h^i\right\|_2 \\
&\overset{\text{(i)}}{\leq} \sum_{i=1}^{t} \left\|A_t^{J_t} \dots A_{i+1}^{J_{i+1}} \left(I - A_i^{J_i}\right)\right\|_2 \left\|\widehat{w}_h^i\right\|_2 \\
&\overset{\text{(ii)}}{\leq} 2\,H\,\sqrt{d_{\boldsymbol{c}}\,|\mathcal{D}^t|} \sum_{i=1}^{t} \left\|A_t^{J_t} \dots A_{i+1}^{J_{i+1}} \left(I - A_i^{J_i}\right)\right\|_2 \\
&\overset{\text{(iii)}}{\leq} 2\,H\,\sqrt{d_{\boldsymbol{c}}\,|\mathcal{D}^t|} \sum_{i=1}^{t} \|A_t\|_2^{J_t} \dots \|A_{i+1}\|_2^{J_{i+1}} \left\|\left(I - A_i^{J_i}\right)\right\|_2 \\
&\overset{\text{(iv)}}{\leq} 2\,H\,\sqrt{d_{\boldsymbol{c}}\,|\mathcal{D}^t|} \sum_{i=1}^{t} \prod_{j=i+1}^{t} \left(1 - \alpha_{\boldsymbol{c}}^{h,j}\,\lambda_{\min}(\Lambda_h^j)\right)^{J_j} \left(\|I\|_2 + \|A_i^{J_i}\|_2\right) \\
&\overset{\text{(v)}}{\leq} 2\,H\,\sqrt{d_{\boldsymbol{c}}\,|\mathcal{D}^t|} \sum_{i=1}^{t} \prod_{j=i+1}^{t} \left(1 - \alpha_{\boldsymbol{c}}^{h,j}\,\lambda_{\min}(\Lambda_h^j)\right)^{J_j} \left(\|I\|_2 + \|A_i\|_2^{J_i}\right) \\
&\overset{\text{(vi)}}{\leq} 2\,H\,\sqrt{d_{\boldsymbol{c}}\,|\mathcal{D}^t|} \sum_{i=1}^{t} \prod_{j=i+1}^{t} \left(1 - \alpha_{\boldsymbol{c}}^{h,j}\,\lambda_{\min}(\Lambda_h^j)\right)^{J_j} \left(1 + \left(1 - \alpha_{\boldsymbol{c}}^{i}\,\lambda_{\min}(\Lambda_h^i)\right)^{J_i}\right) \\
&\leq 2\,H\,\sqrt{d_{\boldsymbol{c}}\,|\mathcal{D}^t|} \sum_{i=1}^{t} \left(\prod_{j=i+1}^{t} \left(1 - \alpha_{\boldsymbol{c}}^{h,j}\,\lambda_{\min}(\Lambda_h^j)\right)^{J_j} + \prod_{j=i}^{t} \left(1 - \alpha_{\boldsymbol{c}}^{h,j}\,\lambda_{\min}(\Lambda_h^j)\right)^{J_j}\right) \\
&\overset{\text{(vii)}}{\leq} 2\,H\,\sqrt{d_{\boldsymbol{c}}\,|\mathcal{D}^t|} \sum_{i=1}^{t} \left(\prod_{j=i+1}^{t} \left(1 - 1/(2\,\kappa_j)\right)^{J_j} + \prod_{j=i}^{t} \left(1 - 1/(2\,\kappa_j)\right)^{J_j}\right),
\end{aligned}
$$

(i) uses the definition of the matrix norm (i.e., $\|A\|_2 \coloneqq \max_x \frac{\|A\,x\|_2}{\|x\|} \implies \|A\,x\|_2 \leq \|A\|_2\,\|x\|_2$). (ii) uses Lemma D.6 and sets $\lambda = 1$. (iii) and (v) come from the submultiplicativity of matrix norm. (iv) and (vi) use the fact that $\|A\|_2 \leq \lambda_{\max}(A)$, and (iv) also uses the triangular inequality. (vii) uses the fact that we set $\alpha_{\boldsymbol{c}}^{h,j} = 1/\left(2\,\lambda_{\max}(\Lambda_h^j)\right)$ and denotes that $\kappa_j = \max_{h\in[H]} \lambda_{\max}(\Lambda_h^j)/\lambda_{\min}(\Lambda_h^j)$.

Using Lemma D.16. we can set $J_j \geq 2\,\kappa_j \log(1/\sigma)$ where $\sigma = 1/\left(4\,H\,(|\mathcal{D}^t|+1)\,\sqrt{d_{\boldsymbol{c}}}\right)$. We can further get that

$$
\begin{aligned}
\|\mu_h^{t,J_t}\|_2 &\leq 2\,H\,\sqrt{d_{\boldsymbol{c}}\,|\mathcal{D}^t|} \sum_{i=1}^{t} \left(\sigma^{t-i} + \sigma^{t-i+1}\right) \\
&\leq 4\,H\,\sqrt{d_{\boldsymbol{c}}\,|\mathcal{D}^t|} \sum_{i=0}^{\infty} \sigma^i \\
&= 4\,H\,\sqrt{d_{\boldsymbol{c}}\,|\mathcal{D}^t|} \left(\frac{1}{1-\sigma}\right) \\
&= \frac{16}{3}\,H\,\sqrt{d_{\boldsymbol{c}}\,|\mathcal{D}^t|}\,.
\end{aligned}
$$

Next, we continue to bound Term (II). Since $\xi_h^{t,J_t} \sim \mathbf{N}(0, \Sigma_h^{t,J_t})$, using Lemma D.11, we have that

$$
\Pr\left(\left\|\xi_h^{t,J_t}\right\|_2 \leq \sqrt{\frac{1}{\delta}\,\mathrm{Tr}\left(\Sigma_h^{t,J_t}\right)}\right) \geq 1 - \delta\,.
$$

Recall from Lemma D.3 that

$$\Sigma_h^{t,J_t} = \sum_{i=1}^{t} \frac{1}{\zeta} A_t^{J_t} \dots A_{i+1}^{J_{i+1}} \left(I - A_i^{2J_i}\right) \left(\Lambda_h^i\right)^{-1} (I + A_i)^{-1} A_{i+1}^{J_{i+1}} \dots A_t^{J_t} .$$

Therefore, we can use Lemma D.13 and derive that

$$\begin{aligned}
\operatorname{Tr}\left(\Sigma_h^{t,J_t}\right) &= \sum_{i=1}^{t} \frac{1}{\zeta} \operatorname{Tr}\left(A_t^{J_t} \dots A_{i+1}^{J_{i+1}} \left(I - A_i^{2J_i}\right) \left(\Lambda_h^i\right)^{-1} (I + A_i)^{-1} A_{i+1}^{J_{i+1}} \dots A_t^{J_t}\right) \\
&\leq \sum_{i=1}^{t} \frac{1}{\zeta} \operatorname{Tr}\left(A_t^{J_t}\right) \dots \operatorname{Tr}\left(A_{i+1}^{J_{i+1}}\right) \operatorname{Tr}\left(I - A_i^{2J_i}\right) \operatorname{Tr}\left(\left(\Lambda_h^i\right)^{-1}\right) \operatorname{Tr}\left((I + A_i)^{-1}\right) \operatorname{Tr}\left(A_{i+1}^{J_{i+1}}\right) \dots \operatorname{Tr}\left(A_t^{J_t}\right) .
\end{aligned}$$

To bound each term, we first have,

$$\begin{aligned}
\operatorname{Tr}\left(A_i^{J_i}\right) &\leq \operatorname{Tr}\left(\left(1 - \alpha_{\boldsymbol{c}}^i \lambda_{\min}(\Lambda_h^i)\right)^{J_i} I\right) \\
&\leq d_{\boldsymbol{c}} \left(1 - \alpha_{\boldsymbol{c}}^i \lambda_{\min}(\Lambda_h^i)\right)^{J_i} \\
&\leq d_{\boldsymbol{c}}\, \sigma \leq 1 ,
\end{aligned}$$

where the first inequality follows from the fact that $A_i^{J_i} \prec (1 - \alpha_{\boldsymbol{c}}^t \lambda_{\min}(\Lambda_h^t))^{J_j} I$. Similarly, since we set $0 < \alpha_{\boldsymbol{c}}^{h,j} < 1/(2\,\lambda_{\max}(\Lambda_j))$, we have $A_i^{J_i} \succ \frac{1}{2^{J_i}} I$ and therefore,

$$\operatorname{Tr}\left(I - A_i^{2J_i}\right) \leq \left(1 - \frac{1}{2^{2J_i}}\right) d_{\boldsymbol{c}} < d_{\boldsymbol{c}} .$$

Similarly, since we set $0 < \alpha_{\boldsymbol{c}}^{h,j} < 1/(2\,\lambda_{\max}(\Lambda_j))$ and thus $I + A_i \succ \frac{3}{2} I$, we have that

$$\operatorname{Tr}\left((I + A_i)^{-1}\right) \leq \frac{2}{3}\, d_{\boldsymbol{c}} .$$

Additionally, since all eigenvalues of Λ_h^i are greater than or equal to 1,

$$\operatorname{Tr}\left((\Lambda_h^i)^{-1}\right) \leq d_{\boldsymbol{c}} \cdot 1 = d_{\boldsymbol{c}} .$$

Finally, we have that

$$\operatorname{Tr}\left(\Sigma_h^{t,J_t}\right) \leq \sum_{i=1}^{t} \frac{1}{\zeta} \cdot \frac{2}{3} \cdot d_{\boldsymbol{c}}^3 = \frac{2\, d_{\boldsymbol{c}}^3}{3\,\zeta}\, t .$$

Therefore, using Lemma D.11, we have that

$$\Pr\left(\left\|\xi_h^{t,J_t}\right\|_2 \leq \sqrt{\frac{1}{\delta} \cdot \frac{2\, d_{\boldsymbol{c}}^3}{3\,\zeta}\, T}\right) \geq \Pr\left(\left\|\xi_h^{t,J_t}\right\|_2 \leq \sqrt{\frac{1}{\delta} \operatorname{Tr}\left(\Sigma_h^{t,J_t}\right)}\right) \geq 1 - \delta .$$

Putting everything together, with probability at least $1 - \delta$, we can obtain that

$$\left\|w_h^{t,m,J_t}\right\|_2 \leq \overline{W}_\delta := \frac{16}{3}\, H \sqrt{d_{\boldsymbol{c}} \left|\mathcal{D}^t\right|} + \sqrt{\frac{2\, d_{\boldsymbol{c}}^3\, t}{3\,\zeta\,\delta}} .$$

This concludes the proof. □

D.2.6 Proof of Lemma D.8

Proof. To start, we decompose the LHS using the triangle inequality,

$$\left|\left\langle \phi(s,a), w_h^{t,m,J_t} - \widehat{w}_h^t \right\rangle\right| \le \underbrace{\left|\left\langle \phi(s,a), w_h^{t,m,J_t} - \mu_h^{t,J_t} \right\rangle\right|}_{\text{(I)}} + \underbrace{\left|\left\langle \phi(s,a), \mu_h^{t,J_t} - \widehat{w}_h^t \right\rangle\right|}_{\text{(II)}},$$

where μ_h^{t,J_t} is defined in Eq. (13). To bound Term (I), we first apply Hölder's inequality and obtain that

$$\left|\left\langle \phi(s,a), w_h^{t,m,J_t} - \mu_h^{t,J_t} \right\rangle\right| \le \left\|\phi(s,a)^\top \left(\Sigma_h^{t,J_t}\right)^{1/2}\right\|_2 \left\|\left(\Sigma_h^{t,J_t}\right)^{-1/2}\left(w_h^{t,m,J_t} - \mu_h^{t,J_t}\right)\right\|_2.$$

Since $w_h^{t,m,J_t} \sim \mathbf{N}(\mu_h^{t,J_t}, \Sigma_h^{t,J_t})$, we know that $\left(\Sigma_h^{t,J_t}\right)^{-1/2}\left(w_h^{t,m,J_t} - \mu_h^{t,J_t}\right) \sim \mathbf{N}(0, I_{d_{\boldsymbol{c}}\times d_{\boldsymbol{c}}})$. Therefore,

$$\Pr\left(\left\|\left(\Sigma_h^{t,J_t}\right)^{-1/2}\left(w_h^{t,m,J_t} - \mu_h^{t,J_t}\right)\right\|_2 \ge 2\sqrt{d_{\boldsymbol{c}}\,\log(1/\delta)}\right) \le \delta^2.$$

Then, we continue to bound $\left\|\phi(s,a)^\top \left(\Sigma_h^{t,J_t}\right)^{1/2}\right\|_2$.

