BEYOND STATE-WISE MIRROR DESCENT: OFFLINE POLICY OPTIMIZATION WITH PARAMETRIC POLICIES

Anonymous authors
Paper under double-blind review

ABSTRACT

We investigate the theoretical aspects of offline reinforcement learning (RL) under general function approximation. While prior works (e.g., Xie et al., 2021) have established the theoretical foundations of learning a good policy from offline data via pessimism, existing algorithms that are computationally tractable (often in an oracle-efficient sense), such as PSPI, only apply to finite and small action spaces. Moreover, these algorithms rely on *state-wise* mirror descent and require actors to be implicitly induced from the critic functions, failing to accommodate standalone policy parameterization which is ubiquitous in practice. In this work, we address these limitations and extend the theoretical guarantees to parameterized policy classes over large or continuous action spaces. When extending mirror descent to parameterized policies, we identify *contextual coupling* as the core difficulty, and show how connecting mirror descent to natural policy gradient leads to novel analyses, guarantees, and algorithmic insights, including a surprising unification between offline RL and imitation learning.

1 INTRODUCTION

Learning a good policy from historical data, a.k.a. offline reinforcement learning (RL), is an important paradigm for bringing RL to real-life domains (Levine et al., 2020). While the statistical aspects of offline RL theory are relatively well-understood (Jiang & Xie, 2025), the information-theoretic algorithms that achieve strong statistical guarantees under general conditions are often not immediately computationally tractable, and addressing the computational challenge typically requires additional assumptions or sacrificing some of the statistical generality. In particular, while the analyses often make mild assumptions on the *critic* (i.e., value function approximator), strong restrictions are often imposed on the *actor* (i.e., policy approximator), limiting the theory's applicability scope and practical relevance (Xie et al., 2021; Cheng et al., 2022).

As a canonical example, consider PSPI by Xie et al. (2021) (Section 3), an actor-critic style algorithm that achieves strong guarantees when critics are modeled via general function approximation. PSPI assumes an oracle that computes a pessimistic critic f_k for policy π_k, and employs *mirror descent* (MD) in the actor update:

$$\pi_{k+1}(a \mid s) \propto \pi_k(a \mid s) \exp\big(\eta f_k(s,a)\big). \tag{1}$$

This brings two issues. First, the guarantee of PSPI depends on the log-cardinality of the action space and thus does not apply to arbitrarily large (including continuous) action spaces that are ubiquitous in control problems (e.g., Gaussian policies in robotics). Secondly, MD update in Eq.(1) is *state-wise*, i.e., the action distribution in each state is updated independently. This means that the actor (policy) is implicitly induced from the critics (f_k) and cannot have its own standalone parameterization (e.g., an actor network that is separate from the critic network).

These limitations not only create a disconnection between theory and practice, but also indicate a gap in theory itself: standalone policy parameterization over continuous actions is not an issue at all when we only pursue statistical guarantees and ignore computational feasibility (Xie et al., 2021, Section 3), but difficulties arise when we aim for computational efficiency.

In this work, we address this gap by systematically studying variants of policy optimization methods and their statistical/computational properties, under the assumption that a standard pessimistic critic oracle is given (Assumption 2). Our contributions are 3-fold:

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

- **PSPI with general action spaces.** In Section 3, we revisit the PSPI algorithm to disentangle critic- and actor-side errors, and present a mild extension of the analysis to arbitrarily large (and even continuous) action spaces.
- **Contextual coupling: hardness and a guiding principle.** In Section 4, we consider standalone policy parameterization. We show that a natural contextualization of PSPI can incur constant regret due to an underlying challenge we coin *contextual coupling*. To figure out the underlying reason, we consider a general policy optimization recipe that helps analyze and identify the problem, which admits a regret decomposition via compatible function approximation (CFA) and can serve as the guiding principle for designing effective actor updates.
- **Actor updates with finite-sample guarantees.** Building on the earlier regret decomposition, we develop and analyze two statistically and computationally efficient actor updates in Section 5: a *least-square regression* update-rule (LSPU) closely related to natural policy gradient, and a *distributionally robust* update-rule (DRPU) that leverages importance weighting and can be more robust to actor-critic incompatibility. Surprisingly, when the offline distribution coincides with that of the comparator policy, our DRPU method reveals a connection to behavior cloning, providing an interesting unification between offline RL and imitation learning.

2 PRELIMINARY

Markov Decision Processes. We formulate RL in infinite-horizon discounted Markov Decision Processes (MDPs), specified by $(\mathcal{S}, \mathcal{A}, P, R, \gamma, d_0)$, where $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action space, $P : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ is the transition kernel, $R : \mathcal{S} \times \mathcal{A} \to [0, R_{\max}]$ is the reward function, $\gamma \in [0, 1)$ is the discount factor, and $d_0 \in \Delta(\mathcal{S})$ is the initial state distribution. We mostly consider finite and discrete $\mathcal{S}$ and $\mathcal{A}$ with possibly arbitrarily large cardinalities, and our results can be easily extended to continuous action spaces as exemplified in Appendix B.2.

A (stochastic) policy $\pi : \mathcal{S} \to \Delta(\mathcal{A})$ specifies a decision-making strategy. Its expected discounted return is denoted by $J(\pi) = \mathbb{E}_\pi[\sum_{t=0}^{\infty} \gamma^t r_t]$, with $s_0 \sim d_0$, $a_t \sim \pi(\cdot|s_t)$, $r_t = R(s_t, a_t)$, and $s_{t+1} \sim P(\cdot|s_t, a_t)$. The Q-function of π is defined by $Q^\pi(s, a) = \mathbb{E}_\pi[\sum_{t=0}^{\infty} \gamma^t r_t | s_0 = s, a_0 = a]$, which is the unique fixed point of its Bellman operator $\mathcal{T}^\pi$: $\forall f \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$, $(\mathcal{T}^\pi f)(s, a) = R(s, a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)}[f(s', \pi)]$, where $f(s', \pi) = \mathbb{E}_{a' \sim \pi(\cdot|s')}[f(s', a')]$. $V^\pi(s) = Q^\pi(s, \pi)$ is the state value function and $A^\pi(s, a) = Q^\pi(s, a) - V^\pi(s)$ is the advantage function. Let d^π denote the discounted state-action occupancy distribution of policy π, i.e., $d^\pi(s, a) = (1 - \gamma) \sum_{t=0}^{\infty} \gamma^t d_t^\pi(s, a)$, where d_t^π is the marginal distribution of (s_t, a_t).[1]

Offline RL. In offline RL, we learn from pre-collected data without environment interaction. For conceptual clarity, we assume access to two datasets: a *critic dataset* with transition tuples (s, a, r, s') for fitting the critic, and an *actor dataset* consisting of only (s, a) pairs for actor updates.

As our focus is on the actor, we do **not** explicitly characterize the critic dataset, since its properties will later be encapsulated in an oracle assumption (Assumption 2). For the actor dataset, we assume access to N i.i.d. (s, a) pairs drawn from a data distribution d^D, which might be arbitrary as long as it satisfies appropriate coverage condition. In practice, they may be extracted from the critic dataset, but our formulation also allows additional data sources, such as expert annotations, which are considered in hybrid setting of imitation learning and offline RL (Mao, 2023; Yang et al., 2023).

We adopt the following standard assumption on data coverage, known as *density coverage* (Jiang & Xie, 2025). The theoretical guarantees in offline RL typically state that we can compete with any comparator policy π_{cp} that is well-covered by data (i.e., C is small).

Assumption 1 (Actor-Side Data Coverage)**.** *Assume* $\sup_{s,a} |d^{\pi_{\text{cp}}}(s, a)/d^D(s, a)| \le \mathsf{C} < \infty$.

To deal with large state and action spaces, we are in the function approximation regime, modeling Q-functions using a function class $\mathcal{F} = \{f : \mathcal{S} \times \mathcal{A} \to [0, V_{\max}]\}$, where $V_{\max} = R_{\max}/(1 - \gamma)$. Typical choices of $\mathcal{F}$ include linear classes or deep neural networks. In the literature, $\mathcal{F}$ generally needs to satisfy expressivity conditions such as realizability or Bellman completeness (Chen & Jiang, 2019), which will also be implicitly captured in Assumption 2.

[1] With a slight abuse of notation, we also use d^π to denote the discounted state occupancy of policy π (after marginalizing out the actions). That is, $d^\pi(s) = \sum_{a \in \mathcal{A}} d^\pi(s, a)$.

Algorithm 1 Pessimistic Soft Policy Iteration (PSPI)

(Xie et al., 2021)

Input: number of iterations K, learning rate η
Initialize policy π_1 as an arbitrary policy (e.g., uniform over $\mathcal{A}$)
for $k = 1, 2, \ldots, K$ **do**
  Critic: compute π_k's pessimistic Q-function f_k using oracle $\mathcal{O}$ satisfying Assumption 2
  Actor: update policy by Eq. (1), i.e., $\pi_{k+1}(a \mid s) \propto \pi_k(a \mid s) \exp\big(\eta f_k(s,a)\big)$
end for
Output: $\hat{\pi} = \mathrm{Unif}[\pi_{1:K}]$

Math notation. We use $A \lesssim B$ or $A = \mathcal{O}(B)$ to denote $A \le cB$ for some constant $c > 0$; $A = \Omega(B)$ is equivalent to $B = \mathcal{O}(A)$. For a function $f : \mathcal{X} \to \mathbb{R}$ for finite $\mathcal{X}$, its supremum norm is $\|f\|_\infty = \max_{x \in \mathcal{X}} |f(x)|$. The KL divergence is $D_{\mathrm{KL}}(p\|q) = \sum_{x \in \mathcal{X}} p(x) \log \frac{p(x)}{q(x)}$ for two distributions with probability mass functions p and q.

3 Pessimistic Soft Policy Iteration as State-wise Mirror Descent

Xie et al. (2021) proposed Pessimistic Soft Policy Iteration (PSPI), an actor-critic method for offline policy optimization (see Algorithm 1). At a high level, PSPI proceeds iteratively: at each iteration k, the critic produces a pessimistic estimate f_k of the value Q^{π_k}, and the actor performs a soft improvement step based on f_k. The final output is a (trajectory-level) uniform mixture of the intermediate policies $\pi_1, \ldots, \pi_K$.

Since our ultimate goal is to bound the suboptimality gap $J(\pi_{\mathrm{cp}}) - J(\hat{\pi})$ between some comparator policy π_{cp} and the output policy $\hat{\pi}$, a key component in the analysis is to decompose it via the generalized performance difference lemma (Lemma 22): $J(\pi_{\mathrm{cp}}) - J(\hat{\pi}) =$

$$\frac{1}{1-\gamma}\Bigg(\underbrace{\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}_{s\sim d^{\pi_{\mathrm{cp}}}}[f_k(s,\pi_{\mathrm{cp}}) - f_k(s,\pi_k)]}_{\text{actor-side error (Eq. (2)), \textbf{focus of this paper}}} + \underbrace{\frac{1}{K}\sum_{k=1}^{K}\Big(\mathbb{E}_{d^{\pi_{\mathrm{cp}}}}[\mathcal{T}^{\pi_k} f_k - f_k] + \mathbb{E}_{d^{\pi_k}}[f_k - \mathcal{T}^{\pi_k} f_k]\Big)}_{\text{critic-side Bellman error, handled by Assumption 2}}\Bigg).$$

The first term corresponds to actor-side optimization error, capturing how well $\pi_1, \ldots, \pi_K$ optimizes against π_{cp} measured by the pessimistic Q-function; the remaining terms are critic-side errors that measure the violation of the Bellman equations. Such decomposition cleanly separates the difficulty of policy optimization (i.e., actor update) from that of value estimation.

To isolate the actor-side difficulty, we abstract the critic as a pessimistic oracle $\mathcal{O}$, as reflected in Assumption 2, which directly controls the latter two Bellman error terms. Such control is central to offline RL theory and essentially an intermediate guarantee achieved in many existing works under standard assumptions (Jin et al., 2020; Xie et al., 2021; Cheng et al., 2022). More detailed discussions and algorithmic realizations of this oracle, e.g., via Bellman error minimization or marginalized importance sampling, are provided in Appendix B.1.

Assumption 2 (Pessimistic Oracle)**.** *We have an efficient oracle $\mathcal{O}$ that, given any policy π as input, it outputs a function f that satisfies the following two conditions:*

1. *(*Pessimism*) The function f is a pessimistic estimation of Q^π (up to some tolerance $\epsilon_r \ge 0$), i.e., $J_f(\pi) - J(\pi) \le \epsilon_r/(1-\gamma)$ with high probability, where $J_f(\pi) = \mathbb{E}_{s\sim d_0}[f(s,\pi)]$.*
2. *(*Bounded "Transferred" Bellman error*) The Bellman error under $d^{\pi_{\mathrm{cp}}}$ is bounded by some $\epsilon_b > 0$. That is, $\mathbb{E}_{d^{\pi_{\mathrm{cp}}}}[\mathcal{T}^\pi f - f] \le \epsilon_b$ with high probability.*

Under Assumption 2, $\epsilon_r + \epsilon_b$ naturally controls the latter two terms in the suboptimality decomposition. Thus, bounding $J(\pi_{\mathrm{cp}}) - J(\hat{\pi})$ reduces to controlling the *actor-side* error, i.e., the first term in the decomposition, which is a notion of ***regret*** we focus on throughout the remainder of the paper:

$$\frac{\mathrm{Reg}_K}{K} := \frac{1}{K}\sum_{k=1}^{K}\mathbb{E}_{s\sim d^{\pi_{\mathrm{cp}}}}\Big[f_k(s,\pi_{\mathrm{cp}}) - f_k(s,\pi_k)\Big]. \tag{2}$$

State-wise regret control. Note that in Eq. (2), the regret is measured under an unknown and often inaccessible state distribution $d^{\pi_{\text{cp}}}$. A natural workaround is to solve this online optimization problem in a *state-wise* manner: if the regret can be controlled uniformly for each state $s \in \mathcal{S}$, then the overall regret bound holds for any $d^{\pi_{\text{cp}}} \in \Delta(\mathcal{S})$. Also, it is worth noting that the computational complexity of doing so does not depend on the size of $\mathcal{S}$, since we only need to lazily run the algorithm on states observed in the data.

Leveraging this observation, the actor update in PSPI can be interpreted as a multiplicative-weights update from online learning (Hazan et al., 2016). In particular, for each state, the policy update in Eq. (1) corresponds to performing ***mirror descent*** (MD) with KL regularizer:

$$\pi_{k+1}(\cdot|s) = \underset{\pi(\cdot|s)\in\Delta(\mathcal{A})}{\arg\max} \left\{ f_k(s,\pi) - \tfrac{1}{\eta} D_{\text{KL}}(\pi(\cdot|s)\|\pi_k(\cdot|s)) \right\}. \tag{3}$$

This perspective leads to the regret guarantee established for PSPI in (Xie et al., 2021), which, however, depends on $|\mathcal{A}|$, the cardinality of $\mathcal{A}$.[2] We next show that their result can be directly extended to remove the explicit $|\mathcal{A}|$ dependence and thus handle arbitrarily large action spaces; we state the result below for finite and discrete $\mathcal{A}$ to be consistent with the setup of the rest of the paper.

Theorem 1 (Regret Bound of Algorithm 1)**.** *Assume that for each iteration k, the critic f_k is bounded in $[0, V_{\max}]$, and that $D_{\text{KL}}(\pi_{\text{cp}}\|\pi_1) < \infty$. Then by choosing $\eta = \sqrt{8D_{\text{KL}}(\pi_{\text{cp}}\|\pi_1)/(KV_{\max}^2)}$, PSPI iterates satisfy*

$$\frac{\text{Reg}_K}{K} \le V_{\max}\sqrt{\frac{D_{\text{KL}}(\pi_{\text{cp}}\|\pi_1)}{2K}},$$

where $D_{\text{KL}}(\pi_{\text{cp}}\|\pi) = \mathbb{E}_{s\sim d^{\pi_{\text{cp}}}}[D_{\text{KL}}(\pi_{\text{cp}}(\cdot|s)\|\pi(\cdot|s))]$ is the expected KL divergence under $d^{\pi_{\text{cp}}}$.

The result is a mild extension of Xie et al. (2021, Theorem 4.1), where we keep $D_{\text{KL}}(\pi_{\text{cp}}\|\pi_1)$ instead of upper-bounding it with $\log|\mathcal{A}|$ as in Xie et al. (2021). This allows the term to be well bounded for some structured policies, even when $\mathcal{A}$ is unbounded in cardinality or even continuous. In Appendix B.2, we show that Theorem 1 can be extended straightforwardly to continuous action space,[3] and for standard forms of continuous policies such as Gaussian policies, or structural action space like convex action space, the KL term remains well bounded; see Appendix B.3 for details.

4 Contextual Coupling in Parameterized Policy Optimization

The state-wise mirror descent of PSPI in Section 3 does not have its own policy parameterization; rather, it is performing policy optimization within a softmax policy class implicitly induced from $\mathcal{F}$: $\{\pi(\cdot|s) \propto \pi_1(\cdot|s)\exp\big(\sum_{i=1}^n f_i(s,\cdot)\big) : k \in [K], f_{1:k} \in \mathcal{F}\}$. In this section, we move beyond this setting and consider policy optimization over a standalone policy class parameterized by some $\theta \in \mathbb{R}^{\mathsf{d}}$, which is ubiquitous in practice:

$$\Pi_\theta := \{\pi_\theta : \theta \in \mathbb{R}^{\mathsf{d}}\}.$$

As standard, we consider gradient-type updates to the actor, which requires the following assumption on policy differentiability and smoothness assumptions in theoretical analyses (Agarwal et al., 2021). Such an assumption is widely adopted in the policy-gradient literature, and holds for many popular policy classes under appropriate norms, including the canonical softmax policies, log-linear policies, neural policies, and even Gaussian policies under mild conditions.

Assumption 3 (Policy Class)**.** *Let π_θ be a differentiable policy parameterized by $\theta \in \mathbb{R}^{\mathsf{d}}$. For all $(s,a) \in \mathcal{S}\times\mathcal{A}$, the log-probability is G-Lipschitz and β-smooth with respect to a chosen norm $\|\cdot\|$ (and its dual norm $\|\cdot\|_*$).*

$$|\log\pi_\theta(a|s) - \log\pi_{\theta'}(a|s)| \le G\|\theta-\theta'\|, \quad \|\nabla\log\pi_\theta(a|s) - \nabla\log\pi_{\theta'}(a|s)\|_* \le \beta\|\theta-\theta'\|.$$

[2] Apart from the regret's dependence on $|\mathcal{A}|$, the computation required by PSPI might also seem to require enumeration over $\mathcal{A}$ since we need to sample from the softmax policy in Eq.(1). This is, however, not always the case, as one can use techniques such as rejection sampling to avoid enumeration, and structured settings such as LQRs admit closed-form expressions for the policy that can be efficiently sampled from; see Appendix B.3 for details. Thus, the PSPI algorithm is also computationally viable for large or even continuous $\mathcal{A}$.

[3] Results in subsequent sections can also be extended to continuous action spaces in a similar manner, which we omit and stick to the discrete setup for readability.

4.1 CONTEXTUAL MIRROR DESCENT CAN FAIL

In Section 3, we explained that PSPI admits a state-wise mirror descent interpretation: each state runs mirror descent and enjoys regret guarantee independently. To deal with the standalone policy class Π_θ, a natural attempt is to *contextualize* mirror descent by coupling these state-wise updates in Eq. (3) through the shared parameter θ and aggregating them across states via some state distribution:

$$\pi_{k+1} = \arg\max_{\pi \in \Pi_\theta} \mathbb{E}_{s \sim d^D} \left[f_k(s, \pi) - \tfrac{1}{\eta} D_{\mathrm{KL}}(\pi(\cdot|s) \| \pi_k(\cdot|s)) \right], \tag{4}$$

where π_k denotes π_{θ_k}. The above form of ***contextual mirror descent*** is defined under the state-marginal of the data distribution d^D, which is the only state distribution directly accessible in the offline setting. In contrast, the regret in Eq. (2) is evaluated under the comparator distribution $d^{\pi_{\mathrm{cp}}}$. As a consequence, the challenge of *distribution mismatch* arises between Eq. (2) and Eq. (4).

Notice that the integrand $f_k(s, \pi_{\mathrm{cp}}) - f_k(s, \pi_k)$ in Eq. (2) is not one-sided and may be positive in some states while negative in others. As a result, errors controlled under d^D do not necessarily translate to $d^{\pi_{\mathrm{cp}}}$ even with the coverage condition (Assumption 1).[4] Therefore, regret bound under d^D does not, in general, imply guarantee under $d^{\pi_{\mathrm{cp}}}$. Indeed, we show via the following hardness result that contextual mirror descent can fail (even if we do not account for finite-sample errors and effectively assume infinite data), and the construction is deferred to Appendix C.1.

Proposition 2 (Failure for Contextual Mirror Descent)**.** *There exist an MDP $\mathcal{M}$, a policy class Π_θ satisfying Assumption 3, a comparator policy $\pi_{\mathrm{cp}} \in \Pi_\theta$, and offline data satisfying Assumption 1, such that the policies $\{\pi_k\}$ produced by contextual MD in Eq.* (4) *incur constant per-step regret:*

$$\mathbb{E}_{s \sim d^{\pi_{\mathrm{cp}}}} \left[f_k(s, \pi_{\mathrm{cp}}) - f_k(s, \pi_k) \right] \geq \frac{1}{4}, \quad \forall k \geq 2.$$

Consequently, it follows that $\mathrm{Reg}_K / K = \Omega(1)$.

We refer to the underlying obstruction as ***contextual coupling***: state-wise policy improvements are coupled through a shared parameterization, while the state distribution used to determine the update may differ from the one under which comparator regret is evaluated.

