Pessimism for Offline Linear Contextual Bandits using ℓ_p Confidence Sets

Gene Li
Toyota Technological Institute at Chicago
gene@ttic.edu

Cong Ma
Department of Statistics
University of Chicago
congm@uchicago.edu

Nathan Srebro
Toyota Technological Institute at Chicago
nati@ttic.edu

Abstract

We present a family $\{\widehat{\pi}_p\}_{p\geq 1}$ of pessimistic learning rules for offline learning of linear contextual bandits, relying on confidence sets with respect to different ℓ_p norms, where $\widehat{\pi}_2$ corresponds to Bellman-consistent pessimism (BCP), while $\widehat{\pi}_\infty$ is a novel generalization of lower confidence bound (LCB) to the linear setting. We show that the novel $\widehat{\pi}_\infty$ learning rule is, in a sense, adaptively optimal, as it achieves the minimax performance (up to log factors) against all ℓ_q-constrained problems, and as such it strictly dominates all other predictors in the family, including $\widehat{\pi}_2$.

1 Introduction

Offline (or batch) reinforcement learning (RL) [17, 18] seeks to learn a good policy from fixed historical data without active interactions with the environment. This offline paradigm has been widely adopted in applications including dialog generation [10], autonomous driving [43], and robotic control [16], etc.

When the offline dataset has insufficient coverage over the state and action spaces, planning via nominal estimates of either the value function or the model may perform poorly—a phenomenon that is observed even in a simple two-armed bandit [24]. This challenge motivates the adoption of the *pessimism principle* for solving offline RL. In essence, the pessimism principle discounts policies that are less represented/supported in the offline dataset, and hence is pessimistic/conservative in outputting a policy. Built on this common principle, a diverse collection of pessimistic learning rules have been proposed in theory and practice [11, 24, 36, 37, 45, 46, 9, 15, 34, 14, 21, 42]. This leads us to the following natural question:

Which pessimistic learning rule should one use for solving offline RL problems?

In this paper, we address the question in the setting of offline linear contextual bandits, in which the expected reward—as a function of the state-action pair—is linear with respect to a known feature mapping that maps state-action pairs to finite-dimensional vectors. Our goal is to make sense of previously proposed learning rules for offline RL, and understand which learning rule is "optimal" in a statistical sense. We present a general family $\{\widehat{\pi}_p\}_{p\geq 1}$ of pessimistic learning rules based on the construction of ℓ_p confidence sets for the unknown linear parameter. We advocate for $\widehat{\pi}_\infty$, a new ℓ_∞ learning rule for offline linear contextual bandits, which we call Pessimism via Uniform Norm Confidence (for short, PUNC).[1] PUNC directly extends the lower confidence bound algorithm proposed in the tabular contextual bandit setting [24]. We show that PUNC (1) achieves a suboptimality guarantee that dominates other $\widehat{\pi}_p$ (up to log factors, which we ignore throughout the introduction), and (2) has an *adaptive minimax optimality* property that is unique among the family $\{\widehat{\pi}_p\}_{p\geq 1}$. In particular, we argue that PUNC dominates prior learning rules which are based on ℓ_2 pessimism (e.g., [37, 45, 11]) and which cannot attain adaptive minimax optimality.

[1] Throughout the paper, we use $\widehat{\pi}_\infty$ and PUNC interchangeably.

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

Roadmap. We first introduce a broad class of pessimistic learning rules in Section 3. The construction of these pessimistic learning rules relies on the observation that any *confidence set* of the linear reward function automatically induces a pessimistic value estimate, and hence a pessimistic learning rule. As concrete examples, for each $p \geq 1$, one can design $\widehat{\pi}_p$, an ℓ_p learning rule, by constructing such a confidence set using the ℓ_p distance metric. We show in Section 3.3 that $\widehat{\pi}_2$ recovers the Bellman-consistent pessimism (BCP) learning rule [37], proposed for offline RL with general function approximation; meanwhile, $\widehat{\pi}_\infty$ generalizes the lower confidence bound (LCB) learning rule, proposed for offline tabular RL, to the linear setting.

Once we have cast pessimistic estimation in this framework, we can study the performance guarantees of the family $\{\widehat{\pi}_p\}_{p\geq 1}$. Employing a notion of *pessimism-validity* (Definition 1) allows us to easily to derive upper bounds on suboptimality for each $\widehat{\pi}_p$ in terms of the dual ℓ_q norm (where $1/p+1/q=1$); see Theorem 1. For $p=2$, the upper bound improves over that provided in the paper [37] for linear contextual bandits. For $p=\infty$, the upper bound matches that proved in the paper [24] for tabular contextual bandits. A key observation regarding the upper bound is that the suboptimality guarantee of $\widehat{\pi}_\infty$ *dominates* all other $\widehat{\pi}_p$ in the general linear setting. This partially showcases the advantage of using PUNC.

To further investigate the advantage of PUNC over other $\widehat{\pi}_p$ (for $p \in [1,\infty)$), we consider the fundamental statistical limits of the offline linear contextual bandit problem in Section 4. Inspired by both the upper bounds we prove and prior work [45, 24, 40], we consider a sequence of norm-constrained classes of contextual bandit instances indexed by the ℓ_q norm ($q \geq 1$). We prove that each $\widehat{\pi}_p$ is minimax rate-optimal within the dual ℓ_q-norm constrained contextual bandit class; see Theorem 2. However, Theorem 2 delivers an even stronger message: PUNC is *adaptively minimax optimal* in the sense that it simultaneously achieves optimality for all ℓ_q-norm constrained classes, as illustrated by Figure 1. We also demonstrate that such adaptivity is unique to PUNC as other values of p (e.g., $p=2$) cannot achieve simultaneous optimality. Instead, $\widehat{\pi}_p$ is only adaptively optimal for ℓ_q-norm constrained classes where $q \geq p/(p-1)$; see Theorem 3.

In summary, our main contributions are the following:

- We introduce a novel learning rule, PUNC, for solving the offline linear contextual bandit problem, whose performance guarantee dominates those of all other $\widehat{\pi}_p$, for finite p (Theorem 1).
- We show minimax lower bounds over norm-constrained classes of contextual bandit instances, which show that each $\widehat{\pi}_p$ is optimal over the dual ℓ_q class, up to log factors in the dimension (Theorem 2).
- We demonstrate that PUNC satisfies the adaptive minimax optimality property (Section 4.2), and show that this property is unique to PUNC by proving a separation result against any other $\widehat{\pi}_p$ (Theorem 3, and see also Figure 1).

2 Problem setup

We begin by introducing the problem of offline learning in linear contextual bandits. Let $\mathcal{S}$ and $\mathcal{A}$ be the state space and the action space, respectively. Let $\phi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$ be a known feature mapping. In the offline setting, we observe a dataset $\mathcal{D} := \{(s_i, a_i, r_i)\}_{i=1}^n$, where the covariates $\{(s_i, a_i)\}_{i=1}^n$ are fixed and the rewards are drawn independently according to $r_i \sim R(s_i, a_i)$, where $R(s,a)$ is the reward distribution associated with the pair (s,a). We assume that $R(s,a)$ is 1-subgaussian for every (s,a) with mean reward $r(s,a) := \mathbb{E}[R(s,a)]$. Furthermore, we assume that the expected reward is linear in the sense that for every (s,a) pair, $r(s,a) = \phi(s,a)^\top \theta^\star$ for some unknown parameter vector $\theta^\star \in \mathbb{R}^d$.

Let $\pi : \mathcal{S} \to \mathcal{A}$ be a deterministic policy. Fixing a (known) test distribution $\rho \in \Delta(\mathcal{S})$, we define the value of the policy π (with respect to ρ) as

$$V(\pi) := \mathbb{E}_{s\sim\rho}\left[r(s,\pi(s))\right] = \mathbb{E}_{s\sim\rho}\left[\phi(s,\pi(s))^\top \theta^\star\right]. \tag{1}$$

Correspondingly, we define the optimal policy $\pi^\star$ as

$$\pi^\star(s) := \arg\max_{a\in\mathcal{A}} \; r(s,a) = \arg\max_{a\in\mathcal{A}} \; \phi(s,a)^\top \theta^\star, \quad \text{for each } s \in \mathcal{S}. \tag{2}$$

The goal of offline learning in linear contextual bandits is to design a learning rule which takes as input a dataset $\mathcal{D}$ and outputs a policy $\widehat{\pi}$ that maximizes the value (1); in this paper we abuse notation and also denote the learning rule as $\widehat{\pi}$. We measure the suboptimality of $\widehat{\pi}$ using $V(\pi^\star) - V(\widehat{\pi})$.

3 Offline learning with pessimism

The pessimism principle has recently gained much attention in offline RL theory and practice. At a high level, pessimistic learning rules first construct a data-dependent estimate $\widehat{V}(\pi)$ of the true value function $V(\pi)$ that is pessimistic, i.e., $\widehat{V}(\pi) \leq V(\pi)$ for all π. Then, the learning rule proceeds to select the policy that maximizes this pessimistic value function, i.e.,

$$\widehat{\pi} \coloneqq \arg\max_{\pi \in \Pi} \widehat{V}(\pi). \tag{3}$$

Here, $\Pi \subseteq \mathcal{A}^{\mathcal{S}}$ is some policy class that contains the optimal policy $\pi^\star$. To see why this choice of policy makes sense, let us decompose the suboptimality of $\widehat{\pi}$ as follows:

$$V(\pi^\star) - V(\widehat{\pi}) = \big(V(\pi^\star) - \widehat{V}(\pi^\star)\big) + \big(\widehat{V}(\pi^\star) - \widehat{V}(\widehat{\pi})\big) + \big(\widehat{V}(\widehat{\pi}) - V(\widehat{\pi})\big). \tag{4}$$

The middle term is non-positive by definition of $\widehat{\pi}$. Due to the pessimistic property of $\widehat{V}$, we also have $\widehat{V}(\widehat{\pi}) - V(\widehat{\pi}) \leq 0$, which yields the suboptimality upper bound

$$V(\pi^\star) - V(\widehat{\pi}) \leq V(\pi^\star) - \widehat{V}(\pi^\star). \tag{5}$$

Consequently, under the selection rule (3), a tight pessimistic value function $\widehat{V}$ induces a policy with small suboptimality.

3.1 Achieving pessimism by building confidence sets

As a general strategy, one can construct the pessimistic value estimator $\widehat{V}$ by building confidence sets for the linear parameter $\theta^\star$. Let $\Theta \subseteq \mathbb{R}^d$ be a *confidence set* that contains the true parameter $\theta^\star$. We can define the corresponding pessimistic value estimator

$$\widehat{V}(\pi) \coloneqq \inf_{\theta \in \Theta} \mathbb{E}_{s\sim\rho}[\phi(s, \pi(s))^\top \theta], \tag{6}$$

and its associated policy learning rule $\widehat{\pi}_\Theta \coloneqq \arg\max_{\pi\in\Pi} \widehat{V}(\pi)$. Here for simplicity we take Π to be the class of all deterministic policies.

In essence, the confidence set Θ captures the amount of uncertainty we have about the ground truth $\theta^\star$. Once Θ is determined, we construct the value estimate $\widehat{V}(\pi)$ via the worst-case value of π among all plausible linear parameters θ in the confidence set Θ. It is immediate to see that under the assumption $\theta^\star \in \Theta$, one has $\widehat{V}(\pi) \leq V(\pi)$ for all π. In other words, the value estimator $\widehat{V}$ is indeed pessimistic. As a result, we can apply the general upper bound (5) to obtain

$$V(\pi^\star) - V(\widehat{\pi}_\Theta) \leq V(\pi^\star) - \widehat{V}(\pi^\star) = \sup_{\theta\in\Theta} \mathbb{E}_{s\sim\rho}[\phi(s, \pi^\star(s))^\top (\theta^\star - \theta)], \tag{7}$$

where the identity follows from the definition (6). Clearly, the "smaller" the confidence set, the smaller the bound on suboptimality. An extreme case is when Θ contains the singleton $\theta^\star$, which yields zero suboptimality. However, since only noisy rewards are observed, we cannot hope to construct such a good confidence set. Given the uncertainty about the rewards, our confidence set has to be "large" enough in order to guarantee that $\theta^\star \in \Theta$ with decent probability.

Below we present a general definition called pessimism-validity that involves both the size of the confidence set and also its confidence level, both of which allow us to bound the suboptimality of the pessimistic learning rule $\widehat{\pi}_\Theta$. Let $\|\cdot\|$ be any norm over $\mathbb{R}^d$ that will be used to measure the size of the confidence set Θ. Let $\delta \in (0, 1)$ be the failure probability. We have the following definition.

Definition 1. *We say the confidence set Θ is (β, δ) pessimism-valid under the norm $\|\cdot\|$ if with probability at least $1 - \delta$, the following two properties hold: (1) $\theta^\star \in \Theta$; (2) $\sup_{\theta\in\Theta} \|\theta^\star - \theta\| \leq \beta$.*

A (β, δ) pessimism-valid confidence set Θ automatically induces a pessimistic learning rule $\widehat{\pi}_\Theta$ with bounded suboptimality, as shown in the following proposition.

Proposition 1. *Suppose that Θ is (β, δ) pessimism-valid under the norm $\|\cdot\|$. Then with probability at least $1 - \delta$, the pessimistic learning rule $\widehat{\pi}_\Theta$ obeys*

$$V(\pi^\star) - V(\widehat{\pi}_\Theta) \leq \beta \cdot \left\| \mathop{\mathbb{E}}_{s\sim\rho} [\phi(s, \pi^\star(s))] \right\|_*,$$

where $\|\cdot\|_*$ is the dual norm of $\|\cdot\|$.

Proposition 1 simply follows from the upper bound (7), the definition of pessimism-validity, and the definition of the dual norm.

3.2 Building ℓ_p confidence sets

In this section, we instantiate the general strategy introduced above for achieving pessimism by constructing an ℓ_p confidence set around the true parameter $\theta^\star$ for some $p \geq 1$. Such constructions using ℓ_p norms include the aforementioned BCP and LCB learning rules (as well as other recently proposed learning rules) as special cases. As we will see, setting up the notion of pessimism-validity allows us to easily bound the suboptimality of the induced policy learning rules.

