Optimistic Actor-Critic with Parametric Policies: Unifying Sample Efficiency and Practicality

Max Qiushi Lin
Simon Fraser University
maxqslin@gmail.com

Reza Asad
Simon Fraser University
rasad@sfu.ca

Kevin Tan
University of Pennsylvania
kevtan@umich.edu

Haque Ishfaq
Mila and McGill University
haque.ishfaq@mail.mcgill.ca

Csaba Szepesvári
Google DeepMind and University of Alberta
szepi@google.com

Sharan Vaswani
Simon Fraser University
vaswani.sharan@gmail.com

Abstract

Although actor-critic (AC) 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 that require complicated algorithmic modifications. Moreover, the AC methods analyzed for finite-horizon MDPs often construct "implicit" policies without explicitly parameterizing them, further exacerbating the mismatch between theory and practice. To that end, we propose an optimistic AC framework with parametric policies that is both practical and equipped with theoretical guarantees for episodic linear MDPs. In particular, we introduce a tractable regression objective for the actor to train log-linear policies. This enables us to control the error between the parameterized actor and the easier-to-analyze implicit policies induced by natural policy gradient. To train the critic, we use approximate Thompson sampling via Langevin Monte Carlo to obtain optimistic value estimates. This results in a principled, yet flexible exploration scheme without any additional assumptions on the MDP. We prove that our algorithm achieves an $\widetilde{\mathcal{O}}(\epsilon^{-4})$ sample complexity in the on-policy setting and an $\widetilde{\mathcal{O}}(\epsilon^{-2})$ complexity in the off-policy setting. Our algorithm matches prior theoretical work in achieving state-of-the-art sample efficiency, while being 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 aligning 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].

2nd Workshop on Aligning Reinforcement Learning Experimentalists and Theorists (ARLET 2025).

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 work on PG methods [Agarwal et al., 2021a, Yuan et al., 2023, Alfano and Rebeschini, 2022, Asad et al., 2025] obtains convergence rates that involve a mismatch ratio between the optimal policy and the initial state 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 work [Hong et al., 2023, Fu et al., 2021, Xu et al., 2020, Cayci et al., 2024] assumes 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 work [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] tackles 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. Consequently, this implementation has a memory complexity that is linear in the number of updates. This drastically deviates from practice, where algorithms typically employ explicitly parameterized complex models as learnable policies and optimize them with gradient descent-based methods. Furthermore, in the off-policy setting, these works require complicated algorithmic modifications, further exacerbating the mismatch between theory and practice. For example, Sherman et al. [2023] adopts a warm-up procedure from Wagenmaker et al. [2022] that does not resemble policy optimization and is difficult to implement in practice. Although Cassel and Rosenberg [2024] can avoid the warm-up phase, it still requires feature contraction techniques that are non-standard in practice, and difficult to extend beyond linear function approximation. The issues above indicate a significant gap between theory and practice for AC methods. Thus, we address the following question:

Can we design a provably sample-efficient, yet practical, actor–critic algorithm with parametric policies for both the on- and off-policy regimes?

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 employs an explicitly parameterized policy. We analyze 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 work [Cassel and Rosenberg, 2024, Sherman et al., 2023, Liu et al., 2023, Zhong and Zhang, 2023], 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 $\widetilde{\mathcal{O}}(1/\epsilon^4)$ 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 $\widetilde{\mathcal{O}}(1/\epsilon^2)$ in the off-policy setting. This matches the results of Sherman et al. [2023], Cassel and Rosenberg [2024], Tan et al. [2025], but with a far less complicated algorithm design.

We thus demonstrate that our optimistic actor-critic method is both practical and sample-efficient.

2 Preliminaries

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

Episodic Linear MDP. An episodic 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 $|\mathcal{A}|$. We note that $\mathbb{P}_h(\cdot \mid s, a) \in \Delta(\mathcal{S})$ 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 given policy π, we define the value function as $V_h^\pi(s) \coloneqq \mathbb{E}[\sum_{\tau=h}^H r_\tau(s_\tau, a_\tau) \mid s_h = s, \pi]$, and the corresponding state-action value function is defined as $Q_h^\pi(s, a) \coloneqq \mathbb{E}_{\pi,\mathbb{P}}[\sum_{\tau=h}^H r_\tau(s_\tau, a_\tau) \mid s_h = s, a_h = a, \pi]$, 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)**.** *An episodic MDP $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathbb{P}, r, H)$ is a linear MDP with a feature map $\phi : \mathcal{S} \times \mathcal{A} \mapsto \mathbb{R}^{d_c}$ if the following holds. There exist H signed measures $\psi_h : \mathcal{S} \to \mathbb{R}^{d_c}$ and $\upsilon_h : \mathbb{R}^{d_c}$ such that $\mathbb{P}_h(s' \mid s, a) = \langle \phi(s, a), \psi_h(s') \rangle$ and $r_h(s, a) = \langle \phi(s, a), \upsilon_h \rangle$. It should also satisfy the following constraints: $\|\phi(s, a)\| \le 1$, $\|\psi_h(s)\| \le \sqrt{d_c}$, and $\|\upsilon_h\| \le \sqrt{d_c}$ for all h, s, and a. Additionally, for any measurable function $V : \mathcal{S} \to [0, 1]$, $\left\|\int_{s\in\mathcal{S}} V(s)\, \psi_h(s)\, \mathrm{d}\, s\right\| \le \sqrt{d_c}$.*

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 $w_h \in \mathbb{R}^{d_c}$ 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 available 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 $\mathrm{Reg(T)} \coloneqq \sum_{\mathrm{t}=1}^{\mathrm{T}}[\mathrm{V}_1^\star(\mathrm{s}_1) - \mathrm{V}_1^{\pi_\mathrm{t}}(\mathrm{s}_1)]$, 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 that0

$$\mathrm{OG(T)} \coloneqq \mathbb{E}\Big[\mathrm{V}_1^\star(\mathrm{s}_1) - \mathrm{V}_1^{\overline{\pi}^\mathrm{T}}(\mathrm{s}_1)\Big] = \frac{\mathrm{Reg(T)}}{\mathrm{T}} \leq \widetilde{\mathcal{O}}(\epsilon)\,,$$

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 $\varphi : \mathcal{S} \times \mathcal{A} \mapsto \mathbb{R}^{d_a}$ are policy features given to the learner. W.l.o.g, we assume that $\|\varphi(s, a)\| \leq 1$ 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.

3 Optimistic Actor-Critic Framework

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 $\pi^1, \ldots, \pi^t$ 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.

Algorithm 1 Optimistic Actor-Critic with Parametric Policies

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

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 each (h, 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 works 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 works use the following equivalent form of the NPG update: $\pi_h^{t+1}(\cdot|s) \propto \pi_h^1(\cdot|s) \exp(\eta \sum_{i=1}^t \widehat{Q}_h^i(s, \cdot))$ and characterize 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 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. Consequently, these existing works require storing *all* the parameterized Q functions, and have a memory complexity linear in $|\mathcal{S}|$ or T. Consequently, 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:

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

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.

Lemma 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 $t \in [T]$,*

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

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

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:

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

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.

[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.

