Offline Actor-Critic for Average Reward MDPs

William Powell
Department of Mathematics
University of Wisconsin-Madison
Madison, WI 53706
wgpowell@wisc.edu

Jeongyeol Kwon
Wisconsin Institute for Discovery
Madison, WI 53706
jeongyeol.kwon@wisc.edu

Qiaomin Xie
Department of Industrial and Systems Engineering
University of Wisconsin-Madison
Madison, WI 53706
qiaomin.xie@wisc.edu

Hanbaek Lyu
Department of Mathematics
University of Wisconsin-Madison
Madison, WI 53706
hlyu@math.wisc.edu

Abstract

We study offline policy optimization for infinite-horizon average-reward Markov decision processes (MDPs) with large or infinite state spaces. Specifically, we propose a pessimistic version of actor-critic methods using a computationally efficient linear function class for value function estimation. At the core of our method is a critic that computes a pessimistic estimate of the average reward under the current policy, as well as the corresponding policy gradient, by solving a fixed-point Bellman equation, rather than solving a successive sequence of regression problems as in finite horizon settings. Due to the nature of our policy-based method, the critic only needs to solve a linear optimization problem with convex quadratic constraints. We show that a very mild data coverage requirement is sufficient for our algorithm to achieve $O(\varepsilon^{-2})$ sample complexity for learning a near-optimal policy up to model misspecification errors. To our knowledge, this is the first result with optimal ε dependence in the offline average reward setting.

1 Introduction

Reinforcement learning (RL) is a sequential decision making framework commonly studied in the online setting where an agent attempts to learn an optimal policy through active interactions with its environment. However, in many relevant applications such as autonomous driving and health care, online learning can be intractable or dangerous [Tang and Wiens, 2021]. In such cases, it is common to resort to *offline* learning, where the agent's goal is to learn a near-optimal policy from a static data set which was collected in a manner known to be safe and efficient for the application.

Without the ability to actively explore the environment, the agent's capability to learn is subject to the quality of the collected data. In particular, it is well known that effective learning is difficult or impossible if the data set does not sufficiently cover the space of states and actions. Thus, a major challenge in offline RL is the design of algorithms with provable guarantees under the weakest possible data coverage requirements. Initial work [Antos et al., 2007, Munos and Szepesvári, 2008] rely on strong uniform coverage assumptions that effectively require the data generating policy to visit the entire state-action space. However, this assumption is often unreasonable, especially for large or infinite state spaces. Accordingly, more recent work [Rashidinejad et al., 2021, Zanette et al., 2021, Zhan et al., 2022, Hong and Tewari, 2024, Li et al., 2024, Gabbianelli et al., 2024, Neu and Okolo, 2025] have focused on provable guarantees depending only on partial coverage conditions

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

where the data set is only required to cover the subset of state-action pairs visited by an optimal policy.

State spaces in contemporary applications of RL can be very large or even infinite. Such applications necessitate the use of some form of function approximation for computational and memory efficient representation of policies and value functions. Empirical results show the promise of both simple linear models as well as more complex forms of function approximation such as neural networks, but there are still important theoretical gaps to be filled. This is especially true for infinite horizon MDPs, where value functions are solutions to a fixed point Bellman equation. As such, classical backward induction techniques that work well for episodic MDPs do not apply.

Our particular interest in this paper is offline RL for infinite horizon average reward MDPs (AMDPs) with linear function approximation. AMDPs are appropriate for continuing tasks such as inventory management [Giannoccaro and Pontrandolfo, 2002] or admission control [Weber et al., 2024], where there is no forced reset as in the episodic setting, or discounting that may lead to a myopic focus on short-term rewards. It is commonly acknowledged that theoretical analysis of algorithms for AMDPs is more challenging than that in the episodic or discounted setting. There are two primary reasons. First, as mentioned earlier, the techniques for episodic cases are not applicable here. Second, unlike discounted MDPs, the Bellman operator for AMDPs is not a contraction. Consequently, methods for discounted MDPs that rely on this contraction property do not carry over to the average reward setting.

In recent years, there have been a number of works investigating algorithms for AMDPs in the online setting [Fruit et al., 2018, Wei et al., 2020, 2021, Hao et al., 2021, Zhang and Xie, 2023, Agrawal and Agrawal, 2025, Hong et al., 2025], and learning from a generative model [Jin and Sidford, 2020, 2021, Zurek and Chen, 2024], as well as from deep RL and optimization perspectives [Zhang and Ross, 2021, Suttle et al., 2023, Agnihotri et al., 2024, Bai et al., 2024]. However, our understanding of the offline setting remains limited, especially with function approximation. The only work we know of is Primal-Dual Offline RL (PDOR) [Gabbianelli et al., 2024], which only achieves a sub-optimal $O(\varepsilon^{-4})$ sample complexity guarantee. Furthermore, PDOR's theoretical results require the strong assumption that the MDP's rewards and transition obey an exact linear structure, which rarely holds in practice. Motivated by this gap in the literature, the main question we ask in this work is the following:

> *Can we design a provably efficient algorithm for offline reinforcement learning in average reward MDPs with function approximation under minimal assumptions?*

We make strides towards an affirmative answer to this question by designing a pessimistic actor-critic algorithm using linear function approximation with a known feature map. Our key observation is that under our policy-based approach, a pessimistic estimate of the Bellman operator's fixed point can be computed by solving a simple linear optimization problem with convex quadratic constraints. We show our algorithm achieves optimal convergence rate guarantees with dependence on a measurement of data coverage used in prior work [Zanette et al., 2021, Gabbianelli et al., 2024, Neu and Okolo, 2025], which is referred to as the *feature coverage ratio*. This quantity is a measurement of how well the dataset aligns with the expected feature vector when following an optimal policy π^*. Our result's dependence on the intrinsic quantities improves over on the best known result for both average reward *and* discounted MDPs [Neu and Okolo, 2025]. Importantly, *we do not require the MDP itself to obey any linear structure*; our results hold under the more general conditions of approximate realizability and Bellman closedness. Furthermore, we only require the transition dynamics to satisfy a mild requirement that is significantly weaker than the uniform ergodicity assumption commonly considered for AMDPs [Wei et al., 2020, 2021, Bai et al., 2024]. Under these conditions, we show that $\tilde{O}(\varepsilon^{-2})$ samples are sufficient to learn a policy which is ε-optimal up to a model misspecification error.

1.1 Additional Related Work

Aside from the previously mentioned work [Zanette et al., 2021], another paper [Jin et al., 2021] also studies offline RL with linear function approximation in episodic MDPs. Their algorithm is a form of pessimistic least-squares value iteration, where pessimism is enforced through an additive bonus as commonly adopted in the online setting. It is not clear, however, whether this form of value iteration can be generalized to infinite horizon MDPs for the reasons discussed in the introduction.

Another line of related work studies offline RL in discounted MDPs with partial coverage and general function approximation [Xie et al., 2021, Cheng et al., 2022]. They also consider approximate realizability and Bellman closedness conditions similar to ours. The information theoretic results by [Xie et al., 2021] are near optimal, but due to their generality, theoretically optimal implementation of these algorithms is intractable. A computationally efficient alternative is presented in [Xie et al., 2021], but with sub-optimal convergence guarantees even when specialized to the linear setting. Similarly, convergence rates from the work [Cheng et al., 2022] are only of order $N^{-1/3}$ for N data samples.

Finally, most closely related to this work are the papers by [Zanette et al., 2021, Gabbianelli et al., 2024, Hong and Tewari, 2024, Neu and Okolo, 2025]. We provide a detailed comparison with these works in Section 5.

2 Notation

For any set $\mathcal{X}$, we let $\Delta(\mathcal{X})$ be the set of all probability distributions on $\mathcal{X}$. Given a state space $\mathcal{S}$, action space $\mathcal{A}$, transition kernel $P : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$, and (stationary) policy $\pi : \mathcal{S} \to \Delta(\mathcal{A})$, the notation $\mathbb{E}_s^\pi[\cdot]$ is the expectation with respect to a Markov chain with transition kernel $P^\pi(s, s') = \sum_a \pi(a|s)P(s'|s, a)$ conditioned on starting in state $s \in \mathcal{S}$. For a function $f \in \mathbb{R}^{\mathcal{S}}$, we also use the notation $P_{s,a}f$ as short hand for the conditional expectation $\mathbb{E}_{s'\sim P(\cdot|s,a)}[f(s')]$. For a policy π and probability measure $\mu \in \Delta(\mathcal{S})$, $\mu \otimes \pi$ is the probability measure on $\mathcal{S} \times \mathcal{A}$ defined by $(\mu \otimes \pi)(s, a) = \mu(s)\pi(a|s)$. Finally, given a symmetric, positive definite matrix $A \in \mathbb{R}^{d\times d}$, $\|\mathbf{x}\|_A$ denotes the norm on $\mathbb{R}^d$ defined by $\sqrt{\mathbf{x}^\top A\mathbf{x}}$.

3 Preliminaries

3.1 Average Reward MDPs

Let $\mathcal{S}$ be a large or possibly infinite state space and $\mathcal{A}$ be a finite space of A actions. We consider an infinite horizon average reward MDP $(\mathcal{S}, \mathcal{A}, P, r)$ with transition kernel $P : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ and reward function $r : \mathcal{S} \times \mathcal{A} \to [0, 1]$. An agent's rule for decision making in the MDP is specified by a (stationary) policy $\pi : \mathcal{S} \to \Delta(\mathcal{A})$ that maps current states to distributions over actions. At each time step t, the agent observes state s_t, takes action $a_t \sim \pi(\cdot|s_t)$, receives reward $r(s_t, a_t)$, then transitions to state $s_{t+1} \sim P(\cdot|s_t, a_t)$. The agents goal is to find a policy maximizing the average reward, which is defined as follows:

$$J^\pi(s) := \lim_{T\to\infty} \frac{1}{T}\mathbb{E}_s^\pi \left[\sum_{t=0}^{T-1} r(s_t, a_t)\right]. \tag{1}$$

For each policy π, we define its associated Bellman operator $T^\pi : \mathbb{R}^{\mathcal{S}\times\mathcal{A}} \to \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$ as

$$T^\pi f(s, a) = r(s, a) + \mathbb{E}_{s'\sim P(\cdot|s,a),a'\sim\pi(\cdot|s')}[f(s, a)], \qquad \forall s \in \mathcal{S}, a \in \mathcal{A}.$$

We will consider the following assumption throughout the paper.

Assumption 3.1. *All (stationary) policies induce a Markov chain that contains a single recurrent class and possibly some transient states (unichain). This implies that $J^\pi(s)$ is a constant independent of the initial state. Furthermore, it implies that for each policy π there exists a function $q^\pi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, unique up to linear translations, satisfying the Bellman equation*

$$q^\pi(s, a) + J^\pi = T^\pi q^\pi(s, a), \qquad \forall s \in \mathcal{S}, a \in \mathcal{A}, \tag{2}$$

which we will call the q-function. In this case, $v^\pi(s) = \mathbb{E}_{a\sim\pi(\cdot|s)}[q^\pi(s, a)]$ is called the value function. We also assume there exists a constant c such that

$$\sup_\pi \|q^\pi\|_{sp} \le c,$$

where $\|q^\pi\|_{sp} = \sup_{s,a} q^\pi(s, a) - \inf_{s,a} q^\pi(s, a)$ is the span semi-norm.

Under Assumption 3.1, for each policy π, there exists a unique stationary measure μ^π satisfying the equation

$$\sum_s \mu^\pi(s) \sum_a \pi(a|s) P(s'|s,a) = \mu^\pi(s'),$$

and we can write

$$J^\pi = \mathbb{E}_{(s,a)\sim\mu^\pi\otimes\pi}[r(s,a)].$$

As will be seen in the sequel, it is crucial for our algorithm that J^π is constant and Assumption 3.1 is sufficient to guarantee this. This assumption also allows us to reliably estimate a bounded q-function to (2), which is necessary for stability in the policy improvement step of our algorithm.

We note that Assumption 3.1 is stronger than the weakly communicating assumption often considered in the online learning literature [Jaksch et al., 2010, Fruit et al., 2018, Wei et al., 2021, Hong et al., 2025]. However, it is important to point out one crucial fact. This assumption does *not* imply exploratory conditions such as uniformly lower bounded stationary measures [Wei et al., 2020] or a uniformly excited features condition [Hao et al., 2021, Wei et al., 2021]. These exploratory conditions imply that to cover the states visited by any policy, it is necessary to cover the entire state space. For more details on this, we refer the reader to Appendix A where we provide a more thorough discussion on the implications of Assumption 3.1 and comparison with other average reward models from the literature.

3.2 Offline RL

In offline RL, the agent only has access to a pre-collected data set $\mathcal{D} = \{(s_i, a_i, r_i, s_i')\}_{i=1}^N$, where $r_i = r(s_i, a_i)$ and each s_i' is sampled from the conditional distribution $P(\cdot|s_i, a_i)$ independently of everything else. We do not need any additional assumptions about the collection of the data in $\mathcal{D}$. As in [Zanette et al., 2021, Neu and Okolo, 2025], the data does not need to be generated from i.i.d. sampling for from following a fixed behavior policy.

3.3 Function Approximation

Function approximation is necessary for learning in MDPs with large state and action spaces, where tabular solution methods are intractable. For actor-critic methods, this typically means using one function class Π to represent policies, and another class $\mathcal{F}$ for value function estimation. Effective learning in the MDP then becomes highly dependent on the ability to accurately represent the true value functions for a given policy $\pi \in \Pi$ using functions in $\mathcal{F}$. To quantify this ability, we introduce the following definition.