4.2 REGRET DECOMPOSITION VIA COMPATIBLE FUNCTION APPROXIMATION

Proposition 2 motivates us to investigate how contextual coupling manifests in the analysis, and what additional actor–critic structure is sufficient to eliminate it. Rather than committing to a particular contextual mirror-descent update, we consider a generic first-order policy optimization scheme:

$$\theta_{k+1} = \theta_k + \eta v_k, \tag{5}$$

where $v_k \in \mathbb{R}^{\mathsf{d}}$ is an update vector that may depend on the current critic f_k, policy π_k, and offline data. This template includes many familiar policy optimization methods. For instance, policy gradient (PG) chooses v_k according to an on-policy gradient estimate. Our goal is therefore to understand directly what property of v_k is sufficient for controlling the actor-side regret in Eq. (2).

Let $A_k(s, a) := f_k(s, a) - f_k(s, \pi_k)$ denote the critic-induced advantage at round k. Under Assumption 3, the change in log-policy caused by Eq. (5) admits a first-order approximation whose leading term is $v_k^\top \nabla_\theta \log \pi_k(a \mid s)$. This suggests comparing the critic advantage A_k with the linearized policy change induced by v_k, which is the notion of ***compatible function approximation*** (CFA) (Sutton et al., 1999; Kakade, 2001). Therefore we define the CFA error by

$$\mathrm{err}_k = \mathbb{E}_{(s,a) \sim d^{\pi_{\mathrm{cp}}}} \left[A_k(s, a) - v_k^\top \nabla_\theta \log \pi_k(a \mid s) \right], \tag{6}$$

The following lemma shows that this quantity directly governs comparator regret for general parameter-space updates, providing a useful lens for understanding contextual coupling through CFA. Related analyses in the on-policy setting appear in (Even-Dar et al., 2009; Agarwal et al., 2021); to our knowledge, our work is the first to connect this perspective to distribution shift in offline RL.

[4] This sharply contrasts with the squared loss in Fitted-Q or Bellman residual minimization (Munos, 2007; Munos & Szepesvári, 2008; Antos et al., 2008), where the non-negative loss allows for error transfer under coverage conditions.

Lemma 3 (Regret Decomposition Lemma). *Under Assumption 3, consider the update in Eq.* (5) *with* $\|v_k\| \le B_L$ *for all* k*. Then, choosing* $\eta = \sqrt{2D_{\mathrm{KL}}(\pi_{\mathrm{cp}}\|\pi_1)/(\beta K B_L^2)}$*, we have*

$$\frac{\mathrm{Reg}_K}{K} \le B_L\sqrt{\frac{2\beta \cdot D_{\mathrm{KL}}(\pi_{\mathrm{cp}}\|\pi_1)}{K}} + \frac{1}{K}\sum_{k=1}^{K}\mathrm{err}_k.$$

Lemma 3 shifts the focus from reproducing the state-wise mirror-descent update to a more direct question: *can we construct an update* v_k *with small CFA error in the sense relevant to comparator regret?* This provides a general design principle for offline actor updates, which we develop in Section 5 into computationally efficient procedures for policy optimization.

5 Constructing Unified Policy Updates in Parameter Space

In this section we develop two *efficient* actor-update strategies for controlling the CFA error: one based on least-squares regression and the other on distributionally robust optimization.

5.1 Least Square Policy Update

Recall that the CFA error, err_k, quantifies how well the policy gradients $\nabla \log \pi_k(a \mid s)$ can linearly approximate the advantage function $A_k(s,a)$ under $d^{\pi_{\mathrm{cp}}}$. Viewing $\nabla \log \pi_k(a \mid s)$ as *features* and $A_k(s,a)$ as the *regression target*, this observation naturally leads to a noiseless linear regression formulation for constructing the update v_k, which we term ***least square policy update*** (LSPU). Specifically, we define the least-square loss at round k as

$$L_k(v) := \mathbb{E}_{(s,a)\sim d^D}\left[(A_k(s,a) - v^\top \nabla \log \pi_k(a \mid s))^2\right],$$

where the expectation is taken on offline data distribution d^D. In each round k, the actor first computes the advantage function A_k, and then obtains the update v_k by minimizing L_k using samples, within the norm constraint $\|v_k\| \le B_L$. In particular, given $\{(s^{(i)}, a^{(i)})\}_{i=1}^N \overset{\text{i.i.d.}}{\sim} d^D$ from the actor dataset, the constrained regression problem is given by:

$$v_k = \operatorname*{arg\,min}_{v:\|v\|\le B_L} \frac{1}{N}\sum_{i=1}^{N}\Big(A_k(s^{(i)}, a^{(i)}) - v^\top \nabla \log \pi_k(a^{(i)}|s^{(i)})\Big)^2. \tag{7}$$

Error transfer through the coverage condition. A key question is why the regression objective can be formed under the data distribution d^D, even though the CFA error are defined under $d^{\pi_{\mathrm{cp}}}$. The reason is that the squared loss L_k is always *non-negative* (c.f. Section 4.1), allowing us to do distribution transfer via the coverage condition:

$$\mathbb{E}_{d^{\pi_{\mathrm{cp}}}}\left[(A_k - v^\top \nabla \log \pi_k)^2\right] \le \mathsf{C} \cdot \mathbb{E}_{d^D}\left[(A_k - v^\top \nabla \log \pi_k)^2\right].$$

Moreover, since this is a linear regression problem, this density coverage (known as *concentrability coefficient*) can be further improved in the compatible case, i.e., when $A_k = w^\top \nabla \log \pi_k$ for some $w \in \mathbb{R}^{\mathsf{d}}$. Let the covariance matrix be $\Sigma_D^\pi = \mathbb{E}_{d^D}[\nabla \log \pi \nabla \log \pi^\top]$. In this case, we can sharpen the density coverage C (Assumption 1) with the notion of *feature coverage* (Jiang & Xie, 2025), though we omit such an improvement in our main theorems for readability (see Appendix E.3 for details): $\mathsf{C}_{\mathrm{feat}} = \max_{k\in[K]} \mathbb{E}_{d^{\pi_{\mathrm{cp}}}}[\nabla \log \pi_k]^\top (\Sigma_D^{\pi_k})^{-1} \mathbb{E}_{d^{\pi_{\mathrm{cp}}}}[\nabla \log \pi_k]$.

Relation to NPG. This least-square interpretation of policy update can be viewed as a form of the *natural policy gradient* (NPG) method under function approximation, which is well-studied in the on-policy setting (Kakade, 2001; Peters & Schaal, 2008; Agarwal et al., 2021). That said, our formulation differs from canonical on-policy (or off-policy) NPG since LSPU is computed on the offline data distribution d^D without any importance-weight correction. Such design choice follows directly from the decomposition in Lemma 3 and tailors specifically to offline RL.

Actor-critic incompatibility. Even with an accurate critic f_k, the linear regression formulation above need **not** be well-specified. In general, the target A_k does not necessarily lie in the linear span of the features $\nabla \log \pi_k$. Hence we are in the *agnostic learning* setting and any regression-based update may inevitably incur an approximation error, even with infinite data. We capture such misspecification via a quantity ϵ_{CFA} defined below.

Assumption 4 (LSPU Approximation Error)**.** *Define v_k^* as the best linear approximator of the least-square loss L_k. Assume that v_k^* satisfies the norm constraint $\|v_k^*\| \le B_L$[5], and its loss is uniformly bounded by some $\epsilon_{\text{CFA}} \ge 0$, i.e., $L(v_k^*) = \min_{v\in\mathbb{R}^d} L_k(v) \le \epsilon_{\text{CFA}}$.*

The quantity ϵ_{CFA} uniformly measures the extent to which $\nabla \log \pi_k$ can compatibly express A_k at each round. To ensure that ϵ_{CFA} is small, the actor and critic must align with each other. Thus we refer to ϵ_{CFA} as ***actor-critic incompatibility***, which vanishes in the compatible case, i.e. when $A_k = w^\top \nabla \log \pi_k$ for some $w \in \mathbb{R}^d$. For instance, the canonical softmax policy class is fully representative, so $\epsilon_{\text{CFA}} = 0$ for any critic class. In linear function approximation, $\epsilon_{\text{CFA}} = 0$ when both Π_θ and $\mathcal{F}$ share the same feature representation (see Appendix D.1); for neural networks, shared actor-critic representations are also prevalent in deep RL practice (Mnih et al., 2016).

This also sheds light on the contextual coupling challenge in Section 4. One source of the difficulty is that a shared policy parameter may fail to represent state-wise improvement directions consistently across states (in contrast to on-policy setting, e.g., PG). Under exact actor–critic alignment, however, this obstruction can disappear. The following proposition revisits contextual mirror descent in the compatible linear setting and shows that, at the population level, it coincides with the LSPU update.

Proposition 4 (Contextual Mirror Descent under Exact Compatibility)**.** *For invertible $\{\Sigma_D^\pi\}$, suppose $\pi_\theta(a \mid s) \propto \exp(\theta^\top \phi(s,a))$ and $f_k(s,a) = w_k^\top \phi(s,a) + c_k(s)$ for some state-dependent c_k. Then the state-wise mirror descent update satisfies $\pi_{k+1} = \pi_{\theta_k + \eta w_k}$ and is therefore also a solution of Eq.* (4)*. Moreover, contextual mirror descent and population LSPU induce the same policy update.*

We now return to present the regret guarantee for LSPU in the following theorem; the proof can be found in Appendix D.3 and D.4.

Theorem 5 (Main Theorem for LSPU)**.** *Let $\lambda_{\min} > 0$ denote (a deterministic lower bound of) the smallest eigenvalue of $\{\Sigma_D^{\pi_k}\}_{k=1}^K$. Under Assumptions 1, 3, 4, and regularity conditions for linear regression (Assumption 6), Algorithm 2 with policy updates $\{v_k\}$ computed in Eq. 7 achieves the following regret bound with probability at least $1-\delta$:*

$$\frac{\text{Reg}_K}{K} \lesssim B_L \sqrt{\frac{\beta D_{\text{KL}}(\pi_{\text{cp}} \| \pi_1)}{K}} + \sqrt{\mathsf{C}\epsilon_{\text{CFA}}} + G\sqrt{\frac{\mathsf{C}\epsilon_{\text{CFA}} \text{Comp}_L}{\lambda_{\min} N}} + \text{h.o.t.},$$

where Comp_L is some complexity measure on the policy and the value class (defined in Eq. (17)*), and "h.o.t." stands for "higher-order terms" of order $\mathcal{O}(1/N)$.*

Theorem 5 shows that the regret bound of LSPU admits a 3-fold decomposition: an *optimization* term due to the update $\theta_{k+1} = \theta_k + \eta v_k$, a *bias* (approximation error) term determined by the actor-critic incompatibility , and a *statistical estimation* term that decays at a rate of $\mathcal{O}(\sqrt{\mathsf{C}/N})$. In the well-specified case $\epsilon_{\text{CFA}} = 0$, exact empirical regression has zero statistical error on the uniform covariance event, so $K = \mathcal{O}(1/\varepsilon^2)$ rounds suffice once Eq. (16) holds (Appendix D.4).

5.2 Distributionally Robust Policy Update

In Section 5.1, we use a squared surrogate to control the CFA error, which enjoys favorable statistical and computational properties of linear regression. However, it is inherently a relaxation: the squared loss does not directly correspond to the linear form of err_k, and may be loose when the approximation error is highly heterogeneous across the state-action space. This naturally raises a question: can we control the linear CFA error directly? Related ideas have been explored in marginalized importance sampling (Liu et al., 2018; Uehara et al., 2020; Xie & Jiang, 2020). The main obstacle is that err_k is defined under the unknown $d^{\pi_{\text{cp}}}$, hence distribution shift remains.

To address this issue, we adopt a *distributionally robust optimization* (DRO) perspective. The key idea is to express the error under $d^{\pi_{\text{cp}}}$ as an importance-weighted expectation under d^D:

$$\left|\mathbb{E}_{d^{\pi_{\text{cp}}}}[A_k - v_k^\top \nabla \log \pi_k]\right| = \left|\mathbb{E}_{d^D}\left[w^*(s,a)(A_k - v_k^\top \nabla \log \pi_k)\right]\right|, \tag{8}$$

[5]Assumption 4 requires that the global minimizer v_k^*, which admits a closed-form solution, must satisfy $\|v_k^*\| \le B_L$. This can be achieved by setting B_L as a uniform upper bound $GV_{\max}/\lambda_{\min}$, where $\lambda_{\min}$ denotes the smallest eigenvalue of the feature covariance matrices $\{\Sigma_D^{\pi_k}\}_{k=1}^K$.

where w^* is some *correction weight* that allows distribution transfer between $d^{\pi_{\text{cp}}}$ and d^D, such as $w^*(s,a) = d^{\pi_{\text{cp}}}(s,a)/d^D(s,a)$, though alternative forms exist which we will discuss later. Since w^* is unknown, we introduce a robust loss that considers the worst-case weight over a weight class $\mathcal{W}$:

$$\ell_k(v) := \max_{w \in \mathcal{W}} \left| \mathbb{E}_{d^D} \left[w(A_k - v^\top \nabla \log \pi_k) \right] \right|.$$

The weight class $\mathcal{W}$ may be specified explicitly or induced implicitly as a nonparametric space. As long as it satisfies *realizability* (i.e., $w^* \in \mathcal{W}$), we have $\text{err}_k \le \ell_k(v_k)$, and minimizing ℓ_k suffices. We refer to the update that leverages DRO as the ***distributionally robust policy update*** (DRPU).

The $\mathcal{W}_\infty$ class and computation. A natural instantiation of the correction weight w^* is the density ratio $d^{\pi_{\text{cp}}}/d^D$ which directly allows the distribution transfer in Eq. (8), i.e., $w^*(s,a) = d^{\pi_{\text{cp}}}(s,a)/d^D(s,a)$. Correspondingly, we can consider the following *bounded-density-ratio* class:

$$\mathcal{W}_\infty = \Big\{ w : \mathcal{S} \times \mathcal{A} \to [0, \mathsf{C}] : \mathbb{E}_{(s,a)\sim d^D}[w(s,a)] = 1 \Big\}. \tag{9}$$

which includes all valid density ratios due to the normalization constraint. This choice aligns with the density coverage condition (Assumption 1). In particular, realizability holds since $\|w^*\|_\infty = \|d^{\pi_{\text{cp}}}/d^D\|_\infty \le \mathsf{C}$. From now on we instantiate $\mathcal{W} = \mathcal{W}_\infty$ and keep the notation ℓ_k for simplicity.

One key advantage of using $\mathcal{W}_\infty$ lies in its *computational* properties. Under $\mathcal{W}_\infty$ class, the robust loss ℓ_k admits an equivalent dual representation (Proposition 15), which is a version of *Conditional Value-at-Risk* (CVaR) objective:

$$\ell_k(v) = \max_{\mathsf{s}\in\{\pm 1\}} \min_{\tau \in \mathbb{R}} \Big\{ \tau + \mathsf{C}\, \mathbb{E}_{d^D} \left[(\mathsf{s}(A_k - v^\top \nabla \log \pi_k) - \tau)_+ \right] \Big\}$$

where $\mathsf{s} \in \{\pm 1\}$ denotes the sign and $(Z)_+ = \max\{0, Z\}$. The algorithm solves for the empirical counterpart of ℓ_k using samples $\{(s^{(i)}, a^{(i)})\}_{i=1}^N \overset{\text{i.i.d.}}{\sim} d^D$:

$$\hat{\ell}_k(v) = \max_{\mathsf{s}\in\{\pm 1\}} \min_{\tau \in \mathbb{R}} \Big\{ \tau + \frac{\mathsf{C}}{N} \sum_{i=1}^N \big(\mathsf{s}(A_k^{(i)} - v^\top \nabla \log \pi_k(a^{(i)}|s^{(i)})) - \tau \big)_+ \Big\}. \tag{10}$$

Note that $\hat{\ell}_k$ is convex in v. This optimization problem can be reformulated as a d-dimensional linear program (or as a SOCP, depending on the norm constraint of the update v_k), thus admitting efficient numerical algorithms; see Appendix E.2.

Guarantee. As in Section 5.1, we allow for actor-critic incompatibility in the actor update and characterize it through the following assumption. The norm bound is set as $B_L = V_{\max}$ in this section.

Assumption 5 (DRPU Approximation Error)**.** *Define $\tilde{v}_k^*$ as the best linear approximator of the robust loss ℓ_k (within the norm constraint $\|\tilde{v}_k^*\| \le V_{\max}$). Assume its loss is uniformly bounded by some $\widetilde{\epsilon}_{\text{CFA}} \ge 0$ for all k, i.e., $\ell_k(\tilde{v}_k^*) = \min_{v:\|v\|\le V_{\max}} \ell_k(v) \le \widetilde{\epsilon}_{\text{CFA}}$.*

It is worth noting that $\widetilde{\epsilon}_{\text{CFA}}$ is closely related to the bias term ϵ_{CFA} associated with LSPU in Assumption 4. In particular, by the Cauchy-Schwarz inequality, one can show that $\widetilde{\epsilon}_{\text{CFA}} \le \sqrt{\mathsf{C} \cdot \epsilon_{\text{CFA}}}$, a relationship that we will explore further in Section 5.3. As a consequence, **DRPU is more robust to actor-critic incompatibility**; but running the algorithm also relies on a reasonably tight knowledge of the coverage constant C, as also discussed in Xie & Jiang (2020, Section 7).

We now present the regret bound for this DRPU method under $\mathcal{W}_\infty$ class given in Eq. (9). The guarantee exhibits a similar 3-fold structure of Theorem 5, showing that err_k scales as $\mathcal{O}(\sqrt{\mathsf{C}/N})$, up to an additional bias term.

Theorem 6 (Main Theorem for DRPU under $\mathcal{W}_\infty$ Class)**.** *Under Assumptions 1, 3, 5 and the finite-cover conditions of Appendix E.1, $\{v_k\}$ minimizing $\{\widehat{\ell}_k\}$ in Eq.* (10) *achieves the following regret bound with probability at least $1-\delta$:*

$$\frac{\text{Reg}_K}{K} \lesssim V_{\max} \sqrt{\frac{\beta D_{\text{KL}}(\pi_{\text{cp}} \| \pi_1)}{K}} + \widetilde{\epsilon}_{\text{CFA}} + V_{\max}(G+1) \sqrt{\frac{\mathsf{C}\text{Comp}_D}{N}} + \text{h.o.t.},$$

where Comp_D is some complexity measure defined in Eq. 21.

Alternative weight classes. $\mathcal{W}_\infty$ in Eq. (9) is one instantiation of the weight class. More generally, the DRO framework allows for alternative characterizations of $\mathcal{W}$, as long as it satisfies the transfer in Eq. (8). Different choices of $\mathcal{W}$ correspond to different structural assumptions on the distribution shift between $d^{\pi_{\text{cp}}}$ and d^D. For instance, in the compatible case, one can leverage the improved coverage notion of feature coverage in Section 5.1 to define the following chi-square weight class: $\mathcal{W}_{\chi^2} = \left\{w : \mathbb{E}_{(s,a)\sim d^D}[w(s,a)^2] \leq \mathsf{C}_{\text{feat}}\right\}$. It suffices to show that such a weight class contains solution to Eq. (8). Indeed, at round k consider $w^*(s,a) = \mathbb{E}_{d^{\pi_{\text{cp}}}}[\nabla \log \pi_k]^\top (\Sigma_D^{\pi_k})^{-1} \nabla \log \pi_k(a|s)$, which satisfies both Eq.(8) and $w^* \in \mathcal{W}_{\chi^2}$; see Appendix E.3 for details.

5.3 Bias Comparison between LSPU and DRPU

As shown in Theorems 5 and 6, the regret bounds are mainly driven by the intrinsic bias terms when optimization and statistical errors vanish (that is, when sample size N and optimization iterations K are sufficiently large). Such terms are defined differently for LSPU (Assumption 4) and DRPU (Assumption 5). As quantified in Section 5.2, DRPU has potential advantage in terms of robustness to actor-critic incompatibility than LSPU. In this section, we use a case study to illustrate this behavior when $d^D = d^{\pi_{\text{cp}}}$. In fact, perhaps surprisingly, **DRPU can be viewed as behavior cloning in this setting**, providing an interesting unification between offline RL and imitation learning.

No-shift case. To isolate the effect of actor-critic incompatibility, we consider the case where $d^D = d^{\pi_{\text{cp}}}$. This rules out the difficulty of distribution shift and naturally arises in hybrid settings of imitation learning and offline RL, where we have access to additional expert annotations (i.e., data from π_{cp}). In this case, we can choose the weight class as realizable $\mathcal{W} = \{\mathbf{1}\}$ since $d^{\pi_{\text{cp}}}/d^D \equiv 1$. The DRPU then reduces to the following *mean-matching* problem:

$$\min_{v:\|v\|\leq V_{\max}} \left|\mathbb{E}_{d^{\pi_{\text{cp}}}}[A_k] - v^\top \mathbb{E}_{d^{\pi_{\text{cp}}}}[\nabla \log \pi_k]\right|. \tag{11}$$

This formulation highlights a fundamental difference between LSPU and DRPU. While LSPU enforces pointwise control by minimizing a squared loss, DRPU only requires agreement in expectation under a certain distribution $d^{\pi_{\text{cp}}}$. Therefore, Eq. (11) can be driven to exactly zero under mild conditions, even when the actor and the critic are incompatible. We empirically illustrate this phenomenon using a simple MDP (details in Appendix F.2), as shown in Figure 1.

The mean-matching update also admits a natural interpretation. Consider the policy update $\theta_{k+1} = \theta_k + \eta v_k$, where v_k solves Eq. (11). This update can be viewed as performing a steepest descent step[6] within $[-V_{\max}, V_{\max}]$ on the expected KL divergence under $d^{\pi_{\text{cp}}}$, $D_{\text{KL}}(\pi_{\text{cp}}\|\pi_\theta)$, which is exactly the objective of *behavior cloning* (BC). In this sense, the mean-matching view reveals an interesting connection between offline RL and imitation learning at the level of policy updates.

6 Conclusion

In this paper, we study the policy optimization in offline RL. With standalone policy class, a naïve contextualization of state-wise mirror descent can fail due to a challenge we coin *contextual coupling*. To investigate this, we propose a unifying framework based on compatible function approximation, encompassing a regression-based policy update (LSPU) and a distributionally robust alternative (DRPU), both enjoying provable statistical and computational efficiency.

