Provable Benefits of Actor-Critic Methods for Offline Reinforcement Learning

Andrea Zanette
Institute for Computational and Mathematical Engineering
Stanford University
zanette@stanford.edu

Martin Wainwright
Department of Electrical Engineering and Computer Sciences
Department of Statistics
University of California at Berkeley
wainwrig@berkeley.edu

Emma Brunskill
Department of Computer Science
Stanford University
ebrun@cs.stanford.edu

Abstract

Actor-critic methods are widely used in offline reinforcement learning practice but are understudied theoretically. In this work we show that the pessimism principle can be naturally incorporated into actor-critic formulations. We create an offline actor-critic algorithm for a linear MDP model more general than the low-rank model. The procedure is both minimax optimal and computationally tractable.

1 Introduction

The problem of learning a near-optimal policy is a core challenge in reinforcement learning (RL). In many settings, it is desired to estimate a good policy using only a pre-collected set of data, and without the possibility of further interaction with the environment; this problem is known as *offline policy learning*. The offline setting has unique challenges due to the incomplete information about the Markov decision process (MDP) encoded in the available dataset. For example, due to the maximization bias, a naive offline algorithm can settle for a policy with a dangerously high estimated value even if such value is highly uncertain. In order to avoid this phenomenon, researchers have introduced the idea of *pessimism* [Liu et al., 2020, Jin et al., 2020b, Buckman et al., 2020, Kumar et al., 2019, Kidambi et al., 2020, Yu et al., 2020]; see Appendix A for additional discussion of the related literature.

Incorporating pessimism prevents algorithms from settling down on uncertain policies whose value might be misleadingly high under the current dataset due to statistical errors. By using pessimism, uncertain policies are penalized and only those robust to statistical errors are returned. The principle can be implemented in at least two different ways: (a) by penalizing policies that are far from the one that generated the dataset; or (b) by penalizing the value functions of policies not well covered by the dataset. In this work, we take the latter avenue.

Preprint. Under review.

Challenges: Implementing pessimism with function approximation is challenging for several reasons. First, uncertainty must be estimated with particular care, because underestimating it may not lead to an effective algorithm and overestimating it leads to policies that are too conservative and thus underperform. Second, the incorporation of pessimism may introduce complex, higher order perturbations into the value function class handled by the algorithm; this phenomenon is similar to adding optimistic bonuses in the exploration . I This increased complexity of the function class often requires additional assumptions on the model, because the new class needs to interact "nicely" with the Bellman operator. Prior art on pessimism with function approximation has bypassed this problem by making strong model assumptions, such as low-rank transitions [Jin et al., 2020b] or algorithm-specific assumptions [Liu et al., 2020].

Actor-critic methods: Most past theoretical work has focused on algorithms that are either model or value-based [Liu et al., 2020, Jin et al., 2020b, Buckman et al., 2020, Kidambi et al., 2020, Yu et al., 2020]; however, in practice, actor-critic methods are widely used [Levine et al., 2020, Wu et al., 2019, Wu et al., 2021, Kumar et al., 2019, Kumar et al., 2020]. An actor-critic method generally consists of an actor that changes the policy in order to maximize its value as estimated by the critic. Given their popularity, it is natural to ask the following question: *do actor-critic methods provably offer any advantage in offline RL?* The main contribution of this paper is to give a positive answer to this question: by separating the policy optimization from the policy evaluation, both tasks become simpler to design and the pessimism principle can be incorporated more naturally.

Contributions: In this paper, we focus on the problem using linear function approximation, and assume that we are given a batch dataset $\mathcal{D}$ of states, actions, rewards and successor states. Using $\mathcal{D}$ we can construct the set $\mathcal{M}$ of statistically plausible MDPs, i.e., a set that contains with high probability the MDP that generated the available dataset.

Our objective is then to find the policy that performs the best in the face of uncertainty. For a statistically plausible MDP $M \in \mathcal{M}$ and a policy π, let $V^{\pi}_{M}(s_1)$ be the value function of π on M at the initial state s_1. With high probability, $\inf_{M\in\mathcal{M}} V^{\pi}_{M}(s_1)$ is a lower bound on the value of π from s_1 on the 'real' MDP that generated the dataset. Thus, the policy with the highest high probability lower bound on its performance is simply

$$\sup_{\pi} \inf_{M\in\mathcal{M}} V^{\pi}_{M}(s_1). \tag{1}$$

Actor-critic methods fit naturally in this framework: the *actor* solves the outer maximization problem over policies which are evaluated in the inner minimization problem by a pessimistic *critic*. This way, each algorithm solves a simple task: 1) the critic provides a pessimistic value function estimate for a fixed policy (the one currently examined by the actor) while 2) the actor ensures online learning-style guarantees with respect to a sequence of pessimistic MDPs implicitly identified by the critic. This is the first algorithmic idea and leads to a computationally tractable implementation.

The second algorithmic idea is to introduce pessimism without altering the prescribed function class (see e.g., [Zanette et al., 2020b]). This is achieved by perturbing the value function (in the critic) within its prescribed functional space without adding pessimistic bonuses or absorbing states. This has two core advantages. First, there are no additional model assumptions compared to the vanilla (i.e., without pessimism) version of our actor-critic method; this is because the original value function class is not modified by the injection of pessimism. Second, the algorithm operates on value functions with the original statistical complexity, enabling the construction of tight confidence intervals and ultimately minimax statistical rates.

Notation: We let $\mathcal{B}_d(r) = \{x \in \mathbb{R}^d \mid \|x\|_2 \le r\}$ denote the Euclidean ball of radius $r \in \mathbb{R}$ in dimension d; we simply write $\mathcal{B}$ when there is no possibility of confusion. We use the standard O (or Ω) notation to denote an upper (or lower) bound that holds up to a universal constant.in the upper and We use the $\widetilde{O}$ to denote an upper bound that holds up to constants and log factors in the input parameters $(\frac{1}{\delta}, d, H, \lambda)$. The notation $\lesssim$ means $\le$ up to a constant while $\lessapprox, \approx, \gtrapprox, \asymp$ are used to highlight dominant terms in the proof sketch without rigorous mathematical definitions. For a vector $x \in \mathbb{R}^d$ we let $[x]_i$ denote its i component.

2 Background

We begin by providing some background on the undiscounted finite-horizon Markov decision processes that we study in this paper; see the book [Puterman, 1994] for more detail. A finite-horizon MDP is specified by a positive integer H, corresponding to the number of stages. The underlying dynamics involve a state-space $\mathcal{S}$, and are controlled by actions taking place in an action set $\mathcal{A}$. For every $h \in [H] = \{1, \ldots, H\}$, there is a reward function $r_h : \mathcal{S} \times \mathcal{A} \rightarrow$, and for every h and state-action pair (s, a), there is a transition function $\mathbb{P}_h(\cdot \mid s, a)$. When at horizon h, if the agent takes action a in state s, it receives a random reward drawn from some distribution $R_h(s, a)$ with mean $r_h(s, a)$, and it then transitions randomly to a next state s^+ drawn from the transition kernel $\mathbb{P}_h(\cdot \mid s, a)$.

For any triple $(s, a, h) \in \mathcal{S} \times \mathcal{A} \times [H]$, a non-stationary policy $\pi = (\pi_1, \ldots, \pi_H)$ is defined as

$$Q_h^\pi(s, a) = r_h(s, a) + \mathbb{E}_{s_\ell \sim \pi \mid (s,a)} \sum_{\ell=h+1}^{H} r_\ell(s_\ell, \pi_\ell(s_\ell)), \tag{2}$$

where the expectation is over the trajectories induced by π upon starting from the pair (s, a). When we omit the starting state-action pair (s, a), the expectation is intended to start from a fixed state denoted by s_1. The value function associated to π is $V_h^\pi(s) = Q_h^\pi(s, \pi_h(s))$. Under some regularity conditions, e.g., [Shreve and Bertsekas, 1978], there always exists an optimal policy $\pi^\star$ whose value and action-value functions are defined as $V_h^\star(s) = V_h^{\pi^\star}(s) = \sup_\pi V_h^\pi(s)$ and $Q_h^\star(s, a) = Q_h^{\pi^\star}(s, a) = \sup_\pi Q_h^\pi(s, a)$. We define the Bellman evaluation operator

$$\mathcal{T}_h^\pi(Q_{h+1})(s, a) = r_h(s, a) + \mathbb{E}_{s' \sim \mathbb{P}_h(s,a)} \mathbb{E}_{A' \sim \pi} Q_{h+1}(s', a')$$

2.1 Assumptions on data generation

In this paper, we study a model in which we receive N_h samples for every timestep, not necessarily from trajectories. Each sample consists of a state, an action, a reward and a next state. For the i sample, we let h_i be the horizon of the state-action and let n_i be the total number of samples collected at level h_i, including sample i. We are given a dataset $\mathcal{D} = \{(s_{h_i n_i}, a_{h_i n_i}, r_{h_i n_i}, s^+_{h_i n_i})\}_{i=1,2,\ldots,\sum_{h=1}^H N_h}$ of $\sum_{h=1}^H N_h$ state-action-reward-next states generated by the underlying MDP, possibly in an adaptive fashion.

Assumption 1 (Data Generation). *Assume that for every sample i in $\mathcal{D}$, conditioned on $(s_{h_i n_i}, a_{h_i n_i})$, the random reward is drawn from a distribution $R_h(s_{h_i n_i}, a_{h_i n_i})$ with mean $r_h(s_{h_i n_i}, a_{h_i n_i})$ that is 1-sub-Gaussian. The dataset $\mathcal{D}$ is such that*

$$r_{h_i n_i} \sim R(s_{h_i n_i}, a_{h_i n_i}), \qquad s^+_{h_i n_i} \sim \mathbb{P}_h(s_{h_i n_i}, a_{h_i n_i}) \tag{3}$$

where each pair $(s_{h_i n_i}, a_{h_i n_i})$ is allowed to depend on all previously sampled quadruples $(s_{h_j n_j}, a_{h_j n_j}, r_{h_j n_j}, s^+_{h_j n_j})$ for $j < i$.

This allows considerable freedom: (a) the dataset may be generated from (mixture) policies or by another mechanism that collects information at different state-actions; and (b) the dataset may be generated by an adversarial procedure that changes the data acquisition strategy as feedback is received.

2.2 Policy and Value Function Class

Next, we define the policy space Π and the action value function space $\mathcal{Q}$ where we seek solutions. For a fixed timestep h (which we omit here for brevity), consider a fixed feature extractor $\phi : \mathcal{S} \times \mathcal{A} \mapsto \mathbb{R}^d$, $\|\phi(\cdot, \cdot)\|_2 \le 1$ and two radii, $r_w \in (0, 1]$, $r_\theta > 0$ for the value function parameter w and for the policy parameter θ.

Definition 1 (Functional Spaces).

$$\mathcal{Q}(\rho^w) = \{(s, a) \mapsto \langle \phi(s, a), w \rangle \mid \|w\|_2 \le \rho^w\}, \qquad \Pi_{soft}(\rho^\theta) \overset{def}{=} \left\{ \frac{\exp[\langle \phi(s, a), \theta \rangle]}{\sum_{a'} \exp[\langle \phi(s, a'), \theta \rangle]} \mid \|\theta\|_2 \le \rho^\theta \right\}.$$

Algorithmically, the actor starts from $\rho^\theta_1 = 0$ (i.e., $\theta_1 = 0$) and implicitly enlarges ρ^θ_k in the k iteration by using the update rule in Line 5 of Algorithm 1. The policy radius can be large $\rho^\theta \gg 1$ but we constrain $\rho^w \le 1$ so that the critic's estimate $Q_w(s,a) = \langle \phi(s,a),\, w\rangle$ is bounded by one, i.e., $\sup_{(s,a,w)} |Q_w(s,a)| \le 1$. In finite horizon problems one can select different feature extractors ϕ_h in every step h; this generates H functional spaces $\mathcal{Q}_1, \dots, \mathcal{Q}_H$ and $\Pi_1, \dots, \Pi_H$. We drop the dependence on the radii when referring to the functional spaces and implicitly assume that the terminal value function is zero.

2.3 Assumptions on Function Class

If we seek to find the policy $\pi \in \Pi$ with the highest value function, it seems reasonable to require that the following representation condition (approximately) holds. We assume a common feature extractor $\phi : \mathcal{S} \times \mathcal{A},\ \|\phi(\cdot,\cdot)\|_2 \le 1$ throughout this section.

Assumption 2 (Linear Q^π)**.** *We say the MDP admits a linear action-value function representation for all policies in Π if Q^π is linear, i.e., for for each policy $\pi \in \Pi$ and $h \in [H]$, there exists a vector w^π_h such that $Q^\pi_h(s,a) = \langle \phi_h(s,a),\, w^\pi_h\rangle$.*

Unfortunately, [Zanette, 2020] establishes that even under such assumption, we might need exponentially many samples to do better than a random policy (see also [Weisz et al., 2020] for a weaker statement using only realizability). This suggests we need even stronger conditions. One such condition is the assumption we make in this work, which allows a classical temporal-difference critic to evaluate the policies in Π.

Assumption 3 (Restricted Closedness)**.** *The policy and value function spaces $(\Pi, \mathcal{Q})$ are closed up to $\nu \in \mathbb{R}^H$ error in the sup-norm if there is a sequence $\{\nu_h\}_{h=1}^H$ such that for each $h \in [H]$, we have*

$$\sup_{\substack{Q_{h+1} \in \mathcal{Q}_{h+1} \\ \pi_{h+1} \in \Pi_{h+1}}} \inf_{Q_h \in \mathcal{Q}_h} \|Q_h - \mathcal{T}_h^{\pi_{h+1}} Q_{h+1}\|_\infty \le \nu_h. \tag{4}$$

The restricted closedness assumption measures how well we can fit the action-value function resulting from the application of the Bellman evaluation operator to an action value function in $\mathcal{Q}$ and for a policy in Π. It enables the analysis of the classical *Least Square Policy Evaluation* (LSPE) [Nedić and Bertsekas, 2003], which will be our starting point when constructing the critic.

A related model assumption is the *low-rank* or *linear* MDP model [Jin et al., 2020a, Yang and Wang, 2020] used by the state of the art for offline RL with pessimismistic guarantees [Jin et al., 2020b] and much of the online RL literature [Agarwal et al., 2020a, Modi et al., 2021, Zanette et al., 2020a].

Assumption 4 (Low-Rank MDP)**.** *We say that an MDP is low-rank if $\forall h \in [H]$ there exist a reward parameter $w^R_h \in \mathbb{R}^{d_h}$ and a component-wise positive mapping $\psi_h : \mathcal{S} \to \mathbb{R}^d_+$ such that $\|\psi_h(s)\|_1 = 1$ for all $s \in \mathcal{S}$ and*

$$r_h(s,a) = \left\langle \phi_h(s,a),\, w^R_h \right\rangle, \qquad \mathbb{P}_h(s' \mid s,a) = \langle \phi_h(s,a),\, \psi_h(s')\rangle, \qquad \forall (s,a,h,s'). \tag{5}$$

We clarify the relation between these assumptions in the following proposition where we assume that the value function parameter $w \in \mathbb{R}^d$ for simplicity; the proof is deferred to Appendix B.

Proposition 1 (Low Rank $\subset$ Restricted Closedness $\subset$ Linear Q^π)**.** *For any fixed state-action space, horizon, and feature extractor:*

(a) The class of low-rank MDPs is a strict subset of the class of MDPs such that restricted closedness holds

(b) The class of MDPs such that restricted closedness holds is a subset of the MDP class where the linear Q^π assumption holds.

Furthermore, all inclusions are strict.

Operating under assumptions stronger than linear Q^π enables polynomial sample complexity. Our algorithm can successfully operate in the low-rank framework as a special case.

3 Intuition and Algorithmic Choices

We next describe in detail the algorithm PACLE (*Pessimistic Actor Critic for Learning without Exploration*). It consists of an actor (Algorithm 1) and a critic (Algorithm 2).

Algorithm 1 ACTOR (MIRROR DESCENT)

```
Input: Dataset $\mathcal{D}$, starting state $s_1$
Set $\theta_1 = (\vec{0}, \dots, \vec{0})$
for $t = 1, 2, \dots, T$ do
    $\underline{w}_t \leftarrow$ CRITIC($\mathcal{D}, \pi_{\theta_t}, s_1$)
    $\theta_{t+1} = \theta_t + \eta \underline{w}_t$
end for
Mixture policy $\pi_{\theta_1}, \dots, \pi_{\theta_T}$
```

Algorithm 2 CRITIC (PLSPE)

```
Input: Dataset $\mathcal{D}$, target policy $\pi$, starting state $s_1$
Solve the optimization program (9)
Return $\underline{w}$
```

The Critic: Pessimistic Least Square Policy Evaluation The purpose of the critic is to provide pessimistic value function estimates corresponding to the policy π under consideration by the actor. Monte Carlo with importance sampling is not appropriate here, as the policy or distribution that generated the dataset might be unknown. This suggests we use a temporal difference method like LSPE, perturbed to return *pessimistic* value function estimates (Algorithm 2); we name this PLSPE.

Our method is based on directly perturbing the regression parameters in the least-square estimate. In contrast to bonus-based approaches, this method has the important advantage of ensuring that the action-value function remains linear. The purpose of the perturbations is to compensate for possible statistical errors in estimating the regression parameter due to poor coverage of the given dataset.

Overall, given a policy $\pi = (\pi_1, \dots, \pi_H)$, the goal of the critic is to minimize the quantity

$$\mathbb{E}_{A' \sim \pi_1} \langle \phi(s_1, a),\, w_1 \rangle = \sum_{a \in \mathcal{A}} \pi_1(a \mid s_1) \langle \phi_1(s_1, a),\, w_1 \rangle \,, \tag{6}$$

which is an estimate of the value function $V^{\pi}(s_1)$ for the policy π at the initial state s_1. The parameter $w_1 \in \mathbb{R}^d$ is a vector to be adjusted, one that is determined by a backwards-running sequence of regression problems from $h = H$ down to $h = 1$.

We introduce the pessimistic perturbations directly to the solution of these regression problems. They involve a norm defined by the cumulative covariance matrix. For each $h \in [H]$ and $n_h \in [N_h]$, we introduce the shorthand notation $\phi_{hn} = \phi_h(s_{hn}, a_{hn})$, and define the *cumulative covariance matrix*

$$\Sigma_h \overset{def}{=} \sum_{n=1}^{N_h} \phi_{hn} \phi_{hn}^{\top} + \lambda I \qquad \text{for each } h \in [H]. \tag{7}$$

Here $\lambda > 0$ is a user-defined regularization parameter. Notice that the cumulative covariance strictly 'grows' with more samples; we do not normalize it by the number of samples so that Σ_h effectively represents the amount of information contained in the batch datset. Since Σ_h is strictly positive definite by construction, it defines a pair of norms

$$\|u\|_{\Sigma_h} \overset{def}{=} \sqrt{u^{\top} \Sigma_h u}, \quad \text{and} \quad \|u\|_{\Sigma_h^{-1}} \overset{def}{=} \sqrt{u^{\top} (\Sigma_h)^{-1} u}. \tag{8}$$

Consider the regression problem that is solved in moving backward from time step $h+1$ to h. Given the weight vector w_{h+1} at time step $h+1$, the regularized least-squares estimate of w_h is given by

$$\widehat{w}_h \overset{def}{=} \Sigma_h^{-1} \sum_{k=1}^{N} \phi_{hk} \Big[r_{hk} + \sum_{a \in \mathcal{A}} \pi_{h+1}(a \mid s_{h+1,k}) \langle \phi_{h+1}(s_{h+1,k}, a),\, w_{h+1} \rangle \Big].$$

We introduce pessimism by directly perturbing the weight vectors themselves—that is, we search for weight vector w_h such that $w_h = \xi_h + \widehat{w}_h$, where the pessimism vector $\xi_h \in \mathbb{R}^d$ satisfies a bound of the form $\|\xi_h\|_{\Sigma_h} \le \alpha_h$, for a user-defined parameter α_h.