Definition 3.2. *Let $\mathcal{F} \subset \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$ be a set of functions used for value function approximation and $\Pi \subset \{\pi : \mathcal{S} \to \Delta(\mathcal{A})\}$ a class of stationary policies.*

(i) We say that $(\mathcal{F}, \Pi)$ satisfies the approximate realizability property with constant $\kappa_{\mathcal{F},\Pi}$ if

$$\sup_{\pi\in\Pi} \inf_{g\in\mathcal{F}, |\lambda|\le 1} \|g + \lambda - T^\pi g\|_\infty \le \kappa_{\mathcal{F},\Pi}. \tag{3}$$

(ii) The tuple $(\mathcal{F}, \Pi)$ is said to satisfy the Bellman-restricted closedness property with constant $\varepsilon_{\mathcal{F},\Pi}$ if

$$\sup_{\pi\in\Pi, f\in\mathcal{F}, |\lambda|\le 1} \inf_{g\in\mathcal{F}} \|g + \lambda - T^\pi f\|_\infty \le \varepsilon_{\mathcal{F},\Pi}.$$

These definitions are average-reward-analogues of those for episodic MDPs [Zanette et al., 2021, Nguyen-Tang and Arora, 2023]. Similar notions appeared in discounted MDPs in Xie et al. [Xie et al., 2021], although our use of the ℓ_∞ norm is slightly stronger than their requirement.

The approximate realizability property states that $\mathcal{F}$ nearly contains q^π for each policy $\pi \in \Pi$. If $\kappa_{\mathcal{F},\Pi} = 0$ then the infimum in (3) is attained by $(g, \lambda) = (q^\pi, J^\pi)$.

However, realizability alone is known to be insufficient for sample efficient learning [Wang et al., 2021]. Therefore, additional conditions are often required. Restricted closedness measures how well

we can perform regression using functions $g \in \mathcal{F}$ when the target is the function resulting from the application of T^π to a function $f \in \mathcal{F}$ plus a reasonable estimate λ of J^π. The addition of λ here is a generalization inspired, in part, by the requirement for linear MDPs that the column span of the feature matrix contains the all-one vector [Wei et al., 2021, Gabbianelli et al., 2024], as well as the development of generalized advantage estimation for the average reward setting in [Zhang and Ross, 2021] which includes a monte-carlo estimate of J^π as part of the regression target.

With these definitions, we can now introduce the function and policy classes considered for our algorithm.

Assumption 3.3. *We consider the use of a linear function class*

$$\mathcal{Q}(B_w) := \left\{ q(s,a) = \phi(s,a)^\top \boldsymbol{w} : \|\boldsymbol{w}\|_2 \leq B_w \right\}$$

where B_w is a user-defined parameter and $\phi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$ is a known d-dimensional feature map with $\|\phi(s,a)\|_2 \leq 1$. We also assume a softmax policy class

$$\Pi := \left\{ \pi(a|s) = \frac{e^{\phi(s,a)^\top \boldsymbol{\theta}}}{\sum_{a'} e^{\phi(s,a')^\top \boldsymbol{\theta}}} : \boldsymbol{\theta} \in \mathbb{R}^d \right\}. \tag{4}$$

We assume that $(\mathcal{Q}(B_w), \Pi)$ satisfies the approximate completeness and Bellman-restricted closedness properties with constants $\kappa_{\mathcal{Q}(B_w),\Pi}$ and $\varepsilon_{\mathcal{Q}(B_w),\Pi}$ respectively.

This assumption is a generalization of the widely studied linear MDP model, which is first introduced for episodic MDPs [Jin et al., 2020] and adapted to the average reward setting [Wei et al., 2021, Hong et al., 2025, Gabbianelli et al., 2024] with the realizability and restricted closedness constants being zero. Note that if the value functions are unbounded, it is unreasonable to assume approximate realizability. One cannot expect the ability to approximate functions in an unbounded set up to uniform error with an bounded function class. This is why we need the bounded span requirement in Assumption 3.1. However, we do not require prior knowledge of the constant c in Assumption 3.1 for our results. Our theoretical guarantees remain true for any choice of B_w provided Assumptions 3.1 and 3.3 hold.

4 Algorithm Details

Our algorithm is a form of pessimistic actor-critic method for the average reward setting. At a high level, it works through an alternating scheme run for a total number of K iterations. First, at iteration k, given π_k, the pessimistic critic first computes the smallest plausible average reward of π_k within a confidence region determined by the dataset $\mathcal{D}$. Then the actor updates the policy to π_{k+1} through a conservative policy improvement step.

4.1 Pessimistic Policy Evaluation

Before giving a more precise description, we start with a motivating discussion of a natural idea inspired by methods in the episodic setting, but which turns out not to work in our setting. Assume for a moment that we have completed k iterations of the algorithm, and we have an estimate J_k of the average reward J^{π_k} and an estimate $\hat{v}_k$ of the value function for policy π_k. Then by the Bellman equation (2), it is natural to estimate the q-function q^{π_k} by solving the ridge regression problem

$$\boldsymbol{w}_k \in \operatorname*{arg\,min}_{\boldsymbol{w} \in \mathbb{R}^d} \left[\sum_{i=1}^N \left(\phi(s_i,a_i)^\top \boldsymbol{w} + J_k - r_i - \hat{v}_k(s_i') \right)^2 + \|\boldsymbol{w}\|_2^2 \right],$$

which has the closed form expression

$$\boldsymbol{w}_k = \hat{\Lambda}^{-1} \sum_{i=1}^N \phi(s_i,a_i) \left(r_i - J_k + \hat{v}_k(s_i') \right). \tag{5}$$

Here, $\hat{\Lambda}$ is the un-normalized empirical covariance matrix

$$\hat{\Lambda} = \sum_{i=1}^N \phi(s_i,a_i)\phi(s_i,a_i)^\top + I,$$

and I is the d-dimensional identity matrix. Let $\hat{q}_k(s,a) = \phi(s,a)^\top \boldsymbol{w}_k$ be our estimate of q^{π_k}. This method is similar to the typical approach for episodic MDPs, where the value function of step $h+1$ is estimated and then used as a regression target to compute the weight vector for step h.

There are two main issues with this approach, however, in our AMDP setting. The first issue is about estimating J^{π_k} in AMDPs, which we will address shortly. The second issue, also shared with discounted MDPs, is that the Bellman equation in our case is a fixed point equation. Consequently, we cannot use backward induction techniques from episodic MDPs. Therefore, it is unclear how to construct the estimated value $\hat{v}_k$ if we haven't already obtained an estimated Q-function $\hat{q}_k$ without resorting to Monte-Carlo methods. However, as pointed out in prior work [Zanette et al., 2021], Monte-Carlo estimation is undesirable in the offline setting: using importance sampling weights to cancel the distribution mismatch requires some knowledge of the data generating distribution, which is not available in most offline settings.

One method to address the second problem is to use $\hat{v}_{k-1}$, i.e. a value function estimate from the previous iteration, as the regression target (e.g. as in the work [Moulin and Neu, 2023]). However, this incurs additional bias and results in an additional term in the sub-optimality guarantee that must be handled in the analysis. This is usually done by showing that the difference in value functions between consecutive policies is small due to the conservative policy update. However, it remains unclear how to address this issue for AMDPs without strengthening Assumption 3.1 to include, for example, a uniform mixing assumption.

We propose to bypass these additional complexities and directly solve for the fixed point equation. Since $\boldsymbol{w}_k$ parametrizes our q-function estimates, we should also have $\hat{v}_k(s) = \mathbb{E}_{a\sim\pi_k(\cdot|s)}[\phi(s,a)^\top \boldsymbol{w}_k]$. Therefore, we replace $\hat{v}_k(s_i')$ in (5) with $\phi^{\pi_k}(s_i')^\top \boldsymbol{w}_k$, where $\phi^{\pi_k}(s) = \mathbb{E}_{a\sim\pi_k(\cdot|s)}[\phi(s,a)]$. Inspired by the ideas of Zanette et al. [2021], we then add a perturbation $\boldsymbol{\xi} \in \mathbb{R}^d$ to the weight vector and solve the following optimization problem:

$$
\begin{aligned}
&(\boldsymbol{w}_k, \boldsymbol{\xi}_k, J_k) \in \operatorname*{arg\,min}_{\boldsymbol{w},\boldsymbol{\xi}\in\mathbb{R}^d, J\in\mathbb{R}} \quad J \\
\text{s.t.} \quad &\boldsymbol{w} = \boldsymbol{\xi} + \hat{\Lambda}^{-1} \sum_{i=1}^{N} \phi(s_i,a_i)\left(r(s_i,a_i) - J + \phi^{\pi_k}(s_i')^\top \boldsymbol{w}\right), \\
&|J| \le 1, \quad \|\boldsymbol{w}\|_2 \le B_w, \quad \text{and } \|\boldsymbol{\xi}\|_{\hat{\Lambda}} \le \beta,
\end{aligned}
\tag{6}
$$

where β is a parameter determined by B_w, K, N, $\kappa_{\mathcal{Q}(B_w),\Pi}$, $\varepsilon_{\mathcal{Q}(B_w),\Pi}$, and a confidence level $\delta \in (0,1)$. Specifically,

$$
\beta = C + (\kappa_{\mathcal{Q}(B_w),\Pi} + \varepsilon_{\mathcal{Q}(B_w),\Pi})\sqrt{N} \tag{7}
$$

where

$$
C = O\left(B_w \sqrt{d \log(KNB_w/\delta)}\right). \tag{8}
$$

This parameter quantifies the uncertainty in the dataset and our knowledge of the true MDP. The addition of $\boldsymbol{\xi}$ is how pessimism is incorporated into the algorithm. The ellipsoid $\|\boldsymbol{\xi}\|_{\hat{\Lambda}} \le \beta$ can be viewed as a confidence set which, with high probability, contains the error due to lack of knowledge of the true transition. In our analysis, we will show that J_k is a nearly pessimistic estimate of J^{π_k} up to misspecification error determined by the constant $\kappa_{\mathcal{Q}(B_w),\Pi}$. This approach is also similar to the FOPO algorithm [Wei et al., 2021] for the online setting, but with one crucial difference: because our algorithm is policy-based, the first constraint in (6) is based on the Bellman equation for a *fixed* policy rather than the Bellman optimality equation. Consequently, the feasible set in (6) is convex, being comprised of linear and convex quadratic constraints. Therefore, approximate solutions to this optimization problem, up to arbitrarily small error, can be computed in polynomial time with interior point methods [Nesterov and Nemirovskii, 1994]. This stands in stark contrast to the FOPO algorithm, where our linear constraint in (6) is replaced by the analogous but nonlinear constraint

$$
\boldsymbol{w} = \boldsymbol{\xi} + \hat{\Lambda}^{-1} \sum_{i=1}^{N} \phi(s_i,a_i)(r(s_i,a_i) - J + \max_a\{\phi(s_i',a)^\top \boldsymbol{w}\}).
$$

The nonlinearity comes from the additional maximization operation. This results in a non-convex constraint set and an optimization problem to be solved at every iteration without a known efficient computation method.

Finally, we remark here that while a fully efficient implementation of our algorithm would involve only approximate solutions to (6), for simplicity we will assume that $(\boldsymbol{w}_k, \boldsymbol{\xi}_k, J_k)$ is an exact solution for the remainder of the paper. We refer the interested reader to Section B of the appendix where we briefly discuss error propagation for a fully efficient implementation of our algorithm where (6) is solved approximately at each step.

4.2 Policy Update

Once the weight vector $\boldsymbol{w}_k$ is computed, the policy parameter is updated via

$$\boldsymbol{\theta}_{k+1} \leftarrow \boldsymbol{\theta}_k + \eta \boldsymbol{w}_k$$

where $\eta = \sqrt{\frac{\log A}{B_w^2 K}}$ is a step-size. Due to the form of policy class (4), this is equivalent to the exponential weights update

$$\pi_{k+1}(a|s) \propto \pi_k(a|s) \exp\left(\eta \hat{q}_k(s,a)\right), \quad \text{where } \hat{q}_k(s,a) = \phi(s,a)^\top \boldsymbol{w}_k$$

which, in turn, is equivalent to one step of mirror ascent with KL divergence regularization. Once the algorithm is terminated, it returns the output policy π_{out}, which is a mixture policy defined the uniform random sampling of policies $\{\pi_1, \ldots, \pi_K\}$. The pseudo code of our algorithm is presented in Algorithm 1.

Algorithm 1 Average Reward Actor-Critic

Input: $\mathcal{D}$ (dataset), B_w (function class parameter), β (uncertainty parameter), η (stepsize)
Form empirical covariance : $\hat{\Lambda} \leftarrow I + \sum_{i=1}^N \phi(s_i, a_i)\phi(s_i, a_i)^\top$
Initialize: $\boldsymbol{\theta}_1 = \mathbf{0}$
for $k = 1, \ldots, K$ **do**
  Let $(\boldsymbol{w}_k, \boldsymbol{\xi}_k, J_k)$ solve (6)
  Update policy parameter: $\boldsymbol{\theta}_{k+1} \leftarrow \boldsymbol{\theta}_k + \eta \boldsymbol{w}_k$.
end for
Output: $\pi_{\text{out}} = \text{Unif}[\pi_1, \ldots, \pi_K]$.