$$\begin{aligned}
&\phi(s,a)^\top \Sigma_h^{t,J_t}\,\phi(s,a) \\
&\quad= \frac{1}{\zeta}\sum_{i=1}^{t} \phi(s,a)^\top A_t^{J_t}\ldots A_{i+1}^{J_{i+1}}\left(I - A^{2J_i}\right)\left(\Lambda_h^i\right)^{-1}\left(I + A_i\right)^{-1} A_{i+1}^{J_{i+1}}\ldots A_t^{J_t}\,\phi(s,a) \\
&\quad\overset{\text{(i)}}{=} \frac{1}{\zeta}\sum_{i=1}^{t} \phi(s,a)^\top \mathscr{A}_{i+1}\left(I - A^{2J_i}\right)\left(\Lambda_h^i\right)^{-1}\left(I + A_i\right)^{-1}\mathscr{A}_{i+1}^\top\,\phi(s,a) \\
&\quad\overset{\text{(ii)}}{\le} \frac{2}{3\zeta}\sum_{i=1}^{t} \phi(s,a)^\top \mathscr{A}_{i+1}\left(\left(\Lambda_h^i\right)^{-1} - A_i^{J_i}\left(\Lambda_h^i\right)^{-1} A_i^{J_i}\right)\mathscr{A}_{i+1}^\top\,\phi(s,a) \\
&\quad= \frac{2}{3\,\zeta}\left(\sum_{i=1}^{t} \phi(s,a)^\top \mathscr{A}_{i+1}\left(\Lambda_h^i\right)^{-1}\mathscr{A}_{i+1}^\top\,\phi(s,a) - \sum_{i=1}^{t} \phi(s,a)^\top \mathscr{A}_i\left(\Lambda_h^i\right)^{-1}\mathscr{A}_i^\top\,\phi(s,a)\right) \\
&\quad\overset{\text{(iii)}}{\le} \frac{2}{3\,\zeta}\sum_{i=1}^{t} \phi(s,a)^\top \mathscr{A}_{i+1}\left(\Lambda_h^i\right)^{-1}\mathscr{A}_{i+1}^\top\,\phi(s,a) \\
&\quad= \frac{2}{3\,\zeta}\left(\|\phi(s,a)\|^2_{(\Lambda_h^i)^{-1}} + \sum_{i=1}^{t-1}\left\|\mathscr{A}_{i+1}^\top\,\phi(s,a)\right\|^2_{(\Lambda_h^i)^{-1}}\right) \\
&\quad\le \frac{2}{3\,\zeta}\,\|\phi(s,a)\|^2_{(\Lambda_h^t)^{-1}} + \frac{2}{3\,\zeta}\sum_{i=1}^{t-1}\prod_{j=i+1}^{t}\left(1 - \alpha_{\boldsymbol{c}}\,\lambda_{\min}(\Lambda_h^j)\right)^{2\,J_j}\|\phi(s,a)\|^2_{(\Lambda_h^i)^{-1}}.
\end{aligned}$$

For (i), we use the denotation that $\mathscr{A}_{i+1} = A_t^{J_t}\ldots A_{i+1}^{J_{i+1}}$. (ii) follows from $I + A_i \succ \frac{3}{2}I$ since we set $\alpha_{\boldsymbol{c}}^{h,j} = 1/\left(2\,\lambda_{\max}(\Lambda_h^j)\right)$. (iii) follows from the fact that $\sum_{i=1}^{t}\phi(s,a)^\top \mathscr{A}_i\left(\Lambda_h^i\right)^{-1}\mathscr{A}_i^\top\,\phi(s,a) > 0$. Therefore,

$$\begin{aligned}
\left\|\phi(s,a)^\top\left(\Sigma_h^{t,J_t}\right)^{1/2}\right\|_2 &= \sqrt{\phi(s,a)^\top\,\Sigma_h^{t,J_t}\,\phi(s,a)} \\
&\overset{\text{(iv)}}{\le} \sqrt{\frac{2}{3\,\zeta}}\|\phi(s,a)\|_{(\Lambda_h^t)^{-1}} + \sqrt{\frac{2}{3\,\zeta}}\sum_{i=1}^{t-1}\prod_{j=i+1}^{t}\left(1 - \alpha_{\boldsymbol{c}}\,\lambda_{\min}(\Lambda_h^j)\right)^{J_j}\|\phi(s,a)\|_{(\Lambda_h^i)^{-1}} \\
&\overset{\text{(v)}}{\le} \sqrt{\frac{2}{3\,\zeta}}\|\phi(s,a)\|_{(\Lambda_h^t)^{-1}} + \sqrt{\frac{2}{3\,\zeta}}\sum_{i=1}^{t-1}\sigma^{t-i}\,\|\phi(s,a)\|_{(\Lambda_h^i)^{-1}} \\
&\overset{\text{(vi)}}{\le} \sqrt{\frac{2}{3\,\zeta}}\|\phi(s,a)\|_{(\Lambda_h^t)^{-1}} + \sqrt{\frac{2\,(|\mathcal{D}^t|+1)}{3\,\zeta}}\left(\sum_{i=1}^{t-1}\sigma^{t-i}\right)\|\phi(s,a)\|_{(\Lambda_h^t)^{-1}}
\end{aligned}$$

$$\overset{\text{(vii)}}{\leq} \sqrt{\frac{2}{3\,\zeta}}\|\phi(s,a)\|_{(\Lambda_h^t)^{-1}} + \sqrt{\frac{2\,(|\mathcal{D}^t|+1)}{3\,\zeta}}\left(\frac{\sigma}{1-\sigma}\right)\|\phi(s,a)\|_{(\Lambda_h^t)^{-1}}$$
$$\leq \sqrt{\frac{2}{3\,\zeta}}\|\phi(s,a)\|_{(\Lambda_h^t)^{-1}} + \frac{1}{4}\sqrt{\frac{2}{3\,\zeta}}\left(\frac{1}{1-\sigma}\right)\|\phi(s,a)\|_{(\Lambda_h^t)^{-1}}$$
$$\leq \left(\sqrt{\frac{2}{3\,\zeta}} + \frac{1}{3}\sqrt{\frac{2}{3\,\zeta}}\right)\|\phi(s,a)\|_{(\Lambda_h^t)^{-1}}$$
$$\leq \frac{4}{3}\sqrt{\frac{2}{3\,\zeta}}\,\|\phi(s,a)\|_{(\Lambda_h^t)^{-1}}\,.$$

(iv) follows from the fact that $\sqrt{a+b} \leq a+b$ for all $a,b>0$. (v) uses Lemma D.16 by setting $J_j \geq \kappa_j \log(1/\sigma)$ where $\sigma = 1/\left(4\,H\,(|\mathcal{D}^t|+1)\,\sqrt{d_{\boldsymbol{c}}}\right)$. (vi) follows from $\|\phi(s,a)\|_{(\Lambda_h^i)^{-1}} \leq \|\phi(s,a)\|_2 \leq \sqrt{|\mathcal{D}^t|+1}\,\|\phi(s,a)\|_{(\Lambda_h^t)^{-1}}$. (vii) follows from $\sum_{i=1}^t \sigma^{t-i} \leq \sum_{i=1}^\infty \sigma^i \leq \sigma/(1-\sigma)$. Therefore, we have

$$\Pr\left(\left|\left\langle \phi(s,a), w_h^{t,m,J_t} - \mu_h^{t,J_t}\right\rangle\right| \geq \frac{8}{3}\sqrt{\frac{2\,d_{\boldsymbol{c}}\,\log(1/\delta)}{3\,\zeta}}\|\phi(s,a)\|_{(\Lambda_h^t)^{-1}}\right)$$
$$\leq \Pr\left(\left\|\phi(s,a)^\top\left(\Sigma_h^{t,J_t}\right)^{1/2}\right\|_2 \left\|\left(\Sigma_h^{t,J_t}\right)^{-1/2}\left(w_h^{t,m,J_t} - \mu_h^{t,J_t}\right)\right\|_2 \geq 2\,\sqrt{d_{\boldsymbol{c}}\,\log(1/\delta)}\left\|\phi(s,a)^\top\left(\Sigma_h^{t,J_t}\right)^{1/2}\right\|_2\right)$$
$$= \Pr\left(\left\|\left(\Sigma_h^{t,J_t}\right)^{-1/2}\left(w_h^{t,m,J_t} - \mu_h^{t,J_t}\right)\right\|_2 \geq 2\,\sqrt{d_{\boldsymbol{c}}\,\log(1/\delta)}\right) = \delta^2 \leq \delta\,.$$

This implies that

$$\Pr\left(\left|\left\langle \phi(s,a), w_h^{t,m,J_t} - \mu_h^{t,J_t}\right\rangle\right| \leq \frac{8}{3}\sqrt{\frac{2\,d_{\boldsymbol{c}}\,\log(1/\delta)}{3\,\zeta}}\|\phi(s,a)\|_{(\Lambda_h^t)^{-1}}\right) \geq 1-\delta\,.$$

Putting everything together, with probability at least $1-\delta$,

$$\left|\left\langle \phi(s,a), w_h^{t,m,J_t} - \widehat{w}_h^t\right\rangle\right| \leq \left|\left\langle \phi(s,a), w_h^{t,m,J_t} - \mu_h^{t,J_t}\right\rangle\right| + \left|\left\langle \phi(s,a), \mu_h^{t,J_t} - \widehat{w}_h^t\right\rangle\right|$$
$$\leq \left(\frac{8}{3}\sqrt{\frac{2\,d_{\boldsymbol{c}}\,\log(1/\delta)}{3\,\zeta}} + \frac{4}{3}\right)\|\phi(s,a)\|_{(\Lambda_h^t)^{-1}}\,.$$

□

D.2.7 Proof of Lemma D.9

Proof. Recall that $\mathbb{P}_h V(s,a) \coloneqq \mathbb{E}_{s'\sim\mathbb{P}_h(\cdot|s,a)}V(s')$ and $\mathbb{P}_h(\cdot \mid s,a) = \langle \phi(s,a), \psi_h(\cdot)\rangle$ due to the linear MDP assumption (Definition 2.1). We also denote that $\widehat{\Psi}_h^t \coloneqq \left\langle \psi_h, \widehat{V}_{h+1}^t\right\rangle_{\mathcal{S}}$ and thus $\mathbb{P}_h\widehat{V}_{h+1}^t(s,a) = \left\langle \phi(s,a), \widehat{\Psi}_h^t\right\rangle$. Then, we have that

$$\mathbb{P}_h\widehat{V}_{h+1}^t(s,a) = \left\langle \phi(s,a), \widehat{\Psi}_h^t\right\rangle$$
$$= \phi(s,a)^\top\left(\Lambda_h^t\right)^{-1}\Lambda_h^t\,\widehat{\Psi}_h^t$$
$$= \phi(s,a)^\top\left(\Lambda_h^t\right)^{-1}\left(\sum_{(s,a,s')\in\mathcal{D}_h^t}\phi(s,a)\,\phi(s,a)^\top + \lambda I\right)\widehat{\Psi}_h^t$$
$$= \phi(s,a)^\top\left(\Lambda_h^t\right)^{-1}\left(\sum_{(s,a,s')\in\mathcal{D}_h^t}\phi(s,a)(\mathbb{P}_h\widehat{V}_{h+1}^t)(s,a) + \lambda\,\widehat{\Psi}_h^t\right).$$

This further implies that

$$\left\langle \phi(s,a), \widehat{w}_h^t\right\rangle - r_h(s,a) - \mathbb{P}_h\widehat{V}_{h+1}^t(s,a)$$

$$
\begin{aligned}
&= \phi(s,a)^\top \big(\Lambda_h^t\big)^{-1} \sum_{(s,a,s')\in\mathcal{D}_h^t} \Big[r_h(s,a) + \widehat{V}_{h+1}^t(s')\Big] \cdot \phi(s,a) - r_h(s,a) \\
&\quad - \phi(s,a)^\top \big(\Lambda_h^t\big)^{-1} \left(\sum_{(s,a,s')\in\mathcal{D}_h^t} \phi(s,a)(\mathbb{P}_h \widehat{V}_{h+1}^t)(s,a) + \lambda\, \widehat{\Psi}_h^t \right) \\
&= \underbrace{\phi(s,a)^\top (\Lambda_h^t)^{-1} \left(\sum_{(s,a,s')\in\mathcal{D}_h^t} \phi(s,a) \left[\widehat{V}_{h+1}^t(s') - \mathbb{P}_h \widehat{V}_{h+1}^t(s,a)\right] \right)}_{\text{(I)}} \\
&\quad \underbrace{+ \phi(s,a)^\top (\Lambda_h^t)^{-1} \left(\sum_{(s,a,s')\in\mathcal{D}_h^t} r_h(s,a)\, \phi(s,a) \right) - r_h(s,a)}_{\text{(II)}} \underbrace{- \lambda\, \phi(s,a)^\top (\Lambda_h^t)^{-1}\, \widehat{\Psi}_h^t}_{\text{(III)}} .
\end{aligned}
$$

