Provable Benefits of Actor-Critic Methods for Offline Reinforcement Learning

Andrea Zanette[*]
University of California, Berkeley
zanette@berkeley.edu

Martin J. Wainwright
University of California, Berkeley
wainwrig@berkeley.edu

Emma Brunskill
Stanford University
ebrun@stanford.edu

Abstract

Actor-critic methods are widely used in offline reinforcement learning practice, but are not so well-understood theoretically. We propose a new offline actor-critic algorithm that naturally incorporates the pessimism principle, leading to several key advantages compared to the state of the art. The algorithm can operate when the Bellman evaluation operator is closed with respect to the action value function of the actor's policies; this is a more general setting than the low-rank MDP model. Despite the added generality, the procedure is computationally tractable as it involves the solution of a sequence of second-order programs. We prove an upper bound on the suboptimality gap of the policy returned by the procedure that depends on the data coverage of any arbitrary, possibly data dependent comparator policy. The achievable guarantee is complemented with a minimax lower bound that is matching up to logarithmic factors.

1 Introduction

The problem of learning a near-optimal policy is a core challenge in reinforcement learning (RL). In many settings, it is beneficial to be able to learn a good policy using only a pre-collected set of data, without further exploration with the environment; this problem is known as *offline or batch 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 maximization bias, a naive offline algorithm can return a policy with a severely overestimated value. In order to avoid such undesirable behavior, researchers have introduced the idea of pessimism under uncertainty, and there is now a growing literature (e.g., Liu et al. (2020); Jin et al. (2020b); Buckman et al. (2020); Kumar et al. (2019); Kidambi et al. (2020); Yu et al. (2020)) on different ways in which pessimism can be incorporated. See Appendix B for additional references and discussion of this body of work.

At a high level, 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 in such a way that only those policies 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 paper, we take the latter avenue.

[*] This work was fully completed while Andrea Zanette was a PhD candidate at Stanford University. Future updates of this work will be available at https://arxiv.org/abs/2108.08812

35th Conference on Neural Information Processing Systems (NeurIPS 2021).

1.1 Overview and our contributions

Implementing pessimism with function approximation is challenging for several reasons. First, uncertainty must be estimated with particular care. On one hand, underestimating it can fail to correct the coverage problem. On the other hand, 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. Similar issues can arise when adding optimistic bonuses in the exploration. The 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 by-passed 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 on offline reinforcement learning on finding with high probability the policy with the highest performance has focused on algorithms that are either model or value-based[2] Liu et al. (2020); Jin et al. (2020b); Buckman et al. (2020); Kidambi et al. (2020); Yu et al. (2020); these often incorporate pessimism into the estimates of the policy performance. Actor-critic methods are a hybrid class of methods that mitigate some deficiencies of methods that are either purely policy or purely value-based Konda and Tsitsiklis (2000, 2003); Heess et al. (2015); Haarnoja et al. (2017, 2018); in modern RL, they are widely used in practice (e.g., Levine et al. (2020); Wu et al. (2019, 2021); Kumar et al. (2019, 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: More specifically, we study the problem of policy learning using linear function approximation in the offline setting. We assume that we are given a batch data set $\mathcal{D}$, in which each sample consists of a quadruple. The first two components are the state-action pair, corresponding to the state in which a given action was taken, and the last two components correspond to a noisy observation of the reward, and a successor state drawn from the appropriate transition function. Our theory allows for a very general dependence structure among the the state-action pairs in these samples; when the data set is ordered according to how the samples were collected (which need not be related to a trajectory), we allow the state-action pair at any given instant to depend on all past samples. This set-up allows from data collected from arbitrary policies, mixtures of policies, generative models or even in adversarial manner.

Given such a data set, our objective is to find the policy that performs best in the face of uncertainty. In particular, we need to account for the fact that the optimal policy π^* for the underlying MDP may not be well covered by the dataset $\mathcal{D}$, in which case the associated uncertainty would be prohibitive. In order to achieve this goal, we design an actor-critic procedure that iteratively optimizes a lower bound on the value of the optimal policy. Suppose that we are interested in optimizing the value function at some given initial s_1. Our strategy works as follows: for any given policy π, we construct a family $\mathcal{M}(\pi)$ of "statistically plausible" MDPs, and use them to define a simple second-order cone program. By solving this convex program, we obtain value function estimate $\underline{V}^{\pi}_{M}(s_1) = \arg\min_{M\in\mathcal{M}(\pi)} V^{\pi}_{M}(s_1)$ that—for an appropriately constructed family $\mathcal{M}(\pi)$—is guaranteed to be a lower bound on the true value function of π in the unknown MDP that generated the dataset. Given a procedure for producing such lower bounds, it is then natural to maximize these lower bounds over some family Π of policies. This combination leads to the saddle-point problem

$$\max_{\pi\in\Pi}\min_{M\in\mathcal{M}(\pi)} V^{\pi}_{M}(s_1). \tag{1}$$

Note that actor-critic methods fit naturally in this framework: the critic provides a pessimistic evaluation of any given policy π, and the actor solves the outer maximization problem over policies.

[2] Exceptions to this include importance-sampling based approaches to selecting among a finite set of policies (e.g. Mandel et al. (2014); Thomas et al. (2015, 2019)); however, such approaches have focused on operating without a Markov assumption and inherently provide much looser guarantees than the ones we and others consider for the Markov setting.

This decoupling lends itself to a computationally tractable implementation, along with an analysis of the procedure. In particular, we show that the actor's sequence of estimated policies enjoys online learning-style guarantees with respect to a sequence of pessimistic MDPs implicitly identified by the critic.

The way in which we introduce pessimism is a second key component of the algorithmic framework. In particular, in line with our previous paper Zanette et al. (2020b), we do so without enlarging the prescribed classes of functions and policies. We do so by a direct perturbation of the value functions examined by the critic; there is no addition of pessimistic bonuses or absorbing states. Since the class of value functions is not altered, this method has two main advantages. First, there are no additional model assumptions compared to the standard—that is non-pessimistic—version of the actor-critic method. Second, the complexity of the underlying classes is not increased, thereby allowing us to construct tight confidence intervals and estimation error bounds that are minimax optimal up to logarithmic factors.

The remainder of this paper is organized as follows. We begin in Section 2 with background on MDPS, and then introduce the modeling assumptions that underlie the analysis of this paper. In Section 3, we introduce the algorithm studied in this paper, namely the Pessimistic Actor Critic for Learning without Exploration (for short, PACLE) algorithm. Section 4 provides statements of our main results and discussion of their consequences, including an upper bound on the PACLE algorithm in Theorem 1, and a minimax lower bound in Theorem 2. In Section A, we provide an outline of the proof of Theorem 1, with various technical details as well as the proof of Theorem 2 deferred to the appendices. We conclude with a discussion in Section 5.

1.2 Notation

For the reader's convenience, we summarize here some notation used throughout the paper. We let $\mathcal{B}_d(r) = \{x \in \mathbb{R}^d \mid \|x\|_2 \leq r\}$ 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. For a vector $x \in \mathbb{R}^d$, we use $[x]_i$ to denote its i^{th} component. We use the $\widetilde{O}$ notation to denote an upper bound that holds up to constants and log factors in the input parameters $(\frac{1}{\delta}, d, H)$. The notation $\lesssim$ means an upper bound that holds up to a constant, with an analogous definition for $\gtrsim$.

2 Background and problem formulation

We begin by providing some background, before introducing the assumptions that underlie our problem formulation.

2.1 Markov decision processes

In this paper, we focus on finite-horizon Markov decision processes, for which we provide a very brief introduction here. See the books Puterman (1994); Bertsekas and Tsitsiklis (1996); Bertsekas (1995); Sutton and Barto (2018) for more background and detail. A finite-horizon MDP is specified by a positive integer H, and events take place over a sequence of stages indexed by the time step $h \in [H] \overset{def}{=} \{1, \ldots, H\}$. The underlying dynamics involve a state space $\mathcal{S}$, and are controlled by actions that take values in some action set $\mathcal{A}$. In this paper, we allow the state space to be arbitrary (continous or discrete), whereas our analysis applies to discrete action spaces. For each time step $h \in [H]$, there is a reward function $r_h : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$, and for every time step 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 a 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 function $\mathbb{P}_h(\cdot \mid s, a)$.

A policy π_h at stage h is a mapping from the state space $\mathcal{S}$ to the action space $\mathcal{A}$. Given a full policy $\pi = (\pi_1, \ldots, \pi_H)$, the state-action value function at time step h is given by

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

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 $V_h^\pi(s) = Q_h^\pi(s, \pi_h(s))$. For a given policy π, we define the Bellman evaluation operator

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

Under some regularity conditions Puterman (1994); Shreve and Bertsekas (1978), there always exists an optimal policy $\pi^\star$ whose value and action-value functions are defined as

$$V_h^\star(s) = V_h^{\pi^\star}(s) = \sup_\pi V_h^\pi(s), \quad \text{and} \quad Q_h^\star(s,a) = Q_h^{\pi^\star}(s,a) = \sup_\pi Q_h^\pi(s,a).$$

2.2 Assumptions on data generation

In this paper, we study a model in which we observe a dataset of the form $\mathcal{D} = \{(s_i, a_i, r_i, s_i^+)\}_{i=1}^n$, where n is the total sample size. For each $i \in [n] = \{1, 2, \ldots, n\}$, the tuple (s_i, a_i) corresponds to a state-action pair associated with some time step h_i. We let $\mathcal{F}_i$ be the σ-field generated by the samples $\{(s_j, a_j, r_j, s_j^+)\}_{j=1}^{i-1}$ that are in the "past" relative to index i. With this notation, we impose the following condition:

Assumption 1 (Data generation). *For each $i \in [n]$, the pair (s_i, a_i) is measurable with respect to $\mathcal{F}_i$. Conditionally on a given pair (s_i, a_i), the random variable r_i is drawn from a reward distribution $R_{h_i}(s_i, a_i)$ that is 1-sub-Gaussian; and the next state s_i^+ is drawn from the distribution $\mathbb{P}_{h_i}(s_i, a_i)$.*

Note that the measurability condition allows the choice of (s_i, a_i) to depend arbitrarily on any of the past data with indices $j < i$. The mild assumption allows for considerable freedom. For example, the state-action pairs may be chosen from (mixture) policies, or they can be generated by an adversarial procedure that changes the data acquisition strategy as feedback is received.

For each $h \in [H]$, we let $\mathcal{I}_h$ denote the subset of observation indices $i \in [n]$ such that $h_i = h$. These index sets define the sub-datasets $\mathcal{D}_h = \{(s_i, a_i, r_i, s_i^+), i \in \mathcal{I}_h\}$ associated with all samples that are based on state-action pairs at time step h. We define $n_h = |\mathcal{D}_h|$, so that our total sample size can be written as $n = \sum_{h=1}^H n_h$.