5 Main Results

Let $\hat{\Lambda}_N = \frac{1}{N}\hat{\Lambda}$ be the normalized covariance matrix. For a given comparator policy π with stationary measure μ^π, let $\phi^{\mu^\pi} = \mathbb{E}_{(s,a)\sim\mu^\pi\otimes\pi}[\phi(s,a)] \in \mathbb{R}^d$. The sub-optimality of the mixture policy π_{out} output by Algorithm 1 with respect to a comparator policy π depends on the random constant

$$\|\phi^{\mu^\pi}\|_{\hat{\Lambda}_N^{-1}},$$

referred to as the *feature coverage ratio* in [Gabbianelli et al., 2024, Neu and Okolo, 2025]. It measures how well the dataset $\mathcal{D}$ covers the feature space visited by π. When the *expected* feature vector ϕ^{μ^π} is aligned with the top eigenvector of the empirical covariance matrix $\hat{\Lambda}$, one should expect this constant to be small—this is the case when $\mathcal{D}$ consists primarily of state-action pairs who's feature vectors are closely aligned with those frequently visited when following policy π. Our converge ratio can be contrasted with that used in prior work by [Jin et al., 2021]

$$\mathbb{E}_{(s,a)\sim\mu^\pi\otimes\pi}[\|\phi(s,a)\|_{\hat{\Lambda}_N^{-1}}],$$

which is no smaller than ours by Jensen's inequality.

Our main result is stated in Theorem 5.1 below.

Theorem 5.1. *Fix any comparator policy π with stationary measure μ^π. If we set β as in (7), then with probability at least $1 - 2\delta$, for any $K \geq \log A$, Algorithm 1 run for K iterations with stepsize $\eta = \sqrt{\frac{\log A}{B_w^2 K}}$ outputs a policy π_{out} satisfying*

$$J^\pi - J^{\pi_{out}} \leq \underbrace{\frac{2C}{\sqrt{N}}\|\phi^{\mu^\pi}\|_{\hat{\Lambda}_N^{-1}}}_{T_1:\textit{Uncertainty}} + \underbrace{2B_w\sqrt{\frac{\log A}{K}}}_{T_2:\textit{Optimization}} + \underbrace{(2\|\phi^{\mu^\pi}\|_{\hat{\Lambda}_N^{-1}} + 1)(\varepsilon_{\mathcal{Q}(B_w),\Pi} + \kappa_{\mathcal{Q}(B_w),\Pi})}_{T_3:\textit{Misspecification}}. \quad (9)$$

The upper bound on the sub-optimality gap (9) consists of three terms. The first term T_1 represents the uncertainty in the dataset. It decays with the reciprocal of the square root of the dataset's size, but increases as the quality of the dataset with respect to the comparator policy degrades, as measured by the coverage ratio. The second term T_2 is the error due to optimization, which can be made small by increasing the number of iterations K. The final term T_3 is an irreducible error due to model misspecification. Note that the T_3 term also decreases as the quality of the dataset improves. Importantly, due to the definition of C in (8), the bound depends only on the feature dimension d rather than on the size of the state space.

As an application of Theorem 5.1, let us consider the optimal policy π^* as the comparator policy π. If $\|\phi^{\mu^{\pi^*}}\|_{\hat{\Lambda}^{-1}}$ is bounded above by some constant C_*, then Theorem 5.1 implies $\tilde{O}(B_w^2 C_*^2 d\varepsilon^{-2})$ samples are sufficient to learn a policy which is ε-optimal up to model misspecification error. One well-studied special case of zero misspecification error is the linear MDP [Jin et al., 2020, Wei et al., 2021]. Under this assumption, combined with Assumption 3.1, we can choose B_w large enough so that the function class $\mathcal{Q}(B_w)$ contains the true value functions with the knowledge of an upper bound on c. Specifically, if $B_w \geq O(c\sqrt{d})$, a straightforward adaptation of our analysis shows that Theorem 5.1 holds with no misspecification error term.

5.1 Comparison with prior work

Our work is inspired by the work on episodic setting [Zanette et al., 2021], particularly the idea of solving a constrained optimization problem at each step. This work is also the first to introduce the definition of coverage ratio that we adopt in this paper. However, despite the algorithmic similarities, the algorithm design and analysis in this work rely crucially on backwards induction methods that are only applicable to the episodic setting. Our work makes a significant contribution by extending the approach to the more challenging infinite horizon setting.

As mentioned in the introduction, the work [Gabbianelli et al., 2024] is the only paper we are aware of that studies offline RL for average reward MDPs with linear function approximation. However, their algorithm only attains $O(\varepsilon^{-4})$ sample complexity guarantees. The main reason for this is their algorithm's double loop structure. Their primal-dual formulation of the offline RL problem involves solving for the saddle point of a certain Lagrangian objective, which is done through multiple rounds of stochastic gradient ascent-descent. They assume that state action pairs in the data set are sampled i.i.d from a fixed distribution. Then in each outer loop of their algorithm, they use $O(\varepsilon^{-2})$ samples to solve a sub-problem nearly exactly. Since they need $O(\varepsilon^{-2})$ outer-loop iterations, this results in a total sample complexity of $O(\varepsilon^{-4})$. In contrast, in our algorithm the data needed to construct the optimization problem (6) only needs to be sampled once and is then re-used in each iteration. This avoids the inner-loop that uses additional samples. The data re-use creates an additional correlation between iterates which is dealt with in the analysis using covering arguments.

The primal-dual algorithm by [Hong and Tewari, 2024] guarantees $O(\varepsilon^{-2})$ sample complexity for discounted MDPs, but their results depend on a weaker definition of coverage. Using our notation, their algorithm requires an upper bound on $\|\phi^{\mu^\pi}\|^2_{\hat{\Lambda}_N^{-2}}$, which is assumed to be known. Finally, [Neu and Okolo, 2025] introduce another primal-dual style algorithm with an $O(\varepsilon^{-2})$ sample complexity guarantee for discounted MDPs. Their suboptimality bounds depend on $\|\phi^{\mu^\pi}\|^2_{\hat{\Lambda}_N^{-1}}$, which is the strongest result we know of in this setting for the discounted case. The authors mention that it would not be difficult to adapt their results to average reward MDPs using the ideas from the work [Gabbianelli et al., 2024], but no additional details are provided. Even so, our work still improves over theirs for two reasons. First, while they use the same definition of coverage ratio as ours, their sub-optimality bounds scale quadratically with this constant in contrast to our linear scaling. Second, their results, like prior work [Gabbianelli et al., 2024, Hong and Tewari, 2024], only cover the more restrictive class of linear MDPs, while we study the more general linear function approximation (cf. Assumption 3.3).

A further extension beyond linear models to more general function approximation would be an interesting future direction. In this case, it is not yet clear how one can efficiently construct confidence sets for the underlying parameters. Even so, as shown in [Xie et al., 2021, Cheng et al., 2022], it is still possible to implement the pessimism principle with general function approximation. The

main difficulty we see in adapting the methods from these papers to the average reward setting is the presence of the additional variable J.

6 Analysis

Let $\hat{q}_k(s,a) := \phi(s,a)^\top \boldsymbol{w}_k$ and $\hat{v}_k(s) := \mathbb{E}_{a\sim\pi_k(\cdot|s)}[\hat{q}_k(s,a)]$ denote the empirical estimates of q^{π_k} and v^{π_k}, respectively, at the end of k-th iteration. Key to our analysis is the following lemma, which allows us to decompose the sub-optimality in Theorem 5.1 into its three main parts as shown in (9).

Lemma 6.1. *Fix policies $\pi, \tilde{\pi} \in \Delta(\mathcal{A})$. Let $\hat{J}^{\tilde{\pi}}$ be an estimate of the true average reward following policy $\tilde{\pi}$, and $\hat{q}^{\tilde{\pi}} \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$ be an estimate of the true Q-function$q^{\tilde{\pi}}$. Then*

$$J^\pi - \hat{J}^{\tilde{\pi}} = \mathbb{E}_{s\sim\mu^\pi}\left[\sum_a (\pi(a|s) - \tilde{\pi}(a|s))\hat{q}^{\tilde{\pi}}(s,a)\right] + \mathbb{E}_{(s,a)\sim\mu^\pi\otimes\pi}\left[T^{\tilde{\pi}}\hat{q}^{\tilde{\pi}}(s,a) - \hat{J}^{\tilde{\pi}} - \hat{q}^{\tilde{\pi}}(s,a)\right].$$

Lemma 6.1 is analogous to the so-called *extended performance difference lemma*, which is commonly used in the analysis of optimistic policy optimization algorithms for the online episodic setting; see, for example, the work [Cai et al., 2020, Shani et al., 2020]. Below we break down the analysis into three main steps.

Step 1: Pessimism. Suppose for the moment that $\kappa_{\mathcal{Q}(B_w),\Pi} = 0$. Then by the definition (3) there would exist some $\boldsymbol{w}_k^*$ such that $\phi(s,a)^\top \boldsymbol{w}_k^*$ solves the Bellman equation for policy π_k, meaning that $\phi(s,a)^\top \boldsymbol{w}_k^* = q^{\pi_k}(s,a)$. If we can show that $(\boldsymbol{w}_k^*, \boldsymbol{\xi}_k^*, J^{\pi_k})$ is feasible for critic's optimization problem (6), we will have $J_k \le J^{\pi_k}$ by the definition of J_k, which has the desired pessimism property. In general though, if $\kappa_{\mathcal{Q}(B_w),\Pi} > 0$, it may be impossible to find such a $\boldsymbol{w}_k^*$. Therefore, we instead define

$$(\boldsymbol{w}_k^*, J_k^*) \in \operatorname*{arg\,min}_{\|\boldsymbol{w}\|_2\le B_w, |J|\le 1} \|\phi(\cdot,\cdot)^\top \boldsymbol{w} + J - T^{\pi_k}(\phi(\cdot,\cdot)^\top \boldsymbol{w})\|_\infty \tag{10}$$

as the best possible weight vector and estimate of J^{π_k} that incurs at most $\kappa_{\mathcal{Q}(B_w),\Pi}$ error by definition. With the help of Lemma 6.1, we then show that $J_k \le J_k^* \le J^{\pi_k} + \kappa_{\mathcal{Q}(B_w),\Pi}$ holds with high probability. This result is summarized in the following lemma.

Lemma 6.2. *With probability at least $1-\delta$, for each $k \in [K]$ there exists $\boldsymbol{\xi}_k^* \in \mathbb{R}^d$ such that $(\boldsymbol{w}_k^*, \boldsymbol{\xi}_k^*, J_k^*)$ is feasible for the optimization problem (6). As a consequence of the definition of J_k,*

$$J_k \le J_k^* \le J^{\pi_k} + \kappa_{\mathcal{Q}(B_w),\Pi}, \quad \forall k \in [K].$$

Step 2: Bounding the estimation error. The next step is to control the error in the estimates $\hat{q}_k$ and J_k. More specifically, in view of the second term on the right hand side of Lemma 6.1, we are interested in bounding

$$\left|\mathbb{E}_{(s,a)\sim\mu^\pi\otimes\pi}\left[\hat{q}_k(s,a) + J_k - T^{\pi_k}\hat{q}_k(s,a)\right]\right|, \tag{11}$$

where π is some comparator policy. In a manner similar to step 1, we define

$$\bar{\boldsymbol{w}}_k \in \operatorname*{arg\,min}_{\|\boldsymbol{w}\|_2\le B_w} \|\phi(\cdot,\cdot)^\top \boldsymbol{w} + J_k - T^{\pi_k}\hat{q}_k\|_\infty,$$

which is the best possible regression parameter with target $J_k - T^{\pi_k}\hat{q}_k$. In this case, we have $\|\phi(\cdot,\cdot)^\top \bar{\boldsymbol{w}}_k + J_k - T^{\pi_k}\hat{q}_k\|_\infty \le \varepsilon_{\mathcal{Q}(B_w),\Pi}$ by Assumption 3.3. If $\varepsilon_{\mathcal{Q}(B_w),\Pi} = 0$, then $\hat{q}_k(s,a) - \phi(s,a)^\top \bar{\boldsymbol{w}}_k$ is exactly equal to quantity inside the expectation in (11). In the general case, (11) can be bounded by

$$|\mathbb{E}_{(s,a)\sim\mu^\pi\otimes\pi}[\hat{q}_k(s,a) - \phi(s,a)^\top \bar{\boldsymbol{w}}_k]| + \varepsilon_{\mathcal{Q}(B_w),\Pi}.$$

A high probability upper bound on the first term above results in the following Lemma 6.3.

Lemma 6.3. *With probability at least $1-\delta$, for all $k \in [K]$ and any policy π,*

$$|\mathbb{E}_{(s,a)\sim\mu^\pi\otimes\pi}[\hat{q}_k(s,a) + J_k - T^{\pi_k}\hat{q}_k(s,a)]| \le 2\beta\|\phi^{\mu^\pi}\|_{\hat{\Lambda}^{-1}} + \varepsilon_{\mathcal{Q}(B_w),\Pi}.$$

Step 3: Completing the proof. Theorem 5.1 holds on the events of Lemmas 6.2 and 6.3, which are true simultaneously with probability at least $1 - 2\delta$. The policy output by Algorithm 1 is a mixture policy with average reward $J^{\pi_{\text{out}}} = \frac{1}{K}\sum_{k=1}^{K} J^{\pi_k}$. Combining Lemma 6.2 with Lemma 6.1 we can show

$$\begin{aligned}\frac{1}{K}\sum_{k=1}^{K} J^{\pi} - J^{\pi_k} &\leq \frac{1}{K}\sum_{k=1}^{K} J^{\pi} - J_k + \kappa_{\mathcal{Q}(B_w),\Pi} \\ &\leq \frac{1}{K}\sum_{k=1}^{K} \mathbb{E}_{s\sim\mu^{\pi}}\left[\sum_{a}(\pi(a|s) - \pi_k(a|s))\hat{q}_k(s,a)\right] \\ &\quad + \frac{1}{K}\sum_{k=1}^{K} \mathbb{E}_{(s,a)\sim\mu^{\pi}\otimes\pi}\left[T^{\pi_k}\hat{q}_k(s,a) - J_k - \hat{q}_k(s,a)\right] + \kappa_{\mathcal{Q}(B_w),\Pi}.\end{aligned}$$

The first term above is bounded above by the optimization error, which is proved through the analysis of mirror descent. The proof is completed by using Lemma 6.3, the definition of β, and rescaling $\hat{\Lambda}^{-1} = \frac{1}{N}\hat{\Lambda}_N^{-1}$.