Let us denote the data matrix $\Phi \in \mathbb{R}^{n \times d}$, where the i-th row of Φ is given by $\phi(s_i, a_i)^\top$. We also define the observed reward vector $r := (r_1, \ldots, r_n)^\top \in \mathbb{R}^n$. Let $\widehat{\theta}_{\mathsf{ols}} := (\Phi^\top \Phi)^{-1} \Phi^\top r$ be the ordinary least-squares estimate for the true parameter $\theta^\star$. Throughout the paper, we assume that the sample "covariance" matrix $\Sigma_{\mathcal{D}} := \frac{1}{n} \sum_{i=1}^n \phi(s_i, a_i) \phi(s_i, a_i)^\top = \frac{1}{n} \Phi^\top \Phi$ is invertible. (The results in the paper can be modified to accomodate the scenario when $\Sigma_{\mathcal{D}}$ is not invertible by considering regularized quantities $\Sigma_{\mathcal{D}} + \lambda I$ for some $\lambda > 0$.) We then consider the confidence sets of the form:

$$\Theta_p := \left\{ \theta \in \mathbb{R}^d \mid \left\| \Sigma_{\mathcal{D}}^{1/2} (\theta - \widehat{\theta}_{\mathsf{ols}}) \right\|_p \leq \beta/2 \right\}, \tag{8}$$

where $\beta > 0$ is a width parameter. In other words, the set Θ_p contains all the θ's that are close to the OLS estimate $\widehat{\theta}_{\mathsf{ols}}$ in ℓ_p distance after the linear transformation $\Sigma_{\mathcal{D}}^{1/2}$. Since $\widehat{\theta}_{\mathsf{ols}}$ is a faithful approximation of the truth $\theta^\star$, we expect that $\theta^\star$ lies in this confidence set Θ_p with an appropriate choice of β. This is indeed true, as the following lemma shows.

Lemma 1. *Fix any $\delta \in (0, 1)$. Set the width parameter $\beta = d^{1/p} \sqrt{\frac{8 \log(d/\delta)}{n}}$. Then the confidence set Θ_p is (β, δ) pessimism-valid with respect to the norm $\|v\| := \|\Sigma_{\mathcal{D}}^{1/2} v\|_p$.*

See Appendix B.1 for the proof of this lemma.

Combining Lemma 1 and Proposition 1 immediately yields the following performance guarantee for the pessimistic learning rule constructed using Θ_p (which for notational brevity we denote as $\widehat{\pi}_p$).

Theorem 1. *For any $p \geq 1$, with probability at least $1 - \delta$, we have*

$$V(\pi^\star) - V(\widehat{\pi}_p) \leq d^{1/p} \sqrt{\frac{8 \log(d/\delta)}{n}} \cdot \left\| \Sigma_{\mathcal{D}}^{-1/2} \mathbb{E}_{s \sim \rho} \left[\phi(s, \pi^\star(s)) \right] \right\|_q,$$

where q is the solution to $1/p + 1/q = 1$.

Several remarks regarding Theorem 1 are in order. First, the performance upper bound has a natural scaling w.r.t. the sample size n, i.e., $V(\pi^\star) - V(\widehat{\pi}_p) \lesssim \sqrt{1/n}$. In addition, Theorem 1 provides a family of upper bounds for each specific choice of $p \geq 1$. Lastly, from an upper bound perspective, the $\widehat{\pi}_\infty$ learning rule (which we call PUNC) dominates all the other $p \in [1, \infty)$, since for any $v \in \mathbb{R}^d$ and $q \in [1, \infty)$, the inequality $\|v\|_1 \leq d^{1-1/q} \|v\|_q$ holds. This partially showcases the benefits of using PUNC over the alternatives. Later in Section 4, we will see a stronger motivation—from the perspective of the lower bound—for using PUNC, in which we show that PUNC is adaptively minimax optimal. We also remark that the max-min form for $\widehat{\pi}_p$ has an equivalent max-only formulation, which will be helpful for our proofs and comparisons to other algorithms:

$$\widehat{\pi}_p = \arg\max_{\pi \in \Pi} \left\{ \mathbb{E}_{s \sim \rho} \left[\phi(s, \pi(s)) \right]^\top \widehat{\theta}_{\mathsf{ols}} - \tfrac{\beta}{2} \cdot \left\| \Sigma_{\mathcal{D}}^{-1/2} \mathbb{E}_{s \sim \rho} \left[\phi(s, \pi(s)) \right] \right\|_q \right\}. \tag{9}$$

3.3 Connections to prior pessimistic learning rules

Now we present several connections to existing methods used for offline linear contextual bandits.

Connection between $\widehat{\pi}_2$ and Bellman-consistent pessimism. Xie et al. [37] proposed the idea of Bellman-consistent pessimism (BCP) for solving offline reinforcement learning with general function approximation. When specialized to linear contextual bandits, the BCP learning rule first forms a

version space that includes all possible linear reward functions with small ℓ_2 prediction error on the observed datasets. Then, BCP defines each policy's pessimistic value as the smallest value the policy can achieve in the version space. Finally, BCP returns the policy that has the highest pessimistic value. In fact, BCP exactly matches the learning rule $\widehat{\pi}_2$ proposed herein. To see this, it suffices to note that the empirical estimate of the Bellman error (cf. Equation (3.1) in the paper [37]) in the linear contextual bandit case is given by

$$\frac{1}{n}\sum_{i=1}^{n}\left(\phi(s_i,a_i)^\top\theta - r_i\right)^2 - \inf_{\theta\in\mathbb{R}^d}\ \frac{1}{n}\sum_{i=1}^{n}\left(\phi(s_i,a_i)^\top\theta - r_i\right)^2 = \left\|\Sigma_{\mathcal{D}}^{1/2}(\theta-\widehat{\theta}_{\mathsf{ols}})\right\|_2^2.$$

Therefore a parameter θ having a small Bellman error is equivalent to having a small ℓ_2 distance to the OLS estimate. Xie et al. [37] prove that BCP enjoys the guarantee (up to log factors) of $\sqrt{d/n}\cdot\mathbb{E}_{s\sim\rho}\left[\|\Sigma_{\mathcal{D}}^{-1/2}\phi(s,\pi^\star(s))\|_2\right]$, which is loose compared to our theoretical guarantee $\sqrt{d/n}\ \cdot \|\Sigma_{\mathcal{D}}^{-1/2}\mathbb{E}_{s\sim\rho}\phi(s,\pi^\star(s))\|_2$, as a consequence of Jensen's inequality and the convexity of the ℓ_2 norm.

Similar ideas using the ℓ_2 confidence set also appear in a recent paper by Zanette et al. [45]; their actor-critic algorithm, PACLE, can be interpreted as providing a computationally efficient way to solve $\widehat{\pi}_2$.[2]

Connection between PUNC and lower confidence bound for tabular contextual bandits. We now discuss how the LCB learning rule for the tabular setting is a specialization of PUNC. The tabular contextual bandit setting is a special case of the linear setting with the feature mapping $\phi(s,a)=e_{sa}$ (the canonical basis vector indexed by (s,a)). For notational convenience, we define $S\coloneqq|\mathcal{S}|$, $A\coloneqq|\mathcal{A}|$, $\widehat{r}(s,a)$ to be the empirical average reward, and $n(s,a)$ to be the number of times the pair (s,a) is seen in the dataset.

Rashidinejad et al. [24] present the following lower confidence bound (LCB) learning rule:

$$\text{for each } s,\qquad \widehat{\pi}_{\mathsf{LCB}}(s)\coloneqq\arg\max_{a\in\mathcal{A}}\ \widehat{r}(s,a)-\beta\cdot\sqrt{\frac{\log(SA/\delta)}{n(s,a)}}. \tag{10}$$

In essence, the quantity $\widehat{r}(s,a)-\beta\cdot\sqrt{\frac{\log(SA/\delta)}{n(s,a)}}$ acts as a lower confidence bound for the true mean reward $r(s,a)$. In every state, LCB picks the action that maximizes this lower confidence bound. It is easy to verify that LCB (10) exactly corresponds to PUNC (with proper choices of β); one just needs to check the max-only formulation in Equation (9) with $p=\infty$ and $q=1$.

In establishing performance guarantees for LCB, Rashidinejad et al. [24] assume that the covariates $\{(s_i,a_i)\}_{i=1}^n$ are drawn i.i.d. from a behavior distribution $\mu\in\Delta(\mathcal{S}\times\mathcal{A})$ (as opposed to our fixed design setting). Nevertheless, it is straightforward to translate our results to this random design case by using Chernoff bounds.

Corollary 1. *In the tabular setting, with probability at least $1-\delta$, the learning rule $\widehat{\pi}_\infty$ with Θ given by Equation* (8) *achieves the suboptimality:*

$$V(\pi^\star)-V(\widehat{\pi}_\infty)\lesssim\sqrt{\frac{\log(SA/\delta)}{n}}\cdot\left(\sum_s\frac{\rho(s)}{\sqrt{\mu(s,\pi^\star(s))}}\right),$$

as long as $n\gtrsim\log(S/\delta)\cdot(\min_s\{\mu(s,\pi^\star(s))\})^{-1}$.

Corollary 1 is proved in Appendix B.2.

Compared to the upper bound in the paper [24], Corollary 1 is more fine-grained, or "problem-dependent", as the suboptimality bound depends on the interaction between the specific behavior distribution μ and test distribution ρ. In contrast, Rashidinejad et al. [24] consider the class of tabular instances with bounded single-policy concentrability coefficient.

Definition 2. *The single-policy concentrability coefficient is defined as $C^\star\coloneqq\sup_{s\in\mathcal{S}}\frac{\rho(s)}{\mu(s,\pi^\star(s))}$.*

Corollary 1 readily recovers the performance guarantee for LCB of $\tilde{O}(\sqrt{SC^\star/n})$ established in the paper [24], which is optimal in the regime where $C^\star\geq 2$.

[2] While we focus on the statistical properties of the family $\{\widehat{\pi}_p\}$ in this work, we believe that the actor-critic approach developed by Zanette et al. [45] can be extended to yield tractable algorithms for general $p\geq 1$.

Connection to pessimistic value iteration. We give another example of how to interpret pessimistic learning rules using the idea of confidence set construction. Consider the pessimistic value iteration (PEVI) learning rule proposed by Jin et al. [11]. PEVI directly extends Equation (10) to the linear setting:

$$\widehat{\pi}_{\mathsf{PEVI}}(s) \coloneqq \arg\max_{a\in\mathcal{A}} \ \phi(s,a)^\top \widehat{\theta}_{\mathsf{ols}} - \beta \cdot \left\| \Sigma_{\mathcal{D}}^{-1/2} \phi(s,a) \right\|_2, \tag{11}$$

where the right hand side still acts as a lower confidence bound for the true mean reward $r(s,a)$. PEVI bears striking resemblance with the max-only formulation (9) (with $p = q = 2$), with the key difference that the max-only formulation is "averaged" over the test distribution ρ, while PEVI directly discounts every (s,a) pair. PEVI does not immediately fit into our confidence set framework. However, if we modify the minimization over confidence sets to minimization over functionals $\theta : \mathcal{S} \to \mathbb{R}^d$, then we can rewrite PEVI as

$$\widehat{\pi}_{\mathsf{PEVI}} \coloneqq \arg\max_{\pi\in\Pi} \inf_{\theta\in\Theta} \ \mathbb{E}_{s\sim\rho}[\phi(s,\pi(s))^\top \theta(s)],$$
$$\text{where } \Theta = \left\{ s \mapsto \theta(s) \mid \left\| \Sigma_{\mathcal{D}}^{1/2} (\theta(s) - \widehat{\theta}_{\mathsf{ols}}) \right\|_2 \le \beta, \text{ for all } s \right\}.$$

In other words, PEVI enlarges the ℓ_2 confidence set by separately picking a pessimistic parameter $\theta(s)$ for each state $s \in \mathcal{S}$. Jin et al. [11] prove the guarantee (up to log factors) of $\sqrt{d^2/n} \cdot \mathbb{E}_{s\sim\rho}\left[\|\Sigma_{\mathcal{D}}^{-1/2}\phi(s,\pi^\star(s))\|_2\right]$, which is loose due to the extra factor of d and the interchanging of the expectation and the norm. However, their guarantee holds for all test distributions—as opposed to a fixed test distribution ρ. This is a consequence of being pessimistic for every state s.

4 Which learning rule should one use?

Having introduced a general strategy for building pessimistic learning rules by constructing ℓ_p confidence sets, it is natural to ask which $\widehat{\pi}_p$ one should use. To enable such comparisons, we investigate the statistical limits of offline linear contextual bandits over constrained sets of problem instances.

4.1 Minimax lower bound for constrained instances

For any feature mapping $\phi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$, sample size $n \in \mathbb{N}$, and two quantities $q \in [1,\infty)$, $\Lambda > 0$, we define a set of linear contextual bandit (CB) instances[3] as follows:

$$\mathsf{CB}_q(\Lambda) \coloneqq \left\{ (\rho, \{(s_i,a_i)\}_{i=1}^n, \theta^\star, R) \mid \left\| \Sigma_{\mathcal{D}}^{-1/2} \mathbb{E}_{s\sim\rho}[\phi(s,\pi^\star(s))] \right\|_q \le \Lambda, \ R \text{ is 1-subgaussian} \right\}.$$

The set $\mathsf{CB}_q(\Lambda)$ includes all the linear contextual bandit instances such that a sort of "complexity measure" $\mathfrak{C}_q \coloneqq \|\Sigma_{\mathcal{D}}^{-1/2}\mathbb{E}_{s\sim\rho}[\phi(s,\pi^\star(s))]\|_q$ is at most Λ. Our motivation to consider the rate of estimation in the CB family $\mathsf{CB}_q(\Lambda)$ are two-fold. First, in view of Theorem 1, the family $\mathsf{CB}_q(\Lambda)$ admits a good learning rule, specifically $\widehat{\pi}_p$ with $1/p + 1/q = 1$, since for every $\mathcal{Q} \in \mathsf{CB}_q(\Lambda)$, w.p. at least $1-\delta$,

$$V_{\mathcal{Q}}^\star - V_{\mathcal{Q}}(\widehat{\pi}_p) \lesssim d^{1/p}\sqrt{\log(d/\delta)/n} \cdot \Lambda, \tag{12}$$

where $V_{\mathcal{Q}}^\star$ denotes the optimal value in instance $\mathcal{Q}$ and $V_{\mathcal{Q}}(\pi)$ denotes the value of policy π in instance $\mathcal{Q}$. Thus, it is natural to view $\mathfrak{C}_q$ as a certain complexity measure for offline learning in linear contextual bandits. Second, prior work [45, 24, 40] has proven various types of lower bounds on offline learning using either the ℓ_2 quantity $\mathfrak{C}_2$ or the ℓ_1 quantity $\mathfrak{C}_1$. We will elaborate more on this point later.

Now we are ready to present the minimax lower bounds for these families of CB instances.