One limitation is that our analysis primarily focuses on explicit stochastic policy classes, such as log-linear policy or Gaussian policy, where log-density-based requirements (e.g., Assumption 3) is well-defined. Extending the theory to deterministic or implicit generative policies (e.g., diffusion policy) remains an important open problem and may require fundamentally different analytical tools, as suggested by recent work on continuous-control learning (Ren et al., 2024; Simchowitz et al., 2025). We view bridging this gap a promising direction for future research.

[6]This holds when $\mathbb{E}_{d^{\pi_{\text{cp}}}}[A_k] > 0$. Roughly speaking, this is the typical situation as it is equivalent to $J(\pi_k) \leq J(\pi_{\text{cp}})$ when the critic is accurate; see Appendix F.1.

AI USE STATEMENT

We used large language model tools (including ChatGPT 5.6) to polish the writing and assist in checking correctness of some theorem statements and proofs. These tools served as auxiliary aids in preparing the manuscript. That said, the authors retain full responsibility for the correctness, originality, and content of the paper.

ETHICS STATEMENT

This is a theoretical work, and we adhere to the ICLR Code of Ethics.

REPRODUCIBILITY STATEMENT

This work is primarily theoretical. Regarding the only one toy experiment in Section 5.3, we have provided details for that in Appendix F.2 and the code in the supplementary material.

REFERENCES

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

András Antos, Csaba Szepesvári, and Rémi Munos. Learning near-optimal policies with bellman-residual minimization based fitted policy iteration and a single sample path. *Machine Learning*, 71 (1):89–129, 2008.

Andrew Bennett, Nathan Kallus, and Miruna Oprescu. Low-rank mdps with continuous action spaces. *arXiv preprint arXiv:2311.03564*, 2023.

Jinglin Chen and Nan Jiang. Information-theoretic considerations in batch reinforcement learning. In *International conference on machine learning*, pp. 1042–1051. PMLR, 2019.

Ching-An Cheng, Tengyang Xie, Nan Jiang, and Alekh Agarwal. Adversarially trained actor critic for offline reinforcement learning. In *International Conference on Machine Learning*, pp. 3852–3878. PMLR, 2022.

Damien Ernst, Pierre Geurts, and Louis Wehenkel. Tree-based batch mode reinforcement learning. *Journal of Machine Learning Research*, 6, 2005.

Eyal Even-Dar, Sham. M. Kakade, and Yishay Mansour. Online markov decision processes. *Mathematics of Operations Research*, 34(3):726–736, 2009.

Amir-massoud Farahmand, Csaba Szepesvári, and Rémi Munos. Error propagation for approximate policy and value iteration. *Advances in neural information processing systems*, 23, 2010.

Aleksandra Faust, Peter Ruymgaart, Molly Salman, Rafael Fierro, and Lydia Tapia. Continuous action reinforcement learning for control-affine systems with unknown dynamics. *IEEE/CAA Journal of automatica Sinica*, 1(3):323–336, 2014.

Maryam Fazel, Rong Ge, Sham Kakade, and Mehran Mesbahi. Global convergence of policy gradient methods for the linear quadratic regulator. In *International conference on machine learning*, pp. 1467–1476. PMLR, 2018.

Scott Fujimoto, David Meger, and Doina Precup. Off-policy deep reinforcement learning without exploration. In *International conference on machine learning*, pp. 2052–2062. PMLR, 2019.

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

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

Daniel Hsu, Sham M Kakade, and Tong Zhang. Random design analysis of ridge regression. In *Conference on learning theory*, pp. 9–1. JMLR Workshop and Conference Proceedings, 2012.

Nan Jiang and Tengyang Xie. Offline reinforcement learning in large state spaces: Algorithms and guarantees. *arXiv preprint arXiv:2510.04088*, 2025.

Nan Jiang, Akshay Krishnamurthy, Alekh Agarwal, John Langford, and Robert E Schapire. Contextual decision processes with low bellman rank are pac-learnable. In *International Conference on Machine Learning*, pp. 1704–1713. PMLR, 2017.

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

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

Nathan Kallus and Masatoshi Uehara. Statistically efficient off-policy policy gradients. In *International Conference on Machine Learning*, pp. 5089–5100. PMLR, 2020.

Walid Krichene, Maximilian Balandat, Claire Tomlin, and Alexandre Bayen. The hedge algorithm on a continuum. In *International Conference on Machine Learning*, pp. 824–832. PMLR, 2015.

Daniel Kuhn, Soroosh Shafiee, and Wolfram Wiesemann. Distributionally robust optimization. *Acta Numerica*, 34:579–804, 2025.

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.

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.

Qiang Liu, Lihong Li, Ziyang Tang, and Dengyong Zhou. Breaking the curse of horizon: Infinite-horizon off-policy estimation. *Advances in neural information processing systems*, 31, 2018.

Yao Liu, Adith Swaminathan, Alekh Agarwal, and Emma Brunskill. Off-policy policy gradient with state distribution correction. *arXiv preprint arXiv:1904.08473*, 2019.

Chenjie Mao. Offline reinforcement learning with additional covering distributions. *arXiv preprint arXiv:2305.12679*, 2023.

Jincheng Mei, Chenjun Xiao, Csaba Szepesvari, and Dale Schuurmans. On the global convergence rates of softmax policy gradient methods. In *International conference on machine learning*, pp. 6820–6829. PMLR, 2020.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin Riedmiller. Playing atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602*, 2013.

Volodymyr Mnih, Adria Puigdomenech Badia, Mehdi Mirza, Alex Graves, Timothy Lillicrap, Tim Harley, David Silver, and Koray Kavukcuoglu. Asynchronous methods for deep reinforcement learning. In *International conference on machine learning*, pp. 1928–1937. PmLR, 2016.

Rémi Munos. Performance bounds in l_p-norm for approximate value iteration. *SIAM journal on control and optimization*, 46(2):541–561, 2007.

Rémi Munos and Csaba Szepesvári. Finite-time bounds for fitted value iteration. *Journal of Machine Learning Research*, 9(5), 2008.

Ofir Nachum, Yinlam Chow, Bo Dai, and Lihong Li. Dualdice: Behavior-agnostic estimation of discounted stationary distribution corrections. *Advances in Neural Information Processing Systems*, 32, 2019.

Francesco Orabona. A modern introduction to online learning. *arXiv preprint arXiv:1912.13213*, 2019.

Jan Peters and Stefan Schaal. Natural actor-critic. *Neurocomputing*, 71(7-9):1180–1190, 2008.

Hamed Rahimian and Sanjay Mehrotra. Distributionally robust optimization: A review. *arXiv preprint arXiv:1908.05659*, 2019.

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

Allen Z Ren, Justin Lidard, Lars L Ankile, Anthony Simeonov, Pulkit Agrawal, Anirudha Majumdar, Benjamin Burchfiel, Hongkai Dai, and Max Simchowitz. Diffusion policy policy optimization. *arXiv preprint arXiv:2409.00588*, 2024.

Max Simchowitz, Daniel Pfrommer, and Ali Jadbabaie. The pitfalls of imitation learning when actions are continuous. *arXiv preprint arXiv:2503.09722*, 2025.

Richard S Sutton, David McAllester, Satinder Singh, and Yishay Mansour. Policy gradient methods for reinforcement learning with function approximation. *Advances in neural information processing systems*, 12, 1999.

Joel A. Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of Computational Mathematics*, 12:389–434, 2012.

Masatoshi Uehara, Jiawei Huang, and Nan Jiang. Minimax weight and q-function learning for off-policy evaluation. In *International Conference on Machine Learning*, pp. 9659–9668. PMLR, 2020.

Hado Van Hasselt and Marco A Wiering. Reinforcement learning in continuous action spaces. In *2007 IEEE International Symposium on Approximate Dynamic Programming and Reinforcement Learning*, pp. 272–279. IEEE, 2007.

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

Tengyang Xie and Nan Jiang. Q* approximation schemes for batch reinforcement learning: A theoretical comparison. In *Conference on Uncertainty in Artificial Intelligence*, pp. 550–559. PMLR, 2020.

Tengyang Xie, Ching-An Cheng, Nan Jiang, Paul Mineiro, and Alekh Agarwal. Bellman-consistent pessimism for offline reinforcement learning. *Advances in neural information processing systems*, 34:6683–6694, 2021.

Lin Yang and Mengdi Wang. Sample-optimal parametric q-learning using linearly additive features. In *International conference on machine learning*, pp. 6995–7004. PMLR, 2019.

Qisen Yang, Shenzhi Wang, Qihang Zhang, Gao Huang, and Shiji Song. Hundreds guide millions: Adaptive offline reinforcement learning with expert guidance. *IEEE Transactions on Neural Networks and Learning Systems*, 35(11):16288–16300, 2023.

Rui Yuan, Simon S Du, Robert M Gower, Alessandro Lazaric, and Lin Xiao. Linear convergence of natural policy gradient methods with log-linear policies. *arXiv preprint arXiv:2210.01400*, 2022.

Yifei Zhou, Ayush Sekhari, Yuda Song, and Wen Sun. Offline data enhanced on-policy policy gradient with provable guarantees. *arXiv preprint arXiv:2311.08384*, 2023.

Algorithm 2 Template for Actor-Critic Policy Optimization

Input: number of iterations K, learning rate η
Initialize policy $\pi_1 = \pi_{\theta_1}$ as any policy in Π_θ
for $k = 1, 2, \ldots, K$ **do**
  Critic: compute f_k using oracle $\mathcal{O}$
  Actor: update policy by $\theta_{k+1} = \theta_k + \eta v_k$
end for
Output: $\hat{\pi} = \text{Unif}[\pi_{1:K}]$

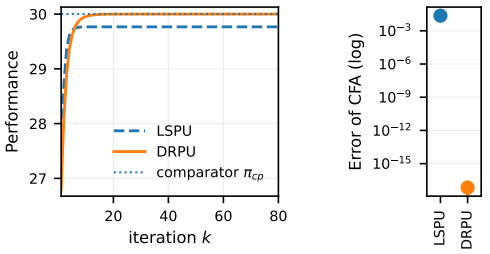


Figure 1: Comparison between LSPU and DRPU under no-shift setting ($d^D = d^{\pi_{\text{cp}}}$). **Left:** performance $J(\pi_k)$ over iterations, where DRPU approaches the comparator performance (*not* optimal), while LSPU plateaus at a worse policy. **Right:** the CFA error, err_k, at iteration $k = 80$ on a log scale, showing that DRPU drives the error close to zero, whereas LSPU incurs a larger bias.

A Related Work

Offline RL studies learning policies from a fixed dataset without further environment interaction (Levine et al., 2020). With function approximation, a dominant line of work builds on value-function estimation and approximate dynamic programming (Ernst et al., 2005; Munos, 2007; Antos et al., 2008; Munos & Szepesvári, 2008; Farahmand et al., 2010; Chen & Jiang, 2019; Xie & Jiang, 2020). This value-centric paradigm underlies many practical algorithms (Mnih et al., 2013) as well as much of the accompanying theory, and provides the standard backdrop for understanding statistical and computational aspects in offline RL. See (Jiang & Xie, 2025) for a comprehensive survey of offline RL theory.

A central difficulty in offline RL is *data coverage*. Existing guarantees typically impose some form of coverage assumption, which can be broadly categorized into (i) *all-policy* coverage, requiring that the dataset sufficiently covers d^π for all policies in a class, and (ii) *single-policy* coverage, which only requires coverage of a particular comparator distribution such as $d^{\pi_{\text{cp}}}$. Algorithmically, two common principles are used to cope with single-policy coverage: behavior regularization, which constrains learned policies to remain close to the behavior distribution (Fujimoto et al., 2019); and pessimism (or uncertainty-aware optimization), which seeks a policy with the best guaranteed performance over models or value functions consistent with the data (Kumar et al., 2020; Jin et al., 2020; Rashidinejad et al., 2021; Xie et al., 2021; Cheng et al., 2022).

Most prior theoretical treatments of offline RL emphasize critic-side learning and treat the actor as an induced object, for example via greedy or soft-greedy policies with respect to learned value functions. In contrast, a smaller but growing literature studies offline RL from an explicit actor–critic or policy-search perspective. The work most closely related to ours is Xie et al. (2021), which induces a soft-greedy policy class with respect to pessimistic value estimates (see Section 3). This structure enables an elegant state-wise mirror-descent interpretation of the actor update, but does not extend to standalone policy classes with arbitrary parameterizations. Subsequent work (Cheng et al., 2022) further develops this actor–critic viewpoint under stronger oracle assumptions on the actor (see Definition 4 of Cheng et al. (2022)) and trains the actor using policy-gradient-style updates based on empirical demonstrations. As we show in this paper, such oracle assumptions are strong, and this formulation can obscure the precise source of actor-side error as well as the role of distribution mismatch.

From a policy-based reinforcement learning perspective, our work also connects to the broader literature on policy gradient (PG) (Williams, 1992) and natural policy gradient (NPG) (Kakade, 2001) methods. On-policy analyses of (natural) policy gradient establish convergence and sample complexity guarantees under smooth parametric policy classes (Fazel et al., 2018; Agarwal et al., 2021; Mei et al., 2020; Yuan et al., 2022), and highlight the importance of distribution shift. In particular, Agarwal et al. (2021) elucidates the connection between NPG, compatible function

approximation (Sutton et al., 1999), and least-squares regression, yielding a regret decomposition into estimation and approximation errors. We build on this policy-optimization viewpoint, but focus on the offline setting, where the distribution mismatch between the data distribution and the visitation distribution of the updated policies fundamentally alters what can be guaranteed. There is also a line of work on off-policy policy gradient methods (Liu et al., 2019; Kallus & Uehara, 2020), including extensions to offline data (Zhou et al., 2023). These approaches are largely rooted in analyses of on-policy policy gradient with importance weighting, whereas our work studies policy optimization directly in the offline setting and explicitly addresses distribution shift through data coverage.

Finally, continuous action spaces further amplify these challenges. Early work such as Antos et al. (2008) extends fitted Q-iteration to continuous actions by explicitly searching over a policy class rather than computing $\arg\max_a Q(s,a)$. Much of the offline RL theory for continuous actions relies on strong structural assumptions, including linear-quadratic regulators, linear MDPs, Gaussian policies, or other restricted dynamics and policy families (Van Hasselt & Wiering, 2007; Faust et al., 2014; Haarnoja et al., 2018; Bennett et al., 2023). In contrast, modern practice often employs flexible standalone policy classes, such as deep neural networks, optimized directly via policy-based methods. This is particularly pronounced in large-scale post-training of language models, where policies are represented by deep networks and updated using PG-type algorithms such as PPO or GRPO. These developments further motivate a principled understanding of policy optimization under distribution shift and function approximation. Our results contribute to this direction by characterizing a fundamental obstruction to naive contextual policy optimization in offline RL and by proposing actor updates with provable guarantees under standalone policy classes.

B Omitted Details for Section 3

B.1 Discussion on the Critic Oracle in Assumption 2

In Assumption 2, we propose a critic oracle $\mathcal{O}$ such that given any policy π, it finds a function $f \in \mathcal{F}$ such that f satisfies the following two conditions:

1. (*Pessimism*) The function f is a pessimistic estimation of Q^π (up to some tolerance $\epsilon_r \geq 0$), i.e., $J_f(\pi) - J(\pi) \leq \epsilon_r/(1-\gamma)$ with high probability, where $J_f(\pi) = \mathbb{E}_{s\sim d_0}[f(s,\pi)]$.
2. (*Bounded Bellman error*) The Bellman error under $d^{\pi_{\text{cp}}}$ (comparator policy's occupancy) is bounded by some $\epsilon_b > 0$. That is, $\mathbb{E}_{d^{\pi_{\text{cp}}}}[\mathcal{T}^\pi f - f] \leq \epsilon_b$ with high probability.

The Bellman error telescoping (Lemma 21) translates the pessimism condition to $\mathbb{E}_{d^\pi}[f - \mathcal{T}^\pi f] = (1-\gamma)(J_f(\pi) - J(\pi)) \leq \epsilon_r$. This means the oracle can control the Bellman error $(f - \mathcal{T}^\pi f)$ under both $d^{\pi_{\text{cp}}}$ and d^π, which corresponds to latter two terms in the generalized performance-difference lemma (Lemma 22). Now we discuss how to realize both of these two conditions if the function class $\mathcal{F}$ satisfies certain structural conditions. Two classical methods include BRM-type (e.g., Xie et al. (2021)) and MQL-type (e.g., Uehara et al. (2020)).

BRM-type critic oracle. The BRM-type oracle leverages classical Bellman residual minimization (BRM) algorithm (Antos et al., 2008) with version-space pessimism. For a policy π and a value function proxy f, define the squared Bellman error (under d^D_{critic}, which is the state-action distribution of the critic-side data) as

$$\mathcal{E}(f;\pi) = \mathbb{E}_{d^D_{\text{critic}}}\left[(f - \mathcal{T}^\pi f)^2\right].$$

By Bellman residual minimization (Antos et al., 2008), this can be written as

$$\mathcal{E}(f;\pi) = \mathcal{L}(f;f,\pi) - \mathcal{L}(\mathcal{T}^\pi f;f,\pi),$$

where

$$\mathcal{L}(f';f,\pi) := \mathbb{E}_{(s,a)\sim d^D_{\text{critic}}, r=R(s,a), s'\sim P(\cdot|s,a)}\left[(f'(s,a) - r - \gamma f(s',\pi))^2\right].$$

Let $\widehat{\mathcal{L}}$ denote the empirical version of $\mathcal{L}$ estimated from the offline dataset $\mathcal{D}$. Under Bellman completeness, i.e., $\mathcal{T}^\pi f \in \mathcal{F}$ for all $f \in \mathcal{F}$, we have that $\mathcal{L}(\mathcal{T}^\pi f;f,\pi) = \min_{g\in\mathcal{F}} \mathcal{L}(g;f,\pi)$. Hence, we can provide an empirical surrogate of the squared Bellman error as

$$\widehat{\mathcal{E}}(f;\pi) = \max_{g\in\mathcal{F}} \widehat{\mathcal{L}}(f;f,\pi) - \widehat{\mathcal{L}}(g;f,\pi).$$

Therefore, the oracle can be given by the following pessimistic estimation of the Bellman error:

$$f^\pi_{\min} = \arg\min_{f \in \mathcal{F}^\pi_{\epsilon_0}} J_f(\pi). \tag{12}$$

The pessimism is taken inside the feasible set $\mathcal{F}^\pi_{\epsilon_0}$, referred as the *version space*, which contains all function $f \in \mathcal{F}$ with small empirical squared Bellman error under d^D (Here $\text{Comp}_{\mathcal{F},\Pi_\theta}$ denotes for some complexity measure of the policy and value class, e.g., $\log(|\mathcal{F}||\Pi_\theta|/\delta)$ for finite classes):

$$\mathcal{F}^\pi_{\epsilon_0} = \left\{ f \in \mathcal{F} : \widehat{\mathcal{E}}(f;\pi) \le \epsilon_0 := \frac{V^2_{\max}}{N} \text{Comp}_{\mathcal{F},\Pi_\theta} \right\}.$$

Since Q^π is the unique fixed point of $\mathcal{T}^\pi$, by standard concentration argument, we have $Q^\pi \in \mathcal{F}^\pi_{\epsilon_0}$ with probability at least $1-\delta$. Therefore, this oracle (Eq. (12)) satisfies the first condition of pessimism ($\epsilon_r = 0$) with high probability:

$$J_f(\pi) = J_{f^\pi_{\min}}(\pi) = \min_{f \in \mathcal{F}^\pi_{\epsilon_0}} J_f(\pi) \le J_{Q^\pi}(\pi) = J(\pi).$$

To see that the oracle (Eq. (12)) also satisfies the second argument of bounded Bellman error under $d^{\pi_{\text{cp}}}$, we need to assume the same coverage condition (Assumption 1) also holds for critic-side data, i.e., $\|d^{\pi_{\text{cp}}}/d^D_{\text{critic}}\| \le \mathsf{C}$ for some finite C. By applying coverage condition (Assumption 1), we can transfer the distribution from $d^{\pi_{\text{cp}}}$ to d^D_{critic} (in addition to some concentration argument to relate $\mathcal{E}(f;\pi)$ with $\widehat{\mathcal{E}}(f;\pi)$): for all $f \in \mathcal{F}^\pi_{\epsilon_0}$,

$$\mathbb{E}_{d^{\pi_{\text{cp}}}}[\mathcal{T}^\pi f - f] \le \sqrt{\mathbb{E}_{d^{\pi_{\text{cp}}}}[(f-\mathcal{T}^\pi f)^2]} \le \sqrt{\mathsf{C} \cdot \mathbb{E}_{d^D_{\text{critic}}}[(f-\mathcal{T}^\pi f)^2]} = \sqrt{\mathsf{C} \cdot \mathcal{E}(f;\pi)} \lesssim \sqrt{\mathsf{C}\epsilon_0} := \epsilon_b.$$

Therefore, the BRM-type oracle in Eq. (12) is a valid oracle that satisfies both conditions in Assumption 2. For computational consideration, we write this as a constrained optimization problem:

$$\min_{f \in \mathcal{F}} J_f(\pi) + \lambda \widehat{\mathcal{E}}(f;\pi),$$

which results in a minimax optimization problem due to the $\max_g$ in the estimation of $\widehat{\mathcal{E}}(f;\pi)$. When $\mathcal{F}$ is a linear function class with respect to some feature $\{\phi_{s,a}\}$, this optimization problem reduces to a quadratic program which can be efficiently solved, as discussed in (Xie et al., 2021).

