More Benefits of Being Distributional: Second-Order Bounds for Reinforcement Learning

Kaiwen Wang [1] **Owen Oertell** [1] **Alekh Agarwal** [2] **Nathan Kallus** [1] **Wen Sun** [1]

Abstract

In this paper, we prove that Distributional Reinforcement Learning (DistRL), which learns the return distribution, can obtain second-order bounds in both online and offline RL in general settings with function approximation. Second-order bounds are instance-dependent bounds that scale with the variance of return, which we prove are tighter than the previously known small-loss bounds of distributional RL. To the best of our knowledge, our results are the first second-order bounds for low-rank MDPs and for offline RL. When specializing to contextual bandits (one-step RL problem), we show that a distributional learning based optimism algorithm achieves a second-order worst-case regret bound, and a second-order gap dependent bound, simultaneously. We also empirically demonstrate the benefit of DistRL in contextual bandits on real-world datasets. We highlight that our analysis with DistRL is relatively simple, follows the general framework of optimism in the face of uncertainty and does not require weighted regression. Our results suggest that DistRL is a promising framework for obtaining second-order bounds in general RL settings, thus further reinforcing the benefits of DistRL.

1. Introduction

The aim of reinforcement learning (RL) is to learn a policy that minimizes the expected cumulative cost along its trajectory. Typically, *squared loss* is used in standard RL algorithms (Mnih et al., 2015; Haarnoja et al., 2018) for learning the value function, the expected cost-to-go from a given state. As an alternative to squared loss, Bellemare et al. (2017) proposed to learn the *whole conditional distribution* of cost-to-go with distributional loss functions such as the negative log-likelihood or the pinball loss (Dabney et al., 2018a). This paradigm is aptly called Distributional RL (DistRL) and has since been empirically validated in numerous real-world tasks (Bellemare et al., 2020; Bodnar et al., 2020; Fawzi et al., 2022; Dabney et al., 2018b), as well as in benchmarks for both online (Yang et al., 2019) and offline RL (Ma et al., 2021). However, there is a lack of understanding for why DistRL often attains stronger performance and sample efficiency (Lyle et al., 2019).

This raises a natural theoretical question: when and why is DistRL better than standard RL? Wang et al. (2023b) recently proved that DistRL based on maximum likelihood estimation (MLE) results in *small-loss bounds*, which are instance-dependent bounds that scale with the minimum possible expected cumulative cost $V^\star$ for the task at hand. If the optimal policy makes few blunders on average, *i.e.*, $V^\star \approx 0$, then small-loss bounds converge at the fast $\mathcal{O}(1/N)$ rate, while standard RL bounds converge at a $\mathcal{O}(1/\sqrt{N})$ rate which is worst-case in nature.

In this paper, we refine the analyses of Wang et al. (2023b) and prove that DistRL actually attains tighter *second-order bounds* in both online and offline settings. Instead of scaling with $V^\star$ as in small-loss bounds, our second-order bounds scale with the variance of the policy's cumulative cost. In offline RL, it is the optimal policy's variance, whilst in online RL, it is the variance of policies played by the algorithm. In both cases, our second-order result is *strictly tighter* than the previously known small-loss bounds (a.k.a. first-order bounds), *i.e.*, second-order implies first-order bounds. In particular, our second-order bounds yield fast $\mathcal{O}(1/N)$ rates in near-deterministic tasks where $V^\star$ may still be far from zero. Our theory applies at the same generality as Wang et al. (2023b). Moreover, in contextual bandits (one-step RL), we prove a novel first and second-order gap-dependent bound that incorporates $V^\star$ and variance into the gap definition. Finally, in contextual bandits, we empirically show that our distributionally optimistic algorithm is efficiently implementable with neural networks via width computation (Feng et al., 2021) and outperforms the same optimistic algorithm with squared loss (Foster et al., 2018).

---

[1]Cornell University [2]Google Research. Correspondence to: Kaiwen Wang <https://kaiwenw.github.io>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

Our contributions are summarized as follows:

1. For online RL, we show that DistRL enjoys second-order bounds in MDPs with low ℓ_1-distributional eluder dimension (Wang et al., 2023b). These are the first second-order bounds in MDPs with function approximation, *e.g.*, low-rank MDPs (Section 5).

2. For offline RL, we show that DistRL enjoys second-order bounds with single-policy coverage, the first of such bounds to our knowledge (Section 6).