2.3 Policy and function classes

Next we define the policy space Π and the action value function space $\mathcal{Q}$ over which we seek solutions. Let $\phi : \mathcal{S} \times \mathcal{A} \mapsto \mathbb{R}^d$ be a d-dimensional feature mapping. We assume throughout that these feature mappings are normalized such that $\|\phi(s,a)\|_2 \leq 1$ uniformly for all (s,a)-pairs. We consider action-value functions that are linear in ϕ, and families of the form

$$\mathcal{Q}(\rho^w) \overset{def}{=} \{(s,a) \mapsto \langle \phi(s,a),\, w \rangle \mid \|w\|_2 \leq \rho^w\}, \tag{3a}$$

where $\rho^w \in (0, 1]$ is a user-defined radius. For policies, we consider the associated soft-max class

$$\Pi_{soft}(\rho^\theta) \overset{def}{=} \left\{ \frac{e^{\langle \phi(s,a),\, \theta \rangle}}{\sum\limits_{a' \in \mathcal{A}} e^{\langle \phi(s,a'),\, \theta \rangle}} \;\mid\; \|\theta\|_2 \leq \rho^\theta \right\}, \tag{3b}$$

where $\rho^\theta > 0$ is a second radius.

In the context of our actor-critic algorithm, the weight radius ρ^w remains fixed for all updates. On the other hand, the actor produces a sequence of soft-max radii $\{\rho_t^\theta\}_{t=1}^T$, indexed by the iterations t of the actor. This sequence is produced via the update rule in Line 5 of Algorithm 1. The policy radius can be large $\rho^\theta \gg 1$ but we constrain $\rho^w \leq 1$ so that the critic's estimate $Q_w(s,a) = \langle \phi(s,a),\, w \rangle$ is bounded by one, i.e., $\sup_{(s,a,w)} |Q_w(s,a)| \leq 1$.

Recall that our MDP consists of sequence of H distinct stages. Our algorithm and theory allows for the possibility of different feature extractors at each step $h \in [H]$, even with possibly different dimensions. Consequently, in implementing and analyzing the algorithm, there are actually H (possibly different) functional spaces $\{\mathcal{Q}_h\}_{h=1}^H$, along with the associated soft-max policy classes $\{\Pi_h\}_{h=1}^H$. So as to simplify notation, we drop the dependence on the radii when referring to the functional spaces, and implicitly assume that the terminal value function is zero.

2.4 A range of function class assumptions

In this section, we discuss a range of assumptions that might be imposed on the class of action-value functions. This discussion serves as motivation for the particular assumption (Bellman restricted closedness—cf. Assumption 3) that underlies our analysis.

We begin with the least restrictive condition, which is a very natural starting point in our given set-up. 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.

Assumption 2 (Linear action-value functions Q^π)**.** *The MDP admits a linear action-value function representation for all policies in* Π*, meaning that for each policy* $\pi \in \Pi$ *and time step* $h \in [H]$*, there exists a vector* w^π_h *such that*

$$Q^\pi_h(s,a) = \langle \phi_h(s,a),\, w^\pi_h \rangle \,. \tag{4}$$

This assumption alone turns out to be inadequate to ensure that effective learning is possible; indeed, the recent papers Zanette (2020); Weisz et al. (2020) establish that even under this condition, there are instances that require exponentially many samples to do better than a random policy.

Given this fact, if one is interested in procedures with polynomial complexity (in both sample size and running time), stronger conditions need to be imposed. In general, the Bellman evaluation operator, even when applied to a linear action-value function, will return a nonlinear value function. The analysis of this paper is based on bounding the Bellman error in the sense of sup-norm deviation from linearity:

Assumption 3 (Bellman 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 non-negative sequence* $\{\nu_h\}_{h=1}^H$ *such that for each* $h \in [H]$*, we have*

$$\sup_{\substack{Q_{h+1} \in \mathcal{Q}_{h+1} \\ \pi_{h+1} \in \Pi_{h+1}}} \inf_{Q_h \in \mathcal{Q}_h} \| Q_h - \mathcal{T}^{\pi_{h+1}}_h Q_{h+1} \|_\infty \le \nu_h. \tag{5}$$

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 least-squares policy evaluation (e.g., Nedić and Bertsekas (2003)), which will be our starting point when constructing the critic.

Finally, for understanding connections to past work, it is relevant to compare to the *low-rank MDP* assumption that has been analyzed in recent work Jin et al. (2020a); Yang and Wang (2020), including in offline RL with pessimismistic guarantees Jin et al. (2020b), as well as in various online settings Agarwal et al. (2020a); Modi et al. (2021); Zanette et al. (2020a).

Assumption 4 (Low-Rank MDP)**.** *An MDP is low-rank if for all* $h \in [H]$*, there exists a reward parameter* $w_h \in \mathbb{R}^d$ *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*

$$r_h(s,a) = \langle \phi_h(s,a),\, w_h \rangle \,, \qquad \mathbb{P}_h(s' \mid s,a) = \langle \phi_h(s,a),\, \psi_h(s') \rangle \,, \qquad \forall (s,a,h,s'). \tag{6}$$

The following proposition explicates the nested relationship between these three conditions, showing that the low-rank MDP condition is the most restrictive:

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 that satisfy Bellman restricted closedness.

(b) The class of MDPs that satisfy Bellman restricted closedness is a strict subset of the linear Q^π *MDP class.*

See Appendix C for the proof of this claim.

Based on Proposition 1, we see that any analysis based on assuming Bellman restricted closedness also *a fortiori* applies to MDPs that satisfy the more stringent low-rank MDP condition.

3 The Pessimistic Actor-Critic

Given the set-up thus far, we are now ready to describe the actor-critic algorithm that we analyze in this paper. We refer to it as the *Pessimistic Actor Critic for Learning without Exploration*, or PACLE for short. We first describe the critic in Section 3.1, and then the actor in Section 3.2. We summarize the actor and critic algorithms, respectively, in pseudocode form in Algorithm 1 and Algorithm 2.

3.1 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) is not desirable in this setting, as the policy or distribution that generated the dataset might be unknown and estimation errors on the distribution can accumulate exponentially with the horizon in IS estimators (see e.g. Liu et al. (2018b)). Instead, we use a least-squares temporal difference method for policy evaluation, but suitably perturbed to return pessimistic estimates—i.e., lower bounds on the true value function of the given policy π. 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.

Let us now give a precise description of the critic. Given a policy $\pi = (\pi_1, \dots, \pi_H)$, the goal of the critic is to minimize the quantity

$$\mathbb{E}_{A' \sim \pi_1} \langle \phi(s_1, A'),\, w_1 \rangle = \sum_{a \in \mathcal{A}} \pi_1(a \mid s_1) \langle \phi_1(s_1, a),\, w_1 \rangle\,, \tag{7}$$

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. Recall that $\mathcal{I}_h$ indexes the subset of observations associated with state-action pairs at time step h. For each $h \in [H]$ and $i \in \mathcal{I}_h$, let us write the associated sample as the quadruple $(s_{hi}, a_{hi}, r_{hi}, s_{h+1,i})$. Introducing the shorthand notation $\phi_{hi} = \phi_h(s_{hi}, a_{hi})$, we define the *cumulative covariance matrix*

$$\Sigma_h \overset{def}{=} \Big(\sum_{i \in \mathcal{I}_h} \phi_{hi} \phi_{hi}^{\top} \Big) + I_{d \times d}, \tag{8}$$

where $I_{d \times d}$ denotes the d-dimensional identity matrix. Notice that the cumulative covariance grows as the number of samples in $\mathcal{I}_h$ increases; we do not normalize it by the local sample size $n_h = |\mathcal{I}_h|$, so that Σ_h effectively represents the amount of information contained in the sub-dataset $\mathcal{D}_h$ at time step h.

Since Σ_h is strictly positive definite by construction, it defines a pair of norms

$$\|u\|_{\Sigma_h} \overset{def}{=} \sqrt{u^{\top} \Sigma_h u}, \quad \text{and} \quad \|u\|_{\Sigma_h^{-1}} \overset{def}{=} \sqrt{u^{\top} (\Sigma_h)^{-1} u}. \tag{9}$$

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

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

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 $\|\xi_h\|_{\Sigma_h} \leq \alpha_h$, 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 $\alpha = (\alpha_1, \ldots, \alpha_H)$, weight radii $\rho^w = (\rho^w_1, \ldots, \rho^w_H)$ with each $\rho^w_h \in (0, 1]$. The optimization variables consist of the regression vectors $w = (w_1, \ldots, w_H) \in (\mathbb{R}^d)^H$ and the pessimism vectors $\xi = (\xi_1, \ldots, \xi_H) \in (\mathbb{R}^d)^H$. The critic then solves the convex program

$$(\xi^\pi, \underline{w}^\pi) \overset{def}{=} \arg \min_{\substack{\xi \in (\mathbb{R}^d)^H \\ w \in (\mathbb{R}^d)^H}} \sum_{a \in \mathcal{A}} \pi_1(a \mid s_1) \langle \phi_1(s_1, a), w_1 \rangle \tag{10a}$$

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

$$w_h = \xi_h + \Sigma_h^{-1} \sum_{k \in \mathcal{I}_h} \phi_{hk} \Big[r_{hk} + \sum_{a \in \mathcal{A}} \pi_{h+1}(a \mid s_{h+1,k}) \langle \phi_{h+1}(s_{h+1,k}, a), w_{h+1} \rangle \Big], \qquad \text{and} \tag{10b}$$

$$\|\xi_h\|^2_{\Sigma_h} \leq \alpha_h^2, \qquad \|w_h\|_2^2 \leq (\rho^w_h)^2 \tag{10c}$$

for all $h \in [H]$. Here the matrices Σ_h were previously defined in equation (8).

The convex program (10) 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.

Algorithm 1 ACTOR (MIRROR DESCENT)

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

Algorithm 2 CRITIC (PLSPE)

```
1: Input: Dataset $\mathcal{D}$, target policy $\pi$, starting state $s_1$, critic radii $\{\rho^w_h\}_{h=1,\ldots,H}$, and parameters $\{\alpha_h\}_{h=1,\ldots,H}$
2: Solve the optimization program (10)
3: Return: Optimal weight vector $\underline{w}$
```

3.2 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 $h \in [H]$, so that the soft-max policy in moving from iteration t to $t + 1$ is updated as

$$\pi_{t+1,h}(a \mid s) \propto \pi_{t,h}(a \mid s) e^{\eta Q_h(s,a)} \qquad \text{for each } (s, a) \in \mathcal{S} \times \mathcal{A}. \tag{11}$$

Here $\eta > 0$ is a stepsize parameter, and our theory specifies a suitable choice.