MQL-type critic oracle. Another implementation of the critic oracle is based on minimax Q-function learning (MQL), a marginalized importance sampling method. It only requires the realizability of function class $\mathcal{F}$ (i.e., $Q^\pi \in \mathcal{F}$ for all $\pi \in \Pi_\theta$) instead of Bellman completeness. The oracle is still given by Eq. (12) but with a different version space $\mathcal{F}^\pi_{\epsilon_0}$ defined as

$$\mathcal{F}^\pi_{\epsilon_0} = \left\{ f \in \mathcal{F} : \max_{w \in \mathcal{W}} \widehat{\ell}(f;w,\pi) \le \epsilon_0' := \sqrt{\frac{V^2_{\max}}{N} \text{Comp}_{\mathcal{F},\Pi_\theta,\mathcal{W}}} \right\},$$

where $\mathcal{W}$ is a weight class similar to what is defined in Section 5.2. And $\widehat{\ell}(f;w,\pi)$ is the empirical version of $\ell(f;w,\pi)$, the average Bellman error of f under weighted distribution $w \cdot d^D_{\text{critic}}$:

$$\ell(f;w,\pi) = \left| \mathbb{E}_{d^D_{\text{critic}}}[w \cdot (f - \mathcal{T}^\pi f)] \right|.$$

Since Q^π is the unique fixed point of $\mathcal{T}^\pi$, $\ell(Q^\pi;w,\pi) = 0$ for all $w \in \mathcal{W}$. By concentration argument, we have $Q^\pi \in \mathcal{F}^\pi_{\epsilon_0}$ with probability at least $1-\delta$. This indicates the pessimism is satisfied when we are taking $\min_{f \in \mathcal{F}^\pi_{\epsilon_0}}$ in Eq. (12), validating the first condition ($\epsilon_r = 0$). Similarly, the second condition of bounded Bellman error under $d^{\pi_{\text{cp}}}$ can be analyzed since $\|d^{\pi_{\text{cp}}}/d^D_{\text{critic}}\|_\infty \le C$ by coverage condition (Assumption 1): for all $f \in \mathcal{F}^\pi_{\epsilon_0}$,

$$\mathbb{E}_{d^{\pi_{\text{cp}}}}[\mathcal{T}^\pi f - f] \le \left| \mathbb{E}_{d^D_{\text{critic}}}[w^*(f - \mathcal{T}^\pi f)] \right| \le \max_{w \in \mathcal{W}} \left| \mathbb{E}_{d^D_{\text{critic}}}[w(f - \mathcal{T}^\pi f)] \right| = \max_{w \in \mathcal{W}} \ell(f;w) \lesssim \epsilon_0' := \epsilon_b,$$

where $w^*(s,a) = d^{\pi_{\text{cp}}}(s,a)/d^D_{\text{critic}}(s,a)$ and $w^* \in \mathcal{W}$ (realizability of the weight class). Therefore, this MQL-type oracle in Eq. (12) also satisfies both conditions in Assumption 2. For computational consideration, it requires an additional weight class $\mathcal{W}$. And this can also be treated as a minimax optimization problem:

$$\min_{f \in \mathcal{F}} \max_{w \in \mathcal{W}} J_f(\pi) + \lambda \widehat{\ell}(f;w,\pi).$$

In special cases like linear function approximation, it can be transformed to some mean-matching problem which can be efficiently solved.

B.2 Settings in Continuous Action Space

For readability, the main text states Section 3 under a finite action space and writes action-wise expectations using sums. In this appendix, we record the corresponding continuous-action formulation of the state-wise mirror-descent argument underlying Theorem 1. This appendix only extends the PSPI analysis in Section 3, the rest of the paper remains stated in the finite-action setting.

Setting. We keep the state space unchanged (since the contextual analysis does not depend on the size of state space) and only generalize the action space. Let $\mathcal{A}$ be a measurable action space equipped with a σ-finite reference measure ν. A policy π is a Markov kernel from states to actions such that, for every state s, the conditional action law $\pi(\cdot \mid s)$ is absolutely continuous with respect to ν. We write its Radon-Nikodym derivative as $\pi(a \mid s)$, so that $\int_{\mathcal{A}} \pi(a \mid s)\nu(\mathrm{d}a) = 1, \forall s \in \mathcal{S}$. Accordingly, whenever a function $g : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is measurable, we write $g(s, \pi) := \int_{\mathcal{A}} g(s, a)\pi(a \mid s)\nu(\mathrm{d}a)$. With this notation, the definition of Bellman operator $(\mathcal{T}^\pi f)(s, a) = R(s, a)+\gamma\mathbb{E}_{s'\sim P(\cdot|s,a)}[f(s', \pi)]$ immediately extends to continuous action spaces as-is.

To avoid ambiguity, it is helpful to distinguish the discounted state occupancy from the discounted state-action occupancy. Denote $d^\pi_{\mathcal{S}}$ the discounted state occupancy measure of π: $d^\pi_{\mathcal{S}}(s) = (1 - \gamma)\sum_{t=0}^{\infty} \gamma^t \Pr_\pi[s_t = s]$ (recall that we still treat states as finite and discrete), and denote by d^π the discounted state-action occupancy measure on $\mathcal{S} \times \mathcal{A}$: $d^\pi(B) = (1 - \gamma)\sum_{t=0}^{\infty} \gamma^t \Pr_\pi[(s_t, a_t) \in B]$ for measurable sets $B \subseteq \mathcal{S} \times \mathcal{A}$. In particular, the regret quantity in Eq. (2) should be interpreted as an expectation over the state marginal $d^{\pi_{\mathrm{cp}}}_{\mathcal{S}}$, namely

$$\frac{\mathrm{Reg}_K}{K} = \frac{1}{K}\sum_{k=1}^{K} \mathbb{E}_{s\sim d^{\pi_{\mathrm{cp}}}_{\mathcal{S}}}[f_k(s, \pi_{\mathrm{cp}}) - f_k(s, \pi_k)].$$

This is the same quantity as in the main text, just written with the state occupancy explicitly separated from the state-action occupancy.

For any two action distributions μ and λ on $\mathcal{A}$ such that μ absolutely continuous with respect to λ, we define their KL-divergence as $D_{\mathrm{KL}}(\mu\|\lambda) = \int_{\mathcal{A}} \log(\mathrm{d}\mu/\mathrm{d}\lambda)\mathrm{d}\mu$. Furthermore, if both μ and λ admit densities p and q, respectively, with respect to ν, the KL reduces to $D_{\mathrm{KL}}(p\|q) = \int_{\mathcal{A}} p(a) \log(p(a)/q(a))\nu(\mathrm{d}a)$. Similar to the main text, we define the expected KL divergence under $d^{\pi_{\mathrm{cp}}}_{\mathcal{S}}$ between policy π_{cp} and π_1 as:

$$D_{\mathrm{KL}}(\pi_{\mathrm{cp}}\|\pi_1) = \mathbb{E}_{s\sim d^{\pi_{\mathrm{cp}}}_{\mathcal{S}}}[D_{\mathrm{KL}}(\pi_{\mathrm{cp}}(\cdot \mid s)\|\pi_1(\cdot \mid s))].$$

Algorithm and guarantee. Under this setup, the PSPI update has the density form

$$\pi_{k+1}(a \mid s) = \frac{\pi_k(a \mid s) \exp(\eta f_k(s, a))}{\int_{\mathcal{A}} \pi_k(a' \mid s) \exp(\eta f_k(s, a'))\nu(\mathrm{d}a')}. \tag{13}$$

This is exactly the continuous analogue of the state-wise multiplicative weights update used in PSPI. The reason this extension is clean is that Section 3 is purely state-wise: once one fixes a state s, the proof is just an online learning argument over the action distribution at that state.

We now state the continuous-action extension of Theorem 1, which exhibits the same structure as that in finite action space. In fact, it strictly subsumes Theorem 1, since the latter can be instantiated by simply choosing the base measure ν as the counting measure over a finite $\mathcal{A}$. Therefore, below we only prove Theorem 7, as Theorem 1 is a direct corollary of it.[7]

Theorem 7 (Theorem 1 with continuous action space)**.** *Assume that for each iteration k, the critic f_k is bounded in $[0, V_{\max}]$, and that $D_{\mathrm{KL}}(\pi_{\mathrm{cp}}\|\pi_1) < \infty$. Then the PSPI iterates satisfy*

$$\frac{\mathrm{Reg}_K}{K} \le \frac{D_{\mathrm{KL}}(\pi_{\mathrm{cp}}\|\pi_1)}{\eta K} + \frac{\eta V^2_{\max}}{8}.$$

By choosing the step size as $\eta = \sqrt{8D_{\mathrm{KL}}(\pi_{\mathrm{cp}}\|\pi_1)/(KV^2_{\max})}$, we have the exactly same regret bound as in Theorem 1 (except here the KL divergence is defined in a continuous manner), i.e., $\mathrm{Reg}_K/K \le V_{\max}\sqrt{D_{\mathrm{KL}}(\pi_{\mathrm{cp}}\|\pi_1)/(2K)}$.

[7] To see that, by choosing the base measure ν as the counting measure in $\mathcal{A}$. Then, all continuous notions defined in this section reduce back to the finite and discrete notions in the main text.

We now give a potential-based proof that utilizes classical online learning tools in the analysis of expert problem (Hazan et al., 2016; Orabona, 2019).

Proof of PSPI regret bound. Let $\xi : L^\infty(\mathcal{A}) \to \mathbb{R}$ be the negative potential function defined as

$$\xi(f) = -\frac{1}{\eta} \log \int_{\mathcal{A}} \pi_1(a \mid s) \exp(\eta f(a)) \nu(\mathrm{d}a).$$

Let $F_k(s,a) = \sum_{j=1}^k f_j(s,a)$ denote the cumulative value function (at each fixed state s) up to round k. We can first bound the difference of the negative potential function as

$$\begin{aligned}
\xi(F_{k+1}(s,\cdot)) - \xi(F_k(s,\cdot)) &= -\frac{1}{\eta} \log \frac{\int_{\mathcal{A}} \pi_1(a \mid s) \exp(\eta \sum_{j=1}^{k+1} f_k(s,a)) \nu(\mathrm{d}a)}{\int_{\mathcal{A}} \pi_1(a \mid s) \exp(\eta \sum_{j=1}^{k} f_k(s,a)) \nu(\mathrm{d}a)} \\
&= -\frac{1}{\eta} \log \int_{\mathcal{A}} \pi_{k+1}(a \mid s) \exp(\eta f_{k+1}(s,a)) \nu(\mathrm{d}a) \\
&= -\frac{1}{\eta} \log \mathbb{E}_{a \sim \pi_{k+1}(\cdot|s)} [\exp(\eta f_{k+1}(s,a))] \\
&\geq -\mathbb{E}_{a \sim \pi_{k+1}(\cdot|s)} [f_{k+1}(s,a)] - \frac{\eta V_{\text{max}}^2}{8},
\end{aligned}$$

where the last step we use the Hoeffding's lemma with random variable $f_{k+1}(s,a)$ (the randomness comes from $a \sim \pi_{k+1}(\cdot \mid s)$) and the range of $f_{k+1}(s,a)$ is $[0, V_{\text{max}}]$. Therefore, by telescoping, we have

$$\sum_{k=1}^K (\xi(F_k(s,\cdot)) - \xi(F_{k-1}(s,\cdot))) \geq -\sum_{k=1}^K \mathbb{E}_{a \sim \pi_k(\cdot|s)} [f_k(s,a)] - \frac{\eta V_{\text{max}}^2}{8} K.$$

Rearranging the terms we have

$$-\sum_{k=1}^K \mathbb{E}_{a \sim \pi_k(\cdot|s)} [f_k(s,a)] \leq \frac{\eta V_{\text{max}}^2}{8} K + \xi(F_K(s,\cdot)),$$

since $\xi(F_0(s,\cdot)) = \xi(0) = -\frac{1}{\eta} \log \int_{\mathcal{A}} \pi_1(a \mid s) \nu(\mathrm{d}a) = -\frac{1}{\eta} \log 1 = 0$ (by setting $\pi_0 \equiv \pi_1$ as the initial policy). By adding $F_K(s, \pi_{\text{cp}}) = \sum_{k=1}^K f_k(s, \pi_{\text{cp}})$ at both sides, we can therefore obtain the regret bound

$$\sum_{k=1}^K (f_k(s, \pi_{\text{cp}}) - f_k(s, \pi_k)) \leq F_K(s, \pi_{\text{cp}}) + \frac{\eta V_{\text{max}}^2}{8} K + \xi(F_K(s,\cdot)).$$

This means we only need to bound the negative log-partition function at round K, i.e., $\xi(F_k(s,\cdot))$. This follows directly via the Gibbs variational principle (Lemma 20) with base measure ν:

$$\xi(F_K(s,\cdot)) = \inf_{u \in \Delta_\nu(\mathcal{A})} \left\{ \frac{1}{\eta} D_{\text{KL}}(u \| \pi_1(\cdot \mid s)) - \mathbb{E}_{a \sim u}[F_K(s,a)] \right\}.$$

Taking $u = \pi_{\text{cp}}(\cdot \mid s)$ and using the fact that $\mathbb{E}_{a \sim \pi_{\text{cp}}(\cdot|s)}[F_K(s,a)] = \sum_{k=1}^K \mathbb{E}_{a \sim \pi_{\text{cp}}(\cdot|s)}[f_k(s,a)]$, we obtain

$$\sum_{k=1}^K (f_k(s, \pi_{\text{cp}}) - f_k(s, \pi_k)) \leq \frac{\eta V_{\text{max}}^2}{8} K + \frac{1}{\eta} D_{\text{KL}}(\pi_{\text{cp}}(\cdot \mid s) \| \pi_1(\cdot \mid s))]).$$

By taking the outer expectation over the designated state distribution $s \sim d_{\mathcal{S}}^{\pi_{\text{cp}}}$ and tuning the step size as $\eta = \sqrt{(8 D_{\text{KL}}(\pi_{\text{cp}} \| \pi_1))/(K V_{\text{max}}^2)}$, we achieve the regret bound in Theorem 1. □

B.3 Examples of Bounded KL Term in Theorem 7

We deliberately keep the KL term in Theorem 1 and Theorem 7, rather than replacing it by a crude worse-case upper bound. In the discrete setting, one often upper-bounds this term by $\log |\mathcal{A}|$, but such a reduction is neither natural nor informative in general continuous action spaces. More importantly, the KL term itself admits elegant and interpretable instantiations for many standard continuous policy classes. Below we provide two examples to further illustrate this point.

Gaussian policies. We first consider Gaussian policies, under which the KL term $D_{\mathrm{KL}}(\pi_{\mathrm{cp}}\|\pi_1)$ admits explicit closed-form expressions and can be further bounded under mild regularity conditions. For the sake of simplicity, we focus on the isotropic Gaussian case, which is widely used in continuous-control reinforcement learning due to its simplicity, numerical stability, and compatibility with policy-gradient methods. In particular, isotropic Gaussian policies arise naturally when actions are modeled as affine functions of the state with additive exploration noise, and they are commonly adopted in practical algorithms (e.g., actor-critic methods with fixed variance) (Haarnoja et al., 2018). Concretely, we assume that for each state s,

$$\pi_i(\cdot \mid s) = \mathcal{N}(\mu_i(s), \sigma^2 I), \qquad i \in \{1, \mathrm{cp}\}.$$

Then, for any fixed s, the KL divergence reduces to a quadratic form in the mean difference:

$$\begin{aligned} D_{\mathrm{KL}}\big(\pi_{\mathrm{cp}}(\cdot \mid s)\|\pi_1(\cdot \mid s)\big) &= \mathbb{E}_{a\sim\pi_{\mathrm{cp}}(\cdot|s)}\left[\log\frac{\pi_{\mathrm{cp}}(a \mid s)}{\pi_1(a \mid s)}\right] \\ &= \mathbb{E}_a\left[-\frac{1}{2}(a-\mu_{\mathrm{cp}}(s))^\top(\sigma^2 I)^{-1}(a-\mu_{\mathrm{cp}}(s)) + \frac{1}{2}(a-\mu_1(s))^\top(\sigma^2 I)^{-1}(a-\mu_1(s))\right] \\ &= \frac{1}{2\sigma^2}\mathbb{E}_a\left[\|a-\mu_1(s)\|_2^2 - \|a-\mu_{\mathrm{cp}}(s)\|_2^2\right] \\ &= \frac{1}{2\sigma^2}\|\mu_{\mathrm{cp}}(s)-\mu_1(s)\|_2^2. \end{aligned}$$

Therefore,

$$D_{\mathrm{KL}}(\pi_{\mathrm{cp}}\|\pi_1) = \mathbb{E}_{s\sim d^{\pi_{\mathrm{cp}}}}\left[D_{\mathrm{KL}}\left(\pi_{\mathrm{cp}}(\cdot \mid s)\|\pi_1(\cdot \mid s)\right)\right] = \frac{1}{2\sigma^2}\mathbb{E}_{s\sim d^{\pi_{\mathrm{cp}}}}\left[\|\mu_{\mathrm{cp}}(s)-\mu_1(s)\|_2^2\right].$$

Moreover, if the mean difference is uniformly bounded, i.e., $\|\mu_{\mathrm{cp}}(s)-\mu_1(s)\|_2 \le B$ for all s, then

$$D_{\mathrm{KL}}(\pi_{\mathrm{cp}}\|\pi_1) \le \frac{B^2}{2\sigma^2}.$$

One important advantage of Gaussian policies is that they admit **efficient computation**. That is, the update does not involve integrating the whole (continuous) action space, but only updates the sufficient statistics.[8] To be specific, we can consider the function class $\mathcal{F}$ as a class of quadratic functions (which is typical in settings like linear quadratic regulators, LQR). Suppose the pessimistic Q-function at round k can be written as (with respect to a)

$$f_k(s,a) = -\frac{1}{2}(a-u_k)^\top Q_k(a-u_k) + c_k,$$

with state-dependent parameters u_k, Q_k, c_k with $Q_k \succ 0$. In this case the multiplicative-weight update in Eq. (1) preserves the Gaussian family: the density $\pi_k(a \mid s)$ remains a multivariate Gaussian $\mathcal{N}(\mu_k, \Sigma_k)$ after the update. Its sufficient statistics are the mean μ_k and covariance Σ_k, which can be equivalently represented by the precision matrix $\Lambda_k = \Sigma_k^{-1}$ and the natural parameter $h_k = \Sigma_k^{-1}\mu_k$. The Hedge update then reduces to

$$\Lambda_{k+1} = \Lambda_k + \eta Q_k, \quad h_{k+1} = h_k + \eta Q_k u_k,$$

so that the updated Gaussian is given by $\mathcal{N}(\mu_{k+1}, \Sigma_{k+1})$ with $\mu_{k+1} = \Lambda_{k+1}^{-1}h_{k+1}$ and $\Sigma_{k+1} = \Lambda_{k+1}^{-1}$.

Convex action space. As claimed in Section 3, under additional structural assumptions on the action space (i.e., convexity) and the function class (i.e., Lipschitzness), we can obtain unified bound that does *not* depend on the KL divergence of any specific policies. The key idea is to leverage the convexity of the action space $\mathcal{A}$, which allows us to construct a sequence of subsets $\mathcal{A}_k \subset \mathcal{A}$ that progressively approximate the deterministic action chosen by π_{cp}. Such a structure can also be extended to notions of *uniform-fatness*, where every action is guaranteed to have sufficient volume in its neighborhood (see Krichene et al. (2015) for more discussion).

[8] This is not restricted to Gaussian policies. In general, for policies that can be expressed as a distribution in exponential family, we can all update them via sufficient statistics.

Theorem 8 (Unified KL Bound with Convex Action Space for Theorem 7)**.** *Suppose the action space $\mathcal{A}$ has finite measure with respect to the Lebesgue measure ν, i.e., $\nu(\mathcal{A}) < \infty$. Furthermore, assume that the action space $\mathcal{A} \subset \mathbb{R}^{\mathsf{d}_\mathcal{A}}$ is convex and compact with diameter B in the $\mathsf{d}_\mathcal{A}$-dimensional Euclidean space $\mathbb{R}^{\mathsf{d}_\mathcal{A}}$, and that $f_k(s,\cdot)$ is L-Lipschitz for all k, then the PSPI update with initial distribution $\pi_1(\cdot \mid s)$ being Lebesgue uniform for all $s \in \mathcal{S}$, i.e., $\pi_1(a \mid s) \equiv 1/\nu(\mathcal{A})$ for all $a \in \mathcal{A}$, and step size $\eta = \sqrt{(8\mathsf{d}_\mathcal{A} \log K)/(KV_{\max}^2)}$ achieves*

$$\frac{\mathrm{Reg}_K}{K} \le V_{\max}\sqrt{\frac{\mathsf{d}_\mathcal{A} \log K}{2K}} + \frac{LB}{K}.$$

Note that the additional $\mathsf{d}_\mathcal{A} \log K$ factor in Theorem 8 (compared to regular $\sqrt{\log|\mathcal{A}|/K}$ regret for Hedge algorithm) can be viewed as the $\log|\mathcal{A}|$ term in the discrete case: the algorithm is effectively *learning from a finite cover* of $\mathcal{A}$. For a $\mathsf{d}_\mathcal{A}$-dimensional set S, its log-covering number (see Definition 1 for formal definition) scales as $\log \mathcal{N}(S,\epsilon) \asymp \mathsf{d}_\mathcal{A} \log(1/\epsilon)$; choosing $\epsilon = 1/\sqrt{K}$ (which is the rate of the no-regret algorithm) yields the $\mathsf{d}_\mathcal{A} \log K$ term. This is why the dimension $\mathsf{d}_\mathcal{A}$ can be viewed as some effective dimension of $\mathcal{A}$, as actions in $\mathcal{A}$ can be embedded as the $\mathsf{d}_\mathcal{A}$-dimensional vector in $\mathbb{R}^{\mathsf{d}_\mathcal{A}}$.

Proof of Theorem 8. Fix an arbitrary state s. Let $F_K(s,a) = \sum_{k=1}^K f_k(s,a)$ denote the cumulative value function. Since each $f_k(s,\cdot)$ is Lipschitz and $\mathcal{A}$ is compact, $F_K(s,\cdot)$ attains its maximum on $\mathcal{A}$. Let $a^* \in \arg\max_{a\in\mathcal{A}} F_K(s,a)$. Then, for any comparator policy $\pi_{\mathrm{cp}}(\cdot \mid s)$, possibly deterministic, we have

$$\begin{aligned}\sum_{k=1}^K (f_k(s,\pi_{\mathrm{cp}}) - f_k(s,\pi_k)) &= F_K(s,\pi_{\mathrm{cp}}) - \sum_{k=1}^K f_k(s,\pi_k) \\ &\le F_K(s,a^*) - \sum_{k=1}^K f_k(s,\pi_k),\end{aligned}$$

where the inequality follows because the expectation of $F_K(s,\cdot)$ under any probability measure is upper bounded by its maximum over $\mathcal{A}$.