4.2 Controlling the Projection Error for Log-Linear Policies

To bound the projection error in Lemma 4.1, one could choose that $\mathrm{Proj}_\Pi(p) = \arg\min_p \mathrm{KL}(u \parallel p) - \mathrm{KL}(u \parallel p^{t+1/2})$, 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: $\frac{1}{2}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \left[\left\langle \varphi(s,a), \theta - \widehat{\theta}_h^t \right\rangle - \eta\, \widehat{Q}_h^t(s,a)\right]^2$, 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 $\mathcal{D}_{\mathrm{exp}} \subset \mathcal{S} \times \mathcal{A}$ along with a corresponding distribution $\rho_{\mathrm{exp}} \in \Delta(\mathcal{D}_{\mathrm{exp}})$ that offers good coverage of the feature space. Given $\mathcal{D}_{\mathrm{exp}}$ and ρ_{exp}, we instantiate the actor loss $\tilde{\ell}_h^t(\theta)$:

$$\widetilde{\ell}_h^t(\theta) = \tfrac{1}{2}\sum_{(s,a)\in\mathcal{D}_{\mathrm{exp}}} \rho_{\mathrm{exp}}(s,a) \left[\left\langle \varphi(s,a), \theta - \widehat{\theta}_h^t \right\rangle - \eta\, \widehat{Q}_h^t(s,a)\right]^2, \tag{6}$$

In order to construct $\mathcal{D}_{\mathrm{exp}}$ and ρ_{exp} and 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)**.** *Suppose φ is the given policy feature for the log-linear policy. Then, it holds that* $\inf_\theta \sup_{t,h,s,a} \left|\langle \varphi(s,a), \theta\rangle - \eta\, \widehat{Q}_h^t(s,a)\right| \le \epsilon_{\mathrm{bias}}$.

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 $\tilde{\ell}_h^t(\theta)$ is sufficiently minimized.

Assumption 4.2 (Optimization Error)**.** *Suppose θ_h^t is obtained by minimizing $\tilde{\ell}_h^t(\theta)$. Let $\widehat{\theta}_h^{t,\star} = \arg\min_\theta \tilde{\ell}_h^t(\theta)$. Then, it holds that* $\sup_{t,h} \left\|\theta_h^t - \widehat{\theta}_h^{t,\star}\right\| \le \epsilon_{\mathrm{opt}}$.

In practice, minimizing $\tilde{\ell}_h^t(\theta)$ by K_t steps of gradient descent ensures that $\epsilon_{\mathrm{opt}} \le \mathcal{O}(\exp(-K_t))$. Given these two mild assumptions, we can then proceed to bound the projection error. Using Lemma 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 4.2. *Under Assumptions 4.1 and 4.2, suppose $\overline{\varphi}_G \coloneqq \sup_{(s,a)\in\mathcal{S}\times\mathcal{A}} \|\varphi(s,a)\|_{G^{-1}}$ where $G \coloneqq \sum_{(s,a)\in\mathcal{D}_{\mathrm{exp}}} \rho_{\mathrm{exp}}(s,a)\, \varphi(s,a)\, \varphi(s,a)^\top$, then, for all (t,h,s),*

$$\left|\epsilon_h^t(s)\right| \le \bar{\epsilon} \coloneqq \sqrt{2}\,(\overline{\varphi}_G + 1)\,\epsilon_{\mathrm{bias}} + \sqrt{2}\,\epsilon_{\mathrm{opt}}\,.$$

The above lemma is true for any choice of $\mathcal{D}_{\mathrm{exp}}$ and ρ_{exp}, and suggests that if we can control $\overline{\varphi}_G$, the projection error can be bounded. Therefore, we would like to construct a suitable $\mathcal{D}_{\mathrm{exp}}$ and ρ_{exp} to bound $\|\varphi(s,a)\|_{G^{-1}}$ for any $(s,a) \in \mathcal{S} \times \mathcal{A}$ and solve the following optimization problem:

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

which fits the form of experimental design. Ideally, we would also like $|\mathcal{D}_{\mathrm{exp}}|$ 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* 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 $|\mathcal{D}_{\mathrm{exp}}| \le \widetilde{\mathcal{O}}(d_{\boldsymbol{a}})$. 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. 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). Given access to such a coreset, we can guarantee that $\bar{\epsilon} \leq \mathcal{O}(d_{\boldsymbol{a}}\, \epsilon_{\text{bias}} + \epsilon_{\text{opt}})$, and optimizing the actor loss 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 bound $\|\varphi(s,a)\|_{G^{-1}}$ and offer similar guarantees. We defer these details to Appendix C.

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]).

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.

Algorithm 2 Actor: Projected NPG

Input: critic parameters w^t, policy optimization learning rate η, number of actor updates K_t, actor learning rate $\alpha^t_{\boldsymbol{a}}$, subset and distribution of the state-action space $\mathcal{D}_{\text{exp}}$ and ρ_{exp}
for $h = 1, 2, \ldots, H$ **do**
  $\widehat{Q}^t_h(\cdot,\cdot) = \text{clip}_{[0,\text{H}-\text{h}+1]}\left\{\max_{\text{m}\in[\text{M}]}\left\langle \phi(\cdot,\cdot), \text{w}_{\text{h}}^{\text{t,m,J}_\text{t}}\right\rangle\right\}$
  Define the actor loss $\tilde{\ell}^t_h(\theta)$ using Eq. (6)
  for $k = 1, \ldots, K_t$ **do**
    $\theta^{t,k}_h \leftarrow \theta^{t,k-1}_h - \alpha^t_{\boldsymbol{a}}\, \nabla_\theta \widetilde{\ell}^t_h(\theta^{t,k-1}_h)$
Return: actor parameters for the policy θ^t

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 $\{\mathcal{D}^t_h\}_{h\in[H]}$, where $\mathcal{D}^t_h$ consists of (s_h, a_h, s_{h+1}) tuples indexed as $\left\{(s^i_h, a^i_h, s^i_{h+1})\right\}_{i=1}^{|\mathcal{D}^t|}$ [2]. The critic

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

loss at episode t and step h uses the estimated value function at step $h+1$, and forms the following ridge regression problem:

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

For each step h, LMC 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 backward 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. Moreover, our proposed framework, instantiated by the projected NPG and LMC, has less space complexity as stated in the following remark.

Remark 5.1. *Existing works (e.g., Liu et al. [2023], Sherman et al. [2023], Cassel and Rosenberg [2024]) that use the standard* NPG *update ($\pi_h^{t+1}(\cdot|s) \propto \pi_h^1(\cdot|s) \exp(\eta \sum_{i=1}^t \widehat{Q}_h^i(s,\cdot))$) with UCB bonuses require storing $\mathcal{D}^t$ and all the historical Q-function for every previous iteration, resulting in the space complexity of $\mathcal{O}(THd)$. Our proposed framework, instantiated by the projected* NPG *and* LMC*, does not have such requirement, and only require $\mathcal{O}(TH + Hd)$, where $\mathcal{O}(TH)$ is for storing $\mathcal{D}^t$ and $\mathcal{O}(Hd)$ is for storing the* LMC *critic parameters.*

Algorithm 3 Critic: LMC

1: **Input**: collected data $\mathcal{D}^t$, policy π^{t-1}, number of critic updates J_t, critic learning rate $\alpha_{\mathbf{c}}^{h,t}$, 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: $w_h^{t,m,j} \leftarrow w_h^{t,m,j-1} - \alpha_{\mathbf{c}}^{h,t} \nabla_w \mathcal{L}_h^t(w_h^{t,m,j-1}) + \sqrt{\alpha_{\mathbf{c}}^{h,t}/\zeta}\, \nu_h^{t,m,j} \quad \forall m \in [M]$
9: $\widehat{Q}_h^t(\cdot,\cdot) = \mathrm{clip}_{[0,H-h+1]} \left\{ \max_{m \in [M]} \left\langle \phi(\cdot,\cdot), w_h^{t,m,J_t} \right\rangle \right\}$
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 $\{w_h^{t,m,J_t}\}_{(m,h) \in [M] \times [H]}$

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 $\iota_h^t(s,a) \coloneqq r_h(s,a) + \mathbb{P}_h \widehat{V}_{h+1}^t(s,a) - \widehat{Q}_h^t(s,a)$.*

The theoretical analyses in existing works [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, $\alpha_{\mathbf{c}}^{h,t}$, 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 $1-\delta$,

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

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 log 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 $\widetilde{\mathcal{O}}(1/\epsilon^4)$ in the on-policy setting, matching the result in Liu et al. [2023].