Theorem 2. *For every $d \ge 2$, there exists a feature mapping ϕ such that the following lower bound holds. For any $p, q \ge 1$ such that $1/p + 1/q = 1$, as long as $\Lambda \ge \sqrt{8} \cdot d^{1/q-1/2}$ and $n \ge d^{2/p}\Lambda^2$, we have*

$$\inf_{\widehat{\pi}} \sup_{\mathcal{Q}\in\mathsf{CB}_q(\Lambda)} \mathbb{E}[V_{\mathcal{Q}}^\star - V_{\mathcal{Q}}(\widehat{\pi})] \ge c \cdot d^{1/p}\sqrt{1/n} \cdot \Lambda,$$

[3] For brevity, we omit the dependence on ϕ and n in the notation $\mathsf{CB}_q(\Lambda)$.

where $c > 0$ is some universal constant. Furthermore, when $p = \infty, q = 1$, the lower bound holds for the extended range of $\Lambda \geq 2$.

The proof can be found in Appendix C. It relies on a reduction to a bound for the minimax regret of the multi-armed bandit problem.

We note that Theorem 2 also consists of a family of lower bounds for each ℓ_q norm constrained CB class. By comparing the lower bound in Theorem 2 with the upper bound (12) obtained by $\widehat{\pi}_p$, we see that for the ℓ_q norm constrained class $\mathsf{CB}_q(\Lambda)$, the learning rule $\widehat{\pi}_p$ with $1/p + 1/q = 1$ is minimax rate-optimal, up to a $\log d$ factor. For instance, over the ℓ_2 class $\mathsf{CB}_2(\Lambda)$, the minimax rate of estimation is $\tilde{\Theta}(\sqrt{d/n} \cdot \Lambda)$, while over the ℓ_1 class $\mathsf{CB}_1(\Lambda)$, the rate is given by $\tilde{\Theta}(\sqrt{1/n} \cdot \Lambda)$.

On a technical front, it would be interesting to extend Theorem 2 to the entire range of $\Lambda \geq 0$. It is unclear whether the same minimax rate of $\Omega(d^{1/p}/\sqrt{n} \cdot \Lambda)$ holds when $\Lambda = O(d^{1/q-1/2})$, or whether we can achieve faster rates in the small Λ regime. In the tabular setting, Rashidinejad et al. [24] recently showed that LCB achieves fast $1/n$ rates when the single policy concentrability coefficient is small; similar results might hold in the linear setting. Several limitations prevent us from extending the range of Λ in Theorem 2; Appendix C.1 provides more technical details.

4.2 Adaptive minimax optimality of PUNC

We point out a even stronger message delivered in Theorem 2: PUNC is *adaptively minimax optimal* for solving the offline linear contextual bandit problem. This is illustrated in Figure 1, where we plot the sample complexity n required in order to achieve constant suboptimality (say, 0.01) for various $\mathsf{CB}_{p/(p-1)}(\Lambda)$. (For sake of illustration, it is more convenient to work with p rather than q on the x-axis.)

As indicated by the red line, Theorem 2 shows that every learning rule must incur sample complexity at least $\Omega(d^{2/p}\Lambda^2)$. Likewise, we can also follow the discussion after Theorem 1 to see that the performance upper bound of $\widehat{\pi}_\infty$ is $d^{1/p}\sqrt{\frac{\log(d/\delta)}{n}} \cdot \mathfrak{C}_q$ for all $p, q \geq 1, 1/p + 1/q = 1$. Thus, PUNC attains the green line in Figure 1; that is, PUNC is *simultaneously* minimax rate-optimal for all ℓ_q-norm constrained classes $\mathsf{CB}_q(\Lambda)$, up to a $\log d$ factor.[4] From worst-case perspective, one should always prefer using PUNC given an unknown CB instance.

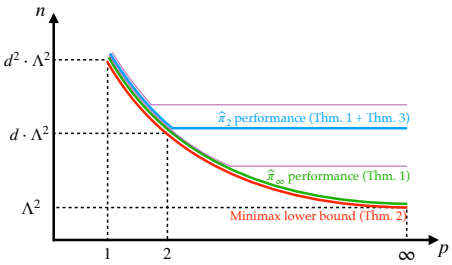


Figure 1: Sample complexity of $\widehat{\pi}_{\widetilde{p}}$ (for various $\widetilde{p}$) over different $\mathsf{CB}_{p/(p-1)}(\Lambda)$ classes. The red line corresponds the minimax lower bound. Other lines correspond to different values of $\widetilde{p}$ and show the number of samples n required to ensure $\sup_{\mathcal{Q} \in \mathsf{CB}_{p/(p-1)}(\Lambda)} \mathbb{E}[V^\star_{\mathcal{Q}} - V_{\mathcal{Q}}(\widehat{\pi})] \leq 0.01$. The blue and green lines correspond to $\widehat{\pi}_2$ and $\widehat{\pi}_\infty$ respectively. Two purple lines correspond to $\widehat{\pi}_{\widetilde{p}}$ for some $\widetilde{p} \in (1, 2)$ and $\widetilde{p} \in (2, \infty)$. PUNC attains minimax optimality over every class, while other $\widehat{\pi}_{\widetilde{p}}$ do not.

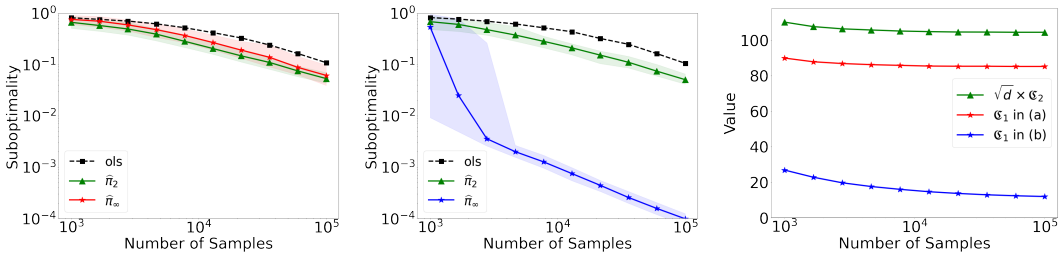
Is this adaptive optimality property unique to PUNC among the family $\{\widehat{\pi}_p\}_{p \geq 1}$ we consider? Below, we answer this question in the positive by presenting a separation result.

Theorem 3 (Informal)**.** *Fix any $p \geq 1$. For sufficiently large n, d, there exists a contextual bandit instance $\mathcal{Q} \in \mathsf{CB}_1(\Lambda)$ with $\Lambda = \sqrt{8d}$, such that with probability at least $1/4$, $\widehat{\pi}_p$ has suboptimality at least $\Omega(d^{1/p}/\sqrt{n} \cdot \Lambda)$.*

Since PUNC attains a suboptimality of $\tilde{O}(1/\sqrt{n} \cdot \Lambda)$ over the class $\mathsf{CB}_1(\Lambda)$, Theorem 3 shows that every other $\widehat{\pi}_p$ is *suboptimal* over the class $\mathsf{CB}_1(\Lambda)$.

A formal statement of Theorem 3 and its proof can be found in Appendix D. The key intuition in the proof is that the ℓ_p confidence sets capture a notion of error that is "averaged" over all directions, while the ℓ_∞ confidence sets separately estimate the error in each direction. In the hard instance we construct, only one direction determines the difficulty of the offline learning problem, so $\widehat{\pi}_p$ is worse

[4] We did not investigate when the $\log d$ factor in Theorem 1 can be removed, so for example, it is possible that $\widehat{\pi}_2$ beats $\widehat{\pi}_\infty$ over $\mathsf{CB}_2(\Lambda)$ by a factor of $\sqrt{\log d}$.



(a) Random rotation (b) Basis-aligned (c) ℓ_1 and ℓ_2 complexities

Figure 2: Comparing the performance of the plug-in rule, $\widehat{\pi}_2$, and $\widehat{\pi}_\infty$ on linear contextual bandit instances with $d = 100$, averaged over 100 trials, with 90% confidence intervals. (a) $\phi_i \sim \mathcal{N}(0, QDQ^\top)$ and $\theta^\star = Qe_{20}$, where Q is a random rotation matrix and D is a diagonal matrix with entries $D_{ii} = i^{-1}/(\sum_i i^{-1})$. (b) $\phi_i \sim \mathcal{N}(0, D)$ and $\theta^\star = e_{20}$. (c) computed average values for $\mathfrak{C}_1$ and $\sqrt{d} \times \mathfrak{C}_2$. The quantity $\mathfrak{C}_2$ is identical in both plots (a) and (b). For (a), $\mathfrak{C}_1 \approx \sqrt{d} \times \mathfrak{C}_2$, while for (b), $\mathfrak{C}_1 \ll \sqrt{d} \times \mathfrak{C}_2$.

by a factor of $d^{1/p}$. There is nothing special about the choice $\Lambda = \sqrt{8d}$, and our construction works for any $\Lambda \geq \Omega(\sqrt{d})$; we pick it to enable comparison with Theorem 2.

For sake of discussion, consider $\widehat{\pi}_2$. Theorem 1 shows that $\widehat{\pi}_2$ attains the rate:

$$V(\pi^\star) - V(\widehat{\pi}_2) \lesssim \begin{cases} d^{1/p}\sqrt{\frac{\log(d/\delta)}{n}} \cdot \left\| \Sigma_{\mathcal{D}}^{-1/2} \mathbb{E}_{s\sim\rho}\left[\phi(s, \pi^\star(s))\right] \right\|_q & \text{when } q \geq 2, \\ \sqrt{\frac{d\log(d/\delta)}{n}} \cdot \left\| \Sigma_{\mathcal{D}}^{-1/2} \mathbb{E}_{s\sim\rho}\left[\phi(s, \pi^\star(s))\right] \right\|_q & \text{when } q \in [1, 2]. \end{cases}$$

Together, Theorem 2 and 3 provide the message that both cases in the upper bound are tight (up to log factors). In the range $p \in [1, 2]$ (or $q \geq 2$), Theorem 2 shows that $\widehat{\pi}_2$ attains the minimax optimal rate (up to log factors) over $\mathsf{CB}_{p/(p-1)}(\Lambda)$, i.e., it is adaptively minimax optimal here. This explains the curved part of the blue line in Figure 1. On the other hand, Theorem 3 shows that $\widehat{\pi}_2$ cannot obtain the minimax rate over $\mathsf{CB}_1(\Lambda)$. Instead, $\widehat{\pi}_2$ may require $\Omega(d \cdot \Lambda^2)$ samples in order to achieve constant suboptimality. Since for any p, the set $\mathsf{CB}_{p/(p-1)}(\Lambda) \supseteq \mathsf{CB}_1(\Lambda)$, we know that $\widehat{\pi}_2$ may require $\Omega(d \cdot \Lambda^2)$ samples for any $\mathsf{CB}_{p/(p-1)}(\Lambda)$. Thus, the second case is tight when $p \geq 2$ (or $q \in [1, 2]$), explaining the flat part of the blue line in Figure 1. In general, for any finite $\widetilde{p}$, the learning rule $\widehat{\pi}_{\widetilde{p}}$ will be adaptively optimal for $\mathsf{CB}_{p/(p-1)}(\Lambda)$ only in the range $p \in [1, \widetilde{p}]$, and afterwards the sample complexity will "flatten out", as illustrated by the purple lines in Figure 1.

Experimental Evidence. In order to further validate this claim, we provide experimental evidence which shows that $\widehat{\pi}_2$ does not adapt to "easy" CB instances. In Figure 2, we consider a simple offline linear contextual bandit in which there is a single state and the feature set is B_2^d; thus the policy learning problem is equivalent to finding a vector $\pi \in \mathbb{S}^{d-1}$ that maximizes $V(\pi) := \pi^\top \theta^\star$. We vary the offline dataset distribution and the hidden parameter $\theta^\star$. When $\theta^\star$ is basis-aligned, we have $\mathfrak{C}_1 \ll \sqrt{d} \times \mathfrak{C}_2$; when $\theta^\star$ is not basis-aligned, the two quantities are on the same order.

4.3 Comparisons with prior lower bounds

There exist several lower bound results for offline reinforcement learning in the literature. In this section, we compare our lower bounds (cf. Theorem 2) with the prior bounds and highlight several improvements offered by our results.

Comparison to lower bounds w.r.t. a single Λ. Our lower bounds are stronger than those provided in the papers [11, 45], which hold for specific choices of $p = q = 2$ and a single fixed Λ. Take Theorem 2 of Zanette et al. [45] for example. Zanette et al. proved that the minimax rate of estimation over $\mathsf{CB}_2(\Lambda = d)$ is given by $d^{3/2}/\sqrt{n}$. Such a lower bound fails to uncover the fundamental scaling on the complexity Λ.[5] Theorem 4.7 of Jin et al. [11] shows a result in similar spirit; their construction essentially shows a minimax lower bound of $1/\sqrt{n}$ over $\mathsf{CB}_2(\Lambda)$ when $\Lambda = \Theta(1)$. Furthermore, their

[5] For instance, their result does not preclude the possibility that the correct lower bound over $\mathsf{CB}_2(\Lambda)$ takes an expression, say, $d^{-98.5}\Lambda^{100}/\sqrt{n}$.

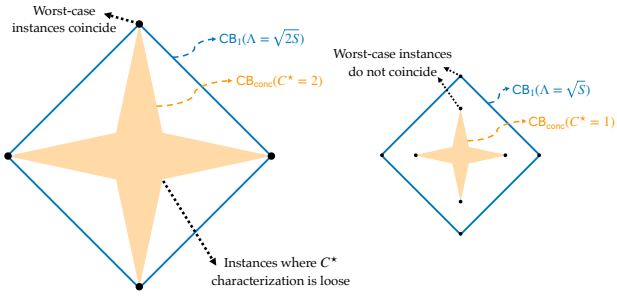


Figure 3: Illustrating the relationship between single policy concentrability and boundedness of $\mathfrak{C}_1$. Left: When $C^\star = 2$, the quantity $\mathfrak{C}_1$ always provides a tighter characterization of the problem difficulty, and the worst-case instances coincide. R: When $C^\star = 1$, the quantity $\mathfrak{C}_1$ does not provide a tight characterization in general.

lower bound is loose by a factor of $\sqrt{d}$ since they reduce to a two-point hypothesis testing problem. In contrast, our lower bound holds for nested families of CB instances with *varying* complexities Λ, which better showcases that the norm quantity is an intrinsic measure of difficulty for offline learning.