The convexity of the action space makes it possible to construct a sequence of action subsets $\mathcal{A}_k \subset \mathcal{A}$ such that

$$\mathcal{A}_k = \{a^* + d_k(a - a^*) : a \in \mathcal{A}\},$$

where $d_k > 0$ is called the *decay rate* (will be specified later). This is because the diameter of $\mathcal{A}_k$ will decay according to d_k, i.e., $\sup_{a,a'\in\mathcal{A}_k} \|a - a'\| = d_k \sup_{a,a'\in\mathcal{A}} \|a - a'\| = d_k B$, where B is the diameter of the action space $\mathcal{A}$.

Notice that all of the $\mathcal{A}_k$ satisfy *realizability*, i.e., $a^* \in \mathcal{A}_k$ for all $k \in [K]$. Specifically we consider $\mathcal{A}_k$, the action subset at the last round. By Lipschitzness of $f_k(s,\cdot)$, for all $a \in \mathcal{A}_k$, we have

$$|f_k(s,a) - f_k(s,a^*)| \le L\|a^* - a\| = d_K LB.$$

This means $f_k(s,a) \ge f_k(s,a^*) - d_K LB$. Summing over $k \in [K]$, we have the cumulative bound

$$F_K(s,a) = \sum_{k=1}^K f_k(s,a) \ge \sum_{k=1}^K f_k(s,a^*) - Kd_K LB = F_K(s,a^*) - Kd_K LB.$$

Recall that in the proof of Theorem 1, we need to control the last-iterate potential function $\xi(F_K(s,\cdot))$:

$$\begin{aligned}\xi(F_K^s) &= -\frac{1}{\eta}\log\int_\mathcal{A} \pi_1(a \mid s)\exp(\eta F_K^s(a))\nu(\mathrm{d}a) \\ &\le -\frac{1}{\eta}\log\int_{\mathcal{A}_k} \pi_1(a \mid s)\exp(\eta F_K(s,a))\nu(\mathrm{d}a) \\ &\le -\frac{1}{\eta}\log\int_{\mathcal{A}_k} \pi_1(a \mid s)\exp(\eta(F_K(s,a^*) - Kd_K LB))\nu(\mathrm{d}a) \\ &= (Kd_K LB - F_K(s,a^*)) - \frac{1}{\eta}\log\int_{\mathcal{A}_k} \pi_1(a \mid s)\nu(\mathrm{d}a).\end{aligned}$$

Therefore, we can get the regret

$$\mathrm{Reg}_K^{\mathrm{s}} \le F_K(s,a^*) + \frac{\eta V_{\max}^2}{8}K + \xi(F_K(s,\cdot)) \le \frac{\eta V_{\max}^2}{8}K + K d_K L B - \frac{1}{\eta}\log \int_{\mathcal{A}_k} \pi_1(a \mid s)\nu(\mathrm{d}a).$$

We only need to control the last term, which is actually the *generalized volume* of $\mathcal{A}_k$. For a set $\mathcal{A}_0 \subset \mathcal{A}$, its generalized volume with respect to distribution $\pi_1(\cdot \mid s) \in \Delta_\nu(\mathcal{A})$ is defined as

$$\mathcal{V}_{\pi_1(\cdot|s)}(\mathcal{A}_0) = \int_{\mathcal{A}_0} \pi_1(a \mid s)\nu(\mathrm{d}a) = \frac{1}{\nu(\mathcal{A})}\int_{\mathcal{A}_k} \nu(\mathrm{d}a) = \frac{\nu(\mathcal{A}_k)}{\nu(\mathcal{A})}$$

where the last equality is due to $\pi_1(a \mid s) = 1/\nu(\mathcal{A})$ is the Lebesgue uniform measure on $\mathcal{A}$. According to our construction of $\mathcal{A}_k$, using change of variables,

$$\begin{aligned}\nu(\mathcal{A}_k) &= \int_{\mathcal{A}_k} \nu(\mathrm{d}a) = \int_{\mathcal{A}} \mathrm{d}\nu(a^* + d_K(a-a^*)) = \int_{\mathcal{A}} |\det(d_K I_{\mathsf{d}_{\mathcal{A}}})|\nu(\mathrm{d}a) \\ &= (d_K)^{\mathsf{d}_{\mathcal{A}}}\int_A \nu(\mathrm{d}a) = (d_K)^{\mathsf{d}_{\mathcal{A}}}\nu(\mathcal{A}).\end{aligned}$$

This means the generalized volume of $\mathcal{A}_k$ is $(d_K)^{\mathsf{d}_{\mathcal{A}}}$. Hence, by setting the decay rate $d_K = 1/K$, we have

$$\sum_{k=1}^{K}(f_k(s,\pi_{\mathrm{cp}}) - f_k(s,\pi_k)) \le \frac{\eta V_{\max}^2}{8}K + LB + \frac{\mathsf{d}_{\mathcal{A}}}{\eta}\log K.$$

Taking the expectation over the designated state distribution $s \sim d_{\mathcal{S}}^{\pi_{\mathrm{cp}}}$ with optimally tuned step size η to be $\eta = \sqrt{8\mathsf{d}_{\mathcal{A}}\log K/(KV_{\max}^2)}$. This leads to the argument in Theorem 8. □

C Omitted Details for Section 4

C.1 Proof of Proposition 2

We construct a two-state contextual bandit (i.e., a one-step MDP with $\gamma = 0$), together with a smooth log-linear policy class and a data-weighted contextual mirror descent update, such that the resulting iterates incur constant per-step comparator regret under $d^{\pi_{\mathrm{cp}}}$.

Proof of Proposition 2. We construct the MDP first. For simplicity, we consider the following contextual bandit problem. Let $\mathcal{S} = \{s_1, s_2\}$ and $\mathcal{A} = \{0, 1\}$. Since $\gamma = 0$, the return equals the immediate reward. Define the reward function

$$R(s,1) = 1, \qquad R(s,0) = 0, \qquad \forall s \in \mathcal{S}.$$

Then for any policy π, the Q-function equals the reward, and we set the oracle outputs to be exact:

$$f_k(s,a) \equiv Q^{\pi_k}(s,a) = R(s,a) \in [0,1], \qquad \forall k \ge 1.$$

Consequently,

$$f_k(s,\pi) = \mathbb{E}_{a\sim\pi(\cdot|s)}[f_k(s,a)] = \pi(1 \mid s), \qquad \forall s \in \mathcal{S},\ \forall k \ge 1.$$

Let the initial distribution be $d_0 = \delta_{s_2}$, so the comparator state distribution is concentrated on s_2:

$$d^{\pi_{\mathrm{cp}}}(s_1) = 0, \qquad d^{\pi_{\mathrm{cp}}}(s_2) = 1.$$

Let the offline data state-marginal distribution be

$$d^D(s_1) = 1 - \varepsilon, \qquad d^D(s_2) = \varepsilon,$$

for some $\varepsilon \in (0, \frac{1}{2})$, and specify $d^D(a \mid s) = 1/2$ for both actions and states. These are arbitrary i.i.d. state–action data as allowed in Section 2, not necessarily trajectories from d_0. For the comparator defined below, $\pi_{\mathrm{cp}}(1 \mid s_2) = 3/4$, so the joint density ratios at $(s_2, 1)$ and $(s_2, 0)$ are $3/(2\varepsilon)$ and $1/(2\varepsilon)$, respectively; both ratios at s_1 are zero. Hence

$$\left\|\frac{d^{\pi_{\mathrm{cp}}}}{d^D}\right\|_\infty = \frac{3}{2\varepsilon}, \qquad \mathsf{C} = \frac{2}{\varepsilon}$$

is a valid conservative coverage bound and DRPU radius.

We consider the one-dimensional log-linear (a.k.a., linear softmax) policy class $\Pi_\theta = \{\pi_\theta : \theta \in \mathbb{R}\}$ defined by

$$\pi_\theta(1 \mid s) = \frac{\exp(\theta x(s))}{\exp(\theta x(s)) + \exp(0)} = \sigma(\theta x(s)), \qquad \pi_\theta(0 \mid s) = 1 - \pi_\theta(1 \mid s),$$

where $x(s_1) = +1$, $x(s_2) = -1$, and $\sigma(u) = \frac{1}{1+e^{-u}}$ is the logistic function. Note that this parameterization induces the coupling identity

$$\pi_\theta(1 \mid s_2) = \sigma(-\theta) = 1 - \sigma(\theta) = 1 - \pi_\theta(1 \mid s_1).$$

Moreover, for any (s, a), $\log \pi_\theta(a \mid s)$ is differentiable with

$$\left|\partial_\theta \log \pi_\theta(a \mid s)\right| \le 1, \qquad \left|\partial_\theta^2 \log \pi_\theta(a \mid s)\right| \le \frac{1}{4},$$

so Π_θ satisfies Assumption 3 (e.g., with $\|\cdot\| = \|\cdot\|_2$, $G = 1$, and $\beta = 1/4$).

Initialize π_1 to be uniform over $\mathcal{A}$, i.e., $\pi_1(1 \mid s) = \frac{1}{2}$ for both states, which corresponds to $\theta_1 = 0$. At each round k, consider the data-weighted contextual mirror descent update obtained by replacing $d^{\pi_{\text{cp}}}$ with d^D *both* in the linear term and in the context-weighted KL regularizer:

$$\pi_{k+1} \in \arg\max_{\pi \in \Pi_\theta} \left\{ \mathbb{E}_{s \sim d^D}\left[f_k(s, \pi)\right] - \frac{1}{\eta} \mathbb{E}_{s \sim d^D}\left[D_{\text{KL}}\left(\pi(\cdot \mid s) \| \pi_k(\cdot \mid s)\right)\right] \right\}. \tag{14}$$

Now we are going to give the regret lower bound. Let

$$p(\theta) := \pi_\theta(1 \mid s_1) = \sigma(\theta), \qquad \pi_\theta(1 \mid s_2) = 1 - p(\theta).$$

Denote $p_k := p(\theta_k)$. Because $D_{\text{KL}}(\text{Ber}(1-p) \| \text{Ber}(1-q)) = D_{\text{KL}}(\text{Ber}(p) \| \text{Ber}(q))$, the objective in Eq. (14) can be written as a concave function of $p \in (0, 1)$:

$$\mathbb{E}_{s \sim d^D}[f_k(s, \pi_\theta)] - \frac{1}{\eta} \mathbb{E}_{s \sim d^D}\left[D_{\text{KL}}\left(\pi_\theta(\cdot \mid s) \| \pi_{\theta_k}(\cdot \mid s)\right)\right] = \varepsilon + (1 - 2\varepsilon)p - \frac{1}{\eta} D_{\text{KL}}\left(\text{Ber}(p) \| \text{Ber}(p_k)\right),$$

where $D_{\text{KL}}(\text{Ber}(p) \| \text{Ber}(q)) = p \log \frac{p}{q} + (1-p) \log \frac{1-p}{1-q}$. Since $-D_{\text{KL}}(\text{Ber}(p) \| \text{Ber}(p_k))$ is concave in p and the first term is linear, the maximizer is characterized by the first-order condition:

$$0 = (1 - 2\varepsilon) - \frac{1}{\eta}\left(\log \frac{p}{p_k} - \log \frac{1-p}{1-p_k}\right) \quad \iff \quad \log \frac{p}{1-p} = \log \frac{p_k}{1-p_k} + \eta(1 - 2\varepsilon).$$

Therefore, writing $\text{logit}(p) = \log \frac{p}{1-p}$, the update satisfies

$$\text{logit}(p_{k+1}) = \text{logit}(p_k) + \eta(1 - 2\varepsilon), \qquad \text{hence} \qquad p_{k+1} = \sigma\left(\text{logit}(p_k) + \eta(1 - 2\varepsilon)\right). \tag{15}$$

Since $\varepsilon \in (0, \frac{1}{2})$ implies $1 - 2\varepsilon > 0$ and $p_1 = 1/2$, Eq. (15) yields

$$p_k \ge \frac{1}{2} \ \forall k \ge 1, \qquad \text{and} \qquad p_k > \frac{1}{2} \ \forall k \ge 2.$$

Let the comparator policy $\pi_{\text{cp}} \in \Pi_\theta$ be $\theta = -\log 3$. This means $\pi_{\text{cp}}(1 \mid s_1) = 1/4$ and $\pi_{\text{cp}}(1 \mid s_2) = 3/4$. Since the comparator occupancy puts all its mass to the state s_2, we have $\mathbb{E}_{s \sim d^{\pi_{\text{cp}}}}[f_k(s, \pi_{\text{cp}})] = f_k(s_2, \pi_{\text{cp}}) = \pi_{\text{cp}}(1 \mid s_2) = 3/4$. While on the other hand, $\mathbb{E}_{s \sim d^{\pi_{\text{cp}}}}[f_k(s, \pi_k)] = f_k(s_2, \pi_k) = \pi_k(1 \mid s_2) = 1 - p_k$. This indicates that

$$\mathbb{E}_{s \sim d^{\pi_{\text{cp}}}}\left[f_k(s, \pi_{\text{cp}}) - f_k(s, \pi_k)\right] = \frac{3}{4} - (1 - p_k) = p_k - \frac{1}{4} \ge \frac{1}{4}.$$

Thus the per-step regret is bounded below by $1/4$ for all $k \ge 2$, and consequently

$$\frac{\text{Reg}_K}{K} = \frac{1}{K} \sum_{k=1}^{K} \mathbb{E}_{s \sim d^{\pi_{\text{cp}}}}\left[f_k(s, \pi_{\text{cp}}) - f_k(s, \pi_k)\right] \ge \Omega(1),$$

which proves the proposition. □

C.2 Proof of Lemma 3

Proof of Lemma 3. With Assumption 3, $\log \pi_\theta(a \mid s)$ is β-smooth in θ with respect to some norm $\|\cdot\|$. That is, for all $(s,a) \in \mathcal{S} \times \mathcal{A}$, we can bound the first-order approximation error between arbitrary $\theta, \theta' \in \mathbb{R}^{\mathsf{d}}$ as

$$-\frac{\beta}{2}\|\theta' - \theta\|^2 \le \log \pi_{\theta'}(a \mid s) - \log \pi_\theta(a \mid s) - \langle \nabla \log \pi_\theta(a \mid s), \theta' - \theta \rangle \le \frac{\beta}{2}\|\theta' - \theta\|^2.$$

Applying the update rule $\theta_{k+1} = \theta_k + \eta v_k$ and substituting $\theta' = \theta_{k+1}$, $\theta = \theta_k$, we obtain (by taking the negative side):

$$\begin{aligned}\log \frac{\pi_{k+1}(a \mid s)}{\pi_k(a \mid s)} &\ge \left(\nabla_\theta \log \pi_k(a \mid s)\right)^\top (\theta_{k+1} - \theta_k) - \frac{\beta}{2}\|\theta_{k+1} - \theta_k\|^2 \\ &\ge \eta \left(\nabla_\theta \log \pi_k(a \mid s)\right)^\top v_k - \frac{\beta}{2}\eta^2 \|v_k\|^2.\end{aligned}$$

Recall the definition of $\mathrm{err}_k = \mathbb{E}_{(s,a)\sim d^{\pi_{\mathrm{cp}}}}\left[A_k(s,a) - \left(\nabla_\theta \log \pi_k(a \mid s)\right)^\top v_k\right]$, we have

$$\mathbb{E}_{(s,a)\sim d^{\pi_{\mathrm{cp}}}}\left[\left(\nabla_\theta \log \pi_k(a \mid s)\right)^\top v_k\right] = \mathbb{E}_{(s,a)\sim d^{\pi_{\mathrm{cp}}}}\left[A_k(s,a)\right] - \mathrm{err}_k.$$

Hence, by the definition of KL divergence $D_{\mathrm{KL}}(p\|q) = \mathbb{E}_p[\log(p/q)]$, we have

$$\begin{aligned}&\mathbb{E}_{s\sim d^{\pi_{\mathrm{cp}}}}\left[D_{\mathrm{KL}}(\pi_{\mathrm{cp}}(\cdot \mid s)\|\pi_k(\cdot \mid s)) - D_{\mathrm{KL}}(\pi_{\mathrm{cp}}(\cdot \mid s)\|\pi_{k+1}(\cdot \mid s))\right] \\ &= \mathbb{E}_{s\sim d^{\pi_{\mathrm{cp}}}}\left[\mathbb{E}_{a\sim\pi_{\mathrm{cp}}(\cdot|s)}\left[\log \frac{\pi_{k+1}(a \mid s)}{\pi_k(a \mid s)}\right]\right] \\ &\ge \mathbb{E}_{(s,a)\sim d^{\pi_{\mathrm{cp}}}}\left[\eta \left(\nabla_\theta \log \pi_k(a \mid s)\right)^\top v_k - \frac{\beta}{2}\eta^2\|v_k\|^2\right] \\ &\ge \eta \mathbb{E}_{(s,a)\sim d^{\pi_{\mathrm{cp}}}}\left[A_k(s,a)\right] - \eta \cdot \mathrm{err}_k - \frac{\beta}{2}\eta^2\|v_k\|^2,\end{aligned}$$

where the first inequality is due to the previous display of the smoothness, and the second inequality is due to the rewrite of err_k. Since $A_k(s,a) = f_k(s,a) - f_k(s,\pi_k)$, rearranging the terms, we get

$$\begin{aligned}\mathbb{E}_{s\sim d^{\pi_{\mathrm{cp}}}}\left[f_k(s,\pi_{\mathrm{cp}}) - f_k(s,\pi_k)\right] = \mathbb{E}_{(s,a)\sim d^{\pi_{\mathrm{cp}}}}[A_k(s,a)] &\le \mathrm{err}_k + \frac{\beta}{2}\eta\|v_k\|^2 \\ &+ \frac{1}{\eta}\mathbb{E}_{s\sim d^{\pi_{\mathrm{cp}}}}\left[D_{\mathrm{KL}}(\pi_{\mathrm{cp}}(\cdot \mid s)\|\pi_k(\cdot \mid s)) - D_{\mathrm{KL}}(\pi_{\mathrm{cp}}(\cdot \mid s)\|\pi_{k+1}(\cdot \mid s))\right].\end{aligned}$$

Telescoping the last term of KL divergence, we get

$$\mathrm{Reg}_K \le \sum_{k=1}^K \mathrm{err}_k + \eta \cdot \frac{\beta}{2}\sum_{k=1}^K \|v_k\|^2 + \frac{1}{\eta} \cdot \mathbb{E}_{s\sim d^{\pi_{\mathrm{cp}}}}\left[D_{\mathrm{KL}}(\pi_{\mathrm{cp}}(\cdot \mid s)\|\pi_1(\cdot \mid s))\right].$$

Since the update direction v_k satisfies $\|v_k\| \le B_L$ for all $k \in [K]$, choosing the step size optimally as $\eta = \sqrt{\frac{2D_{\mathrm{KL}}(\pi_{\mathrm{cp}}\|\pi_1)}{\beta K B_L^2}}$, yields the regret bound

$$\frac{\mathrm{Reg}_K}{K} \le \frac{1}{K}\sum_{k=1}^K \mathrm{err}_k + B_L\sqrt{\frac{2\beta \cdot D_{\mathrm{KL}}(\pi_{\mathrm{cp}}\|\pi_1)}{K}},$$

where $D_{\mathrm{KL}}(\pi_{\mathrm{cp}}\|\pi_1) = \mathbb{E}_{s\sim d^{\pi_{\mathrm{cp}}}}[D_{\mathrm{KL}}(\pi_{\mathrm{cp}}(\cdot \mid s)\|\pi_1(\cdot \mid s))]$ is the expected KL divergence under the comparator occupancy $d^{\pi_{\mathrm{cp}}}$. □

D Omitted Details for Section 5.1

In this section, we establish the performance guarantee of Algorithm 2 along with LSPU update. We will decompose the error of compatible function approximation (CFA) into the bias term ϵ_{CFA} and the estimation error term ϵ_{stat}. And then we utilize results from Hsu et al. (2012) to give a sharp analysis on linear regression, which controls the estimation error ϵ_{stat}. Finally, we will give a more rigorous analysis that is based on log-covering number (i.e., metric entropy) rather than log-cardinality of the function or policy class.

D.1 Two Cases with Zero Actor-Critic Incompatibility

Taking a detour, we first show that the intrinsic bias term ϵ_{CFA} vanishes in two special cases: (i) under the canonical softmax policy class with any function class $\mathcal{F}$, and (ii) in the linear function approximation setting (Jin et al., 2020; Yang & Wang, 2019; Jiang et al., 2017) under the log-linear policy class. In general, if we achieves "compatible" function approximation (i.e., the actor and the critic aligns compatibly), then $\epsilon_{\text{CFA}} = 0$.

Proposition 9 (No Bias under Canonical Softmax Policy)**.** *Consider the softmax policy class*

$$\pi_\theta(a \mid s) = \frac{\exp(\theta(s,a))}{\sum_{a' \in \mathcal{A}} \exp(\theta(s,a'))},$$

where $\theta(s,a)$ denotes the (s,a)-th component of the parameter $\theta \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$. Then, for any advantage function $A_k(s,a)$, there exists a parameter update direction v_k^{n} such that

$$A_k(s,a) = (v_k^{\mathsf{n}})^\top \nabla_\theta \log \pi_k(a \mid s) \quad \textit{for all } (s,a) \in \mathcal{S} \times \mathcal{A}.$$

Consequently, the function approximation is well-specified, and the intrinsic bias term in Assumption 4 satisfies $\epsilon_{\text{CFA}} = 0$.