Theorem 6.1. *Under Assumptions 4.1 and 4.2, consider Algorithm 1 in the on-policy setting with the* LMC *critic (Algorithm 3) and the projected* NPG *actor (Algorithm 2). For an appropriate choice of the actor and critic parameters, $N = d_{\boldsymbol{c}}^3\, T/(H^2 \log|\mathcal{A}|)$ and $\delta \in (0,1)$, if $\bar{\epsilon}$ is the projection error in the actor, then, with probability at least $1-\delta$,*

$$\mathrm{OG(T)} \leq \widetilde{\mathcal{O}}\left(\frac{\mathrm{H}^2\sqrt{\log|\mathcal{A}|}}{\sqrt{\mathrm{T}}} + \mathrm{H}^2\sqrt{\bar{\epsilon}}\right).$$

Hence, for any $\epsilon > 0$, by setting $T = H^4 \log|\mathcal{A}|/\epsilon^2$, Algorithm 1 returns an $(\epsilon + H^2\sqrt{\bar{\epsilon}})$-optimal mixture policy, and therefore requires $T \times N = \widetilde{\mathcal{O}}(1/\epsilon^4)$ 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, the lower-bound in Lemma D.2 can be instantiated as $-\iota_h^t(s,a) \leq \mathcal{O}\left(\sqrt{d_{\boldsymbol{c}}^3\, H^4\, T\, \log^2(N/\delta)/N}\right)$ in the on-policy setting. Putting everything together with the chosen value of N leads to the stated sample complexity. □

6.2 Off-Policy Setting

Next, we show that, in the off-policy setting, Algorithm 1 can achieve $\widetilde{\mathcal{O}}(1/\epsilon^2)$ sample complexity, matching Sherman et al. [2023], Cassel and Rosenberg [2024].

Theorem 6.2 (Off-Policy Sample Efficiency)**.** *For an appropriate choice of the actor and critic parameters, $\delta \in (0,1)$, if $\bar{\epsilon}$ is the projection error in the actor, then, with probability at least $1-\delta$,*

$$\mathrm{OG(T)} \leq \widetilde{\mathcal{O}}\left(\frac{\mathrm{d_c^2}\,\max\{\mathrm{d_a},\mathrm{d_c}\}\,\mathrm{H}^2\sqrt{\log|\mathcal{A}|}}{\sqrt{\mathrm{T}}} + \mathrm{H}^2\,\sqrt{\bar{\epsilon}}\right).$$

Hence, for any $\epsilon > 0$, by setting $T = H^4\,\log|\mathcal{A}|/\epsilon^2$, Algorithm 1 returns a $(\epsilon + H^2\sqrt{\bar{\epsilon}})$-optimal mixture policy, and therefore requires $T \times 1 = \widetilde{\mathcal{O}}(1/\epsilon^2)$ 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. We use the uniform concentration argument in Jin et al. [2020] to obtain that

$$-\iota_h^t(s,a) \leq \mathcal{O}\left(\sqrt{d_{\boldsymbol{c}}^3\,H^4\,T\,\log(T/\delta)}\left[\sqrt{\log(\mathcal{N}_\Delta(\mathcal{V}))} + T^2\Delta^2\right]\right),$$

which involves a bound on $\log(\mathcal{N}_\Delta(\mathcal{V}))$, the log covering number of the value function class. The log-covering number is a measure of the complexity of the space of value functions. We show that for an actor with log-linear policies, we can easily bound the log 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 $\mathsf{V} := d_{\boldsymbol{c}}\,\log\left(1 + \frac{4\,\overline{W}+4\,H\,\sqrt{2\,\overline{Z}}}{\Delta}\right) + d_{\boldsymbol{a}}\,\log\left(1+\frac{4\,H\,\sqrt{2\,\overline{Z}}}{\Delta}\right)$.*

In particular, we can show $\overline{W} \leq \mathcal{O}(\sqrt{T})$ (Lemma D.7), and $\overline{Z} \leq \mathcal{O}(\bar{\epsilon}\,T)$ (Lemma F.1). Putting everything together and setting $\Delta = \mathcal{O}(1/T^2)$ yields that

$$-\iota_h^t(s,a) \leq \widetilde{\mathcal{O}}(\sqrt{d_{\boldsymbol{c}}^2\,\max\{d_{\boldsymbol{a}},d_{\boldsymbol{c}}\}\,H^4\,T}\,.$$

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

In order to control this log 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 learns a parametric policy at each iteration, the log covering number of the policy class is bounded by $\mathcal{O}(d_{\boldsymbol{a}}\log(T))$ without any bespoke tricks.

Furthermore, our proposed framework is also compatible with other policy optimization or policy evaluation methods. For the actor, we can replace the projected NPG with other policy mirror descent-based methods (e.g., SPMA [Asad et al., 2025]) that can provide a similar bound for the policy optimization error as in Lemma 4.1 (details in Appendix B.3). We can also instantiate the critic with the UCB bonuses that can provide a similar bound for the policy evaluation error as in Lemma 5.1. In this case, we can recover the same guarantees for sample complexity as Liu et al. [2023] for the on-policy setting and as Sherman et al. [2023], Cassel and Rosenberg [2024] for the off-policy setting.

7 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 framework can simultaneously achieve sample efficiency and practicality.

For future work, we aim to investigate the actor-critic methods in more practical 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 explore the practical implementations of the proposed method and evaluate their performance across standard benchmarks.

Acknowledgments

We would like to thank Xingtu Liu and Yunxiang Li for helpful feedback on the paper. 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

Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. *Advances in Neural Information Processing Systems*, 24, 2011.

Yasin Abbasi-Yadkori, Peter Bartlett, Kush Bhatia, Nevena Lazic, Csaba Szepesvari, and Gellért Weisz. Politex: Regret bounds for policy iteration using expert prediction. In *International Conference on Machine Learning*, pages 3692–3702. PMLR, 2019.

Milton Abramowitz and Irene A Stegun. *Handbook of mathematical functions with formulas, graphs, and mathematical tables*, volume 55. US Government Printing Office, 1948.

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

Alekh Agarwal, Yujia Jin, and Tong Zhang. 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, 2023.

Naman Agarwal, Syomantak Chaudhuri, Prateek Jain, Dheeraj Nagaraj, and Praneeth Netrapalli. Online target q-learning with reverse experience replay: Efficiently finding the optimal policy for linear mdps. *arXiv preprint arXiv:2110.08440*, 2021b.

Carlo Alfano and Patrick Rebeschini. Linear convergence for natural policy gradient with log-linear policy parametrization. *arXiv preprint arXiv:2209.15382*, 2022.

Carlo Alfano, Rui Yuan, and Patrick Rebeschini. A novel framework for policy mirror descent with general parameterization and linear convergence. *Advances in Neural Information Processing Systems*, 36:30681–30725, 2023.

Reza Asad, Reza Babanezhad Harikandeh, Issam H Laradji, Nicolas Le Roux, and Sharan Vaswani. Fast convergence of softmax policy mirror ascent. In *International Conference on Artificial Intelligence and Statistics*, pages 3943–3951. PMLR, 2025.

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

Shalabh Bhatnagar, Richard S Sutton, Mohammad Ghavamzadeh, and Mark Lee. Natural actor–critic algorithms. *Automatica*, 45(11):2471–2482, 2009.

Qi Cai, Zhuoran Yang, Chi Jin, and Zhaoran Wang. Provably efficient exploration in policy optimization. In *International Conference on Machine Learning*, pages 1283–1294. PMLR, 2020.

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

Semih Cayci, Niao He, and R. Srikant. Finite-time analysis of entropy-regularized neural natural actor-critic algorithm. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL `https://openreview.net/forum?id=BkEqk7pS1I`.