In detail, the critic takes as input the dataset $\mathcal{D}$, a policy π, a sequence of tolerance parameters $\alpha = (\alpha_1, \dots, \alpha_H)$, weight radii $\rho^w = (\rho^w_1, \dots, \rho^w_H)$ with each $\rho^w_h \in (0, 1]$, and a regularization parameter $\lambda > 0$. The optimization variables consist of the regression vectors $w = (w_1, \dots, w_H) \in (\mathbb{R}^d)^H$ and the pessimism vectors $\xi = (\xi_1, \dots, \xi_H) \in (\mathbb{R}^d)^H$. The critic then solves the convex program

$$(\xi^\pi, \underline{w}^\pi) \overset{def}{=} \arg \min_{\substack{\xi \in (\mathbb{R}^d)^H \\ w \in (\mathbb{R}^d)^H}} \sum_{a \in \mathcal{A}} \pi_1(a \mid s_1) \langle \phi_1(s_1, a), \, w_1 \rangle \tag{9a}$$

with the terminal condition $w_{H+1} = 0$, and subject to the constraints

$$w_h = \xi_h + \Sigma_h^{-1} \sum_{k=1}^N \phi_{hk} \Big[r_{hk} + \sum_{a \in \mathcal{A}} \pi_{h+1}(a \mid s_{h+1,k}) \langle \phi_{h+1}(s_{h+1,k}, a), \, w_{h+1} \rangle \Big], \qquad \text{and} \tag{9b}$$

$$\|\xi_h\|^2_{\Sigma_h} \leq \alpha_h, \qquad \|w_h\|^2_2 \leq \rho^w_h \tag{9c}$$

for all $h \in [H]$. Here the matrix $\Sigma_h \equiv \Sigma_h(\lambda)$ was previously defined (7).

The convex program (9) consists of a linear objective subject to quadratic constraints; it is a special case of a second order cone program, and can be efficiently solved with standard convex solvers.

The Actor: Mirror Descent We now turn to the behavior of the actor. It applies the mirror descent algorithm based on the Kullback Leibler (KL) divergence [Bubeck, 2014]; this combination leads to the exponentiated gradient update rule in every timestep $h \in [H]$

$$\pi_{t+1,h}(a \mid s) \propto \pi_{t,h}(a \mid s) e^{\eta Q_h(s,a)} \qquad \text{for each } (s, a) \in \mathcal{S} \times \mathcal{A}, \tag{10}$$

where $\eta > 0$ is a stepsize parameter. If the Q-value above from the critic lives in $\mathcal{Q}$, then it is possible to show that $\pi_{t+1,h} \in \Pi_h$ and the update rule takes a much simpler and computationally more efficient form (cf. Line 5 of Algorithm 1), where $\underline{w}_t$ is the gradient of the value function on the pessimistic MDP implicitly identified by the critic. In this case, the spaces $(\mathcal{Q}, \Pi)$ are said to be *compatible* [Sutton et al., 1999, Kakade, 2001, Agarwal et al., 2020b, Raskutti and Mukherjee, 2015] and the resulting algorithm is often called the *Natural Policy Gradient* (NPG) (see also [Geist et al., 2019, Shani et al., 2020]). By construction, the critic maintains a linear action value function even after pessimistic perturbations. As a consequence, the actor policy space is the simple softmax policy class Π and the easier update rule can be used. As we explain in the analysis, this has important statistical benefits.

After T rounds of updates, the mirror descent algorithm that we use here readily achieves online regret rates (in the optimization setting with exact feedback) $\sim 1/T$ or $\sim 1/\sqrt{T}$ depending on the analysis [Agarwal et al., 2020b] and the learning rate, although we mention that these rates could potentially be improved [Khodadadian et al., 2021, Lan, 2021, Bhandari and Russo, 2020].

4 Main results

We now turn to the statement of a bound on the performance of the policy π_{ALG} returned by PACLE. This upper bound involves three terms: an optimization error, an uncertainty term, and a model mis-specification term. The *optimization error* is given by $\mathcal{C}(T) \overset{def}{=} 4H\sqrt{\frac{\log|\mathcal{A}|}{T}}$; it captures the rate at which the error decreases as a function of the iterations of the actor. The *mis-specification error* $\mathcal{E}_{\text{msp}}(\nu) \overset{def}{=} \sum_{h=1}^H \nu_h$ is simply the sum of all the stage-wise mis-specification errors; notice that the mis-specification error does depend on the choice of the radii for the critic $\rho^w_1, \dots, \rho^w_H$ in a problem dependent way (cf. Assumption 3). Finally, for each h, define the vector $\bar{\phi}^\pi_h \overset{def}{=} \mathbb{E}_{(S_h, A_h) \sim \pi}[\phi_h(S_h, A_h)]$, where the expectation is over the state-action (S_h, A_h) encountered at timestep h upon following policy π. In terms of these vectors, the *uncertainty error* is given by

$$\mathcal{U}(\pi, \alpha, \lambda) \overset{def}{=} 2 \sum_{h=1}^H \sqrt{\alpha_h} \|\bar{\phi}^\pi_h\|_{\Sigma_h^{-1}} = 2 \sum_{h=1}^H \sqrt{\alpha_h (\bar{\phi}^\pi_h)^\top \Sigma_h^{-1} \bar{\phi}^\pi_h}, \tag{11}$$

where the cumulative covariance matrix $\Sigma_h \equiv \Sigma_h(\lambda)$ was defined in equation (7).

The amount of information from the dataset $\mathcal{D}$ is fully encoded in the uncertainty function $\mathcal{U}$ through the sequence of cumulative covariance matrices $\{\Sigma_h\}_{h=1}^H$ and parameters $\{\alpha_h\}_{h=1}^H$. Both Σ_h and α_h depend on λ in opposite way; to simplify the presentation, we consider the setting $\lambda = 1$. The more data are available, the more positive definite Σ_h is and the smaller the uncertainty function $\mathcal{U}(\pi)$ becomes for a fixed policy π. If the sampling distribution that generates the dataset is fixed, then we can write $\mathcal{U}(\pi) \lessapprox C/\sqrt{n}$ where C does not depend on n and can be interpreted as the coverage of the sampling distribution with respect to policy π.

4.1 A guarantee for Pacle

Our main result holds under Assumption 1 (*Data Generation*), when the actor uses the learning rate $\eta = \sqrt{\log|\mathcal{A}|/T}$, the radii $\rho_1^w, \dots, \rho_H^w$ for the action value function[1] parameters are in $(0, 1]$, the regularization is $\lambda = 1$ and the number of iterations is $T \geq \log|\mathcal{A}|$; Π_{all} is the class of all stochastic policies. We let $\sqrt{\alpha_h} = \widetilde{O}(\sqrt{d_h + d_{h+1}}) + \nu_h\sqrt{N} + \sqrt{\lambda}$, where d_h is the dimensionality of the feature map at timestep h; we also highlight that we obtain a family of results, function of the critic's radii $\rho_1^w, \dots, \rho_H^w$. The choice of the radii is a modeling choice: increasing the radii increases both the approximation power of the function class $\mathcal{Q}_h$ used for regression, but also increases the complexity of the function class $\mathcal{Q}_{h+1}$ to represent (cf Assumption 3); thus, the choice of the radii affects the approximation error $\mathcal{E}_{\text{msp}}(\nu)$ in a problem dependent way.

Theorem 1 (An achievable guarantee)**.** *Under assumption Assumption 1 (*Data Generation*), and given parameters $(T, \lambda, \eta, \{\alpha_h, \rho_h^w\}_{h=1}^H)$ as described above,* Pacle *returns a policy π_{ALG} such that we have*

$$V_1^\pi(s_1) - V_1^{\pi_{\text{ALG}}}(s_1) \leq \mathcal{U}(\pi, \alpha) + \mathcal{E}_{msp}(\nu) + \mathcal{C}(T) \qquad \text{uniformly over all policies } \pi \in \Pi_{all} \tag{12}$$

with probability exceeding $1 - \delta$.

The result provides a family of upper bounds on the sub-optimality of the learned policy π_{ALG}, indexed by the choice of comparator policy π, and embodies a tradeoff between the sub-optimality of the comparator π and its uncertainty $\mathcal{U}(\pi)$. As a special case, if we set $\pi = \pi^\star$, then we obtain that the learned policy satisfies a bound of the form

$$V_1^{\pi_{\text{ALG}}}(s_1) \geq V_1^{\pi^\star}(s_1) - \mathcal{U}(\pi^\star) - \mathcal{C}(T) \tag{13}$$

with probability at least $1 - \delta$. Note that the optimization error $\mathcal{C}(T)$ can be reduced arbitrarily, while α (and thus $\mathcal{U}$) increase only logarithmically with T. Consequently, the guarantees (13) and (12) are satisfying ones whenever the remaining uncertainty term $\mathcal{U}(\pi^\star)$ is small.

Ignoring the optimization error, regularization and misspecification and assuming $d_h = d, \forall h \in [H]$ we obtain with high probability that the sub-optimality gap satisfies

$$V_1^\pi(s_1) - V_1^{\pi_{\text{ALG}}}(s_1) \preceq \sqrt{d}\sum_{h=1}^H \|\phi_h^\pi\|_{\Sigma_h^{-1}},$$

uniformly over all choices of policies $\pi \in \Pi_{\text{all}}$. Such a guarantee is significantly stronger as Pacle competes with all comparator policies simultaneously; these policies are not necessarily in the prescribed policy class Π. To highlight the strength of our formulation (see also [Yu et al., 2020, Liu et al., 2020] for results in a similar form), suppose that the optimal policy is not well covered, i.e., $\mathcal{U}(\pi^\star)$ infinite, but there exists a near-optimal policy π^+ i.e., such that $V_1^{\pi^+}(s_1) \geq V_1^\star(s_1) - \epsilon$ for some small ϵ, which is well covered by the dataset, i.e., $\mathcal{U}(\pi^+) \approx 0$. In this case, Theorem 1 ensures with high probability $V_1^{\text{ALG}}(s_1) \gtrapprox V_1^\star(s_1) - \epsilon$. In contrast, traditional analyses that use only $\pi^\star$ as comparator cannot return meaningful guarantees.

4.2 A lower bound

The result is complemented by a matching worst-case lower bound on the quality of the returned policy, excluding constants and log factors. The lower bound already arises in a setting that is easier

[1] This represents a setting where both the reward and the value function can be as large as 1 in absolute value. One easily recovers the setting with value functions in $[0, H]$ using a rescaling argument.

for the learner, as it holds (1) when the MDP is *low-rank* (thus it applies when Assumption 3 (*Restricted Closedness*) holds), and (2) when the mechanism that generates the dataset is *non-adaptive* (thus it applies when Assumption 1 (*Data Generation*) holds).

We assume $d_h = d, \forall h \in [H]$ and $\nu_h = 0$ for simplicity, as well as $\lambda = 1$ when referring to the uncertainty function $\mathcal{U}$; $\mathbb{E}_M$ indicates that the expectation is with respect to MDP M.

Theorem 2 (Information-theoretic limit)**.** *Fix any choice of horizon H and of dimension d and choose a large enough number of samples N to collect at each timestep. There exists an MDP class $\mathcal{M}$ function of d, H, N and a universal constant c such that for any estimator $\widehat{\pi}_{\text{ALG}}$, we have*

$$\sup_{M \in \mathcal{M}} \left\{ V^{\pi}_{1M}(s_1) - \mathbb{E}_M[V^{\widehat{\pi}_{\text{ALG}}}_{1M}(s_1)] \right\} \geq \frac{c}{\log\left(\frac{N}{\delta}\right)} \mathcal{U}(\pi) \qquad \textit{uniformly over all } \pi \in \Pi_{all}. \tag{14}$$

4.3 Comparison to related work

Theorem 1 automatically implies the typical bound $\mathbb{P}[V^{\pi_{\text{ALG}}}_1(s_1) \geq V^\star_1(s_1) - \mathcal{U}(\pi^\star)] \geq 1 - \delta$ when the comparator policy is the optimal policy $\pi^\star$, e.g., [Jin et al., 2020b, Rashidinejad et al., 2021, Kidambi et al., 2020, Kumar et al., 2019, Buckman et al., 2020]. The guarantee can be written as $V_1^{\pi}{}^{\text{ALG}}(s_1) \gtrapprox V^\star_1(s_1) - C/\sqrt{n}$ where n is the number of samples and C is the (scaled) condition number of Σ_h^{-1}. One could interpret C as a concentrability coefficient that expresses the coverage of dataset — through Σ_h — with respect to the average direction in feature space $\mathbb{E}_{(s_h,a_h)\sim\pi^\star_h}\phi(s_h, a_h)$ of the optimal policy $\pi^\star$. As in the paper [Jin et al., 2020b], such factor C can be small even when traditional concentrability coefficients are large because they depend on state-action visit ratios (see the literature in Appendix A, e.g., [Chen and Jiang, 2019]).

Even ignoring the concentrability coefficient, the form of our result is significantly stronger as our algorithm competes with all comparator policies simultaneously; these policies are not necessarily in the prescribed policy class Π. To highlight the strength of our formulation (see also [Yu et al., 2020, Liu et al., 2020] for results in a similar form), suppose that the optimal policy is not well covered, i.e., $\mathcal{U}(\pi^\star)$ infinite, but there exists a near-optimal policy π^+ i.e., such that $V^{\pi^+}_1(s_1) \geq V^\star_1(s_1) - \epsilon$ for some small ϵ, which is well covered by the dataset, i.e., $\mathcal{U}(\pi^+) \approx 0$. In this case, Theorem 1 ensures with high probability $V^{\text{ALG}}_1(s_1) \gtrapprox V^\star_1(s_1) - \epsilon$. In contrast, traditional analyses that use only $\pi^\star$ as comparator cannot return meaningful guarantees.

The work closest to ours is [Jin et al., 2020b]; our work directly improves on theirs by closing the dH gap between their upper and lower bound while working under the more permissive Assumption 3 (*Restricted Closedness*) which includes low-rank MDPs. A $\sqrt{d}$-improvement is due to the algorithm we use and the remaining is due to a more refined analysis and construction to certify optimality in Theorem 2 (notice that our upper and lower bounds differ from theirs by a factor of H due to a different normalization in the value function). The result of [Liu et al., 2020] can be specialized to the low-rank MDP setting but would give a suboptimal bound while additionally requiring density estimates.

Deriving a computationally tractable model-free algorithm without low-rank dynamics but subject to value function perturbations (e.g., optimistic or pessimistic perturbations) is an open problem even in the more heavily studied exploration setting: there the current state of the art [Zanette et al., 2020b, Jin et al., 2021, Du et al., 2021, Jiang et al., 2017] only present computationally *intractable* algorithms with the exception of [Zanette et al., 2020c] for a PAC setting with low inherent Bellman error which however requires an additional "explorability" condition.

5 Proof Sketch

Our analysis has three main ingredients: the online learning guarantees of the actor, the pessimistic estimates from the critic, and the concept of induced MDP that connects the actor to the critic. The proof sketch follows a bottom-up approach: (1) starting with the critic, we explain the benefits of working within the prescribed value function space, (2) we introduce the concept of MDPs induced by the actor's policy to interpret the pessimistic value function given by the critic as an exact value function on an adversarially chosen MDP and (3) we conclude by giving online-style learning guarantees for the actor.

The proof sketch is specialized to the well specified case ($\nu = 0$), it ignores the bias due to regularization ($\lambda = 0$) and it assumes we are working in a high probability 'good event' to simplify the analysis.

Additional notation We denote with $\pi_t = \pi_{\theta_t}$ the policy of the actor at iteration t; we use M_t to denote the corresponding induced MDP (the definition will be given later). When the critic is invoked with any actor policy π, it solves the convex program (9), thereby obtaining the solution $(\underline{\xi}^\pi, \underline{w}^\pi)$ that determines the (action) value functions

$$(s,a) \mapsto \underline{Q}^\pi_h(s,a) \overset{def}{=} \langle \phi(s,a), \underline{w}^\pi_h \rangle, \qquad \text{and} \quad s \mapsto \underline{V}^\pi_h(s) \overset{def}{=} \mathbb{E}_{A' \sim \pi_h(\cdot|s)} \underline{Q}^\pi_h(s, A'). \tag{15}$$

Using the policy class Π_h and action value function $\mathcal{Q}_h$ from Definition 1 (*Functional Spaces*), we define the induced value function class

$$\mathcal{V}_h \overset{def}{=} \big\{ s \mapsto \mathbb{E}_{A' \sim \pi_h(\cdot|s)} Q(s, A') \mid Q \in \mathcal{Q}, \pi \in \Pi \big\}. \tag{16}$$

5.1 Critic's Pessimistic Guarantees

We start by analyzing the value function error. For each $h \in [H]$ and policy π, we define the *statistical error in parameter space* as

$$\varepsilon^\pi_h \overset{def}{=} \Sigma_h^{-1} \sum_{k=1}^{N} \phi_{hk} \Big[r_{hk} + \underline{V}^\pi_{h+1}(s_{h+1,k}) - (\mathcal{T}^\pi_h \underline{Q}^\pi_{h+1})(s_{hk}, a_{hk}) \Big], \tag{17a}$$

where the reader should recall the definitions (15) of $\underline{Q}$ and $\underline{V}$. This statistical error interacts with the pessimism vector $\underline{\xi}^\pi_h$ to determine the *aggregated perturbation function* b^π_h given by

$$b^\pi_h(s,a) \overset{def}{=} \Big\langle \phi_h(s,a), \, \underline{\xi}^\pi_h + \varepsilon^\pi_h \Big\rangle \qquad \text{for all } (s,a) \in \mathcal{S} \times \mathcal{A}. \tag{17b}$$

Our first lemma bounds the error in the estimate $\underline{V}^\pi_1(s_1)$ of the value function at the initial state:

Lemma 1 (PLSPE Errors). *For any input policy $\pi \in \Pi$, the output of the critic (Algorithm 2) ensures that*

$$\underbrace{(\underline{V}^\pi_1(s_1) - V^\pi_1(s_1))}_{\textit{Error at the initial state}} = \sum_{h=1}^{H} \mathbb{E}_{(S_h, A_h) \sim \pi} \big[b^\pi_h(S_h, A_h) \big]. \tag{18}$$

The above statement is stated as Eq. (41) in appendix. The proof of this lemma exploits the least-squares constraint (9b) along with the Bellman closure condition Assumption 3 (*Restricted Closedness*). We highlight that such assumption already arises in the analysis of standard LSPE when $\underline{\xi}^\pi_h = 0$. Allowing $\underline{\xi}^\pi_h \neq 0$ is equivalent to experiencing a different noise perturbation, and therefore no further model assumption is needed. This is a benefit of working in parameter space in terms of modeling assumptions.

Ideally we would set $\underline{\xi}^\pi_h = -\varepsilon^\pi_h$ in the program (9) leading to no value function error $\underline{V}^\pi_1 = V^\pi_1$. Since ε^π_h is unknown, the program finds a pessimistic value by allowing $\|\underline{\xi}^\pi_h\|_{\Sigma_h}$ to be of similar size as the noise error $\|\varepsilon^\pi_h\|_{\Sigma_h}$, which is bounded below.

Let $\log \mathcal{N}(\mathcal{V}_{\text{ALG},h+1})$ be the log covering number of the value function class for $\underline{V}^\pi_{h+1}$ in ∞ norm and for some appropriate discretization step (see Lemma 9 in appendix for additional details)

Lemma 2 (Statistical Error). *Under Assumption 1, we have*

$$\|\varepsilon^\pi_h\|_{\Sigma_h} \leq \sqrt{\alpha_h} \overset{def}{=} \widetilde{O}\left(\sqrt{d_h + \log \mathcal{N}(\mathcal{V}_{\text{ALG},h+1}) + \log \frac{1}{\delta}} \right) \qquad \textit{for all } \pi_h \in \Pi_h \textit{ and } h \in [H]$$

with probability at least $1 - \delta$.

A computation of the covering number of the next-state value function class gives $\log \mathcal{N}(\mathcal{V}_{\text{ALG},h+1}) \approx d_{h+1}$; plugging this into the above expression gives the value for α that we choose in Theorem 1. The proof of the above lemma can be found in Lemma 8 in appendix.