7 Conclusion

In this paper, we have introduced a pessimistic actor critic algorithm for offline learning in infinite horizon average reward MDPs with linear function approximation. Our results show that our algorithm is sample efficient, with provable guarantees under only partial data coverage. One limitation of our algorithm is that we require the MDP to be unichain and value functions to have uniformly bounded span. Aside from an extension to general function approximation, potential future work could also include weakening Assumption 3.1 to include all weakly communicating MDPs.

Acknowledgments and Disclosure of Funding

We thank all reviewers for their helpful suggestions and comments. The research of WP was supported in part by NSF Award DMS-2023239. JK was partially funded by AFOSR/AFRL grant no. FA9550-18-1-0166. HL was partially supported by NSF Award DMS-2206296. QX was supported in part by NSF grants CNS-1955997, ECCS-2339794, and ECCS-2432546.

References

Yasin Abbasi-Yadkori, Peter Bartlett, Kush Bhatia, Nevena Lazic, Csaba Szepesvari, and Gellert Weisz. POLITEX: Regret bounds for policy iteration using expert prediction. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 3692–3702. PMLR, 2019.

Akhil Agnihotri, Rahul Jain, and Haipeng Luo. ACPO: A policy optimization algorithm for average MDPs with constraints. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 397–415. PMLR, 2024.

Priyank Agrawal and Shipra Agrawal. Optimistic q-learning for average reward and episodic reinforcement learning extended abstract. In *Proceedings of Thirty Eighth Conference on Learning Theory*, volume 291 of *Proceedings of Machine Learning Research*, pages 1–1. PMLR, 2025.

András Antos, Csaba Szepesvári, and Rémi Munos. Fitted q-iteration in continuous action-space mdps. In *Advances in Neural Information Processing Systems*, volume 20. Curran Associates, Inc., 2007.

Qinbo Bai, Washim Uddin Mondal, and Vaneet Aggarwal. Regret analysis of policy gradient algorithm for infinite horizon average reward markov decision processes. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(10):10980–10988, Mar. 2024.

Qi Cai, Zhuoran Yang, Chi Jin, and Zhaoran Wang. Provably efficient exploration in policy optimization. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 1283–1294. PMLR, 2020.

Ching-An Cheng, Tengyang Xie, Nan Jiang, and Alekh Agarwal. Adversarially trained actor critic for offline reinforcement learning. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 3852–3878. PMLR, 17–23 Jul 2022.

Ronan Fruit, Matteo Pirotta, Alessandro Lazaric, and Ronald Ortner. Efficient bias-span-constrained exploration-exploitation in reinforcement learning. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 1578–1586. PMLR, 2018.

Germano Gabbianelli, Gergely Neu, Matteo Papini, and Nneka M Okolo. Offline primal-dual reinforcement learning for linear MDPs. In *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pages 3169–3177. PMLR, 2024.

Ilaria Giannoccaro and Pierpaolo Pontrandolfo. Inventory management in supply chains: a reinforcement learning approach. *International Journal of Production Economics*, 78(2):153–161, 2002.

Botao Hao, Nevena Lazic, Yasin Abbasi-Yadkori, Pooria Joulani, and Csaba Szepesvari. Adaptive approximate policy iteration. In *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pages 523–531. PMLR, 2021.

Kihyuk Hong and Ambuj Tewari. A primal-dual algorithm for offline constrained reinforcement learning with linear MDPs. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 18711–18737. PMLR, 2024.

Kihyuk Hong, Woojin Chae, Yufan Zhang, Dabeen Lee, and Ambuj Tewari. Reinforcement learning for infinite-horizon average-reward linear mdps via approximation by discounted-reward mdps. In *Proceedings of The 28th International Conference on Artificial Intelligence and Statistics*, volume 258 of *Proceedings of Machine Learning Research*, pages 2989–2997. PMLR, 2025.

Thomas Jaksch, Ronald Ortner, and Peter Auer. Near-optimal regret bounds for reinforcement learning. *Journal of Machine Learning Research*, 11(51):1563–1600, 2010.

Chi Jin, Zhuoran Yang, Zhaoran Wang, and Michael I Jordan. Provably efficient reinforcement learning with linear function approximation. In *Proceedings of Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pages 2137–2143. PMLR, 2020.

Ying Jin, Zhuoran Yang, and Zhaoran Wang. Is pessimism provably efficient for offline rl? In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 5084–5096. PMLR, 18–24 Jul 2021.

Yujia Jin and Aaron Sidford. Efficiently solving MDPs with stochastic mirror descent. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 4890–4900. PMLR, 2020.

Yujia Jin and Aaron Sidford. Towards tight bounds on the sample complexity of average-reward mdps. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 5055–5064. PMLR, 2021.

Gen Li, Laixi Shi, Yuxin Chen, Yuejie Chi, and Yuting Wei. Settling the sample complexity of model-based offline reinforcement learning. *The Annals of Statistics*, 52(1):233 – 260, 2024.

Antoine Moulin and Gergely Neu. Optimistic planning by regularized dynamic programming. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 25337–25357. PMLR, 2023.

Rémi Munos and Csaba Szepesvári. Finite-time bounds for fitted value iteration. *Journal of Machine Learning Research*, 9(27):815–857, 2008.

Yurii Nesterov and Arkadii Nemirovskii. *Interior-Point Polynomial Algorithms in Convex Programming*. Society for Industrial and Applied Mathematics, 1994.

Gergely Neu and Nneka Okolo. Offline RL via feature-occupancy gradient ascent. In *The 28th International Conference on Artificial Intelligence and Statistics*, 2025.

Thanh Nguyen-Tang and Raman Arora. On sample-efficient offline reinforcement learning: Data diversity, posterior sampling and beyond. In *Advances in Neural Information Processing Systems*, volume 36, pages 61115–61157. Curran Associates, Inc., 2023.

Martin L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley I& Sons, 2005.

Paria Rashidinejad, Banghua Zhu, Cong Ma, Jiantao Jiao, and Stuart Russell. Bridging offline reinforcement learning and imitation learning: A tale of pessimism. In *Advances in Neural Information Processing Systems*, volume 34, pages 11702–11716. Curran Associates, Inc., 2021.

Lior Shani, Yonathan Efroni, Aviv Rosenberg, and Shie Mannor. Optimistic policy optimization with bandit feedback. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 8604–8613. PMLR, 2020.

Uri Sherman, Alon Cohen, Tomer Koren, and Yishay Mansour. Rate-optimal policy optimization for linear Markov decision processes. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 44815–44837. PMLR, 21–27 Jul 2024.

Alexander L. Strehl and Michael L. Littman. An analysis of model-based interval estimation for markov decision processes. *Journal of Computer and System Sciences*, 74(8):1309–1331, 2008.

Wesley A Suttle, Amrit Bedi, Bhrij Patel, Brian M. Sadler, Alec Koppel, and Dinesh Manocha. Beyond exponentially fast mixing in average-reward reinforcement learning via multi-level Monte Carlo actor-critic. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 33240–33267. PMLR, 2023.

Shengpu Tang and Jenna Wiens. Model selection for offline reinforcement learning: Practical considerations for healthcare settings. In *Proceedings of the 6th Machine Learning for Healthcare Conference*, volume 149 of *Proceedings of Machine Learning Research*, pages 2–35. PMLR, 2021.

Ruosong Wang, Dean Foster, and Sham M. Kakade. What are the statistical limits of offline RL with linear function approximation? In *International Conference on Learning Representations*, 2021.

Lucas Weber, Ana Busic, and Jiamin Zhu. Reinforcement learning and regret bounds for admission control. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 52403–52427. PMLR, 2024.

Chen-Yu Wei, Mehdi Jafarnia Jahromi, Haipeng Luo, Hiteshi Sharma, and Rahul Jain. Model-free reinforcement learning in infinite-horizon average-reward Markov decision processes. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 10170–10180. PMLR, 2020.

Chen-Yu Wei, Mehdi Jafarnia Jahromi, Haipeng Luo, and Rahul Jain. Learning infinite-horizon average-reward mdps with linear function approximation. In *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pages 3007–3015. PMLR, 2021.

Tengyang Xie, Ching-An Cheng, Nan Jiang, Paul Mineiro, and Alekh Agarwal. Bellman-consistent pessimism for offline reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 34, pages 6683–6694. Curran Associates, Inc., 2021.

Andrea Zanette, Alessandro Lazaric, Mykel Kochenderfer, and Emma Brunskill. Learning near optimal policies with low inherent Bellman error. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 10978–10989. PMLR, 2020.

Andrea Zanette, Martin J Wainwright, and Emma Brunskill. Provable benefits of actor-critic methods for offline reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 34, pages 13626–13640. Curran Associates, Inc., 2021.

Wenhao Zhan, Baihe Huang, Audrey Huang, Nan Jiang, and Jason Lee. Offline reinforcement learning with realizability and single-policy concentrability. In *Proceedings of Thirty Fifth Conference on Learning Theory*, volume 178 of *Proceedings of Machine Learning Research*, pages 2730–2775. PMLR, 02–05 Jul 2022.

Yiming Zhang and Keith W Ross. On-policy deep reinforcement learning for the average-reward criterion. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 12535–12545. PMLR, 18–24 Jul 2021.

Zihan Zhang and Qiaomin Xie. Sharper model-free reinforcement learning for average-reward markov decision processes. In *Proceedings of Thirty Sixth Conference on Learning Theory*, volume 195 of *Proceedings of Machine Learning Research*, pages 5476–5477. PMLR, 2023.

Matthew Zurek and Yudong Chen. Span-based optimal sample complexity for weakly communicating and general average reward mdps. In *Advances in Neural Information Processing Systems*, volume 37, pages 33455–33504. Curran Associates, Inc., 2024.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction clearly describe the main contribution of this work, which is an algorithm for offline learning in average reward MDPs with linear function approximation and partial coverage.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The assumptions required for our algorithm are clearly stated in section 3, namely bounded Bellman closedness and realizability error as well as unichain MDPs. The paper also clearly discusses the limitation of our algorithm to linear function approximation.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.

- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: All assumptions are stated in Section 3. A rigorous proof of the main theorem is given in the appendix.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [NA]

   Justification: The paper does not include experiments.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example

(a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
(b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
(c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [NA]

   Justification: The paper does not include experiments requiring code.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [NA]

   Justification: The paper does not include experiments.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [NA]

   Justification: The paper does not include experiments.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [NA]

   Justification: The paper does not include experiments.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: We have reviewed the NeurIPS Code of Ethics and verified that our paper conforms with it.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

    Justification: This paper is theoretical in nature and not directly related to any practical application with the potential for societal impact.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: The paper poses no such risks.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [NA]

    Justification: The paper does not use existing assets.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: The paper does not release new assets.

    Guidelines:

    - The answer NA means that the paper does not release new assets.

- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

    Justification: The core method development in this research does not involve LLMs as any important, original, or non-standard components.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (https://neurips.cc/Conferences/2025/LLM) for what should or should not be described.

A Further Discussion of Assumption 3.1

Here we give a more detailed discussion of Assumption 3.1 and how it compares to other average reward MDP models used in the literature.

The weakest assumption frequently made is the requirement that the MDP be weakly communicating [Jaksch et al., 2010, Fruit et al., 2018, Wei et al., 2020, 2021, Zhang and Xie, 2023, Hong et al., 2025, Zurek and Chen, 2024]. Aside from [Zurek and Chen, 2024] who study learning with access to a generative model, all of the works cited above study *online* learning in weakly communicating MDPs. We are not yet aware of any papers studying offline learning in weakly communicating MDPs.

An MDP is weakly communicating if the state space can be divided into two classes. One class consists of states that are transient for every policy. The other, called the communicating class, consists of a set of states with the following property: for each pair of states (s, s') there exists a policy π such that s' is reachable from s when following π [Puterman, 2005]. As shown in [Jaksch et al., 2010], the weakly communicating assumption is necessary for efficient online learning. The property of being weakly communicating is sufficient to guarantee the existence of a solution (q^*, J^*) to the Bellman optimality equation

$$q^*(s,a) + J^* = r(s,a) + \mathbb{E}_{s' \sim P(\cdot|s,a)}[\max_a q^*(s',a)]$$

where J^* is constant. However, this is not enough to guarantee that for a fixed policy π, $J^\pi(s)$ as defined in (1) is constant. In this case, solutions q^π to the Bellman equation

$$q^\pi(s,a) + J^\pi(s) = T^\pi q^\pi(s,a)$$

may exist, but may only be unique modulo a subspace of dimension greater than one. This is because, in weakly communicating MDPs, there can be policies that induce Markov chains with multiple recurrence classes. As such, the chain's stationary measure is not unique: there is a different one for each recurrence class. Moreover, while q^* is bounded, there is no guaranteed uniform upper bound on the size of q^π as opposed to discounted and episodic MDPs where $\frac{1}{1-\gamma}$ or H provide a natural upper bound. This seems to make learning with policy optimization style algorithms difficult. It is not clear how to accurately estimate $q^\pi(s,a)$ in the weakly communicating case, and lack of a clear upper bound can destabilize the algorithm.