We first start by bounding Term (I). With probability at least $1-\delta$, it holds that

$$
\begin{aligned}
&\phi(s,a)^\top (\Lambda_h^t)^{-1} \left(\sum_{(s,a,s')\in\mathcal{D}_h^t} \phi(s,a) \Big[\widehat{V}_{h+1}^t(s') - \mathbb{P}_h \widehat{V}_{h+1}^t(s,a)\Big] \right) \\
&\quad \overset{\text{(i)}}{\leq} \left\| \sum_{(s,a,s')\in\mathcal{D}_h^t} \phi(s,a) \Big[\widehat{V}_{h+1}^t(s') - \mathbb{P}_h \widehat{V}_{h+1}^t(s,a)\Big] \right\|_{(\Lambda_h^t)^{-1}} \|\phi(s,a)\|_{(\Lambda_h^t)^{-1}} \\
&\quad \overset{\text{(ii)}}{\leq} C_\delta\, H \sqrt{d_{\boldsymbol{c}}}\, \|\phi(s,a)\|_{(\Lambda_h^t)^{-1}} ,
\end{aligned}
$$

where (i) follows from the Cauchy-Schwarz inequality, and (ii) follows from the good event defined in Lemma D.1. Next, we continue to bound Term (II). We observe that

$$
\begin{aligned}
&\phi(s,a)^\top (\Lambda_h^t)^{-1} \left(\sum_{(s,a,s')\in\mathcal{D}_h^t} r_h(s,a)\, \phi(s,a) \right) - r_h(s,a) \\
&\quad \overset{\text{(iii)}}{=} \phi(s,a)^\top (\Lambda_h^t)^{-1} \left(\sum_{(s,a,s')\in\mathcal{D}_h^t} r_h(s,a)\, \phi(s,a) \right) - \phi(s,a)^\top \theta_h \\
&\quad = \phi(s,a)^\top (\Lambda_h^t)^{-1} \left(\sum_{(s,a,s')\in\mathcal{D}_h^t} r_h(s,a)\, \phi(s,a) - \Lambda_h^t \theta_h \right) \\
&\quad = \phi(s,a)^\top (\Lambda_h^t)^{-1} \left(\sum_{(s,a,s')\in\mathcal{D}_h^t} r_h(s,a)\, \phi(s,a) - \sum_{(s,a,s')\in\mathcal{D}_h^t} \phi(s,a)\, \phi(s,a)^\top \theta_h - \lambda\, \theta_h \right) \\
&\quad \overset{\text{(iv)}}{=} \phi(s,a)^\top (\Lambda_h^t)^{-1} \left(\sum_{(s,a,s')\in\mathcal{D}_h^t} r_h(s,a)\, \phi(s,a) - \sum_{(s,a,s')\in\mathcal{D}_h^t} \phi(s,a)\, r_h(s,a) - \lambda\, \theta_h \right) \\
&\quad = -\lambda\, \phi(s,a)^\top (\Lambda_h^t)^{-1}\, \theta_h \\
&\quad \overset{\text{(v)}}{\leq} \lambda\, \|\phi(s,a)\|_{(\Lambda_h^t)^{-1}} \|\theta_h\|_{(\Lambda_h^t)^{-1}} \\
&\quad \overset{\text{(vi)}}{\leq} \sqrt{\lambda\, d_{\boldsymbol{c}}}\, \|\phi(s,a)\|_{(\Lambda_h^t)^{-1}} .
\end{aligned}
$$

(iii) and (iv) follow from the definition $r_h(s,a) = \langle \phi(s,a), \theta_h \rangle$. (v) applies the Cauchy-Schwarz inequality. (vi) follows from $\|\phi(s,a)\|_{(\Lambda_h^t)^{-1}} \leq \sqrt{1/\lambda}\, \|\phi(s,a)\|_2$ and $\|\theta_h\|_2 \leq \sqrt{d_{\boldsymbol{c}}}$ (Definition 2.1).

Lastly, we derive the bound for Term (III).

$$\begin{aligned}
\lambda\,\phi(s,a)^\top\,(\Lambda_h^t)^{-1}\,\widehat{\Psi}_h^t &\overset{\text{(vii)}}{\leq} \lambda\,\|\phi(s,a)\|_{(\Lambda_h^t)^{-1}}\,\left\|\widehat{\Psi}_h^t\right\|_{(\Lambda_h^t)^{-1}} \\
&\overset{\text{(viii)}}{\leq} \sqrt{\lambda}\,\|\phi(s,a)\|_{(\Lambda_h^t)^{-1}}\,\left\|\widehat{\Psi}_h^t\right\|_2 \\
&\leq \sqrt{\lambda}\,\|\phi(s,a)\|_{(\Lambda_h^t)^{-1}}\left\|\left\langle \psi_h, \widehat{V}_{h+1}^t\right\rangle_{\mathcal{S}}\right\|_2 \\
&= H\,\sqrt{\lambda}\,\|\phi(s,a)\|_{(\Lambda_h^t)^{-1}}\left\|\int_{s\in\mathcal{S}} \psi_h(s)\Big(\widehat{V}_{h+1}^t(s)/H\Big)\,\mathrm{d}\,s\right\|_2 \\
&\overset{\text{(xiv)}}{\leq} H\,\sqrt{\lambda\,d_{\boldsymbol{c}}}\,\|\phi(s,a)\|_{(\Lambda_h^t)^{-1}}\,.
\end{aligned}$$

(vii) applies the Cauchy-Schwarz inequality. (viii) follows from $\left\|\widehat{\Psi}_h^t\right\|_{(\Lambda_h^t)^{-1}} \leq \sqrt{\lambda}\left\|\widehat{\Psi}_h^t\right\|_2$. (xiv) comes from the assumption that $\left\|\int_{s\in\mathcal{S}} \psi_h(s)\Big(\widehat{V}_{h+1}^t(s)/H\Big)\,\mathrm{d}\,s\right\|_2 \leq \sqrt{d_{\boldsymbol{c}}}$ (Definition 2.1).
Putting everything together and setting $\lambda = 1$, we have with probability at least $1-\delta$,

$$\begin{aligned}
&\left|\left\langle \phi(s,a), \widehat{w}_h^t\right\rangle - r_h(s,a) - \mathbb{P}_h\widehat{V}_{h+1}^t(s,a)\right| \\
&\quad\leq \left(C_\delta\,H\,\sqrt{d_{\boldsymbol{c}}} + \sqrt{\lambda\,d_{\boldsymbol{c}}} + H\,\sqrt{\lambda\,d_{\boldsymbol{c}}}\right)\|\phi(s,a)\|_{(\Lambda_h^t)^{-1}} \\
&\quad= 3\,C_\delta\,H\,\sqrt{d_{\boldsymbol{c}}}\,\|\phi(s,a)\|_{(\Lambda_h^t)^{-1}}\,.
\end{aligned}$$

This concludes the proof. □

D.3 Technical Tools

Lemma D.10 (Jin et al. 2020, Lemma D.1). *Let $\Lambda = \lambda I + \sum_{i=1}^t \phi_i\phi_i^\top$, where $\phi_i \in \mathbb{R}^d$ and $\lambda > 0$. Then,*

$$\sum_{i=1}^t \phi_i^\top(\Lambda)^{-1}\phi_i \leq d\,.$$

Lemma D.11 (Ishfaq et al. 2024a, Lemma E.1). *Given a multivariate normal distribution $X \sim \mathbf{N}(0, \Sigma_{d\times d})$, for any $\delta \in (0,1]$, it hold that*

$$\Pr\left(\|X\|_2 \leq \sqrt{\frac{1}{\delta}\operatorname{Tr}(\Sigma)}\right) \geq 1-\delta\,.$$

Lemma D.12 (Abramowitz and Stegun 1948). *Suppose X is a Gaussian random variable $X \sim \mathbf{N}(\mu, \sigma^2)$, where $\sigma > 0$. For $z \in [0,1]$, it holds that*

$$\Pr(X > \mu + z\,\sigma) \geq \frac{e^{-z^2/2}}{\sqrt{8\pi}} \quad \textit{and} \quad \Pr(X < \mu - z\,\sigma) \geq \frac{e^{-z^2/2}}{\sqrt{8\pi}}\,.$$

Additionally, for any $z \geq 1$,

$$\frac{e^{-z^2/2}}{2\,z\,\sqrt{\pi}} \leq \Pr(|X-\mu| > z\,\sigma) \leq \frac{e^{-z^2/2}}{z\,\sqrt{\pi}}\,.$$

Lemma D.13. *If A and B are positive semi-definite square matrices of the same size, then*

$$[\operatorname{Tr}(AB)]^2 \leq \operatorname{Tr}(A^2)\operatorname{Tr}(B^2) \leq [\operatorname{Tr}(A)]^2[\operatorname{Tr}(B)]^2\,.$$

Lemma D.14. *Given two symmetric positive semi-definite square matrices A and B such that $A \succeq B$, it holds that $\|A\|_2 \geq \|B\|_2$.*

Proof. Note that $A - B$ is also positive semi-definite. Then, we have that

$$\|B\|_2 = \sup_{\|x\|=1} x^\top B x \leq \sup_{\|x\|=1} \left(x^\top B x + x^\top (A-B)x\right) = \sup_{\|x\|=1} x^\top A x = \|A\|_2 .$$

□

Lemma D.15. *Let $A \in \mathbb{R}^{d\times d}$ be a positive definite matrix where its largest eigenvalue $\lambda_{\max}(A) \leq \lambda$. Given that $v_1, \ldots, v_n$ are n vectors in $\mathbb{R}^d$, it holds that*

$$\left\| A \sum_{i=1}^n v_i \right\| \leq \sqrt{\lambda\, n \sum_{i=1}^n \|v_i\|_A^2} .$$

Lemma D.16. *Let Λ be a positive definite matrix and $\kappa = \frac{\lambda_{\max}(\Lambda)}{\lambda_{\min}(\Lambda)}$ be the condition number of Λ. If $\Lambda \succ I$ and $J \geq 2\,\kappa \log(1/\sigma)$, then, for any $\sigma > 0$,*

$$(1 - 1/(2\,\kappa))^J < \sigma .$$

Proof. The statement is equivalent to proving that

$$J \geq \frac{\log(1/\sigma)}{\log\left(\frac{1}{1-1/(2\,\kappa)}\right)} .$$

Since $\kappa \geq 1$ and for any $x \in (0,1)$, $e^{-x} > 1 - x$, we have that

$$e^{-1/(2\,\kappa)} > 1 - 1/(2\,\kappa) \implies \log\left(\frac{1}{1-1/(2\,\kappa)}\right) \geq \frac{1}{2\,\kappa} .$$

Therefore, we have that

$$J \geq 2\,\kappa \log(1/\sigma) \geq \frac{\log(1/\sigma)}{\log\left(\frac{1}{1-1/(2\,\kappa)}\right)} .$$

This concludes the proof. □

E SAMPLE COMPLEXITY IN THE ON-POLICY SETTING

E.1 Proof of Good Event

Lemma E.1. *Consider Algorithm 1 in the on-policy setting with $\lambda = 1$. Then, for any $\delta \in (0,1)$, with probability at least $1-\delta$, it holds that*

$$\left\| \sum_{(s,a,s')\in\mathcal{D}_h^t} \phi(s,a)\left[\widehat{V}_{h+1}^t(s') - \mathbb{P}_h \widehat{V}_{h+1}^t(s,a)\right] \right\|_{(\Lambda_h^t)^{-1}} \leq C_\delta^{\mathrm{on}}\, H \sqrt{d_{\mathbf{c}}} ,$$

where $C_\delta^{\mathrm{on}} = \log(N/\delta)$.