As $\xi^\pi_h = -\varepsilon^\pi_h$ is now feasible for the program (9), this choice produces no value function error, it must follow that $\underline{V}^\pi_1(s_1) \leq V^\pi_1(s_1)$. Thus, we have informally established the following gurantee:

Proposition 2 (Critic's Pessimistic Guarantees)**.** *With high probability, we have the upper bound* $\underline{V}_1^{\pi}(s_1) \le V_1^{\pi}(s_1)$.

It remains to bound $\sqrt{\alpha_h}$ by computing the log covering number for the action value function class $\mathcal{V}_{\text{ALG},h+1}$ used by the algorithm. Thankfully, our choice of making perturbations within the function class ensures that $\mathcal{V}_{\text{ALG},h+1} = \mathcal{V}_{h+1}$ (defined in Eq. (16)); this function class is relatively simple as reflected by its small covering number. This is where we strongly benefit from working in parameter space in terms of statistical efficiency. In order to cover the value function (15), it suffices to cover the action value function class $\mathcal{Q}$ and the policy class Π.

First, for a fixed policy π, since the agent's action value function is enforced to be linear $\underline{Q}^{\pi} \in \mathcal{Q}$ even after perturbations, the union bound only needs to be done over the linear class $\mathcal{Q}$; we thus avoid a potentially more costly union bound over a much larger function class $\mathcal{Q}' =$ [linear functions] + [complex bonus], e.g. [Jin et al., 2020b].

Second, the union bound needs to be extended to all policies that the actor can use to invoke the critic. Since the action value function $\underline{Q}$ that the critic identifies is linear, the update takes a simple closed-form expression, because the softmax policy class Π and the linear action value function class $\mathcal{Q}$ are compatible [Kakade, 2001, Agarwal et al., 2020b]. Precisely, this gives the simple update rule in Line 1 of Algorithm 1. If the action value function $\underline{Q}$ was perturbed by bonuses, linearity of the critic's value function would be lost and its functional space $\neq \mathcal{Q}$ would not be compatible with the policy space Π. The more general form of the exponential gradient update in Eq. (10) would need to be used. However, the resulting policy would no longer live in Π and instead would generate a more complex (i.e., with a larger covering number) policy class for the actor. The space complexity would increase too: our update (cf. Line 1 in Algorithm 1) requires a simple vector addition while the exponential update rule in Eq. (10) should be performed in all (s, a) where the policy is needed.

A simple computation now gives $\log \mathcal{N}(\mathcal{V}_{h+1}) \asymp d$ and thus $\sqrt{\alpha} = \widetilde{O}(\sqrt{d})$: this means that our union bounds are small enough that the resulting confidence intervals for the error ε_h^{π} are about the same size as those arising from linear bandit regression where no union bound is needed. Ultimately, this is where we can save a $\sqrt{d}$ factor compared to [Jin et al., 2020b].

5.2 Induced MDP

At each round t, the critic is given the actor's policy π_t, and is designed to extract pessimistically biased $\underline{Q}^{\pi_t}$ values. As we show here, such Q-values can be interpreted as being the exact Q-values for a perturbed MDP M_t. This connection is useful in relating the bias from the critic to the online learning guarantees from the actor.

More precisely, recall the aggregated perturbation function $b_h^{\pi_t}(s,a) = \left\langle \phi_h(s,a), \underline{\xi}_h^{\pi_t} + \varepsilon_h^{\pi_t} \right\rangle$, as previously defined in equation (17b). The perturbed MDP M_t is equivalent to the original M, *except* that its reward function $\widetilde{r}^{\pi_t}$ is given by

$$\widetilde{r}^{\pi_t}(s,a) = r(s,a) + b_h^{\pi_t}(s,a) \qquad \text{for each } h \in [H]. \tag{19}$$

Note that this perturbation, in addition to the statistical error $\varepsilon_h^{\pi_t}$, also includes an *adversarial component*, since the vectors $\underline{\xi}_h^{\pi_t}$ were chosen by the critic to minimize the value of the actor's policies. The perturbed reward functions (19) are useful in our analysis because they allow us to evaluate arbitrary policies in the critic's pessimistic world.

Proposition 3 (Value of Policies on Induced MDP)**.** *Given the actor's policy* π_t *at round* t*, we have*

$$Q_{M_t}^{\pi_t} = \underline{Q}^{\pi_t} \quad \textit{and} \quad V_{M_t}^{\pi_t} = \underline{V}^{\pi_t}. \tag{20}$$

I.e., the critic's pessimistic $\underline{Q}^{\pi_t}$ *function equals the action value function* $Q_{M_t}^{\pi_t}$ *in the induced MDP* M_t*. Furthermore, we have the following guarantees on the values of policies on* M_t

$$V_{1,M_t}^{\pi_t}(s_1) \le V_1^{\pi_t}(s_1) \qquad \textit{for the actor's policy } \pi_t\textit{, and} \tag{21a}$$

$$V_{1,M_t}^{\pi}(s_1) \le V_1^{\pi}(s_1) + \mathcal{U}(\pi) \qquad \textit{for any } \pi. \tag{21b}$$

The statements above follow directly from the definition of induced MDP through the additional reward function b_h; please see Eqs. (43a) and (43b) in appendix for additional details. The proposition states three important facts: 1) the policy π_t that induces the MDP M_t is 'special', as its Q function on M_t equals the critic's $\underline{Q}^{\pi_t}$ function (Eq. (20)); 2) as a result, the value of π_t on M_t is pessimistic compared to its value on the original MDP (first equation in Eq. (21)); 3) the value of any other policy on M_t and M differs by at most the uncertainty function $\mathcal{U}(\cdot)$ (second equation in Eq. (21)).

We highlight that the induced MDP M_t is unknown to the learner, because constructing M_t requires knowledge of the original reward and transition function in addition to the statistical error $\varepsilon_h^{\pi_t}$; the induced MDP is used only in the analysis.

5.3 Actor's Analysis

In this section, we present the actor's guarantees on the sequence of adversarial MDPs M_t identified by the critic. In order to do so, we modify an analysis of the natural policy gradient algorithm [Agarwal et al., 2020b] to derive online learning-style guarantees. The proposition below holds under additional preconditions that are satisfied during the execution of the actor; see Proposition 6 in the appendix for details.

We begin by observing that our development thus far ensures that at each round t, the actor receives a sequence of weight vectors $\underline{w}_t = (\underline{w}_{1t}, \ldots, \underline{w}_{Ht})$ corresponding to the action value function of π_t on M_t—that is, we have the equality

$$Q^{\pi_t}_{h,M_t}(s,a) = \langle \phi_h(s,a), \, \underline{w}_{ht} \rangle \,, \qquad \text{for all triples } (s,a,h). \tag{22}$$

This is indeed the case: the critic returns the parameter $\underline{w}_t$ such that

$$Q^{\pi_t}_{h,M_t}(s,a) \overset{(i)}{=} \underline{Q}^{\pi_t}_h(s,a) \overset{(ii)}{=} \langle \phi_h(s,a), \, \underline{w}_{hk} \rangle \,, \tag{23}$$

where equality (ii) follows from the definition (15) of $\underline{Q}^{\pi_t}_h$; and equation (i) is a consequence of the $\underline{Q}$-value preserving property (20) of the induced MDP (cf. Proposition 3). Given this property, we have the following guarantee for the actor.

Proposition 4 (Actor's Analysis)**.** *The sequence of actor policies $\{\pi_t\}_{t=1}^T$ satisfies the bound*

$$\frac{1}{T}\sum_{t=1}^{T} \left\{ V^{\pi}_{1,M_t}(s_1) - V^{\pi_t}_{1,M_t}(s_1) \right\} \leq \mathcal{C}(T) \lesssim H\sqrt{\frac{\log|\mathcal{A}|}{T}}, \tag{24}$$

valid for any *fixed comparator policy π.*

To be clear, the fixed comparator policy need not be in Π. This fact is important, as it allows us to derive bounds relative to an arbitrary comparator.

5.4 Combining the ingredients

We now have all the ingredients to prove the upper bound on $V^{\pi}_1(s_1) - V^{\pi_{\text{ALG}}}_1(s_1)$ stated in Theorem 1. The following reasoning holds under a high probability event (event $\mathcal{G}$ in appendix, which arises from Lemma 2) and for any comparator π, not necessarily in Π.

By construction, the actor's final policy π_{ALG} is a weighted mixture of the collection $\{\pi_t\}_{t=1}^T$, so that by definition, we have

$$V^{\pi}_1(s_1) - V^{\pi_{\text{ALG}}}_1(s_1) = \frac{1}{T}\sum_{t=1}^{T}\left\{ V^{\pi}_1(s_1) - V^{\pi_t}_1(s_1) \right\}.$$

From the bound (68b) in Proposition 4, the actor provides control on the sub-optimality gaps of the policies $\{\pi_t\}_{t=1}^T$ relative to the value functions of the perturbed MDPs $\{M_t\}_{t=1}^T$. But from the bounds Eq. (21) in Proposition 3, these original value functions can be bounded by these perturbed value functions, plus the uncertainty term, which yields the bound

$$\frac{1}{T}\sum_{t=1}^{T}\left\{ V^{\pi}_1(s_1) - V^{\pi_t}_1(s_1) \right\} \leq \frac{1}{T}\sum_{t=1}^{T}\left\{ V^{\pi}_{1,M_t}(s_1) - V^{\pi_t}_{1,M_t}(s_1) \right\} + \mathcal{U}(\pi).$$

Finally, applying the on-line regret bound (68b) and putting together the pieces, we obtain

$$V^{\pi}_1(s_1) - V^{\pi_{\text{ALG}}}_1(s_1) \leq \mathcal{U}(\pi) + \mathcal{C}(T),$$

as claimed.

6 Discussion

A key idea of this paper is to introduce pessimism while remaining in the prescribed function class. Doing so allows us to avoid making additional model assumptions, and achieves minimax optimality. Similar ideas have appeared before in the exploration setting (e.g., [Zanette et al., 2020b, Jin et al., 2021, Du et al., 2021]) with similar advantages (batch-style assumptions + minimax regret) *but at the expense of computational tractability*.

Fortunately, the offline RL setting differs from the online setting and we are able to maintain computational tractability by clearly separating the actor's update from the critic evaluation. In this way, *each algorithm solves a simpler task*, and computational tractability is retained.

The numerical evaluation of PACLE and the extension to more general function classes are important next steps, and it will be interesting to see if any of these ideas can be translated to the more challenging exploration setting.

References

[Abbasi-Yadkori et al., 2011] Abbasi-Yadkori, Y., Pal, D., and Szepesvari, C. (2011). Improved algorithms for linear stochastic bandits. In *Advances in Neural Information Processing Systems (NIPS)*.

[Agarwal et al., 2020a] Agarwal, A., Kakade, S., Krishnamurthy, A., and Sun, W. (2020a). Flambe: Structural complexity and representation learning of low rank mdps. *arXiv preprint arXiv:2006.10814*.

[Agarwal et al., 2020b] Agarwal, A., Kakade, S. M., Lee, J. D., and Mahajan, G. (2020b). Optimality and approximation with policy gradient methods in markov decision processes. In *Conference on Learning Theory*, pages 64–66.

[Agarwal et al., 2020c] Agarwal, R., Schuurmans, D., and Norouzi, M. (2020c). An optimistic perspective on offline reinforcement learning. In *International Conference on Machine Learning*, pages 104–114. PMLR.

[Antos et al., 2007] Antos, A., Munos, R., and Szepesvári, C. (2007). Fitted q-iteration in continuous action-space mdps.

[Antos et al., 2008] Antos, A., Szepesvári, C., and Munos, R. (2008). Learning near-optimal policies with bellman-residual minimization based fitted policy iteration and a single sample path. *Machine Learning*, 71(1):89–129.

[Bhandari and Russo, 2020] Bhandari, J. and Russo, D. (2020). A note on the linear convergence of policy gradient methods. *arXiv preprint arXiv:2007.11120*.

[Bubeck, 2014] Bubeck, S. (2014). Convex optimization: Algorithms and complexity. *arXiv preprint arXiv:1405.4980*.

[Buckman et al., 2020] Buckman, J., Gelada, C., and Bellemare, M. G. (2020). The importance of pessimism in fixed-dataset policy optimization. *arXiv preprint arXiv:2009.06799*.

[Chen and Jiang, 2019] Chen, J. and Jiang, N. (2019). Information-theoretic considerations in batch reinforcement learning. In *International Conference on Machine Learning*, pages 1042–1051.

[de la Pena et al., 2009] de la Pena, V. H., Lai, T. L., and Shao, Q. M. (2009). *Self-normalized processes*. Springer.

[Du et al., 2021] Du, S. S., Kakade, S. M., Lee, J. D., Lovett, S., Mahajan, G., Sun, W., and Wang, R. (2021). Bilinear classes: A structural framework for provable generalization in rl. *arXiv preprint arXiv:2103.10897*.

[Duan et al., 2021] Duan, Y., Jin, C., and Li, Z. (2021). Risk bounds and rademacher complexity in batch reinforcement learning. *arXiv preprint arXiv:2103.13883*.

[Duan and Wang, 2020] Duan, Y. and Wang, M. (2020). Minimax-optimal off-policy evaluation with linear function approximation. *arXiv preprint arXiv:2002.09516*.

[Fan et al., 2020] Fan, J., Wang, Z., Xie, Y., and Yang, Z. (2020). A theoretical analysis of deep q-learning. In *Learning for Dynamics and Control*, pages 486–489. PMLR.

[Farahmand et al., 2016] Farahmand, A.-m., Ghavamzadeh, M., Szepesvári, C., and Mannor, S. (2016). Regularized policy iteration with nonparametric function spaces. *The Journal of Machine Learning Research*, 17(1):4809–4874.

[Farahmand et al., 2010] Farahmand, A.-m., Szepesvári, C., and Munos, R. (2010). Error propagation for approximate policy and value iteration. In *Advances in Neural Information Processing Systems (NIPS)*.

[Farajtabar et al., 2018] Farajtabar, M., Chow, Y., and Ghavamzadeh, M. (2018). More robust doubly robust off-policy evaluation. In *International Conference on Machine Learning*, pages 1447–1456. PMLR.

[Fu et al., 2020] Fu, Z., Yang, Z., and Wang, Z. (2020). Single-timescale actor-critic provably finds globally optimal policy. *arXiv preprint arXiv:2008.00483*.

[Geist et al., 2019] Geist, M., Scherrer, B., and Pietquin, O. (2019). A theory of regularized markov decision processes. In *International Conference on Machine Learning*, pages 2160–2169. PMLR.

[Hao et al., 2021] Hao, B., Ji, X., Duan, Y., Lu, H., Szepesvári, C., and Wang, M. (2021). Bootstrapping statistical inference for off-policy evaluation. *arXiv preprint arXiv:2102.03607*.

[Jaques et al., 2019] Jaques, N., Ghandeharioun, A., Shen, J. H., Ferguson, C., Lapedriza, A., Jones, N., Gu, S., and Picard, R. (2019). Way off-policy batch deep reinforcement learning of implicit human preferences in dialog. *arXiv preprint arXiv:1907.00456*.

[Jiang and Huang, 2020] Jiang, N. and Huang, J. (2020). Minimax value interval for off-policy evaluation and policy optimization. *arXiv preprint arXiv:2002.02081*.

[Jiang et al., 2017] Jiang, N., Krishnamurthy, A., Agarwal, A., Langford, J., and Schapire, R. E. (2017). Contextual decision processes with low Bellman rank are PAC-learnable. In Precup, D. and Teh, Y. W., editors, *International Conference on Machine Learning (ICML)*, volume 70 of *Proceedings of Machine Learning Research*, pages 1704–1713, International Convention Centre, Sydney, Australia. PMLR.

[Jiang and Li, 2016] Jiang, N. and Li, L. (2016). Doubly robust off-policy value evaluation for reinforcement learning. In *International Conference on Machine Learning*, pages 652–661. PMLR.

[Jin et al., 2021] Jin, C., Liu, Q., and Miryoosefi, S. (2021). Bellman eluder dimension: New rich classes of rl problems, and sample-efficient algorithms. *arXiv preprint arXiv:2102.00815*.

[Jin et al., 2020a] Jin, C., Yang, Z., Wang, Z., and Jordan, M. I. (2020a). Provably efficient reinforcement learning with linear function approximation. In *Conference on Learning Theory*.

[Jin et al., 2020b] Jin, Y., Yang, Z., and Wang, Z. (2020b). Is pessimism provably efficient for offline rl? *arXiv preprint arXiv:2012.15085*.

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

[Kakade et al., 2003] Kakade, S. M. et al. (2003). *On the sample complexity of reinforcement learning*. PhD thesis, University of London London, England.

[Kallus and Uehara, 2019] Kallus, N. and Uehara, M. (2019). Efficiently breaking the curse of horizon in off-policy evaluation with double reinforcement learning. *arXiv preprint arXiv:1909.05850*.

[Khodadadian et al., 2021] Khodadadian, S., Jhunjhunwala, P. R., Varma, S. M., and Maguluri, S. T. (2021). On the linear convergence of natural policy gradient algorithm. *arXiv preprint arXiv:2105.01424*.

[Kidambi et al., 2020] Kidambi, R., Rajeswaran, A., Netrapalli, P., and Joachims, T. (2020). Morel: Model-based offline reinforcement learning. *arXiv preprint arXiv:2005.05951*.

[Kumar et al., 2019] Kumar, A., Fu, J., Tucker, G., and Levine, S. (2019). Stabilizing off-policy q-learning via bootstrapping error reduction. *arXiv preprint arXiv:1906.00949*.

[Kumar et al., 2020] Kumar, A., Zhou, A., Tucker, G., and Levine, S. (2020). Conservative q-learning for offline reinforcement learning. *arXiv preprint arXiv:2006.04779*.

[Lan, 2021] Lan, G. (2021). Policy mirror descent for reinforcement learning: Linear convergence, new sampling complexity, and generalized problem classes. *arXiv preprint arXiv:2102.00135*.

[Laroche et al., 2019] Laroche, R., Trichelair, P., and Des Combes, R. T. (2019). Safe policy improvement with baseline bootstrapping. In *International Conference on Machine Learning*, pages 3652–3661. PMLR.

[Levine et al., 2020] Levine, S., Kumar, A., Tucker, G., and Fu, J. (2020). Offline reinforcement learning: Tutorial, review, and perspectives on open problems. *arXiv preprint arXiv:2005.01643*.

[Liao et al., 2020] Liao, P., Qi, Z., and Murphy, S. (2020). Batch policy learning in average reward markov decision processes. *arXiv preprint arXiv:2007.11771*.

[Liu et al., 2018] Liu, Q., Li, L., Tang, Z., and Zhou, D. (2018). Breaking the curse of horizon: Infinite-horizon off-policy estimation. In *Advances in Neural Information Processing Systems*, pages 5356–5366.

[Liu et al., 2020] Liu, Y., Swaminathan, A., Agarwal, A., and Brunskill, E. (2020). Provably good batch reinforcement learning without great exploration. *arXiv preprint arXiv:2007.08202*.

[Mannor et al., 2012] Mannor, S., Mebel, O., and Xu, H. (2012). Lightning does not strike twice: Robust mdps with coupled uncertainty. *arXiv preprint arXiv:1206.4643*.

[Modi et al., 2021] Modi, A., Chen, J., Krishnamurthy, A., Jiang, N., and Agarwal, A. (2021). Model-free representation learning and exploration in low-rank mdps. *arXiv preprint arXiv:2102.07035*.

[Munos, 2003] Munos, R. (2003). Error bounds for approximate policy iteration. In *ICML*, volume 3, pages 560–567.

[Munos, 2005] Munos, R. (2005). Error bounds for approximate value iteration. In *AAAI Conference on Artificial Intelligence (AAAI)*.

[Nachum et al., 2019a] Nachum, O., Chow, Y., Dai, B., and Li, L. (2019a). Dualdice: Behavior-agnostic estimation of discounted stationary distribution corrections. *arXiv preprint arXiv:1906.04733*.

[Nachum and Dai, 2020] Nachum, O. and Dai, B. (2020). Reinforcement learning via fenchel-rockafellar duality. *arXiv preprint arXiv:2001.01866*.