Proof. Under the softmax parameterization, the log-policy gradient takes the form

$$\nabla_\theta \log \pi_k(a \mid s) = \nabla_\theta \theta(s,a) - \mathbb{E}_{a' \sim \pi_k(\cdot|s)}[\nabla_\theta \theta(s,a')].$$

Since $\theta(s,a)$ can be any real-valued function over (s,a), the gradient basis $\nabla_\theta \log \pi_k(a \mid s)$ spans all zero-mean functions with respect to $\pi_k(\cdot \mid s)$. The advantage function $A_k(s,a)$ also satisfies $\mathbb{E}_{a\sim\pi_k(\cdot|s)}[A_k(s,a)] = 0$ by definition. Therefore, A_k lies exactly in the span of $\nabla_\theta \log \pi_k(a \mid s)$, implying that there exists a vector v_k^{n} achieving zero regression error. Moreover, any function of the form $A_k(s,a) + c_k$, where c_k is a state-dependent constant, is also a minimizer, since

$$\mathbb{E}_{a\sim\pi_k(\cdot|s)}[\nabla_\theta \log \pi_k(a \mid s)] = \sum_{a\in\mathcal{A}} \pi_k(a \mid s) \nabla_\theta \log \pi_k(a \mid s) = \nabla_\theta \sum_{a\in\mathcal{A}} \pi_k(a \mid s) = \nabla_\theta 1 = 0.$$

Hence, under the canonical softmax policy class, we have $\epsilon_{\text{CFA}} = 0$. □

Proposition 10 (No Bias under Compatible Log-Linear Policy)**.** *Assume a log-linear policy*

$$\pi_\theta(a \mid s) = \frac{\exp(\theta^\top \phi(s,a))}{\sum_{a' \in \mathcal{A}} \exp(\theta^\top \phi(s,a'))},$$

and a linear function class $\mathcal{F} = \{f_w(s,a) = \phi(s,a)^\top w : w \in \mathbb{R}^{\mathsf{d}}\}$ with the same feature map $\phi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^{\mathsf{d}}$. At iteration k, let $f_k \in \mathcal{F}$ and define $A_k(s,a)$. Then there exists a parameter update direction $v_k^{\mathsf{n}} \in \mathbb{R}^{\mathsf{d}}$ such that

$$A_k(s,a) = v_k^{\mathsf{n}\top} \nabla_\theta \log \pi_k(a \mid s) \quad \textit{for all } (s,a) \in \mathcal{S} \times \mathcal{A},$$

and hence the function approximation is realizable with $\epsilon_{\text{CFA}} = 0$ in Assumption 4.

Proof. Fix an iteration k and a state s. Define $Z(s,\theta) = \sum_{a' \in \mathcal{A}} \exp(\theta^\top \phi(s,a'))$. Then $\log \pi_\theta(a \mid s) = \theta^\top \phi(s,a) - \log Z(s,\theta)$, whose gradient is

$$\nabla_\theta \log \pi_\theta(a \mid s) = \phi(s,a) - \mathbb{E}_{a' \sim \pi_\theta(\cdot|s)}[\phi(s,a')].$$

Substituting $\theta = \theta_k$ gives $\nabla_\theta \log \pi_k(a \mid s)$. On the critic side, $f_k(s,a) = \phi(s,a)^\top w_k$ for some $w_k \in \mathbb{R}^{\mathsf{d}}$, so the advantage

$$A_k(s,a) = f_k(s,a) - f_k(s,\pi_k) = w_k^\top \left(\phi(s,a) - \mathbb{E}_{a' \sim \pi_k(\cdot|s)}[\phi(s,a')]\right).$$

Comparing both expressions shows $A_k(s,a) = w_k^\top \nabla_\theta \log \pi_k(a \mid s)$ for all (s,a). Hence the regression model is exactly realizable with $v_k^{\mathsf{n}} = w_k$, producing zero residual and $\epsilon_{\text{CFA}} = 0$. □

D.2 Proof of Proposition 4

Proof of Proposition 4. For any fixed state s, the state-wise mirror-descent update has the closed-form solution

$$\pi_{k+1}(a \mid s) \propto \pi_k(a \mid s) \exp\big(\eta f_k(s,a)\big).$$

Since $\pi_k(a \mid s) \propto \exp(\theta_k^\top \phi(s,a))$ and $f_k(s,a) = w_k^\top \phi(s,a) + c_k(s)$, we have

$$\pi_{k+1}(a \mid s) \propto \exp\big((\theta_k + \eta w_k)^\top \phi(s,a)\big),$$

where the state-dependent term $\exp(\eta c_k(s))$ cancels under normalization. Hence $\pi_{k+1} = \pi_{\theta_k + \eta w_k}$. Since this policy maximizes the mirror-descent objective pointwise for every s and belongs to Π_θ, it also maximizes their d^D-weighted average in Eq. (4).

It remains to connect this update to LSPU. For the log-linear policy,

$$\nabla \log \pi_k(a \mid s) = \phi(s,a) - \mathbb{E}_{a' \sim \pi_k(\cdot|s)}[\phi(s,a')],$$

while

$$A_k(s,a) = f_k(s,a) - f_k(s,\pi_k) = w_k^\top \nabla \log \pi_k(a \mid s).$$

Therefore the population least-square loss satisfies $L_k(w_k) = 0$. Since the feature covariance matrix $\Sigma_D^{\pi_k}$ is invertible, the population minimizer is unique, so LSPU chooses $v_k = w_k$. Its parameter update is therefore $\theta_{k+1} = \theta_k + \eta w_k$, which induces exactly the same policy update as contextual mirror descent. □

D.3 Decomposed Regret Bound and Bounding the Estimation Error

In this section we prepare for proving Theorem 5. First, we can decompose the CFA error into a bias term ϵ_{CFA} (in Assumption 4) and an estimation error term ϵ_{stat}, multiplied by some coverage constant. Recall that the least-square loss L_k at round k is given by

$$L_k(v) = \mathbb{E}_{(s,a)\sim d^D}\big[(A_k(s,a) - v^\top \nabla \log \pi_k(a \mid s))^2\big].$$

And the bias term and the estimation error term in Section 5.1 are respectively defined as

$$L_k(v_k^*) = \min_{v:\|v\| \le B_L} L_k(v) \le \epsilon_{\text{CFA}}, \qquad L_k(v_k) - L_k(v_k^*) \le \epsilon_{\text{stat}}, \qquad \forall k \in [K].$$

The following lemma formalizes this fact of decomposition.

Lemma 11 (Decomposed Regret Bound for LSPU)**.** *Under Assumptions 1, 3, and 4, the update in Eq.* (5) *with step size* $\eta = \sqrt{2D_{\text{KL}}(\pi_{\text{cp}}\|\pi_1)/(\beta K B_L^2)}$ *achieves the following regret bound:*

$$\frac{\text{Reg}_K}{K} \le B_L \sqrt{\frac{2\beta D_{\text{KL}}(\pi_{\text{cp}}\|\pi_1)}{K}} + \sqrt{\mathsf{C}}\Big(\sqrt{\epsilon_{\text{CFA}}} + \sqrt{\epsilon_{\text{stat}}}\Big).$$

Proof. Using the regret decomposition lemma (Lemma 3), we have

$$\frac{\text{Reg}_K}{K} \le B_L \sqrt{\frac{2\beta D_{\text{KL}}(\pi_{\text{cp}}\|\pi_1)}{K}} + \frac{1}{K}\sum_{k=1}^K \text{err}_k.$$

We make the following decomposition of err_k:

$$\text{err}_k = \mathbb{E}_{(s,a)\sim d^{\pi_{\text{cp}}}}\left[A_k(s,a) - (v_k^*)^\top \nabla_\theta \log \pi_k(a \mid s)\right] + \mathbb{E}_{(s,a)\sim d^{\pi_{\text{cp}}}}\left[(v_k^* - v_k)^\top \nabla_\theta \log \pi_k(a \mid s)\right].$$

By the coverage assumption that $\|d^{\pi_{\text{cp}}}/d^D\|_\infty \le \mathsf{C}$, we can translate the error to the offline data distribution d^D and bound the first term with

$$\begin{aligned}
\mathbb{E}_{(s,a)\sim d^{\pi_{\text{cp}}}}\left[A_k(s,a) - (v_k^*)^\top \nabla_\theta \log \pi_k(a \mid s)\right] &\le \sqrt{\mathbb{E}_{(s,a)\sim d^{\pi_{\text{cp}}}}\left[\left(A_k(s,a) - (v_k^*)^\top \nabla_\theta \log \pi_k(a \mid s)\right)^2\right]} \\
&\le \sqrt{\mathsf{C} \cdot \mathbb{E}_{(s,a)\sim d^D}\left[\left(A_k(s,a) - (v_k^*)^\top \nabla_\theta \log \pi_k(a \mid s)\right)^2\right]} \\
&= \sqrt{\mathsf{C} \cdot L_k(v_k^*)} \le \sqrt{\mathsf{C} \cdot \epsilon_{\text{CFA}}}.
\end{aligned}$$

For the second term, we note that v_k^* is the minimizer of $L_k(v)$ over the set $\mathcal{V} = \{v : \|v\| \le B_L\}$ (since we assume $\|v_k^*\| \le B_L$ in Assumption 4). Hence for any v such that $\|v\| \le B_L$, the first-order optimality condition for v_k^* imply that

$$(v - v_k^*)^\top \nabla_v L_k(v_k^*) \ge 0.$$

Therefore, for any v such that $\|v\| \le B_L$, we have

$$\begin{aligned}
L_k(v) - L_k(v_k^*) &= \mathbb{E}_{d^D}\left[(A_k - \phi_k^\top v)^2\right] - \mathbb{E}_{d^D}\left[(A_k - \phi_k^\top v_k^*)^2\right] \\
&= \mathbb{E}_{d^D}\left[(A_k - \phi_k^\top v_k^* + \phi_k^\top v_k^* - \phi_k^\top v)^2\right] - \mathbb{E}_{d^D}\left[(A_k - \phi_k^\top v_k^*)^2\right] \\
&= \mathbb{E}_{d^D}\left[\left(\phi_k^\top (v_k^* - v)\right)^2\right] - 2\mathbb{E}_{d^D}\left[\phi_k^\top (v - v_k^*)(A_k - \phi_k^\top v_k^*)\right] \\
&= \|v_k^* - v\|_{\Sigma_D}^2 + (v - v_k^*)^\top \nabla_v L_k(v_k^*) \\
&\ge \|v_k^* - v\|_{\Sigma_D}^2,
\end{aligned}$$

where we use ϕ_k to denote the feature $\nabla_\theta \log \pi_k$, and $\Sigma_D = \mathbb{E}_{d^D}[\phi_k \phi_k^\top]$. The last inequality is due to the first-order optimality condition stated before. By taking $v = v_k$ and the coverage condition, we have

$$\begin{aligned}
\mathbb{E}_{(s,a)\sim d^{\pi_{\text{cp}}}}\left[(v_k^* - v_k)^\top \nabla_\theta \log \pi_k(a \mid s)\right] &\le \sqrt{\mathbb{E}_{(s,a)\sim d^{\pi_{\text{cp}}}}\left[\left((v_k^* - v_k)^\top \nabla_\theta \log \pi_k(a \mid s)\right)^2\right]} \\
&\le \sqrt{\mathsf{C} \cdot \mathbb{E}_{(s,a)\sim d^{D}}\left[\left((v_k^* - v_k)^\top \nabla_\theta \log \pi_k(a \mid s)\right)^2\right]} \\
&= \sqrt{\mathsf{C} \cdot \|v_k^* - v_k\|_{\Sigma_D}^2} \\
&\le \sqrt{\mathsf{C} \cdot (L_k(v_k) - L_k(v_k^*))} \le \sqrt{\mathsf{C} \cdot \epsilon_{\text{stat}}}.
\end{aligned}$$

Combining those two terms, we get that $\text{err}_k \le \sqrt{\mathsf{C}}(\sqrt{\epsilon_{\text{CFA}}} + \sqrt{\epsilon_{\text{stat}}})$. Substituting into the regret bound, we get Lemma 11:

$$\frac{\text{Reg}_K}{K} \le B_L \sqrt{\frac{2\beta D_{\text{KL}}(\pi_{\text{cp}} \| \pi_1)}{K}} + \sqrt{\mathsf{C}}\left(\sqrt{\epsilon_{\text{CFA}}} + \sqrt{\epsilon_{\text{stat}}}\right).$$

□

We establish a single uniform event for the actor dataset, allowing the same N i.i.d. pairs from d^D to be reused at every round. Throughout this analysis, norms are Euclidean (and matrix norms are the induced operator norms). Write $\mathbb{E} = \mathbb{E}_{d^D}$ and $\widehat{\mathbb{E}}$ for its empirical average, and, for a critic–parameter pair, define

$$A_{f,\theta} = f - f(\cdot, \pi_\theta), \qquad \phi_\theta = \nabla \log \pi_\theta, \qquad \Sigma_\theta = \mathbb{E}[\phi_\theta \phi_\theta^\top].$$

Fix a deterministic initialization and step size, and let $\Theta_0 = \{\theta : \|\theta - \theta_1\|_2 \le \eta B_L K\}$. Bounded updates imply that every iterate lies in this region; no projection of the policy parameter is required. Assumption 3 applies on this region and its line segments. We assume $D_{\text{KL}}(\pi_{\text{cp}} \| \pi_1) < \infty$.

Assumption 6 (Data Model for Uniform Linear Regression)**.** *Fix a deterministic $\lambda_{\min} > 0$ and let $\Theta_\lambda = \{\theta \in \Theta_0 : \Sigma_\theta \succeq \lambda_{\min} I\}$. There is a deterministic class $\mathcal{U} \subseteq \mathcal{F} \times \Theta_\lambda$ containing every generated pair (f_k, θ_k). For every $(f, \theta) \in \mathcal{U}$, the unconstrained population solution and its residual satisfy*

$$\begin{aligned}
&v_{f,\theta}^* = \Sigma_\theta^{-1} \mathbb{E}[\phi_\theta A_{f,\theta}], \qquad \epsilon_{f,\theta} = A_{f,\theta} - (v_{f,\theta}^*)^\top \phi_\theta, \\
&\|v_{f,\theta}^*\|_2 \le B_L, \qquad \mathbb{E}[\epsilon_{f,\theta}^2] \le \epsilon_{\text{CFA}}, \qquad \|\epsilon_{f,\theta}\|_\infty \le b.
\end{aligned}$$

Here $b \le B := V_{\max} + GB_L$ is deterministic; taking $b = B$ always suffices. All classes are measurable and separable in the displayed uniform metrics, and the covering numbers used below are finite.

This is the uniform form of the feasible unconstrained optimum in Assumption 4. It imposes nondegeneracy only on the admissible parameter set, not on every policy. Normal equations give $\mathbb{E}[\phi_\theta \epsilon_{f,\theta}] = 0$. The advantage has no additional Monte Carlo label noise; the argument does not assume it is determined by the score feature alone.

Lemma 12 (Estimation Error Decomposition)**.** *Fix an admissible pair, suppress its subscripts, and suppose $\widehat{\Sigma} \succeq \Sigma/2$. If $\widehat{v}$ is feasible and its empirical squared loss is within $\zeta \geq 0$ of the constrained minimum, then*

$$\|\widehat{v} - v^*\|_\Sigma \leq \frac{4}{\sqrt{\lambda_{\min}}}\|\widehat{\mathbb{E}}[\phi\epsilon]\|_2 + \sqrt{2\zeta}, \qquad L(\widehat{v}) - L(v^*) = \|\widehat{v} - v^*\|_\Sigma^2.$$

Proof. Let $\Delta = \widehat{v} - v^*$ and $q = \widehat{\mathbb{E}}[\phi\epsilon]$. Feasibility of v^* and empirical near-optimality imply $\|\Delta\|_{\widehat{\Sigma}}^2 \leq 2\Delta^\top q + \zeta$. Thus $\|\Delta\|_{\widehat{\Sigma}} \leq 2\|q\|_{\widehat{\Sigma}^{-1}} + \sqrt{\zeta}$. Since $\|\Delta\|_\Sigma \leq \sqrt{2}\|\Delta\|_{\widehat{\Sigma}}$ and $\|q\|_{\widehat{\Sigma}^{-1}} \leq \sqrt{2/\lambda_{\min}}\|q\|_2$, the first bound follows. Expanding the population loss and using $\mathbb{E}[\phi\epsilon] = 0$ proves the equality. □

Lemma 13 (Uniform Covariance Bound)**.** *Let M_Σ be the size of a Euclidean $\lambda_{\min}/(16G\beta)$-cover of Θ_λ with centers in that set; when $\beta = 0$, take $M_\Sigma = 1$. If*

$$N \geq \frac{32G^2}{\lambda_{\min}} \log \frac{2\mathsf{d}M_\Sigma}{\delta}, \tag{16}$$

then, with probability at least $1 - \delta/2$, $\widehat{\Sigma}_\theta \succeq \Sigma_\theta/2$ simultaneously for every $\theta \in \Theta_\lambda$.

Proof. At a fixed center $\widetilde{\theta}$, the whitened feature $Y = \Sigma_{\widetilde{\theta}}^{-1/2}\phi_{\widetilde{\theta}}$ satisfies $\mathbb{E}[YY^\top] = I$ and $\|Y\|_2^2 \leq G^2/\lambda_{\min}$. Lemma 23 gives failure probability at most $\mathsf{d}\exp\{-N\lambda_{\min}/(32G^2)\}$ for $\widehat{\Sigma}_{\widetilde{\theta}} \succeq 3\Sigma_{\widetilde{\theta}}/4$. Union-bound over the centers. For $\|\theta - \widetilde{\theta}\|_2 \leq r$, smoothness gives

$$\|\phi_\theta\phi_\theta^\top - \phi_{\widetilde{\theta}}\phi_{\widetilde{\theta}}^\top\|_{\mathrm{op}} \leq 2G\beta r.$$

This bounds both covariance perturbations, so $\widehat{\Sigma}_\theta \succeq 3\Sigma_\theta/4 - (7/2)G\beta r I \succeq \Sigma_\theta/2$ at the stated radius. When $\beta = 0$, the score is constant on the connected region Θ_0, so one center suffices. □

D.4 Proof of Theorem 5

Definition 1 (ε-covering number)**.** *An ε-cover of a metric space $(\mathcal{G}, d)$ is a finite subset of $\mathcal{G}$ within distance ε of every member. Its smallest cardinality is $\mathcal{N}_d(\mathcal{G}, \varepsilon)$. For critics we use the uniform norm; for parameters we use the parameter norm; for vector-valued functions we use $\|h\|_{\infty,2} = \sup_{s,a}\|h(s,a)\|_2$.*

Define the deterministic centered moment class and its entropy by

$$\mathcal{H}_L = \{\phi_\theta\epsilon_{f,\theta} : (f,\theta) \in \mathcal{U}\}, \qquad \xi = GB/N^2, \qquad \mathrm{Comp}_L = \log \frac{2\mathcal{N}_{\infty,2}(\mathcal{H}_L, \xi)}{\delta}. \tag{17}$$

Every member h satisfies $\mathbb{E}[h] = 0$, $\|h\|_{\infty,2} \leq Gb$, and $\mathbb{E}\|h\|_2^2 \leq G^2\epsilon_{\mathrm{CFA}}$. Set

$$T_L = G\sqrt{\frac{\epsilon_{\mathrm{CFA}}}{N}}(1 + \sqrt{8\mathrm{Comp}_L}) + \frac{4Gb\mathrm{Comp}_L}{3N} + \xi. \tag{18}$$

Lemma 14 (Uniform LSPU Estimation Error)**.** *Under Assumptions 3, 6 and Eq.* (16)*, with probability at least $1 - \delta$, simultaneously for all admissible pairs and all empirical ζ-minimizers,*

$$\sqrt{L(\widehat{v}) - L(v^*)} \leq \frac{4T_L}{\sqrt{\lambda_{\min}}} + \sqrt{2\zeta}.$$

Proof. Apply Lemma 24 at each center of the deterministic ξ-cover of $\mathcal{H}_L$, with vector bound Gb and total second moment $NG^2\epsilon_{\mathrm{CFA}}$. A union bound gives, with probability at least $1 - \delta/2$,

$$\sup_{h \in \mathcal{H}_L} \|\widehat{\mathbb{E}}[h]\|_2 \leq G\sqrt{\frac{\epsilon_{\mathrm{CFA}}}{N}}(1 + \sqrt{8\mathrm{Comp}_L}) + \frac{4Gb\mathrm{Comp}_L}{3N} + \xi = T_L.$$

The last term extends the bound from the cover to the whole class. Intersect with Lemma 13, then apply Lemma 12. This event is uniform before the data-dependent critics and policies are selected; empirical optimizers are never transferred to a different regression problem. □

Explicit critic and policy entropy. The complexity in Eq. (17) can be bounded using separate critic and parameter covers. For admissible pairs within distances $\varepsilon_f, \varepsilon_\theta$, differentiating the policy along the parameter segment gives $\sup_s \|\pi_\theta(\cdot|s) - \pi_{\widetilde{\theta}}(\cdot|s)\|_1 \le G\varepsilon_\theta$. Consequently,

$$\|A - \widetilde{A}\|_\infty \le a_0 := 2\varepsilon_f + V_{\max} G\varepsilon_\theta, \qquad \|\phi - \widetilde{\phi}\|_{\infty,2} \le \beta\varepsilon_\theta, \qquad \|\Sigma - \widetilde{\Sigma}\|_{\mathrm{op}} \le 2G\beta\varepsilon_\theta.$$

The normal equations yield $\Sigma(v^* - \widetilde{v}^*) = \mathbb{E}[\phi A - \widetilde{\phi}\widetilde{A}] + (\widetilde{\Sigma} - \Sigma)\widetilde{v}^*$, hence

$$\|v^* - \widetilde{v}^*\|_2 \le \frac{Ga_0 + \beta(V_{\max} + 2GB_L)\varepsilon_\theta}{\lambda_{\min}} =: d_v, \qquad \|\epsilon - \widetilde{\epsilon}\|_\infty \le a_0 + B_L\beta\varepsilon_\theta + Gd_v.$$

Thus $\|h - \widetilde{h}\|_{\infty,2} \le L_f\varepsilon_f + L_\theta\varepsilon_\theta$, where

$$L_f = 2G(1+G^2/\lambda_{\min}), \qquad L_\theta = b\beta + G\left[V_{\max}G + B_L\beta + \frac{G}{\lambda_{\min}}(V_{\max}G^2 + \beta V_{\max} + 2GB_L\beta)\right].$$

Cover $\mathcal{F}$ and Θ_λ at radii $\varepsilon_f/2$ and $\varepsilon_\theta/2$, and select one admissible representative from each nonempty product cell. Representatives remain in $\mathcal{U}$, so their moments are centered and obey the same bounds. Taking $\varepsilon_f = \xi/(2L_f)$ and $\varepsilon_\theta = \xi/(2L_\theta)$ gives

$$\mathcal{N}_{\infty,2}(\mathcal{H}_L, \xi) \le \mathcal{N}_\infty(\mathcal{F}, \varepsilon_f/2)\mathcal{N}_2(\Theta_\lambda, \varepsilon_\theta/2).$$

A zero coefficient requires no corresponding cover. These formulas specify the metrics and scales; boundedness of $\mathcal{F}$ alone does not imply finite uniform entropy.