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 $\rho_1^w, \dots, \rho_H^w$ in a problem dependent way (cf. Assumption 3). Finally, for each h, define the vector $\bar{\phi}_h^{\pi} \overset{def}{=} \mathbb{E}_{(S_h, A_h)\sim\pi}[\phi_h(S_h, A_h)]$, 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

$$\mathcal{U}(\pi; \alpha) \overset{def}{=} 2\sum_{h=1}^{H} \alpha_h \|\bar{\phi}_h^{\pi}\|_{\Sigma_h^{-1}} = 2\sum_{h=1}^{H} \alpha_h \sqrt{(\bar{\phi}_h^{\pi})^\top \Sigma_h^{-1} \bar{\phi}_h^{\pi}}, \tag{12}$$

where the cumulative covariance matrix Σ_h was defined in equation (8).

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 $\{\alpha_h\}_{h=1}^{H}$. The more data are available, the more positive definite Σ_h is and the smaller the uncertainty function $\mathcal{U}(\pi; \alpha)$ becomes for a fixed policy π. If the sampling distribution that generates the dataset is fixed, then we can write $\mathcal{U}(\pi; \alpha) \lessapprox c/\sqrt{n}$ 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 on the data collection process. It is based on radii $\{\rho_h^w\}_{h=1}^{H}$ for the action value function[3] that lie in the interval $(0, 1]$, and it provides a guarantee relative to the class Π_{all} of all stochastic policies.

Theorem 1 (An achievable guarantee)**.** *Suppose that we are given a data set $\mathcal{D}$ collected in a way that respects Assumption 1. Then there are pessimism vectors bounded as $\alpha_h = \widetilde{O}(\sqrt{d\ \log(1/\delta)}) + \nu_h\sqrt{n_h}$ such that, after running $T \geq \log|\mathcal{A}|$ rounds of the actor with stepsize $\eta = \sqrt{\frac{\log|\mathcal{A}|}{T}}$, the* PACLE *procedure returns a policy π_{ALG} for which*

$$V_1^{\pi}(s_1) - V_1^{\pi_{\text{ALG}}}(s_1) \leq \mathcal{U}(\pi; \alpha) + \underbrace{\sum_{h=1}^{H} \nu_h}_{\mathcal{E}_{msp}(\nu)} + \underbrace{4H\sqrt{\frac{\log|\mathcal{A}|}{T}}}_{\mathcal{C}(T)} \qquad \text{uniformly over all } \pi \in \Pi_{all} \tag{13}$$

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 $\mathcal{U}(\pi; \alpha)$. Note that the optimization error $\mathcal{C}(T)$ can be reduced arbitrarily, while α (and thus $\mathcal{U}(\pi; \alpha)$) increase only logarithmically with T. As a special case, if we set $\pi = \pi^\star$ and assume that there is no mis-specification error, then we obtain that the learned policy satisfies a bound of the form

$$V_1^{\pi^\star}(s_1) - V_1^{\pi_{\text{ALG}}}(s_1) \leq \mathcal{U}(\pi^\star; \alpha) + \mathcal{C}(T) \tag{14}$$

with probability at least $1 - \delta$. Since $\mathcal{C}(T)$ is well-controlled, this guarantee is satisfied whenever the uncertainty term $\mathcal{U}(\pi^\star; \alpha)$ is small.

More generally, the guarantee (13) is significantly stronger than most prior work as PACLE competes not just with the optimal policy $\pi^\star$, but with all comparator policies simultaneously. Such comparator policies need not necessarily be in the prescribed policy class Π. To highlight the strength of this

[3] 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.

generality, suppose that the uncertainty $\mathcal{U}(\pi^\star; \alpha)$ of the optimal $\pi^\star$ is *not* small—it could in fact be infinite. In this case, the bound (14) would not be useful.

However, suppose that there exists a near-optimal policy—meaning a policy π^+ such that $V_1^{\pi^+}(s_1) \ge V_1^\star(s_1) - \epsilon$ for some small ϵ—that is well-covered by the dataset (i.e., for which $\mathcal{U}(\pi^+; \alpha) \approx 0$). In this case, Theorem 1 ensures with high probability $V_1^{\mathrm{ALG}}(s_1) \gtrsim V_1^\star(s_1) - \epsilon$. In contrast, traditional analyses that use only the optimal policy $\pi^\star$ as a comparator—as opposed to also allowing near-optimal policies—cannot return meaningful guarantees. We note also that the papers Yu et al. (2020); Liu et al. (2020); Kidambi et al. (2020) provide results of a similar flavor. These types of guarantees are also provided by some concurrent works Uehara and Sun (2021); Xie et al. (2021).

It should also be noted that Theorem 1 provides a family of results indexed by the choice of the critic's radii $\{\rho_h^w\}_{h=1}^H$. This choice is a modeling decision: 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.

4.2 A lower bound

Thus far, we have stated an upper bound on the quality of the returned policy for a given procedure. Central to this upper bound is the uncertainty function $\mathcal{U}(\pi; \alpha)$. In this section, we show that a term of this form is unavoidable for any procedure. In particular, working within the well-specified setting, we prove a lower bound in terms of the quantity $\mathcal{U}(\pi; \sqrt{d}) = \sqrt{d} \sum_{h=1}^H \|\bar{\phi}_h^\pi\|_{\Sigma_h^{-1}}$. Recalling that our choice of α scales with $\sqrt{d}$ (along with other logarithmic factors), this lower bound shows that our result is tight up to logarithmic factors.

We show that the lower bound actually holds in a setting that is easier for the learner, in the sense that (1) we restrict to low-rank MDPs, where there is no mis-specification error; and (2) the mechanism that generates the dataset is non-adaptive, and so certainly satisfies Assumption 1.

Theorem 2 (Information-theoretic lower bound)**.** *For a given horizon H and dimension d, consider a sample size $n \ge 2d^3H^3$. There is a class $\mathcal{M}$ of low-rank MDPs and a data generating procedure satisfying Assumption 1 such that for any policy $\widehat{\pi}_{\mathrm{ALG}}$, we have*

$$\sup_{M \in \mathcal{M}} \mathbb{E}_M \left[V_{1M}^{\pi}(s_1) - V_{1M}^{\widehat{\pi}_{\mathrm{ALG}}}(s_1) \right] \ge c\, \mathcal{U}(\pi; \sqrt{d}) \qquad \textit{uniformly over all } \pi \in \Pi_{\textit{all}}, \tag{15}$$

where $c > 0$ is a universal constant.

When $H = 1$ the above result gives a sample complexity lower bound for learning a near optimal policy from batch data in a linear bandit instance.

4.3 Comparison to related work

Theorem 1 automatically implies the typical bound $\mathbb{P}[V_1^{\pi_{\mathrm{ALG}}}(s_1) \ge V_1^\star(s_1) - \mathcal{U}(\pi^\star; \alpha)] \ge 1 - \delta$ 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 $V_1^{\pi_{\mathrm{ALG}}}(s_1) \gtrsim V_1^\star(s_1) - C/\sqrt{n}$ 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 $\mathbb{E}_{(S_h, A_h) \sim \pi_h^\star}[\phi(S_h, A_h)]$ of the optimal policy $\pi^\star$. As in the paper Jin et al. (2020b), such a factor can be small even when traditional concentrability coefficients are large because they depend on state-action visit ratios (see the literature in Appendix B, e.g., Chen and Jiang (2019)).

With reference to the results in the paper Jin et al. (2020b), our work provides improvements in two distinct ways. First, their upper and lower bounds exhibit a gap of the order dH, which our analysis closes. Second, our analysis holds under the more permissive Assumption 3 *(Bellman Restricted Closedness)* which includes low-rank MDPs. Of this improvement, a factor of $\sqrt{d}$ is due to the algorithm that we use, and the remainder is due to a more refined construction to certify optimality in Theorem 2. To be clear, our upper and lower bounds differ from theirs by a factor of H due to a different normalization in the value function). We also note that the result of Liu et al.Liu et al.

(2020) can be specialized to the low-rank MDP setting; however, even in this simpler setting, the results would be sub-optimal and also require additional 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 online 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. Due to space constraints, the proof outline is deferred to Appendix A.

5 Discussion

In this paper, we have developed and analyzed an actor-critic method procedure, designed for finding near-optimal policies in the offline setting. The PACLE procedure introduces pessimism into the critic's evaluation of a given policy's value function, thereby ensuring that, under suitable parameter choices and assumptions, it maintains (with high probability) a lower bound on the true value function. The actor then performs a form of mirror ascent so as to maximize the value of these lower bounds.

An important feature of our method is that it introduces pessimism via direct perturbations of the parameter vectors in a linear function approximation scheme. In this way, we avoid having to impose additional model assumptions; moreover, the pessimism does *not* substantially increase the complexity of our under value/policy classes, which allows us to provide minimax-optimal guarantees. We note that similar approaches have appeared before in the exploration setting; for example, see the recent papers Zanette et al. (2020b); Jin et al. (2021); Du et al. (2021). These methods enjoy similar advantages in terms of theoretical guarantees, but at the expense of computational tractability. In contrast, the method of this paper entails solving a low-dimensional second-order cone program, a simple class of convex programs for which there exist many polynomial-time algorithms. We enjoy this advantage due to some key differences between the offline and online settings of RL. In the offline setting, it is possible to keep the actor's update cleanly separated from the evaluation step of the critic, as we have done here; this separation underlies the computational tractability.

Our work leaves open a number of interesting questions for future work. First, it would be interesting to provide some numerical studies of the PACLE's performance, so as to understand its practical behavior relative to the theoretical guarantees provided here. Also, our analysis here has focused purely on approximation using linear basis expansions; extension to more general function classes is an important next step. Finally, it will be interesting to see to what extent these ideas can be translated to the more challenging setting of exploration.

Acknowledgements

This work was partially supported by NSF-DMS grant 2015454, NSF-IIS grant 1909365, and NSF-FODSI grant 2023505 to MJW, a Stanford Artificial Intelligence Laboratory Toyota gift to EB, and a Office of Naval Research grant DOD-ONR-N00014-18-1-2640 to MJW.

References

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

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

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.

Bertsekas, D. P. (1995). *Dynamic programming and stochastic control*, volume 1. Athena Scientific, Belmont, MA.

Bertsekas, D. P. and Tsitsiklis, J. N. (1996). *Neuro-dynamic programming*. Athena Scientific.