[Nachum et al., 2019b] Nachum, O., Dai, B., Kostrikov, I., Chow, Y., Li, L., and Schuurmans, D. (2019b). Algaedice: Policy gradient from arbitrary experience. *arXiv preprint arXiv:1912.02074*.

[Nair et al., 2020] Nair, A., Dalal, M., Gupta, A., and Levine, S. (2020). Accelerating online reinforcement learning with offline datasets. *arXiv preprint arXiv:2006.09359*.

[Nedić and Bertsekas, 2003] Nedić, A. and Bertsekas, D. P. (2003). Least squares policy evaluation algorithms with linear function approximation. *Discrete Event Dynamic Systems*, 13(1):79–110.

[Puterman, 1994] Puterman, M. L. (1994). *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons, Inc., New York, NY, USA.

[Rashidinejad et al., 2021] Rashidinejad, P., Zhu, B., Ma, C., Jiao, J., and Russell, S. (2021). Bridging offline reinforcement learning and imitation learning: A tale of pessimism. *arXiv preprint arXiv:2103.12021*.

[Raskutti and Mukherjee, 2015] Raskutti, G. and Mukherjee, S. (2015). The information geometry of mirror descent. *IEEE Transactions on Information Theory*, 61(3):1451–1457.

[Shani et al., 2020] Shani, L., Efroni, Y., and Mannor, S. (2020). Adaptive trust region policy optimization: Global convergence and faster rates for regularized mdps. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 5668–5675.

[Shreve and Bertsekas, 1978] Shreve, S. E. and Bertsekas, D. P. (1978). Alternative theoretical frameworks for finite horizon discrete-time stochastic optimal control. *SIAM Journal on control and optimization*, 16(6):953–978.

[Siegel et al., 2020] Siegel, N. Y., Springenberg, J. T., Berkenkamp, F., Abdolmaleki, A., Neunert, M., Lampe, T., Hafner, R., Heess, N., and Riedmiller, M. (2020). Keep doing what worked: Behavioral modelling priors for offline reinforcement learning. *arXiv preprint arXiv:2002.08396*.

[Sutton et al., 1999] Sutton, R. S., McAllester, D. A., Singh, S. P., Mansour, Y., et al. (1999). Policy gradient methods for reinforcement learning with function approximation. In *NIPs*, volume 99, pages 1057–1063. Citeseer.

[Tang et al., 2019] Tang, Z., Feng, Y., Li, L., Zhou, D., and Liu, Q. (2019). Doubly robust bias reduction in infinite horizon off-policy estimation. *arXiv preprint arXiv:1910.07186*.

[Thomas and Brunskill, 2016] Thomas, P. and Brunskill, E. (2016). Data-efficient off-policy policy evaluation for reinforcement learning. In *International Conference on Machine Learning*, pages 2139–2148.

[Tsybakov, 2009] Tsybakov, A. B. (2009). *Introduction to Nonparametric Estimation*. Springer, New York.

[Uehara et al., 2020] Uehara, M., Huang, J., and Jiang, N. (2020). Minimax weight and q-function learning for off-policy evaluation. In *International Conference on Machine Learning*, pages 9659–9668. PMLR.

[Voloshin et al., 2021] Voloshin, C., Jiang, N., and Yue, Y. (2021). Minimax model learning. In *International Conference on Artificial Intelligence and Statistics*, pages 1612–1620. PMLR.

[Wainwright, 2019] Wainwright, M. J. (2019). *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press.

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

[Wang et al., 2020] Wang, Z., Novikov, A., Żołna, K., Springenberg, J. T., Reed, S., Shahriari, B., Siegel, N., Merel, J., Gulcehre, C., Heess, N., et al. (2020). Critic regularized regression. *arXiv preprint arXiv:2006.15134*.

[Weisz et al., 2020] Weisz, G., Amortila, P., and Szepesvári, C. (2020). Exponential lower bounds for planning in mdps with linearly-realizable optimal action-value functions. *arXiv preprint arXiv:2010.01374*.

[Wu et al., 2019] Wu, Y., Tucker, G., and Nachum, O. (2019). Behavior regularized offline reinforcement learning. *arXiv preprint arXiv:1911.11361*.

[Wu et al., 2021] Wu, Y., Zhai, S., Srivastava, N., Susskind, J., Zhang, J., Salakhutdinov, R., and Goh, H. (2021). Uncertainty weighted actor-critic for offline reinforcement learning. *arXiv preprint arXiv:2105.08140*.

[Xie and Jiang, 2020a] Xie, T. and Jiang, N. (2020a). Batch value-function approximation with only realizability. *arXiv preprint arXiv:2008.04990*.

[Xie and Jiang, 2020b] Xie, T. and Jiang, N. (2020b). Q* approximation schemes for batch reinforcement learning: A theoretical comparison. volume 124 of *Proceedings of Machine Learning Research*, pages 550–559, Virtual. PMLR.

[Xie et al., 2019] Xie, T., Ma, Y., and Wang, Y.-X. (2019). Towards optimal off-policy evaluation for reinforcement learning with marginalized importance sampling. In *Advances in Neural Information Processing Systems*, pages 9668–9678.

[Yang and Wang, 2020] Yang, L. F. and Wang, M. (2020). Reinforcement leaning in feature space: Matrix bandit, kernels, and regret bound. In *International Conference on Machine Learning (ICML)*.

[Yang et al., 2020] Yang, M., Nachum, O., Dai, B., Li, L., and Schuurmans, D. (2020). Off-policy evaluation via the regularized lagrangian. *arXiv preprint arXiv:2007.03438*.

[Yin et al., 2020] Yin, M., Bai, Y., and Wang, Y.-X. (2020). Near optimal provable uniform convergence in off-policy evaluation for reinforcement learning. *arXiv preprint arXiv:2007.03760*.

[Yin and Wang, 2020] Yin, M. and Wang, Y.-X. (2020). Asymptotically efficient off-policy evaluation for tabular reinforcement learning. In *International Conference on Artificial Intelligence and Statistics*, pages 3948–3958. PMLR.

[Yu et al., 2020] Yu, T., Thomas, G., Yu, L., Ermon, S., Zou, J., Levine, S., Finn, C., and Ma, T. (2020). Mopo: Model-based offline policy optimization. *arXiv preprint arXiv:2005.13239*.

[Zanette, 2020] Zanette, A. (2020). Exponential lower bounds for batch reinforcement learning: Batch rl can be exponentially harder than online rl. *arXiv preprint arXiv:2012.08005*.

[Zanette et al., 2020a] Zanette, A., Brandfonbrener, D., Pirotta, M., and Lazaric, A. (2020a). Frequentist regret bounds for randomized least-squares value iteration. In *AISTATS*.

[Zanette et al., 2020b] Zanette, A., Lazaric, A., Kochenderfer, M., and Brunskill, E. (2020b). Learning near optimal policies with low inherent bellman error. In *International Conference on Machine Learning (ICML)*.

[Zanette et al., 2020c] Zanette, A., Lazaric, A., Kochenderfer, M. J., and Brunskill, E. (2020c). Provably efficient reward-agnostic navigation with linear value iteration. In *Advances in Neural Information Processing Systems*.

[Zhang et al., 2020a] Zhang, J., Koppel, A., Bedi, A. S., Szepesvari, C., and Wang, M. (2020a). Variational policy gradient method for reinforcement learning with general utilities. *arXiv preprint arXiv:2007.02151*.

[Zhang et al., 2020b] Zhang, R., Dai, B., Li, L., and Schuurmans, D. (2020b). Gendice: Generalized offline estimation of stationary values. *arXiv preprint arXiv:2002.09072*.

Contents

A Additional Literature

For empirical studies on offline RL, see the papers [Laroche et al., 2019, Jaques et al., 2019, Wu et al., 2019, Agarwal et al., 2020c, Wang et al., 2020, Siegel et al., 2020, Nair et al., 2020] in addition to those presented in the main text. Several works have investigated offline policy learning, where concentrability coefficients are introduced to account for the non-uniform error propagation [Munos, 2003, Munos, 2005, Antos et al., 2007, Antos et al., 2008, Farahmand et al., 2010, Farahmand et al., 2016, Chen and Jiang, 2019, Xie and Jiang, 2020a, Xie and Jiang, 2020b, Duan et al., 2021]. For additional literature, see also the papers [Zhang et al., 2020a, Liao et al., 2020, Fan et al., 2020, Fu et al., 2020, Wang et al., 2019]. Concentrability coefficients or density ratios also appears in the off-policy evaluation problem, which is distinct from the policy learning problem that we consider here [Zhang et al., 2020b, Thomas and Brunskill, 2016, Farajtabar et al., 2018, Liu et al., 2018, Xie et al., 2019, Yang et al., 2020, Nachum et al., 2019b, Yin et al., 2020, Yin and Wang, 2020, Duan and Wang, 2020, Uehara et al., 2020, Jiang and Huang, 2020, Kallus and Uehara, 2019, Tang et al., 2019, Nachum and Dai, 2020, Nachum et al., 2019a, Jiang and Li, 2016, Uehara et al., 2020, Voloshin et al., 2021, Jiang and Huang, 2020, Hao et al., 2021].

B Proof of Proposition 1

First, let us define an MDP class indexed by N; we will use this MDP class to show that each inclusion is strict. At a high level, this MDP class has a starting state 0 where the agent can choose to go left (action -1) or right (action $+1$); after that, it will keep going left or right until the leftmost or rightmost terminal state is reached. The reward is non-zero only at the terminal states.

For a fixed N, let the horizon be $H = N + 1$ and consider the following chain MDP, where the state space is

$$\mathcal{S} = \{N, -(N-1), \dots, -1, 0, +1, \dots, N-1, N\}.$$

The starting state is 0, and there the agent can choose among two actions (-1 and $+1$). In states $s \neq 0$ only one action is available. Formally

$$\mathcal{A}_s = \begin{cases} \{-1\} & \text{if } s < 0 \\ \{-1, +1\} & \text{if } s = 0 \\ \{+1\} & \text{if } s > 0. \end{cases} \tag{25}$$

The reward is everywhere zero except in the terminal states $-N, +N$ where it is $-1, +1$, respectively, for the only action available there. The transition function is deterministic, and the successor state is always $s' = s + a$ (e.g., action $+1$ in state $+2$ leads to state $+3$). In other words, if the agent is a state s with positive value, it will move to $s+1$, and if s has negative value it will move to $s-1$.

Low Rank $\subseteq$ Restricted Closedness

We first prove that a low-rank MDP must satisfy the restricted closedness assumption. Assume the MDP is low rank. Then for any $Q_{h+1} \in \mathcal{Q}_{h+1}$ and $\pi \in \Pi$, we have

$$\begin{aligned} \mathcal{T}_h^\pi Q_{h+1} &= \langle \phi_h(s,a),\, w_h^R \rangle + \left\langle \phi_h(s,a),\, \int_{s'} \mathbb{E}_{a' \sim \pi} Q_{h+1}(s', a') d\psi(s') \right\rangle \\ &= \left\langle \phi_h(s,a),\, w_h^R + \int_{s'} \mathbb{E}_{a' \sim \pi} Q_{h+1}(s', a') d\psi(s') \right\rangle \\ &= \langle \phi_h(s,a),\, w \rangle \end{aligned}$$

for some $w \in \mathbb{R}^d$. Thus, we have $(\mathcal{T}_h^\pi Q_{h+1}) \in \mathcal{Q}_h$ for all $Q_{h+1} \in \mathcal{Q}_{h+1}$ and $\pi \in \Pi$—i.e., if the MDP is low rank then it satisfies the restricted closedness condition.

To show the strict inclusion, consider the MDP described at the beginning of the proof with the following feature extractor:

$$\phi(s,a) = \begin{cases} +1 & \text{if } a = +1 \\ -1 & \text{if } a = -1. \end{cases} \tag{26}$$

The MDP with this feature map is not low rank. For example, we must have

$$1 = \mathbb{P}(N \mid N-1, +1) = \phi(N-1, +1)^\top \psi(N) = \psi(N)$$

which implies $\psi(-N) = 0$ for ψ to be a measure. However, this means we won't be able to represent all transitions correctly, as we would need to have

$$1 = \mathbb{P}(-N \mid -(N-1), -1) = \phi(-(N-1), -1)^\top \psi(-N) = -\psi(-N) = 0.$$

This means the MDP is not low rank. However, we show that it still satisfies the restricted closedness assumption. Notice that it is enough to verify the condition in the reachable space, which is $|s| + 1 = h$ at timestep h. If the reward is zero it suffices to verify that for all choices of θ_{h+1} we can find θ_h such that

$$\langle \phi(h-1, +1),\, \theta_h \rangle = \langle \phi(h, +1),\, \theta_{h+1} \rangle \tag{27}$$

$$\langle \phi(-(h-1), -1),\, \theta_h \rangle = \langle \phi(-h, -1),\, \theta_{h+1} \rangle. \tag{28}$$

Notice that in all cases there is only one policy available at the successor states; for any choice of θ_{h+1}, just set $\theta_h = \theta_{h+1}$. It is easy to verify that at the last step $h = H = N+1$ the reward function is either $+1$ or -1, depending on the state, and can be represented by $\theta_h = +1$:

$$\langle \phi(H-1, +1),\, \theta_H \rangle = +1 \tag{29}$$

$$\langle \phi(-(H-1), -1),\, \theta_H \rangle = -1. \tag{30}$$

Restricted Closedness $\subseteq$ Linear Q^π. We first show that every MDP that satisfies restricted closedness satisfies the linear Q^π assumption. For any timestep $h \in H$, and for a given policy $\pi \in \Pi$, if restricted closedness holds, choose $Q_{h+1} = Q^\pi_{h+1}$ in the definition of restricted closedness and use the Bellman equations to obtain

$$Q^\pi_h \overset{def}{=} \mathcal{T}^\pi_h Q^\pi_{h+1} \in \mathcal{Q}_h. \tag{31}$$

Thus, the linear Q^π assumption is automatically satisfied.

In order to show the strict inclusion, consider again the MDP described at the beginning of the proof, but with a different feature map. The map reads

$$\phi(s,a) = \begin{cases} [+1, 0] & \text{if } a = +1, s \neq 0 \\ [0, +1] & \text{if } a = -1, s \neq 0, \end{cases} \tag{32}$$

and at the start state

$$\phi(0,a) = \begin{cases} +1 & \text{if } a = +1 \\ -1 & \text{if } a = -1. \end{cases} \tag{33}$$

Notice that we only need to verify that restricted closedness does not hold at some timestep. When $\theta_2 = [+1, +1]$, there is no θ_1 such that

$$+\theta_1 = \langle \phi(0, +1), \, \theta_1 \rangle = \langle \phi(1, 1), \, \theta_2 \rangle = 1 \tag{34}$$

$$-\theta_1 = \langle \phi(0, -1), \, \theta_1 \rangle = \langle \phi(-1, -1), \, \theta_2 \rangle = 1. \tag{35}$$

The MDP however satisfies the linear Q^π assumption with $\theta_1 = 1$ and $\theta_h = [+1, -1]$ for $h \geq 2$.

C Proof of Theorem 1

For each iteration $t \in [T]$, let $\pi_t \overset{def}{=} \pi_{\theta_t}$ be the policy chosen by the actor, and let $M_t = M_{\pi_t}$ be the corresponding induced MDP.

Given the "good" event $\mathcal{G}$ defined in equation (47), Lemma 5 guarantees that it occurs with probability at least $1-\delta$. Conditioned on the occurrence of $\mathcal{G}$, the bounds (43a) and (43b) ensure that for any comparator $\widetilde{\pi}$, we have

$$\begin{aligned} V_1^{\widetilde{\pi}}(s_1) - V_1^{\pi_t}(s_1) &\le V_{1,M_t}^{\widetilde{\pi}}(s_1) - V_{1,M_t}^{\pi_t}(s_1) + 2\sum_{h=1}^{H}\left[\nu_h + \sqrt{\alpha_h}\|\mathbb{E}_{(S_h,A_h)\sim\widetilde{\pi}}\phi(S_h,A_h)\|_{\Sigma_h^{-1}}\right] \\ &= V_{1,M_t}^{\widetilde{\pi}}(s_1) - V_{1,M_t}^{\pi_t}(s_1) + \mathcal{E}_{\text{msp}}(\nu) + \mathcal{U}(\widetilde{\pi}). \end{aligned}$$

Now average over the iterations $t \in [T]$; Lemma 3 ensures that the actor in every iteration k receives a vector $\underline{w}_t$ which identifies the action value function for π_t on the MDP M_t it induces, i.e., such that

$$Q_{h,M_t}^{\pi_t}(s,a) \overset{Lem. 3}{=} \underline{Q}_h^{\pi_t}(s,a) = \langle \phi_h(s,a), \underline{w}_{hk} \rangle \,. \tag{36}$$

In other words, the action value function $\underline{Q}^{\pi_t}$ that the actor implicitly receives through $\underline{w}_t$ is the action value function of π_t on its induced MDP M_t, i.e., $Q_{M_t}^{\pi_t}$. Then Proposition 6 *(Actor's Analysis)* gives

$$\frac{1}{T}\sum_{t=1}^{T}\left[V_{1,M_t}^{\widetilde{\pi}}(s_1) - V_{1,M_t}^{\pi_t}(s_1)\right] \le \mathcal{C}(T).$$

Combining with the prior display yields

$$V_1^{\widetilde{\pi}}(s_1) - \frac{1}{T}\sum_{t=1}^{T} V_1^{\pi_t}(s_1) \le \mathcal{C}(T) + \mathcal{E}_{\text{msp}}(\nu) + \mathcal{U}(\widetilde{\pi}).$$

Notice that the policy returned by the agent π_{ALG} is the mixture policy of the policies $\pi_1, \dots, \pi_T$ and its value function is $V^{\pi_{\text{ALG}}} = \frac{1}{T}\sum_{t=1}^{T} V^{\pi_t}$.

Since the above statement holds under the good event $\mathcal{G}$ for any comparator policy $\widetilde{\pi}$, rearranging and taking $\sup$ over all comparator policies (not necessarily in Π) concludes the proof.

D Critic's Analysis

Given a policy π, the critic returns the pair $(\underline{\xi}^\pi, \underline{w}^\pi) = \{(\underline{\xi}_h^\pi, \underline{w}_h^\pi)\}_{h=1}^H$ if a feasible solution is found. The weight vectors induce the estimated value functions

$$\underline{Q}_h^\pi(s,a) \overset{def}{=} \langle \phi(s,a), \underline{w}_h^\pi \rangle, \qquad \text{and} \quad \underline{V}_h^\pi(s) \overset{def}{=} \mathbb{E}_{A' \sim \pi_h(\cdot|s)} \underline{Q}_h^\pi(s, A'). \tag{37}$$

Our ultimate goal is to relate the critic-estimated value functions to the true value functions $\{Q_h^\pi\}_{h=1}^H$.

D.1 Induced MDP and critic's guarantee

Essential to our analysis is an object that provides the essential link between the critic's output and the actor's input. In particular, it is helpful to understand the critic in the following way: when given a policy π as input, the critic computes the estimates $\{\underline{Q}_h^\pi\}_{h=1}^H$, and uses them form a new MDP $\hat{M}(\pi)$, which we refer to as the *induced MDP*. This new MDP shares the same state/action space and transition dynamics with the original MDP M, differing only in the perturbation of the reward function. In particular, for each $h \in [H]$, we define the *perturbed reward function*

$$\widehat{r}_h^\pi(s,a) \overset{def}{=} r_h(s,a) + \underline{Q}_h^\pi(s,a) - \mathcal{T}_h^\pi(\underline{Q}_{h+1}^\pi)(s,a). \tag{38}$$

The induced MDP $\hat{M}(\pi)$ is simply the original MDP that uses this perturbed reward function.