3. For contextual bandits, our online algorithm further achieves a novel first/second-order gap-dependent bound (Section 4.2). Finally, we empirically evaluate our distributional contextual bandit algorithm and show it outperforms the squared loss baseline (Section 7).[1]

2. Related Works

Theory of DistRL. Rowland et al. (2018; 2023a) showed that DistRL algorithms such as C51 and QR-DQN converges asymptotically with a tabular representation. This unfortunately does not imply finite-sample statistical improvements over standard RL, which is our focus. Recently, Rowland et al. (2023b) showed that quantile temporal-difference (QTD) learning may have smaller bounded variance in each update step than temporal-difference (TD) learning, which may have unbounded variance. While this finding may explain improved training stability, it does not affirmatively imply that QTD obtains better finite-sample regret, which is our focus. For off-policy evaluation (OPE), Wu et al. (2023) showed that fitted likelihood estimation can learn the true return distribution up to errors in total variation and Wasserstein distance. We focus on online and offline RL rather than OPE.

Small-loss Bounds from DistRL. The closest work to ours is Wang et al. (2023b) which showed that MLE-based DistRL can achieve small-loss bounds in online RL and offline RL under distributional Bellman completeness, building on the earlier contextual bandit results of Foster & Krishnamurthy (2021). While Wang et al. (2023b) gave the first small-loss bounds in low-rank MDPs and in offline RL, we prove that their DistRL algorithms can actually achieve tighter, second-order bounds *under identical assumptions*. Our bounds are strictly more general than small-loss (a.k.a. first-order) bounds as shown by the following theorem.

Theorem 2.1 (Informal)**.** *In online and offline RL, a second-order bound implies a first-order bound (with a worse universal constant). This is formalized in Theorem D.2.*

[1]Code is available at `https://github.com/Cornell-RL/DistUCB`

Other second-order bounds. Variance-dependent (a.k.a. second-order bounds) are known in tabular MDPs (Zanette & Brunskill, 2019; Zhou et al., 2023; Zhang et al., 2023), linear mixture MDPs (Zhao et al., 2023), and linear contextual bandits (Ito et al., 2020; Olkhovskaya et al., 2023). These prior works mainly use variance weighted regression, and their analysis does not easily extend beyond linear function approximation. Surprisingly, we show that by simply learning the return distribution with MLE, one can obtain general variance-dependent bounds, by leveraging the tool of triangular discrimination that was first leveraged in Foster & Krishnamurthy (2021). In other words, DistRL is an attractive alternative to variance weighted regression for obtaining sharp second-order bounds in RL.

3. Preliminaries

Contextual Bandits (CB). We first consider CBs with context space $\mathcal{X}$, finite action space $\mathcal{A}$ of size A and normalized conditional costs $C : \mathcal{X} \times \mathcal{A} \to \Delta([0,1])$, where $\Delta([0,1])$ is the set of all distributions on $[0,1]$ that are absolutely continuous with respect to some dominating measure λ, *e.g.*, Lebesgue for continuous or counting for discrete. We identify such a distribution via its density with respect to λ, hence we write $(C(x,a))(y)$ or $C(y \mid x,a)$ for the density of $C(x,a)$ at y. The CB proceeds over K episodes as follows: at episode $k \in [K] = \{1,\dots,K\}$, the learner observes a context $x_k \in \mathcal{X}$, takes an action $a_k \sim \mathcal{A}$, and receives a cost $c_k \sim C(x_k, a_k)$. We do not require that contexts are sampled from a fixed distribution; they may be arbitrarily chosen by an adaptive adversary. The goal is to minimize the regret, defined as

$$\mathrm{Reg}_{\mathsf{CB}}(K) := \sum_{k=1}^{K} \bar{C}(x_k, a_k) - \min_{a \in \mathcal{A}} \bar{C}(x_k, a),$$

where the bar denotes the mean of the distribution, *i.e.*, $\bar{f} = \int y f(y) \mathrm{d}\lambda(y)$ for any $f \in \Delta([0,1])$. We'll also use $\mathrm{Var}(f) = \int (y - \bar{f})^2 f(y) \mathrm{d}\lambda(y)$ to denote the variance.

Reinforcement Learning (RL). We now consider a Markov Decision Process (MDP) with observation space $\mathcal{X}$, finite action space $\mathcal{A}$ of size A, horizon H, transition kernels $P_h : \mathcal{X} \times \mathcal{A} \to \Delta(\mathcal{X})$, and normalized cost distributions $C_h : \mathcal{X} \times \mathcal{A} \to \Delta([0,1])$ at each step $h \in [H]$. Given a policy $\pi : \mathcal{X} \to \Delta(\mathcal{A})$ and an initial state $x_1 \sim \mathcal{X}$, the "roll in" process occurs as follows: for each step $h = 1, 2, \dots, H$, the policy π samples an action a_h based on the current state x_h, incurs a cost c_h from the cost distribution, and transitions to the next state x_{h+1}. The return is the cumulative cost from this random process $Z^\pi(x_1) := \sum_{h=1}^{H} c_h$. The value is the expected return $V^\pi(x_1) := \mathbb{E}[Z^\pi(x_1)]$. We use subscript h to denote cost-to-go from a particular step: $Z_h^\pi(x_h) = \sum_{t=h}^{H} c_t$ and $V_h^\pi(x_h) = \mathbb{E}[Z_h^\pi(x_h)]$. We use $Z^\star, V^\star$ to denote these quantities for the optimal, min-cost policy $\pi^\star$. We use $Z_h^\pi(x_h, a_h)$ to denote the random cost-to-go conditioned on rolling in π from x_h, a_h, and so $Q_h^\pi(x_h, a_h) := \mathbb{E}[Z_h^\pi(x_h, a_h)]$. Without loss of generality, we assume cumulative costs $\sum_{h=1}^H c_h$ are normalized in $[0, 1]$ almost surely, to avoid artificial scaling in H (Jiang & Agarwal, 2018).