At the other end of the spectrum, the strongest assumption made in the literature is uniform ergodicity. The "ergodicity" part of this assumption requires that each policy π induce an irreducible, aperiodic Markov chain. This ensures that the stationary measure is unique for each policy and positive everywhere. The "uniform" part of this assumption means that that the worst case mixing time

$$t_{\text{mix}} = \sup_\pi \inf\{t \geq 1 : \max_s \|(P^\pi)^t(s,\cdot) - \mu^\pi\|_{TV} < \tfrac{1}{4}\} \tag{12}$$

is finite and all stationary measures are uniformly bounded away from zero:

$$\inf_{\pi,s} \mu^\pi(s) \geq \sigma > 0. \tag{13}$$

When learning with linear function approximation, this second condition is often replaced with

$$\mathbb{E}_{(s,a)\sim\mu^\pi\otimes\pi}\left[\phi(s,a)\phi(s,a)^\top\right] \succeq \lambda I, \tag{14}$$

meaning the true covariance matrix for each policy is uniformly positive definite. This is referred to as a "uniformly excited features" assumption in [Wei et al., 2021, Hao et al., 2021].

In the online setting, uniform ergodicity has been assumed in e.g. [Abbasi-Yadkori et al., 2019, Wei et al., 2020, 2021, Bai et al., 2024]. Under this assumption, the Bellman equation is solved by

$$q^\pi(s,a) = \sum_{t=0}^{\infty} \mathbb{E}^\pi_{s,a}\left[r(s_t,a_t) - J^\pi\right] \tag{15}$$

and $|q^\pi(s,a)| \leq O(t_{\text{mix}})$. This is convenient for online learning, especially with policy optimization, because the q functions are bounded and (13) and (14) imply that each policy is self-exploratory.

In the offline setting, assumptions such as (13) and (14) would greatly weaken single policy coverage results such as our Theorem 5.1. As discussed in the main body, effectively covering any policy requires covering the entire state space in the case of (14) or the entire feature space in the case of (14).

Because of this, it is important for us to emphasize that our assumptions do *not* imply conditions such as (13) or (14). Our Assumption 3.1, which is the same assumption made by [Gabbianelli et al., 2024], can be thought of as being somewhere in the middle of the two extremes mentioned above. It is a stronger than weakly communicating, but weaker than uniform ergodicity. Indeed, the uniform mixing condition (12) by itself implies Assumption 3.1 but is not necessary. Another sufficient condition is the existence of a single state $\bar{s}$ visited in

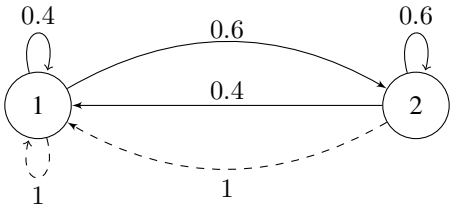

Figure 1: Two state river swim model. The dashed edges represent transition probabilities when taking action L and the solid edges show the transition probabilities when taking action R.

expected time at most h started from any other state as assumed in [Agrawal and Agrawal, 2025]. With this assumption, the span of the q-function is bounded by h.

We end this section with a simple example of an MDP which satisfies Assumption 3.1 but is not uniformly ergodic. Figure 1 diagrams a simple two state river-swim [Strehl and Littman, 2008] MDP where $\mathcal{S} = \{1, 2\}$ and $\mathcal{A} = \{L, R\}$. Conditioned on being in state 1, the agent stays in state 1 almost surely if action L is chosen. If action R is chosen, the agent will stay in state 1 with probability 0.4 and transition to state 2 with probability 0.6. On the other hand, if in state 2, the agent deterministically transitions to state 1 if action L is chosen and will stay in state 2 with probability 0.6 if action R is chosen. Otherwise, it transitions to state 1. A simple linear algebra computation shows that the stationary measure is given by

$$[\mu^\pi(1), \mu^\pi(2)] = \left[\frac{1 - 0.6\pi(R|2)}{1 + 0.6(\pi(R|1) - \pi(R|2))}, \frac{0.6\pi(R|1)}{1 + 0.6(\pi(R|1) - \pi(R|2))}\right].$$

for any π. If the agent never takes action R in state 1, i.e. $\pi(R|1) = 0$, then $\mu^\pi(2) = 0$ so condition (13) fails. It is also easy to see that the expected covariance matrix, as in (14), is rank 1 and so cannot be positive definite if $d \geq 2$. However, it is not hard to show that for any initial distribution ν and policy π,

$$|(\nu P^\pi)^{t+1}(j) - \mu^\pi(j)| \leq 0.6|(\nu P^\pi)^t(j) - \mu^\pi(j)|$$

for $j \in \{1, 2\}$. This gives a upper bound on t_{mix} of $\frac{5 \log 4}{2}$, which by (15) implies an upper bound on the span of q^π.

B A Fully Computationally Efficient Implementation

As discussed in Section 4, Algorithm 1 as written in the main body is not fully computationally efficient. This is because Line 5 assumes that the parameters $(\boldsymbol{w}_k, \boldsymbol{\xi}_k, J_k)$ are an exact solution to the optimization problem (6). It is not guaranteed that an exact solution to this problem can be computed efficiently. However, approximate solutions to (6) can be computed in polynomial time using interior point methods [Nesterov and Nemirovskii, 1994]. In this section, we briefly show how a fully computationally efficient implementation of Algorithm 1 where (6) is solved only approximately in step 5 results only in a small additive error proportional to the accuracy of the solution.

Specifically, for some tolerance parameter $\eta > 0$, let us assume that $(\boldsymbol{w}_k, \boldsymbol{\xi}_k, J_k)$ are a η-approximate solution to (6). This means that J_k is within η of the optimal solution and the magnitude of the constraint violation is at most η when measured in the ℓ_2 norm. Then the inequality in Lemma 6.2 may be updated to read

$$J_k \leq J_k^* + \eta \leq J^{\pi_k} + \kappa_{\mathcal{Q}(B_w),\Pi} + \eta.$$

The error in the constraint violation will appear in Lemma 6.3. If we approximately solve (6) so that $(\boldsymbol{w}_k, \boldsymbol{\xi}_k, J_k)$ satisfy

$$\left\| \boldsymbol{w}_k - \boldsymbol{\xi}_k - \hat{\Lambda}^{-1} \sum_{i=1}^N \phi(s_i, a_i) \left(r(s_i, a_i) - J + \phi^{\pi_k}(s_i')^\top \boldsymbol{w}_k \right) \right\|_2 \leq \eta,$$

$$|J_k| \leq 1 + \eta, \quad \|\boldsymbol{w}_k\|_2 \leq B_w + \eta, \quad \text{and } \|\boldsymbol{\xi}_k\|_{\hat{\Lambda}} \leq \beta + \eta,$$

then a quick inspection of the proof of Lemma 6.3 (more specifically Lemma C.3 below) shows that the result can be updated to read

$$|\mathbb{E}_{(s,a)\sim\mu^\pi \otimes \pi}[\hat{q}_k(s,a) + J_k - T^{\pi_k}\hat{q}_k(s,a)]| \leq 2(\beta + \eta)\|\phi^{\mu^\pi}\|_{\hat{\Lambda}^{-1}} + \varepsilon_{\mathcal{Q}(B_w),\Pi} + \eta.$$

Altogether, this only results an additional error of

$$2\eta\|\phi^{\mu^\pi}\|_{\hat{\Lambda}^{-1}} + 2\eta$$

added to the right hand side of (9) in Theorem 5.1.

C Proofs of Key Lemmas and Theorem 5.1

In this section, we will prove the key Lemmas from section 6 and also give a rigorous proof of Theorem 5.1. We start with Lemma 6.1.

Proof of Lemma 6.1. Define $\hat{v}^{\tilde{\pi}}(s) = \sum_a \tilde{\pi}(a|s)\hat{q}^{\tilde{\pi}}(s,a)$. Adding and subtracting $r(s,a) - \hat{J}^{\pi} + P_{s,a}\hat{v}^{\tilde{\pi}}$ and then using $J^{\pi} = \mathbb{E}_{(s,a)\sim\mu^{\pi}\otimes\pi}[r(s,a)]$ we have

$$\begin{aligned}
&\mathbb{E}_{s\sim\mu^{\pi}}\left[\sum_a \pi(a|s)\hat{q}^{\tilde{\pi}}(s,a)\right] \\
&= \mathbb{E}_{s\sim\mu^{\pi}}\left[\sum_a \pi(a|s)\left(\hat{q}^{\tilde{\pi}}(s,a) + \hat{J}^{\tilde{\pi}} - r(s,a) - P_{s,a}\hat{v}^{\tilde{\pi}}\right)\right] \\
&\qquad + \mathbb{E}_{s\sim\mu^{\pi}}\left[\sum_a \pi(a|s)\left(r(s,a) - \hat{J}^{\tilde{\pi}} + P_{s,a}\hat{v}^{\tilde{\pi}}\right)\right] \\
&= \mathbb{E}_{(s,a)\sim\mu^{\pi}\otimes\pi}\left[\hat{q}^{\tilde{\pi}}(s,a) + \hat{J}^{\tilde{\pi}} - T^{\tilde{\pi}}\hat{q}(s,a)\right] + J^{\pi} - \hat{J}^{\tilde{\pi}} + \mathbb{E}_{(s,a)\sim\mu^{\pi}\otimes\pi}\left[P_{s,a}\hat{v}^{\tilde{\pi}}\right]
\end{aligned}$$

Now, since μ^{π} is the stationary measure for policy π,

$$\mathbb{E}_{(s,a)\sim\mu^{\pi}\otimes\pi}\left[P_{s,a}\hat{v}^{\tilde{\pi}}\right] = \mathbb{E}_{s\sim\mu^{\pi}}[\hat{v}^{\tilde{\pi}}(s)] = \mathbb{E}_{s\sim\mu^{\pi}}\left[\sum_a \tilde{\pi}(a|s)\hat{q}^{\tilde{\pi}}(s,a)\right].$$

Inserting this into the last line above and re-arranging completes the proof. □

C.1 Proofs of Lemmas 6.2 and 6.3

Before proving Lemmas 6.2 and 6.3, we need to introduce some additional notation and one auxillary result. Define the restricted policy class

$$\Pi(B_{\boldsymbol{\theta}}) = \left\{\frac{e^{\phi(s,a)^{\top}\boldsymbol{\theta}}}{\sum_{a'} e^{\phi(s,a')^{\top}\boldsymbol{\theta}}} : \|\boldsymbol{\theta}\|_2 \le B_{\boldsymbol{\theta}}\right\}$$

and the value function class

$$\mathcal{V}(B_{\boldsymbol{\theta}}, B_w) := \{v(s;\pi) = \langle \pi(\cdot|s), q(s,\cdot)\rangle : \pi \in \Pi(B_{\boldsymbol{\theta}}), q \in \mathcal{Q}(B_w)\}. \tag{16}$$

The number $B_{\boldsymbol{\theta}}$ represents the maximum size of any policy parameter $\boldsymbol{\theta}_k$ used during the course of K iterations of Algorithm 1. Notice that, based on the policy update in Algorithm 1 we have

$$\|\boldsymbol{\theta}_k\|_2 \le \eta \sum_{k=1}^{K} \|\boldsymbol{w}_k\|_2 \le \eta K B_w$$

so we take $B_{\boldsymbol{\theta}} = \eta K B_w$.

The proofs of Lemmas 6.2 and 6.3 rely on the following uniform concentration inequality. Its proof is based on standard arguments from the literature first used in [Jin et al., 2020]. The proof of this lemma is delayed to Appendix D.

Lemma C.1. *Let $\{v_k\}_{k=1}^{K}$ be any, possibly random, collection from the function class $\mathcal{V}(B_{\boldsymbol{\theta}}, B_w)$. With probability at least $1-\delta$ we have, for all $k \in [K]$,*

$$\left\|\sum_{i=1}^{N} \phi(s_i,a_i)\left(P_{s_i,a_i}v_k - v_k(s_i')\right)\right\|^2_{\tilde{\Lambda}^{-1}} \le \Gamma^2(B_w, N, \delta, K, d).$$

where

$$\Gamma^2(B_w, N, \delta, K, d) = 4B_w^2\left(\frac{d}{2}\log\left(\frac{K(N+1)}{\delta}\right) + d\log(1+4NB_w) + d\log(1+16NB_wB_{\boldsymbol{\theta}})\right) + 8.$$

We define C, the constant that appears in (8), as

$$C = \Gamma(B_w, N, \delta, K, d) + B_w.$$

The constant Γ is derived from the well-known concentration of self-normalized processes (Lemma E.4) and the log-covering number of the class $\mathcal{V}(B_{\boldsymbol{\theta}}, B_w)$.

The next lemma shows that J_k^* defined in (10) is within $\kappa_{\mathcal{Q}(B_w),\Pi}$ of J^{π_k}.

Lemma C.2. *Under Assumption 3.3 we have*

$$|J_k^* - J^{\pi_k}| \leq \kappa_{\mathcal{Q}(B_w),\Pi}$$

for every $k \in [K]$.

Proof. Let $q_k^*(s,a) = \phi(s,a)^\top \boldsymbol{w}_k^*$. By Lemma 6.1 and Assumption 3.3

$$\begin{aligned}
|J^{\pi_k} - J_k^*| &= \left|\mathbb{E}_{(s,a)\sim\mu^{\pi_k}\otimes\pi_k}\left[q_k^*(s,a) + J_k^* - T^{\pi_k} q_k^*(s,a)\right]\right| \\
&\leq \|q_k^* + J_k^* - T^{\pi_k} q_k^*\|_\infty \\
&\leq \kappa_{\mathcal{Q}(B_w),\Pi}.
\end{aligned}$$

□

We are now ready to prove Lemma 6.2. Throughout the proof, let

$$q_k^*(s,a) = \phi(s,a)^\top \boldsymbol{w}_k^* \quad v_k^*(s) = \mathbb{E}_{a\sim\pi_k(\cdot|s)}\left[q_k^*(s,a)\right]$$

be the approximate q and value functions corresponding to the optimal parameter $\boldsymbol{w}_k^*$.