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

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

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

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, V. H., Lai, T. L., and Shao, Q. M. (2009). *Self-normalized processes*. Springer.

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

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

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

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.

Haarnoja, T., Tang, H., Abbeel, P., and Levine, S. (2017). Reinforcement learning with deep energy-based policies. In *International Conference on Machine Learning*, pages 1352–1361. PMLR.

Haarnoja, T., Zhou, A., Abbeel, P., and Levine, S. (2018). Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor.

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*.

Heess, N., Wayne, G., Silver, D., Lillicrap, T., Tassa, Y., and Erez, T. (2015). Learning continuous control policies by stochastic value gradients.

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

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

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

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

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

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

Konda, V. R. and Tsitsiklis, J. N. (2000). Actor-critic algorithms. In *Advances in neural information processing systems*, pages 1008–1014.

Konda, V. R. and Tsitsiklis, J. N. (2003). On actor-critic algorithms. *SIAM Jour. Opt. Control*, 42(4):1143–1166.

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

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

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

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

Liu, Y., Gottesman, O., Raghu, A., Komorowski, M., Faisal, A. A., Doshi-Velez, F., and Brunskill, E. (2018b). Representation balancing mdps for off-policy policy evaluation. *Advances in Neural Information Processing Systems*, 31:2644–2653.

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

Mandel, T., Liu, Y.-E., Levine, S., Brunskill, E., and Popovic, Z. (2014). Offline policy evaluation across representations with applications to educational games. In *AAMAS*, volume 1077.

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

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

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

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

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

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

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, 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, 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, R. S. and Barto, A. G. (2018). *Reinforcement Learning: An Introduction*. MIT Press, Cambridge, MA, 2nd edition.

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

Thomas, P., Theocharous, G., and Ghavamzadeh, M. (2015). High confidence policy improvement. In *International Conference on Machine Learning*, pages 2380–2388.

Thomas, P. S., da Silva, B. C., Barto, A. G., Giguere, S., Brun, Y., and Brunskill, E. (2019). Preventing undesirable behavior of intelligent machines. *Science*, 366(6468):999–1004.

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

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.

Uehara, M. and Sun, W. (2021). Pessimistic model-based offline rl: Pac bounds and posterior sampling under partial coverage.

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

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

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

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, T., Cheng, C.-A., Jiang, N., Mineiro, P., and Agarwal, A. (2021). Bellman-consistent pessimism for offline reinforcement learning.

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

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

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

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

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

A Proof Outline

In this section, we provide an outline of the proof of Theorem 1. The main components of the proof are guarantees for the pessimistic estimates produced by the critic, and online learning guarantees for the updates taken by the actor. These two guarantees are coupled together via the notion of an induced MDP.

The proof outline given here follows a bottom-up approach: (a) starting with the critic in Section A.1, we first introduce the notion of induced MDP that links the critic's output to the actor's input (see Section A.1.1), and then discuss how suitable choices of the pessimism parameters α allow us to guarantee that the critic underestimates the true value function (see Sections A.1.2 and A.1.3); (b) next in Section A.2, we provide online-style learning guarantees for the actor, again using the notion of induced MDP to link these guarantees back to the critic; and (c) in Section A.3, we put together the pieces to prove the theorem itself.

A.1 Critic's Analysis

Given a policy π and pessimism parameters α for which the convex program (10) is feasible, the critic returns the pair $(\underline{\xi}^\pi, \underline{w}^\pi) = \{(\underline{\xi}_h^\pi, \underline{w}_h^\pi)\}_{h=1}^H$. These weight vectors induce the estimated value functions

$$\underline{Q}_h^\pi(s,a) \overset{def}{=} \langle \phi(s,a),\, \underline{w}_h^\pi \rangle\,, \qquad \text{and} \quad \underline{V}_h^\pi(s) \overset{def}{=} \mathbb{E}_{A' \sim \pi_h(\cdot|s)} \underline{Q}_h^\pi(s, A'). \tag{16}$$

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

A.1.1 Induced MDP

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*

$$\widehat{r}_h^\pi(s,a) \overset{def}{=} r_h(s,a) + \underline{Q}_h^\pi(s,a) - \mathcal{T}_h^\pi(\underline{Q}_{h+1}^\pi)(s,a). \tag{17}$$

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

One important property of the induced MDP—which motivates the definition (17)—is that the estimates (16) returned by the critic correspond to the *exact value functions* of policy π in the induced MDP. We summarize in the following:

Lemma 1 (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*

$$\underline{Q}_h^\pi = Q_{h,\hat{M}(\pi)}^\pi, \quad \textit{and} \tag{18a}$$

$$\underline{V}_h^\pi = V_{h,\hat{M}(\pi)}^\pi \qquad \textit{for all } h \in [H], \tag{18b}$$

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.1 for the proof of this claim.

Moreover, since the induced MDP differs from the original MDP only in terms of the reward perturbation (17), 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

$$V_{1,\hat{M}(\pi)}^{\widetilde{\pi}}(s_1) - V_1^{\widetilde{\pi}}(s_1) = \sum_{h=1}^H \mathbb{E}_{(S_h,A_h)\sim\widetilde{\pi}}\Big[\widehat{r}_h^\pi(S_h,A_h) - r_h(S_h,A_h)\Big], \tag{19}$$

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

A.1.2 Critic's guarantee under a "good" event

We now show that there is a "good event"—call it $\mathcal{G}(\alpha)$—under which the critic's value function estimates have some additional desirable properties. Once this event is defined, the core of our proof involves determining the smallest choice of pessimism parameters under which it holds with probability at least $1 - \delta$.

We begin with some notation required to define the good event. Let $\mathcal{F}$ denote the space of all real-valued functions on $\mathcal{S} \times \mathcal{A}$. The *regression operator* is a mapping from $\mathcal{F}$ to $\mathbb{R}^d$, given by

$$\mathcal{R}_h^\pi(F) \overset{def}{=} \Sigma_h^{-1} \sum_{k=1}^{T} \phi_{hk} \big\{ r_{hk} + \mathbb{E}_{A' \sim \pi(\cdot | s_{hk})} F(s_{h+1,k}, A') \big\}, \tag{20a}$$

where $F \in \mathcal{F}$. To appreciate the relevance of the regression operator, note that by definition of the critic, we have the equivalence

$$\underline{w}_h^\pi = \underline{\xi}_h^\pi + \mathcal{R}_h^\pi(\underline{Q}_{h+1}^\pi). \tag{20b}$$

We also define the *sup-norm projection operator* (for the definition of $\mathcal{B}$ please see Section 1.2)

$$\mathcal{P}_h^\pi(F) \overset{def}{=} \arg \min_{w_h \in \mathcal{B}(\rho_h^w)} \sup_{(s,a)} \Big| \langle \phi(s,a),\, w_h \rangle - \left(\mathcal{T}_h^\pi F\right)(s,a) \Big|. \tag{20c}$$

Note that $\mathcal{P}_h^\pi$ 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 $\mathcal{T}_h^\pi(F)$.

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

$$\mathcal{E}_h^\pi(F) \overset{def}{=} \mathcal{R}_h^\pi(F) - \mathcal{P}_h^\pi(F). \tag{21}$$

For a given sequence $\alpha = (\alpha_1, \ldots, \alpha_H)$ of pessimism parameters, we define the *good event*

$$\mathcal{G}(\alpha) \overset{def}{=} \Big\{ \sup_{\substack{Q_{h+1} \in \mathcal{Q}_{h+1} \\ \pi_{h+1} \in \Pi_{h+1}}} \| \mathcal{E}_h^{\pi_{h+1}}(Q_{h+1}) \|_{\Sigma_h} \leq \alpha_h \quad \text{for all } h \in [H] \Big\}. \tag{22}$$

Some intuition: Why is this event relevant for guaranteeing good performance of the critic? In order to gain intuition, let us consider the special case in which there is no approximation error, so that the exact state-action value functions are actually linear. Letting w_h^π denote the parameter associated with the linear action-value function at step h, when the good event holds, our choice of α allows us to set

$$\underline{\xi}_h^\pi = -\mathcal{E}_h^{\pi_{h+1}}(Q_{h+1}^\pi) \;=\; w_h^\pi - \mathcal{R}_h^\pi(Q_{h+1}^\pi) \qquad \text{for each } h \in [H],$$

in the constraints (10b). In this way, at each step h the vector $\underline{\xi}_h^\pi$ can perfectly compensate the noise error $\mathcal{E}_h^{\pi_{h+1}}(Q_{h+1}^\pi)$ ensuring that the action-value function Q_h^π (compactly encoded in the parameter w_h^π) can be perfectly represented. In other words, our choice guarantees that the feasible set for (10) contains the 'true' solution w_h^π. Since the convex program involves minimizing over value functions, this feasibility underlies showing the critic returns an underestimate of the true value function for π along with some approximation error in the general setting; see equation (23a) below for a precise statement. We highlight that such underestimates is only guaranteed at the initial state s_1 and timestep $h = 1$ as encoded in the objective of the program in equation (10).

On the other hand, for other policies $\widetilde{\pi}$, we can use the relation (19) to control the difference between the value function $V_{1,\hat{M}(\pi)}^{\widetilde{\pi}}$ in the induced MDP, and the exact value function $V_1^{\widetilde{\pi}}$; see equation (23b) for a precise statement of our conclusion. We summarize all of our findings thus far in the following:

Proposition 2. *Conditionally on the event $\mathcal{G}(\alpha)$, when given as input 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

$$V_{1,\hat{M}(\pi)}^\pi(s_1) \leq V_1^\pi(s_1) + \sum_{h=1}^{H} \nu_h. \tag{23a}$$

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

$$\left|V^{\widetilde{\pi}}_{1,\hat{M}(\pi)}(s_1) - V^{\widetilde{\pi}}_1(s_1)\right| \le 2\sum_{h=1}^{H} \alpha_h \, \|\bar{\phi}^{\widetilde{\pi}}_h\|_{\Sigma_h^{-1}} + \sum_{h=1}^{H} \nu_h, \tag{23b}$$

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

See Appendix D.2 for the proof.

A.1.3 Choice of pessimism parameters

Based on Proposition 2, our problem is now reduced to determining a choice of α for which the good event (22) holds with probability at least $1-\delta$. The bulk of our effort in analyzing the critic is devoted to the technical details of this step; we provide only a high-level summary here.