The *Online RL* problem proceeds over K episodes: at episode $k \in [K]$, the learner executes a policy $\pi^k : \mathcal{X} \to \Delta(\mathcal{A})$ from an initial state $x_{1,k}$. We do not require that $x_{1,k}$ are sampled from a fixed distribution; they may be chosen by an adaptive adversary. The goal is to minimize regret,

$$\mathrm{Reg}_{\mathsf{RL}}(K) := \sum_{k=1}^K V^{\pi^k}(x_{1,k}) - V^\star(x_{1,k}).$$

In *Offline RL*, the learner is directly given *i.i.d.* samples of transitions drawn from unknown distributions $\nu_1, \ldots, \nu_H$, and the goal is to learn a policy with a lower cost than any other policy whose behavior is covered by the dataset, similar to prior best-effort guarantees in offline RL (Liu et al., 2020; Xie et al., 2021). Concretely, the learner receives a dataset $\mathcal{D} = (\mathcal{D}_1, \mathcal{D}_2, \ldots, \mathcal{D}_H)$, where each $\mathcal{D}_h$ contains N *i.i.d.* samples $(x_{h,i}, a_{h,i}, c_{h,i}, x'_{h,i})$ such that $(x_{h,i}, a_{h,i}) \sim \nu_h, c_{h,i} \sim C_h(x_{h,i}, a_{h,i}), x'_{h,i} \sim P_h(x_{h,i}, a_{h,i})$. Unlike the online setting where initial states can be adversarial, we assume for offline RL that initial states are sampled from a fixed and known distribution d_1.

Distributional RL. The Bellman operator acts on a function $f : \mathcal{X} \times \mathcal{A} \to [0, 1]$ as follows: $\mathcal{T}_h^\pi f(x, a) = \bar{C}_h(x, a) + \mathbb{E}_{x' \sim P_h(x,a), a' \sim \pi(x')}[f(x', a')]$. Analogously, the distributional Bellman operator (Bellemare et al., 2017) acts on a conditional distribution $d : \mathcal{X} \times \mathcal{A} \to \Delta([0, 1])$ as follows: $\mathcal{T}_h^{\pi,D} d(x, a) \overset{D}{=} C_h(x, a) * d(x', a')$, where $x' \sim P_h(x, a), a' \sim \pi(x')$ and $*$ denotes convolution. Another sampling view of the distributional Bellman operator is that $z \sim \mathcal{T}_h^{\pi,D} d(x, a)$ is equivalent to: $c \sim C_h(x, a), x' \sim P_h(x, a), a' \sim \pi(x'), y \sim d(x', a')$ and $z := c + y$. Also recall the optimality operator $\mathcal{T}_h^\star$ and its distributional variant $\mathcal{T}_h^{\star,D}$ are defined as follows: $\mathcal{T}_h^\star f(x, a) = \bar{C}_h(x, a) + \mathbb{E}_{x' \sim P_h(x,a)}[\min_{a \in \mathcal{A}} f(x', a')]$ and $\mathcal{T}_h^{\star,D} d(x, a) \overset{D}{=} C_h(x, a) + d(x', a')$ where $x' \sim P_h(x, a), a' = \arg\min_a \bar{d}(x', a)$.

Triangular Discrimination. For any distributions $f, g \in L^2(\lambda)$, their triangular discrimination (Topsoe, 2000) is defined as $D_\triangle(f \parallel g) := \int \frac{(f(y)-g(y))^2}{f(y)+g(y)} \mathrm{d}\lambda(y)$, which is equivalent to the squared Hellinger distance up to universal constants. Please see Table 2 for an index of notations.