Proof of Lemma 6.2. By the definition (10), $\boldsymbol{w}_k^*$ and J_k^* are always feasible variables. Thus, we need to show that there is some $\boldsymbol{\xi}_k^* \in \mathbb{R}^d$ with $\|\boldsymbol{\xi}_k^*\|_{\hat{\Lambda}} \leq \beta$ such that

$$\boldsymbol{w}_k^* = \boldsymbol{\xi}_k^* + \hat{\Lambda}^{-1} \sum_{i=1}^N \phi(s_i,a_i)\left(r(s_i,a_i) - J_k^* + \phi^{\pi_k}(s_i')^\top \boldsymbol{w}_k^*\right).$$

To this end, note that

$$\begin{aligned}
\boldsymbol{w}_k^* &= \hat{\Lambda}^{-1}\hat{\Lambda}\boldsymbol{w}_k^* \\
&= \hat{\Lambda}^{-1} \sum_{i=1}^N \phi(s_i,a_i)\phi(s_i,a_i)^\top \boldsymbol{w}_k^* + \hat{\Lambda}^{-1}\boldsymbol{w}_k^* \\
&= \hat{\Lambda}^{-1} \sum_{i=1}^N \phi(s_i,a_i)\left(T^{\pi_k} q_k^*(s,a) - J_k^*\right) + \hat{\Lambda}^{-1} \sum_{i=1}^N \phi(s_i,a_i)\Delta_k(s_i,a_i) + \hat{\Lambda}^{-1}\boldsymbol{w}_k^*. \qquad (17)
\end{aligned}$$

where we define

$$\Delta_k(s,a) = q_k^*(s,a) + J_k^* - T^{\pi_k} q_k^*(s,a).$$

Adding and subtracting $v_k^*(s_i')$ from the first term in (17) we have

$$\begin{aligned}
&\hat{\Lambda}^{-1} \sum_{i=1}^N \phi(s_i,a_i)\left(T^{\pi_k} q_k^*(s_i,a_i) - J_k^*\right) \\
&= \hat{\Lambda}^{-1} \sum_{i=1}^N \phi(s_i,a_i)\left(r(s_i,a_i) - J_k^* + v_k^*(s_i')\right) + \hat{\Lambda}^{-1} \sum_{i=1}^N \phi(s_i,a_i)\left(P_{s_i,a_i} v_k^* - v_k^*(s_i')\right).
\end{aligned}$$

Now $v_k^*(s_i') = \mathbb{E}_{a\sim\pi_k(\cdot|s_i')}[q_k^*(s_i',a)] = \phi^{\pi_k}(s_i')^\top \boldsymbol{w}_k^*$. So, plugging this back into (17) we have

$$\boldsymbol{w}_k^* = \boldsymbol{\xi}_k^* + \hat{\Lambda}^{-1} \sum_{i=1}^N \phi(s_i,a_i)\left(r(s_i,a_i) - J_k^* + \phi^{\pi_k}(s_i')^\top \boldsymbol{w}_k^*\right)$$

where

$$\boldsymbol{\xi}_k^* = \hat{\Lambda}^{-1} \sum_{i=1}^N \phi(s_i,a_i)(P_{s_i,a_i} v_k^* - v_k^*(s_i')) + \hat{\Lambda}^{-1} \sum_{i=1}^N \phi(s_i,a_i)\Delta_k(s_i,a_i) + \hat{\Lambda}^{-1}\boldsymbol{w}_k^*. \qquad (18)$$

To complete the proof, we only need to show that $\|\boldsymbol{\xi}_k^*\|_{\hat{\Lambda}} \leq \beta$ holds with high probability. Since $v_k^* \in \mathcal{V}(B_w, B_{\boldsymbol{\theta}})$ for all $k \in [K]$, by Lemma C.1 we have

$$\begin{aligned}
\left\|\hat{\Lambda}^{-1} \sum_{i=1}^N \phi(s_i,a_i)(P_{s_i,a_i} v_k^* - v_k^*(s_i'))\right\|_{\hat{\Lambda}} &= \left\|\sum_{i=1}^N \phi(s_i,a_i)(P_{s_i,a_i} v_k^* - v_k^*(s_i'))\right\|_{\hat{\Lambda}^{-1}} \\
&\leq \Gamma(B_w, N, \delta, K, d)
\end{aligned}$$

holds for all k with probability at least $1-\delta$. Next, using the projection bound, Lemma E.3, and $|\Delta_k(s_i,a_i)| \le \kappa_{\mathcal{Q}(B_w),\Pi}$ implied by Assumption 3.3, we have

$$\left\|\hat{\Lambda}^{-1}\sum_{i=1}^{N}\phi(s_i,a_i)\Delta_k(s_i,a_i)\right\|_{\hat{\Lambda}} = \left\|\sum_{i=1}^{N}\phi(s_i,a_i)\Delta_k(s_i,a_i)\right\|_{\hat{\Lambda}^{-1}} \le \kappa_{\mathcal{Q}(B_w),\Pi}\sqrt{N}.$$

Finally, since $\hat{\Lambda}^{-1} \preceq I$,

$$\|\hat{\Lambda}^{-1}\boldsymbol{w}_k^*\|_{\hat{\Lambda}} = \|\boldsymbol{w}_k^*\|_{\hat{\Lambda}^{-1}} \le \|\boldsymbol{w}_k^*\|_2 \le B_w.$$

Therefore, by (18) and the triangle inequality,

$$\|\boldsymbol{\xi}_k^*\|_{\hat{\Lambda}} \le \Gamma(B_w,N,\delta,K,d) + \kappa_{\mathcal{Q}(B_w),\Pi}\sqrt{N} + B_w \le \beta.$$

holds with probability at least $1-\delta$. The proof is complete. □

We now move on to the proof of Lemma 6.3. Recall

$$\hat{v}_k(s) = \mathbb{E}_{a\sim\pi_k(\cdot|s)}[\hat{q}_k(s,a)]$$

as the definition of the empirical value function estimated by the critic. Since $\hat{v}_k(s) = \phi^{\pi_k}(s)^\top \boldsymbol{w}_k$, it follows from the definition of $(J_k, \boldsymbol{w}_k, \boldsymbol{\xi}_k)$ as the solution to the optimization problem (6) that

$$\boldsymbol{w}_k = \boldsymbol{\xi}_k + \hat{\Lambda}^{-1}\sum_{i=1}^{N}\phi(s_i,a_i)(r(s_i,a_i) - J_k + \hat{v}_k(s_i')). \tag{19}$$

We will need the following statement.

Lemma C.3. *With probability at least $1-\delta$, for each $k \in [K]$ we have*

$$|\mathbb{E}_{(s,a)\sim\mu}[\hat{q}_k(s,a) - \phi(s,a)^\top\bar{\boldsymbol{w}}_k]| \le 2\beta\|\phi^\mu\|_{\hat{\Lambda}^{-1}}$$

for any probability measure $\mu \in \Delta(\mathcal{S}\times\mathcal{A})$.

Proof. For this proof, we define

$$\bar{\Delta}_k(s,a) = \phi(s,a)^\top\bar{\boldsymbol{w}}_k + J_k - T^{\pi_k}\hat{q}_k(s,a).$$

We observe that

$$\begin{aligned}
\bar{\boldsymbol{w}}_k &= \hat{\Lambda}^{-1}\hat{\Lambda}\bar{\boldsymbol{w}}_k \\
&= \hat{\Lambda}^{-1}\sum_{i=1}^{N}\phi(s_i,a_i)\phi(s_i,a_i)^\top\bar{\boldsymbol{w}}_k + \hat{\Lambda}^{-1}\bar{\boldsymbol{w}}_k \\
&= \hat{\Lambda}^{-1}\sum_{i=1}^{N}\phi(s_i,a_i)(T^{\pi_k}\hat{q}_k(s_i,a_i) - J_k) + \hat{\Lambda}^{-1}\sum_{i=1}^{N}\phi(s_i,a_i)\bar{\Delta}_k(s_i,a_i) + \hat{\Lambda}^{-1}\bar{\boldsymbol{w}}_k.
\end{aligned}$$

Using (19), this allows us to write the difference in the parameters $\boldsymbol{w}_k$ and $\bar{\boldsymbol{w}}_k$ as

$$\begin{aligned}
\boldsymbol{w}_k - \bar{\boldsymbol{w}}_k &= \boldsymbol{\xi}_k + \hat{\Lambda}^{-1}\sum_{i=1}^{N}\phi(s_i,a_i)\left(r(s_i,a_i) - J_k + \hat{v}_k(s_i')\right) - \hat{\Lambda}^{-1}\sum_{i=1}^{N}\phi(s_i,a_i)(T^{\pi_k}\hat{q}_k(s_i,a_i) - J_k) \\
&\quad - \hat{\Lambda}^{-1}\sum_{i=1}^{N}\phi(s_i,a_i)\bar{\Delta}_k(s_i,a_i) - \hat{\Lambda}^{-1}\bar{\boldsymbol{w}}_k \\
&= \boldsymbol{\xi}_k + \hat{\Lambda}^{-1}\sum_{i=1}^{N}\phi(s_i,a_i)\left(\hat{v}_k(s_i') - P_{s_ia_i}\hat{v}_k\right) - \hat{\Lambda}^{-1}\sum_{i=1}^{N}\phi(s_i,a_i)\bar{\Delta}_k(s_i,a_i) - \hat{\Lambda}^{-1}\bar{\boldsymbol{w}}_k.
\end{aligned}$$

So we have

$$\begin{aligned}
\phi(s,a)^\top(\boldsymbol{w}_k - \bar{\boldsymbol{w}}_k) &= \phi(s,a)^\top\boldsymbol{\xi}_k + \phi(s,a)^\top\hat{\Lambda}^{-1}\sum_{i=1}^{N}\phi(s_i,a_i)(\hat{v}_k(s_i') - P_{s_i,a_i}\hat{v}_k) \\
&\quad - \phi(s,a)^\top\hat{\Lambda}^{-1}\sum_{i=1}^{N}\phi(s_i,a_i)\bar{\Delta}_k(s_i,a_i) - \phi(s,a)^\top\hat{\Lambda}^{-1}\bar{\boldsymbol{w}}_k.
\end{aligned}$$

Denote $\phi^\mu = \mathbb{E}_{(s,a)\sim\mu}[\phi(s,a)]$. Taking an expectation followed by an absolute value we get

$$\begin{aligned}|\mathbb{E}_{(s,a)\sim\mu}[\hat{q}_k(s,a) - \phi(s,a)^\top \bar{\boldsymbol{w}}_k]| &= |(\phi^\mu)^\top(\boldsymbol{w}_k - \bar{\boldsymbol{w}}_k)| \\ &\le |(\phi^\mu)^\top \boldsymbol{\xi}_k| + \left|(\phi^\mu)^\top \hat{\Lambda}^{-1} \sum_{i=1}^N \phi(s_i,a_i)(\hat{v}_k(s_i') - P_{s_i,a_i}\hat{v}_k)\right| \\ &\quad + \left|(\phi^\mu)^\top \hat{\Lambda}^{-1} \sum_{i=1}^N \phi(s_i,a_i)\bar{\Delta}_k(s_i,a_i)\right| + |(\phi^\mu)^\top \hat{\Lambda}^{-1}\bar{\boldsymbol{w}}_k|. \end{aligned} \tag{20}$$

Let us consider each of the above terms on the right hand side. First, we have by Holder's inequality

$$|(\phi^\mu)^\top \xi_k| \le \|\phi^\mu\|_{\hat{\Lambda}^{-1}} \|\boldsymbol{\xi}_k\|_{\hat{\Lambda}} \le \beta \|\phi^\mu\|_{\hat{\Lambda}^{-1}}$$

where the second inequality is due to the constraint $\|\boldsymbol{\xi}\|_{\hat{\Lambda}} \le \beta$ in the optimization problem (6). For the second term, using $\hat{v}_k \in \mathcal{V}(B_w, B_{\boldsymbol{\theta}})$ for all k, we appeal to Lemma C.1 to get

$$\begin{aligned}\left|(\phi^\mu)^\top \hat{\Lambda}^{-1} \sum_{i=1}^N \phi(s_i,a_i)(\hat{v}_k(s_i') - P_{s_i,a_i}\hat{v}_k)\right| &\le \|\phi^\mu\|_{\hat{\Lambda}^{-1}} \left\|\sum_{i=1}^N \phi(s_i,a_i)(\hat{v}_k(s_i') - P_{s_i,a_i}\hat{v}_k)\right\|_{\hat{\Lambda}^{-1}} \\ &\le \Gamma(B_w, N, \delta, K, d)\|\phi^\mu\|_{\hat{\Lambda}^{-1}}\end{aligned}$$

with probability at least $1-\delta$ for all k. For the third term, we use $|\bar{\Delta}_k(s,a)| \le \varepsilon_{\mathcal{Q}(B_w),\Pi}$ by Assumption 3.3 and Lemma E.3 to get

$$\begin{aligned}\left|(\phi^\mu)^\top \hat{\Lambda}^{-1} \sum_{i=1}^N \phi(s_i,a_i)\bar{\Delta}_k(s_i,a_i)\right| &\le \|\phi^\mu\|_{\hat{\Lambda}^{-1}} \left\|\sum_{i=1}^N \phi(s_i,a_i)\bar{\Delta}_k(s_i,a_i)\right\|_{\hat{\Lambda}^{-1}} \\ &\le \varepsilon_{\mathcal{Q}(B_w),\Pi}\sqrt{N}\|\phi^\mu\|_{\hat{\Lambda}^{-1}}.\end{aligned}$$