The event (22) needs to hold uniformly over the value function and policy classes used by the algorithm. Our analysis involves deriving an upper bound R on the ℓ_2-radius of the actor parameter over all T iterations of the algorithm, as follows:

$$\rho^\theta = \|\theta_T\|_2 = \|\sum_{t=1}^{T} \eta w_t\|_2 \le \sum_{t=1}^{T} \eta \|w_t\|_2 \le T\eta \overset{def}{=} R.$$

For such choice of R and failure probability $\delta \in (0,1)$, suppose that we set

$$\alpha_h(\delta) \overset{def}{=} 1+\sqrt{n_h}\nu_h+c\left\{1 + d\log\left(1+\tfrac{T}{d}\right) + d\log\left(1+8\sqrt{T}\right) + d\log\left(1+16R\sqrt{T}\right) + \log\tfrac{H}{\delta}\right\}^{1/2} \tag{24}$$

for a suitably large universal constant c. Central to our analysis is the following lemma:

Lemma 2. *For any $\delta \in (0,1)$, given the choice of pessimism vector $\alpha(\delta)$ in equation* (24)*, we have*

$$\mathbb{P}\big[\mathcal{G}(\alpha(\delta))\big] \ge 1-\delta. \tag{25}$$

See Section D.3 for the proof of this claim.

In our proof of Lemma 2, we benefit from the fact that our procedure injects its pessimism by direct perturbations of the parameter vectors. Indeed, one key step in the proof is bounding certain metric entropies defined by classes $\mathcal{Q}$ of linear action-value functions, and policy classes $\Pi_{soft}(R)$ used in the actor's iterations.

First, for any fixed policy π, since the agent's action value function $\underline{Q}^\pi$ is enforced to be linear $\underline{Q}^\pi \in \mathcal{Q}$ even after perturbations, the relevant action-value class $\mathcal{Q}$ is also linear. Thus, we need only control metric entropy (and perform union bounds over the resulting covering) for a linear function class; in this way, we avoid a potentially more costly union bound over the much larger function class obtained by adding complex bonuses to linear functions, as in past work Jin et al. (2020b). In this way, we achieve a guarantee that is sharper by a factor of $\sqrt{d}$.

Second, the union bound needs to be extended to all policies that the actor can use to invoke the critic. Recall that the critic returns a linear action-value function $\underline{Q}$, which is compatible Kakade (2001); Agarwal et al. (2020b) with the soft-max policy class Π_{soft}. Consequently, the actor's updates take the simple form (5) of Algorithm 1. If the action-value function $\underline{Q}$ were perturbed by bonuses, then linearity of the critic's value function would be lost.

A.2 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

$$Q^{\pi_t}_{h,M_t}(s,a) \overset{def}{=} \langle \phi_h(s,a),\, w_{ht} \rangle\,, \qquad \text{for all } (s,a) \text{ and } h \in [H]. \tag{26a}$$

In particular, the value of w_{ht} is the value $\underline{w}_{ht}$ identified by the critic (see equation (16)) corresponding to policy π_t, so that $Q^{\pi_t}_{M_t} = \underline{Q}^{\pi_t}$. Define the value function $V^{\pi_t}_{h,M_t}(s) = \mathbb{E}_{A' \sim \pi_t}\big[Q^{\pi_t}_{h,M_t}(s,A')\big]$ along with the advantage function

$$G^{\pi_t}_{h,M_t}(s,a) \overset{def}{=} Q^{\pi_t}_{h,M_t}(s,a) - V^{\pi_t}_{h,M_t}(s). \tag{26b}$$

Proposition 3 (Actor's Analysis)**.** *Suppose that the actor takes $T \geq \log|\mathcal{A}|$ steps using a stepsize $\eta \in (0,1)$, 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*

$$\frac{1}{T}\sum_{t=1}^{T}\big\{V^{\pi}_{1,M_t}(s_1) - V^{\pi_t}_{1,M_t}(s_1)\big\} \leq H\left[\frac{\log|\mathcal{A}|}{\eta T} + \eta\right]. \tag{27a}$$

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

$$\frac{1}{T}\sum_{t=1}^{T}\big\{V^{\pi}_{1,M_t}(s_1) - V^{\pi_t}_{1,M_t}(s_1)\big\} \leq \underbrace{2H\sqrt{\frac{\log|\mathcal{A}|}{T}}}_{=\mathcal{C}(T)}. \tag{27b}$$

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

A.3 Combining the pieces

We are now ready to combine the pieces so as to prove 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. Recall that Lemma 2, stated in Section D.2, guarantees that the "good" event $\mathcal{G}$ from equation (22) occurs with probability at least $1-\delta$. Conditioned on the occurrence of $\mathcal{G}$, the bounds (23a) and (23b) ensure that for any comparator $\widetilde{\pi}$, we have

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

We now average over the iterations $t \in [T]$. The equality (18a) from Lemma 1 ensures for each iteration t, the actor receives as an input a vector $\underline{w}_t$ such that

$$Q^{\pi_t}_{h,M_t}(s,a) \overset{Lem.1}{=} \underline{Q}^{\pi_t}_h(s,a) = \langle \phi_h(s,a),\, \underline{w}_{hk} \rangle\,. \tag{28}$$

Consequently, the action-value function $\underline{Q}^{\pi_t}$ that provided as input to the actor via $\underline{w}_t$ is the action-value function of π_t on the associated induced MDP M_t, i.e., $Q^{\pi_t}_{M_t}$. Applying the bound (27b) from Proposition 3 yields $\frac{1}{T}\sum_{t=1}^{T}\Big[V^{\widetilde{\pi}}_{1,M_t}(s_1) - V^{\pi_t}_{1,M_t}(s_1)\Big] \leq \mathcal{C}(T)$. Combining with the prior display yields

$$V_1^{\widetilde{\pi}}(s_1) - \frac{1}{T}\sum_{t=1}^{T}V_1^{\pi_t}(s_1) \leq \mathcal{C}(T) + \mathcal{E}_{\text{msp}}(\nu) + \mathcal{U}(\widetilde{\pi};\alpha). \tag{29}$$

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

Note that under the good event $\mathcal{G}$, the bound (29) holds for any comparator policy $\widetilde{\pi}$, which was the claim of the theorem.

B 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, 2005); Antos et al. (2007, 2008); Farahmand et al. (2010, 2016); Chen and Jiang (2019); Xie and Jiang (2020a,b); 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. (2018a); 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).

C 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, we define

$$\mathcal{A}_s = \begin{cases} \{-1\} & \text{if } s < 0 \\ \{-1, +1\} & \text{if } s = 0 \\ \{+1\} & \text{if } s > 0. \end{cases} \tag{30}$$

The reward is everywhere zero except in the terminal states $-N$ and $+N$, for which it takes the values -1 and $+1$, 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 $s-1$.

C.1 Proof of part (a): Low Rank $\subset$ Restricted Closed

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.

In order to establish the strict inclusion, consider the MDP described at the beginning of the proof with the following feature extractor:

$$\phi(s,a) = \begin{cases} +1 & \text{if } a = +1 \\ -1 & \text{if } a = -1. \end{cases} \tag{31}$$

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

$$\langle \phi(h-1, +1), \, \theta_h \rangle = \langle \phi(h, +1), \, \theta_{h+1} \rangle \tag{32}$$

$$\langle \phi(-(h-1), -1), \, \theta_h \rangle = \langle \phi(-h, -1), \, \theta_{h+1} \rangle \, . \tag{33}$$

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 $\theta_h = +1$:

$$\langle \phi(H-1, +1), \, \theta_H \rangle = +1 \tag{34}$$

$$\langle \phi(-(H-1), -1), \, \theta_H \rangle = -1. \tag{35}$$

C.2 Proof of part (b): Restricted Closedness ⊂ Linear Q^π

We first show that every MDP that satisfies restricted closedness satisfies the linear Q^π assumption. For any time step $h \in H$, 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

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

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

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

and at the start state

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

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

$$+\theta_1 = \langle \phi(0, +1), \, \theta_1 \rangle = \langle \phi(1, 1), \, \theta_2 \rangle = 1$$
$$-\theta_1 = \langle \phi(0, -1), \, \theta_1 \rangle = \langle \phi(-1, -1), \, \theta_2 \rangle = 1.$$

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

D Proofs for the critic

In this section, we collect together the statements and proofs of various technical results that underlie the critic's analysis in Section A.1. In Section D.1, we prove Lemma 1 that guarantees exactness of the critic on the induced MDP, whereas Section D.2 is devoted to proving our main guarantee for the critic, namely Proposition 2.

Let us introduce some additional notation that plays an important role in the proof. Recall the regression operator $\mathcal{R}^\pi_h$ and sup-norm projection operator $\mathcal{P}^\pi_h$ that were previously defined in equations (20a) and (20c), respectively. In addition to these two operators, our proof also makes use of the *approximation error operator*

$$\mathcal{A}^\pi_h(F)(s, a) \overset{def}{=} \langle \phi(s, a), \, \mathcal{P}^\pi_h(F) \rangle - \left(\mathcal{T}^\pi_h F\right)(s, a), \tag{36}$$

which is a mapping from $\mathcal{F}$ to itself.

D.1 Proof of Lemma 1

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

$$Q^{\pi}_{h,\hat{M}(\pi)}(s,a) - Q^{\pi}_{h}(s,a) = \sum_{\ell=h}^{H} \mathbb{E}_{(S_\ell,A_\ell)\sim\pi|(s,a)} \left[\widehat{r}^{\pi}_{h}(S_\ell, A_\ell) - r_h(S_\ell, A_\ell)\right]. \tag{37a}$$

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

$$\underline{Q}^{\pi}_{h}(s,a) - Q^{\pi}_{h}(s,a) = \sum_{\ell=h}^{H} \mathbb{E}_{(S_\ell,A_\ell)\sim\pi|(s,a)} \Big[\widehat{r}^{\pi}_{h}(S_\ell, A_\ell) - r_h(S_\ell, A_\ell)\Big] \tag{37b}$$

Subtracting equation (37b) from equation (37a) yields the claim.

D.2 Proof of Proposition 2

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

D.2.1 Proof of the bound (23a)

We begin by proving the bound on the critic's estimate for the value function of the input policy π.