Connections with single-policy concentrability. Our lower bound shares a similar flavor as that established in the paper [24], with the key difference lying in the class of CB instances under consideration: Rashidinejad et al. [24] consider the contextual bandit instances $\mathsf{CB}_{\mathrm{conc}}(C^\star)$ with bounded single-policy concentrability coefficient $C^\star$ (cf. Definition 2), while we consider the instances with bounded complexity $\mathfrak{C}_1$. These two quantities are intimately related, and we illustrate the relationship in Figure 3. As we have alluded to in Section 3.3, one has the inclusion

$$\mathsf{CB}_{\mathrm{conc}}(C^\star) \subseteq \mathsf{CB}_1(\sqrt{SC^\star}).$$

When $C^\star \geq 2$, the minimax rate of estimation over $\mathsf{CB}_{\mathrm{conc}}(C^\star)$ exactly matches that over $\mathsf{CB}_1(\sqrt{SC^\star})$, which implies that the hard instances for $\mathsf{CB}_{\mathrm{conc}}(C^\star)$ are also the hard instances in $\mathsf{CB}_1(\sqrt{SC^\star})$. However, this no longer holds when $C^\star \in [1, 2)$. Take $C^\star = 1$ as an example. Rashidinejad et al. show that the optimal rate over $\mathsf{CB}_{\mathrm{conc}}(C^\star = 1)$ is S/n, while Theorem 2 indicates that the optimal rate over $\mathsf{CB}_1(\Lambda = \sqrt{S})$ is $\sqrt{S/n}$. There is no contradiction, since the hard instances we construct for $\mathsf{CB}_1(\sqrt{S})$ satisfy $C^\star \geq 2$. This shows that when $C^\star < 2$, we "lose something" by working with the larger $\mathsf{CB}_1(\sqrt{SC^\star})$ class, as we are no longer able to achieve the fast rates possible over $\mathsf{CB}_{\mathrm{conc}}(C^\star)$.

On the flip side, the quantity $\mathfrak{C}_1$ can give tighter suboptimality guarantees than the $C^\star$ bound for a given instance. Consider the tabular instance where $\rho = \mathrm{Unif}(\mathcal{S})$ and $\mu(1, \pi^\star(1)) = 1/S^3$, while $\mu(s, \pi^\star(s)) = 1/S$ for all $s \geq 2$. This instance has $C^\star = S^2$, implying a guarantee of $S^{3/2}/\sqrt{n}$, while $\mathfrak{C}_1 = O(\sqrt{S})$, implying a better guarantee of $\sqrt{S/n}$.

4.4 A better complexity measure?

Our results lend support to the claim that we should always use PUNC, since it is simultaneously minimax rate-optimal over all the ℓ_q norm-constrained contextual bandit classes. Furthermore, the ℓ_1 quantity $\mathfrak{C}_1$ can be thought of as a "complexity measure" that dominates other ℓ_q "complexity measures" $\mathfrak{C}_q$ for $q > 1$. To see this, consider the following thought experiment. Suppose before solving the linear contextual bandit problem, an oracle told us that the instance satisfies $\mathfrak{C}_q \leq \Lambda$. The results herein show that we do not lose anything by assuming that the instance satisfies the weaker condition $\mathfrak{C}_1 \leq d^{1/p}\Lambda$; using PUNC will give us the optimal rate of $d^{1/p}/\sqrt{n} \cdot \Lambda$.

However this is certainly not the complete answer to guiding question of "which pessimistic learning rule should one use for offline linear contextual bandits?". One piece of evidence comes from the comparison with the single policy concentrability assumption: in the regime where $C^\star \in [1, 2)$, we do "lose something" when we assume the instance satisfies the weaker condition $\mathfrak{C}_1 \leq \sqrt{SC^\star}$. Below we discuss another drawback associated with using $\mathfrak{C}_1$ as the complexity measure.

Rotation ambiguity. One drawback of the complexity $\mathfrak{C}_1$ (as well as the learning rule PUNC) lies in the fact that it is not rotation invariant. (In fact, $\mathfrak{C}_2$ is the only rotational invariant complexity!)

To see this, let $U \in \mathbb{R}^{d \times d}$ be a fixed rotation matrix. Suppose that the features are rotated from ϕ to $U\phi$, which yields a different ℓ_1 complexity $\mathfrak{C}_1(U) := \|U\Sigma_{\mathcal{D}}^{-1/2}\mathbb{E}_{s\sim\rho}[\phi(s,\pi^\star(s))]\|_1$, where $\Sigma_{\mathcal{D}}$ is defined using the old feature mapping. Since the ℓ_1 norm is not rotation invariant, the $\mathfrak{C}_1(U)$ varies for differing choices of U, by up to a $\sqrt{d}$ factor. Thus, we cannot claim that any "complexity measure" $\mathfrak{C}_1(U)$ dominates others. A naive attempt to modify the ℓ_1 set to be rotationally invariant by taking a minimization over U also fails; observe that:

$$\Theta_1^{\min} := \left\{\theta \in \mathbb{R}^d \mid \inf_U \left\|U\Sigma_{\mathcal{D}}^{1/2}(\theta - \widehat{\theta}_{\mathsf{ols}})\right\|_1 \le \beta\right\} = \left\{\theta \in \mathbb{R}^d \mid \left\|\Sigma_{\mathcal{D}}^{1/2}(\theta - \widehat{\theta}_{\mathsf{ols}})\right\|_2 \le \beta\right\} =: \Theta_2,$$

that is, we recover $\widehat{\pi}_2$. A similar equivalence holds if we take the $\max$ inside the confidence set; we will recover the confidence set with an extra factor of $\sqrt{d}$.

Instance-dependent optimality? Arguably, the strongest possible support for PUNC would be an instance-dependent lower bound which shows that for *every specific* linear contextual bandit instance, the performance achieved by PUNC is not improvable. Instance-dependent optimality results have been shown for related problems such as policy evaluation [23, 12] and optimal value estimation [13] in tabular MDPs; the recent work [7] also employs the local minimax method for online bandit and RL problems. For offline bandits, the paper [36] shows how a particular definition of instance optimality cannot be achieved by any algorithm. Establishing instance-dependent guarantees for offline learning is an important direction for future research.

Recent work [40] establishes the local minimax rate for offline learning in terms of the complexity $\mathfrak{C}_1$ for tabular contextual bandits. However, the theorem seems incorrect; we provide a counterexample in Appendix F to demonstrate—via explicit construction—that the complexity $\mathfrak{C}_1$ cannot characterize the local minimax risk for a two-armed bandit instance. The key observation is that the reduction used in the proof of the paper [40] from offline policy learning to optimal value estimation is invalid; if the gap in rewards for different actions is large, offline policy learning is fundamentally easier than optimal value estimation. This in turn allows us to break the claimed parametric $1/\sqrt{n}$ rate.

5 Conclusion

In this paper, we introduce a family $\{\widehat{\pi}_p\}_{p\ge1}$ of pessimistic learning rules that include a number of prior works as special cases for the problem of offline learning in linear contextual bandits. We prove upper bounds for each learning rule $\widehat{\pi}_p$ and show matching minimax lower bounds over appropriately defined constrained instance classes. Our results highlight the benefits of using PUNC, the $\widehat{\pi}_\infty$ learning rule: namely (1) the guarantee for PUNC dominates all others; (2) PUNC is the sole learning rule with an adaptive minimax property. In particular, our results demonstrate that prior learning rules based on ℓ_2 pessimism can be suboptimal (by a factor of $\sqrt{d}$).

Below we list several interesting directions for future investigation.

- *Extending to MDPs.* The MDP setting is more difficult due to the long horizon and transition dynamics. Extending the results of this paper to the MDP setting is an interesting future direction. One possible approach is to modify the PACLE algorithm [45] to solve for any ℓ_p learning rule.
- *Gap-dependent bounds.* In online RL, there is a wealth of results which adapt to easy instances which are characterized by gap structure in the rewards, see, e.g., [6, 32]. Obtaining tight gap-dependent bounds for the offline setting is an interesting direction for future work.
- *Offline RL with general function approximation.* In this paper, we focus on offline RL with linear function approximation. What is the right extension of these ℓ_p learning rules to general function approximation? While $\widehat{\pi}_2$ has the natural interpretation of defining a version space with small squared prediction error, no such interpretation exists for PUNC. It would be interesting to establish an analog for PUNC for general function classes.

Acknowledgments and Disclosure of Funding

This work is supported by funding from the Institute for Data, Econometrics, Algorithms, and Learning (IDEAL).

References

[1] Léon Bottou, Jonas Peters, Joaquin Quiñonero-Candela, Denis X Charles, D Max Chickering, Elon Portugaly, Dipankar Ray, Patrice Simard, and Ed Snelson. Counterfactual reasoning and learning systems: The example of computational advertising. *Journal of Machine Learning Research*, 14(11), 2013.

[2] Jinglin Chen and Nan Jiang. Information-theoretic considerations in batch reinforcement learning. *arXiv preprint arXiv:1905.00360*, 2019.

[3] Jinglin Chen and Nan Jiang. Offline reinforcement learning under value and density-ratio realizability: the power of gaps. *arXiv preprint arXiv:2203.13935*, 2022.

[4] Ching-An Cheng, Tengyang Xie, Nan Jiang, and Alekh Agarwal. Adversarially trained actor critic for offline reinforcement learning. *arXiv preprint arXiv:2202.02446*, 2022.

[5] Yaqi Duan, Chi Jin, and Zhiyuan Li. Risk bounds and rademacher complexity in batch reinforcement learning. *arXiv preprint arXiv:2103.13883*, 2021.

[6] Dylan J Foster, Alexander Rakhlin, David Simchi-Levi, and Yunzong Xu. Instance-dependent complexity of contextual bandits and reinforcement learning: A disagreement-based perspective. *arXiv preprint arXiv:2010.03104*, 2020.

[7] Dylan J Foster, Sham M Kakade, Jian Qian, and Alexander Rakhlin. The statistical complexity of interactive decision making. *arXiv preprint arXiv:2112.13487*, 2021.

[8] Dylan J Foster, Akshay Krishnamurthy, David Simchi-Levi, and Yunzong Xu. Offline reinforcement learning: Fundamental barriers for value function approximation. *arXiv preprint arXiv:2111.10919*, 2021.

[9] Scott Fujimoto, David Meger, and Doina Precup. Off-policy deep reinforcement learning without exploration. In *International Conference on Machine Learning*, pages 2052–2062. PMLR, 2019.

[10] Natasha Jaques, Asma Ghandeharioun, Judy Hanwen Shen, Craig Ferguson, Agata Lapedriza, Noah Jones, Shixiang Gu, and Rosalind Picard. Way off-policy batch deep reinforcement learning of implicit human preferences in dialog. *arXiv preprint arXiv:1907.00456*, 2019.

[11] Ying Jin, Zhuoran Yang, and Zhaoran Wang. Is pessimism provably efficient for offline rl? In *International Conference on Machine Learning*, pages 5084–5096. PMLR, 2021.

[12] Koulik Khamaru, Ashwin Pananjady, Feng Ruan, Martin J Wainwright, and Michael I Jordan. Is temporal difference learning optimal? an instance-dependent analysis. *arXiv preprint arXiv:2003.07337*, 2020.

[13] Koulik Khamaru, Eric Xia, Martin J Wainwright, and Michael I Jordan. Instance-optimality in optimal value estimation: Adaptivity via variance-reduced q-learning. *arXiv preprint arXiv:2106.14352*, 2021.

[14] Rahul Kidambi, Aravind Rajeswaran, Praneeth Netrapalli, and Thorsten Joachims. MOReL: Model-based offline reinforcement learning. *arXiv preprint arXiv:2005.05951*, 2020.

[15] Aviral Kumar, Justin Fu, George Tucker, and Sergey Levine. Stabilizing off-policy Q-learning via bootstrapping error reduction. *arXiv preprint arXiv:1906.00949*, 2019.

[16] Aviral Kumar, Aurick Zhou, George Tucker, and Sergey Levine. Conservative q-learning for offline reinforcement learning. *Advances in Neural Information Processing Systems*, 33: 1179–1191, 2020.

[17] Sascha Lange, Thomas Gabel, and Martin Riedmiller. Batch reinforcement learning. In *Reinforcement learning*, pages 45–73. Springer, 2012.

[18] Sergey Levine, Aviral Kumar, George Tucker, and Justin Fu. Offline reinforcement learning: Tutorial, review, and perspectives on open problems. *arXiv preprint arXiv:2005.01643*, 2020.

[19] Gen Li, Laixi Shi, Yuxin Chen, Yuejie Chi, and Yuting Wei. Settling the sample complexity of model-based offline reinforcement learning. *arXiv preprint arXiv:2204.05275*, 2022.

[20] Lihong Li, Wei Chu, John Langford, and Robert E Schapire. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th international conference on World wide web*, pages 661–670, 2010.

[21] Yao Liu, Adith Swaminathan, Alekh Agarwal, and Emma Brunskill. Provably good batch reinforcement learning without great exploration. *arXiv preprint arXiv:2007.08202*, 2020.

[22] Horia Mania, Stephen Tu, and Benjamin Recht. Certainty equivalence is efficient for linear quadratic control. *Advances in Neural Information Processing Systems*, 32, 2019.

[23] Ashwin Pananjady and Martin J Wainwright. Instance-dependent ℓ_∞-bounds for policy evaluation in tabular reinforcement learning. *IEEE Transactions on Information Theory*, 67(1): 566–585, 2020.

[24] Paria Rashidinejad, Banghua Zhu, Cong Ma, Jiantao Jiao, and Stuart Russell. Bridging offline reinforcement learning and imitation learning: A tale of pessimism. *arXiv preprint arXiv:2103.12021*, 2021.

[25] Laixi Shi, Gen Li, Yuting Wei, Yuxin Chen, and Yuejie Chi. Pessimistic q-learning for offline reinforcement learning: Towards optimal sample complexity. *arXiv preprint arXiv:2202.13890*, 2022.

[26] Max Simchowitz and Dylan Foster. Naive exploration is optimal for online lqr. In *International Conference on Machine Learning*, pages 8937–8948. PMLR, 2020.

[27] Alex Strehl, John Langford, Sham Kakade, and Lihong Li. Learning from logged implicit exploration data. *arXiv preprint arXiv:1003.0120*, 2010.

[28] Adith Swaminathan and Thorsten Joachims. Counterfactual risk minimization: Learning from logged bandit feedback. In *International Conference on Machine Learning*, pages 814–823. PMLR, 2015.

[29] Adith Swaminathan and Thorsten Joachims. The self-normalized estimator for counterfactual learning. *advances in neural information processing systems*, 28, 2015.

[30] Masatoshi Uehara and Wen Sun. Pessimistic model-based offline reinforcement learning under partial coverage. *arXiv preprint arXiv:2107.06226*, 2021.