Proof of Lemma E.1. Recall that $\mathbb{P}_h \widehat{V}_{h+1}^t(s,a) = \mathbb{E}_{s'\sim\mathbb{P}_h}\left[\widehat{V}_{h+1}^t(s')\right]$. Thus, $\mathbb{E}[\widehat{V}_{h+1}^t(s') - \mathbb{P}_h \widehat{V}_{h+1}^t(s,a)] = 0$. Also, $\left|\widehat{V}_{h+1}^t(s') - \mathbb{P}_h \widehat{V}_{h+1}^t(s,a)\right| \leq H$. Therefore, $\widehat{V}_{h+1}^t(s') - \mathbb{P}_h \widehat{V}_{h+1}^t(s,a)$ is zero-mean and H-sub Gaussian. Given that, we can invoke Lemma E.3.

$$\left\| \sum_{(s,a,s')\in\mathcal{D}_h^t} \phi(s,a)\left[\widehat{V}_{h+1}^t(s') - \mathbb{P}_h \widehat{V}_{h+1}^t(s,a)\right] \right\|_{(\Lambda_h^t)^{-1}}$$

$$
\begin{aligned}
&\leq \sqrt{2}\, H \sqrt{\log\left[\frac{\det(\Lambda_h^t)^{1/2}\det(\Lambda_h^0)^{-1/2}}{\delta}\right]} \\
&= \sqrt{2}\, H \sqrt{\log\left[\left(\frac{N+\lambda}{\lambda}\right)^{d/2}\right] - \log(\delta)} \\
&= \sqrt{2}\, H \sqrt{\frac{d_{\boldsymbol{c}}}{2}\log(N/\delta)} \\
&= H\sqrt{d_{\boldsymbol{c}}\,\log(N/\delta)}\,,
\end{aligned}
$$

where the first equality follows from Lemma E.4, and the second equality holds by setting $\lambda = 1$. This concludes the proof. □

E.2 Proof of Theorem 6.1

Using Lemma E.1, we can instantiate Lemma D.2 in the on-policy setting with

$$
\begin{aligned}
\Gamma_{\mathsf{LMC}}^{\mathrm{on}} &= C_\delta^{\mathrm{on}}\, H\,\sqrt{d_{\boldsymbol{c}}} + \frac{4}{3}\sqrt{\frac{2\,d_{\boldsymbol{c}}\log(1/\delta)}{3\,\zeta}} + \frac{4}{3} \\
&= H\sqrt{d_{\boldsymbol{c}}\,\log(N/\delta)} + \frac{4}{3}\sqrt{\frac{2\,d_{\boldsymbol{c}}\log(1/\delta)}{3\,\zeta}} + \frac{4}{3}\,.
\end{aligned}
$$

Proof of Theorem 6.1. The optimal gap for the mixture policy can be written as

$$
\mathbb{E}\Big[V_1^\star(s_1) - V_1^{\overline{\pi}^T}(s_1)\Big] = \frac{1}{T}\sum_{t=1}^{T}\Big(V_1^\star(s_1) - V_1^{\pi^t}(s_1)\Big)\,.
$$

Then, to decompose the above summation, we have that

$$
\sum_{t=1}^{T}\Big(V_1^\star(s_1) - V_1^{\pi^t}(s_1)\Big) = \sum_{t=1}^{T}\Big(V_1^\star(s_1) - \widehat{V}_1^t(s_1)\Big) + \sum_{t=1}^{T}\Big(\widehat{V}_1^t(s_1) - V_1^{\pi^t}(s_1)\Big)\,.
$$

We can further decompose the first term by invoking Lemma E.2 with $\pi = \pi^\star$ and obtain that

$$
V_1^\star(s_1) - \widehat{V}_1^t(s_1) = \sum_{h=1}^{H}\mathbb{E}_{\pi^\star}\Big[\Big\langle \pi_h^\star(\cdot\mid s) - \pi_h^t(\cdot\mid s), \widehat{Q}_h^t(s,\cdot)\Big\rangle\Big] + \sum_{h=1}^{H}\mathbb{E}_{\pi^\star}\Big[r_h(s,a) + \mathbb{P}_h\widehat{V}_{h+1}^t(s,a) - \widehat{Q}_h^t(s,a)\Big]\,.
$$

Similarly, we can decompose the second term by invoking Lemma E.2 with $\pi = \pi^t$ and get that

$$
\begin{aligned}
\widehat{V}_1^t(s_1) - V_1^{\pi^t}(s_1) &= \sum_{h=1}^{H}\mathbb{E}_{\pi^t}\Big[\Big\langle \pi_h^t(\cdot\mid s) - \pi_h^t(\cdot\mid s), \widehat{Q}_h^t(s,\cdot)\Big\rangle\Big] - \sum_{h=1}^{H}\mathbb{E}_{\pi^t}\Big[r_h(s,a) + \mathbb{P}_h\widehat{V}_{h+1}^t(s,a) - \widehat{Q}_h^t(s,a)\Big] \\
&= -\sum_{h=1}^{H}\mathbb{E}_{\pi^t}\Big[r_h(s,a) + \mathbb{P}_h\widehat{V}_{h+1}^t(s,a) - \widehat{Q}_h^t(s,a)\Big]\,.
\end{aligned}
$$

Therefore, using the definition of the model prediction error ι in Definition 5.1, we have

$$
\begin{aligned}
&\sum_{t=1}^{T}\Big(V_1^\star(s_1) - V_1^{\pi^t}(s_1)\Big) \\
&\quad= \underbrace{\sum_{t=1}^{T}\sum_{h=1}^{H}\mathbb{E}_{\pi^\star}\Big[\Big\langle \pi_h^\star(\cdot\mid s) - \pi_h^t(\cdot\mid s), \widehat{Q}_h^t(s,\cdot)\Big\rangle\Big]}_{\text{(I) policy optimization (actor) error}} + \underbrace{\sum_{t=1}^{T}\sum_{h=1}^{H}\big(\mathbb{E}_{\pi^\star}[\iota_h^t(s,a)] - \mathbb{E}_{\pi^t}[\iota_h^t(s,a)]\big)}_{\text{(II) policy evaluation (critic) error}}\,.
\end{aligned}
$$

Policy optimization error. We first start by bounding Term (I), the policy optimization (actor) error.

$$\begin{aligned}
\text{Term (I)} &= \sum_{t=1}^{T}\sum_{h=1}^{H}\mathbb{E}_{s\sim\pi^\star}\left[\left\langle \pi_h^\star(\cdot\mid s)-\pi_h^t(\cdot\mid s),\widehat{Q}_h^t(s,\cdot)\right\rangle\right]\\
&= \sum_{h=1}^{H}\mathbb{E}_{s\sim\pi^\star}\left(\sum_{t=1}^{T}\left\langle \pi_h^\star(\cdot\mid s)-\pi_h^t(\cdot\mid s),\widehat{Q}_h^t(s,\cdot)\right\rangle\right)\\
&\le H\max_{(h,s)\in[H]\times\mathcal{S}}\left(\sum_{t=1}^{T}\left\langle \pi_h^\star(\cdot\mid s)-\pi_h^t(\cdot\mid s),\widehat{Q}_h^t(s,\cdot)\right\rangle\right)\\
&\overset{\text{(i)}}{\le} H\left(\frac{\log|\mathcal{A}|+\sum_{t=1}^{T}\|\epsilon_h^t(\cdot)\|_\infty}{\eta}+\frac{\eta\,H^2\,T}{2}\right)\\
&\overset{\text{(ii)}}{\le} H^2\sqrt{(\log|\mathcal{A}|+\bar{\epsilon}\,T)/2}\,\sqrt{T}\\
&\overset{\text{(iii)}}{\le} \mathcal{O}\Big(H^2\sqrt{\log|\mathcal{A}|}\,\sqrt{T}+H^2\sqrt{\bar{\epsilon}}\,T\Big).
\end{aligned}$$

(i) follows from Theorem 4.1 with $u=\pi_h^*(\cdot\mid s)$. (ii) is obtained by setting $\eta=\frac{\sqrt{2\,(\log|\mathcal{A}|+\bar{\epsilon}\,T)}}{H\sqrt{T}}$. (iii) is based on that for all $a,b\ge 0$, $\sqrt{a+b}\le\sqrt{a}+\sqrt{b}$.

Policy evaluation error. Then, we continue to bound Term (II), the policy evaluation (critic) error.

$$\begin{aligned}
\text{Term (II)} &= \sum_{t=1}^{T}\sum_{h=1}^{H}\left(\mathbb{E}_{\pi^\star}[\iota_h^t(s,a)]-\mathbb{E}_{\pi^t}[\iota_h^t(s,a)]\right)\\
&\overset{\text{(iv)}}{\le} -\sum_{t=1}^{T}\sum_{h=1}^{H}\mathbb{E}_{\pi^t}[\iota_h^t(s,a)]\\
&\overset{\text{(v)}}{\le} \Gamma^{\text{on}}_{\texttt{LMC}}\sum_{t=1}^{T}\sum_{h=1}^{H}\mathbb{E}_{\pi^t}\left[\|\phi(s,a)\|_{(\Lambda_h^t)^{-1}}\right]\\
&\le \Gamma^{\text{on}}_{\texttt{LMC}}\,T\max_{t\in[T]}\sum_{h=1}^{H}\mathbb{E}_{\pi^t}\left[\|\phi(s,a)\|_{(\Lambda_h^t)^{-1}}\right].
\end{aligned}$$

(iv) and (v) both follow from Lemma D.2, where (iv) is based on the optimism guarantee (RHS of Eq. (10)), while (v) is based on the error bound (LHS of Eq. (10)).

Bounding the sum of bonuses. Since $\Gamma^{\text{on}}_{\texttt{LMC}}$ is bounded, it suffices to bound $\mathbb{E}_{\pi^t}\left[\|\phi(s,a)\|_{(\Lambda_h^t)^{-1}}\right]$. Note that $\Lambda_h^t=\sum_{(s,a,s')\in\mathcal{D}_h^t}\phi(s,a)\,\phi(s,a)^\top+\lambda I$, and $\mathcal{D}_h^t$ only depends on π^t in the on-policy setting. (This is not true for the off-policy setting since Λ_h^t would depend on $\{\pi^1,\dots,\pi^t\}$.) We then index each data point in $\mathcal{D}_h^t$ as $\big\{(s_h^i,a_h^i,s_{h+1}^i)\big\}_{i\in[N]}$. Let $\Lambda_h^{t,i}=\Big(\sum_{j=1}^{i}\phi(s_h^j,a_h^j)\,\phi(s_h^j,a_h^j)^\top+\lambda I\Big)$. Then, we have that

$$\begin{aligned}
&\sum_{h=1}^{H}\mathbb{E}_{\pi^t}\left[\|\phi(s,a)\|_{(\Lambda_h^t)^{-1}}\right]\\
&\quad\overset{\text{(vi)}}{\le}\frac{1}{N}\sum_{i=1}^{N}\sum_{h=1}^{H}\mathbb{E}_{\pi^t}\left[\|\phi(s,a)\|_{\left(\Lambda_h^{t,i}\right)^{-1}}\right]\\
&\quad=\frac{1}{N}\sum_{i=1}^{N}\sum_{h=1}^{H}\|\phi(s_h^i,a_h^i)\|_{\left(\Lambda_h^{t,i}\right)^{-1}}+\frac{1}{N}\sum_{i=1}^{N}\sum_{h=1}^{H}\underbrace{\mathbb{E}_{\substack{s_h\sim\mathbb{P}(\cdot\mid s_{h-1}^t a_{h-1}^t)\\ a_h\sim\pi_h^t(\cdot\mid s_h)}}\left[\|\phi(s,a)\|_{\left(\Lambda_h^{t,i}\right)^{-1}}\right]-\|\phi(s_h^i,a_h^i)\|_{\left(\Lambda_h^{t,i}\right)^{-1}}}_{:=\mathcal{M}_{i,h}^{\text{on}}},
\end{aligned}\tag{15}$$

where (vi) follows from the fact that $\Lambda_h^{t,i}\preceq\Lambda_h^t$.

Applying the elliptical potential lemma. For the first term of Eq. (15), we have that

$$\begin{aligned}
&\frac{1}{N}\sum_{i=1}^{N}\sum_{h=1}^{H}\left\|\phi(s_h^i,a_h^i)\right\|_{(\Lambda_h^{t,i})^{-1}} \\
&\quad=\frac{1}{N}\sum_{h=1}^{H}\sum_{i=1}^{N}\left\|\phi(s_h^i,a_h^i)\right\|_{(\Lambda_h^{t,i})^{-1}} \\
&\quad\overset{\text{(vii)}}{\leq}\frac{1}{N}\sum_{h=1}^{H}\sqrt{N}\left(\sum_{i=1}^{N}\left\|\phi(s_h^i,a_h^i)\right\|^2_{(\Lambda_h^{t,i})^{-1}}\right)^{1/2} \\
&\quad\overset{\text{(viii)}}{\leq}\mathcal{O}\left(\sqrt{\frac{d_{\mathbf{c}}\,H^2\log(N/\delta)}{N}}\right).
\end{aligned}$$

(vii) applies the Cauchy-Schwarz inequality, and (viii) follows the elliptical argument from Lemma E.5.