Proof of Theorem 5. On the event of Lemma 14, density coverage and Cauchy–Schwarz give, simultaneously over rounds,

$$\mathrm{err}_k \le \sqrt{\mathsf{C}\epsilon_{\mathrm{CFA}}} + \sqrt{\mathsf{C}}\|v_k - v_k^*\|_{\Sigma_{\theta_k}} \le \sqrt{\mathsf{C}\epsilon_{\mathrm{CFA}}} + 4\sqrt{\frac{\mathsf{C}}{\lambda_{\min}}}T_L + \sqrt{2\mathsf{C}\zeta_k}.$$

For any fixed $\eta > 0$, the KL-potential calculation in Appendix C.2 therefore gives the explicit bound

$$\frac{\mathrm{Reg}_K}{K} \le \frac{D_{\mathrm{KL}}(\pi_{\mathrm{cp}}\|\pi_1)}{\eta K} + \frac{\beta\eta B_L^2}{2} + \sqrt{\mathsf{C}\epsilon_{\mathrm{CFA}}} + 4\sqrt{\frac{\mathsf{C}}{\lambda_{\min}}}T_L + \frac{1}{K}\sum_{k=1}^{K}\sqrt{2\mathsf{C}\zeta_k}. \tag{19}$$

With exact empirical minimization ($\zeta_k = 0$), the step size of Lemma 3, $b \le B$, and Eq. (18), this proves the main-text bound. Both the Comp_L/N and the covering remainder are retained. If $\epsilon_{\mathrm{CFA}} = 0$, separability makes all admissible residuals zero on a common set of full d^D measure. Then $\widehat{\mathbb{E}}[\phi\epsilon] = 0$ simultaneously, and exact empirical minimization on the uniform covariance event gives $v_k = v_k^*$ and zero statistical error. □

E Omitted Details for Section 5.2

Using the generic regret decomposition lemma (Lemma 3), we can bound the regret term as (by taking $B_L = V_{\max}$):

$$\frac{\mathrm{Reg}_K}{K} \le V_{\max}\sqrt{\frac{2\beta D_{\mathrm{KL}}(\pi_{\mathrm{cp}}\|\pi_1)}{K}} + \widetilde{\epsilon}_{\mathrm{CFA}} + \widetilde{\epsilon}_{\mathrm{stat}}, \tag{20}$$

where $\widetilde{\epsilon}_{\mathrm{CFA}}$ and $\widetilde{\epsilon}_{\mathrm{stat}}$ are similarly the approximation error and the estimation error. Recall:

$$\ell_k(\tilde{v}_k^*) = \min_{v:\|v\|\le V_{\max}} \ell_k(v) \le \widetilde{\epsilon}_{\mathrm{CFA}}, \qquad \ell_k(v_k) - \ell_k(\tilde{v}_k^*) \le \widetilde{\epsilon}_{\mathrm{stat}}, \qquad \forall k \in [K].$$

Since $\widetilde{\epsilon}_{\mathrm{CFA}}$ is given in Assumption 5, our goal is to give a non-asymptotic control of $\widetilde{\epsilon}_{\mathrm{stat}}$.

E.1 Proof of Theorem 6

We first show that under $\mathcal{W}_\infty$ class, the robust optimization problem can essentially be transformed as an CVaR problem, which would validate that DRPU under $\mathcal{W}_\infty$ in Eq. (9) by setting $\alpha = 1/\mathsf{C}$.

Proposition 15. *Let Z be an integrable random variable and $\alpha \in (0, 1]$ be a tail probability level. Then the following equivalence holds:*

$$\sup_{\substack{w:\mathbb{E}[w]=1 \\ 0\le w\le 1/\alpha}} \mathbb{E}[wZ] = \inf_{\tau\in\mathbb{R}} \left\{ \tau + \frac{1}{\alpha}\mathbb{E}[(Z-\tau)_+] \right\} =: \mathrm{CVaR}_{1-\alpha}(Z).$$

Proof. For any feasible w and $\tau \in \mathbb{R}$, we have

$$wZ = w\tau + w(Z-\tau) \le w\tau + \frac{1}{\alpha}(Z-\tau)_+,$$

which is due to $0 \le w \le 1/\alpha$. Taking expectations and using $\mathbb{E}[w] = 1$,

$$\mathbb{E}[wZ] \le \tau + \frac{1}{\alpha}\mathbb{E}[(Z-\tau)_+].$$

Since this holds for every feasible w and every τ,

$$\sup_{\substack{w:\mathbb{E}[w]=1 \\ 0\le w\le 1/\alpha}} \mathbb{E}[wZ] \le \inf_{\tau\in\mathbb{R}} \left\{ \tau + \frac{1}{\alpha}\mathbb{E}[(Z-\tau)_+] \right\}.$$

For $\alpha = 1$, the only feasible weight is $w = 1$, and $\tau + \mathbb{E}[(Z-\tau)_+] = \mathbb{E}[Z] + \mathbb{E}[(\tau - Z)_+] \downarrow \mathbb{E}[Z]$ as $\tau \to -\infty$ by integrability. Thus the identity holds, although a finite minimizing threshold need not exist. For $\alpha \in (0, 1)$, take τ^* to be the $(1-\alpha)$ quantile of Z such that $\Pr(Z > \tau^*) \le \alpha \le \Pr(Z \ge \tau^*)$, and the corresponding w^* is chosen as

$$w^* = \frac{1}{\alpha}\mathbf{1}\{Z > \tau^*\} + \frac{\lambda}{\alpha}\mathbf{1}\{Z = \tau^*\},$$

with $\lambda \in [0, 1]$ chosen so that $\mathbb{E}[w^*] = 1$. That is, if we let $q_1 = \Pr(Z = \tau^*)$, $q_2 = \Pr(Z > \tau^*)$, then $q_2 \le \alpha \le q_1 + q_2$. Since $q_1, q_2 \in [0, 1]$, there must exist some $\lambda \in [0, 1]$ such that $\lambda q_1 + q_2 = \alpha$. Using this construction,

$$\mathbb{E}[w^*Z] = \frac{1}{\alpha}\mathbb{E}[Z \cdot \mathbf{1}\{Z > \tau^*\}] + \frac{\lambda}{\alpha}\mathbb{E}[Z \cdot \mathbf{1}\{Z = \tau^*\}].$$

For the first term, we have $Z \cdot \mathbf{1}\{Z > \tau^*\} = (Z - \tau^*)\mathbf{1}\{Z > \tau^*\} + \tau^*\mathbf{1}\{Z > \tau^*\}$. Taking expectation,

$$\mathbb{E}[Z \cdot \mathbf{1}\{Z > \tau^*\}] = \mathbb{E}[(Z-\tau^*)_+] + \tau^* \Pr(Z > \tau^*) = \mathbb{E}[(Z-\tau^*)_+] + \tau^* q_2.$$

For the second term, it's easy to see that $\mathbb{E}[Z \cdot \mathbf{1}\{Z = \tau^*\}] = \tau^* \Pr(Z = \tau^*) = \tau^* q_1$. Combining the two terms,

$$\mathbb{E}[w^*Z] = \frac{1}{\alpha}\Big(\tau^* q_2 + \mathbb{E}[(Z-\tau^*)_+]\Big) + \frac{\lambda}{\alpha}\tau^* q_1 = \frac{1}{\alpha}\mathbb{E}[(Z-\tau^*)_+] + \frac{\tau^*}{\alpha}(\lambda q_1 + q_2) = \frac{1}{\alpha}\mathbb{E}[(Z-\tau^*)_+] + \tau^*.$$

Therefore, the proposition holds with exact equality.

□

We now give a uniform finite-sample proof. Write $\mathbb{E} = \mathbb{E}_{d^D}$ and $\widehat{\mathbb{E}}$ for its empirical average. Let $\Theta = \{\theta : \|\theta - \theta_1\| \le \eta V_{\max} K\}$ for deterministic initialization and step size, and put $B = V_{\max}(1+G) > 0$ (the case $V_{\max} = 0$ is trivial). Assume finite initial comparator KL. Bounded updates ensure that all iterates lie in Θ. Define the deterministic residual class

$$\mathcal{Z} = \{f - f(\cdot, \pi_\theta) - v^\top \nabla \log \pi_\theta : f \in \mathcal{F},\ \theta \in \Theta,\ \|v\| \le V_{\max}\} \subseteq [-B, B]^{\mathcal{S}\times\mathcal{A}}.$$

Assume the classes are measurable and separable in the uniform metrics below, with finite covers. For $N \ge 1$, set

$$r = \frac{B}{4\mathsf{C}N}, \qquad M_Z = \mathcal{N}_\infty(\mathcal{Z}, r), \qquad M_\tau = \lceil 8\mathsf{C}N \rceil + 1, \qquad \mathrm{Comp}_D = \log \frac{8 M_Z M_\tau}{\delta}. \tag{21}$$

Cover centers belong to the relevant classes. For completeness, Assumption 3 implies

$$\|Z_{f,\theta,v} - Z_{\widetilde{f},\widetilde{\theta},\widetilde{v}}\|_\infty \le 2\|f - \widetilde{f}\|_\infty + V_{\max}(G+\beta)\|\theta - \widetilde{\theta}\| + G\|v - \widetilde{v}\|.$$

Indeed, integrating the score bound along the parameter segment gives $\sup_s \|\pi_\theta(\cdot|s) - \pi_{\widetilde{\theta}}(\cdot|s)\|_1 \le G\|\theta - \widetilde{\theta}\|$, which controls the moving expectation $f(s, \pi_\theta)$. Thus, with $r_f = r/6$, $r_\theta = r/[3V_{\max}(G+\beta)]$ and $r_v = r/(3G)$,

$$M_Z \le \mathcal{N}_\infty(\mathcal{F}, r_f)\mathcal{N}_{\|\cdot\|}(\Theta, r_\theta)(1 + 2V_{\max}/r_v)^{\mathsf{d}}.$$

A zero coefficient needs no corresponding cover. This specifies the critic, parameter, update-vector and threshold contributions, including both signs through Eq. (21). No separate union bound over rounds is needed.

Lemma 16 (Uniform DRPU Estimation Error Bound)**.** *Under the preceding boundedness and finite-cover conditions, with probability at least $1-\delta$, simultaneously for all f, θ, v in the residual class,*

$$|\ell_{f,\theta}(v) - \widehat{\ell}_{f,\theta}(v)| \le \Delta_D := \frac{B}{N} + 5B\sqrt{\frac{\mathsf{CComp}_D}{N}} + 8B\frac{\mathsf{CComp}_D}{N}. \tag{22}$$

Consequently, if v_k is an empirical ζ_k-minimizer over $\|v\| \le V_{\max}$, then $\ell_k(v_k) - \inf_{\|v\|\le V_{\max}} \ell_k(v) \le 2\Delta_D + \zeta_k$ simultaneously for all rounds.

The empirical loss uses the constraint $\widehat{\mathbb{E}}[w] = 1$, whereas the population loss uses $\mathbb{E}[w] = 1$. The proof compares these two normalized CVaR functionals directly.

Lemma 17 (Quantile Characterization of Tail Probability)**.** *For any $Z \in \mathcal{Z}$ and either sign $\mathsf{s} \in \{\pm 1\}$, choose population and empirical minimizing thresholds $\tau^*, \widehat{\tau} \in [-B, B]$ for $\tau + \mathsf{C}\mathbb{E}[(\mathsf{s}Z - \tau)_+]$ and its empirical counterpart. They satisfy*

$$\Pr(\mathsf{s}Z > \tau^*) \le 1/\mathsf{C} \le \Pr(\mathsf{s}Z \ge \tau^*), \qquad \widehat{\Pr}(\mathsf{s}Z > \widehat{\tau}) \le 1/\mathsf{C} \le \widehat{\Pr}(\mathsf{s}Z \ge \widehat{\tau}).$$

Proof. The population subdifferential is $[1 - \mathsf{C}\Pr(\mathsf{s}Z \ge \tau),\ 1 - \mathsf{C}\Pr(\mathsf{s}Z > \tau)]$ and contains zero at a minimizer; replace Pr by $\widehat{\Pr}$ for the empirical statement. Boundedness permits thresholds in $[-B, B]$, including $\tau = -B$ when $\mathsf{C} = 1$. The inequalities handle atoms and ties without requiring equality of tail mass. □

For example, if $z_{(1)} \ge \cdots \ge z_{(N)}$ are the observations of $\mathsf{s}Z$, let $m = \lfloor N/\mathsf{C} \rfloor$ and $q = N/\mathsf{C} - m$. The empirical CVaR is $\frac{\mathsf{C}}{N}(\sum_{j=1}^m z_{(j)} + q z_{(m+1)})$, with the last term omitted when $q = 0$. This includes the fractional boundary weight when N/C is not an integer.

Proof of Lemma 16. Take an r-cover $\{Z_j\}$ of $\mathcal{Z}$ and a grid $\{\tau_l\}$ of $[-B, B]$ with M_τ points and spacing at most r. The normalized hinges

$$\mathcal{H}_0 = \left\{ \frac{(\mathsf{s}Z_j - \tau_l)_+}{2B} : \mathsf{s} \in \{\pm 1\},\ j \le M_Z,\ l \le M_\tau \right\}$$

form a deterministic class of at most $2M_Z M_\tau$ functions in $[0, 1]$. Lemma 25 applies with $\Gamma = \mathrm{Comp}_D$. For every normalized hinge $H = (\mathsf{s}Z - \tau)_+/(2B)$, a cover member H_0 satisfies

$$\|H - H_0\|_\infty \le \nu := \frac{1}{4\mathsf{C}N}, \qquad \|H^2 - H_0^2\|_\infty \le 2\nu.$$

Thus on a single event of probability at least $1 - \delta$, simultaneously for all Z, s, τ,

$$|(\mathbb{E} - \widehat{\mathbb{E}})H| \le 2\nu + \sqrt{\frac{2(\mathbb{E}H^2 + 2\nu)\mathrm{Comp}_D}{N}} + \frac{2\mathrm{Comp}_D}{3N},$$

$$|(\mathbb{E} - \widehat{\mathbb{E}})H| \le 2\nu + 2\sqrt{\frac{(\widehat{\mathbb{E}}H^2 + 2\nu)\mathrm{Comp}_D}{N}} + \frac{4\mathrm{Comp}_D}{N}.$$

This uniform extension is deterministic, so it remains valid for sample-dependent critics, policies, update vectors and thresholds. At a population minimizer, $\mathbb{E}H^2 \le 1/\mathsf{C}$; at an empirical minimizer, $\widehat{\mathbb{E}}H^2 \le 1/\mathsf{C}$, by Lemma 17. Importantly, these are different certificates.

Write $T(Z) = \inf_\tau\{\tau + \mathsf{C}\mathbb{E}[(Z-\tau)_+]\}$ and $\widehat{T}(Z)$ for its empirical counterpart. Optimality gives the two comparisons

$$T(Z) - \widehat{T}(Z) \le \mathsf{C}(\mathbb{E} - \widehat{\mathbb{E}})(Z - \widehat{\tau})_+, \qquad \widehat{T}(Z) - T(Z) \le \mathsf{C}(\widehat{\mathbb{E}} - \mathbb{E})(Z - \tau^*)_+.$$

Use the empirical certificate for the first and the population certificate for the second. Both absolute deviations are bounded by

$$4B\mathsf{C}\nu + 4B\sqrt{\frac{\mathsf{CComp}_D}{N}(1+2\mathsf{C}\nu)} + 8B\frac{\mathsf{CComp}_D}{N} \le \frac{B}{N} + 5B\sqrt{\frac{\mathsf{CComp}_D}{N}} + 8B\frac{\mathsf{CComp}_D}{N}.$$

The same holds for $-Z$. Since $\ell(v) = \max\{T(Z), T(-Z)\}$, the maximum is 1-Lipschitz in its two arguments and Eq. (22) follows. Finally, for any feasible v, empirical near-optimality yields $\ell_k(v_k) \le \widehat{\ell}_k(v_k) + \Delta_D \le \widehat{\ell}_k(v) + \zeta_k + \Delta_D \le \ell_k(v) + \zeta_k + 2\Delta_D$. Taking the infimum proves the excess-loss bound. □

Proof of Theorem 6. Fix a chosen radius $\mathsf{C} \ge 1$ satisfying Assumption 1. The density ratio lies in $\mathcal{W}_\infty$, so $\mathrm{err}_k \le \ell_k(v_k) \le \widetilde{\epsilon}_{\mathrm{CFA}} + 2\Delta_D + \zeta_k$. With finite initial comparator KL, the same KL-potential calculation as in Lemma 3 gives, for any fixed $\eta > 0$, on the uniform event of Lemma 16,

$$\frac{\mathrm{Reg}_K}{K} \le \frac{D_{\mathrm{KL}}(\pi_{\mathrm{cp}}\|\pi_1)}{\eta K} + \frac{\beta\eta V_{\max}^2}{2} + \widetilde{\epsilon}_{\mathrm{CFA}} + 2\Delta_D + \frac{1}{K}\sum_{k=1}^K \zeta_k. \tag{23}$$

For exact minimization and the step size of Lemma 3 with $B_L = V_{\max}$, this proves the main-text theorem. Only when $N \ge \mathsf{CComp}_D$ may the linear term be absorbed into the square-root term. Without this regime the full bound remains valid. The bound need not vanish at finite N when $\widetilde{\epsilon}_{\mathrm{CFA}} = 0$; this is an upper-bound statement, not a claim that DRPU necessarily incurs nonzero estimation error. □

E.2 Computation

In Section 5.2 we mentioned that minimizing the loss $\hat{\ell}_k(v)$ can be viewed as a linear program (or more generally, a SOCP)[9], which can be efficiently solved by any convex solver. Now we give the specific realization of this program. Recall the CVaR expression of $\hat{\ell}_k$ is given by

$$\hat{\ell}_k(v) = \max\left\{\min_{\tau\in\mathbb{R}}\left\{\tau + \frac{\mathsf{C}}{N}\sum_{i=1}^N\left(\epsilon_v^{(i)} - \tau\right)_+\right\}, \min_{\tau\in\mathbb{R}}\left\{\tau + \frac{\mathsf{C}}{N}\sum_{i=1}^N\left(-\epsilon_v^{(i)} - \tau\right)_+\right\}\right\},$$

where $\epsilon_v^{(i)} = A_k^{(i)} - v^\top \nabla \log \pi_k(a^{(i)}|s^{(i)})$. Since the outer max of the two CVaRs with an epigraph variable (similarly for the inner τ, we can write down the corresponding d-dimensional convex program of $\min_{v\in\mathbb{R}^{\mathsf{d}}:\|v\|\le V_{\max}} \hat{\ell}_k(v)$:

$$\begin{aligned}
\min_{\substack{v,z,\tau_+,\tau_-,\\ \{c_+^{(i)}\}_{i=1}^N,\{c_-^{(i)}\}_{i=1}^N}} \quad & z \\
\text{s.t.} \quad & z \ge \tau_+ + \frac{\mathsf{C}}{N}\sum_{i=1}^N c_+^{(i)}, \\
& c_+^{(i)} \ge A_k^{(i)} - v^\top\phi_k^{(i)} - \tau_+, \qquad c_+^{(i)} \ge 0, \quad i = 1,\dots,N, \\
& z \ge \tau_- + \frac{\mathsf{C}}{N}\sum_{i=1}^N c_-^{(i)}, \\
& c_-^{(i)} \ge -A_k^{(i)} + v^\top\phi_k^{(i)} - \tau_-, \qquad c_-^{(i)} \ge 0, \quad i = 1,\dots,N, \\
& \|v\| \le V_{\max}.
\end{aligned}$$

[9] The specific type of the program depends on the norm constraint of v : $\|v\| \le V_{\max}$. That is, if it's $\|\cdot\|_\infty$ or $\|\cdot\|_1$, then it's a linear program (LP); if it's $\|\cdot\|_2$, then it's a second-order cone program (SOCP)

E.3 Feature Coverage and Chi-Square Weight Class

In this section, we discuss the feature coverage condition in Section 5.1 and clarify its relation to the chi-square weight class introduced in Section 5.2.

Feature coverage condition. Recall that in the compatible case, the advantage function $A_k(s,a)$ lies in the linear span of the policy-gradient *features*, which we denote by $\phi_k(s,a) = \nabla \log \pi_k(a \mid s)$. Let $\Sigma_D^\pi = \mathbb{E}_{d^D}[(\nabla \log \pi)(\nabla \log \pi)^\top]$ denote the feature covariance matrix under the data distribution d^D (suppose it's invertible), and let $\mu_{\text{cp}}^\pi = \mathbb{E}_{(s,a)\sim d^{\pi_{\text{cp}}}}[\nabla \log \pi(a \mid s)]$ denote the mean feature under $d^{\pi_{\text{cp}}}$. Then the feature coverage condition C_{feat} can be written as

$$(\mu_{\text{cp}}^{\pi_k})^\top (\Sigma_D^{\pi_k})^{-1} \mu_{\text{cp}}^{\pi_k} \le \mathsf{C}_{\text{feat}}, \qquad \forall k \in [K].$$

Unlike density coverage, which controls distribution shift uniformly over all (s,a) pairs, feature coverage only controls transfer on the restricted function class induced by the policy-gradient features. That is, C_{feat} requires data only to cover a single direction $\mu_{\text{cp}}^{\pi_k}$, which is the mean feature direction under $d^{\pi_{\text{cp}}}$. Since the CFA error is exactly a linear error in the feature class, the feature coverage is the known tightest coverage notion under linear function approximation(Jiang & Xie, 2025).