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

Gabriel Dulac-Arnold, Daniel Mankowitz, and Todd Hester. Challenges of real-world reinforcement learning. *arXiv preprint arXiv:1904.12901*, 2019.

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

Zuyue Fu, Zhuoran Yang, and Zhaoran Wang. Single-timescale actor-critic provably finds globally optimal policy. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=pqZV_srUVmK.

Scott Fujimoto, Herke Hoof, and David Meger. Addressing function approximation error in actor-critic methods. In *International conference on machine learning*, pages 1587–1596. PMLR, 2018.

Mudit Gaur, Amrit Singh Bedi, Di Wang, and Vaneet Aggarwal. On the global convergence of natural actor-critic with two-layer neural network parametrization. *arXiv preprint arXiv:2306.10486*, 2023.

Mudit Gaur, Amrit Bedi, Di Wang, and Vaneet Aggarwal. 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, 2024.

Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *International conference on machine learning*, pages 1861–1870. Pmlr, 2018.

Botao Hao, Tor Lattimore, Csaba Szepesvári, and Mengdi Wang. Online sparse reinforcement learning. In *International Conference on Artificial Intelligence and Statistics*, pages 316–324. PMLR, 2021.

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

Jiafan He, Heyang Zhao, Dongruo Zhou, and Quanquan Gu. Nearly minimax optimal reinforcement learning for linear markov decision processes. In *International Conference on Machine Learning*, pages 12790–12822. PMLR, 2023.

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

Haque Ishfaq, Qiwen Cui, Viet Nguyen, Alex Ayoub, Zhuoran Yang, Zhaoran Wang, Doina Precup, and Lin Yang. Randomized exploration in reinforcement learning with general value function approximation. In *International Conference on Machine Learning*, pages 4607–4616. PMLR, 2021.

Haque Ishfaq, Qingfeng Lan, Pan Xu, A Rupam Mahmood, Doina Precup, Anima Anandkumar, and Kamyar Azizzadenesheli. Provable and practical: Efficient exploration in reinforcement learning via langevin monte carlo. In *The Twelfth International Conference on Learning Representations*, 2024a.

Haque Ishfaq, Yixin Tan, Yu Yang, Qingfeng Lan, Jianfeng Lu, A Rupam Mahmood, Doina Precup, and Pan Xu. More efficient randomized exploration for reinforcement learning via approximate sampling. In *Reinforcement Learning Conference*, 2024b.

Haque Ishfaq, Guangyuan Wang, Sami Nur Islam, and Doina Precup. Langevin soft actor-critic: Efficient exploration through uncertainty-driven critic learning. In *The Thirteenth International Conference on Learning Representations*, 2025.

Chi Jin, Zhuoran Yang, Zhaoran Wang, and Michael I Jordan. Provably efficient reinforcement learning with linear function approximation. In *Conference on learning theory*, pages 2137–2143. PMLR, 2020.

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

Emilio Jorge, Christos Dimitrakakis, and Debabrota Basu. Isoperimetry is all we need: Langevin posterior sampling for rl with sublinear regret. *arXiv preprint arXiv:2412.20824*, 2024.

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

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

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

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

Vijay Konda and John Tsitsiklis. Actor-critic algorithms. *Advances in Neural Information Processing Systems*, 12, 1999.

Tor Lattimore and Csaba Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020.

Timothy P Lillicrap, Jonathan J Hunt, Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David Silver, and Daan Wierstra. Continuous control with deep reinforcement learning. *arXiv preprint arXiv:1509.02971*, 2015.

Qinghua Liu, Gellért Weisz, András György, Chi Jin, and Csaba Szepesvári. Optimistic natural policy gradient: a simple efficient policy optimization framework for online rl. *Advances in Neural Information Processing Systems*, 36:3560–3577, 2023.

Eric Mazumdar, Aldo Pacchiano, Yian Ma, Michael Jordan, and Peter Bartlett. On approximate thompson sampling with langevin algorithms. In *international conference on machine learning*, pages 6797–6807. PMLR, 2020.

Gergely Neu, Anders Jonsson, and Vicenç Gómez. A unified view of entropy-regularized markov decision processes. *arXiv preprint arXiv:1705.07798*, 2017.

Ian Osband, Yotam Doron, Matteo Hessel, John Aslanides, Eren Sezener, Andre Saraiva, Katrina McKinney, Tor Lattimore, Csaba Szepesvari, Satinder Singh, et al. Behaviour suite for reinforcement learning. *arXiv preprint arXiv:1908.03568*, 2019.

Jan Peters, Sethu Vijayakumar, and Stefan Schaal. Natural actor-critic. In *European Conference on Machine Learning*, pages 280–291. Springer, 2005.

John Schulman, Sergey Levine, Pieter Abbeel, Michael Jordan, and Philipp Moritz. Trust region policy optimization. In *International conference on machine learning*, pages 1889–1897. PMLR, 2015.

John Schulman, Xi Chen, and Pieter Abbeel. Equivalence between policy gradients and soft q-learning. *arXiv preprint arXiv:1704.06440*, 2017a.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017b.

Uri Sherman, Alon Cohen, Tomer Koren, and Yishay Mansour. Rate-optimal policy optimization for linear markov decision processes. *arXiv preprint arXiv:2308.14642*, 2023.

Richard S Sutton, David McAllester, Satinder Singh, and Yishay Mansour. Policy gradient methods for reinforcement learning with function approximation. *Advances in Neural Information Processing Systems*, 12, 1999.

Csaba Szepesvári. *Algorithms for reinforcement learning*. Springer nature, 2022.

Kevin Tan, Wei Fan, and Yuting Wei. Actor-critics can achieve optimal sample efficiency. *arXiv preprint arXiv:2505.03710*, 2025.

Michael J Todd. *Minimum-volume ellipsoids: Theory and algorithms*. SIAM, 2016.

Victor Uc-Cetina, Nicolás Navarro-Guerrero, Anabel Martin-Gonzalez, Cornelius Weber, and Stefan Wermter. Survey on reinforcement learning for language processing. *Artificial Intelligence Review*, 56(2):1543–1575, 2023.

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

Andrew J Wagenmaker, Yifang Chen, Max Simchowitz, Simon Du, and Kevin Jamieson. Reward-free rl is no harder than reward-aware rl in linear markov decision processes. In *International Conference on Machine Learning*, pages 22430–22456. PMLR, 2022.

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

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

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

Pan Xu, Hongkai Zheng, Eric V Mazumdar, Kamyar Azizzadenesheli, and Animashree Anandkumar. Langevin monte carlo for contextual bandits. In *International Conference on Machine Learning*, pages 24830–24850. PMLR, 2022.

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

Rui Yuan, Simon S Du, Robert M Gower, Alessandro Lazaric, and Lin Xiao. Linear convergence of natural policy gradient methods with log-linear policies. In *International Conference on Learning Representations*, 2023.