The key property of the induced MDP—which motivates the definition (38)—is that the estimates (37) returned by the critic correspond to the *exact value functions* of policy π in the induced MDP:

Lemma 3 (Critic exactness in induced MDP)**.** *Given a policy π as input, the critic returns a sequence $\{\underline{V}_h^\pi\}_{h=1}^H$ such that*

$$\underline{Q}_h^\pi = Q_{h,\hat{M}(\pi)}^\pi \tag{39}$$

$$\underline{V}_h^\pi = V_{h,\hat{M}(\pi)}^\pi \qquad \text{for all } h \in [H], \tag{40}$$

where $V_{h,\hat{M}(\pi)}^\pi$ is the exact value function of policy π in the induced MDP $\hat{M}(\pi)$.

See Section D.2 for the proof of this claim.

Moreover, since the induced MDP differs from the original MDP only in terms of the reward perturbation (38), we have the following convenient property: for any policy $\widetilde{\pi}$—which need not be of the soft-max form—the definition of value functions ensures that

$$V_{1,\hat{M}(\pi)}^{\widetilde{\pi}}(s_1) - V_1^{\widetilde{\pi}}(s_1) = \sum_{h=1}^H \mathbb{E}_{(S_h,A_h)\sim\widetilde{\pi}} \Big[\widehat{r}_h^\pi(S_h,A_h) - r_h(S_h,A_h) \Big], \tag{41}$$

where $V_{1,\hat{M}(\pi)}^{\widetilde{\pi}}$ is the value function of $\widetilde{\pi}$ in the induced MDP. This simple relation will allow us to use the induced MDP to relate arbitrary policies to their exact value functions.

With these two properties in mind, let us state our main guarantee for the critic. In the following proposition, $R > 0$ is an upper bound on the ℓ_2-radius of the actor parameter. When the number of actor iterations N is fixed, the maximum ℓ_2-norm of the actor's parameter θ (i.e., the value of R) is also fixed (since the learning rate η is fixed and $\|w_h\|_2 \leq 1, \forall h \in [H]$). If N is not know, one can easily perform an additional union bound in the proposition below.

Proposition 5. *For any fixed $R > 0$, given a failure probability $\delta \in (0,1)$, suppose that we set*

$$\begin{aligned} \sqrt{\alpha_h(\delta)} &\overset{def}{=} \sqrt{\lambda} + \sqrt{N}\nu_h \\ &\quad + c\left\{1 + d_h \log\left(1 + \tfrac{N}{d_h\lambda}\right) + \log\left(1 + 8\sqrt{N}\right) + d\log\left(1 + \tfrac{16R}{\epsilon}\right) + \log\frac{H}{\delta}\right\}^{1/2} \end{aligned} \tag{42}$$

for a suitably large universal constant c. Then with probability at least $1-\delta$, uniformly for any policy π in the soft-max class $\Pi_{soft}(R)$, the critic returns an induced MDP $\hat{M}(\pi)$ such that:

(a) *For the given policy π, we have*

$$V^{\pi}_{1,\hat{M}(\pi)}(s_1) \leq V^{\pi}_1(s_1) + \sum_{h=1}^{H} \nu_h. \tag{43a}$$

(b) *For any policy $\widetilde{\pi}$, not necessarily in the soft-max class Π, we have*

$$\left| V^{\widetilde{\pi}}_{1,\hat{M}(\pi)}(s_1) - V^{\widetilde{\pi}}_1(s_1) \right| \leq 2 \sum_{h=1}^{H} \sqrt{\alpha_h(\delta)}\; \|\bar{\phi}^{\widetilde{\pi}}_h\|_{\Sigma_h^{-1}} + \sum_{h=1}^{H} \nu_h, \tag{43b}$$

where $\bar{\phi}^{\widetilde{\pi}}_h \overset{def}{=} \mathbb{E}_{(S_h, A_h) \sim \widetilde{\pi}}[\phi_h(S_h, A_h)]$.

The remainder of this section is devoted to the proof of these two claims.

D.2 Proof of Lemma 3

Before proving Proposition 5, let us quickly prove Lemma 3. By definition, the induced MDP differs from the original MDP only by the perturbation of the reward function. Thus, by definition of value functions, we can write

$$Q^{\pi}_{h,\hat{M}(\pi)}(s,a) - Q^{\pi}_h(s,a) = \sum_{\ell=h}^{H} \mathbb{E}_{(S_\ell, A_\ell) \sim \pi | (s,a)} \left[\widehat{r}^{\pi}_h(S_\ell, A_\ell) - r_h(S_\ell, A_\ell) \right]. \tag{44a}$$

On the other hand, using the definition of $\underline{Q}^{\pi}_h$ and the Bellman conditions, we have

$$\begin{aligned} \underline{Q}^{\pi}_h(s,a) - Q^{\pi}_h(s,a) &= \langle \phi(s,a),\, \underline{w}^{\pi}_h \rangle - \mathcal{T}^{\pi}_h(Q^{\pi}_{h+1})(s,a) \\ &= \Big\{ \langle \phi(s,a),\, \underline{w}^{\pi}_h \rangle - \mathcal{T}^{\pi}_h(\underline{Q}^{\pi}_{h+1})(s,a) \Big\} + \Big\{ \mathcal{T}^{\pi}_h(\underline{Q}^{\pi}_{h+1})(s,a) + \mathcal{T}^{\pi}_h(Q^{\pi}_{h+1})(s,a) \Big\} \\ &= \widehat{r}^{\pi}_h(s,a) - r_h(s,a) + \mathbb{E}_{S' \sim \mathbb{P}_h(s,a)} \mathbb{E}_{A' \sim \pi(\cdot | S')} (\underline{Q}^{\pi}_{h+1} - Q^{\pi}_{h+1})(S', A') \end{aligned}$$

Applying this argument recursively to $\ell = h+1, \ldots, H$, we find that

$$\underline{Q}^{\pi}_h(s,a) - Q^{\pi}_h(s,a) = \sum_{\ell=h}^{H} \mathbb{E}_{(S_\ell, A_\ell) \sim \pi | (s,a)} \Big[\widehat{r}^{\pi}_h(S_\ell, A_\ell) - r_h(S_\ell, A_\ell) \Big] \tag{44b}$$

Subtracting equation (44b) from equation (44a) yields the claim.

D.3 Proof of Proposition 5

Let $\alpha = (\alpha_1, \ldots, \alpha_H)$ denote an arbitrary vector of non-negative pessimism parameters. Underlying Proposition 5 is a "good" event $\mathcal{G}(\alpha)$ to be defined momentarily. Our proof consists of two parts:

(i) First, we show that conditionally on $\mathcal{G}(\alpha)$, the two bounds in Proposition 5 hold.

(ii) Second, we show that with the choice of $\alpha(\delta)$ given in equation (42), the event $\mathcal{G}(\alpha)$ holds with probability at least $1 - \delta$.

We now define the good event $\mathcal{G}(\alpha)$. In order to do so, we need to introduce some auxiliary operators that play a key role in our analysis. Let $\mathcal{F}$ denote the space of all real-valued functions on $\mathcal{S} \times \mathcal{A}$. For an arbitrary $F \in \mathcal{F}$, we define the *sup-norm projection operator*

$$\mathcal{P}^{\pi}_h(F) \overset{def}{=} \arg \min_{w_h \in \mathcal{B}(\rho^w_h)} \sup_{(s,a)} \Big| \langle \phi(s,a),\, w_h \rangle - \left(\mathcal{T}^{\pi}_h F \right)(s,a) \Big|. \tag{45a}$$

Note that $\mathcal{P}^{\pi}_h$ is a mapping from $\mathcal{F}$ to $\mathbb{R}^d$; it returns the weight vector of the best-fitting linear function to the Bellman update $\mathcal{T}^{\pi}_h(F)$. Associated with this projection operator is the *approximation error operator*

$$\mathcal{A}^{\pi}_h(F)(s,a) \overset{def}{=} \langle \phi(s,a),\, \mathcal{P}^{\pi}_h(F) \rangle - \left(\mathcal{T}^{\pi}_h F \right)(s,a), \tag{45b}$$

which is a mapping from $\mathcal{F}$ to itself. We also define the *regression operator*

$$\mathcal{R}_h^\pi(F) \overset{def}{=} \Sigma_h^{-1} \sum_{k=1}^{N} \phi_{hk} \big\{ r_{hk} + \mathbb{E}_{A' \sim \pi(\cdot | s_{hk})} F(s_{h+1,k}, A') \big\}, \tag{45c}$$

which is another mapping from $\mathcal{F}$ to $\mathbb{R}^d$. To appreciate the relevance of the regression operator, note that we have $\underline{w}_h^\pi = \underline{\xi}_h^\pi + \mathcal{R}_h^\pi(\underline{Q}_{h+1}^\pi)$, by definition of the critic.

Our good event is defined in terms of the *parameter error operators* $\mathcal{E}_h^\pi : \mathcal{F} \to \mathbb{R}^d$ given by

$$\mathcal{E}_h^\pi(F) \overset{def}{=} \mathcal{R}_h^\pi(F) - \mathcal{P}_h^\pi(F). \tag{46}$$

With this set-up, we have the following:

Definition 2 (A "good" event)**.** *Given a sequence $\alpha = (\alpha_1, \ldots, \alpha_H)$ of pessimism parameters, define*

$$\mathcal{G}(\alpha) \overset{def}{=} \Big\{ \sup_{\substack{Q_{h+1} \in \mathcal{Q}_{h+1} \\ \pi_{h+1} \in \Pi_{h+1}}} \|\mathcal{E}_h^{\pi_{h+1}}(Q_{h+1})\|_{\Sigma_h} \leq \sqrt{\alpha_h}, \qquad \textit{for all } h \in [H] \Big\}. \tag{47}$$

Given this event, the proof of Proposition 5 can be reduced to proving the following two lemmas.

Lemma 4. *For any vector $\alpha \in \mathbb{R}^H$ of non-negative weights, conditionally on the event $\mathcal{G}(\alpha)$, the bounds (43a) and (43b) hold.*

Lemma 5. *For any $\delta \in (0, 1)$, given the choice of pessimism vector $\alpha(\delta)$ in equation (42), we have*

$$\mathbb{P}\big[\mathcal{G}(\alpha(\delta))\big] \geq 1 - \delta. \tag{48}$$

Note that the result of Proposition 5 follows as a direct consequence of these two claims. Thus, the remainder of our effort is devoted to prove these auxiliary results, with Sections D.4 and D.5 devoted to the proofs of Lemmas 4 and 5, respectively.

D.4 Proof of Lemma 4

We split the proof into two parts, corresponding to the two bounds.

D.4.1 Proof of Lemma 4(a)

We first prove the bound (43a) stated in part (a).

High-level roadmap: We begin by outlining the main steps in the proof. Our first step is to define a sequence of weight vectors $\widehat{w} \overset{def}{=} \{\widehat{w}_h^\pi\}_{h=1}^H$ such that

$$\Big| \sum_{a_1 \in \mathcal{A}} \pi(a_1 \mid s_1) \, \langle \phi_1(s_1, a_1), \, \widehat{w}_1^\pi \rangle - V_1^\pi(s_1) \Big| \leq \sum_{h=1}^{H} \nu_h. \tag{49a}$$

Our second step is to show that conditioned on $\mathcal{G}(\alpha)$, the sequence $\widehat{w}$ is feasible for the critic's convex program; this feasibility, combined with the optimality of $\underline{w}$, implies that

$$V_{1,\hat{M}(\pi)}^\pi(s_1) \overset{(i)}{=} \sum_{a_1 \in \mathcal{A}} \pi(a_1 \mid s_1) \, \langle \phi_1(s_1, a_1), \, \underline{w}_1^\pi \rangle \;\leq\; \sum_{a_1 \in \mathcal{A}} \pi(a_1 \mid s_1) \, \langle \phi_1(s_1, a_1), \, \widehat{w}_1^\pi \rangle \,. \tag{49b}$$

Here step (i) follows from Lemma 3, which guarantees that the estimated value functions $\underline{V}_h^\pi$ of the critic are exact in the induced MDP. Combining the two bounds (49a) and (49b) yields $V_{1,\hat{M}(\pi)}^\pi(s_1) \leq V_1^\pi(s_1) + \sum_{h=1}^H \nu_h$, as claimed in equation (43a).

It remains to prove our two auxiliary claims (49a) and (49b).

Proof of claim (49a): Given a policy π, we use backwards induction to define the sequence $\{\widehat{w}^\pi\}_{h=1}^H$ by first setting $\widehat{w}^\pi_{H+1} = 0$, and then defining

$$\widehat{w}^\pi_h \overset{def}{=} \mathcal{P}^\pi_h(\widehat{Q}^\pi_{h+1}) \qquad \text{for } h = H, H-1, \ldots, 1, \tag{50}$$

where $\widehat{Q}^\pi_{h+1}(s,a) \overset{def}{=} \langle \phi_{h+1}(s,a),\, \widehat{w}^\pi_{h+1} \rangle$. Notice that by construction, we have the bound $\|\widehat{w}^\pi_h\|_2 \leq \rho^w_h$ for all $h \in [H]$. The following lemma bounds the sup-norm distance between the induced linear Q-value function estimate, and the actual Q^π-value function.

Lemma 6 (Accuracy of Best Predictor)**.** *The functions $\{\widehat{Q}^\pi_h\}_{h=1}^H$ defined by the best-predictor sequence $\{\widehat{w}^\pi_h\}_{h=1}^H$ from equation (50) satisfy the bound*

$$\left|\widehat{Q}^\pi_h(s,a) - Q^\pi_h(s,a)\right| \leq \sum_{\ell=h}^H \nu_\ell \qquad \textit{for all } h \in [H]. \tag{51}$$

Proof. Introduce the shorthand $\Delta_h(s,a) \overset{def}{=} \widehat{Q}^\pi_h(s,a) - Q^\pi_h(s,a)$ for the error at stage h to be bounded. Since $Q^\pi_h = \mathcal{T}^\pi_h(Q^\pi_{h+1})$, we can write

$$\begin{aligned}
\Delta_h(s,a) &= \widehat{Q}^\pi_h(s,a) - Q^\pi_h(s,a) \\
&= \widehat{Q}^\pi_h(s,a) - (\mathcal{T}^\pi_h \widehat{Q}^\pi_{h+1})(s,a) + (\mathcal{T}^\pi_h \widehat{Q}^\pi_{h+1})(s,a) - \mathcal{T}^\pi_h(Q^\pi_{h+1})(s,a) \\
&= \widehat{Q}^\pi_h(s,a) - (\mathcal{T}^\pi_h \widehat{Q}^\pi_{h+1})(s,a) + \mathbb{E}_{S' \sim \mathbb{P}_h(s,a)} \mathbb{E}_{A' \sim \pi(\cdot \mid S')}\Big[\widehat{Q}^\pi_{h+1}(S',A') - Q^\pi_{h+1}(S',A')\Big] \\
&= \sum_{\ell=h}^H \mathbb{E}_{(S_\ell, A_\ell) \sim \pi \mid (s,a)}\Big[\widehat{Q}^\pi_\ell(S_\ell, A_\ell) - \mathcal{T}^\pi_\ell(\widehat{Q}^\pi_{\ell+1})(S_\ell, A_\ell)\Big],
\end{aligned}$$

where the final equality follows by induction.

From the definition (50) of $\widehat{w}$ and the function estimate $\widehat{Q}^\pi_\ell(s,a) = \langle \phi_\ell(s,a),\, \widehat{w}^\pi_\ell \rangle$, combined with the Bellman approximation condition, we have

$$\left|\widehat{Q}^\pi_\ell(s,a) - (\mathcal{T}^\pi_\ell \widehat{Q}^\pi_{\ell+1})(s,a)\right| \leq \mathcal{A}^\pi_\ell(\widehat{Q}^\pi_{\ell+1}) \;\leq\; \nu_\ell,$$

uniformly over all ℓ, and over all state-action pairs (s,a). Summing these bounds completes the proof. □

Proof of claim (49b): In order to prove this claim, we need to exhibit a sequence $\xi = (\widehat{\xi}_1, \ldots, \widehat{\xi}_H)$ such that the pair $(\widehat{\xi}, \widehat{w})$ are feasible for the critic's convex program (9). In particular, we need to ensure the following three conditions:

(a) $\|\widehat{w}^\pi_h\|_2 \leq \rho^w_h$ for all $h \in [H]$

(b) $\|\widehat{\xi}_h\|_{\Sigma_h} \leq \sqrt{\alpha_h}$ for all $h \in [H]$.

(c) We have $\widehat{w}^\pi_h = \widehat{\xi}^\pi_h + \mathcal{R}^\pi_h(\widehat{Q}^\pi_{h+1})$ for all $h \in [h]$.

Note that condition (a) is automatically satisfied by the definition (50) of $\widehat{w}$, since the projection $\mathcal{P}^\pi_h$ imposes this Euclidean norm bound.

It remains to exhibit a choice of $\widehat{\xi}$ such that conditions (b) and (c) hold. Since $\widehat{w}^\pi_h = \mathcal{P}^\pi_h(\widehat{Q}^\pi_h)$ by definition, condition (c) forces us to set

$$\widehat{\xi}^\pi_h = \mathcal{P}^\pi_h(\widehat{Q}^\pi_{h+1}) - \mathcal{R}^\pi_h(\widehat{Q}^\pi_{h+1}) \;=\; -\mathcal{E}^\pi_h(\widehat{Q}^\pi_{h+1}).$$

But since the event $\mathcal{G}(\alpha)$ holds by assumption, we have

$$\|\widehat{\xi}^\pi_h\|_{\Sigma_h} = \|\mathcal{E}^\pi_h(\widehat{Q}^\pi_{h+1})\|_{\Sigma_h} \leq \sqrt{\alpha_h},$$

showing that this choice of $\widehat{\xi}$ satisfies condition (b).

D.4.2 Proof of Lemma 4(b)

Here we prove the bound (43b) stated in part (b) of the lemma.

Our proof is based on establishing an auxiliary result that implies the claim. In particular, we first show that for any policy $\widetilde{\pi}$, we have

$$\left|V^{\widetilde{\pi}}_{1,\hat{M}(\pi)}(s_1) - V^{\widetilde{\pi}}_1(s_1)\right| \leq \sum_{h=1}^{H} \|\bar{\phi}^{\widetilde{\pi}}_h\|_{\Sigma_h^{-1}} \left\{\sqrt{\alpha_h} + \|\mathcal{E}^{\pi}_h(\underline{Q}^{\pi}_{h+1})\|_{\Sigma_h}\right\} + \sum_{h=1}^{H} \nu_h, \tag{52}$$

where $\bar{\phi}^{\widetilde{\pi}}_h \overset{def}{=} \mathbb{E}_{(S_h,A_h)\sim\widetilde{\pi}}[\phi(S_h, A_h)]$. Since $\|\mathcal{E}^{\pi}_h(\underline{Q}^{\pi}_{h+1})\|_{\Sigma_h} \leq \sqrt{\alpha_h}$ conditioned on $\mathcal{G}(\alpha)$, this implies the claim.

Let us now prove the auxiliary claim (52). First, we observe that by definition, the perturbation in the reward can be written as

$$\begin{aligned}
\widehat{r}^{\pi}_h(s,a) - r_h(s,a) &\overset{(i)}{=} \langle \phi_h(s,a),\, \underline{w}^{\pi}_h \rangle - \mathcal{T}^{\pi}_h(\underline{Q}^{\pi}_{h+1})(s,a) \\
&\overset{(ii)}{=} \Big\langle \phi_h(s,a),\, \underline{\xi}^{\pi}_h \Big\rangle + \Big\langle \phi_h(s,a),\, \mathcal{R}^{\pi}_h(\underline{Q}^{\pi}_{h+1}) \Big\rangle - \mathcal{T}^{\pi}_h(\underline{Q}^{\pi}_{h+1})(s,a) \\
&\overset{(iii)}{=} \Big\langle \phi_h(s,a),\, \underline{\xi}^{\pi}_h \Big\rangle + \Big\langle \phi_h(s,a),\, \mathcal{E}^{\pi}_h(\underline{Q}^{\pi}_{h+1}) \Big\rangle + \mathcal{A}^{\pi}_h(\underline{Q}^{\pi}_{h+1})(s,a),
\end{aligned}$$

where step (i) uses the definition $\underline{Q}^{\pi}_h(s,a) = \langle \phi_h(s,a), \underline{w}^{\pi}_h \rangle$; step (ii) uses the relation $\underline{w}^{\pi}_h = \underline{\xi}^{\pi}_h + \mathcal{R}^{\pi}_h(\underline{Q}^{\pi}_{h+1})$; and step (iii) involves adding and subtracting $\Big\langle \phi_h(s,a),\, \mathcal{P}^{\pi}_h(\underline{Q}^{\pi}_{h+1}) \Big\rangle$, and using the definitions of the approximation error (45b) and the error operator (46).