The next proposition shows that feature coverage is always implied by density coverage, and can therefore be viewed as a refined notion of coverage.

Proposition 18 (Density Coverage Implies Feature Coverage)**.** *Assume Assumption 1 holds, namely* $\|w^*\|_\infty \le \mathsf{C}$ *for* $w^* = d^{\pi_{\text{cp}}}/d^D$. *Then for every* $k \in [K]$ *and every* $f \in \text{span}(\phi_k)$,

$$\big(\mathbb{E}_{d^{\pi_{\text{cp}}}}[f]\big)^2 \le \mathsf{C}\, \mathbb{E}_{d^D}[f^2].$$

Consequently, for invertible $\Sigma_D^{\pi_k}$, *this is equivalent to* $(\mu_{\text{cp}}^{\pi_k})^\top (\Sigma_D^{\pi_k})^{-1} \mu_{\text{cp}}^{\pi_k} \le \mathsf{C}$. *In particular, the feature coverage constant always satisfies* $\mathsf{C}_{\text{feat}} \le \mathsf{C}$.

Proof. Fix any $k \in [K]$ and any $f \in \text{span}(\phi_k)$. Since $w^* = d^{\pi_{\text{cp}}}/d^D$, we have $\mathbb{E}_{d^{\pi_{\text{cp}}}}[f] = \mathbb{E}_{d^D}[w^* f]$. Therefore,

$$\big(\mathbb{E}_{d^{\pi_{\text{cp}}}}[f]\big)^2 = \big(\mathbb{E}_{d^D}[w^* f]\big)^2 \le \mathbb{E}_{d^D}[(w^*)^2]\mathbb{E}_{d^D}[f^2] \le \mathsf{C}\, \mathbb{E}_{d^D}[f^2], \qquad \forall f \in \text{span}(\phi_k),$$

where the first inequality is by Cauchy-Schwarz, and the second inequality leverages the "low-variance" property of the importance weight. That is, for w^* such that $\mathbb{E}_{d^D}[w^*] = 1$ and $\|w^*\|_\infty \le \mathsf{C}$, we have $\mathbb{E}_{d^D}[(w^*)^2] \le \|w^*\|_\infty \mathbb{E}_{d^D}[w^*] \le \mathsf{C}$.

Now write any $f \in \text{span}(\phi_k)$ as $f_\theta(s,a) = \theta^\top \phi_k(s,a)$ for some $\theta \in \mathbb{R}^d$. Then $\mathbb{E}_{d^{\pi_{\text{cp}}}}[f_\theta] = \theta^\top \mu_{\text{cp}}^{\pi_k}$ and $\mathbb{E}_{d^D}[f_\theta^2] = \theta^\top \Sigma_D^{\pi_k} \theta$. Hence,

$$\sup_{\theta \ne 0} \frac{(\theta^\top \mu_{\text{cp}}^{\pi_k})^2}{\theta^\top \Sigma_D^{\pi_k} \theta} \le \mathsf{C}, \quad \forall k \in [K].$$

Since $\Sigma_D^{\pi_k}$ is invertible, the left-hand side is the Rayleigh quotient

$$\sup_{\theta \ne 0} \frac{(\theta^\top \mu_{\text{cp}}^{\pi_k})^2}{\theta^\top \Sigma_D^{\pi_k} \theta} = (\mu_{\text{cp}}^{\pi_k})^\top (\Sigma_D^{\pi_k})^{-1} \mu_{\text{cp}}^{\pi_k},$$

which implies

$$(\mu_{\text{cp}}^{\pi_k})^\top (\Sigma_D^{\pi_k})^{-1} \mu_{\text{cp}}^{\pi_k} \le \mathsf{C}, \quad \forall k \in [K].$$

Taking the supremum over $k \in [K]$ gives $\mathsf{C}_{\text{feat}} \le \mathsf{C}$. □

Proposition 18 shows that feature coverage is never worse than density coverage. More importantly, it can be *substantially smaller*. The reason is that density coverage controls the worst-case reweighting error over the entire state-action space, whereas feature coverage only measures mismatch along the specific feature directions relevant to the actor update.

Chi-square weight class. We now connect the above discussion to the chi-square weight class in Section 5.2. *In the compatible linear case, one need not use the true density ratio for distribution transfer*. Instead, it suffices to use a feature-based correction function that is exact on the linear residual class. Define

$$w_k^*(s,a) := (\mu_{\text{cp}}^{\pi_k})^\top (\Sigma_D^{\pi_k})^{-1} \phi_k(s,a), \tag{24}$$

and consider the chi-square weight class

$$\mathcal{W}_{\chi^2} := \{w : \mathbb{E}_{d^D}[w^2] \le \mathsf{C}_{\text{feat}}\}.$$

Proposition 19. *Assume the compatible case, so that $A_k(s,a) \in \text{span}(\phi_k)$. Then the feature-based correction weight w_k^* defined above satisfies the following two properties:*

*(i) (**Distribution transfer**) for any $f \in \text{span}(\phi_k)$, $\mathbb{E}_{d^{\pi_{\text{cp}}}}[f] = \mathbb{E}_{d^D}[w_k^* f]$.*

*(ii) (**Realizability**) $\mathbb{E}_{d^D}[(w_k^*)^2] = (\mu_{\text{cp}}^{\pi_k})^\top (\Sigma_D^{\pi_k})^{-1} \mu_{\text{cp}}^{\pi_k} \le \mathsf{C}_{\text{feat}}$, and hence $w_k^* \in \mathcal{W}_{\chi^2}$.*

Proof. For any $f \in \text{span}(\phi_k)$, there exists $\theta \in \mathbb{R}^d$ such that $f(s,a) = \theta^\top \phi_k(s,a)$. Then

$$\mathbb{E}_{d^D}[w_k^* f] = (\mu_{\text{cp}}^{\pi_k})^\top (\Sigma_D^{\pi_k})^{-1} \mathbb{E}_{d^D}[\phi_k \phi_k^\top]\theta = (\mu_{\text{cp}}^{\pi_k})^\top \theta = \mathbb{E}_{d^{\pi_{\text{cp}}}}[f],$$

which proves (i). For (ii),

$$\mathbb{E}_{d^D}[(w_k^*)^2] = (\mu_{\text{cp}}^{\pi_k})^\top (\Sigma_D^{\pi_k})^{-1} \mathbb{E}_{d^D}[\phi_k \phi_k^\top] (\Sigma_D^{\pi_k})^{-1} \mu_{\text{cp}}^{\pi_k} = (\mu_{\text{cp}}^{\pi_k})^\top (\Sigma_D^{\pi_k})^{-1} \mu_{\text{cp}}^{\pi_k} \le \mathsf{C}_{\text{feat}}.$$

Thus $w_k^* \in \mathcal{W}_{\chi^2}$. □

Proposition 19 shows that, in the compatible linear case, the chi-square weight class can be instantiated directly by feature coverage: the correction function w_k^* exactly transfers expectations on the linear residual class, and its second moment under d^D is precisely the feature coverage quantity. In this sense, $\mathcal{W}_{\chi^2}$ provides a feature-based analogue of the density-ratio weight class that is tailored to the actor update.

Remark: feature correction weight. It is important to note that the feature-based correction weight w_k^* in Eq. (24) should NOT be interpreted as a density ratio. In general, it need not be nonnegative and need not satisfy the normalization condition $\mathbb{E}_{d^D}[w_k^*] = 1$. This is not an issue for our purpose, since the role of w_k^* is only to realize exact transfer on the residual class $\text{span}(\phi_k)$, rather than to represent a valid change of measure.

The normalization issue can also be handled by a simple centering trick. Let $\mu_D^{\pi_k} = \mathbb{E}_{d^D}[\phi_k]$, $\mu_{\text{cp}}^{\pi_k} = \mathbb{E}_{d^{\pi_{\text{cp}}}}[\phi_k]$, and $\bar{\Sigma}_D^{\pi_k} = \mathbb{E}_{d^D}[(\phi_k - \mu_D^{\pi_k})(\phi_k - \mu_D^{\pi_k})^\top]$. Define

$$\bar{w}_k(s,a) = 1 + (\mu_{\text{cp}}^{\pi_k} - \mu_D^{\pi_k})^\top (\bar{\Sigma}_D^{\pi_k})^{-1} (\phi_k(s,a) - \mu_D^{\pi_k}).$$

Then $\mathbb{E}_{d^D}[\bar{w}_k] = 1$, and the same exact transfer property holds for every $f \in \text{span}(\phi_k)$. Moreover,

$$\mathbb{E}_{d^D}[(\bar{w}_k - 1)^2] = (\mu_{\text{cp}}^{\pi_k} - \mu_D^{\pi_k})^\top (\bar{\Sigma}_D^{\pi_k})^{-1} (\mu_{\text{cp}}^{\pi_k} - \mu_D^{\pi_k}).$$

Thus, normalization can be kept by replacing the uncentered 2nd-moment geometry with a centered covariance geometry. Nonnegativity, however, is a separate requirement and is not needed for restricted transfer over the linear residual class.

This observation highlights a broader distinction between our correction-weight perspective and the traditional DRO or importance-weighting literature. In classical DRO (Kuhn et al., 2025; Rahimian & Mehrotra, 2019), the uncertainty set $\mathcal{W}$ is typically chosen as an f-divergence ball around the data distribution d^D, so its elements correspond to valid probability reweightings. Similarly, importance weighting methods, including marginalized importance sampling (MIS) (Liu et al., 2018; Nachum et al., 2019), use correction weights that are valid density ratios. In contrast, our feature-based correction weights need not define a valid change of measure; they only need to realize the desired transfer over the function class relevant to the actor update. In this sense, correction weights are more general than density ratios.

F Omitted Details for Section 5.3

F.1 Recovery of Behavior Cloning

Recall that with known $d^{\pi_{\text{cp}}}$, let $m_k = \mathbb{E}_{(s,a)\sim d^{\pi_{\text{cp}}}}[A_k(s,a)]$ and $\mu_k = \mathbb{E}_{(s,a)\sim d^{\pi_{\text{cp}}}}[\nabla \log \pi_k(a \mid s)]$, we solve for $\min_{v:\|v\|\le V_{\max}} |m_k - v^\top \mu_k|$ in Eq. (11). There exists a closed-form solution which can be expressed as

$$v_k = \begin{cases} \frac{m_k}{\|\mu_k\|_*} u_k, & \text{if} \quad |m_k| \le V_{\max} \cdot \|\mu_k\|_*, \\ V_{\max} \cdot u_k, & \text{if} \quad m_k > V_{\max} \cdot \|\mu_k\|_*, \\ -V_{\max} \cdot u_k, & \text{if} \quad m_k < -V_{\max} \cdot \|\mu_k\|_*, \end{cases} \tag{25}$$

where $u_k = \arg\max_{u:\|u\|\le 1} u^\top \mu_k$ aligns with the direction of μ_k. This can be viewed as running steepest descent on $\Phi(\theta)$, defined as the expected KL divergence under $d^{\pi_{\text{cp}}}$:

$$\Phi(\theta) = \mathbb{E}_{s\sim d^{\pi_{\text{cp}}}}[D_{\text{KL}}(\pi_{\text{cp}}(\cdot \mid s) \| \pi_\theta(\cdot \mid s))].$$

This is because that the negative gradient of Φ corresponds to feature mean (at round k) μ_k:

$$\begin{aligned} \nabla_\theta \Phi(\theta_k) &= \nabla_\theta \mathbb{E}_{s\sim d^{\pi_{\text{cp}}}} \left[\mathbb{E}_{a\sim\pi_{\text{cp}}(\cdot|s)} \left[\log \frac{\pi_{\text{cp}}(a \mid s)}{\pi_k(a \mid s)} \right] \right] \\ &= \nabla_\theta \mathbb{E}_{(s,a)\sim d^{\pi_{\text{cp}}}} [\log \pi_{\text{cp}}(a \mid s) - \log \pi_k(a \mid s)] \\ &= -\mathbb{E}_{(s,a)\sim d^{\pi_{\text{cp}}}} [\nabla_\theta \log \pi_k(a \mid s)] = -\mu_k. \end{aligned}$$

Sign of advantage. There is a subtlety regarding the sign of the advantage term: we are only updating along the direction of $-\nabla\Phi(\theta_k)$ to minimize the BC objective if $m_k = \mathbb{E}_{d^{\pi_{\text{cp}}}}[A_k] \ge 0$; this is the BC regime described in Section 5.3. However, when $m_k < 0$: the update will *increase* $\Phi(\theta)$ and move away from the comparator. To understand the condition $m_k \ge 0$, consider the case where f_k is accurate, i.e., $f_k = Q^{\pi_k}$, then

$$m_k \ge 0 \quad \Longleftrightarrow \quad \mathbb{E}_{s\sim d^{\pi_{\text{cp}}}}[A^{\pi_k}(s,a)] \ge 0 \quad \Longleftrightarrow \quad J(\pi_{\text{cp}}) \ge J(\pi_k),$$

where the second equivalence is due to the performance-difference lemma (PDL). This means $m_k \ge 0$ corresponds to $J(\pi_k) \le J(\pi_{\text{cp}})$, whereas $m_k < 0$ indicates that the current policy is already outperforming the comparator (at least from the critic's perspective). Therefore the update deliberately moves towards an "improving" direction which depends whether the current policy is better than the comparator π_{cp}. This means that DRPU remains aligned with the BC descent direction in typical settings when the learned policy is unlikely to exceed the performance of π_{cp}.

Finally, this sign issue does not conflict with our theoretical guarantees. The reason is that the guarantees in the paper are stated in terms of comparator regret, not monotonic decrease of the KL objective. Recall that in Section 3 we introduce the actor-side quantity as

$$\frac{\text{Reg}_K}{K} = \frac{1}{K} \sum_{k=1}^{K} \mathbb{E}_{s\sim d^{\pi_{\text{cp}}}}[f_k(s, \pi_{\text{cp}}) - f_k(s, \pi_k)] = \frac{1}{K} \sum_{k=1}^{K} m_k,$$

which is precisely the average advantage with respect to the comparator policy. Thus, rounds with $m_k < 0$ are not harmful for the regret analysis; if anything, they help, because they indicate that the current policy already outperforms the comparator. What matters in the analysis is not the iterates remain close to π_{cp}, but rather that the mean-matching error is controlled, which is where the norm constraint $\|v_k\| \le V_{\max}$ plays an important stabilizing role here.

F.2 Numerical Result Setting

We consider a simple discounted MDP with finite state and action spaces. The state space is $\mathcal{S} = \{1, 2, 3\}$ and the action space is $\mathcal{A} = \{a_1, a_2\}$. The discount factor is $\gamma = 0.9$, and the initial state distribution d_0 is uniform over $\mathcal{S}$. The transition dynamics are state-absorbing, so the state does not change over time. As a result, the problem reduces to a contextual bandit with discounted returns, while still admitting a well-defined occupancy measure d^π.

The reward function is deterministic and state-dependent. For action a_1, the reward is given by $r(s, a_1) = (1, 4, 4)$ for states $s = 1, 2, 3$, respectively. For action a_2, the reward is constant across

states, given by $r(s, a_2) = (2, 2, 2)$. This construction ensures that the advantage function varies across states and actions, while remaining smooth and bounded. In particular, the policy that always selects a_1 is not optimal, since action a_2 yields higher reward in state $s = 1$.

We consider a one-dimensional softmax policy class $\Pi_\theta = \{\pi_\theta : \theta \in \mathbb{R}\}$ parameterized by

$$\pi_\theta(a_1 \mid s) = \frac{\exp(\theta c_s)}{\exp(\theta c_s) + \exp(-\theta c_s)}, \qquad \pi_\theta(a_2 \mid s) = 1 - \pi_\theta(a_1 \mid s),$$

where the state-dependent coefficients are $c = (1, 2, 3)$. The corresponding score function $\nabla \log \pi_\theta(a \mid s)$ is bounded and one-dimensional, which makes model misspecification effects transparent.

The comparator policy π_{cp} is chosen as the deterministic policy that always selects action a_1 in every state. Importantly, π_{cp} is *not* an optimal policy for this MDP. In the experiments, π_{cp} is approximated within the policy class by a large parameter value $\theta_{\text{cp}} = 100$. The data distribution is set to $d^D = d^{\pi_{\text{cp}}}$, corresponding to the no-shift setting as in imitation learning with expert-generated data.

All methods are initialized at $\theta_0 = 0$ and run for a fixed number of iterations. Performance is evaluated by tracking the policy value and the CFA error err_k under $d^{\pi_{\text{cp}}}$ at each iteration. Since $d^D = d^{\pi_{\text{cp}}}$, there is no distribution shift in this experiment, and any observed performance gap reflects model misspecification alone.

G Technical Lemmas

Lemma 20 (Gibbs Variational Principle)**.** *For any measurable function $\phi : \mathcal{A} \to \mathbb{R}$ and any distribution $u \in \Delta_\nu(\mathcal{A})$ that is absolutely continuous w.r.t. the base measure ν, we have*

$$-\frac{1}{\eta} \log \int_{\mathcal{A}} \exp(-\eta\phi(a))\nu(\mathrm{d}a) = \inf_{u \in \Delta_\nu(\mathcal{A})} \left\{ \mathbb{E}_{a\sim u}[\phi(a)] + \frac{1}{\eta} D_{\text{KL}}(u\|\nu) \right\}.$$

Moreover, the infimum is attained at the Gibbs (softmax) distribution

$$u^\star(a) = \frac{\exp(-\eta\phi(a))}{\int_{\mathcal{A}} \exp(-\eta\phi(a'))\nu(da')}.$$

Lemma 21 (Bellman Error Telescoping(Jiang & Xie, 2025, Lemma 2))**.** *For any $\pi : \mathcal{S} \to \Delta(\mathcal{A})$, and any $f \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$, the performance gap using f as an estimate of Q^π is given by*

$$J_f(\pi) - J(\pi) = \frac{1}{1-\gamma} \mathbb{E}_{d^\pi}[f - \mathcal{T}^\pi f],$$

where $J_f(\pi) = \mathbb{E}_{s\sim d_0}[f(s, \pi)]$, and d^π is the discounted state-action occupancy of π.

Lemma 22 (Generalized Performance-Difference Lemma (Jiang & Xie, 2025, Lemma 6))**.** *For any $f \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$, and policies $\pi, \pi' : \mathcal{S} \to \Delta(\mathcal{A})$, the difference between there expected return is given by*

$$J(\pi') - J(\pi) = \frac{1}{1-\gamma} \left(\mathbb{E}_{s\sim d^{\pi'}} \left[f(s, \pi') - f(s, \pi) \right] + \mathbb{E}_{d^{\pi'}}[\mathcal{T}^\pi f - f] + \mathbb{E}_{d^\pi}[f - \mathcal{T}^\pi f] \right).$$

Lemma 23 (Matrix Chernoff Bound (Tropp, 2012, Corollary 5.2))**.** *Let $X_1, \ldots, X_n$ be independent random vectors in $\mathbb{R}^{\mathsf{d}}$ with $\mathbb{E}[X_i X_i^\top] = I$ and $\|X_i\|_2 \le b$ almost surely. For $t \in (0, 1)$,*

$$\Pr\left\{ \lambda_{\min}\left(\frac{1}{n} \sum_{i=1}^n X_i X_i^\top \right) < 1 - t \right\} \le \mathsf{d} \exp\left(-\frac{nt^2}{2b^2} \right).$$

Lemma 24 (Freedman Inequality for Vector-Valued Martingales (Hsu et al., 2012, Lemma 15))**.** *Let $X_1, \ldots, X_n$ be a martingale difference vector sequence in a Euclidean space (i.e., $\mathbb{E}[X_i \mid X_{1:i-1}] = 0$ for all $i = 1, \ldots, n$) such that for all i,*

$$\sum_{i=1}^n \mathbb{E}\left[\|X_i\|^2 \mid X_{1:i-1} \right] \le v, \qquad \|X_i\| \le b,$$

almost surely. Then for any $\delta \in (0, 1)$, with probability at least $1 - \delta$,

$$\left\| \sum_{i=1}^n X_i \right\| \le \sqrt{v} \left(1 + \sqrt{8 \log \frac{1}{\delta}} \right) + \frac{4}{3} b \log \frac{1}{\delta}.$$

Lemma 25 (Uniform Empirical-Second-Moment Bernstein Bound)**.** *Let $\mathcal{H}_0$ be a deterministic finite class of m measurable functions into $[0,1]$, evaluated on N i.i.d. observations. Put $\Gamma = \log(4m/\delta)$ and $a = \Gamma/N$. With probability at least $1-\delta$, simultaneously for all $h \in \mathcal{H}_0$,*

$$|(\mathbb{E} - \widehat{\mathbb{E}})h| \le \sqrt{2(\mathbb{E}h^2)a} + 2a/3, \qquad |(\mathbb{E} - \widehat{\mathbb{E}})h| \le 2\sqrt{(\widehat{\mathbb{E}}h^2)a} + 4a.$$

Proof. Two-sided scalar Bernstein applied to h and h^2, followed by a union bound over m functions, gives the first bound and $q := \mathbb{E}h^2 \le \widehat{\mathbb{E}}h^2 + \sqrt{2qa} + 2a/3$, since $\mathrm{Var}(h^2) \le \mathbb{E}h^4 \le q$. Using $\sqrt{2qa} \le q/2 + a$ gives $q \le 2\widehat{\mathbb{E}}h^2 + 10a/3$. Substitution into the first bound proves the second, because $\sqrt{20/3} + 2/3 < 4$. There are at most $4m$ one-sided failure events, each of probability at most $e^{-\Gamma}$. □