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

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

Contents of Appendix

A Notation

Notation for Problem Setting and Algorithm Design

Table 1: Notation for Problem Setting and Algorithm Design

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 probability
$\phi : \mathcal{S} \times \mathcal{A} \mapsto \mathbb{R}^{d_c}$	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
$\mathcal{D}^t, \mathcal{D}_h^t$	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
$\alpha_{\boldsymbol{c}}$	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
$\mathcal{D}_{\text{exp}}, \rho_{\text{exp}}$	subset of $\mathcal{S} \times \mathcal{A}$, distribution over $\mathcal{D}_{\text{exp}}$
$\theta \in \mathbb{R}^{d_a}$	learnable actor parameters
K	number of actor updates
$\alpha_{\boldsymbol{a}}$	actor learning rate
η	policy optimization learning rate

Additional Notation 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_h^t 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., $V^t := \{V_h^t\}_{h\in[H]}$). $|\mathcal{D}^t|$ represents the number of trajectories in $\mathcal{D}^t$ or the number of (s_h, a_h, s_{h+1}) tuples in $\mathcal{D}_h^t$. 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 Analyses for the Actor

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

Proof of Lemma 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 4.2)

Proof of Lemma 4.2. 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 $s \in \mathcal{S}$,

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

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

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

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

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

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

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

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 $(s,a) \in \mathcal{S} \times \mathcal{A}$,

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

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

Putting everything together, we have that

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

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 $s \in \mathcal{S}$,

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

This concludes the proof. □

B.3 Instantiating the Actor with SPMA

Lemma 4.2 can not only be applied to NPG but also other mirror descent-based policy optimization methods such as TRPO Schulman et al. (2015), AMPO (Alfano et al., 2023), 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 the following:

$$\begin{aligned}
\tilde{\ell}_h^t(\theta) &= \frac{1}{2} \sum_{(s,a) \in \mathcal{D}_{\text{exp}}} \rho_{\text{exp}}(s,a) \left[\langle \varphi(s,a), \theta \rangle - \widehat{Z}_h^t(s,a) \right]^2, \\
&\text{where } \widehat{Z}_h^t(s,a) \coloneqq \left\langle \varphi(s,a), \widehat{\theta}_h^t \right\rangle + \log\left(1 + \eta\, A^{\pi^t}(s,\cdot) \right).
\end{aligned}$$

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 Lemma 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 $\epsilon^t := \mathrm{KL}(u \,\|\, p^{t+1}) - \mathrm{KL}(u \,\|\, p^{t+1/2})$ 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

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

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

(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., $\langle u, v\rangle \le \|u\|_1 \|v\|_\infty \le \|u\|_1^2/2 + \|v\|_\infty^2/2$), and (vi) applies the Pinkster's inequality and $\|d^t\|_\infty \le H$.

Moreover, we have that

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

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

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

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

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

(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

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

This concludes the proof. □

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

Furthermore, we introduce the alternative version of Assumption 4.1.

Assumption B.1 (Bias). *Let φ be the policy feature. Then,*

$$\inf_\theta \sup_{t,h,s,a} \left|\langle \varphi(s,a), \theta\rangle - \log(1 + \eta\, A_h^{\pi^t}(s,a))\right| \le \epsilon_{\text{bias}} .$$

Therefore, under Assumptions 4.2 and B.1, we can easily prove that Lemma 4.2 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 $y, z \in \mathrm{int} X$,*

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

C Constructing $\mathcal{D}_{\text{exp}}$ via Experimental Design

In this section, we introduce various methods of experimental design to bound $\overline{\varphi}_G$ defined in Lemma 4.2. The experimental design problem can be written as

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

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*

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

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 4.2 as $\bar{\epsilon} \le \mathcal{O}(d_{\boldsymbol{a}}\,\epsilon_{\text{bias}} + \epsilon_{\text{opt}})$.

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 $\sup_{\pi \in \Pi} \lambda_{\min}\left(\mathbb{E}_{\pi}\left[\phi(s,a)\,\phi(s,a)^{\top}\right]\right) \geq \epsilon_{\mathcal{M}}$. Running K rounds of* `CoverTraj` *to collect $\mathcal{D}_{\text{exp}} = \{(s_h^\tau, a_h^\tau)\}_{\tau=1}^{K}$ where*

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

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 Analyses 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,*

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

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 where C_δ will be set to C_δ^{on} in Lemma E.1 and C_δ^{off} in Lemma F.2.

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

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.

$$\nabla L_h^t(w_h) = \Lambda_h^t w_h - b_h^t, \text{ where} \tag{11}$$
$$\Lambda_h^t \coloneqq \sum_{(s,a)\in\mathcal{D}_h^t} \phi(s,a)\,\phi(s,a)^\top + \lambda\, I\,,$$
$$b_h^t \coloneqq \sum_{(s,a,s')\in\mathcal{D}_h^t} \Big[r_h(s,a) + \widehat{V}_{h+1}^t(s')\Big]\phi(s,a)\,.$$

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 $(s,a)\in\mathcal{S}\times\mathcal{A}$,*

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

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*

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

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 $1-\delta$,*

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

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* $1 - \delta$*,*

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

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*

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

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 $\widehat{Q}_h^t(s,a) = \min\left\{\left\langle \phi(s,a), w_h^{t,m,J_t}\right\rangle, H - h + 1\right\}$. Since $r_h(s,a) + \mathbb{P}_h \widehat{V}_{h+1}^t(s,a) \leq H - h + 1$, when $\left\langle \phi(s,a), w_h^{t,m,J_t}\right\rangle > H - h + 1$, the statement is trivially true. Thus, we only need to consider the case when $\left\langle \phi(s,a), w_h^{t,m,J_t}\right\rangle \leq H - h + 1$ and thus $\widehat{Q}_h^t(s,a) = \left\langle \phi(s,a), w_h^{t,m,J_t}\right\rangle$.

Based on the mean and covariance matrix defined in Lemma D.3, we have that $\left\langle \phi(s,a), w_h^{t,m,J_t}\right\rangle$ 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 $X_t := \frac{r_h(s,a) + \mathbb{P}_h \widehat{V}_{h+1}^t(s,a) - \left\langle \phi(s,a), \mu_h^{t,J_t}\right\rangle}{\sqrt{\phi(s,a)^\top\, \Sigma_h^{t,J_t}\, \phi(s,a)}}$ and will next show that $|X_t| \leq 1$. First, we have that

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

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

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

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

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

$$
\begin{aligned}
&= \Pr\left(\frac{\left\langle \phi(s,a), w_h^{t,m,J_t} \right\rangle - \left\langle \phi(s,a), \mu_h^{t,J_t} \right\rangle}{\sqrt{\phi(s,a)^\top \, \Sigma_h^{t,J_t} \, \phi(s,a)}} \geq X_t \right) \\
&\geq \frac{1}{2\sqrt{2\pi}} \exp\left(-X_t^2/2\right) \\
&\geq \frac{1}{2\sqrt{2e\pi}} \, .
\end{aligned}
$$

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

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

This further implies that

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

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

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

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

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

$$
\begin{aligned}
-\iota_h^t(s,a) &= \widehat{Q}_h^t(s,a) - r_h(s,a) - \mathbb{P}_h \widehat{V}_{h+1}^t(s,a) \\
&= \min\left\{ \max_{m \in [M]} \left\langle \phi(s,a), w_h^{t,m,J_t} \right\rangle, H - h + 1 \right\}^+ - r_h(s,a) - \mathbb{P}_h \widehat{V}_{h+1}^t(s,a) \\
&\leq \max_{m \in [M]} \left\langle \phi(s,a), w_h^{t,m,J_t} \right\rangle - r_h(s,a) - \mathbb{P}_h \widehat{V}_{h+1}^t(s,a) \\
&= \max_{m \in [M]} \left\langle \phi(s,a), w_h^{t,m,J_t} \right\rangle - \left\langle \phi(s,a), \widehat{w}_h^t \right\rangle + \left\langle \phi(s,a), \widehat{w}_h^t \right\rangle - r_h(s,a) - \mathbb{P}_h \widehat{V}_{h+1}^t(s,a)
\end{aligned}
$$

$$\leq \left| \max_{m\in[M]} \left\langle \phi(s,a), w_h^{t,m,J_t} \right\rangle - \left\langle \phi(s,a), \widehat{w}_h^t \right\rangle \right| + \left| \left\langle \phi(s,a), \widehat{w}_h^t \right\rangle - r_h(s,a) - \mathbb{P}_h \widehat{V}_{h+1}^t(s,a) \right|$$

$$\overset{\text{(iii)}}{\leq} \left(C_\delta \, H \, \sqrt{d_{\boldsymbol{c}}} + \frac{4}{3} \sqrt{\frac{2\, d_{\boldsymbol{c}} \log\left(1/\delta\right)}{3\,\zeta}} + \frac{4}{3} \right) \|\phi(s,a)\|_{(\Lambda_h^t)^{-1}} ,$$