A martingale difference sequence. For the second term of Eq. (15), since for a fixed $i\in[N]$, $\left\{\mathcal{M}_{i,h}^{\text{on}}\right\}_{h\in[H]}$ forms a martingale sequence adapted to the filtration,

$$\mathcal{F}_{i,h}^{\text{on}}=\left\{(s_\tau^i,a_\tau^i)\right\}_{\tau\in[h-1]},$$

such that $\mathbb{E}\left[\mathcal{M}_{i,h}^{\text{on}}\mid\mathcal{F}_{i,h}^{\text{on}}\right]=0$, where the expectation is with respect to the randomness in the policy and the environment at step h. Since $|\mathcal{M}_{i,h}^{\text{on}}|\leq 1$, we can apply the Azuma–Hoeffding inequality and obtain that

$$\Pr\left(\sum_{i=1}^{N}\sum_{h=1}^{H}\mathcal{M}_{i,h}^{\text{on}}\geq m\right)\geq\exp\left(\frac{-m^2}{2\,H\,N}\right).$$

Setting $m=\sqrt{2\,H\,N\,\log(1/\delta)}$ and using a union bound over $i\in[N]$, with probability at least $1-\delta$, it holds that

$$\frac{1}{N}\sum_{i=1}^{N}\sum_{h=1}^{H}\mathcal{M}_{i,h}^{\text{on}}\leq\sqrt{\frac{2\,H\log(1/\delta)}{N}}\leq\mathcal{O}\left(\sqrt{\frac{H\,\log(1/\delta)}{N}}\right).$$

Putting everything together. Therefore, we have that

$$\sum_{h=1}^{H}\mathbb{E}_{\pi^t}\left[\left\|\phi(s,a)\right\|_{(\Lambda_h^t)^{-1}}\right]=\frac{1}{N}\sum_{i=1}^{N}\sum_{h=1}^{H}\left\|\phi(s_h^i,a_h^i)\right\|_{(\Lambda_h^{t,i})^{-1}}+\frac{1}{N}\sum_{i=1}^{N}\sum_{h=1}^{H}\mathcal{M}_{i,h}^{\text{on}}\leq\mathcal{O}\left(\sqrt{\frac{d_{\mathbf{c}}\,H^2\log(N/\delta)}{N}}\right).$$

It further implies that, with probability at least $1-\delta$,

$$\begin{aligned}
\text{Term (II)}&\leq\Gamma_{\texttt{LMC}}^{\text{on}}\,T\max_{t\in[T]}\sum_{h=1}^{H}\mathbb{E}_{\pi^t}\left[\left\|\phi(s,a)\right\|_{(\Lambda_h^t)^{-1}}\right] \\
&\overset{\text{(ix)}}{\leq}\mathcal{O}\left(\sqrt{\frac{d_{\mathbf{c}}^3\,H^4\,\log^2(N/\delta)}{N}}\,T\right) \\
&\leq\widetilde{\mathcal{O}}\left(H^2\,\sqrt{d_{\mathbf{c}}^3\,\log|\mathcal{A}|}\,\sqrt{T}\right),
\end{aligned}$$

where (ix) comes from setting $N=\frac{T}{\log|\mathcal{A}|}=\frac{H^4}{\epsilon^2}$.

Finally, putting everything together, with probability at least $1-\delta$,

$$\mathbb{E}\left[V_1^\star(s_1)-V_1^{\overline{\pi}^T}(s_1)\right]=\frac{1}{T}(\text{Term (I)}+\text{Term (II)})=\widetilde{\mathcal{O}}\left(\frac{H^2\sqrt{d_{\mathbf{c}}^3\,\log|\mathcal{A}|}}{\sqrt{T}}+H^2\,\sqrt{\epsilon}\right).$$

This concludes the proof. □

E.3 Technical Tools

Lemma E.2 (Extended Value Difference). *Given any $\pi, \pi' \in \Delta(\mathcal{A} \mid \mathcal{S}, H)$ and any Q-function $\widehat{Q} \in \mathbb{R}^{H \times |\mathcal{S}| \times |\mathcal{A}|}$, we define $\widehat{V}_h(\cdot) = \mathbb{E}_{a \sim \pi'_h(s,\cdot)} \widehat{Q}_h(\cdot, a)$ for any $h \in [H]$. Then,*

$$\begin{aligned}
&\widehat{V}_1(s_1) - V_1^{\pi}(s_1) \\
&\quad = \sum_{h=1}^{H} \mathbb{E}_{s \sim \pi} \Big[\Big\langle \pi'_h(s,\cdot) - \pi_h(s,\cdot), \widehat{Q}_h(s,\cdot) \Big\rangle \Big] \\
&\qquad + \sum_{h=1}^{H} \mathbb{E}_{(s,a) \sim \pi} \left[\widehat{Q}_h(s,a) - r_h(s,a) - \sum_{s' \in \mathcal{S}} \mathbb{P}_h(s' \mid s,a) \widehat{V}_{h+1}(s') \right].
\end{aligned}$$

Lemma E.3 (Concentration of Self-Normalized Processes (Abbasi-Yadkori et al., 2011, Theorem 1)). *Let $\{x_t\}_{t=1}^{\infty}$ be a real-valued stochastic process with the correspond filtration $\{\mathcal{F}_t\}_{t=0}^{\infty}$ such that x_t is $\mathcal{F}_{t-1}$-measurable, and x_t is conditionally σ-sub-Gaussian for some $\sigma > 0$, i.e.,*

$$\forall \lambda \in \mathbb{R}, \quad \mathbb{E}[\exp(\lambda\, x_t) \mid \mathcal{F}_{t-1}] = \exp(\lambda^2\, \sigma^2 / 2)\,.$$

Let $\{\phi_t\}_{t=1}^{\infty}$ be an $\mathbb{R}^d$-valued stochastic process such that ϕ_t is $\mathcal{F}_{t-1}$-measurable. Assume Λ_0 is a $d \times d$ positive definite matrix, and let $\Lambda_t = \Lambda_0 + \sum_{i=1}^{t} \phi_i\, \phi_i^{\top}$. Then, for any $\delta > 0$, with probability at least $1 - \delta$, for all $t \geq 0$, it holds that

$$\left\| \sum_{i=1}^{t} \phi_i x_i \right\|_{\Lambda_t^{-1}}^2 \leq 2\, \sigma^2 \log \left[\frac{\det(\Lambda_t)^{1/2}\, \det(\Lambda_0)^{-1/2}}{\delta} \right].$$

Lemma E.4 (Determinant-Trace Inequality (Abbasi-Yadkori et al., 2011, Lemma 10)). *Suppose $X_1, X_2, \ldots, X_t \in \mathbb{R}^d$ and for any $s \in [t]$, $\|X\|_2 \leq L$. Let $\Lambda_t = \lambda\, I + \sum_{s=1}^{t} X_s\, X_s^{\top}$ for some $\lambda > 0$. Then, for all t, it holds that*

$$\det(\Lambda_t) \leq (\lambda + t\, L^2 / d)^d\,.$$

Lemma E.5 (Abbasi-Yadkori et al. 2011, Lemma 11). *Suppose $X_1, X_2, \ldots, X_t \in \mathbb{R}^d$ and for any $s \in [t]$, $\|X\|_2 \leq L$. Let $\Lambda_t = \Lambda_0 + \sum_{s=1}^{t} X_s\, X_s^{\top}$ and $\lambda_{\min}(\Lambda_0) \geq \max\{1, L^2\}$. Then, for all t, it hold that*

$$\log\left(\frac{\det(\Lambda_t)}{\det(\Lambda_0)} \right) \leq \sum_{s=1}^{t} \|X_t\|_{(\Lambda_t)^{-1}}^2 \leq 2\, \log\left(\frac{\det(\Lambda_t)}{\det(\Lambda_0)} \right).$$

F SAMPLE COMPLEXITY IN THE OFF-POLICY SETTING

F.1 Covering Number (Proof of Lemma 6.1)

We first present a bound for the norm of the logit.

Lemma F.1. *Consider Algorithm 1 with the* NPG *actor in Algorithm 2. Then, under Assumptions 4.1 and 4.2, for all $(t, h, s, a) \in [T] \times [H] \times \mathcal{S} \times \mathcal{A}$, it holds that*

$$\left| \left\langle \varphi(s,a), \theta_h^{t, K_t}(s,a) \right\rangle \right| \leq (\bar{\epsilon} + \eta\, H)\, t\,,$$

where $\bar{\epsilon}$ is defined in Lemma 4.1.

Proof of Lemma F.1. We will prove this by induction. When $t = 0$, since we set $\theta_h^0 = \mathbf{0}$, the statement is trivially true. For $t \geq 1$, assume that the statement stands true for $t - 1$. Since Algorithm 1 optimizes the actor loss up to some errors that are assumed to be bounded, using the triangular inequality, we have that

$$\begin{aligned}
&\left| \left\langle \varphi(s,a), \theta_h^{t, K_t}(s,a) \right\rangle \right| \\
&\quad = \left| \left\langle \varphi(s,a), \theta_h^{t, K_t} - \widehat{\theta}_h^{t, \star} \right\rangle \right| + \left| \left\langle \varphi(s,a), \widehat{\theta}_h^{t, \star}(s,a) \right\rangle \right| \\
&\quad \leq \epsilon_{\text{opt}} + \left| \left\langle \varphi(s,a), \widehat{\theta}_h^{t, \star}(s,a) \right\rangle \right|
\end{aligned}$$

$$
\begin{aligned}
&\leq \epsilon_{\text{opt}} + \left|\left\langle \varphi(s,a), \widehat{\theta}_h^{t,\star}(s,a) - \theta_h^{t-1,K_{t-1}}(s,a)\right\rangle - \eta\, \widehat{Q}_h^t(s,a)\right| \\
&\quad + \left|\left\langle \varphi(s,a), \theta_h^{t-1,K_{t-1}}(s,a)\right\rangle + \eta\, \widehat{Q}_h^t(s,a)\right| \\
&\leq \epsilon_{\text{opt}} + \epsilon_{\text{bias}} + \left|\left\langle \varphi(s,a), \theta_h^{t-1,K_{t-1}}(s,a)\right\rangle + \eta\, \widehat{Q}_h^t(s,a)\right| \\
&\leq \overline{\epsilon} + \left|\left\langle \varphi(s,a), \theta_h^{t-1,K_{t-1}}(s,a)\right\rangle\right| + \left|\eta\, \widehat{Q}_h^t(s,a)\right| \\
&\leq (\overline{\epsilon} + \eta\, H)\, t\,,
\end{aligned}
$$

where $\widehat{\theta}_h^{t,\star}$ denotes the optimal actor parameters when optimizing over $\mathcal{D}_{\text{exp}}$ and ρ_{exp}, $\left\langle \varphi(s,a), \theta_h^{t-1,K_{t-1}}(s,a)\right\rangle + \eta\, \widehat{Q}_h^t(s,a)$ is the optimization target in the actor loss of the projected NPG, and the last inequality uses the inductive hypothesis. This concludes the proof. □

Proof of Lemma 6.1. Consider any $Q, Q' \in \mathcal{Q}$ such that $Q(\cdot,\cdot) = \min\{\langle \phi(\cdot,\cdot), w\rangle, H\}^+$ and $Q'(\cdot,\cdot) = \min\{\langle \phi(\cdot,\cdot), w'\rangle, H\}^+$. Therefore, we have that

$$
\begin{aligned}
\sup_{(s,a)\in\mathcal{S}\times\mathcal{A}} |Q(s,a) - Q'(s,a)| &\leq \sup_{(s,a)\in\mathcal{S}\times\mathcal{A}} |\langle \phi(s,a), w - w'\rangle| \\
&\leq \sup_{(s,a)\in\mathcal{S}\times\mathcal{A}} \|\phi(s,a)\| \|w - w'\| \\
&\leq 2\,\overline{W}\,,
\end{aligned}
$$

where the first inequality uses the Cauchy-Schwarz inequality, and the second inequality uses Definition 2.1, the triangular inequality, and the definition of $\overline{W}$.

Consider any $\pi, \pi' \in \Pi_{\text{lin}}$ such that $\pi(\cdot \mid s) \propto \exp(\langle \phi(s,\cdot), \theta\rangle)$ and $\pi'(\cdot \mid s) \propto \exp(\langle \phi(s,\cdot), \theta'\rangle)$. By invoking Lemma F.6 and using Lemma F.1, we can observe that for a fixed $s \in \mathcal{S}$,

$$
\sup_{a\in\mathcal{A}} |\pi(s,a) - \pi'(s,a)| \leq \|\pi(s,\cdot) - \pi'(s,\cdot)\|_1 \leq 2\sqrt{\sup_a |\langle \varphi(s,a), \theta - \theta'\rangle|} \leq 2\sqrt{2\,\overline{Z}}\,.
$$

Taking the sup over $\mathcal{S}$, we get that

$$
\sup_{(s,a)\in\mathcal{S}\times\mathcal{A}} |\pi(s,a) - \pi'(s,a)| \leq 2\sqrt{2\,\overline{Z}}\,.
$$

Therefore, we can bound the logarithm of the covering number of the value function class as follows.

$$
\begin{aligned}
\log \mathcal{N}_\Delta(\mathcal{V}) &\leq \log \mathcal{N}_{\Delta/2}(\mathcal{Q}) + \log \mathcal{N}_{\Delta/(2H)}(\Pi_{\text{lin}}) \\
&\leq d_{\boldsymbol{c}} \log\left(1 + \frac{4\,\overline{W}}{\Delta}\right) + d_{\boldsymbol{a}} \log\left(1 + \frac{8\,H\,\sqrt{2\,\overline{Z}}}{\Delta}\right),
\end{aligned}
$$

where the first inequality follows from Lemma F.3, and the second inequality uses Lemma F.5. This concludes the proof. □