[31] Masatoshi Uehara, Xuezhou Zhang, and Wen Sun. Representation learning for online and offline rl in low-rank mdps. *arXiv preprint arXiv:2110.04652*, 2021.

[32] Andrew Wagenmaker and Kevin Jamieson. Instance-dependent near-optimal policy identification in linear mdps via online experiment design. *arXiv preprint arXiv:2207.02575*, 2022.

[33] Ruosong Wang, Dean P Foster, and Sham M Kakade. What are the statistical limits of offline rl with linear function approximation? *arXiv preprint arXiv:2010.11895*, 2020.

[34] Yifan Wu, George Tucker, and Ofir Nachum. Behavior regularized offline reinforcement learning. *arXiv preprint arXiv:1911.11361*, 2019.

[35] Yihong Wu. Lecture notes on: information-theoreticl methods for high-dimensional statistics. 2020.

[36] Chenjun Xiao, Yifan Wu, Jincheng Mei, Bo Dai, Tor Lattimore, Lihong Li, Csaba Szepesvari, and Dale Schuurmans. On the optimality of batch policy optimization algorithms. In *International Conference on Machine Learning*, pages 11362–11371. PMLR, 2021.

[37] Tengyang Xie, Ching-An Cheng, Nan Jiang, Paul Mineiro, and Alekh Agarwal. Bellman-consistent pessimism for offline reinforcement learning. *arXiv preprint arXiv:2106.06926*, 2021.

[38] Tengyang Xie, Nan Jiang, Huan Wang, Caiming Xiong, and Yu Bai. Policy finetuning: Bridging sample-efficient offline and online reinforcement learning. *Advances in neural information processing systems*, 34, 2021.

[39] Yuling Yan, Gen Li, Yuxin Chen, and Jianqing Fan. The efficacy of pessimism in asynchronous q-learning. *arXiv preprint arXiv:2203.07368*, 2022.

[40] Ming Yin and Yu-Xiang Wang. Towards instance-optimal offline reinforcement learning with pessimism. *arXiv preprint arXiv:2110.08695*, 2021.

[41] Ming Yin, Yaqi Duan, Mengdi Wang, and Yu-Xiang Wang. Near-optimal offline reinforcement learning with linear representation: Leveraging variance information with pessimism. *arXiv preprint arXiv:2203.05804*, 2022.

[42] Tianhe Yu, Garrett Thomas, Lantao Yu, Stefano Ermon, James Zou, Sergey Levine, Chelsea Finn, and Tengyu Ma. MOPO: Model-based offline policy optimization. *arXiv preprint arXiv:2005.13239*, 2020.

[43] Ekim Yurtsever, Jacob Lambert, Alexander Carballo, and Kazuya Takeda. A survey of autonomous driving: Common practices and emerging technologies. *IEEE Access*, 8:58443–58469, 2020.

[44] Andrea Zanette. Exponential lower bounds for batch reinforcement learning: Batch rl can be exponentially harder than online rl. In *International Conference on Machine Learning*, pages 12287–12297. PMLR, 2021.

[45] Andrea Zanette, Martin J Wainwright, and Emma Brunskill. Provable benefits of actor-critic methods for offline reinforcement learning. *arXiv preprint arXiv:2108.08812*, 2021.

[46] Wenhao Zhan, Baihe Huang, Audrey Huang, Nan Jiang, and Jason D Lee. Offline reinforcement learning with realizability and single-policy concentrability. *arXiv preprint arXiv:2202.04634*, 2022.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes]
   (c) Did you discuss any potential negative societal impacts of your work? [N/A] This is a theoretical work which does not have immediate societal impacts.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] We describe the experiment we ran in sufficient detail; it can be replicated in a few dozen lines of code.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [N/A] Experiments were ran on a laptop.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [N/A]
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

A Additional literature

In this section, we mention a few additional related works which are not discussed in the main text.

First, we mention that offline contextual bandits have an extensive history dating back to early work in recommender systems [20, 1]. A popular approach is to use importance-reweighting to estimate the values of policies [27, 1, 28, 29]. A key aspect of this approach is that it requires one to estimate or know the behavior policy which generated the offline data. In contrast, our work relies on the principle of pessimism and directly bounds the suboptimality in terms of the quality of the data coverage.

We highlight additional works which employ the principle of pessimism in order to address the issue of dataset coverage in offline RL. In the tabular setting, several additional works study the single policy concentrability assumption [38, 25, 40, 39, 19]. Several additional works study offline RL with function approximation, e.g., [41, 30, 31, 4, 46, 3].

We also highlight several lower bounds for the offline RL problem. Most relevant to our work are lower bounds for the single policy concentrability assumption in the tabular setting [24, 38]. In addition, there is a line of work for offline RL with function approximation which studies the interplay between data coverage and representation power [2, 8, 44, 33]. These results generally show lower bounds with exponential dependence on H. Since we focus on the linear contextual bandit setting ($H = 1$), these results are tangential to our discussion.

B Proofs for Section 3

This section gathers the proofs for the results in Section 3.

B.1 Proof of Lemma 1

We decompose the proof into two steps: (1) proving $\theta^\star \in \Theta_p$, and (2) proving $\sup_{\theta\in\Theta} \|\theta^\star - \theta\| \le \beta$.

Step 1: proving $\theta^\star \in \Theta_p$. It is easy to see that $\Sigma_{\mathcal{D}}^{1/2}(\theta^\star - \widehat{\theta}_{\mathsf{ols}}) = \frac{1}{\sqrt{n}}(\Phi^\top\Phi)^{-1/2}\Phi^\top\eta$, where $\eta \in \mathbb{R}^n$ has i-th entry equal to $r_i - \mathbb{E}[R(s_i, a_i)]$. Since each row of $(\Phi^\top\Phi)^{-1/2}\Phi^\top$ has unit ℓ_2 norm, $[(\Phi^\top\Phi)^{-1/2}\Phi^\top\eta]_j$ is a 1-subgaussian random variable for each $1 \le j \le d$. By the standard tail bound for subgaussian random variables and the union bound, we have

$$\mathbb{P}_\eta\left[\sup_{1\le j\le d}[(\Phi^\top\Phi)^{-1/2}\Phi^\top\eta]_j \ge \sqrt{2\log\frac{d}{\delta}}\right] \le \delta.$$

As a result, we have with probability at least $1-\delta$

$$\left\|\Sigma_{\mathcal{D}}^{1/2}(\theta^\star - \widehat{\theta}_{\mathsf{ols}})\right\|_p = \frac{1}{\sqrt{n}}\left\|(\Phi^\top\Phi)^{-1/2}\Phi^\top\eta\right\|_p \le d^{1/p}\sqrt{\frac{2\log(d/\delta)}{n}},$$

which implies $\theta^\star \in \Theta_p$.

Step 2: proving $\sup_{\theta\in\Theta} \|\theta^\star - \theta\| \le \beta$. By the triangle inequality, we have

$$\sup_{\theta\in\Theta}\|\theta^\star - \theta\| \le \sup_{\theta\in\Theta}\left\|\widehat{\theta}_{\mathsf{ols}} - \theta\right\| + \left\|\widehat{\theta}_{\mathsf{ols}} - \theta^\star\right\| \le \frac{\beta}{2} + \frac{\beta}{2} = \beta.$$

Here the second inequality relies on the definition of Θ, and the consequence $\theta_\star \in \Theta$ of Step 1.

Combining both steps finishes the proof of this lemma.

B.2 Proof of Corollary 1

We state and prove a more general version of Corollary 1 that provides a guarantee for each $\widehat{\pi}_p$.

Corollary 2. *In the tabular setting, the learning rule $\widehat{\pi}_p$ with Θ given by Equation* (8) *achieves the suboptimality with probability at least $1-\delta$:*

$$V(\pi^\star) - V(\widehat{\pi}_p) \lesssim (SA)^{1/p}\sqrt{\frac{\log(SA/\delta)}{n}} \cdot \left(\sum_s \frac{\rho^q(s)}{(n(s,\pi^\star(s))/n)^{q/2}}\right)^{1/q}.$$

as long as $n \gtrsim \log(S/\delta) \cdot (\min_s\{\mu(s,\pi^\star(s))\})^{-1}$.

Proof. We specialize Theorem 1 to the tabular setting. In the tabular setting, we have $d = SA$. The empirical second moment matrix is a diagonal matrix with values $n(s,a)/n$. The vector $\mathbb{E}_{s\sim\rho}[\phi(s,\pi^\star(s))]$ takes the value in the (s,a)-th coordinate of $\rho(s)\mathbb{1}\{a=\pi^\star(s)\}$. Thus, we can derive the guarantee that with probability at least $1-\delta/2$,

$$V(\pi^\star) - V(\widehat{\pi}_p) \lesssim (SA)^{1/p}\sqrt{\frac{\log(SA/\delta)}{n}} \cdot \left(\sum_s \frac{\rho^q(s)}{(n(s,\pi^\star(s))/n)^{q/2}}\right)^{1/q}.$$

In order to convert this to a guarantee in terms of the behavior distribution, we use the following lemma.

Lemma 2. *If $n \geq 8\log(SA/\delta) \cdot (\min_s\{\mu(s,\pi^\star(s))\})^{-1}$, then with probability at least $1-\delta$:*

$$n(s,\pi^\star(s)) \geq \frac{n \cdot \mu(s,\pi^\star(s))}{2}, \text{ for all } s.$$

Applying the lemma concludes the proof of the first part of the statement. □

It remains to prove the lemma.

Proof of Lemma 2. Fix any $s \in \mathcal{S}$. Using Chernoff bounds, one has

$$\mathbb{P}\left[n(s,\pi^\star(s)) \leq \frac{n}{2} \cdot \mu(s,\pi^\star(s)\right] \leq \exp(-n\mu(s,\pi^\star(s))/8).$$

Thus we have by union bound,

$$\mathbb{P}\left[\text{exists } s : n(s,\pi^\star(s)) \leq \frac{n}{2} \cdot \mu(s,\pi^\star(s))\right] \leq S \cdot \exp(-n\mu_{\min}/8),$$

where $\mu_{\min} := \min_s\{\mu(s,\pi^\star(s))\}$. By setting the RHS to δ, we prove the result. □

B.3 Tighter analysis for $\widehat{\pi}_2$ in the tabular setting

In this section, we point out an issue with adapting the general result of Theorem 1 to the tabular setting. This reveals a flaw with pessimism-validity, namely the requirement that $\theta^\star \in \Theta$.

Consider the single policy concentrability assumption of Definition 2. It is straightforward to compute that for any $q \geq 1$, we have:

$$\sum_s \frac{\rho^q(s)}{\mu^{q/2}(s,\pi^\star(s))} \leq (C^\star)^{q/2} \cdot \sum_s \rho^{q/2}(s) \leq (C^\star)^{q/2} \cdot S^{1-q/2}.$$

Therefore, in the tabular setting, the $\widehat{\pi}_p$ recovers the guarantee of

$$V(\pi^\star) - V(\widehat{\pi}_p) \lesssim A^{1/p}\sqrt{\frac{SC^\star \log(SA/\delta)}{n}}.$$

Therefore, under the single policy concentrability assumption, all $\widehat{\pi}_p$ attain the minimax-optimal rate of $\tilde{O}(\sqrt{SC^\star/n})$ in the regime $C^\star \geq 2$ [24], up to the additional dependence on $A^{1/p}$. The polynomial dependence on A can be removed by considering a smaller confidence set Θ for the tabular setting and directly analyzing the suboptimality. We illustrate how to do this for $p=2$; it is trivial to extend to all p.

Proposition 2. *Fix any $\delta \in (0, 1/2)$. In the tabular setting, the learning rule $\widehat{\pi}_2$ configured with $\beta \coloneqq \sqrt{\frac{16S\log(SA/\delta)}{n}}$ yields the suboptimality with probability at least $1-\delta$:*

$$V(\pi^\star) - V(\widehat{\pi}_2) \lesssim \sqrt{\frac{SC^\star \log(SA/\delta)}{n}},$$

as long as $n \gtrsim \log(S/\delta) \cdot (\min_s\{\mu(s, \pi^\star(s))\})^{-1}$.

Before we prove the theorem, we remark that this observation implies that the proof techniques relying on assuming $\theta^\star \in \Theta$ (i.e., the results in this paper, as well as extensions to general function classes [37]) can be *fundamentally loose*. Indeed one can check that we will not have $\theta^\star \in \Theta$ for the above learning rule with high probability. We leave developing a tighter analysis of pessimism via confidence sets that bypasses this assumption for general function approximation to future work.

Proof of Proposition 2. By Equation (9), we can write the $\widehat{\pi}_2$ in the tabular setting as follows:

$$\widehat{\pi}_2 \coloneqq \arg\max_{\pi\in\Pi} \widehat{V}(\pi), \quad \text{where } \widehat{V}(\pi) = \left(\sum_s \rho(s)\widehat{r}(s,\pi(s))\right) - \frac{\beta}{2}\cdot\sqrt{\sum_s \frac{\rho^2(s)}{n(s,\pi(s))/n}}.$$

We will directly bound the suboptimality in view of Equation (4). We claim that

$$V(\pi^\star) - \widehat{V}(\pi^\star) \lesssim \sqrt{\frac{SC^\star\log(SA/\delta)}{n}}; \tag{13a}$$

$$\widehat{V}(\widehat{\pi}_2) - V(\widehat{\pi}_2) \le 0. \tag{13b}$$

We will prove the two inequalities conditioned on the following event:

$$\mathcal{E} \coloneqq \left\{\text{for all } s\in\mathcal{S}, \quad r(s,\pi^\star(s)) - \widehat{r}(s,\pi^\star(s)) \le \sqrt{\frac{4\log(SA/\delta)}{n(s,\pi^\star(s))}}\right\}.$$

By Hoeffding's inequality and union bound, with probability at least $1-\delta$, $\mathcal{E}$ holds.