Since the induced MDP differs from the original only by the reward perturbation, we have

$$\begin{aligned}
\left|V^{\widetilde{\pi}}_{1,\hat{M}(\pi)}(s_1) - V^{\widetilde{\pi}}_1(s_1)\right| &= \left|\sum_{h=1}^{H} \mathbb{E}_{(S_h,A_h)\sim\widetilde{\pi}}\left[\widehat{r}^{\pi}_h(S_h,A_h) - r_h(S_h,A_h)\right]\right| \\
&= \left|\sum_{h=1}^{H} \mathbb{E}_{(S_h,A_h)\sim\widetilde{\pi}}\left[\Big\langle \phi_h(S_h,A_h),\, \underline{\xi}^{\pi}_h + \mathcal{E}^{\pi}_h(\underline{Q}^{\pi}_{h+1}) \Big\rangle + \mathcal{A}^{\pi}_h(\underline{Q}^{\pi}_{h+1})(S_h,A_h)\right]\right|.
\end{aligned}$$

We now observe that $|\mathcal{A}^{\pi}_h(\underline{Q}^{\pi}_{h+1})(S_h,A_h)| \leq \nu_h$ by the Bellman closure assumption. As for the first term, introducing the shorthand $\bar{\phi}^{\widetilde{\pi}}_h \overset{def}{=} \mathbb{E}_{(S_h,A_h)\sim\widetilde{\pi}}\big[\phi_h(S_h,A_h)\big]$, we have

$$\begin{aligned}
\mathbb{E}_{(S_h,A_h)\sim\widetilde{\pi}}\left[\Big\langle \phi_h(S_h,A_h),\, \underline{\xi}^{\pi}_h + \mathcal{E}^{\pi}_h(\underline{Q}^{\pi}_{h+1}) \Big\rangle\right] &\leq \|\bar{\phi}^{\widetilde{\pi}}_h\|_{\Sigma_h^{-1}}\, \|\underline{\xi}^{\pi}_h + \mathcal{E}^{\pi}_h(\underline{Q}^{\pi}_{h+1})\|_{\Sigma_h} \\
&\leq \|\bar{\phi}^{\widetilde{\pi}}_h\|_{\Sigma_h^{-1}} \Big\{\sqrt{\alpha_h} + \|\mathcal{E}^{\pi}_h(\underline{Q}^{\pi}_{h+1})\|_{\Sigma_h}\Big\},
\end{aligned}$$

where the final step combines the triangle inequality, with the fact that $\|\underline{\xi}^{\pi}_h\|_{\Sigma_h} \leq \sqrt{\alpha_h}$, since $\underline{\xi}^{\pi}_h$ must be feasible for the critic's convex program (9). Putting together the pieces yields the claim (52).

D.5 Proof of Lemma 5

Recall from equation (46) that for any pair (Q, π), the parameter error is given by $\mathcal{E}^{\pi}_h(Q) = \mathcal{R}^{\pi}_h(Q) - \mathcal{P}^{\pi}_h(Q)$. We begin with a simple lemma that decomposes this error into three terms. In order to state the lemma, we introduce two forms of error variables: statistical and approximation-theoretic. The first noise variables take the form

$$\eta_{hk}(Q,\pi) \overset{def}{=} r_{hk} + \mathbb{E}_{A'\sim\pi(\cdot|s_{hk})} Q(s_{h+1,k}, A') - (\mathcal{T}^{\pi}_h Q)(s_{hk}, a_{hk}), \tag{53a}$$

defined for each $h \in [H]$ and $k \in [N]$. Note that conditionally on the pair (s_{hk}, a_{hk}), our sampling model and the definition of the Bellman operator $\mathcal{T}^{\pi}_h$ ensures that each η_{hk} is zero-mean random variable, corresponding to a form of statistical error. Our analysis also involves some approximation error terms, in particular via the quantities

$$\Delta_{hk}(Q,\pi) \overset{def}{=} -\mathcal{A}^{\pi}_h(Q)(s_{hk}, a_{hk}) \;=\; (\mathcal{T}^{\pi}_h Q)(s_{hk}, a_{hk}) - \langle \phi_h(s_{hk}, a_{hk}),\, \mathcal{P}^{\pi}_h(Q) \rangle \tag{53b}$$

With these definitions, we have the following guarantee:

Lemma 7 (Decomposition of $\mathcal{E}^\pi_h(Q)$)**.** *For any pair (Q, π), we have the decomposition*

$$\mathcal{E}^\pi_h(Q) = e^\eta_h(Q,\pi) + e^\lambda_h(Q,\pi) + e^\Delta_h(Q,\pi), \tag{54}$$

where the three error terms are given by

$$e^\eta_h(Q,\pi) \overset{def}{=} \Sigma_h^{-1} \sum_{k=1}^N \phi_{hk} \eta_{hk}(Q,\pi), \qquad \textit{(Statistical estimation error)} \tag{55a}$$

$$e^\lambda_h(Q,\pi) \overset{def}{=} -\lambda \Sigma_h^{-1} \mathcal{P}^\pi_h(Q), \qquad \textit{(Regularization error), \quad and} \tag{55b}$$

$$e^\Delta_h(Q,\pi) \overset{def}{=} \Sigma_h^{-1} \sum_{k=1}^N \phi_{hk} \Delta_{hk}(Q;\pi) \qquad \textit{(Approximation error).} \tag{55c}$$

See Section D.5.1 for the proof of this claim.

The remainder of our analysis is focused on bounding these three terms. Analysis of the regularization error and approximation error terms is straightforward, whereas bounding the statistical estimation error requires more technical effort. We begin with the two easy terms.

Regularization error: Beginning with the definition (55b), we have

$$\|e^\lambda_h(Q,\pi)\|_{\Sigma_h} = \lambda \|\mathcal{P}^\pi_h(Q)\|_{\Sigma_h^{-1}} \overset{(i)}{\leq} \sqrt{\lambda} \|\mathcal{P}^\pi_h(Q)\|_2 \overset{(ii)}{\leq} \sqrt{\lambda}, \tag{56}$$

where step (i) follows since $\Sigma_h \succeq \lambda I$; and inequality (ii) follows from the bound $\|\mathcal{P}^\pi_h(Q)\|_2 \leq \rho^w_h \leq 1$, guaranteed by the definition of $\mathcal{P}^\pi_h$.

Approximation error: By definition, we have $\|e^\Delta_h(Q,\pi)\|_{\Sigma_h} = \| \sum_{k=1}^N \phi_{hk} \Delta_{hk}(Q,\pi)\|_{\Sigma_h^{-1}}$. By the Bellman approximation condition, we have $|\Delta_{hk}(Q,\pi)| \leq \nu_h$ uniformly over all k. Consequently, applying Lemma 8 (Projection Bound) from the paper [Zanette et al., 2020b] guarantees that

$$\|e^\Delta_h(Q,\pi)\|_{\Sigma_h} \leq \sqrt{N} \nu_h. \tag{57}$$

Statistical estimation error: Lastly, we turn to the analysis of the statistical estimation error. In particular, we prove the following guarantee:

Lemma 8. *There is a universal constant $c > 0$ such that*

$$\|e^\eta_h(Q,\pi)\|^2_{\Sigma_h} \leq c \left\{ 1 + d_h \log\left(1 + \tfrac{N}{d_h \lambda}\right) + \log\left(1 + 8\sqrt{N}\right) + d \log\left(1 + \tfrac{16R}{\epsilon}\right) + \log \frac{H}{\delta} \right\} \tag{58}$$

uniformly over all $Q \in \mathcal{Q}_h$, $\pi \in \Pi_{soft}(R)$ and $h \in [H]$ with probability at least $1 - \delta$.

See Section D.5.2 for the proof of this claim.

Putting together the pieces: By combining our three bounds—namely, equations (56), (57) and (58), we conclude that with the choice

$$\sqrt{\alpha_h(\delta)} \overset{def}{=} \sqrt{\lambda} + \sqrt{N}\nu_h + \\ c \left\{ 1 + d_h \log\left(1 + \tfrac{N}{d_h \lambda}\right) + \log\left(1 + 8\sqrt{N}\right) + d \log\left(1 + \tfrac{16R}{\epsilon}\right) + \log \frac{H}{\delta} \right\}^{1/2},$$

the good event $\mathcal{G}(\delta)$ holds with probability at least $1 - \delta$. This completes the proof of Lemma 5.

It remains to prove the two auxiliary lemmas that we stated: namely, Lemma 7 that gave a decomposition of the parameter error, and Lemma 8 that bounded the statistical error. We do so in Sections D.5.1 and D.5.2, respectively.

D.5.1 Proof of Lemma 7

Starting with the definition (45c) of the regression operator $\mathcal{R}_h^\pi$, we have

$$\begin{aligned}
\mathcal{R}_h^\pi(Q) &\overset{def}{=} \Sigma_h^{-1} \sum_{k=1}^N \phi_{hk}[r_{hk} + \mathbb{E}_{A' \sim \pi(\cdot \mid s_{hk})} Q(s_{h+1,k}, A')] \\
&\overset{(i)}{=} \Sigma_h^{-1} \sum_{k=1}^N \phi_{hk}[(\mathcal{T}_h^\pi Q)(s_{hk}, a_{hk})] + \underbrace{\Sigma_h^{-1} \sum_{k=1}^N \phi_{hk} \eta_{hk}(Q, \pi)}_{= e_h^\eta(Q,\pi)}
\end{aligned}$$

where equality (i) follows by adding and subtracting terms, and using the definition (53a) of η_{hk}. Next we use the definition (53b) of the approximation error terms Δ_{hk} to find that

$$\mathcal{R}_h^\pi(Q) = \xi_h + \Sigma_h^{-1} \left(\sum_{k=1}^N \phi_{hk} \big[\langle \phi_{hk}, \, \mathcal{P}_h^\pi(Q) \rangle + \Delta_{hk}(Q, \pi) \big] \right) + e_h^\eta(Q, \pi)$$

Since $\Sigma_h = \sum_{k=1}^N \phi_{hk} \phi_{hk}^\top + \lambda I$, we can write

$$\begin{aligned}
w_h(Q, \pi, \xi_h) &= \xi_h + \Sigma_h^{-1} \Big\{ \Sigma_h w_h^\star(Q, \pi) + \sum_{k=1}^N \phi_{hk} \Delta_{hk}(Q, \pi) - \lambda w_h^\star(Q, \pi) \Big\} + e_h^\eta \\
&= \xi_h + w_h^\star(Q, \pi) + \Sigma_h^{-1} \left(\sum_{k=1}^N \phi_{hk} \Delta_{hk}(Q, \pi) - \lambda w_h^\star(Q, \pi) \right) + e_h^\eta \\
&= \xi_h + w_h^\star(Q, \pi) + e_h^\eta + e_h^\lambda + e_h^\Delta,
\end{aligned}$$

which completes the proof.

D.5.2 Proof of Lemma 8

Recall the definition $\mathcal{Q}$ in Definition 1 for the linear action value function with a prescribed choice for the radii $\{\rho_h^w\}_{h=1}^H$ such that $\rho_h^w \in [0, 1]$ $\forall h \in [H]$. In this section, the policy functional space Π_{soft} in Definition 1 (*Functional Spaces*) is identified by a fixed sequence of radii $\{\rho^\theta\}_{h=1}^H$ for the ℓ_2-norm of the actor parameter θ. The upper bound on ρ_h^θ only depends on the number of actor iteration T selected by the user and the learning rate η, according to the relation $\|\theta_{t,h}\|_2 = \| \sum_{t=1}^T \eta w_{t,h} \|_2 \leq \eta \sum_{t=1}^T \|w_{t,h}\|_2 \leq \eta T \rho_h^w \leq \eta T$.

We make use of a discretization argument to control the associated empirical process. Let $N_\infty(\epsilon; \mathcal{Q})$ denote the cardinality of the smallest ϵ-covering of $\mathcal{Q}$ in the sup-norm—that is, a collection $\{Q^i\}_{i=1}^N$ such that for all $Q \in \mathcal{Q}$, we can find some $i \in [N]$ such that $\|Q - Q^i\|_\infty = \sup_{(s,a)} |Q(s, a) - Q^i(s, a)| \leq \epsilon$. Similarly, we let $N_{\infty,1}(\epsilon; \Pi(R))$ denote an ϵ-cover of $\Pi(R)$ when measuring distances with the norm

$$\|\pi - \pi'\|_{\infty,1} \overset{def}{=} \sup_s \sum_{a \in \mathcal{A}} \big| \pi(a \mid s) - \pi'(a \mid s) \big|. \tag{59}$$

We have the following bounds on these covering numbers:

Lemma 9 (Covering number bounds)**.** *For any* $\epsilon \in (0, 1)$*, we have*

$$\log N_\infty(\epsilon; \mathcal{Q}) \leq d \log \left(1 + \tfrac{2}{\epsilon}\right) \qquad \textit{and} \tag{60a}$$

$$\log N_{\infty,1}(\epsilon; \Pi(R)) \leq d \log \left(1 + \tfrac{16R}{\epsilon}\right). \tag{60b}$$

See Section D.5.3 for the proofs of these claims.

For any $\epsilon \in (0, 1)$, we define

$$\beta(\epsilon) \overset{def}{=} d_h \log \left(1 + \tfrac{N}{d_h \lambda}\right) + \log N_\infty(\epsilon; \mathcal{Q}) + \log N_{\infty,1}(\epsilon; \Pi_{soft}) + \log \frac{H}{\delta} \tag{61}$$

Given this definition and the bounds from Lemma 9, the proof of Lemma 8 is reduced to showing that for any $\epsilon \in (0,1)$, there is a universal constant c such that

$$\max_{h \in [H]} \sup_{\substack{Q \in \mathcal{Q}_h \\ \pi \in \Pi_{soft}}} \|e^{\eta}_h(Q,\pi)\|_{\Sigma_h} \leq c\sqrt{\beta(\epsilon)} + 4\sqrt{N}\epsilon \tag{62}$$

with probability at least $1-\delta$. The claim stated in Lemma 8 follows from the choice $\epsilon = \frac{1}{4\sqrt{N}}$. The remainder of our proof is devoted to the proof of this claim.

Proof of the claim (62): Let us recall the definition

$$\eta_{hk}(Q,\pi) = r_{hk} + \mathbb{E}_{A' \sim \pi_h(\cdot|s_{hk})} Q(s_{h+1,k}, A') - (\mathcal{T}^{\pi}_h Q)(s_{hk}, a_{hk}).$$

Consequently, by starting with the definition of e^{η}_h and applying the triangle inequality, we obtain the upper bound $\|e^{\eta}_h(Q,\pi)\|_{\Sigma_h} = \|\sum_{k=1}^N \phi_{hk}\eta_{hk}(Q,\pi)\|_{\Sigma_h^{-1}} \leq Z_1 + Z_2(Q,\pi)$, where

$$Z_1 \overset{def}{=} \|\sum_{k=1}^{N} \phi_{hk} [\underbrace{r_{hk} - r(s_{hk}, a_{hk})}_{\overset{def}{=} Y_{hk}}]\|_{\Sigma_h^{-1}} \quad \text{and}$$

$$Z_2(Q,\pi) \overset{def}{=} \Big\| \sum_{k=1}^{N} \phi_{hk}[Q(s_{h+1,k},\pi) - \mathbb{E}_{S' \sim \mathbb{P}(\cdot|s_{hk},a_{hk})} Q(S',\pi)] \Big\|_{\Sigma_h^{-1}}$$

For a fixed (π, Q) and conditioned on the sampling history, both Z_1 and Z_2 are mean zero. Note that Z_1 is independent of the pair (Q,π), so that its analysis does not require discretization techniques. On the other hand, analyzing $Z_2(Q,\pi)$ does require a reduction step via discretization, with which we begin.

Introducing the shorthand $N = N(\epsilon, \mathcal{Q})$, let $\{Q^i\}_{i=1}^N$ be an ϵ-cover of the set $\mathcal{Q}$ in the sup-norm. Similarly, with the shorthand $J = N(\epsilon, \Pi)$, let $\{\pi^j\}_{j=1}^J$ be an ϵ-cover of Π in the norm (59). For a given Q, let Q^i denote the member of the cover such that $\|Q - Q^i\|_\infty \leq \epsilon$. With this choice, we have

$$Z_2(Q,\pi) = Z_2(Q^i,\pi) + \{Z_2(Q,\pi) - Z_2(Q^i,\pi)\}.$$

Similarly, let π^m be a member of the cover such that $\|\pi(\cdot \mid s) - \pi^m(\cdot \mid s)\|_1 \leq \epsilon$ for all s. With this choice, we have

$$Z_2(Q,\pi) \leq Z_2(Q^i,\pi^m) + \underbrace{\{Z_2(Q^i,\pi) - Z_2(Q^i,\pi^m)\}}_{D^\pi} + \underbrace{\{Z_2(Q,\pi) - Z_2(Q^i,\pi)\}}_{D^Q}.$$

We begin by bounding the two discretization errors. By the triangle inequality, we have

$$D^Q \leq \Big\| \sum_{k=1}^{N} \phi_{hk} \underbrace{[Q(s_{h+1,k},\pi) - Q^i(s_{h+1,k},\pi) + \mathbb{E}_{S' \sim p(s_{hk},a_{hk})}(Q(S',\pi) - Q^i(S',\pi))]}_{\overset{def}{=} E^i_{hk}(Q,\pi)} \Big\|_{\Sigma_h^{-1}}.$$

Our choice of discretization ensures that $|E^i_{hk}(Q,\pi)| \leq 2\epsilon$ uniformly for all (h,k) and (Q,π). Applying Lemma 8 (Projection Bound) from the paper [Zanette et al., 2020b] ensures that $D^Q \leq 2\epsilon\sqrt{N}$. To be clear, this is a deterministic claim; it holds uniformly over the choices of Q, Q^i, and π. A similar argument yields that $D^\pi \leq 2\epsilon\sqrt{N}$.

Putting togther the pieces yields that for any (Q,π), we have the bound

$$Z_2(Q,\pi) \leq \max_{\substack{i \in [N] \\ j \in [M]}} Z_2(Q^i,\pi^j) + 4\sqrt{N}\epsilon. \tag{63}$$

We now need to bound Z_1 along with $Z_2(Q^i,\pi^j)$ for a fixed pair (Q^i,π^j). In order to do so, we apply known self-normalized tail bounds [de la Pena et al., 2009], which apply to sums of the form

$\| \sum_{k=1}^{N} \phi_{hk} V_{hk} \|_{\Sigma_h^{-1}}$, where the V_{hk} form a martingale difference sequence with conditionally sub-Gaussian tails. Note that Z_1 is of this general form with $V_{hk} = Y_{hk}$, which is a 1-sub-Gaussian variable by assumption. On the other hand, the variable $Z_2(Q^i, \pi^j)$ is of this form with

$$V_{hk} = Q^i(s_{h+1,k}, \pi^j) - \mathbb{E}_{S' \sim p(s_{hk}, a_{hk})} Q^i(S', \pi^j).$$

Since $|V_{hk}| \leq 1$ due to the uniform boundedness of Q^i, this is a 1-sub-Gaussian variable as well.

Consequently, Theorem 1 from the paper [Abbasi-Yadkori et al., 2011] ensures that

$$\mathbb{P}\left(\max\{Z_1, Z_2(Q^i, \pi^j)\} \geq \log \frac{\det \Sigma_h}{\det \lambda I} + 2 \log \frac{1}{\delta} \right) \leq \delta.$$

Note that $\det \lambda I = \lambda^{d_h}$. Moreover, Lemma 10 (Determinant-Trace Inequality) in [Abbasi-Yadkori et al., 2011] yields $\log \det \Sigma_h \leq d_h \log \left(\lambda + \frac{N}{d_h} \right)$.