F.2 Proof of Good Event

Lemma F.2. *Consider Algorithm 1 in the off-policy setting with $\lambda = 1$. Then, for any $\delta \in (0,1)$, with probability at least $1-\delta$, it holds that*

$$
\left\| \sum_{(s,a)\in\mathcal{D}_h^t} \phi(s,a)\left[\widehat{V}_{h+1}^t(s) - \mathbb{P}_h \widehat{V}_{h+1}^t(s,a)\right] \right\|_{(\Lambda_h^t)^{-1}} \leq C_\delta^{\text{off}}\, H\, \sqrt{d_{\boldsymbol{c}}}\,,
$$

where

$$
C_\delta^{\text{off}} = 3\sqrt{\frac{1}{2}\log(T+1) + \log\left(\frac{2\sqrt{2}T}{H}\right) + \log\frac{2}{\delta} + \mathsf{V}}\,,
$$

$$\mathsf{V} = d_{\boldsymbol{c}} \log\left(1 + \frac{4\,\overline{W} + 4\,H\,\sqrt{2\,\overline{Z}}}{\Delta}\right) + d_{\boldsymbol{a}} \log\left(1 + \frac{4\,H\,\sqrt{2\,\overline{Z}}}{\Delta}\right),$$
$$\overline{W} = \frac{16}{3}\,H\,\sqrt{d_{\boldsymbol{c}}\,T} + \sqrt{\frac{2\,d_{\boldsymbol{c}}^3\,T}{3\,\zeta\,\delta}}, \quad \overline{Z} = (\overline{\epsilon} + \eta\,H)\,T\,.$$

Proof of Lemma F.2. Since $\widehat{V}(\cdot) \in [0.H]$, we can invoke Lemma F.4. Then, we have that for any $\Delta > 0$, with probability at least $1-\delta$,

$$\begin{aligned}
&\left\| \sum_{(s,a,s') \in \mathcal{D}_h^t} \phi(s,a) \left[\widehat{V}_{h+1}^t(s') - \mathbb{P}_h \widehat{V}_{h+1}^t(s,a) \right] \right\|_{(\Lambda_h^t)^{-1}} \\
&\quad \le \left(4\,H^2 \left[\frac{d_{\boldsymbol{c}}}{2} \log\left(\frac{T+\lambda}{\lambda}\right) + d_{\boldsymbol{c}} \log\left(\frac{\mathcal{N}_\Delta(\mathcal{V})}{\Delta}\right) + \log\frac{2}{\delta} \right] + \frac{8\,T^2\Delta^2}{\lambda} \right)^{1/2} \\
&\quad \le 2H \left[\frac{d_{\boldsymbol{c}}}{2} \log\left(\frac{T+\lambda}{\lambda}\right) + d_{\boldsymbol{c}} \log\left(\frac{\mathcal{N}_\Delta(\mathcal{V})}{\Delta}\right) + \log\frac{2}{\delta} \right]^{1/2} + \frac{2\sqrt{2}T\Delta}{\sqrt{\lambda}}\,.
\end{aligned}$$

Setting $\lambda = 1$, $\Delta = \frac{H}{2\sqrt{2}\,T}$, we have that with probability at least $1-\delta$,

$$\begin{aligned}
&\left\| \sum_{(s,a,s') \in \mathcal{D}_h^t} \phi(s,a) \left[\widehat{V}_{h+1}^t(s') - \mathbb{P}_h \widehat{V}_{h+1}^t(s,a) \right] \right\|_{(\Lambda_h^t)^{-1}} \\
&\quad \le 2\,H\,\sqrt{d_{\boldsymbol{c}}} \left[\frac{1}{2}\log(T+1) + \log\left(\frac{\mathcal{N}_\Delta(\mathcal{V})}{\frac{H}{2\sqrt{2}T}}\right) + \log\frac{2}{\delta} \right]^{1/2} + H \\
&\quad \le 3\,H\,\sqrt{d_{\boldsymbol{c}}} \left[\frac{1}{2}\log(T+1) + \log\left(\frac{2\sqrt{2}T}{H}\right) + \log\frac{2}{\delta} + \mathsf{V} \right]^{1/2},
\end{aligned}$$

where the last inequality uses Lemma 6.1 to bound the logarithm of the covering number. This concludes the proof. □

F.3 Proof of Theorem 6.2

We first instantiate Lemma D.2 in the off-policy setting. Given the above good event, we have that

$$\Gamma_{\texttt{LMC}}^{\text{off}} \le \mathcal{O}\Big(C_\delta^{\text{off}}\,H\,d_{\boldsymbol{c}}\,\sqrt{\log(1/\delta)}\Big) = \widetilde{\mathcal{O}}\Big(H\,\sqrt{d_{\boldsymbol{c}}^3\,\max\{d_{\boldsymbol{c}}, d_{\boldsymbol{a}}\}}\Big)\,.$$

Proof of Theorem 6.2. Following the proof of Theorem 6.1 (Appendix E.2), we can use the same regret decomposition as follows.

$$\begin{aligned}
&\sum_{t=1}^{T} \Big(V_1^\star(s_1) - V_1^{\pi^t}(s_1) \Big) \\
&\quad = \underbrace{\sum_{t=1}^{T}\sum_{h=1}^{H} \mathbb{E}_{\pi^\star}\left[\left\langle \pi_h^\star(\cdot \mid s) - \pi_h^t(\cdot \mid s), \widehat{Q}_h^t(s,\cdot) \right\rangle \right]}_{\text{(I) policy optimization (actor) error}} + \underbrace{\sum_{t=1}^{T}\sum_{h=1}^{H} \big(\mathbb{E}_{\pi^\star}[\iota_h^t(s,a)] - \mathbb{E}_{\pi^t}[\iota_h^t(s,a)] \big)}_{\text{(II) policy evaluation (critic) error}}\,.
\end{aligned}$$

Term (I) can be bounded the same way as the proof in Appendix E.2. Hence, it suffices to only bound Term (II).

$$\text{Term (II)} = \sum_{t=1}^{T}\sum_{h=1}^{H} \mathbb{E}_{\pi^\star}[\iota_h^t(s,a)] - \mathbb{E}_{\pi^t}[\iota_h^t(s,a)]$$

$$\overset{\text{(i)}}{\leq} -\sum_{t=1}^{T}\sum_{h=1}^{H}\mathbb{E}_{\pi^t}[\iota_h^t(s,a)]$$
$$\overset{\text{(ii)}}{\leq} \Gamma^{\text{off}}_{\texttt{LMC}}\sum_{t=1}^{T}\sum_{h=1}^{H}\mathbb{E}_{\pi^t}\left[\left\|\phi(s_h^t,a_h^t)\right\|_{(\Lambda_h^t)^{-1}}\right].$$

(i) and (ii) both follow from Lemma D.2, where (i) is based on the optimism guarantee (RHS of Eq. (10)), while (ii) is based on the error bound (LHS of Eq. (10)).

Bounding the sum of bonuses. Since $\Gamma^{\text{off}}_{\texttt{LMC}}$ is bounded, it suffices to bound $\mathbb{E}_{\pi^t}\left[\left\|\phi(s,a)\right\|_{(\Lambda_h^t)^{-1}}\right]$. We then index each data point in $\mathcal{D}_h^t$ as $\left\{(s_h^i,a_h^t,s_{h+1}^t)\right\}_{t\in[T]}$ and get that $\Lambda_h^t=\sum_{t=1}^{T}\phi(s_h^t,a_h^t)\phi(s_h^t,a_h^t)^\top+\lambda I$. Then, we have that

$$\sum_{t=1}^{T}\sum_{h=1}^{H}\mathbb{E}_{\pi^t}\left[\left\|\phi(s,a)\right\|_{(\Lambda_h^t)^{-1}}\right]$$
$$=\sum_{i=1}^{T}\sum_{h=1}^{H}\left\|\phi(s_h^t,a_h^t)\right\|_{(\Lambda_h^{t,i})^{-1}}+\sum_{t=1}^{T}\sum_{h=1}^{H}\underbrace{\mathbb{E}_{\substack{s_h\sim\mathbb{P}(\cdot|s_{h-1}^t a_{h-1}^t)\\ a_h\sim\pi_h^t(\cdot|s_h)}}\left[\left\|\phi(s_h,a_h)\right\|_{(\Lambda_h^t)^{-1}}\right]-\left\|\phi(s_h^t,a_h^t)\right\|_{(\Lambda_h^t)^{-1}}}_{:=\mathcal{M}_{t,h}^{\text{off}}}. \tag{16}$$

Applying the elliptical potential lemma. For the first term of Eq. (16), we have that

$$\sum_{t=1}^{T}\sum_{h=1}^{H}\left\|\phi(s_h^t,a_h^t)\right\|_{(\Lambda_h^t)^{-1}}=\sum_{h=1}^{H}\sum_{t=1}^{T}\left\|\phi(s_h^t,a_h^t)\right\|_{(\Lambda_h^t)^{-1}}$$
$$\overset{\text{(iii)}}{\leq}\sum_{h=1}^{H}\sqrt{T}\left(\sum_{t=1}^{T}\left\|\phi(s_h^t,a_h^t)\right\|^2_{(\Lambda_h^t)^{-1}}\right)^{1/2}$$
$$\overset{\text{(iv)}}{\leq}\mathcal{O}\left(\sqrt{d_{\boldsymbol{c}}\,H^2\,T\log(T/\delta)}\right).$$

(iii) applies the Cauchy-Schwarz inequality, and (iv) follows the elliptical potential argument from Lemma E.5.

A martingale difference sequence. For the second term of Eq. (16), since $\left\{\mathcal{M}_{t,h}^{\text{off}}\right\}_{(t,h)\in[T]\times[H]}$ forms a martingale sequence adapted to the filtration,

$$\mathcal{F}_{t,h}^{\text{off}}=\left\{(s_\tau^i,a_\tau^i)\right\}_{(i,\tau)\in[t-1]\times[H]}\cup\left\{(s_\tau^t,a_\tau^t)\right\}_{\tau\in[h-1]},$$

such that $\mathbb{E}\left[\mathcal{M}_{t,h}^{\text{off}}\mid\mathcal{F}_{t,h}^{\text{off}}\right]=0$. Since $|\mathcal{M}_{i,h}^{\text{off}}|\leq 1$, we can apply the Azuma–Hoeffding inequality and obtain that

$$\Pr\left(\sum_{t=1}^{T}\sum_{h=1}^{H}\mathcal{M}_{t,h}^{\text{off}}\geq m\right)\geq\exp\left(\frac{-m^2}{2\,H\,T}\right).$$

Setting $m=\sqrt{2\,H\,T\,\log(1/\delta)}$, with probability at least $1-\delta$, it holds that

$$\sum_{t=1}^{T}\sum_{h=1}^{H}\mathcal{M}_{t,h}^{\text{off}}\leq\sqrt{2\,H\,T\,\log(1/\delta)}\leq\mathcal{O}\left(\sqrt{H\,T\,\log(1/\delta)}\right).$$

Therefore, we have that

$$\sum_{t=1}^{T}\sum_{h=1}^{H}\mathbb{E}_{\pi^t}\left[\left\|\phi(s,a)\right\|_{(\Lambda_h^t)^{-1}}\right]=\sum_{t=1}^{T}\sum_{h=1}^{H}\left\|\phi(s_h^i,a_h^i)\right\|_{(\Lambda_h^{t,i})^{-1}}+\sum_{t=1}^{T}\sum_{h=1}^{H}\mathcal{M}_{i,h}^{\text{off}}\leq\mathcal{O}\left(\sqrt{d_{\boldsymbol{c}}\,H^2\,T\log(T/\delta)}\right).$$

It further implies that, with probability at least $1-\delta$,

$$\text{Term (II)}\leq\Gamma^{\text{off}}_{\texttt{LMC}}\sum_{t=1}^{T}\sum_{h=1}^{H}\mathbb{E}_{\pi^t}\left[\left\|\phi(s,a)\right\|_{(\Lambda_h^t)^{-1}}\right]\leq\widetilde{\mathcal{O}}\left(\sqrt{d_{\boldsymbol{c}}^3\,\max\{d_{\boldsymbol{c}},d_{\boldsymbol{a}}\}\,H^4\,T}\right).$$

Putting everything together. Therefore, we have that with probability at least $1-\delta$,

$$\mathbb{E}\Big[V_1^\star(s_1) - V_1^{\overline{\pi}^T}(s_1)\Big] = \frac{1}{T}(\text{Term (I)} + \text{Term (II)}) = \widetilde{\mathcal{O}}\left(\frac{H^2\sqrt{d_{\boldsymbol{c}}^3\,\max\{d_{\boldsymbol{c}}, d_{\boldsymbol{a}}\}\,\log|\mathcal{A}|}}{\sqrt{T}} + H^2\sqrt{\epsilon}\right).$$

This concludes the proof. □

F.4 Technical Tools

Lemma F.3 (Zhong and Zhang, 2023, Lemma B.1). *Consider the value function class $\mathcal{V} = \{\langle Q(\cdot,\cdot), \widehat{\pi}(\cdot \mid \cdot)\rangle_{\mathcal{A}} \mid Q \in \mathcal{Q}, \widehat{\pi} \in \Pi\}$. Then, it holds that*

$$\mathcal{N}_\Delta(\mathcal{V}) \le \mathcal{N}_{\Delta/2}(\mathcal{Q}) \cdot \mathcal{N}_{\Delta/(2\,H)}(\Pi)\,.$$

Lemma F.4 (Value-Aware Uniform Concentration (Jin et al., 2020, Lemma D.4)). *Let $\{s_t\}_{t=1}^\infty$ be a stochastic process on the state space $\mathcal{S}$ with the correspond filtration $\{\mathcal{F}_t\}_{t=0}^\infty$ such that s_t is $\mathcal{F}_{t-1}$-measurable. Let $\{\phi_t\}_{t=1}^\infty$ be an $\mathbb{R}^d$-valued stochastic process such that ϕ_t is $\mathcal{F}_{t-1}$-measurable, and $\|\phi_t\| \le 1$. Let $\Lambda_t = I + \sum_{s=1}^t \phi_s\,\phi_s^\top$. Assume $\mathcal{V}$ is a value function class such that $\sup_{s\in\mathcal{S}}|V(s)| \le H$. Then, for any $\delta > 0$, with probability at least $1-\delta$, for all $t \ge 0$ and any $V \in \mathcal{V}$, it holds that*

$$\left\|\sum_{i=1}^t \phi_i\,\{V(s_i) - \mathbb{E}[V(s_i) \mid \mathcal{F}_{i-1}]\}\right\|_{\Lambda_t^{-1}}^2 \le 4\,H^2\left[\frac{d}{2}\log\left(\frac{t+\lambda}{\lambda}\right) + \log\left(\frac{\mathcal{N}_\Delta}{\delta}\right)\right] + \frac{8\,t^2\,\Delta^2}{\lambda}\,,$$

where $\mathcal{N}_\Delta$ represents the Δ-covering number of $\mathcal{V}$ with the distance measured by dist$(V, V') = \sup_{s\in\mathcal{S}}|V(s) - V'(s)|$.