Proof of Equation (13a). We compute the bound that

$$\begin{aligned}
V(\pi^\star) - \widehat{V}(\pi^\star) &= \sum_s \rho(s)\left(r(s,\pi^\star(s)) - \widehat{r}(s,\pi^\star(s))\right) + \frac{\beta}{2}\cdot\sqrt{\sum_s \frac{\rho^2(s)}{n(s,\pi^\star(s))/n}} \\
&\overset{(i)}{\le} \sum_s \rho(s)\sqrt{\frac{4\log(SA/\delta)}{n(s,\pi^\star(s))}} + \frac{\beta}{2}\cdot\sqrt{\sum_s \frac{\rho^2(s)}{n(s,\pi^\star(s))/n}} \\
&\overset{(ii)}{\le} \sqrt{\sum_s \rho^2(s)\cdot\frac{4S\log(SA/\delta)}{n(s,\pi^\star(s))}} + \frac{\beta}{2}\cdot\sqrt{\sum_s \frac{\rho^2(s)}{n(s,\pi^\star(s))/n}} \\
&\overset{(iii)}{=} \sqrt{\sum_s \rho^2(s)\cdot\frac{16S\log(SA/\delta)}{n(s,\pi^\star(s))}}.
\end{aligned}$$

Inequality (i) follows by Hoeffding's inequality, (ii) is due to Cauchy-Schwarz, and (iii) is the definition of β. After applying Chernoff bounds and applying the assumption of single policy concentrability, we see that with probability at least $1-\delta$,

$$V(\pi^\star) - \widehat{V}(\pi^\star) \lesssim \sqrt{\frac{SC^\star\log(SA/\delta)}{n}}.$$

Proof of Equation (13b). The proof for this inequality follows a similar outline. Fix any policy π. We can compute that with probability at least $1-\delta$:

$$\begin{aligned}
\widehat{V}(\pi) - V(\pi) &= \sum_s \rho(s)\left(\widehat{r}(s,\pi(s)) - r(s,\pi(s))\right) - \frac{\beta}{2}\cdot\sqrt{\sum_s \frac{\rho^2(s)}{n(s,\pi(s))/n}} \\
&\le \sqrt{\sum_s \rho^2(s)\cdot\frac{4S\log(SA/\delta)}{n(s,\pi(s))}} - \frac{\beta}{2}\cdot\sqrt{\sum_s \frac{\rho^2(s)}{n(s,\pi(s))/n}} \le 0,
\end{aligned}$$

again using Hoeffding's inequality, Cauchy-Schwarz, and the definition of β.

By combining the two inequalities we prove the proposition. □

C Proof of Theorem 2

The proof of the lower bound relies on a reduction to a bound on the minimax regret of the offline multi-armed bandit problem [36]. We will utilize the following lemma, which lower bounds the minimax regret for tabular contextual bandit instances.

Lemma 3. *Let S be arbitrary and $A \geq 2$. Fix any test distribution $\rho \in \Delta(\mathcal{S})$ and counts $\mathbf{n} = \{n(s,a)\}_{s\in\mathcal{S}, a\in\mathcal{A}}$. Define the set of contextual bandit instances:*

$$\mathsf{CB}_{\rho,\mathbf{n}} := \{R : R(s,a) \text{ is 1-subgaussian for all } (s,a)\}.$$

Then there exists a universal constant $c > 0$ such that

$$\inf_{\widehat{\pi}} \sup_{\mathcal{Q}\in\mathsf{CB}_{\rho,\mathbf{n}}} \mathbb{E}[V^\star_{\mathcal{Q}} - V_{\mathcal{Q}}(\widehat{\pi})] \geq c \cdot \sum_s \left(\rho(s) \max_{a\in\{a^{(1)},a^{(2)}\}} \frac{1}{\sqrt{n(s,a)}} \right). \tag{14}$$

Proof. Theorem 1 of Xiao et al. [36] proves a lower bound on the Bayes suboptimality for any multi-armed bandit instance. Specifically, they show that for any A-armed bandit problem, one can define a collection of instances $\mathcal{B}$; for any sequence of counts $\mathbf{n} = \{n(a)\}_{a\in\mathcal{A}}$, the Bayes suboptimality is at least:

$$\inf_{\widehat{\pi}} \mathbb{E}_{\mathcal{Q}\sim\mathrm{Unif}(\mathcal{B})}[r^\star_{\mathcal{Q}} - r_{\mathcal{Q}}(\widehat{\pi})] \gtrsim \max_{a\in\mathcal{A}} \frac{1}{\sqrt{n(a)}}. \tag{15}$$

In order to prove a guarantee for the *tabular contextual bandit problem*, we will tensorize this result. Intuitively, we can treat the estimation of the policy for each state as a separate multi-armed bandit policy estimation problem. Formally, the Bayes suboptimality bound of Equation (15) allows us to do this. Note that the suboptimality for any policy π can be written as

$$V^\star_{\mathcal{Q}} - V_{\mathcal{Q}}(\pi) = \sum_s \rho(s) \cdot (r_{\mathcal{Q}}(s,\pi^\star) - r_{\mathcal{Q}}(s,\pi)).$$

Therefore, we can lower bound the minimax suboptimality of the contextual bandit problem as the sum of the Bayes suboptimalities for the S individual bandit problems:

$$\inf_{\widehat{\pi}} \sup_{\mathcal{Q}\in\mathsf{CB}_{\rho,\mathbf{n}}} \mathbb{E}[V^\star_{\mathcal{Q}} - V_{\mathcal{Q}}(\widehat{\pi})] \geq \sum_s \rho(s) \cdot \left(\inf_{\widehat{\pi}_s} \mathbb{E}_{\mathcal{Q}_s\sim\mathrm{Unif}(\mathcal{B}_s)} \left[r^\star_{\mathcal{Q}_s} - r_{\mathcal{Q}_s}(\widehat{\pi}_s) \right] \right).$$

Here, $\widehat{\pi}_s$ is taken to be any estimation procedure that looks only at the subset of the dataset over state s; the collection $\mathcal{B}_s$ is the set of bandit instances defined by Xiao et al. [36] (for the counts $\{n(s,a)\}_{a\in\mathcal{A}}$). The inequality follows by the lower bound for the minimax risk of a tensor product (see, e.g., [35], Theorem 3.1). The proof concludes by applying the bounds on the Bayes risk for each state in (15) to the previous display. □

With Lemma 3 in hand, we now prove the theorem.

Proof of Theorem 2. The family of hard instances we construct are tabular contextual bandits. We begin by describing the state and action spaces. Define $S = d/2$ (in the case where d is odd, we set $S = \lfloor d/2 \rfloor$ and add a "dummy" coordinate to the tabular feature mapping, which does not affect the rest of the proof). We set $\mathcal{S} = \{1, 2, \ldots, S\}$ and $\mathcal{A} = \{a^{(1)}, a^{(2)}\}$. Thus, we use the tabular feature mapping, i.e., $\phi(s,a) = e_{s,a}$.

In view of Lemma 3, our task is as follows. For each tuple (Λ, q, n) we need to supply a test distribution ρ and counts $\mathbf{n}$ which satisfies two properties:

1. $\mathsf{CB}_{\rho,\mathbf{n}} \subseteq \mathsf{CB}_q(\Lambda)$. In other words, we need ρ and $\mathbf{n}$ to satisfy the inequality

$$\|\Sigma_{\mathcal{D}}^{-1/2} \mathbb{E}_{s\sim\rho}[\phi(s,\pi^\star_v(s))]\|_q = \left(\sum_s \rho^q(s) \max_{a\in\{a^{(1)},a^{(2)}\}} \frac{1}{(n(s,a)/n)^{q/2}} \right)^{1/q} \leq \Lambda.$$

2. The RHS of Equation (14) is sufficiently large:

$$\sum_s \rho(s) \max_{a \in \{a^{(1)}, a^{(2)}\}} \frac{1}{\sqrt{n(s,a)/n}} \gtrsim d^{1/p}\Lambda.$$

Hard instance construction. Fix any value of $\Lambda \geq \sqrt{8} d^{1/2-1/p}$. We will utilize the construction from Rashidinejad et al. [24]. Let us set:

$$\rho = \mathrm{Unif}(\{1, \ldots, S\}), \quad \text{and} \quad \begin{array}{l} n(s, a^{(1)}) = \lfloor n/S \cdot (1/\Gamma) \rfloor, \\ n(s, a^{(2)}) = \lfloor n/S \cdot (1 - 1/\Gamma) \rfloor. \end{array}$$

The parameter Γ is set to be $\Gamma = S^{2/p-1}\Lambda^2/2$; without loss of generality we will require $\Gamma \geq 2$ so that $n(s, a^{(1)}) \leq n(s, a^{(2)})$. Note that as long as $n \geq S^{2/p}\Lambda^2 = 2S\Gamma$ (which holds whenever $n \geq d^{2/p}\Lambda^2$), we have $n(s, a^{(1)}) \geq 1/2 \cdot n/S \cdot (1/\Gamma)$ and $n(s, a^{(2)}) \geq 1/2 \cdot n/S \cdot (1 - 1/\Gamma)$. It is easy to calculate the bounds that:

$$\sum_s \rho^q(s) \max_{a \in \{a^{(1)}, a^{(2)}\}} \frac{1}{(n(s,a)/n)^{q/2}} \leq \Lambda^q,$$

$$\sum_s \rho(s) \max_{a \in \{a^{(1)}, a^{(2)}\}} \frac{1}{\sqrt{n(s,a)/n}} \geq S^{1/p}\Lambda/\sqrt{2} \gtrsim d^{1/p}\Lambda,$$

thus the conditions are satisfied. Lastly, we note that in order to ensure $\Gamma \geq 2$, we need $\Lambda \geq 2S^{1/2-1/p}$; using the fact that $S = d/2$, a sufficient condition is $\Lambda \geq \sqrt{8} d^{1/2-1/p}$. By reduction to the contextual bandit lower bound, this proves the result.

Improving the range of Λ for $q = 1$. We show how we can prove the theorem for an extended range of Λ when $q = 1$. Fix any value of $\Lambda \geq 2$. We will slightly tweak the previous construction to be T-sparse for some $T \in \{1, \ldots, S\}$. Let us set:

$$\rho = \mathrm{Unif}(\{1, \ldots, T\}), \quad \text{and} \quad \begin{array}{l} n(s, a^{(1)}) = \lfloor n/T \cdot \Gamma \rfloor, \\ n(s, a^{(2)}) = \lfloor n/T \cdot (1 - 1/\Gamma) \rfloor. \end{array}$$

We will pick parameters T and Γ such that $\Gamma \geq 2$ and $T := \lfloor \Lambda^2/(2\Gamma) \rfloor$. As long as $\Lambda \geq 2$, such a choice is indeed valid because when $\Gamma = 2$ (the smallest possible choice for Γ), we have $\Lambda^2/\Gamma \geq 1 \Rightarrow T \geq 1$; also, we can always pick Γ sufficiently large so that $T \leq S$. In addition, we note that under the sample complexity requirement $n \geq \Lambda^2 \geq 2T\Gamma$, we have $n(s, a^{(1)}) \geq 1/2 \cdot n/T \cdot 1/\Gamma$ and $n(s, a^{(2)}) \geq 1/2 \cdot n/T \cdot (1 - 1/\Gamma)$.

Now it remains to verify the conditions. We calculate that

$$\Lambda \lesssim \sqrt{\Gamma T} \leq \sum_s \rho(s) \max_{a \in \{a^{(1)}, a^{(2)}\}} \frac{1}{\sqrt{n(s,a)/n}} \leq \Lambda.$$

Thus, in the case where $p = \infty$, $q = 1$, we have proven that the instance satisfies both conditions. □

C.1 Limitations of the lower bound

We discuss in more detail the technical challenges with strengthening the lower bound by proving that the condition holds over a larger range of Λ. Since we argue that Λ roughly corresponds to the difficulty of the offline policy learning problem, it would be desirable to show that the lower bound still holds even when Λ is small. We are able to do this for the $p = \infty, q = 1$ case, but extending this construction to finite p seems challenging.

Fundamental limitation of tabular design. First, we observe that there is already a fundamental limitation in the reduction to contextual bandits; namely, we cannot hope to prove Theorem 2 for any value of complexity parameter $\Lambda \leq O(1/d^{1/p})$. Observe that for any tabular contextual bandit, the key ℓ_1 quantity takes the form:

$$\sum_s \rho(s) \max_{a \in \{a^{(1)}, a^{(2)}\}} \frac{1}{\sqrt{n(s,a)/n}}.$$

Since $\rho \in \Delta(\mathcal{S})$ and $n(s,a) \leq n$ for all (s,a) pairs, this quantity must be $\Omega(1)$. Using the identity $\|v\|_1 \leq d^{1/p} \|v\|_q$, we see that it is not even possible to construct tabular contextual bandit instances unless $\Lambda \geq \Omega(d^{-1/p})$. Thus we must look beyond tabular lower bound constructions.

Extending Λ range for $q \in (1,2]$. We provide an example of how extend the lower bound to a larger range of Λ when $q \in (1,2]$ using a more complicated tabular construction. The catch is that the sample complexity increases from $n \geq d^{2/p}\Lambda^2$ to $n \geq d^{2/q}\Lambda^2$. We conjecture that this is an artifact of our analysis, and that it is possible to prove the lower bound under the requirement $n \geq d^{2/p}\Lambda^2$ (which would allow us to show a lower bound that takes all values $d^{1/p}\Lambda/\sqrt{n} \in (0, O(1)]$).

Proposition 3. *For every $d \geq 2$, there exists a feature mapping ϕ such that the following lower bound holds. Fix any $q \in (1,2]$. As long as $\Lambda \geq \sqrt{12}$ and $n \geq d^{2/q}\Lambda^2$, we have*

$$\inf_{\widehat{\pi}} \sup_{\mathcal{Q} \in \mathsf{CB}_q(\Lambda)} \mathbb{E}[V^\star_{\mathcal{Q}} - V_{\mathcal{Q}}(\widehat{\pi})] \geq c \cdot d^{1/p}\sqrt{\frac{1}{n}} \cdot \Lambda,$$

where $c > 0$ is some universal constant.

Proof. We will modify the construction. We redefine the state space to be $\{0, 1, \ldots, S\}$ (adding an extra state). This only changes everything by a constant at most but makes the definitions simpler. State 0 is a special state with:

$$\rho(0) = \rho_0, \quad \text{and} \quad \begin{array}{l} n(0, a^{(1)}) = \lfloor n\rho_0/\Gamma_0 \rfloor, \\ n(0, a^{(2)}) = \lfloor n\rho_0 \left(1 - 1/\Gamma_0\right) \rfloor. \end{array}$$

The other states $s \geq 1$ are set as follows:

$$\rho(s) = (1-\rho_0)/S, \quad \text{and} \quad \begin{array}{l} n(s, a^{(1)}) = \lfloor n\frac{1-\rho_0}{S} \cdot \frac{1}{\Gamma_1} \rfloor, \\ n(s, a^{(2)}) = \lfloor n\frac{1-\rho_0}{S} \left(1 - \frac{1}{\Gamma_1}\right) \rfloor. \end{array}$$

Again, as long as $n \geq 2\max\left(\Gamma_0/\rho_0, S\Gamma_1/(1-\rho_0)\right)$, we have

$$\begin{aligned} n(0, a^{(1)}) &\geq 1/2 \cdot (n\rho_0/\Gamma_0), & n(0, a^{(2)}) &\geq 1/2 \cdot n\rho_0 \left(1 - 1/\Gamma_0\right), \\ n(s, a^{(1)}) &\geq 1/2 \cdot n\frac{1-\rho_0}{S} \cdot \frac{1}{\Gamma_1}, & n(s, a^{(2)}) &\geq n\frac{1-\rho_0}{S}\left(1 - \frac{1}{\Gamma_1}\right). \end{aligned}$$

We will now pick the values of $\rho_0 \in [0,1]$ and $\Gamma_0, \Gamma_1 \geq 2$ in order to satisfy the properties enumerated in the proof of Theorem 2.