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

$$\Big|\sum_{a_1\in\mathcal{A}} \pi(a_1 \mid s_1)\,\langle \phi_1(s_1,a_1),\, \widehat{w}^{\pi}_{1}\rangle - V^{\pi}_{1}(s_1)\Big| \le \sum_{h=1}^{H} \nu_h. \tag{38a}$$

Our second step is to show that conditioned on the good event $\mathcal{G}(\alpha)$ from equation (22), the sequence $\widehat{w}$ is feasible for the critic's convex program; this feasibility, combined with the optimality of $\underline{w}$, implies that

$$V^{\pi}_{1,\hat{M}(\pi)}(s_1) \overset{(i)}{=} \sum_{a_1\in\mathcal{A}} \pi(a_1 \mid s_1)\,\langle \phi_1(s_1,a_1),\, \underline{w}^{\pi}_{1}\rangle \;\le\; \sum_{a_1\in\mathcal{A}} \pi(a_1 \mid s_1)\,\langle \phi_1(s_1,a_1),\, \widehat{w}^{\pi}_{1}\rangle. \tag{38b}$$

Here step (i) follows from Lemma 1, which guarantees that the estimated value functions $\underline{V}^{\pi}_{h}$ of the critic are exact in the induced MDP. Combining the two bounds (38a) and (38b) yields $V^{\pi}_{1,\hat{M}(\pi)}(s_1) \le V^{\pi}_{1}(s_1) + \sum_{h=1}^{H} \nu_h$, as claimed in equation (23a).

It remains to prove our two auxiliary claims (38a) and (38b).

Proof of claim (38a)**:** 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

$$\widehat{w}^{\pi}_{h} \overset{def}{=} \mathcal{P}^{\pi}_{h}(\widehat{Q}^{\pi}_{h+1}) \qquad \text{for } h = H, H-1, \ldots, 1, \tag{39}$$

where $\widehat{Q}^{\pi}_{h+1}(s,a) \overset{def}{=} \big\langle \phi_{h+1}(s,a),\, \widehat{w}^{\pi}_{h+1}\big\rangle$. By construction, we have the bound $\|\widehat{w}^{\pi}_{h}\|_2 \le \rho^{w}_{h}$ 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 3. *The functions* $\{\widehat{Q}^\pi_h\}_{h=1}^H$ *defined by the best-predictor sequence* $\{\widehat{w}^\pi_h\}_{h=1}^H$ *from equation* (39) *satisfy the bound*

$$\left|\widehat{Q}^\pi_h(s,a) - Q^\pi_h(s,a)\right| \leq \sum_{\ell=h}^H \nu_\ell \qquad \textit{for all } h \in [H]. \tag{40}$$

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 (39) 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 (38b)**:** 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 (10). 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) $\|\widehat{\xi}_h\|_{\Sigma_h} \leq \alpha_h$ 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 (39) 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

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

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

D.2.2 Proof of part (b)

Here we prove the bound (23b) stated in part (b) of the lemma, which provides an inequality on the value function error for an arbitrary policy.

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

$$\left|V^{\widetilde{\pi}}_{1,\hat{M}(\pi)}(s_1) - V^{\widetilde{\pi}}_1(s_1)\right| \leq \sum_{h=1}^H \|\bar{\phi}^{\widetilde{\pi}}_h\|_{\Sigma_h^{-1}} \left\{\alpha_h + \|\mathcal{E}^\pi_h(\underline{Q}^\pi_{h+1})\|_{\Sigma_h}\right\} + \sum_{h=1}^H \nu_h, \tag{41}$$

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

Let us now prove the auxiliary claim (41). First, we observe that by definition, the perturbation in the reward can be written as

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

where step (i) uses the definition $\underline{Q}_h^\pi(s,a) = \langle \phi_h(s,a),\, \underline{w}_h^\pi\rangle$; step (ii) uses the relation $\underline{w}_h^\pi = \underline{\xi}_h^\pi + \mathcal{R}_h^\pi(\underline{Q}_{h+1}^\pi)$; and step (iii) involves adding and subtracting $\Big\langle \phi_h(s,a),\, \mathcal{P}_h^\pi(\underline{Q}_{h+1}^\pi)\Big\rangle$, and using the definitions of the approximation error (36) and the error operator (21).

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

$$\begin{aligned}
\Big| V_{1,\hat{M}(\pi)}^{\widetilde{\pi}}(s_1) - V_1^{\widetilde{\pi}}(s_1)\Big| &= \Big| \sum_{h=1}^H \mathbb{E}_{(S_h,A_h)\sim\widetilde{\pi}}\Big[\widehat{r}_h^\pi(S_h,A_h) - r_h(S_h,A_h)\Big]\Big| \\
&= \Big| \sum_{h=1}^H \mathbb{E}_{(S_h,A_h)\sim\widetilde{\pi}}\Big[\,\Big\langle \phi_h(S_h,A_h),\, \underline{\xi}_h^\pi + \mathcal{E}_h^\pi(\underline{Q}_{h+1}^\pi)\Big\rangle + \mathcal{A}_h^\pi(\underline{Q}_{h+1}^\pi)(S_h,A_h)\Big]\Big|.
\end{aligned}$$

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

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

where the final step combines the triangle inequality, with the fact that $\|\underline{\xi}_h^\pi\|_{\Sigma_h} \leq \alpha_h$, since $\underline{\xi}_h^\pi$ must be feasible for the critic's convex program (10). Putting together the pieces yields the claim (41).

D.3 Proof of Lemma 2

We now prove Lemma 2, which asserts that the good event $\mathcal{G}(\delta)$, as defined in equation (22), holds with high probability when the pessimism parameters are chosen according to equation (24).

Recall from equation (21) that for any pair (Q, π), the associated parameter error is given by the difference $\mathcal{E}_h^\pi(Q) = \mathcal{R}_h^\pi(Q) - \mathcal{P}_h^\pi(Q)$. 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.

Recall that $\mathcal{I}_h$ denotes the subset of indices associated with time step h. The first noise variables take the form

$$\eta_{hk}(Q,\pi) \overset{def}{=} r_{hk} + \mathbb{E}_{A'\sim\pi(\cdot|s_{hk})} Q(s_{h+1,k}, A') - (\mathcal{T}_h^\pi Q)(s_{hk}, a_{hk}), \tag{42a}$$

defined for each $h \in [H]$ and $k \in \mathcal{I}_h$. Note that conditionally on the pair (s_{hk}, a_{hk}), our sampling model and the definition of the Bellman operator $\mathcal{T}_h^\pi$ 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

$$\Delta_{hk}(Q,\pi) \overset{def}{=} -\mathcal{A}_h^\pi(Q)(s_{hk}, a_{hk}) \;=\; (\mathcal{T}_h^\pi Q)(s_{hk}, a_{hk}) - \langle \phi_h(s_{hk}, a_{hk}),\, \mathcal{P}_h^\pi(Q)\rangle \tag{42b}$$

With these definitions, we have the following guarantee:

Lemma 4 (Decomposition of $\mathcal{E}_h^\pi(Q)$)**.** *For any pair (Q,π), we have the decomposition*

$$\mathcal{E}_h^\pi(Q) = e_h^\eta(Q,\pi) + e_h^\lambda(Q,\pi) + e_h^\Delta(Q,\pi), \tag{43}$$

where the three error terms are given by

$$e_h^{\eta}(Q,\pi) \overset{def}{=} \Sigma_h^{-1} \sum_{k\in\mathcal{I}_h} \phi_{hk}\eta_{hk}(Q,\pi), \qquad \textit{(Statistical estimation error)} \tag{44a}$$

$$e_h^{\lambda}(Q,\pi) \overset{def}{=} -\lambda\Sigma_h^{-1}\mathcal{P}_h^{\pi}(Q), \qquad \textit{(Regularization error), and} \tag{44b}$$

$$e_h^{\Delta}(Q,\pi) \overset{def}{=} \Sigma_h^{-1} \sum_{k\in\mathcal{I}_h} \phi_{hk}\Delta_{hk}(Q;\pi) \qquad \textit{(Approximation error).} \tag{44c}$$

See Section D.3.1 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 (44b), we have

$$\|e_h^{\lambda}(Q,\pi)\|_{\Sigma_h} = \lambda\|\mathcal{P}_h^{\pi}(Q)\|_{\Sigma_h^{-1}} \overset{(i)}{\leq} \sqrt{\lambda}\|\mathcal{P}_h^{\pi}(Q)\|_2 \overset{(ii)}{\leq} \sqrt{\lambda}, \tag{45}$$

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

Approximation error: By definition, we have $\|e_h^{\Delta}(Q,\pi)\|_{\Sigma_h} = \|\sum_{k\in\mathcal{I}_h} \phi_{hk}\Delta_{hk}(Q,\pi)\|_{\Sigma_h^{-1}}$. 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

$$\|e_h^{\Delta}(Q,\pi)\|_{\Sigma_h} \leq \sqrt{n_h}\nu_h. \tag{46}$$

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

Lemma 5. *There is a universal constant $c > 0$ such that*

$$\|e_h^{\eta}(Q,\pi)\|_{\Sigma_h}^2 \leq c\left\{1 + d_h\log\left(1+\tfrac{T}{d_h\lambda}\right) + d_h\log\left(1+8\sqrt{T}\right) + d_h\log\left(1+16R\sqrt{T}\right) + \log\frac{H}{\delta}\right\} \tag{47}$$

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.3.2 for the proof of this claim.

Putting together the pieces: By combining our three bounds—namely, equations (45), (46) and (47), we conclude that with the choice

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

the good event $\mathcal{G}(\delta)$ holds with probability at least $1-\delta$. This completes the proof of Lemma 2.

It remains to prove the two auxiliary lemmas that we stated: namely, Lemma 4 that gave a decomposition of the parameter error, and Lemma 5 that bounded the statistical error. We do so in Sections D.3.1 and D.3.2, respectively.

D.3.1 Proof of Lemma 4

Starting with the definition (20a) of the regression operator $\mathcal{R}_h^\pi$, we have

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

where equality (i) follows by adding and subtracting terms, and using the definition (42a) of η_{hk}.

Next we use the definition (42b) of the approximation error terms Δ_{hk} to find that

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

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

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

which completes the proof.

D.3.2 Proof of Lemma 5

From the definition (3a), we need to study the constrained class of linear action-value functions based on radii $\rho_h^w \in (0, 1]$ for all $h \in [H]$. As for the constraint defining the soft-max policy class (3b), let us upper bound how large the ℓ_2-norm of the actor's parameter vector can be over T iterations.

Based on the actor's updates, we have the bound

$$\|\theta_{t,h}\|_2 = \| \sum_{t=1}^T \eta w_{t,h} \|_2 \leq \eta \sum_{t=1}^T \|w_{t,h}\|_2 \overset{(i)}{\leq} \eta T \rho_h^w \overset{(ii)}{\leq} \eta T,$$

where step (i) follows from the definition of the critic's program (10), and step (ii) follows from the assumption $\rho_h^w \in (0, 1]$. Thus, we are assured that $R = \eta T$ is an upper bound on this ℓ_2-norm.