Finally, for the last term on the right hand side of (20) we simply observe that

$$|(\phi^\mu)^\top \hat{\Lambda}^{-1}\bar{\boldsymbol{w}}_k| \le \|\phi^\mu\|_{\hat{\Lambda}^{-1}}\|\bar{\boldsymbol{w}}_k\|_{\hat{\Lambda}^{-1}} \le \|\phi^\mu\|_{\hat{\Lambda}^{-1}}\|\bar{\boldsymbol{w}}_k\|_2 \le B_w\|\phi^\mu\|_{\hat{\Lambda}^{-1}}$$

Combining everything shows that with probability at least $1-\delta$,

$$\begin{aligned}|\mathbb{E}_{(s,a)\sim\mu}[\hat{q}_k(s,a) - \phi(s,a)^\top \bar{\boldsymbol{w}}_k]| &\le \left(\beta + \Gamma(B_w, B, \delta, K, d) + \varepsilon_{\mathcal{Q}(B_w),\Pi}\sqrt{N} + B_w\right)\|\phi^\mu\|_{\hat{\Lambda}^{-1}} \\ &\le 2\beta\|\phi^\mu\|_{\hat{\Lambda}^{-1}}\end{aligned}$$

as desired. □

Proof of Lemma 6.3. Adding and subtracting $\phi(s,a)^\top \bar{\boldsymbol{w}}_k$, we can decompose the Bellman error into two terms as

$$\hat{q}_k(s,a) + \hat{J}_k - T^{\pi_k}\hat{q}_k(s,a) = \hat{q}_k(s,a) - \phi(s,a)^\top \bar{\boldsymbol{w}}_k + \phi(s,a)^\top \bar{\boldsymbol{w}}_k + \hat{J}_k - T^{\pi_k}\hat{q}_k(s,a).$$

Therefore, by Lemma C.3 and the definition of $\bar{\boldsymbol{w}}_k$,

$$\begin{aligned}|\mathbb{E}_{(s,a)\sim\mu^\pi\otimes\pi}[\hat{q}_k(s,a) + \hat{J}_k - T^{\pi_k}\hat{q}_k(s,a)]| &\le |\mathbb{E}_{(s,a)\sim\mu^\pi\otimes\pi}[\hat{q}_k(s,a) - \phi(s,a)^\top \bar{\boldsymbol{w}}_k]| \\ &\quad + \|\phi(\cdot,\cdot)^\top \bar{\boldsymbol{w}}_k + \hat{J}_k - T^{\pi_k}\hat{q}_k\|_\infty \\ &\le 2\beta\|\phi^{\mu^\pi}\|_{\hat{\Lambda}^{-1}} + \varepsilon_{\mathcal{Q}(B_w),\Pi}.\end{aligned}$$

□

C.2 Proof of Theorem 5.1

We are now ready to prove Theorem 5.1. We will need the following Lemma analyzing the actor's optimization procedure.

Lemma C.4. *Fix a comparator policy $\pi \in \Delta(\mathcal{A})$. We have*

$$\mathbb{E}_{s\sim\mu^\pi}\left[\frac{1}{K}\sum_{k=1}^K \sum_a (\pi(a|s) - \pi_k(a|s))\hat{q}_k(s,a)\right] \le 2B_w\sqrt{\frac{\log A}{K}}.$$

Proof. Notice that the policy parameter update in Line 7 of Algorithm 1 can be written as the exponential weight update

$$\pi_{k+1}(a|s) \propto \pi_k(a|s)\exp\left(\eta\hat{q}_k(s,a)\right).$$

We have $|\hat{q}_k(s,a)| = |\phi(a,s)^\top \boldsymbol{w}_k| \le B_w$ by Algorithm 1's constraint on $\boldsymbol{w}_k$ and the normalization of $\phi(s,a)$ in Assumption 3.3. We then appeal to standard analysis of exponential weights (Lemma E.1). If we choose $\eta \le \frac{1}{B_w}$ then the conditions of the Lemma E.1 are satisfied and we have

$$\begin{aligned}\sum_{k=1}^{K}(\pi(a|s) - \pi_k(a|s))\hat{q}_k(s,a) &\le \frac{\log A}{\eta} + \eta \sum_{k=1}^{K}\sum_{a\in\mathcal{A}} \pi_k(a|s)|\hat{q}_k(s,a)|^2 \\ &\le \frac{\log A}{\eta} + \eta K B_w^2.\end{aligned}$$

We choose $\eta = \sqrt{\frac{\log A}{KB_w^2}}$, which is $\le \frac{1}{B_w}$ whenever $K \ge \log A$, to get

$$\sum_{k=1}^{K}(\pi(a|s) - \pi_k(a|s))\hat{q}_k(s,a) \le 2B_w\sqrt{K\log A}.$$

The proof is completed by dividing both sides by K, and taking an expectation with respect to $s \sim \mu^\pi$. □

Proof of Theorem 5.1. We work on the events of Lemmas 6.2 and Lemma 6.3 which, by a union bound, hold simultaneously with probability at least $1 - 2\delta$.

By Lemma 6.1 we have

$$\begin{aligned}\frac{1}{K}\sum_{j=1}^{K} J^\pi - \hat{J}_k = \frac{1}{K}\sum_{k=1}^{K} &\mathbb{E}_{s\sim\mu^\pi}\left[(\pi(a|s) - \pi_k(a|s))\hat{q}_k(s,a)\right] \\ &+ \frac{1}{K}\sum_{k=1}^{K}\mathbb{E}_{(s,a)\sim\mu^\pi\otimes\pi}\left[T^{\pi_k}\hat{q}_k(s,a) - \hat{J}_k - \hat{q}_k(s,a)\right].\end{aligned}$$

By Lemma C.4, the first term on the right hand side bounded by $2B_w\sqrt{\frac{\log A}{K}}$. For the second term, by Lemma 6.3

$$\begin{aligned}\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}_{(s,a)\sim\mu^\pi\otimes\pi}\left[T^{\pi_k}\hat{q}_k(s,a) - \hat{J}_k - \hat{q}_k(s,a)\right] &\le \frac{1}{K}\sum_{k=1}^{K}\left(2\beta\|\phi^{\mu^\pi}\|_{\hat{\Lambda}^{-1}} + \varepsilon_{B_w,B_{\boldsymbol{\theta}}}\right) \\ &= 2\beta\|\phi^{\mu^\pi}\|_{\hat{\Lambda}^{-1}} + \varepsilon_{B_w,B_{\boldsymbol{\theta}}}.\end{aligned}$$

We next observe that Lemma 6.2 implies

$$\begin{aligned}\frac{1}{K}\sum_{j=1}^{K} J^\pi - J^{\pi_k} &\le \frac{1}{K}\sum_{k=1}^{K} J^\pi - \hat{J}_k + \kappa_{B_w,B_{\boldsymbol{\theta}}} \\ &\le 2B_w\sqrt{\frac{\log A}{K}} + 2\beta\|\phi^{\mu^\pi}\|_{\hat{\Lambda}^{-1}} + \varepsilon_{B_w,B_{\boldsymbol{\theta}}} + \kappa_{B_w,B_{\boldsymbol{\theta}}}.\end{aligned}$$

To complete the proof, we use the definition of β and replace $\hat{\Lambda}^{-1}$ with $\frac{1}{N}\hat{\Lambda}_N^{-1}$. □

D Proof of Lemma C.1

This section is dedicated to the proof of Lemma C.1. We start with the following general result whose purpose will be to show that policies in Π are Lipchitz continuous in the parameter $\boldsymbol{\theta}$. This is similar to what is done in [Sherman et al., 2024].

Lemma D.1. *Let $\Theta \subset \mathbb{R}^d$ be a convex set and $\{f_\theta : \theta \in \Theta\}$ be a class of parameterized functions from $\mathbb{R}^d \to \mathbb{R}$. Suppose that the map $\theta \mapsto f_\theta(x)$ is continuously differentiable and that $\sup_{x\in\mathbb{R}^d, \|x\|_2\le 1, \theta\in\Theta}\|\nabla_\theta f_\theta(x)\|_2 \le L$ for some constant L. Then for any $\theta_1, \theta_2 \in \Theta$ and $s \in \mathcal{S}$,*

$$\|\pi_{\theta_1}(\cdot|s) - \pi_{\theta_2}(\cdot|s)\|_1 \le 2L\|\theta_1 - \theta_2\|_2$$

where

$$\pi_\theta(a|s) = \frac{e^{f_\theta(\phi(s,a))}}{\sum_{a'} e^{f_\theta(\phi(s,a))}}$$

and $\phi(s,a)$ is as in Assumption 3.3.

Proof. Note that

$$\sup_{(s,a),\boldsymbol{\theta}\in\boldsymbol{\Theta}} \|\nabla_{\boldsymbol{\theta}} f_{\boldsymbol{\theta}}(\phi(s,a))\|_2 \le L$$

since $\|\phi(s,a)\| \le 1$ under Assumption 3.3. Fix θ_1, θ_2 and a state $s \in \mathcal{S}$. Let $m = |\mathcal{A}|$ and let

$$J_\theta(s) = \begin{bmatrix} \nabla_\theta \pi_\theta(a_1|s) \\ \dots \\ \nabla_\theta \pi_\theta(a_m|s) \end{bmatrix} \in \mathbb{R}^{m\times d}$$

be the Jacobian of $\theta \mapsto \pi_\theta(\cdot|s)$. A straight forward computation shows that

$$\nabla_\theta \pi_\theta(a|s) = \pi_\theta(a|s)\left(\nabla_\theta f_\theta(\phi(s,a)) - \sum_{a'} \nabla_\theta f_\theta(\phi(s,a'))\pi_\theta(a'|s)\right). \tag{21}$$

Fix $t \in [0,1]$ and define $\theta_t = (1-t)\theta_1 + t\theta_2$. Let J_{θ_1,θ_2} be the matrix $\int_0^1 J_{\theta_t} dt$. Then by the fundamental theorem of calculus

$$\|\pi_{\theta_1}(\cdot|s) - \pi_{\theta_2}(\cdot|s)\|_1 = \|J_{\theta_1,\theta_2}(\theta_1 - \theta_2)\|_1.$$

Now let a be fixed. Using (21), we can bound the absolute value of the a-th entry of the vector $J_{\theta_1,\theta_2}(\theta_1-\theta_2)$ as

$$\begin{aligned}
|J_{\theta_1,\theta_2}(\theta_1-\theta_2)_a| &\le \left|\int_0^1 \pi_{\theta_t}(a|s)\nabla_\theta f_{\theta_t}(\phi(s,a))^\top(\theta_1-\theta_2)\right| \\
&\qquad + \left|\int_0^1 \pi_{\theta_t}(a|s)\sum_{a'}\nabla_\theta f_{\theta_t}(\phi(s,a'))^\top(\theta_1-\theta_2)\pi_{\theta_t}(a'|s)\right| \\
&\le \int_0^1 \pi_{\theta_t}(a|s)|\nabla_\theta f_{\theta_t}(\phi(s,a))^\top(\theta_1-\theta_2)|dt \\
&\qquad + \int_0^1 \pi_{\theta_t}(a|s)\sum_{a'}|\nabla_\theta f_{\theta_t}(\phi(s,a'))^\top(\theta_1-\theta_2)|\pi_{\theta_t}(a'|s)dt \\
&\le L\|\theta_1-\theta_2\|_2\int_0^1 \pi_{\theta_t}(a|s)dt + L\|\theta_1-\theta_2\|_2\int_0^1 \pi_{\theta_t}(a|s)\sum_{a'}\pi_{\theta_t}(a'|s)dt \\
&= 2L\|\theta_1-\theta_2\|_2\int_0^1 \pi_{\theta_t}(a|s)dt
\end{aligned}$$

where the second to last line we used Cauchy-Schwartz and $\|\nabla_\theta f_\theta(\phi(s,a))\| \le L$. Therefore

$$\|\pi_{\theta_1}(\cdot|s) - \pi_{\theta_2}(\cdot|s)\| = \sum_a |J_{\theta_1,\theta_2}(\theta_1-\theta_2)_a| \le 2L\|\theta_1-\theta_2\|_2\int_0^1\sum_a \pi_{\theta_t}(a|s)dt = 2L\|\theta_1-\theta_2\|_2.$$

□

Lemma D.1 can be applied to the restricted policy class $\Pi(B_{\boldsymbol{\theta}})$ by taking $\boldsymbol{\Theta}$ to be the d-dimensional Euclidean ball of radius $B_{\boldsymbol{\theta}}$ and the family $\{f_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$ to be the set of functions $f_{\boldsymbol{\theta}}(\phi) = \phi^\top \boldsymbol{\theta}$ where $\|\phi\|_2 \le 1$. Clearly, we have $\|\nabla_{\boldsymbol{\theta}} f_{\boldsymbol{\theta}}(\phi)\|_2 \le 1$. Thus, Lemma D.1 shows that

$$\|\pi_{\boldsymbol{\theta}_1}(\cdot|s) - \pi_{\boldsymbol{\theta}_2}(\cdot|s)\|_1 \le 2\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|_2 \tag{22}$$

for all $\pi_{\boldsymbol{\theta}_1}, \pi_{\boldsymbol{\theta}_2} \in \Pi(B_{\boldsymbol{\theta}})$.

With this we can now bound the ε-covering number of the function class $\mathcal{V}$.