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

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

(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

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

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

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

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,

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

This further implies that

$$\begin{aligned}
\Sigma_h^{t,m,J_t} &= \sum_{i=1}^{t} \alpha_{\mathbf{c}}^i\,\zeta^{-1}\,\mathscr{A}_{i+1}\left(\sum_{l=0}^{J_i-1} A_i^{2l}\right)\mathscr{A}_{i+1}^\top \\
&= \sum_{i=1}^{t} \alpha_{\mathbf{c}}^i\,\zeta^{-1}\,A_t^{J_t} \ldots A_{i+1}^{J_{i+1}} \left(\sum_{l=0}^{J_i-1} A_i^{2l}\right) A_{i+1}^{J_{i+1}} \ldots A_t^{J_t} \\
&\overset{\text{(v)}}{=} \sum_{i=1}^{t} \alpha_{\mathbf{c}}^i\,\zeta^{-1}\,A_t^{J_t} \ldots A_{i+1}^{J_{i+1}} \left(I - A_i^{2J_i}\right)\left(I - A_i^2\right)^{-1} A_{i+1}^{J_{i+1}} \ldots A_t^{J_t} \\
&= \sum_{i=1}^{t} \alpha_{\mathbf{c}}^i\,\zeta^{-1}\,A_t^{J_t} \ldots A_{i+1}^{J_{i+1}} \left(I - A_i^{2J_i}\right)\left(\Lambda_h^i\right)\left(\Lambda_h^i\right)^{-1}(I - A_i)^{-1}\,(I + A_i)^{-1} A_{i+1}^{J_{i+1}} \ldots A_t^{J_t} \\
&\overset{\text{(vi)}}{=} \sum_{i=1}^{t} \zeta^{-1}\,A_t^{J_t} \ldots A_{i+1}^{J_{i+1}} \left(I - A_i^{2J_i}\right)\left(\Lambda_h^i\right)^{-1}(I + A_i)^{-1}\,A_{i+1}^{J_{i+1}} \ldots A_t^{J_t}\,.
\end{aligned}$$

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

This concludes the proof. □

D.2.2 Proof of Lemma D.4

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

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

This implies that

$$\begin{aligned}
&\left|\left\langle \phi(s,a), \left(\mu_h^{t,J_t} - \widehat{w}_h^t\right)\right\rangle\right| \\
&\quad = \phi(s,a)^\top A_t^{J_t} \ldots A_1^{J_1} \left(w_h^{1,m,0} - \widehat{w}_h^1\right) + \phi(s,a)^\top \sum_{i=1}^{t-1} A_t^{J_t} \ldots A_{i+1}^{J_{i+1}} \left(\widehat{w}_h^i - \widehat{w}_h^{i+1}\right)
\end{aligned}$$

$$
\begin{aligned}
&\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}} \left| \mathcal{D}^t \right| / \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(\left| \mathcal{D}^t \right| + 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(\left| \mathcal{D}^t \right| + 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(\left| \mathcal{D}^t \right| + 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}} \\
&\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) \|\phi(s,a)\|_{\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) \|\phi(s,a)\|_{\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) \|\phi(s,a)\|_{\left(\Lambda_h^t\right)^{-1}} \\
&= \left(\frac{1}{1-\sigma} \right) \|\phi(s,a)\|_{\left(\Lambda_h^t\right)^{-1}} \\
&\leq \frac{4}{3} \|\phi(s,a)\|_{\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 $A_i \prec (1 - \alpha_{\boldsymbol{c}}^{h,j} \lambda_{\min} \left(\Lambda_h^j \right)) I$ 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

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

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

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 trivial that A and $(\Lambda_h^t)^{-1}$ are commuting matrices. Hence,

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

(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,

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

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

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

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

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

$$
\begin{aligned}
&\left|\phi(s,a)^\top \mathscr{A}_{i+1}\left((\Lambda_h^i)^{-1} - (\Lambda_h^{i+1})^{-1}\right)\mathscr{A}_{i+1}^\top\,\phi(s,a)\right| \\
&\quad\le \left|\phi(s,a)^\top \mathscr{A}_{i+1}\,(\Lambda_h^i)^{-1}\,\mathscr{A}_{i+1}^\top\,\phi(s,a)\right| \\
&\qquad+ \left|\left\langle \phi(s,a), \mathscr{A}_{i+1}\,(\Lambda_h^{i+1})^{-1}\,\mathscr{A}_{i+1}^\top\,\phi(s,a)\right\rangle\right| \\
&\quad\le \left\|\phi(s,a)^\top \mathscr{A}_{i+1}\,(\Lambda_h^i)^{-1/2}\right\|^2 + \left\|\phi(s,a)^\top \mathscr{A}_{i+1}\,(\Lambda_h^{i+1})^{-1/2}\right\|^2 \\
&\quad= \prod_{j=i+1}^t \left(1 - \alpha_{\mathbf{c}}^{h,j}\,\lambda_{\min}(\Lambda_h^j)\right)^{2\,J_j}\left(\|\phi(s,a)\|^2_{(\Lambda_h^i)^{-1}} + \|\phi(s,a)\|^2_{(\Lambda_h^{i+1})^{-1}}\right) \\
&\quad\le 2\prod_{j=i+1}^t \left(1 - \alpha_{\mathbf{c}}^{h,j}\,\lambda_{\min}(\Lambda_h^j)\right)^{2\,J_j}\|\phi(s,a)\|_2^2\,,
\end{aligned}
$$

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

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

$$= \frac{1}{24\,\zeta}\,\|\phi(s,a)\|^2_{(\Lambda_h^t)^{-1}}\,,$$

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

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

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

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

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

$$
\begin{aligned}
&\overset{\text{(v)}}{\leq} 2\,H\,\sqrt{d_{\boldsymbol{c}}\,|\mathcal{D}^t|}\sum_{i=1}^{t}\prod_{j=i+1}^{t}\left(1-\alpha_{\boldsymbol{c}}^{h,j}\,\lambda_{\min}(\Lambda_h^j)\right)^{J_j}\left(\|I\|_2+\|A_i\|_2^{J_i}\right)\\
&\overset{\text{(vi)}}{\leq} 2\,H\,\sqrt{d_{\boldsymbol{c}}\,|\mathcal{D}^t|}\sum_{i=1}^{t}\prod_{j=i+1}^{t}\left(1-\alpha_{\boldsymbol{c}}^{h,j}\,\lambda_{\min}(\Lambda_h^j)\right)^{J_j}\left(1+\left(1-\alpha_{\boldsymbol{c}}^{i}\,\lambda_{\min}(\Lambda_h^i)\right)^{J_i}\right)\\
&\leq 2\,H\,\sqrt{d_{\boldsymbol{c}}\,|\mathcal{D}^t|}\sum_{i=1}^{t}\left(\prod_{j=i+1}^{t}\left(1-\alpha_{\boldsymbol{c}}^{h,j}\,\lambda_{\min}(\Lambda_h^j)\right)^{J_j}+\prod_{j=i}^{t}\left(1-\alpha_{\boldsymbol{c}}^{h,j}\,\lambda_{\min}(\Lambda_h^j)\right)^{J_j}\right)\\
&\overset{\text{(vii)}}{\leq} 2\,H\,\sqrt{d_{\boldsymbol{c}}\,|\mathcal{D}^t|}\sum_{i=1}^{t}\left(\prod_{j=i+1}^{t}\left(1-1/(2\,\kappa_j)\right)^{J_j}+\prod_{j=i}^{t}\left(1-1/(2\,\kappa_j)\right)^{J_j}\right),
\end{aligned}
$$

(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 $\sigma = 1/\left(4\,H\,(|\mathcal{D}^t|+1)\,\sqrt{d_{\boldsymbol{c}}}\right)$ and 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