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

$$\|\pi - \pi'\|_{\infty,1} \overset{def}{=} \sup_s \sum_{a \in \mathcal{A}} \big| \pi(a \mid s) - \pi'(a \mid s) \big|. \tag{48}$$

We have the following bounds on these covering numbers:

Lemma 6 (Covering number bounds)**.** *For any $\epsilon \in (0, 1)$, we have*

$$\log N_\infty(\epsilon; \mathcal{Q}) \leq d \log \left(1 + \tfrac{2}{\epsilon}\right) \qquad \textit{and} \tag{49a}$$

$$\log N_{\infty,1}(\epsilon; \Pi(R)) \leq d \log \left(1 + \tfrac{16R}{\epsilon}\right). \tag{49b}$$

See Section D.3.3 for the proofs of these claims.

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

$$\beta(\epsilon) \overset{def}{=} d \log\left(1 + \tfrac{T}{d\lambda}\right) + \log N_{\infty}(\epsilon; \mathcal{Q}) + \log N_{\infty,1}(\epsilon; \Pi_{soft}) + \log \frac{H}{\delta} \tag{50}$$

Given this definition and the bounds from Lemma 6, the proof of Lemma 5 is reduced to showing that for any $\epsilon \in (0, 1)$, there is a universal constant c such that

$$\max_{h \in [H]} \sup_{\substack{Q \in \mathcal{Q}_h \\ \pi \in \Pi_{soft}}} \|e^{\eta}_h(Q, \pi)\|_{\Sigma_h} \leq c \sqrt{\beta(\epsilon)} + 4\sqrt{T}\epsilon \tag{51}$$

with probability at least $1 - \delta$. The claim stated in Lemma 5 follows from the choice $\epsilon = \frac{1}{4\sqrt{T}}$. The remainder of our proof is devoted to the proof of this claim.

Proof of the claim (51): 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 $\|e^{\eta}_h(Q, \pi)\|_{\Sigma_h} = \| \sum_{k \in \mathcal{I}_h} \phi_{hk} \eta_{hk}(Q, \pi)\|_{\Sigma^{-1}_h} \leq Z_1 + Z_2(Q, \pi)$, where

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

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

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 (48). 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

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

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 $D^Q \leq 2\epsilon\sqrt{T}$. 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 $D^\pi \leq 2\epsilon\sqrt{T}$.

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

$$Z_2(Q, \pi) \leq \max_{\substack{i \in [N] \\ j \in [M]}} Z_2(Q^i, \pi^j) + 4\sqrt{T}\epsilon. \tag{52}$$

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 $\| \sum_{k \in \mathcal{I}_h} \phi_{hk} V_{hk} \|_{\Sigma_h^{-1}}$, 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 $\log \det \Sigma_h \leq d_h \log \left(\lambda + \frac{T}{d_h} \right)$.

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

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

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 (51).

D.3.3 Proof of Lemma 6

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 (49a) follows from standard results on coverings of Euclidean balls (cf. Example 5.8 in the book Wainwright (2019)).

As for the bound (49b), we claim that

$$\sum_{a \in \mathcal{A}} \left|\pi_{\theta'}(a \mid s) - \pi_\theta(a \mid s)\right| \leq 8\|\theta - \theta'\|_2, \qquad \text{for all } s \in \mathcal{S}. \tag{53}$$

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 (49b).

It remains to prove the claim (53).

Proof of the claim (53): Let us state and prove the claim (53) 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 7 (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*

$$\sum_{a \in \mathcal{A}} \left|\pi_{\theta'}(a \mid s) - \pi_\theta(a \mid s)\right| \leq 8\|\theta - \theta'\|_2, \tag{54}$$

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

$$T \leq e^{\|\theta'-\theta\|_2} \sum_{a''\in\mathcal{A}} \pi_{\theta'}(a'' \mid s) e^{\|\theta-\theta'\|_2} \overset{(i)}{=} e^{2\|\theta-\theta'\|_2} \overset{(ii)}{\leq} 1 + 4\|\theta-\theta'\|_2, \tag{55}$$

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 (55) 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

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

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: Proof of Proposition 3

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

$$Q(s,a) - F(s) = \frac{1}{\eta}\left[\log\frac{\pi_{\theta^+}(s,a)}{\pi_\theta(s,a)} + \log\left(\sum_{a'\in\mathcal{A}} \pi_\theta(s,a') e^{\eta\left(Q(s,a') - F(s)\right)}\right)\right], \tag{56}$$

valid for all pairs (s,a).

See Section E.1 for the proof of this claim.

Turning to the proof of the proposition, we have

$$V^{\pi}_{1,M_t}(s_1) - V^{\pi_t}_{1,M_t}(s_1) \overset{(i)}{=} \sum_{h=1}^{H} \mathbb{E}_{(S_h,A_h)\sim\pi}\left[G^{\pi_t}_{h,M_t}(S_h,A_h)\right] \overset{(ii)}{=} \frac{1}{\eta}\sum_{h=1}^{H} X_{h,t}, \tag{57}$$

where we have introduced the shorthand

$$X_{h,t} \overset{def}{=} \mathbb{E}_{(S_h,A_h)\sim\pi}\left[\log\frac{\pi_{\theta_{t+1}}(S_h,A_h)}{\pi_{\theta_t}(S_h,A_h)} + \log\left(\mathbb{E}_{A'_h\sim\pi_t(\cdot|S_h)}\left[e^{\eta G^{\pi_t}_{h,M_t}(S_h,A'_h)}\right]\right)\right]. \tag{58}$$

Here step (i) follows from the simulation lemma (e.g., Kakade et al. (2003)), and step (ii) makes use of Lemma 8 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 (58) 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

$$\begin{aligned}\sum_{a_h \in \mathcal{A}} \pi(a_h \mid s_h) \log \frac{\pi_{t+1}(s_h, a_h)}{\pi_t(s_h, a_h)} &= \sum_{a_h} \pi(a_h \mid s_h) \Big[\log \frac{\pi_{t+1}(s_h, a_h)}{\pi(s_h, a_h)} - \log \frac{\pi_t(s_h, a_h)}{\pi(s_h, a_h)} \Big] \\ &= -D_{s_h}(\pi; \pi_{t+1}) + D_{s_h}(\pi; \pi_t). \end{aligned} \tag{59}$$

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 $\mathbb{E}_{A'_h \sim \pi_t}\left[G^{\pi_t}_{h,M_t}(s_h, A'_h)\right] = 0$, so that we have

$$\log\Big(\mathbb{E}_{A'_h \sim \pi_t} e^{\eta G^{\pi_t}_{h,M_t}(s_h, A'_h)}\Big) \leq \log\left(1 + 4\eta^2\right) \leq 4\eta^2. \tag{60}$$

Combining the pieces: Combining the bounds (59) and (60) 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

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

thereby establishing the claim (27a).

Finally, the bound (27b) follows by making the particular stepsize choice $\eta = \sqrt{\frac{\log |\mathcal{A}|}{T}}$. Note that the assumed lower bound $T \geq \log |\mathcal{A}|$ ensures that $\eta \leq 1$, as required to apply the bound (27a).

E.1 Proof of Lemma 8

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 Proof of Theorem 2

We now turn to the proof of the lower bound stated in Theorem 2. In Section F.1, we describe the class of MDPs used in the construction, along with the data generating procedure. Section F.2 provides the core argument, which involves three auxiliary lemmas. These lemmas are proved in Sections F.3, F.4 and F.5, respectively.

F.1 MDP class and data collection

For a given horizon H and dimension d, we define a family of MDPs that are parameterized by a Boolean vector $u = (u_1, \ldots, u_H) \in \{-1,+1\}^{dH}$, where each $u_h \in \{-1,+1\}^d$. For a given Boolean vector u, the associated MDP M_u has the following structure:

State space and transition: At each time step h, there is only one state—viz. $\mathcal{S} = \{s\}$. Since there is a single state, the transition is deterministic into the same state.

Action space: At each time step h, the action space is given by $\mathcal{A} = \{-1, 0, +1\}^d$.

Feature map: At each time step h, the feature map $\phi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^{d+1}$ takes the form

$$\phi(s,a) = \left[\tfrac{a}{\sqrt{2d}}, \tfrac{1}{\sqrt{2}}\right]. \tag{61}$$

Notice that by construction, we have the bound $\|\phi(s,a)\|_2 = \sqrt{\frac{\|a\|_2^2}{2d} + \frac{1}{2}} \leq 1$ for any state-action pair.

Reward mean: The mean reward at time step h is proportional to the inner product $\langle a,\, u_h \rangle$, where $u_h \in \{-1,1\}^d$ is the sub-vector associated with time step h. More precisely, we have

$$r_h(s,a) = \langle \phi_h(s,a),\, [\delta u_h \quad 0] \rangle \;=\; \tfrac{\delta}{2\sqrt{d}} \langle a,\, u_h \rangle, \tag{62}$$

where $\delta > 0$ is a parameter to be specified in the proof.

Low-rank MDP model: It is easy to verify that the MDP so defined is low-rank; here we only verify explicitly the regularity conditions about the size of the radii so that the setting for the lower bound matches the setting that PACLE can handle. We need to verify explicitly that we can represent the action value function for any policy π, namely that there exists w_h^π such that the action value function $Q_h^\pi(s,a) = \langle \phi_h(s,a),\, w_h^\pi \rangle$ with $\|w_h^\pi\|_2 \leq (H-h+1)/(2H)$. One can verify that for any policy π we have

$$w_h^\pi = [\delta u_h, \sqrt{2} V_{h+1}^\pi], \qquad \forall h \in [H]. \tag{63}$$

A sufficient condition for the regularity conditions to be satisfied is when

$$\delta \|u_h\|_2 \leq 1/(2H) \to \delta \leq \frac{1}{2\sqrt{d}H}, \tag{64}$$

which implies $|V_h^\pi| \leq (H-h+1)/(2H)$ and hence

$$\|w_h^\pi\|_2 \leq \delta_2 \|u_h\|_2 + \sqrt{2}|V_{h+1}^\pi| \leq (H-h+1)/(2H), \qquad \forall h \in [H]. \tag{65}$$

In Lemma 10 we choose $\delta = \frac{d\sqrt{H}}{\sqrt{2n}}$ which implies the lemma holds when

$$\frac{d\sqrt{H}}{\sqrt{2n}} \leq \frac{1}{2\sqrt{d}H} \to n \geq 2d^3H^3. \tag{66}$$

Reward observations: We observe the mean reward contaminated by additive Gaussian noise, so that the reward distribution has the form

$$R_h(s,a) \sim \mathcal{N}\left(\frac{\delta}{\sqrt{2d}} \langle a,\, u_h \rangle, 1\right). \tag{67}$$

Data collection: We assume that the n samples are collected according to the following non-adaptive process.