Lemma D.2. *Let $\mathcal{V}$ be the function class (16). Then*

$$\log \mathcal{N}_\varepsilon(\mathcal{V}) \le d\log\left(1 + 4B_w/\varepsilon\right) + d\log\left(1 + 16B_w B_{\boldsymbol{\theta}}/\varepsilon\right).$$

where $\mathcal{N}_\varepsilon(\mathcal{V})$ is the ε-covering number of $\mathcal{V}$ with respect to the norm $\|\cdot\|_\infty$

Proof. First, consider two functions $q(\cdot,\cdot;\boldsymbol{w})$ and $q(\cdot,\cdot;\boldsymbol{w}')$ in $\mathcal{Q}(B_w)$. Using the normalization in Assumption 3.3

$$|q(s,a;\boldsymbol{w}) - q(s,a;\boldsymbol{w}')| = |\langle\phi(s,a), \boldsymbol{w} - \boldsymbol{w}'\rangle| \le \|\boldsymbol{w} - \boldsymbol{w}'\|_2$$

for any (s, a). So for any fixed policy $\pi \in \Pi(B_{\boldsymbol{\theta}})$ we have

$$\begin{aligned}|v(s;\pi,\boldsymbol{w}) - v(s;\pi,\boldsymbol{w}')| &\le \max_s \left|\sum_a \pi(a|s)(q(s,a,\boldsymbol{w}) - q(s,a,\boldsymbol{w}'))\right| \\ &\le \max_{s,a} |q(s,a,\boldsymbol{w}) - q(s,a,\boldsymbol{w}')| \\ &\le \|\boldsymbol{w} - \boldsymbol{w}'\|_2.\end{aligned}$$

On the other hand, for a fixed $\boldsymbol{w}$, and separate policies $\pi_{\boldsymbol{\theta}}, \pi_{\boldsymbol{\theta}'} \in \Pi$,

$$\begin{aligned}|v(s,\pi,\boldsymbol{w}) - v(s,\pi',\boldsymbol{w})| &\le \max_s \left|\sum_a (\pi_{\boldsymbol{\theta}}(a|s) - \pi_{\boldsymbol{\theta}'}(a|s)) q(s,a;\boldsymbol{w})\right| \\ &\le B_w \max_s \|\pi_{\boldsymbol{\theta}}(\cdot|s) - \pi_{\boldsymbol{\theta}'}(\cdot|s)\|_1 \\ &\le 4B_w \|\boldsymbol{\theta} - \boldsymbol{\theta}'\|_2\end{aligned}$$

where the second line used $|q(s,a);\boldsymbol{w})| \le B_w$ and the last line used (22). Thus it holds for any $v(\cdot,\pi,\boldsymbol{w})$, $v(\cdot,\pi',\boldsymbol{w}') \in \Pi$,

$$\begin{aligned}|v(s,\pi,\boldsymbol{w}) - v(s,\pi',\boldsymbol{w}')| &\le |v(s,\pi,\boldsymbol{w}) - v(s,\pi,\boldsymbol{w}')| + |v(s,\pi,\boldsymbol{w}') - v(s,\pi',\boldsymbol{w}')| \\ &\|\boldsymbol{w} - \boldsymbol{w}'\|_2 + 4B_w \|\boldsymbol{\theta} - \boldsymbol{\theta}'\|_2.\end{aligned} \tag{23}$$

Now, using a standard result concerning the covering number of the d-dimensional Euclidian ball (Lemma E.2), we can construct an $\frac{\varepsilon}{2}$ covering of the euclidean ball d-dimensional euclidean ball of radius B_w with cardinality at most $(1 + 4B_w/\varepsilon)^d$ and an $\frac{\varepsilon}{8B_w}$ covering of the Euclidean ball of radius $B_{\boldsymbol{\theta}}$ with cardinality not exceeding $(1 + 16B_w B_{\boldsymbol{\theta}}/\varepsilon)^d$. Let $\mathcal{V}_\varepsilon$ be the members of $\mathcal{V}$ parameterized by members $(\boldsymbol{w}', \boldsymbol{\theta}')$ of the Cartesian product of these two coverings. Then

$$\log \mathcal{N}_\varepsilon(\mathcal{V}) = \log |\mathcal{V}_\varepsilon| \le d \log\left(1 + 4B_w/\varepsilon\right) + d \log\left(1 + 16 B_w B_{\boldsymbol{\theta}}/\varepsilon\right),$$

and by (23), for any $v(\cdot;\pi_{\boldsymbol{\theta}},\boldsymbol{w}) \in \mathcal{V}$ we can find $v(\cdot;\pi_{\boldsymbol{\theta}'},\boldsymbol{w}') \in \mathcal{V}_\varepsilon$ with

$$\begin{aligned}|v(s;\pi_{\boldsymbol{\theta}},\boldsymbol{w}) - v(s,\pi_{\boldsymbol{\theta}'},\boldsymbol{w}')| &\le \|\boldsymbol{w} - \boldsymbol{w}'\|_2 + 4B_w \|\boldsymbol{\theta} - \boldsymbol{\theta}'\|_2 \\ &\le \frac{\varepsilon}{2} + 4B_w \cdot \frac{\varepsilon}{8B_w} = \varepsilon.\end{aligned}$$

□

Proof of Lemma C.1. Fix $\{v_k\}_{k=1}^K \subset \mathcal{V}(B_w, B_{\boldsymbol{\theta}})$ Appealing to the uniform concentration of self-normalized processes (Lemma E.5) and using that $\|v\|_\infty \le B_w$ for any $v \in \mathcal{V}$ we have, for fixed k,

$$\left\|\sum_{i=1}^N \phi(s_i,a_i)\left(P_{s_i,a_i} v_k - v_k(s_i')\right)\right\|^2_{\hat{\Lambda}^{-1}} \le 4B_w^2 \left(\frac{d}{2}\log\left(\frac{K(N+1)}{\delta}\right) + \log \mathcal{N}_\varepsilon(\mathcal{V})\right) + 8N^2\varepsilon^2$$

holds with probability $1 - \frac{\delta}{K}$. Now, substituting the bound for $\log \mathcal{N}_\varepsilon(\mathcal{V})$ from Lemma D.2, the upper bound becomes

$$4B_w^2 \left(\frac{d}{2}\log\left(\frac{K(N+1)}{\delta}\right) + d\log(1 + 4B_w/\varepsilon) + d\log(1 + 16 B_w B_{\boldsymbol{\theta}}/\varepsilon)\right) + 8N^2\varepsilon^2.$$

Taking $\varepsilon = 1/N$ followed by a union bound for over $k = 1, \ldots, K$ we complete the proof. □

E Auxiliary Lemmas

Lemma E.1. *Let $\{X_k\}_{k\ge1}$ be a sequence of vectors in $\mathbb{R}^A$. Set $\pi_1(a) = \frac{1}{A}$ for all $a \in \mathcal{A}$ and for each $k \ge 2$,*

$$\pi_{k+1}(a) = \frac{\pi_k(a)\exp(\eta X_k(a))}{\sum_{a' \in \mathcal{A}} \pi_k(a')\exp(\eta X_k(a'))}$$

for some positive stepsize η satisfying $\eta X_k(a) \ge -1$ for all k and $a \in \mathcal{A}$. Then for any fixed $\pi^* \in \Delta(\mathcal{A})$,

$$\sum_{k=1}^K \langle \pi^* - \pi_k, X_k\rangle \le \frac{\log A}{\eta} + \eta \sum_{k=1}^K \sum_{a\in\mathcal{A}} \pi_k(a) X_k(a)^2.$$

Proof. Define $Z_k = \sum_{a' \in \mathcal{A}} \pi_k(a') \exp(\eta X_k(a'))$. Then using the inequality $e^x \leq 1 + x + x^2$ which holds for all $x \leq 1$ followed by $\log(1+x) \leq x$ we have

$$\log Z_k = \log\left(\sum_{a' \in \mathcal{A}} \pi_k(a') \exp(\eta X_k(a'))\right) \leq \log\left(1 + \sum_{a \in \mathcal{A}} \pi_k(a)\eta X_k(a) + \sum_{a \in \mathcal{A}} \pi_k(a)\eta^2 X_k(a)^2\right)$$
$$\leq \eta \sum_{a \in \mathcal{A}} \pi_k(a) X_k(a) + \eta^2 \sum_{a \in \mathcal{A}} \pi_k(a) X_k(a)^2.$$

Thus,

$$D_{KL}(\pi^*||\pi_{k+1}) - D_{KL}(\pi^*||\pi_k) = \sum_{a \in \mathcal{A}} \pi^*(a) \log\left(\frac{\pi^*(a)}{\pi_{k+1}(a)}\right) - \sum_{a \in \mathcal{A}} \pi^*(a) \log\left(\frac{\pi^*(a)}{\pi_k(a)}\right)$$
$$= \sum_{a \in \mathcal{A}} \pi^*(a) \log\left(\frac{\pi^*(a) Z_k \exp(-\eta X_k(a))}{\pi_k(a)}\right) - \sum_{a \in \mathcal{A}} \pi^*(a) \log\left(\frac{\pi^*(a)}{\pi_k(a)}\right)$$

$$= \log Z_k - \eta \sum_{a \in \mathcal{A}} \pi^*(a) X_k(a)$$
$$\leq \eta \sum_{a \in \mathcal{A}} \pi_k(a) X_k(a) + \eta^2 \sum_{a \in \mathcal{A}} \pi_k(a) X_k(a)^2 - \eta \sum_{a \in \mathcal{A}} \pi^*(a) X_k(a).$$

Rearranging and summing from $k = 1$ to K,

$$\eta \sum_{k=1}^{K} \sum_{a \in \mathcal{A}} (\pi^*(a) - \pi_k(a)) X_k(a) \leq \sum_{k=1}^{K} D_{KL}(\pi^*||\pi_k) - D_{KL}(\pi^*||\pi_{k+1}) + \eta^2 \sum_{k=1}^{K} \sum_{a \in \mathcal{A}} \pi_k(a) X_k(a)^2$$
$$\leq D_{KL}(\pi^*||\pi_1) + \eta^2 \sum_{k=1}^{K} \sum_{a \in \mathcal{A}} \pi_k(a) X_k(a)^2.$$

Since $\pi_1(a) = \frac{1}{A}$ for all a,

$$D_{KL}(\pi^*||\pi_1) = \sum_{a \in \mathcal{A}} \pi^*(a) \log\left(A\pi^*(a)\right) \leq \log A.$$

Plugging this bound in above and dividing both sides by η we complete the proof. □

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

Proof. See [Jin et al., 2020] Lemma D.5. □

Lemma E.3 (Projection Bound)**.** *Let $\{\phi_i\}_{i \geq 1}$ be a sequence in $\mathbb{R}^d$ with $\|\phi_i\|_2 \leq 1$ and $\{a_i\}_{i \geq 1}$ be a sequence of real numbers with $|a_i| \leq A$. Define*

$$\Lambda_n = \sum_{i=1}^{n} \phi_i \phi_i^\top + I.$$

We have

$$\left\|\sum_{i=1}^{n} a_i \phi_i\right\|_{\Lambda_n^{-1}} \leq A\sqrt{n}.$$

Proof. See [Zanette et al., 2020] Lemma 8. □

Lemma E.4 (Concentration of Self-Normalized Processes)**.** *Let $(X_t)_t$ be a real-valued martingale difference sequence adapted to filtration $(\mathcal{F}_t)_t$. Suppose that X_t is σ-subgaussian conditioned on $\mathcal{F}_{t-1}$ i.e.,*

$$\log \mathbb{E}[e^{\lambda X_t} | \mathcal{F}_{t-1}] \leq \frac{\lambda^2 \sigma^2}{2}.$$

Let $(\phi_t)_t$ be an $\mathbb{R}^d$-valued predictable process. Assume $\Lambda_0 \in \mathbb{R}^{d \times d}$ is positive definite and let $\Lambda_t = \Lambda_0 + \sum_{s=1}^{t} \phi_s \phi_s^\top$. Then for any $\delta > 0$, with probability at least $1 - \delta$ we have

$$\left\|\sum_{s=1}^{t} \phi_s X_s\right\|_{\Lambda_t^{-1}}^2 \leq 2\sigma^2 \log\left(\frac{\sqrt{det(\Lambda_t) det(\Lambda_0)}}{\delta}\right).$$

Proof. See [Jin et al., 2020] Lemma D.3. □

Lemma E.5 (Lemma D.4 in [Jin et al., 2020])**.** *Let $\{x_t\}_{t=1}^{\infty}$ be a stochastic process on state space $\mathcal{S}$ with corresponding filtration $\{\mathcal{F}_t\}_{t=1}^{\infty}$. Let $\{\phi_t\}_{t=1}^{\infty}$ be an $\mathbb{R}^d$-valued stochastic process where $\phi_t \in \mathcal{F}_{t-1}$ and $\|\phi_t\| \leq 1$. Let $\Lambda_k = \sum_{t=1}^{k} \phi_t \phi_t^\top$. Then for any $\delta > 0$, with probability at least $1-\delta$, for all $k \geq 0$, and any $V \in \mathcal{V}$ so that $\|V\|_\infty \leq H$, we have:*

$$\left\| \sum_{t=1}^{k} \phi_t \left(V(x_t) - \mathbb{E}[V(x_t)|\mathcal{F}_{t-1}] \right) \right\|_{\Lambda_k^{-1}}^2 \leq 4H^2 \left(\frac{d}{2} \log\left(\frac{k+\lambda}{\lambda} \right) + \log \frac{\mathcal{N}_\varepsilon(\mathcal{V})}{\delta} \right) + \frac{8k^2\varepsilon^2}{\lambda},$$

where $\mathcal{N}_\varepsilon(\mathcal{V})$ is the ε-covering number of $\mathcal{V}$ with respected to the distance dist$(V, V') = \sup_x |V(x) - V'(x)|$.