Lemma F.5 (Covering Number of Euclidean Ball). *For any $\Delta > 0$, the Δ-covering number, $\mathcal{N}_\Delta$, of the Euclidean ball of radius $B > 0$ in $\mathbb{R}^d$ satisfies that*

$$\mathcal{N}_\Delta \le \left(1 + \frac{2B}{\Delta}\right)^d.$$

Lemma F.6 (Zhong and Zhang 2023, Lemma B.3). *For $\pi, \pi' \in \Delta(\mathcal{A})$ and $Z, Z' : \mathcal{A} \to \mathbb{R}^+$, if $\pi(\cdot) \propto \exp(Z(\cdot))$ and $\pi'(\cdot) \propto \exp(Z'(\cdot))$, then it holds that*

$$\|\pi - \pi'\|_1 \le 2\sqrt{\|Z - Z'\|_\infty}\,.$$

G EXPERIMENTS

In this section, we evaluate the performance of our proposed algorithm with other methods on various benchmarks. In Appendix G.1, we test our proposed algorithm in two specific environments of linear MDPs. In Appendix G.2, we further conduct some ablation studies in the same two environments. In Appendix G.3, we test our proposed algorithm in large-scale deep RL applications (Atari (Mnih et al., 2013)) and compare its performance to two commonly used deep RL algorithms, PPO (Schulman et al., 2017b) in the on-policy setting and DQN (Mnih et al., 2015) in the off-policy setting.

G.1 Experiments in Linear MDPs

First, we test our proposed algorithm in the randomly generated linear MDPs (Random MDP) and the linear MDP version of the Deep Sea (Osband et al., 2019).

G.1.1 Environment Setup

Our experimental setup is an extension of Ishfaq et al. (2024a). In particular, we extend the prior off-policy setting to test our proposed algorithm in the linear MDP version of the Deep Sea (Osband et al., 2019) and the Random MDP. In both experiments, we use the linear MDP features as the policy features (i.e., $\phi = \varphi$), and we set $d \coloneqq d_{\boldsymbol{c}} = d_{\boldsymbol{a}}$ to represent the feature dimension for both the actor and the critic parameters.

For the Random MDP environment, we consider 15 states and 5 actions. For each state $s \in \mathcal{S}$, we generate $\psi_h(s) \in \mathbb{R}^d$ uniformly at random in $[0, 1]$ and construct tile coded features. The agent always starts from state 0, receiving a small reward of 0.1 upon taking action 0, and obtains the maximum reward when reaching the final

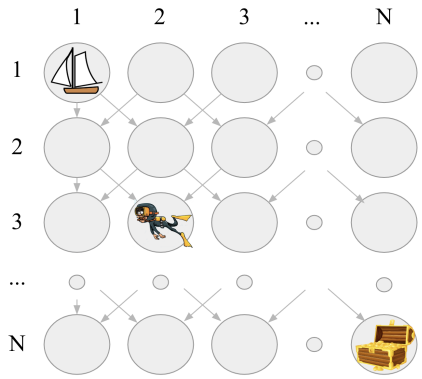

Figure 2: Example of the Deep Sea environment from Osband et al. (2019).

state and taking action 1. All other state-action pairs yield zero reward. Given this reward function and a randomly generated transition kernel, we solve for ψ_h and υ_h via minimizing least square errors following Definition 2.1 and use the corresponding $\mathbb{P}_h$ and r_h to set up the environment.

For the Deep Sea environment, we use a $N \times N$ grid with $N = 10$ where the agent always starts at $(1, 1)$ and can move either bottom-right or bottom-left, receiving rewards of 0 and $-0.01/N$ respectively. Reaching the bottom-right corner yields a reward of 1. Furthermore, we generate the actor and critic features by projecting each state-action pair uniformly between $[0, d-1]$, which recovers one-hot encoded features when $d = |\mathcal{S}| \times |\mathcal{A}|$. Given the true transition probabilities and rewards, we solve for ψ_h and υ_h via minimizing least square errors following Definition 2.1 and use the corresponding $\mathbb{P}_h$ and r_h to ensure the linearity of the MDP.

G.1.2 Coreset Construction

To implement our proposed algorithm, we need to conduct the experimental design to obtain $\mathcal{D}_{\text{exp}}$ and ρ_{exp}. For this, we follow the offline G-Experimental design outlined in Algorithm 4 to construct a coreset. In particular, in each iteration, this greedy iterative algorithm traverses the entire state-action space and adds a data point to the coreset that has the highest marginal gain $g(s, a) = \|\varphi(s, a)\|_{G^{-1}}$. For a specific threshold ϵ_G, the algorithm only terminates when $g_{\max} = \max_{s,a \in (\mathcal{S} \times \mathcal{A})} g(s, a) \leq \epsilon_G$, hence giving us direct control over $\sup_{(s,a) \in \mathcal{S} \times \mathcal{A}} \|\varphi(s, a)\|_{G^{-1}}$. In practice, we find that it often selects too many data points, so we cap the coreset at 80% of the total data.

Algorithm 4 Coreset Construction Using G-Experimental Design

Input: features $\varphi : \mathcal{S} \times \mathcal{A} \mapsto \mathbb{R}^{d_a}$, threshold $\epsilon_G \in \mathbb{R}$
Initialize: $G = I_{d_a \times d_a}$, $\mathcal{D}_{\text{exp}} = \emptyset$, $g_{\max} = \infty$
while $g_{\max} > \epsilon_G$ **do**
  $g_{\max} = 0$
  for $(s, a) \in \mathcal{S} \times \mathcal{A}$ **do**
    $g(s, a) = \|\varphi(s, a)\|_{G^{-1}}$
    if $g_{\max} < g(s, a)$ **then**
      $(s^\star, a^\star) = (s, a)$
      $g_{\max} = g(s, a)$
    $\mathcal{D}_{\text{exp}} = \mathcal{D}_{\text{exp}} \cup \{(s^\star, a^\star)\}$
    $G = G + \varphi(s^\star, a^\star)\, \varphi(s^\star, a^\star)^\top$

G.1.3 Algorithms and Hyperparameters

We denote by `LMC-NPG-EXP` our proposed algorithm with an explicit log-linear policy parameterization that uses `LMC` for policy evaluation and projected `NPG` for policy optimization over the obtained coreset. We denote by `LMC-NPG-IMP` an idealized variant of `NPG` that does not have an explicit policy parameterization and maintains an implicit policy by storing all parameterized Q functions (and hence requires significantly more memory). As a baseline, we also consider the value-based algorithm `LMC` (Ishfaq et al., 2024a).

In Table 3, we list the hyperparameters used across all experiments. For log-linear policies, the actor loss in Eq. (6)

admits a closed-form solution, allowing us to avoid tuning of the actor learning rate α_a and the number of actor updates K_t, by minimizing the objective exactly. In general, for non-linear models, inexact optimization (e.g., stochastic gradient descent) is usually required to optimize the actor loss.

Hyperparameter	LMC	LMC-NPG-IMP	LMC-NPG-EXP
Policy Optimization Learning Rate (η)	✗	$[0.1, 1, 10, 100]$	$[0.1, 1, 10, 100]$
Inverse Temperature (ζ^{-1})	$[10^{-2}, 10^{-3}, 10^{-4}, 10^{-5}]$		
Number of Critic Updates (J_t)	100		
Critic Learning Rate (α_c)	$[10^{-2}, 10^{-3}, 10^{-4}, 10^{-5}]$		
Number of Episodes (T)	600		
Horizon Length (H)	100		

Table 3: Hyperparameter search space for our experiments in linear MDPs.

G.1.4 Experimental Results

Following the protocol of Ishfaq et al. (2024a), each algorithm is run with 20 random seeds. We sweep the hyperparameters as shown in Table 3 and report the best performance with 95% confidence intervals.

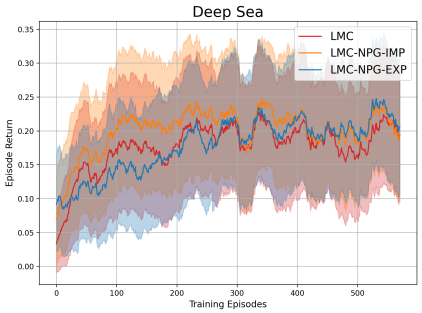

Figure 3: Comparison of `LMC-NPG-EXP` (our proposed framework), `LMC-NPG-IMP` (memory-intensive variant), and `LMC` (value-based baseline) in the Deep Sea environment.

For the Random MDP, Figure 1 indicates that `LMC-NPG-EXP` closely matches `LMC-NPG-IMP` while outperforming the value-based baseline, `LMC`. For the linear MDP version of Deep Sea, Figure 3 showcases that `LMC-NPG-EXP` can achieve comparable performance with `LMC-NPG-IMP` and `LMC`.

G.2 Ablation Studies

G.2.1 Ablation on Exploration

To study the impact of the exploration mechanism, `LMC`, in our proposed algorithm, we perform an ablation in the linear MDP variant of Deep Sea. For the baseline without exploration, we consider the same algorithm design as `LMC-NPG-EXP` and simply do not inject any noise into the LMC update. The results in Figure 4 indicate that when the feature dimensions of the critic and the actor are relatively small, the exploration mechanism is crucial for our proposed algorithm to achieve decent performance.

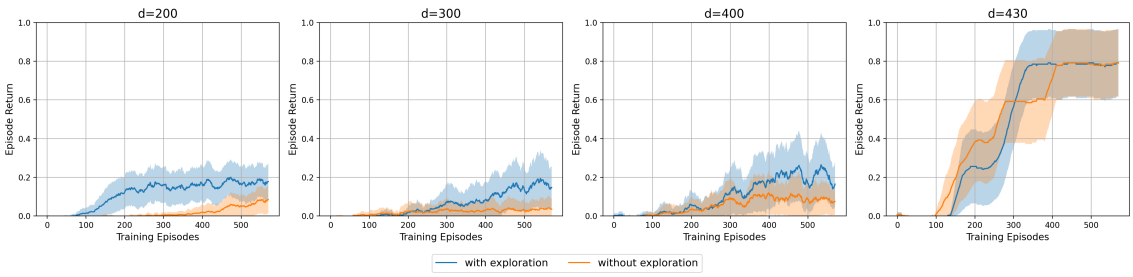

Figure 4: Ablation of the exploration mechanism for `LMc-NPG-EXP`.

G.2.2 Ablation on Feature Dimensions

We also study the effect of the feature dimensions. We use the same feature for the MDP environment and the policy (i.e., $\phi = \varphi$), and we denote that $d := d_{\boldsymbol{c}} = d_{\boldsymbol{a}}$. The results in Figure 5 show that larger feature dimensions d for both the actor and the critic lead to greater performance of the proposed algorithm.

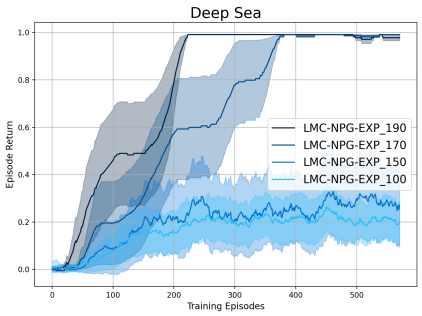

Figure 5: Effect of feature dimension d in Deep Sea.

G.2.3 Sensitivity to Inverse Temperature (ζ^{-1})

We study the sensitivity of our algorithm to the inverse temperature hyperparameter, ζ^{-1}, in the Deep Sea environment. As shown Figure 6, we observe relatively robust performance for different choices of this hyperparameter across a range of different feature dimensions.