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

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

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

To bound each term, we first have,

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

$$\leq d_{\boldsymbol{c}}\,\sigma \leq 1\,,$$

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,

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

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

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

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

$$\mathrm{Tr}\big((\Lambda_h^i)^{-1}\big) \leq d_{\boldsymbol{c}} \cdot 1 = d_{\boldsymbol{c}}\,.$$

Finally, we have that

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

Therefore, using Lemma D.11, we have that

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

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

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

This concludes the proof. □

D.2.6 Proof of Lemma D.8

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

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

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

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

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

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

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

$$\begin{aligned}
&\phi(s,a)^\top\,\Sigma_h^{t,J_t}\,\phi(s,a) \\
&\quad = \frac{1}{\zeta}\sum_{i=1}^{t} \phi(s,a)^\top\,A_t^{J_t}\ldots A_{i+1}^{J_{i+1}}\left(I - A^{2J_i}\right)\left(\Lambda_h^i\right)^{-1}(I+A_i)^{-1}\,A_{i+1}^{J_{i+1}}\ldots A_t^{J_t}\,\phi(s,a)
\end{aligned}$$

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

For (i), we use the denotation that $\mathscr{A}_{i+1} = A_t^{J_t} \dots A_{i+1}^{J_{i+1}}$. (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,

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

(iv) follows from the fact that $\sqrt{a+b} \leq a + b$ for all $a, b > 0$. (v) uses Lemma D.16 by setting $J_j \geq \kappa_j \log\left(1/\sigma\right)$ 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 $\sum_{i=1}^{t} \sigma^{t-i} \leq \sum_{i=1}^{\infty} \sigma^i \leq \sigma/(1-\sigma)$. Therefore, we have

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

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

This implies that

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

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

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

□

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

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

This further implies that

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

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

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

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

$$\overset{\text{(ii)}}{\leq} C_\delta\, H\,\sqrt{d_{\boldsymbol{c}}}\,\|\phi(s,a)\|_{(\Lambda_h^t)^{-1}}\,,$$

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).

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

(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 $\left\|\int_{s\in\mathcal{S}}\psi_h(s)\left(\widehat{V}_{h+1}^t(s)/H\right)\mathrm{d}\,s\right\|_2 \leq \sqrt{d_{\boldsymbol{c}}}$ (Definition 2.1).

Putting everything together and setting $\lambda = 1$, we have with probability at least $1-\delta$,

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

$$\leq \left(C_\delta\, H\, \sqrt{d_{\boldsymbol{c}}} + \sqrt{\lambda\, d_{\boldsymbol{c}}} + H\, \sqrt{\lambda\, d_{\boldsymbol{c}}}\right) \|\phi(s,a)\|_{(\Lambda_h^t)^{-1}}$$
$$= 3\, C_\delta\, H\, \sqrt{d_{\boldsymbol{c}}}\, \|\phi(s,a)\|_{(\Lambda_h^t)^{-1}}\,.$$

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

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

which 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*

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

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.

$$\begin{aligned}
&\left\| \sum_{(s,a,s') \in \mathcal{D}_h^t} \phi(s,a) \left[\widehat{V}_{h+1}^t(s') - \mathbb{P}_h \widehat{V}_{h+1}^t(s,a) \right] \right\|_{(\Lambda_h^t)^{-1}} \\
&\qquad \leq \sqrt{2}\, H \sqrt{\log\left[\frac{\det(\Lambda_h^t)^{1/2}\det(\Lambda_h^0)^{-1/2}}{\delta}\right]} \\
&\qquad = \sqrt{2}\, H \sqrt{\log\left[\left(\frac{N+\lambda}{\lambda}\right)^{d/2}\right] - \log(\delta)} \\
&\qquad = \sqrt{2}\, H \sqrt{\frac{d_{\boldsymbol{c}}}{2}\log(N/\delta)} \\
&\qquad = H \sqrt{d_{\boldsymbol{c}}\,\log(N/\delta)}\,,
\end{aligned}$$

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

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

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

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

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

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

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

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

(i) follows from Lemma 4.1 with $u = \pi_h^\star(\cdot \mid s)$. (ii) is obtained by setting $\eta = \frac{\sqrt{2\,(\log|\mathcal{A}| + \overline{\epsilon}\,T)}}{H\sqrt{T}}$. (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 error.

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

(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 $\{\pi^1,\ldots,\pi^t\}$.) We then index each data point in $\mathcal{D}_h^t$ as $\left\{(s_h^i,a_h^i,s_{h+1}^i)\right\}_{i\in[N]}$. Let $\Lambda_h^{t,i} = \left(\sum_{j=1}^{i}\phi(s_h^j,a_h^j)\,\phi(s_h^j,a_h^j)^\top + \lambda I\right)$. Then, we have that

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

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

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

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

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

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

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

Putting everything together. Therefore, we have that

$$\begin{aligned}
&\sum_{h=1}^{H}\mathbb{E}_{\pi^t}\left[\|\phi(s,a)\|_{(\Lambda^t_h)^{-1}}\right] \\
&\quad= \frac{1}{N}\sum_{i=1}^{N}\sum_{h=1}^{H}\left\|\phi(s^i_h,a^i_h)\right\|_{(\Lambda^{t,i}_h)^{-1}} + \frac{1}{N}\sum_{i=1}^{N}\sum_{h=1}^{H}\mathcal{M}^{\text{on}}_{i,h} \leq \mathcal{O}\left(\sqrt{\frac{d_{\mathbf{c}}\,H^2\log(N/\delta)}{N}}\right).
\end{aligned}$$

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

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

where (ix) comes from setting $N = d^3_{\mathbf{c}}\,T/\log|\mathcal{A}|$.

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

$$\mathbb{E}\Big[V^{\star}_1(s_1) - V^{\overline{\pi}^T}_1(s_1)\Big] = \frac{1}{T}(\text{Term (I)} + \text{Term (II)}) = \widetilde{\mathcal{O}}\left(\frac{H^2\sqrt{\log|\mathcal{A}|}}{\sqrt{T}} + H^2\,\sqrt{\epsilon}\right).$$

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,*

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

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*

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

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 4.2.

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

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

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

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

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 $s \in \mathcal{S}$,

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

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

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

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

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

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*

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

where

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

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 $1-\delta$,

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

$$\leq \left(4\,H^2\left[\frac{d_{\boldsymbol{c}}}{2}\log\left(\frac{T+\lambda}{\lambda}\right)+d_{\boldsymbol{c}}\,\log\left(\frac{\mathcal{N}_\Delta(\mathcal{V})}{\Delta}\right)+\log\frac{2}{\delta}\right]+\frac{8\,T^2\Delta^2}{\lambda}\right)^{1/2}$$
$$\leq 2H\left[\frac{d_{\boldsymbol{c}}}{2}\log\left(\frac{T+\lambda}{\lambda}\right)+d_{\boldsymbol{c}}\,\log\left(\frac{\mathcal{N}_\Delta(\mathcal{V})}{\Delta}\right)+\log\frac{2}{\delta}\right]^{1/2}+\frac{2\sqrt{2}T\Delta}{\sqrt{\lambda}}\,.$$

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

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

where the last inequality uses Lemma 6.1 to bound the log 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

$$\Gamma_{\mathsf{LMC}}^{\mathrm{off}}=C_\delta^{\mathrm{off}}\,H\,\sqrt{d_{\boldsymbol{c}}}+\frac{4}{3}\sqrt{\frac{2\,d_{\boldsymbol{c}}\log(1/\delta)}{3\,\zeta}}+\frac{4}{3}\leq\widetilde{\mathcal{O}}(H\,d_{\boldsymbol{c}})\,.$$