Putting together the pieces, taking a union bound over the two covers yields that, for each fixed $h \in [H]$, we have

$$\|e_h^\eta(Q, \pi)\|_{\Sigma_h^{-1}} \leq d_h \log\left(1 + \tfrac{N}{d_h \lambda}\right) + \log N_\infty(\epsilon; \mathcal{Q}) + \log N_{\infty,1}(\epsilon; \Pi) + \log\left(\tfrac{1}{\delta}\right) + 4\sqrt{N}\epsilon$$

with probability at least $1 - \delta$. Finally, we take a union bound over all $h \in [H]$, which forces us to redefine δ to $\frac{\delta}{H}$ in the above bound. This completes the proof of the uniform bound (62).

D.5.3 Proof of Lemma 9

Since $\|\phi(s,a)\|_2 \leq 1$, for any pair of weight vectors $w, w' \in \mathbb{R}^d$, we have $\sup_{(s,a)} |\langle \phi(s,a), \, w - w' \rangle \|_2 \leq \|w - w'\|_2$. Thus, the bound (60a) follows from standard results on coverings of Euclidean balls (cf. Example 5.8 in the book [Wainwright, 2019]).

As for the bound (60b), we claim that

$$\sum_{a \in \mathcal{A}} \left| \pi_{\theta'}(a \mid s) - \pi_\theta(a \mid s) \right| \leq 8 \|\theta - \theta'\|_2, \qquad \text{for all } s \in \mathcal{S}. \tag{64}$$

Taking this claim as given for the moment, it suffices to obtain an $\epsilon/8$-cover of the ball $\mathcal{B}(R)$ in the ℓ_2-norm, and applying the same standard results yields the claimed bound (60b).

It remains to prove the claim (64).

Proof of the claim (64): Let us state and prove the claim (64) more formally as a lemma. It applies to the softmax policy $\pi_\theta(a \mid s) = \frac{\exp\{\langle \phi(s,a), \, \theta \rangle\}}{\sum_{a' \in \mathcal{A}} \exp(\langle \phi(s,a'), \, \theta \rangle)}$.

Lemma 10 (Nearby Policies)**.** *Consider a feature mapping $\phi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$ such that $\|\phi(s,a)\|_2 \leq 1$ uniformly for all pairs (s,a). Then for all $s \in \mathcal{S}$, we have*

$$\sum_{a \in \mathcal{A}} \left| \pi_{\theta'}(a \mid s) - \pi_\theta(a \mid s) \right| \leq 8 \|\theta - \theta'\|_2, \tag{65}$$

valid for any pair $\theta, \theta' \in \mathbb{R}^d$ such that $\|\theta - \theta'\|_2 \leq \frac{1}{2}$.

Proof. Dividing $\pi_{\theta'}(s,a)$ by $\pi_\theta(s,a)$ yields

$$\begin{aligned}
T \overset{def}{=} \frac{\pi_{\theta'}(a \mid s)}{\pi_\theta(a \mid s)} &= \frac{e^{\langle \phi(s,a), \, \theta' \rangle}}{e^{\langle \phi(s,a), \, \theta \rangle}} \times \frac{\sum_{a''} e^{\langle \phi(s,a''), \, \theta \rangle}}{\sum_{\tilde{a}} e^{\langle \phi(s,\tilde{a}), \, \theta' \rangle}} \\
&= e^{\langle \phi(s,a), \, \theta' - \theta \rangle} \times \sum_{a''} \left(e^{\langle \phi(s,a''), \, \theta - \theta' \rangle} \times \frac{e^{\langle \phi(s,a''), \, \theta' \rangle}}{\sum_{\tilde{a}} e^{\langle \phi(s,\tilde{a}), \, \theta' \rangle}} \right) \\
&= e^{\langle \phi(s,a), \, \theta' - \theta \rangle} \times \sum_{a''} \pi_{\theta'}(a'' \mid s) e^{\langle \phi(s,a''), \, \theta - \theta' \rangle}.
\end{aligned}$$

By Cauchy-Schwarz and the assumption on ϕ, we have the bound $|\langle\theta(s,a),\, \gamma\rangle| \leq \|\gamma\|_2$, valid for any vector γ. Monotonicity of the exponential allows us to exponentiate this inequality. Combined with the fact that $\pi_{\theta'}(a'' \mid s) \geq 0$, we find that

$$T \leq e^{\|\theta'-\theta\|_2} \sum_{a'' \in \mathcal{A}} \pi_{\theta'}(a'' \mid s) e^{\|\theta-\theta'\|_2} \overset{(i)}{=} e^{2\|\theta-\theta'\|_2} \overset{(ii)}{\leq} 1 + 4\|\theta-\theta'\|_2, \tag{66}$$

where step (i) uses the fact that π_θ is a probability distribution over the action space; and step (ii) follows by combining the elementary inequality $e^x \leq 1 + 2x$, valid for all $x \in [0,1]$, with our assumption that $\|\theta - \theta'\|_2 \leq 1/2$.

Recalling that $T = \frac{\pi_{\theta'}(a|s)}{\pi_\theta(a|s)}$, re-arranging the inequality (66) yields the bound

$$\pi_{\theta'}(a \mid s) - \pi_\theta(a \mid s) \leq 4\pi_\theta(a \mid s)\ \|\theta - \theta'\|_2,$$

valid uniformly over all pairs (s, a). We can apply the same argument with the roles of θ and θ' reversed, and combining the two bounds yields

$$|\pi_{\theta'}(a \mid s) - \pi_\theta(a \mid s)| \leq 4\|\theta - \theta'\|_2 \max\{\pi_\theta(a \mid s),\ \pi_{\theta'}(a \mid s)\},$$

again uniformly over all pairs (s, a). Now summing over the actions a, we find that

$$\begin{aligned}\sum_{a \in \mathcal{A}} \left|\pi_{\theta'}(a \mid s) - \pi_\theta(a \mid s)\right| &\leq 4 \sum_{a \in \mathcal{A}} \max\left\{\pi_\theta(a \mid s), \pi_{\theta'}(a \mid s)\right\}\ \|\theta - \theta'\|_2 \\ &\leq 4 \sum_{a \in \mathcal{A}} \left\{\pi_\theta(a \mid s) + \pi_{\theta'}(a \mid s)\right\} \|\theta - \theta'\|_2 \\ &= 8\|\theta - \theta'\|_2,\end{aligned}$$

where the last step uses the fact that π_θ and $\pi_{\theta'}$ are probability distributions over the action space. Note that this inequality holds for all states s, as claimed. □

E Actor's Analysis

In this section, we analyze the mirror descent algorithm—that is, the actor in Algorithm 1. Our analysis exploits the methods in the paper [Agarwal et al., 2020b], with some small changes to accommodate our framework; in particular, while our analysis assumes no error in the critic's evaluation, it does involve a sequence of time-varying MDPs.

Given a sequence of MDPs $\{M_t\}_{t=1}^T$, let V_t^π be the value function associated with policy π on MDP M_t. Given the initialization $\theta_1 = 0$, let $\{\theta_t\}_{t=1}^T$ be parameter sequence generated by the actor, and let $\pi_t = \pi_{\theta_t}$ be the policy associated with parameter θ_t. For each t, there is a sequence $w_t = \{w_{ht}\}_{h=1}^H$ such that $\|w_{ht}\|_2 \leq \rho_h^w$ for all $h \in [H]$, and

$$Q^{\pi_t}_{h,M_t}(s,a) \overset{def}{=} \langle \phi_h(s,a),\, w_{ht} \rangle\,, \qquad \text{for all } (s,a) \text{ and } h \in [H]. \tag{67a}$$

In particular, the value of w_{ht} is the value $\underline{w}_{ht}$ identified by the critic (see Eq. (37)) corresponding to policy π_t, so that $Q^{\pi_t}_{M_t} = \underline{Q}^{\pi_t}$. Define the value function $V^{\pi_t}_{h,M_t}(s) = \mathbb{E}_{A' \sim \pi_t}\left[Q^{\pi_t}_{h,M_t}(s,A')\right]$ along with the advantage function

$$G^{\pi_t}_{h,M_t}(s,a) \overset{def}{=} Q^{\pi_t}_{h,M_t}(s,a) - V^{\pi_t}_{h,M_t}(s). \tag{67b}$$

Proposition 6 (Actor's Analysis)**.** *Suppose that the actor takes $T \geq \log|\mathcal{A}|$ steps using a stepsize $\eta \lesssim 1$, and the advantage function at each iteration t is uniformly bounded as $|G^{\pi_t}_{h,M_t}(s,a)| \leq 2$ for all (s,a). Then for any fixed policy π, we have*

$$\frac{1}{T}\sum_{t=1}^T \left\{V^\pi_{1,M_t}(s_1) - V^{\pi_t}_{1,M_t}(s_1)\right\} \leq H\left[\frac{\log|\mathcal{A}|}{\eta T} + \eta\right]. \tag{68a}$$

In particular, setting $\eta \simeq \sqrt{\frac{\log|\mathcal{A}|}{T}}$ yields the bound

$$\frac{1}{T}\sum_{t=1}^T \left\{V^\pi_{1,M_t}(s_1) - V^{\pi_t}_{1,M_t}(s_1)\right\} \leq \underbrace{2H\sqrt{\frac{\log|\mathcal{A}|}{T}}}_{=\mathcal{C}(T)}. \tag{68b}$$

E.1 Proof of Proposition 6

In order to prove this claim, we require an auxiliary result that re-expresses the mirror update rule. Given the Q-value function $Q(s,a) \overset{def}{=} \langle \phi(s,a),\, w\rangle$, consider the linear update $\theta^+ \overset{def}{=} \theta + \eta w$, and the induced soft-max policy π_{θ^+}. The following auxiliary result extracts a useful property of this update:

Lemma 11 (Update in Natural Policy Gradient)**.** *For any function $F : \mathcal{S} \to \mathbb{R}$, we have*

$$Q(s,a) - F(s) = \frac{1}{\eta}\left[\log\frac{\pi_{\theta^+}(s,a)}{\pi_\theta(s,a)} + \log\left(\sum_{a' \in \mathcal{A}} \pi_\theta(s,a') e^{\eta\left(Q(s,a') - F(s)\right)}\right)\right], \tag{69}$$

valid for all pairs (s,a).

See Section E.2 for the proof of this claim.

Turning to the proof of the proposition, we have

$$V^\pi_{1,M_t}(s_1) - V^{\pi_t}_{1,M_t}(s_1) \overset{(i)}{=} \sum_{h=1}^H \mathbb{E}_{(S_h,A_h)\sim\pi}\left[G^{\pi_t}_{h,M_t}(S_h,A_h)\right] \overset{(ii)}{=} \frac{1}{\eta}\sum_{h=1}^H X_{h,t} \tag{70a}$$

where we have introduced the shorthand

$$X_{h,t} \overset{def}{=} \mathbb{E}_{(S_h,A_h)\sim\pi}\left[\log\frac{\pi_{\theta_{t+1}}(S_h,A_h)}{\pi_{\theta_t}(S_h,A_h)} + \log\left(\mathbb{E}_{A'_h \sim \pi_t(\cdot|S_h)}\left[e^{\eta G^{\pi_t}_{h,M_t}(S_h,A'_h)}\right]\right)\right]. \tag{70b}$$

Here step (i) follows from the simulation lemma (e.g., [Kakade et al., 2003]), and step (ii) makes use of Lemma 11 with $F(s) = V^{\pi_t}_{h,M_t}(s)$, along with the definition of the advantage function—namely, $G^{\pi_t}_{h,M_t}(s,a) = Q^{\pi_t}_{h,M_t}(s,a) - V^{\pi_t}_{h,M_t}(s)$.

For each $h \in [H]$ and $t \in [T]$, we now bound the two terms within the definition (70b) of $X_{h,t}$ separately. In particular, we derive a telescoping relationship for the first term, and a uniform bound on the second term.

First term: For any pair of policies $\pi, \widetilde{\pi}$ and s, we introduce the shorthand

$$D_s(\pi; \widetilde{\pi}) \overset{def}{=} KL\left(\pi(\cdot \mid s) \| \widetilde{\pi}(\cdot \mid s)\right).$$

From the definition of KL divergence, for each s_h, we have

$$\begin{aligned}\sum_{a_h \in \mathcal{A}} \pi(a_h \mid s_h) \log \frac{\pi_{t+1}(s_h,a_h)}{\pi_t(s_h,a_h)} &= \sum_{a_h} \pi(a_h \mid s_h) \Big[\log \frac{\pi_{t+1}(s_h,a_h)}{\pi(s_h,a_h)} - \log \frac{\pi_t(s_h,a_h)}{\pi(s_h,a_h)} \Big] \\ &= -D_{s_h}(\pi; \pi_{t+1}) + D_{s_h}(\pi; \pi_t). \qquad (71a)\end{aligned}$$

Second term: We begin with the elementary inequality $e^x \leq 1 + x + x^2$ valid for all $x \in [0,1]$. By assumption, we have $|\eta G^{\pi_t}_{h,M_t}(s,a)| \leq 2\eta \leq 1$ for any pair (s,a), and hence

$$e^{\eta G^{\pi_t}_{h,M_t}(s,a)} \leq 1 + \Big(\eta G^{\pi_t}_{h,M_t}(s,a)\Big) + \Big(\eta G^{\pi_t}_{h,M_t}(s,a)\Big)^2 \leq 1 + \Big(\eta G^{\pi_t}_{h,M_t}(s,a)\Big) + 4\eta^2.$$

By definition of the advantage function, we have $\mathbb{E}_{A'_h \sim \pi_t}\Big[G^{\pi_t}_{h,M_t}(s_h, A'_h)\Big] = 0$, so that we have

$$\log\Big(\mathbb{E}_{A'_h \sim \pi_t} e^{\eta G^{\pi_t}_{h,M_t}(s_h,A'_h)}\Big) \leq \log\left(1 + 4\eta^2\right) \;\leq\; 4\eta^2. \qquad (71b)$$

Combining the pieces: Combining the bounds (71a) and (71b) yields

$$\frac{1}{\eta} X_{h,t} \leq \frac{1}{\eta} \mathbb{E}_{(S_h) \sim \pi}\left[-D_{S_h}(\pi; \pi_{t+1}) + D_{S_h}(\pi; \pi_t)\right] + 4\eta.$$

Averaging this bound over all $t \in [T]$ and exploiting the telescoping of the terms yields

$$\begin{aligned}\frac{1}{\eta T} \sum_{t=1}^{T} X_{h,t} &\leq \frac{1}{\eta T} \mathbb{E}_{S_h \sim \pi}\left[-D_{S_h}(\pi; \pi_{t+1}) + D_{S_h}(\pi; \pi_1)\right] + 4\eta \\ &\overset{(i)}{\leq} \frac{1}{\eta T} \mathbb{E}_{(S_h) \sim \pi} D_{S_h}(\pi; \pi_1) + 4\eta \\ &\overset{(ii)}{\leq} \frac{1}{\eta T} \log(|\mathcal{A}|) + 4\eta,\end{aligned}$$

where step (i) follows by non-negativity of the KL divergence; and step (ii) uses the fact that the KL divergence is at most $\log(|\mathcal{A}|)$. Summing these bounds over $h \in [H]$ yields

$$\frac{1}{T} \sum_{t=1}^{T} \left\{ V^{\pi}_{1,M_t}(s_1) - V^{\pi_t}_{1,M_t}(s_1) \right\} = \frac{1}{\eta T} \sum_{t=1}^{T} \sum_{h=1}^{H} X_{h,t} \;\leq\; H \left\{ \frac{1}{\eta T} \log(|\mathcal{A}|) + 4\eta \right\},$$

thereby establishing the claim (68a).

Finally, the bound (68b) follows by making the particular stepsize choice $\eta \simeq \sqrt{\frac{\log |\mathcal{A}|}{T}}$. Note that the assumed lower bound $T \geq \log |\mathcal{A}|$ ensures that $\eta \lesssim 1$, as required to apply the bound (68a).

E.2 Proof of Lemma 11

By definition of the soft-max policy, we have $\pi_{\theta^+}(s,a) = \frac{\exp(\langle \phi(s,a),\, \theta^+ \rangle)}{\sum_{a' \in \mathcal{A}} e^{\langle \phi(s,a'),\, \theta^+ \rangle}}$. Since $\theta_+ = \theta + \eta w$, we can write

$$\begin{aligned}
\pi_{\theta^+}(s,a) = \frac{e^{\langle \phi(s,a),\, \theta + \eta w \rangle}}{\sum_{a' \in \mathcal{A}} e^{\langle \phi(s,a'),\, \theta + \eta w \rangle}} &= \frac{e^{\langle \phi(s,a),\, \theta \rangle} e^{\eta \langle \phi(s,a),\, w \rangle}}{\sum_{a' \in \mathcal{A}} e^{\langle \phi(s,a'),\, \theta \rangle} e^{\eta \langle \phi(s,a'),\, w \rangle}} \\
&= \frac{e^{\langle \phi(s,a),\, \theta \rangle}}{\sum_{\tilde{a} \in \mathcal{A}} e^{\langle \phi(s,\tilde{a}),\, \theta \rangle}} \times \frac{e^{\eta \langle \phi(s,a),\, w \rangle}}{\sum_{a' \in \mathcal{A}} \frac{e^{\langle \phi(s,a'),\, \theta \rangle}}{\sum_{\tilde{a} \in \mathcal{A}} e^{\langle \phi(s,\tilde{a}),\, \theta \rangle}} e^{\eta \langle \phi(s,a'),\, w \rangle}} \\
&= \pi_\theta(s,a) \times \frac{e^{\eta \langle \phi(s,a),\, w \rangle}}{\sum_{a' \in \mathcal{A}} \pi_\theta(s,a') e^{\eta \langle \phi(s,a'),\, w \rangle}} \\
&= \pi_\theta(s,a) \times \frac{e^{\eta Q(s,a)}}{\sum_{a' \in \mathcal{A}} \pi_\theta(s,a') e^{\eta Q(s,a')}}
\end{aligned}$$

where the last step uses the definition of Q. Multiplying both sides by $e^{-F(s)}$ and re-arranging yields

$$\frac{\pi_{\theta^+}(s,a)}{\pi_\theta(s,a)} \sum_{a' \in \mathcal{A}} \pi_\theta(s,a') e^{\eta [Q(s,a') - F(s)]} = e^{\eta [Q(s,a) - F(s)]},$$

which is equivalent to the claim.

F Lower Bound

In this section, we prove the lower bound stated in Theorem 2. We first describe the MDP class in Appendix F.1 (*MDP Class*) which consists of a sequence of linear bandits. Then in Appendix F.2 (*Policy disagreement and suboptimality gap*) we derive a useful result on the distance of policies; this will be useful for the applications of Assouad's method to derive the lower bound. We describe the interaction process in Appendix F.3 (*Interaction Process*) and finally we provide the proof for the lower bound in Appendix F.4 (*Proof of Theorem 2*).

F.1 MDP Class

Fix the horizon H and the total dimension $d_{\text{tot}} = \sum_{h=1}^{H} d_h$ across all time steps. While our result will be presented for $d = d_1 = \cdots = d_h$, we conduct the analysis with different feature dimension $d_1, \dots, d_h$. The MDP class that we introduce is parameterized by Boolean vectors $u \in \{-1,+1\}^{d_{\text{tot}}}$.

We describe the MDP class $\mathcal{M} = \{M_u \mid u \in \{-1,+1\}^{\}}$ by describing each MDP M_u. For notational convenience we identify an index of the vector by hi, i.e., entry $[u]_{hi} \in \{-1,+1\}$ where the first index $h \in [H]$ represents the horizon and for a fixed h the second index $i \in [d_h]$ represents a direction in $\mathbb{R}^{d_h}$.

State space: There is only one state—viz. $\mathcal{S} = \{s\}$.

Action space: For every time step h and state s, we have the action space $\mathcal{A} = \{-1, 0, +1\}^{d_h}$. Note that the action space has cardinaltiy 3^{d_h}.