First, we compute expressions for the bounds. We see that

$$\sum_s \rho(s) \max_{a \in \{a^{(1)}, a^{(2)}\}} \frac{1}{\sqrt{n(s,a)/n}} \geq \sqrt{\rho_0\Gamma_0} + \sqrt{S(1-\rho_0)\Gamma_1},$$

$$\left(\sum_s \rho^q(s) \max_{a \in \{a^{(1)}, a^{(2)}\}} \frac{1}{(n(s,a)/n)^{q/2}}\right)^{1/q} \leq \left((2\rho_0\Gamma_0)^{q/2} + S^{1-q/2}(2(1-\rho_0)\Gamma_1)^{q/2}\right)^{1/q},$$

We pick

$$\rho_0 = 1 - S^{1-2/q}/4, \quad \Gamma_0 = \Lambda^2/(8\rho_0), \quad \Gamma_1 = \Lambda^2/2.$$

One can calculate that

$$\sum_s \rho^q(s) \max_{a \in \{a^{(1)}, a^{(2)}\}} \frac{1}{(n(s,a)/n)^{q/2}} \leq 2(\Lambda/2)^q \leq \Lambda^q,$$

$$\sum_s \rho(s) \max_{a \in \{a^{(1)}, a^{(2)}\}} \frac{1}{n(s,a)/n} \geq \Lambda/2 + S^{1/p}\Lambda/2 \gtrsim S^{1/p}\Lambda.$$

When $q \in (1,2]$, we have the bound that $\rho_0 \in (3/4, 1]$. In order for Γ_0 and Γ_1 to satisfy the requirement that $\Gamma_0, \Gamma_1 \geq 2$, we require $\Lambda \geq \sqrt{12}$. In terms of sample complexity, we require

$$n \geq \max\left\{\frac{4}{9}\Lambda^2, 2S^{2/q}\Lambda^2\right\},$$

which is satisfied whenever $n \geq d^{2/q}\Lambda^2$. □

D Separations between pessimistic learning rules

We begin by stating a formal version of Theorem 3.

Theorem 4. *Fix any $p \geq 1$. Fix any dimension $d \geq 20$, sample size $n \geq 9d^3$, and $\xi(d, n)$, where $\xi : \mathbb{Z}_+ \times \mathbb{Z}_+ \to \mathbb{R}_+$ is a functional that returns a positive real that satisfies $\xi \geq K_\xi$ for an absolute numerical constant $K_\xi > 0$. Then there exists a contextual bandit instance $\mathcal{Q} \in \mathsf{CB}_1(\sqrt{8d})$ such that:*

1. *With probability at least $1/4$, the learning rule $\widehat{\pi}_p$ configured with $\beta = \xi \cdot d^{1/p}/\sqrt{n}$ has suboptimality at least*

$$V(\pi^\star) - V(\widehat{\pi}_p) \geq \frac{K_\xi}{\sqrt{8}} \cdot \frac{d^{1/p+1/2}}{\sqrt{n}}.$$

   This implies a lower bound on the expected suboptimality of $\Omega(K_\xi \cdot d^{1/p+1/2}/\sqrt{n})$.

2. *The learning rule $\widehat{\pi}_\infty$ configured with $\beta = \sqrt{\frac{8\log(K_\xi d^{5/2}/\sqrt{8})}{n}}$ has expected suboptimality at most*

$$\mathbb{E}_{\mathcal{D}}\left[V(\pi^\star) - V(\widehat{\pi}_\infty)\right] \leq c \cdot \sqrt{\frac{d\log(K_\xi d)}{n}},$$

   where $c > 0$ is an absolute numerical constant.

The rest of the section is devoted to the proof of Theorem 4.

Construction of the instance. The instance we construct is a tabular contextual bandit instance with the canonical basis as the feature mapping. We set $S = d/2$[6] and let the state space be $\mathcal{S} = \{1, 2, \ldots, S\}$. Also we choose the action space be $\mathcal{A} = \{a^{(1)}, a^{(2)}\}$. For simplicity, we assume that the sample size n is an integer multiple of S^3.[7] For each state s, we define $\rho(s) = \frac{1}{S}$. Now we will define the empirical counts for each (s, a) pair:

$$n(s, a^{(1)}) = \begin{cases} \frac{n}{9S^3}, & \text{if } s = 1, \\ \frac{n}{S}, & \text{otherwise,} \end{cases} \quad \text{and} \quad n(s, a^{(2)}) = \begin{cases} \frac{n}{S} - \frac{n}{9S^3}, & \text{if } s = 1, \\ 0, & \text{otherwise.} \end{cases}$$

Let $\gamma > 0$ be a *gap parameter* that we will specify later. We set the reward distributions as follows:

$$R(1, a) = \begin{cases} \mathcal{N}(\gamma, 1), & a = a^{(1)}, \\ \mathcal{N}(0, 1), & a = a^{(2)}, \end{cases} \quad \text{and} \quad R(s, a) = \begin{cases} \frac{1}{\sqrt{n}}, & a = a^{(1)}, \\ 0, & a = a^{(2)}, \end{cases} \quad \text{for } s \in \{2, 3, \ldots, S\}.$$

It is seen that for this instance, $\pi^\star(s) = a^{(1)}$ for all $s \in \mathcal{S}$. The difficulty of offline learning lies solely in selecting the optimal action for $s = 1$.

Lastly, we verify that the instance indeed lies in $\mathsf{CB}_1(\Lambda)$ for $\Lambda = \sqrt{8d}$. Algebraic manipulation yields

$$\|\Sigma_{\mathcal{D}}^{-1/2}\mathbb{E}_{s\sim\rho}[\phi(s, \pi^\star(s))]\|_1 = \sum_s \frac{\rho(s)}{\sqrt{n(s, \pi^\star(s))/n}} = \frac{1}{S}\left(3S^{3/2} + (S-1)S^{1/2}\right) = 4S^{1/2} - S^{-1/2}. \tag{16}$$

Plugging in $S = d/2$ proves that $\|\Sigma_{\mathcal{D}}^{-1/2}\mathbb{E}_{s\sim\rho}[\phi(s, \pi_v^\star(s))]\|_1 \leq \Lambda$, i.e., the CB instance is in $\mathsf{CB}_1(\Lambda)$ for $\Lambda = \sqrt{8d}$.

Lower bounding the performance of $\widehat{\pi}_p$. Now we will prove that $\widehat{\pi}_p$ achieves expected suboptimality which scales with $d^{1/p+1/2}$. For the tabular setting, $\widehat{\pi}_p$ can be rewritten as (cf. Equation (9)):

$$\widehat{\pi}_p = \operatorname*{arg\,max}_{\pi:\mathcal{S}\to\mathcal{A}} \left\{ \sum_s \rho(s)\widehat{r}(s, \pi(s)) - \beta\left(\sum_s \frac{\rho^p(s)}{(n(s, \pi(s))/n)^{p/2}}\right)^{1/p} \right\}, \quad \text{where } \beta := \xi\frac{(2S)^{1/p}}{\sqrt{n}}.$$

[6] Again, if d is odd then we let $S = \lfloor \frac{d}{2} \rfloor$ which does not affect the rest of the proof.

[7] Otherwise we can pad the samples with at most S^3 dummy state-action pairs.

By the definition of $\widehat{r}(s,a)$ and $n(s,a)$, we know that $\widehat{\pi}_p(s) = \pi^\star(s) = a^{(1)}$ for $s \geq 2$. Therefore, we have that the suboptimality $V(\pi^\star) - V(\widehat{\pi}_p) \geq \gamma/S$ under the event $\left\{\widehat{\pi}_p(1) = a^{(2)}\right\}$. Therefore our goal boils down to lower bounding the probability $\mathbb{P}_{\mathcal{D}}\left[\widehat{\pi}_p(1) = a^{(2)}\right]$. We can rewrite the learning rule on $s = 1$ as follows:

$$\widehat{\pi}_p(1) = \operatorname*{arg\,max}_{a \in \{a^{(1)}, a^{(2)}\}} \quad \widehat{r}(1,a) - \xi(2S)^{1/p}\Big(\tfrac{1}{n(1,a)^{p/2}} + \sum_{s\geq 2} \tfrac{1}{n(s,a^{(1)})^{p/2}}\Big)^{1/p} \quad =: \operatorname*{arg\,max}_{a \in \{a^{(1)}, a^{(2)}\}} \quad \tilde{r}(1,a),$$

where we use the observation that $\widehat{\pi}_p(s) = \pi^\star(s) = a^{(1)}$ for $s \geq 2$. As a result, we have the identity

$$\mathbb{P}_{\mathcal{D}}\left[\widehat{\pi}_p(1) = a^{(2)}\right] = \mathbb{P}_{\mathcal{D}}\left[\tilde{r}(1,a^{(2)}) > \tilde{r}(1,a^{(1)})\right].$$

The lower bound of the right hand side is provided in the following lemma, whose proof is deferred to the end of this section.

Lemma 4. *Setting the gap parameter $\gamma = K_\xi \frac{S^{1/p+3/2}}{\sqrt{n}}$, one has $\mathbb{P}_{\mathcal{D}}\left[\tilde{r}(1,a^{(2)}) > \tilde{r}(1,a^{(1)})\right] \geq \frac{1}{4}$.*

Putting together the previous claims yields the performance guarantee stated in part (1).

Upper bounding the performance of $\widehat{\pi}_\infty$. By Theorem 1, under the choice $\beta := \sqrt{\frac{8\log(2S/\delta)}{n}}$, with probabilty at least $1-\delta$, the learning rule $\widehat{\pi}_\infty$ achieves the guarantee

$$V(\pi^\star) - V(\widehat{\pi}_\infty) \leq \sqrt{\frac{8\log(d/\delta)}{n}} \cdot \|\Sigma_{\mathcal{D}}^{-1/2}\mathbb{E}_{s\sim\rho}[\phi(s,\pi_v^\star(s))]\|_1 \leq \sqrt{\frac{32S\log(2S/\delta)}{n}},$$

where the inequality uses Equation (16). Thus the expected suboptimality is at most

$$\mathbb{E}_{\mathcal{D}}\left[V(\pi^\star) - V(\widehat{\pi}_\infty)\right] \leq (1-\delta)\cdot\sqrt{\frac{32S\log(2S/\delta)}{n}} + \delta\cdot\left(\frac{K_\xi S^{3/2}}{\sqrt{n}} + \frac{1}{\sqrt{n}}\right).$$

The second term in the expression comes from the fact that on this instance, the worst suboptimality any policy can incur is at most:

$$\frac{1}{S}\left(\gamma + (S-1)\cdot\frac{1}{\sqrt{n}}\right) \leq \frac{K_\xi S^{1/p+1/2}}{\sqrt{n}} + \frac{1}{\sqrt{n}}.$$

Now we pick $\delta = 1/(K_\xi S^{3/2})$, which shows that when $\beta = \sqrt{\frac{8\log(2S^{5/2}K_\xi)}{n}}$, we achieve the guarantee in expectation

$$\mathbb{E}_{\mathcal{D}}\left[V(\pi^\star) - V(\widehat{\pi}_\infty)\right] \leq c\cdot\sqrt{\frac{d\log(K_\xi d)}{n}},$$

for some absolute numerical constant $c > 0$.

D.1 Proof of Lemma 4

We calculate that

$$\begin{aligned}
\mathbb{P}_{\mathcal{D}}\left[\tilde{r}(1,a^{(2)}) > \tilde{r}(1,a^{(1)})\right] &= \mathbb{P}_{\mathcal{D}}\Bigg[\widehat{r}(1,a^{(2)}) - \xi(2S)^{1/p}\Big(\tfrac{1}{n(1,a^{(2)})^{p/2}} + \sum_{s\geq 2}\tfrac{1}{n(s,a^{(1)})^{p/2}}\Big)^{1/p} \\
&\qquad > \widehat{r}(1,a^{(1)}) - \xi(2S)^{1/p}\Big(\tfrac{1}{n(1,a^{(1)})^{p/2}} + \sum_{s\geq 2}\tfrac{1}{n(s,a^{(1)})^{p/2}}\Big)^{1/p}\Bigg] \\
&= \mathbb{P}_{\mathcal{D}}\Bigg[\widehat{r}(1,a^{(2)}) - \xi\frac{(2S)^{1/p}}{\sqrt{n}}\Big(\big(\tfrac{1}{S} - \tfrac{1}{9S^3}\big)^{-p/2} + (S-1)S^{p/2}\Big)^{1/p} \\
&\qquad > \widehat{r}(1,a^{(1)}) - \xi\frac{(2S)^{1/p}}{\sqrt{n}}\Big((9S^3)^{p/2} + (S-1)S^{p/2}\Big)^{1/p}\Bigg].
\end{aligned}$$

Note that the empirical reward $\widehat{r}(1, a^{(1)})$ is distributed as a mean-γ Gaussian, so we can lower bound the previous display as

$$\begin{aligned}
&\mathbb{P}_{\mathcal{D}}\left[\tilde{r}(1, a^{(2)}) > \tilde{r}(1, a^{(1)})\right] \\
&\quad \geq \frac{1}{2} \cdot \mathbb{P}_{\mathcal{D}}\Bigg[\widehat{r}(1, a^{(2)}) > \gamma \\
&\qquad + \xi \frac{(2S)^{1/p}}{\sqrt{n}} \left\{ \left(\left(\tfrac{1}{S} - \tfrac{1}{9S^3}\right)^{-p/2} + (S-1)S^{p/2} \right)^{1/p} - \left((9S^3)^{p/2} + (S-1)S^{p/2} \right)^{1/p} \right\} \Bigg].
\end{aligned}$$

Now observe that under the choice of gap parameter $\gamma = K_\xi \frac{S^{1/p+3/2}}{\sqrt{n}}$, for sufficiently large S (say $S \geq 10$) we have:

$$\begin{aligned}
&\gamma + \xi \frac{(2S)^{1/p}}{\sqrt{n}} \left\{ \left(\left(\tfrac{1}{S} - \tfrac{1}{9S^3}\right)^{-p/2} + (S-1)S^{p/2} \right)^{1/p} - \left((9S^3)^{p/2} + (S-1)S^{p/2} \right)^{1/p} \right\} \\
&\quad < \gamma + K_\xi \frac{(2S)^{1/p}}{\sqrt{n}} \left\{ \left(\left(\tfrac{1}{S} - \tfrac{1}{9S^3}\right)^{-p/2} + (S-1)S^{p/2} \right)^{1/p} - \left((9S^3)^{p/2} + (S-1)S^{p/2} \right)^{1/p} \right\} \\
&\quad < 0.
\end{aligned}$$

Using the fact that $\widehat{r}(1, a^{(2)})$ is distributed as a mean-zero Gaussian, we can further lower bound the probability as $P_{\mathcal{D}}\left[\tilde{r}(1, a^{(2)}) > \tilde{r}(1, a^{(1)})\right] \geq \frac{1}{4}$, thus proving the lemma.