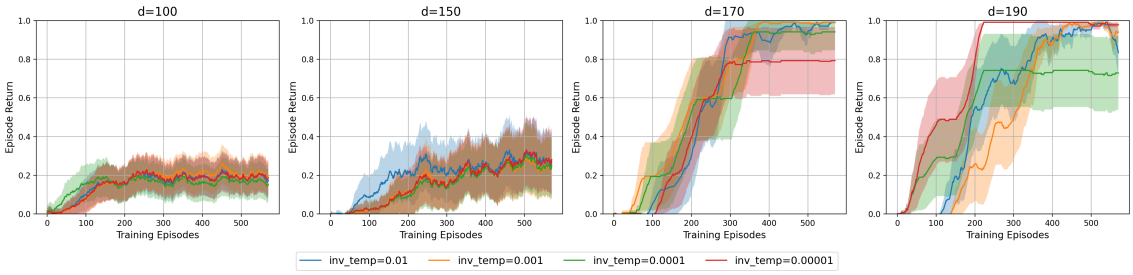

Figure 6: `LMC-NPG-EXP` exhibits robustness to different choices of ζ^{-1}.

G.2.4 Sensitivity to the Number of Critic Samples (M)

We further study the effect of the number of critic samples, M, in the Deep Sea environment. We vary M across a range of different feature dimensions. Similarly, as shown in Figure 7, we find that the performance of our proposed algorithm remains relatively stable across the tested range of different feature dimensions.

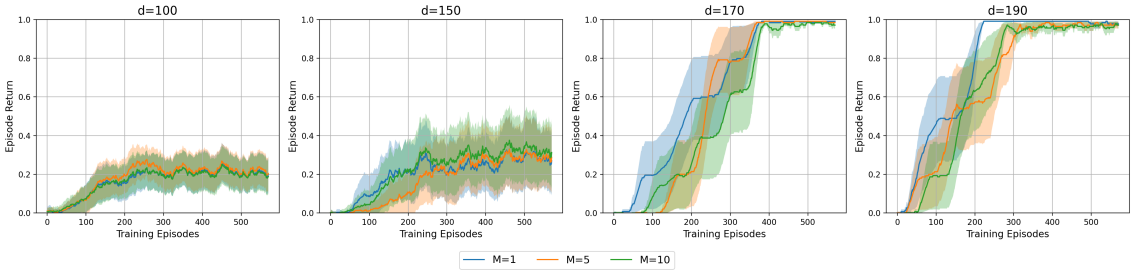

Figure 7: `LMC-NPG-EXP` exhibits robustness to different choices of M.

G.3 Experiments Beyond Linear MDPs: Atari

G.3.1 Extension to Deep RL Applications

Compared to our theoretical results in the finite-horizon linear MDP setting, the Atari benchmark requires handling discounted problems with complex nonlinear function approximation. Consequently, starting from our theoretically principled algorithm, we make three mild changes and follow the same protocol as the practical version of `LMC` in Ishfaq et al. (2024a). Since we are not in the finite-horizon setting, we compute the value functions and Q functions in a forward fashion ($h = 1, \ldots$) rather than backwards ($h = H, \ldots, 1$), which changes how the critic loss is formed in Algorithm 3. Moreover, instead of using stochastic gradient descent to optimize the `LMC` critic loss, we follow Ishfaq et al. (2024a) and use the practical version of `LMC` that integrates Adam. Finally, rather than relying on the experimental design, at iteration t, we roll out the current policy to store state–action pairs in a buffer $\mathcal{D}^t$ and minimize the actor objective over (a subset of) it. Following such changes, we can extend our proposed algorithm beyond the linear function approximation and empirically test its performance against other commonly used algorithms.

G.3.2 Environment Setups and Hyperparameters

In the Atari experiments, we use the following setups. In the on-policy setting, we adopt the recommended hyperparameters from Raffin (2020) for `PPO` and our algorithm. In the off-policy setting, following prior work (Tomar et al., 2020), we use the default hyperparameters from stable baselines (Raffin et al., 2021) for `DQN` and our algorithm. These setups are motivated by two considerations. First, we aim to evaluate the effect of different objectives without performing an extensive hyperparameter search. Second, the CNN-based actor and the critic architectures make large grid searches over multiple hyperparameters (e.g., framestack, λ in GAE, horizon length, and discount factor) computationally prohibitive. A complete list of hyperparameters used in the on-policy and off-policy Atari experiments is provided in Tables 4 and 5. Additionally, for the policy optimization learning rate η in our algorithm, we perform a grid search over $\{0.01, 0.1, 1.0\}$.

G.3.3 Experimental Results

As illustrated in Figure 8, our algorithm can achieve comparable or even better performance than `PPO` in the on-policy setting. Similarly, in the off-policy setting, our algorithm's performance is comparable to or exceeds `DQN` in most considered games, as shown in Figure 9. These results underscore that our theoretically grounded approach holds significant practical value for large-scale deep RL applications.

Hyperparameter	LMC-NPG-EXP	PPO
Reward normalization	✗	✗
Observation normalization	✗	✗
Orthogonal weight initialization	✓	✓
Value function clipping	✗	✗
Gradient clipping	✗	✓
Probability ratio clipping	✗	✓
Clip range	✗	0.1
Entropy coefficient	0	0.01
Number of inner loop updates (m)	5	4
Adam step-size	3×10^{-4}	2.5×10^{-4}
Value Function Coefficient	✗	0.5
Minibatch size	256	
Framestack	4	
Number of environment copies	8	
GAE (λ)	0.95	
Horizon (T)	128	
Discount factor	0.99	
Total number of timesteps	10^7	
Number of runs for plot averages	5	
Confidence interval for plot runs	$\sim 95\%$	

Table 4: Hyperparameters for the Atari experiments in the on-policy setting.

Hyperparameter	LMC-NPG-EXP	DQN
Reward normalization	✗	✗
Observation normalization	✗	✗
Orthogonal weight initialization	✗	✗
Value function clipping	✗	✗
Gradient clipping	✗	✗
Probability ratio clipping	✗	✗
Exploration	LMC	ϵ-greedy
Adam step-size	3×10^{-4}	
Buffer size	10^6	
Minibatch size	256	
Framestack	4	
Number of environment copies	8	
Discount factor	0.99	
Total number of timesteps	10^7	
Number of runs for plot averages	5	
Confidence interval for plot runs	$\sim 95\%$	

Table 5: Hyperparameters for the Atari experiments in the off-policy setting.

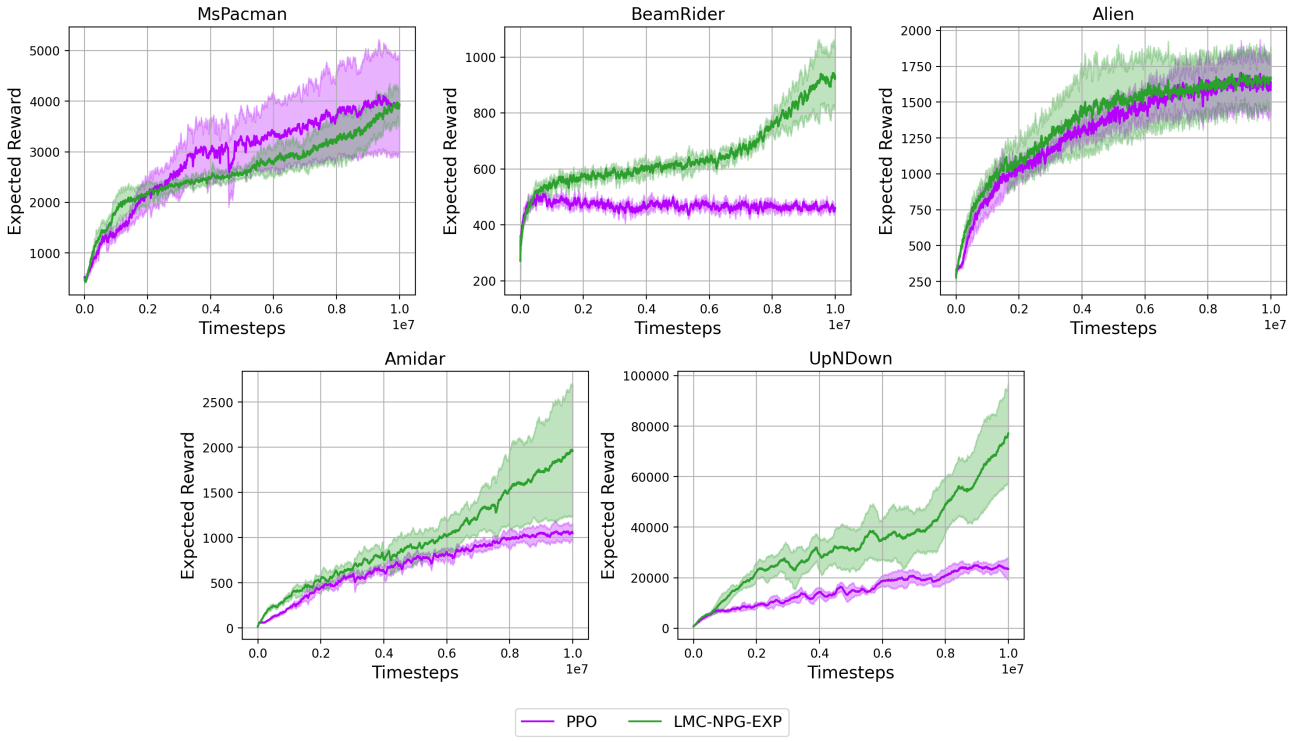

Figure 8: In the on-policy setting, LMC-NPG-EXP achieves comparable or better performance compared to PPO.

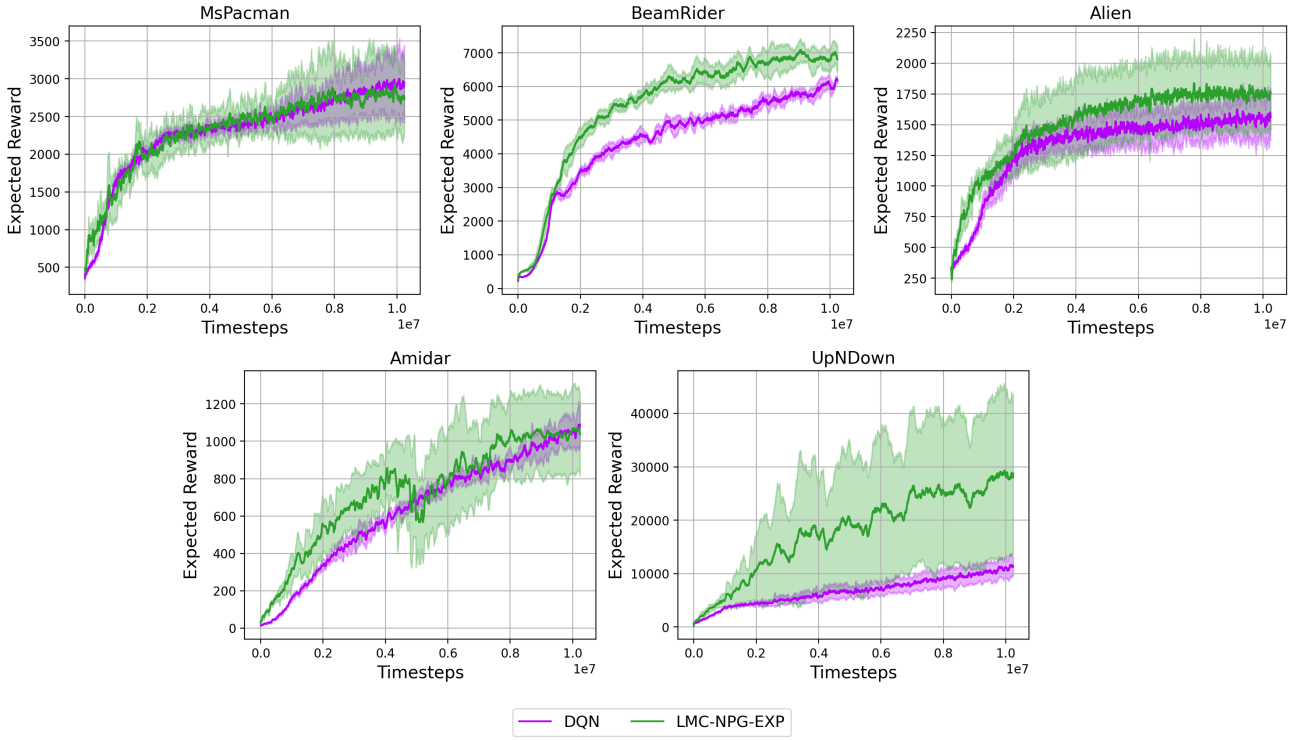

Figure 9: In the off-policy setting, LMC-NPG-EXP achieves comparable or better performance compared to DQN.