Feature extractor: The feature map simply returns the action selected (with a rescaling factor) chained with a bias term at the end, i.e.,

$$\phi_h(s,a) = \left[\frac{a}{\sqrt{2d_h}}, \frac{1}{\sqrt{2}}\right] \in \mathbb{R}^{d_h+1}, \qquad \text{for all triples } (s,a,h). \tag{72}$$

Notice that by construction, for any pair (s,a), we have $\|\phi(s,a)\|_2 = \sqrt{\frac{\|a\|_2^2}{2d_h} + \frac{1}{2}} \le 1$.

Transition function: Since there is a single state, the transition is deterministic into the same state.

Reward function: The reward function is linear with additive Gaussian noise, and it distinguishes different MDPs. It depends on vector u_h (a subset of the entries of u) through a scaling factor δ_h to be determined later.

$$R_h(s,a) = \frac{a^\top}{\sqrt{2d_h}}(\delta u_h) + \mathcal{N}(0,1), \qquad\qquad u_h = [u_{h1}, \dots, u_{hd_h}] \in \mathbb{R}^{d_h}. \tag{73}$$

Verifying the Low-Rank Assumption It is easy to see that the MDP satisfies Assumption 4 (*Low-Rank MDP*). The transition matrix is the 1×1 identity and the reward function is by definition linear:

$$\forall (s,a,h,s'): \qquad r_h(s,a) = \left\langle \phi_h(s,a), \underbrace{[\delta u_h, 0]}_{w_h^R} \right\rangle$$

$$\mathbb{P}_h(s,a) = \langle \phi_h(s,a), \psi_h \rangle = [1], \quad \text{where } \psi_h \overset{def}{=} \begin{bmatrix} 0 & 0 & \cdots & \sqrt{2} \end{bmatrix}.$$

We will later set $\delta \lessapprox 1/\sqrt{n_h}$; therefore, for n_h large enough the regularity condition on the reward parameter and on the action value function boundness are satisfied.

F.2 Policy disagreement and suboptimality gap

In order to evaluate the agent's performance, we first need to find a way to measure the distance between policies. Doing so will be useful to reduce the policy learning problem to hypothesis testing where standard statistical methods can be applied.

Let $\pi^\star_u$ be the optimal policy on M_u and let $V^\star_u$ the optimal value function. We assess the quality of the learned policy via the cumulative sign disagreement

$$\rho(\pi, \pi') \overset{def}{=} \sum_{h=1}^{H} \sum_{i=1}^{d_h} \mathbb{1}\{\text{sign}[\mathbb{E}_{A\sim\pi_h} \langle A,\, e_i\rangle] \neq \text{sign}[\mathbb{E}_{A\sim\pi'_h} \langle A,\, e_i\rangle]\}. \tag{74}$$

Clearly, for any π, π' we have $\rho(\pi, \pi) = 0$, $\rho(\pi, \pi') \geq 0$ and $\rho(\pi, \pi') = \rho(\pi', \pi)$, so ρ is at least a seminorm.

Suboptimality gap as function of policy disagreement: Let $u \in \mathcal{U}$ and consider the MDP M_u as described before. Since the optimal action at timestep h on M_u is u_h, by inspection, the associated suboptimality of π on M_u compared to the optimal policy on M_u is

$$\begin{aligned}
V^\star_u - V^\pi_u &= \sum_{h=1}^{H} \frac{1}{\sqrt{2d_h}} \Big[u_h^\top (\delta_h u_h) - \mathbb{E}_{a\sim\pi_h} a^\top (\delta_h u_h) \Big] \\
&= \sum_{h=1}^{H} \sum_{i=1}^{d_h} \frac{\delta_h}{\sqrt{2d_h}} \Big[[u]_{hi}[u]_{hi} - [\mathbb{E}_{a\sim\pi_h} a]_i [u]_{hi} \Big] \\
&= \sum_{h=1}^{H} \sum_{i=1}^{d_h} \frac{\delta_h}{\sqrt{2d_h}} \Big([u]_{hi} - [\mathbb{E}_{a\sim\pi_h} a]_i \Big) [u]_{hi} \\
&= \sum_{h=1}^{H} \sum_{i=1}^{d_h} \frac{\delta_h}{\sqrt{2d_h}} \Big| [u]_{hi} - [\mathbb{E}_{a\sim\pi_h} a]_i \Big| \\
&\geq \sum_{h=1}^{H} \sum_{i=1}^{d_h} \frac{\delta_h}{\sqrt{2d_h}} \Big| [u]_{hi} - [\mathbb{E}_{a\sim\pi_h} a]_i \Big| \mathbb{1}\{\text{sign}(\mathbb{E}_{a\sim\pi_h} a^\top e_i) \neq [u]_{hi}\} \\
&\geq \sum_{h=1}^{H} \sum_{i=1}^{d_h} \frac{\delta_h}{\sqrt{2d_h}} \mathbb{1}\{\text{sign}(\mathbb{E}_{a\sim\pi_h} a^\top e_i) \neq [u]_{hi}\}.
\end{aligned} \tag{75}$$

This is useful as it will allow use to satisfy the assumptions of Assouad's method.

F.3 Interaction Process

MJW COMMENT: Please see my concern in the Slack channel about this restriction of the data collection process in step (c), which is really something that should be under control of the estimator.

Consider the following process:

(a) An algorithm π is selected together with a sampling budget $n_1, \dots, n_h$ where in particular n_h is the number of samples to be allocated to timestep h. For simplicity, let n_h be a multiple of d_h.

(b) The environment selects vector $u \in \mathcal{U}$ (and thus the MDP M_u)

(c) The dataset $\mathcal{D}$ is generated by playing each action in $\{e_1, \dots, e_{d_h}, \vec{0}\}$ exactly n_h/d_h times. (Note that playing the null action does affect the covariance matrix because of the bias term and simplifies the computations later).

(d) The algorithm is required to return a time-dependent stochastic policy using the dataset $\mathcal{D}$.

In particular, $\mathcal{D} = (\mathcal{D}_1, \dots, \mathcal{D}_h)$ where $\mathcal{D}_h$ contains all the information pertaining to timestep h: $\mathcal{D}_h = \{(s, , r_{hk}, s)\}_{k=1}$, where r_{hk} are the sampled rewards with distribution indicated in Eq. (73) and the action follows the process just described.

F.4 Proof of Theorem 2

For this proof, we consider the MDP class described in Appendix F.1 (*MDP Class*) along with the interaction process described in Appendix F.3 (*Interaction Process*). We make use of Appendix F.2 (*Policy disagreement and suboptimality gap*) when applying Assouad's method.

Let $\mathbb{Q}_u$ denote the distribution of the data when the sampling process is applied to the MDP M_u, and let $\mathbb{E}_u$ denote expectations under this distribution.

If we define $[u_{\text{ALG}}]_{hi} = \text{sign}(\mathbb{E}_{A \sim \pi_{\text{ALG},h}} \langle A,\, e_i \rangle)$ and let $d_1 = \cdots = d_h \overset{def}{=} d$, then equation (75) ensures that

$$\inf_{\text{ALG}} \sup_{u \in \mathcal{U}} \mathbb{E}_u[V^\star_u - V^{\pi_{\text{ALG}}}_u] \geq \frac{\delta}{\sqrt{2d}} \inf_{\text{ALG}} \sup_{u \in \mathcal{U}} \mathbb{E}_u \underbrace{\sum_{h=1}^{H} \sum_{i=1}^{d} \mathbb{1}\{\text{sign}(\mathbb{E}_{A \sim \pi_{\text{ALG},h}} \langle A,\, e_i \rangle) \neq [u]_{hi}\}}_{(u_{\text{ALG}},u)}$$

MJW COMMENT: There was a very unfortunate notation class between H for horizon and H for Hamming distance. Another reason for macros...fixed now where (v, u) is the Hamming distance between the binary sequences v and u. Using Assouad's method (cf. Lemma 2.12 in the book [Tsybakov, 2009]), we continue the lower bound above to obtain

$$\geq \frac{\delta}{\sqrt{2d}} \frac{dH}{2} n_{u,u':(u,u')=1} \inf_{\psi} \Big[\mathbb{P}_u(\psi \neq 0) + \mathbb{P}_{u'}(\psi \neq 1)\Big] \tag{76}$$

where $\inf_\psi$ denotes the minimum over all test functions taking values in $\{-1, +1\}$. A further lower bound that uses the KL divergence is given by Theorem 2.12 in [Tsybakov, 2009]

It remains to bound the the Kullback-Leibler divergence of the distributions $\mathbb{Q}_u$ and $\mathbb{Q}_{u'}$ that generate the samples in the dataset where u and u' only differ in one coordinate.

Note that the only stochasticity in the dataset lies in the rewards. For any given u, equation (73) implies that the distribution over rewards has the product form

$$\mathbb{Q}_u = \prod_{h=1}^{H} \prod_{i=1}^{d} \prod_{j=1}^{\frac{n_h}{d}} \mathcal{N}\left(\frac{e_i^\top}{\sqrt{2d_h}}(\delta u_h), 1\right). \tag{77}$$

Notice that each normal distribution in the above display for $\mathbb{Q}_u$ is identical to the corresponding factor in $\mathbb{Q}_{u'}$ except for the single index in which the vectors u and u' differ. Thus, applying the chain rule for KL divergence yields

$$\begin{aligned} D_{\text{KL}}(\mathbb{Q}_u \| \mathbb{Q}_{u'}) &= \sum_{k=1}^{\frac{n_h}{d}} D_{\text{KL}}(\mathcal{N}\left(\frac{\delta}{\sqrt{2d}}, 1\right) \| \mathcal{N}\left(\frac{-\delta}{\sqrt{2d}}, 1\right)) \\ &= \frac{n_h}{2d}\left(2\frac{\delta}{\sqrt{2d}}\right)^2 \\ &= \frac{n_h \delta^2}{d^2}, \end{aligned}$$

valid for any pair u, u' differing in a single coordinate.

Plugging back, we obtain

$$\inf_{\text{ALG}} \sup_{u \in \mathcal{U}} \mathbb{E}_u[V^\star_u - V^{\pi_{\text{ALG}}}_u] \geq \frac{\delta}{\sqrt{2d}} \frac{dH}{2}\left(1 - \sqrt{\frac{1}{2}\frac{n_h \delta^2}{d^2}}\right).$$

If we set $\delta = \frac{d}{\sqrt{2n_h}}$ the bound becomes

$$\begin{aligned}\inf_{\textsc{Alg}} \sup_{u \in \mathcal{U}} \mathbb{E}_u[V_u^\star - V_u^{\pi_{\textsc{Alg}}}] &\geq \frac{\delta}{\sqrt{2d}} \frac{dH}{2}\left(1 - \frac{1}{2}\right) \\ &= \frac{\sqrt{d}H\delta}{4\sqrt{2}} \\ &= \frac{dH}{8}\sqrt{\frac{d}{n_h}}.\end{aligned}$$

We can recast this as

$$\sup_{\textsc{Alg}} \inf_{u \in \mathcal{U}} \mathbb{E}_u[V_u^{\pi_{\textsc{Alg}}} - V_u^\star] = -\inf_{\textsc{Alg}} \sup_{u \in \mathcal{U}} \mathbb{E}_u[V_u^\star - V_u^{\pi_{\textsc{Alg}}}] \leq -\frac{dH}{8}\sqrt{\frac{d}{n_h}}$$

and so

$$\sup_{\textsc{Alg}} \inf_{u \in \mathcal{U}} \mathbb{E}_u[V_u^{\pi_{\textsc{Alg}}}] \leq V_u^\star - \frac{dH}{8}\sqrt{\frac{d}{n_h}}$$

where the u on the right hand side is intended to be the same u that appears on the left hand side. This is the first part of the statement in Theorem 2 (*Information-theoretic limit*).

To prove the second part of the statement, we start from the right hand side in the theorem's statement and lower bound it to obtain the middle term. Consider any $u \in \mathcal{U}$; the idea is to start from

$$\sup_\pi \left[V_{1u}^\pi(s_1) - \frac{\Omega(1)}{\log\left(\frac{1}{\delta}, K, \lambda\right)} \times \mathcal{U}(\pi)\right] \geq V_{1u}^\star(s_1) - \frac{\Omega(1)}{\log\left(\frac{1}{\delta}, K, \lambda\right)} \times \mathcal{U}(\pi^\star)$$

where the uncertainty function U was defined in Eq. (11). As mentioned previously, here we use $\lambda = 1$. For the rest of the proof, we upper bound the uncertainty function:

$$\mathcal{U}(\pi^\star) \leq \sup_\pi \mathcal{U}(\pi) = \sqrt{\alpha}\sum_{h=1}^H \sup_\pi \|\phi_h^\pi\|_{\Sigma_h^{-1}}.$$

Now denote with $[x]_{1:p}$ the first p components of the vector x. Using the triangle inequality we can write

$$\|\phi_h^\pi\|_{\Sigma_h^{-1}} \leq \|\left[[\phi_h^\pi]_{1:d}, 0\right]\|_{\Sigma_h^{-1}} + \|\left[0, [\phi_h^\pi]_{d+1}\right]\|_{\Sigma_h^{-1}}.$$

Next, we use a technical lemma to compute the inverse of Σ_h. By construction Σ_h is an arrowhead matrix, i.e., can be written as

$$\Sigma_h = \begin{bmatrix} D & v \\ v^\top & b \end{bmatrix}$$

where we let the normalization constants inside of ϕ in Eq. (72) to be

$$\gamma = \frac{1}{\sqrt{2d_h}}, \qquad c = \frac{1}{\sqrt{2}}$$

to define $D \in \mathbb{R}^{d \times d}$ as a diagonal matrix with entries

$$[D]_{ii} = \gamma^2 \frac{n_h}{d} + \lambda$$

and $v \in \mathbb{R}^d$ is a vector with entries

$$[v]_i = \gamma c \frac{n_h}{d}$$

and $b \in \mathbb{R}$ is a scalar

$$b = c^2\left(n_h + \frac{n_h}{d}\right) + \lambda.$$

The inverse of Σ_h can then be computed explicitly using known formulas for block matrices or arrowhead matrices. We arrive to

$$\Sigma_h^{-1} = \begin{bmatrix} D' & v' \\ v'^\top & b' \end{bmatrix}$$

where we define the entries in a second. First, the inverse of the Schur complement is

$$b' \overset{def}{=} (b - v^\top D^{-1} v)^{-1} = \left(c^2 \left(n_h + \frac{n_h}{d} \right) + \lambda - \sum_{i=1}^{d} \frac{\left(\gamma c \frac{n_h}{d}\right)^2}{\gamma^2 \frac{n_h}{d} + \lambda} \right)^{-1}.$$

Our goal is to show that this is positive, which helps in simplifying the final expression. Notice that

$$\sum_{i=1}^{d} \frac{\left(\gamma c \frac{n_h}{d}\right)^2}{\gamma^2 \frac{n_h}{d} + \lambda} < \sum_{i=1}^{d} \frac{\left(\gamma c \frac{n_h}{d}\right)^2}{\gamma^2 \frac{n_h}{d}} = dc^2 \frac{n_h}{d} = c^2 n_h.$$

Thus

$$(b')^{-1} = \left(c^2 \left(n_h + \frac{n_h}{d} \right) + \lambda - \sum_{i=1}^{d} \frac{\left(\gamma c \frac{n_h}{d}\right)^2}{\gamma^2 \frac{n_h}{d} + \lambda} \right) > c^2 \frac{n_h}{d} + \lambda > 0.$$

These facts imply that the inverse of the above quantity is bounded as

$$b' < \frac{d}{c^2 n_h + d\lambda} < \frac{d}{c^2 n_h}.$$

Continuing the construction of the inverse, we obtain

$$D' = \underbrace{D^{-1}}_{\overset{def}{=} D_1'} + \underbrace{D^{-1} v b' v^\top D^{-1}}_{\overset{def}{=} D_2'}$$

Noice that D_1' is symmetric positive definite with positive diagonal elements and D_2' is also symmetric positive semidefinite:

$$0 \prec D_1' = D^{-1} = \left(\gamma^2 \frac{n_h}{d} + \lambda \right)^{-1} I \prec \frac{d}{\gamma^2 n_h} I$$

$$D_2' = \underbrace{b'}_{\geq 0} \underbrace{D^{-1} v}_{y} \underbrace{v^\top D^{-1}}_{y^\top} = b' y y^\top \succcurlyeq 0.$$

We now use the above block expressions for Σ_h^{-1} to bound

$$\|\phi_h^\pi\|_{\Sigma_h^{-1}} \leq \| \left[[\phi_h^\pi]_{1:d}, 0 \right] \|_{\Sigma_h^{-1}} + \| \left[\vec{0}, [\phi_h^\pi]_{d+1} \right] \|_{\Sigma_h^{-1}}.$$

By construction, $[\phi_h^\pi]_{1:d}$ only interacts with the D' block in Σ_h^{-1}; using this and

$$\|x\|_{D'}^2 = x^\top (D_1' + D_2') x \leq \|x\|_2 \left(\|D_1'\|_2 + \|D_2'\|_2 \right) \|x\|_2$$

we can write

$$\| \left[[\phi_h^\pi]_{1:d}, 0 \right] \|_{\Sigma_h^{-1}} = \| [\phi_h^\pi]_{1:d} \|_{D'} \leq \| [\phi_h^\pi]_{1:d} \|_2 \sqrt{\|D_1'\|_2 + \|D_2'\|_2}$$

Likewise,

$$\| \left[0, [\phi_h^\pi]_{d+1} \right] \|_{\Sigma_h^{-1}} = \| [\phi_h^\pi]_{d+1} \|_{b'}.$$

We now bound all norms:

$$\| [\phi_h^\pi]_{1:d} \|_2 \leq \frac{\| \mathbb{1} \|_2}{\sqrt{2d}} \leq \frac{1}{\sqrt{2}}$$

$$\|D_1'\|_2 = \|D^{-1}\|_2 \lesssim \frac{2d^2}{n_h}$$

$$\|D_2'\|_2 \leq b' \|D^{-1}\|_2 \|v\|_2 \|v\|_2 \|D^{-1}\|_2 \lesssim \underbrace{\frac{d}{n_h}}_{b'} \underbrace{\left(\gamma \frac{n_h}{d} \right)^2 \| \mathbb{1} \|_2^2}_{\|v\|_2^2} \underbrace{\frac{d^4}{n_h^2}}_{\|D^{-1}\|_2^2} \lesssim \frac{d^2}{n_h}$$

Plugging back we conclude

$$\|\left[[\phi_h^\pi]_{1:d}, 0\right]\|_{\Sigma_h^{-1}} \lesssim \frac{d}{\sqrt{n_h}}$$

Similarly

$$\|[\phi_h^\pi]_{d+1}\|_{b'} = \sqrt{\frac{1}{\sqrt{2}} b' \frac{1}{\sqrt{2}}} \leq \sqrt{\frac{1}{2} \frac{d}{c^2 n_h}} \lesssim \frac{\sqrt{d}}{\sqrt{n_h}}.$$

Plugging back, we obtain with $\nu = 0, \lambda = 1$:

$$\sup_\pi \mathcal{U}(\pi) \lesssim \sqrt{\alpha} \sum_{h=1}^{H} \sup_\pi \|\phi_h^\pi\|_{\Sigma_h^{-1}} \lesssim \sqrt{\alpha} \frac{dH}{\sqrt{n_h}} \lesssim dH \sqrt{\frac{d}{n_h}} \times \frac{\log\left(\frac{1}{\delta}, K, \lambda\right)}{\Omega(1)}.$$

In summary, we have shown that

$$\begin{aligned} \sup_\pi \left[V_{1u}^\pi(s_1) - \frac{\Omega(1)}{\log\left(\frac{1}{\delta}, K, \lambda\right)} \times \mathcal{U}(\pi)\right] &\geq V_{1u}^\star(s_1) - \frac{\Omega(1)}{\log\left(\frac{1}{\delta}, K, \lambda\right)} \times \mathcal{U}(\pi^\star) \\ &\geq V_{1u}^\star(s_1) - \Omega(1) \times dH \sqrt{\frac{d}{n_h}}. \end{aligned}$$

i.e., we have proved the right inequality in Theorem 2 (*Information-theoretic limit*).