E Plug-in estimation performance

In this section, we discuss the performance of the simple plug-in rule:

$$\widehat{\pi}_{\mathsf{plug}}(s) \coloneqq \arg\max_{a \in \mathcal{A}} \phi(s, a)^\top \widehat{\theta}_{\mathsf{ols}}.$$

This is a natural baseline to study. It is an instantiation of the so-called *certainty equivalence principle* from control theory, which is shown to be optimal for some online control problems [see, e.g., 22, 26].

E.1 Performance upper bound

Proposition 4 (Folklore; see, e.g., [5], pg. 50)**.** *Assume that $\|\phi(s,a)\|_2 \leq B$. Then with probability at least $1-\delta$, $\widehat{\pi}_{\mathsf{plug}}$ achieves a suboptimality*

$$V(\pi^\star) - V(\widehat{\pi}_{\mathsf{plug}}) \leq \sqrt{\frac{8d\log(d/\delta)}{n}} \cdot B \cdot \lambda_{\min}(\Sigma_{\mathcal{D}})^{-1/2}.$$

Proof. Let us denote $\widehat{V}(\pi) \coloneqq \phi(s,a)^\top \widehat{\theta}_{\mathsf{ols}}$. We use Equation (4) and note that

$$\begin{aligned}
V(\pi^\star) - \widehat{V}(\pi^\star) &= \mathop{\mathbb{E}}_{s\sim\rho}\left[\phi(s, \pi^\star(s))^\top (\theta^\star - \widehat{\theta}_{\mathsf{ols}})\right], \\
\widehat{V}(\widehat{\pi}_{\mathsf{plug}}) - V(\widehat{\pi}_{\mathsf{plug}}) &= \mathop{\mathbb{E}}_{s\sim\rho}\left[\phi(s, \widehat{\pi}_{\mathsf{plug}}(s))^\top (\widehat{\theta}_{\mathsf{ols}} - \theta^\star)\right].
\end{aligned}$$

Note that for any (s,a):

$$\phi(s,a)^\top(\theta^\star - \widehat{\theta}_{\mathsf{ols}}) \leq B \left\|\theta^\star - \widehat{\theta}_{\mathsf{ols}}\right\|_2 \leq \left\|\Sigma_{\mathcal{D}}^{1/2}(\theta^\star - \widehat{\theta}_{\mathsf{ols}})\right\|_2 \cdot B \cdot \lambda_{\min}(\Sigma_{\mathcal{D}})^{-1/2}.$$

Lastly, the ℓ_2 norm term is bounded as a consequence of Lemma 1, which shows that:

$$\left\|\Sigma_{\mathcal{D}}^{1/2}(\theta^\star - \widehat{\theta}_{\mathsf{ols}})\right\|_2 \leq \sqrt{\frac{2d\log(d/\delta)}{n}}.$$

(The $\log d$ factor can be removed with a more involved argument.) This proves the result. □

Note that Proposition 4 gives a dependence on $B \cdot \lambda_{\min}(\Sigma_{\mathcal{D}})^{-1/2}$, which is always worse than the guarantee for $\widehat{\pi}_2$ learning rule (where the relevant "complexity measure" is $\mathfrak{C}_2$). The plug-in rule requires the covariates $\{(\phi(s_i, a_i))\}_{i=1}^n$ to cover all directions well in order to obtain low suboptimality, even if the optimal policy is well-covered by the dataset. The benefit of pessimism is that when the features cover the optimal policy well, one can achieve better performance.

E.2 Separation between plug-in estimation and pessimism

Now we state a separation which shows when pessimism is prefered over the plug-in rule. This result is included for completeness; a version of it appears in prior work, see, e.g., Proposition 1 from the paper [24]. It is stated for the random design setting (where we assume the covariates (s, a) are drawn i.i.d. from a behavior distribution μ). This implies a fixed design result (in line with the other results of this paper).

Proposition 5. *Fix any $A \geq 8$. There exists a multi-armed bandit instance with A actions such that*

1. *For any $n \leq 2^A$, the plug-in rule obeys*

$$\mathbb{E}_{\mathcal{D}}\left[V(\pi^\star) - V(\widehat{\pi}_{\mathsf{plug}})\right] \geq c_1.$$

2. *As long as $n \geq 200$, the learning rule $\widehat{\pi}_\infty$ achieves the guarantee*

$$\mathbb{E}_{\mathcal{D}}\left[V(\pi^\star) - V(\widehat{\pi}_\infty)\right] \leq c_2 \sqrt{\frac{\log(An)}{n}}.$$

Here $c_1, c_2 > 0$ are two universal constants.

Proof. **Proof of part (1).** We specify the behavior distribution as follows

$$\mu(a_k) = 2^{-k} \quad \text{for } 1 \leq k \leq A-1, \quad \text{and} \quad \mu(a_A) = 2^{-A+1}.$$

In addition, the reward distribution is chosen to be

$$R(a_1) = 0.99, \quad \text{and} \quad R(a_i) = \mathsf{Ber}(1/2) \text{ for all } k \geq 2.$$

Fix any sample count $8 \leq n \leq 2^A$. There must exist some $1 \leq k \leq A$ such that $\mu(a_k) \in [\frac{1}{n}, \frac{2}{n}]$. Consider the event $\mathcal{E}_k := \{n(k) = 1\}$. Then we have

$$\mathbb{P}[\mathcal{E}_k] = n \cdot (1 - \mu(a_k))^{n-1} \mu(a_k) \overset{(i)}{\geq} \left(1 - \frac{2}{n}\right)^{n-1} \overset{(ii)}{\geq} 0.1.$$

Here step (i) follows from the fact that $\mu(a_k) \in [\frac{1}{n}, \frac{2}{n}]$, and step (ii) is a result of the assumption $n \geq 8$. Thus we can lower bound the expected suboptimality of the plug-in rule as

$$\mathbb{E}_{\mathcal{D}}\left[V(\pi^\star) - V(\widehat{\pi}_{\mathsf{plug}})\right] \geq 0.49 \cdot \mathbb{P}\left[\widehat{\pi}_{\mathsf{plug}} \neq 1\right] \geq 0.49 \cdot \mathbb{P}\left[\mathcal{E}_k \cap \{\widehat{r}(k) = 1\}\right] \geq 0.02.$$

This proves part (1).

Proof of part (2). We use Corollary 1 to prove part (2). In the multi-armed bandit instance we construct, we have $S = 1$ and $\mu(\pi^\star) = 1/2$. We set $\delta = 1/n$ to get the in-expectation guarantee:

$$\mathbb{E}_{\mathcal{D}}\left[V(\pi^\star) - V(\widehat{\pi}_\infty)\right] \lesssim (1-\delta) \cdot \sqrt{\frac{\log(A/\delta)}{n}} + \delta \cdot 0.49 \lesssim \sqrt{\frac{\log(An)}{n}}.$$

It remains to check the sample complexity requirement. Under the choice of $\delta = 1/n$, the sample complexity holds whenever $n \gtrsim \log n$, which is true if n is sufficiently large (say, $n \geq 200$).

□

The proof of Proposition 5 illustrates that pessimistic learning rules can outperform plug-in estimation when the optimal (or near-optimal) actions are well-represented in the dataset. In this case, the optimal action appears $1/2$ the time in expectation. Proposition 5 *does not show* whether pessimism is better in a minimax sense. Indeed, the recent paper [36] shows the intriguing result that pessimism, plug-in estimation, and even optimism (an algorithmic paradigm used in the *online* setting) are equally "minimax-optimal", and there exist instances where any one of the three outperforms the other two.

F Counterexample to Theorem 4.3 in the paper [40]

In this section, we present a counterexample to showcase that the statement in Theorem 4.3 in the paper [40] cannot hold in general.

We consider the simple two-armed bandit problem with Bernoulli reward distributions. Fix the instance $\mathcal{Q}_1$ to be $\mu(a_1) = \mu(a_1) = 1/2$, $R(a_1) = \mathsf{Ber}(1)$, and $R(a_2) = \mathsf{Ber}(0)$. We make the following two observations. First, the optimal action in this instance $\mathcal{Q}_1$ is a_1. Second, the quantity $\mathfrak{C}_1$ is given by

$$\frac{1}{\sqrt{\mu(a_1)}} = \sqrt{2}.$$

To demonstrate that Theorem 4.3 in the paper [40] cannot hold in general, it suffices to prove the following proposition.

Proposition 6. *For any alternative instance $\mathcal{Q}_2$, there exists a learning rule $\widehat{\pi}$ such that*

$$\max_{\mathcal{Q}\in\mathcal{Q}_1,\mathcal{Q}_2} \mathbb{E}_{\mathcal{D}}\left[V^\star_{\mathcal{Q}} - V_{\mathcal{Q}}(\widehat{\pi})\right] \le e^{-cn} \tag{17}$$

holds for some constant $c > 0$.

Proof. In general, we can parametrize the alternative instance $\mathcal{Q}_2$ using the following three quantities

$$\begin{aligned}
\mu_{\mathcal{Q}_2}(a_1) &= p;\\
R_{\mathcal{Q}_2}(a_1) &= \mathsf{Ber}(\alpha);\\
R_{\mathcal{Q}_2}(a_2) &= \mathsf{Ber}(\beta),
\end{aligned}$$

where p, α, β are all between 0 and 1.

We make a key observation that: if $\alpha \ge \beta$ (i.e., the alternative instance $\mathcal{Q}_2$ has the same optimal action a_1 as for $\mathcal{Q}_1$), then we could take $\widehat{\pi} = a_1$. The desired claim (17) follows trivially since such $\widehat{\pi}$ will incur zero suboptimality in both instances.

Consequently, we only need to focus on the case when $\alpha < \beta$. For this case, we will explicitly construct an estimated policy $\widehat{\pi}$ such that the claim (17) is true. Let $\widehat{\pi}_{\mathsf{LR}}$ be the optimal likelihood ratio test—based on the data $\mathcal{D}$—between the two instances $\mathcal{Q}_1$, and $\mathcal{Q}_2$. We claim that

$$\max_{\mathcal{Q}\in\mathcal{Q}_1,\mathcal{Q}_2} \mathbb{E}_{\mathcal{D}}\left[V^\star_{\mathcal{Q}} - V_{\mathcal{Q}}(\widehat{\pi}_{\mathsf{LR}})\right] \le e^{-cn}.$$

To see this, we observe that

$$\begin{aligned}
\max_{\mathcal{Q}\in\mathcal{Q}_1,\mathcal{Q}_2} \mathbb{E}_{\mathcal{D}}\left[V^\star_{\mathcal{Q}} - V_{\mathcal{Q}}(\widehat{\pi}_{\mathsf{LR}})\right] &\le \mathbb{E}_{\mathcal{D}}\left[V^\star_{\mathcal{Q}_1} - V_{\mathcal{Q}_1}(\widehat{\pi}_{\mathsf{LR}})\right] + \mathbb{E}_{\mathcal{D}}\left[V^\star_{\mathcal{Q}_2} - V_{\mathcal{Q}_2}(\widehat{\pi}_{\mathsf{LR}})\right]\\
&= \mathbb{P}_{\mathcal{Q}_1}(\widehat{\pi}_{\mathsf{LR}} = a_2) + (\beta - \alpha)\cdot \mathbb{P}_{\mathcal{Q}_2}(\widehat{\pi}_{\mathsf{LR}} = a_1)\\
&\le \mathbb{P}_{\mathcal{Q}_1}(\widehat{\pi}_{\mathsf{LR}} = a_2) + \mathbb{P}_{\mathcal{Q}_2}(\widehat{\pi}_{\mathsf{LR}} = a_1).
\end{aligned}$$

Here, the equality follows from the definition, and the last inequality uses the fact that $\beta - \alpha \le 1$.

By definition of $\widehat{\pi}_{\mathsf{LR}}$, we have

$$\mathbb{P}_{\mathcal{Q}_1}(\widehat{\pi}_{\mathsf{LR}} = a_2) + \mathbb{P}_{\mathcal{Q}_2}(\widehat{\pi}_{\mathsf{LR}} = a_1) = 1 - \mathsf{TV}(\mathcal{Q}_1^n \| \mathcal{Q}_2^n).$$

Therefore it suffices to prove that

$$1 - \mathsf{TV}(\mathcal{Q}_1^n \| \mathcal{Q}_2^n) \le e^{-cn}. \tag{18}$$

Proof of the bound (18)**.** By the relation between the total variation distance and the Hellinger distance, we have

$$1 - \mathsf{TV}(\mathcal{Q}_1^n \| \mathcal{Q}_2^n) \le 1 - \mathsf{Hel}^2(\mathcal{Q}_1^n \| \mathcal{Q}_2^n) = \left[1 - \mathsf{Hel}^2(\mathcal{Q}_1 \| \mathcal{Q}_2)\right]^n,$$

where the inequality arises from the property of the Hellinger distance. By constructions of $\mathcal{Q}_1$, $\mathcal{Q}_2$, and the definition of the Hellinger distance, one has

$$\mathsf{Hel}^2(\mathcal{Q}_1 \| \mathcal{Q}_2) = \frac{1}{2}\left\{\left(\frac{1}{\sqrt{2}} - \sqrt{p\alpha}\right)^2 + p(1-\alpha) + (1-p)\beta + \left(\frac{1}{\sqrt{2}} - \sqrt{(1-p)(1-\beta)}\right)^2\right\}.$$

Note that

$$\inf_{p,\alpha,\beta:\beta>\alpha} \mathsf{Hel}^2(\mathcal{Q}_1 \| \mathcal{Q}_2) = 1 - \frac{1}{\sqrt{2}}.$$

We can then combine the previous relations to finish the proof of the bound (18). □