- Each time step $h \in [H]$ is allocated $n_H \overset{def}{=} n/H$ samples (assumed to be an integer for simplicity).
- For each h, the dataset $\mathcal{D}_h$ is generated by playing each action $a \in \{e_1, \ldots, e_d, 0\}$ exactly $n_H/(d+1)$ times, where $e_j \in \{0,1\}^d$ denotes the standard basis vector with a single one in index j.

F.2 Main argument

With this set-up, we now introduce the three lemmas that form the core of the proof. For any given $u \in \{-1,+1\}^{dH}$, let $\mathbb{Q}_u$ denote the distribution of the data $\mathcal{D}$ when the sampling process is applied to the MDP M_u, and let $\mathbb{E}_u$ denote expectations under this distribution. Our first lemma exploits the Assouad construction so as to reduce the problem of finding a good policy to a family of testing problems.

Lemma 9 (Reduction to testing)**.** *For any estimated policy π_{ALG}, we have*

$$\sup_{u \in \mathcal{U}} \mathbb{E}_u[V_u^\star - V_u^{\pi_{\text{ALG}}}] \geq \frac{\delta}{\sqrt{2d}} \frac{dH}{2} \min_{\substack{u,u' \in \mathcal{U} \\ D_H(u;u')=1}} \inf_{\psi} \Big[\mathbb{Q}_u(\psi(\mathcal{D}) \neq u) + \mathbb{Q}_{u'}(\psi(\mathcal{D}) \neq u')\Big], \tag{68}$$

where a test function ψ is a measurable function of the data taking values in $\{u, u'\}$.

See Section F.3 for the proof.

Our second lemma involves further lower bounding the testing error in the bound (69). In particular, we prove the following:

Lemma 10 (Lower bound on testing error)**.** *For the given family of distributions $\{\mathbb{Q}_u, u \in \mathcal{U}\}$, we have*

$$\min_{\substack{u,u' \in \mathcal{U} \\ D_H(u;u')=1}} \inf_{\psi} \Big[\mathbb{Q}_u(\psi(\mathcal{D}) \neq u) + \mathbb{Q}_{u'}(\psi(\mathcal{D}) \neq u')\Big] \geq \left(1 - \sqrt{\frac{1}{2}\frac{n_H \delta^2}{d^2}}\right). \tag{69}$$

Thus, the testing error is lower bounded by $\frac{1}{2}$ with the choice $\delta = \frac{d}{\sqrt{2n_H}}$.

See Section F.4 for the proof.

Combining the claims of Lemmas 9 and 10, along with the choice $\delta = \frac{d}{\sqrt{2n_H}}$, yields the lower bound

$$\sup_{u \in \mathcal{U}} \mathbb{E}_u[V_u^\star - V_u^{\pi_{\text{ALG}}}] \geq \frac{\delta}{\sqrt{2d}} \frac{dH}{2} \frac{1}{2} \geq \frac{1}{8} dH \sqrt{\frac{d}{n_H}}. \tag{70}$$

Thus, the only remaining step is to relate this lower bound to the uncertainty function $\mathcal{U}(\pi; \sqrt{d})$ associated with our family of MDPs. More precisely, we prove the following:

Lemma 11. *There is a universal constant such that*

$$\sup_{\pi} \mathcal{U}(\pi; \sqrt{d}) \leq c dH \sqrt{\frac{d}{n_H}} \tag{71}$$

See Section F.5 for the proof.

Combining Lemma 11 with the lower bound (70) concludes the proof of the theorem.

It remains to prove our auxiliary lemmas, and we do so in the following subsections.

F.3 Proof of Lemma 9

For a given $u \in \mathcal{U}$, let $\pi_u^\star$ be the optimal policy on M_u and let $V_u^\star$ the optimal value function. For any estimated policy π, we define the estimated sign vector $u^\pi \in \{-1, 1\}^{dH}$ with entries $[u^\pi]_{hi} \overset{def}{=} \operatorname{sign}(\mathbb{E}_{a\sim\pi_h} a_i)$.

With this set-up, we prove the lemma in two steps:

(a) First, we show that the value function gap $V_u^\star - V_u^\pi$ can be lower bounded in terms of the Hamming distance

$$V_u^\star - V_u^\pi \geq \frac{\delta}{\sqrt{2d}} D_{\mathrm{H}}(u^\pi; u). \tag{72}$$

(b) We use Assaoud's method to lower bound the estimation error in the Hamming distance.

Step (a): 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

$$\begin{aligned}
V_u^\star - V_u^\pi &= \frac{1}{\sqrt{2d}} \sum_{h=1}^{H} \Big[\langle u_h,\, \delta u_h \rangle - \mathbb{E}_{a\sim\pi_h} \langle a,\, \delta u_h \rangle \Big] \\
&= \frac{\delta}{\sqrt{2d}} \sum_{h=1}^{H} \sum_{i=1}^{d} \Big[[u]_{hi}[u]_{hi} - [\mathbb{E}_{a\sim\pi_h} a]_i [u]_{hi} \Big] \\
&= \frac{\delta}{\sqrt{2d}} \sum_{h=1}^{H} \sum_{i=1}^{d} \Big([u]_{hi} - [\mathbb{E}_{a\sim\pi_h} a]_i \Big) [u]_{hi} \\
&= \frac{\delta}{\sqrt{2d}} \sum_{h=1}^{H} \sum_{i=1}^{d} \Big| [u]_{hi} - [\mathbb{E}_{a\sim\pi_h} a]_i \Big|.
\end{aligned}$$

Now recalling that $[u^\pi]_{hi} \overset{def}{=} \operatorname{sign}(\mathbb{E}_{a\sim\pi_h} a_i)$, we have the lower bound

$$\begin{aligned}
V_u^\star - V_u^\pi &\geq \frac{\delta}{\sqrt{2d}} \sum_{h=1}^{H} \sum_{i=1}^{d} \Big| [u]_{hi} - [\mathbb{E}_{a\sim\pi_h} a]_i \Big| \mathbb{1}\{u^\pi_{hi} \neq [u]_{hi}\} \\
&\geq \frac{\delta}{\sqrt{2d}} \sum_{h=1}^{H} \sum_{i=1}^{d} \mathbb{1}\{u^\pi_{hi} \neq [u]_{hi}\} \\
&= \frac{\delta}{\sqrt{2d}} D_{\mathrm{H}}(u^\pi; u),
\end{aligned}$$

which establishes the lower bound (72).

Step (b): We can now apply Assouad's method (cf. Lemma 2.12 in the book Tsybakov (2009)), so as to conclude that for any estimated policy π, we have

$$\sup_{u\in\mathcal{U}} \mathbb{E}_u \big[D_{\mathrm{H}}(u^\pi; u) \big] \geq \frac{dH}{2} \min_{u,u' \mid D_{\mathrm{H}}(u;u')=1} \inf_{\psi} \Big[\mathbb{P}_u(\psi \neq u) + \mathbb{P}_{u'}(\psi \neq u') \Big] \tag{73}$$

where $\inf_\psi$ denotes the minimum over all test functions taking values in $\{u, u'\}$.

F.4 Proof of Lemma 10

We begin by observing that the testing error can be lower bounded in terms of the KL divergence as

$$\min_{\substack{u,u'\in\mathcal{U} \\ D_{\mathrm{H}}(u;u')=1}} \inf_{\psi} \Big[\mathbb{P}_u(\psi \neq u) + \mathbb{P}_{u'}(\psi \neq u') \Big] \geq 1 - \Big(\tfrac{1}{2} \max_{\substack{u,u'\in\mathcal{U} \\ D_{\mathrm{H}}(u;u')=1}} D_{\mathrm{KL}}(\mathbb{Q}_u \| \mathbb{Q}_{u'}) \Big)^{1/2}. \tag{74}$$

For instance, see Theorem 2.12 in Tsybakov (2009).

Thus, in order to prove Lemma 10, it remains to bound the Kullback-Leibler divergence of the distributions $\mathbb{Q}_u$ and $\mathbb{Q}_{u'}$ for pairs $u, u' \in \{-1, +1\}^{dH}$ that differ only in a single coordinate.

By construction, the only stochasticity in the dataset lies in the rewards. For any given u, equation (67) implies that the distribution over rewards has the product form

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

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

$$D_{\mathrm{KL}}(\mathbb{Q}_u \| \mathbb{Q}_{u'}) = \sum_{k=1}^{\frac{n_H}{d}} D_{\mathrm{KL}}(\mathcal{N}\left(\frac{\delta}{\sqrt{2d}}, 1\right) \| \mathcal{N}\left(\frac{-\delta}{\sqrt{2d}}, 1\right)) = \frac{n_H}{2d}\left(2\frac{\delta}{\sqrt{2d}}\right)^2$$
$$= \frac{n_H \delta^2}{d^2},$$

valid for any pair u, u' differing in a single coordinate. Substituting back into the lower bound (74) yields the claim.

F.5 Proof of Lemma 11

Recall that by definition, we have $\mathcal{U}(\pi; \sqrt{d}) = \sqrt{d}\sum_{h=1}^{H} \|\phi_h^\pi\|_{\Sigma_h^{-1}}$. Consequently, in order to establish the claim, it suffices to show there is a universal constant c such that

$$\sup_{\pi \in \Pi} \|\phi_h^\pi\|_{\Sigma_h^{-1}} \leq c\frac{d}{\sqrt{n_H}} \qquad \text{for each } h \in [H]. \tag{75}$$

Now denote with $[x]_{1:p}$ the first p components of the vector x, and with $[x]_p$ the p component of 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. (61) 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

$$\|x\|^2_{D'} = x^\top (D'_1 + D'_2) x \leq \|x\|_2 \left(\|D'_1\|_2 + \|D'_2\|_2 \right) \|x\|_2$$

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:

$$\| [\phi_h^\pi]_{1:d} \|_2 \leq \frac{\| \mathbb{1} \|_2}{\sqrt{2d}} \leq \frac{1}{\sqrt{2}}$$
$$\|D'_1\|_2 = \|D^{-1}\|_2 \lesssim \frac{2d^2}{n_H}$$
$$\|D'_2\|_2 \leq b' \|D^{-1}\|_2 \|v\|_2 \|v\|_2 \|D^{-1}\|_2 \lesssim \underbrace{\frac{d}{n_H}}_{b'} \underbrace{\left(\gamma \frac{n_H}{d} \right)^2 \| \mathbb{1} \|_2^2}_{\|v\|_2^2} \underbrace{\frac{d^4}{n_h^2}}_{\|D^{-1}\|_2^2} \lesssim \frac{d^2}{n_H}$$

Substituting back yields the bound

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

Similarly, we have

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

Putting together the pieces yields the claim (75).