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

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

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

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

(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 $\Gamma_{\mathsf{LMC}}^{\mathrm{off}}$ is bounded, it suffices to bound $\mathbb{E}_{\pi^t}\left[\|\phi(s,a)\|_{(\Lambda_h^t)^{-1}}\right]$. 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

$$\sum_{t=1}^{T}\sum_{h=1}^{H}\mathbb{E}_{\pi^t}\left[\|\phi(s,a)\|_{(\Lambda_h^t)^{-1}}\right]$$

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

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

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

(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

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

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

$$\text{Term (II)} \leq \Gamma_{\texttt{LMC}}^{\text{off}}\sum_{t=1}^{T}\sum_{h=1}^{H}\mathbb{E}_{\pi^t}\left[\left\|\phi(s,a)\right\|_{(\Lambda_h^t)^{-1}}\right] \leq \widetilde{\mathcal{O}}\left(\sqrt{d_{\boldsymbol{c}}^2\,\max\{d_{\boldsymbol{c}}, d_{\boldsymbol{a}}\}\,H^4\,T}\right).$$

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

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

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*

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

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* $\|\phi_t\| \leq 1$. *Let* $\Lambda_t = I + \sum_{s=1}^{t} \phi_s\, \phi_s^{\top}$. *Assume* $\mathcal{V}$ *is a value function class such that* $\sup_{s\in\mathcal{S}} |V(s)| \leq H$. *Then, for any* $\delta > 0$*, with probability at least* $1-\delta$*, for all* $t \geq 0$ *and any* $V \in \mathcal{V}$*, it holds that*

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

where $\mathcal{N}_{\Delta}$ *is 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*

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

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*

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

G Experiments

In this section, we evaluate our actor-critic algorithm on a Random MDP and a linear MDP version of the Deep Sea (Osband et al., 2019) in the off-policy setting.

G.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 actor-critic framework on a Random MDP and a linear MDP version of the Deep Sea (Osband et al., 2019). In particular, we use the linear MDP features as the policy features, and $d := d_{\boldsymbol{c}} = d_{\boldsymbol{a}}$ represents the feature dimension for both the actor and the critic parameters.

For the Deep Sea environment, we use a $N \times N$ grid with $N = 10$ where the agent always starts at $(0, 0)$ 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 for $d = |\mathcal{S}| \times |\mathcal{A}|$. Given the true transition probabilities and rewards, and following the linear MDP assumption in Definition 2.1, we solve for ψ_h and $v_h^{\star}$ via least squares and obtain the corresponding $\mathbb{P}_h$ and r_h.

For the Random MDP environment, we consider 15 states and 5 actions and set $d = 30$. 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. Following Definition 2.1 and using least squares (similar to Deep Sea), we obtain the probability transitions. 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 state and taking action 1. All other state-action pairs yield zero reward. Using the same procedure for Deep Sea, we compute r_h and ensure linearity of the MDP.

G.2 Coreset Construction

To construct the coreset, we follow the offline G-experimental design outlined in Algorithm 4. 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 via G-Experimental Design

Input: features $\varphi : \mathcal{S} \times \mathcal{A} \mapsto \mathbb{R}^{d_{\boldsymbol{a}}}$, threshold $\epsilon_G \in \mathbb{R}$
Initialize: $G = I_{d_{\boldsymbol{a}} \times d_{\boldsymbol{a}}}$, $\mathcal{D}_{\text{exp}} = \emptyset$, $g_{\max} = \infty$
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.3 Hyperparameters

In Table 2, we list the hyperparameters we tested across all experiments. For log-linear policies, the actor loss in Eq. (6) admits a closed-form solution, allowing us to avoid tuning $\alpha_{\boldsymbol{a}}$ and K_t by minimizing the objective exactly. We swept the hyperparameters and picked the best combination for each of the considered methods to report the results.

Table 2: Hyperparameter sweep in our experiments.

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 ($\alpha_{\boldsymbol{c}}$)	$[10^{-2}, 10^{-3}, 10^{-4}, 10^{-5}]$		
Number of Episodes (T)	600		
Horizon Length (H)	100		

G.4 Experimental Results

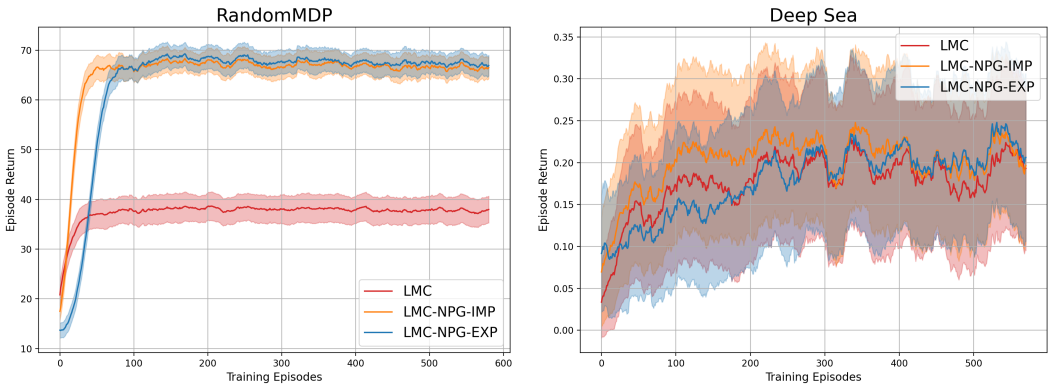

Figure 1: Comparison of `LMC-NPG-EXP` (our proposed framework), `LMC-NPG-IMP` (memory-intensive variant), and `LMC` (value-based baseline) in the Random MDP and the Deep Sea.

We denote by `LMC-NPG-EXP` our proposed framework with an explicit log-linear policy parameterization that uses `LMC` for policy evaluation and projected `NPG` for policy optimization on a coreset. We denote by `LMC-NPG-IMP` an idealized implicit variant of `NPG` that does not have an explicit actor parameterization and maintains an implicit policy by storing all parameterized Q functions (and hence requires significantly more memory). As a baseline, we include the value-based algorithm `LMC` (Ishfaq et al., 2024a). Following the protocol of Ishfaq et al. (2024a), each algorithm is run with 20 random seeds. We sweep hyperparameters as discussed in Appendix G.3 and report the best performance with 95% confidence intervals.

For the random linear MDP, Figure 1(left) indicates that LMC-NPG-EXP closely matches LMC-NPG-IMP while outperforming the value-based baseline, LMC. For the Deep Sea, Figure 1(right) showcases that LMC-NPG-EXP can achieve comparable performance with LMC-NPG-IMP and LMC.

G.5 Additional Results

Ablation on Feature Dimensions. When using LMC-NPG-EXP, which employs LMC for policy evaluation to optimize an explicitly parameterized log-linear policy over a coreset, we study the impact of the feature dimension ($d := d_{\boldsymbol{c}} = d_{\boldsymbol{a}}$). The results in Figure 2 show that larger feature dimensions d for both the actor and the critic can substantially improve the performance of the proposed framework.

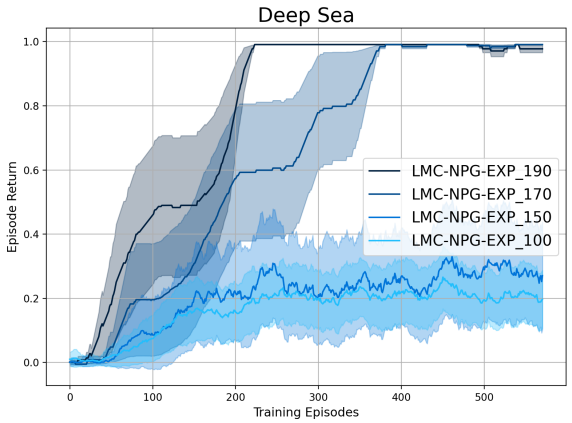

Figure 2: Effect of feature dimension d in the Deep Sea..