Algorithm 1 DISTUCB (O-DISCO at $H = 1$)

Input: no. episodes K, distribution class $\mathcal{F}$
Init $\mathcal{D}_0 \leftarrow \emptyset$ and $\mathcal{F}_0 \leftarrow \mathcal{F}$.
for episode $k = 1, 2, \ldots, K$ **do**
  Observe context x_k.
  Play $a_k = \arg\min_{a \in \mathcal{A}} \min_{f \in \mathcal{F}_{k-1}} \bar{f}(x_k, a)$.
  Observe cost $c_k \sim C(x_k, a_k)$.
  $\mathcal{D}_k \leftarrow \mathcal{D}_{k-1} \cup \{(x_k, a_k, c_k)\}$, $\mathcal{F}_k \leftarrow \mathsf{CS}_{\mathsf{CB}}(\mathcal{D}_k)$.
end for

4. Warmup: Second-Order Bounds for CBs

As a warmup, we consider contextual bandits and prove that distributional UCB (DISTUCB) attains second-order regret. The distributional confidence set is the main construct that is optimized over to ensure optimism. To construct it, we need a dataset of state, action, costs, $D = \{x_i, a_i, c_i\}_{i \in [N]}$, a threshold β to be specified later, as well as a function class $\mathcal{F} \subset \mathcal{X} \times \mathcal{A} \to \Delta([0, 1])$ containing the true conditional cost distribution $C(\cdot \mid x, a)$. Then, the confidence set is

$$\mathsf{CS}_{\mathsf{CB}}(D) = \left\{ f \in \mathcal{F} : \mathcal{L}_{\mathsf{CB}}(f, D) \geq \max_{g \in \mathcal{F}} \mathcal{L}_{\mathsf{CB}}(g, D) - \beta \right\},$$

where $\mathcal{L}_{\mathsf{CB}}(f, D) := \sum_{i=1}^N \log f(c_i \mid x_i, a_i)$ is the log-likelihood of f on D. In words, $\mathsf{CS}_{\mathsf{CB}}(\mathcal{F}, D)$ contains all functions $f \in \mathcal{F}$ that are β-near-optimal according to the log-likelihood. Then, DISTUCB simply selects the action with the minimum lower confidence bound (LCB) induced by the current confidence set.

Theorem 4.1. *Suppose $C \in \mathcal{F}$ (realizability). For any $\delta \in (0, 1)$, w.p. at least $1 - \delta$, running* DISTUCB *with $\beta = \log(K|\mathcal{F}|/\delta)$ enjoys the regret bound,*

$$\mathrm{Regret}_{\mathsf{CB}}(K) \leq \widetilde{\mathcal{O}}\left(\sqrt{d_{\mathsf{CB}} \beta \cdot \sum_{k=1}^K \mathrm{Var}(C(x_k, a_k)) + d_{\mathsf{CB}} \beta} \right),$$

where d_{CB} is the ℓ_1-eluder dimension (Liu et al., 2022) of $\{(x, a) \mapsto D_\triangle(f(x, a) \parallel C(x, a)) : f \in \mathcal{F}\}$ at threshold K^{-1}. This is a special case of the distributional eluder dimension (Definition 5.2) where $\mathfrak{D} = \{\delta_z : z \in \mathcal{X} \times \mathcal{A}\}$.

The dominant term scales with $\sqrt{\sum_{k=1}^K \mathrm{Var}(C(x_k, a_k))}$ which is sharper than the $\sqrt{K}$ bound of RegCB (Foster et al., 2018), the squared loss variant of DISTUCB. For example, in deterministic settings, our variance-dependent regret scales as $\widetilde{\mathcal{O}}(d_{\mathsf{CB}})$, which is tight in K up to log factors. Nonetheless, confidence-set based strategies like DISTUCB and RegCB are not minimax-optimal as the eluder dimension may scale linearly in $\mathcal{F}$ (Foster et al., 2018, Proposition 1). It would be interesting to derive second-order regret with inverse-gap weighting (Foster & Rakhlin, 2020).

Practical considerations. We note that DistUCB is amenable to practical implementation since conditional on x_k and a, the LCB can be computed efficiently via binary search (Foster & Rakhlin, 2020) or disagreement computation (Feng et al., 2021). We include implementation pseudocode and empirical results in Section 7 and the Appendix.

4.1. Proof of Theorem 4.1

Our first step is to bound the difference of means by variances multiplied by the triangular discrimination.

Lemma 4.2. *For $f, g \in L^2(\lambda)$ s.t. $D_\triangle(f \parallel g) \leq \frac{1}{2}$,*

$$\left|\bar{f} - \bar{g}\right| \leq 2\sqrt{(\mathrm{Var}(f) + \mathrm{Var}(g))D_\triangle(f \parallel g)}. \quad (1)$$

This lemma tightens Eq.(Δ_1) of Wang et al. (2023b) so that variances of f and g appear in the RHS instead of the means. Note that Eq.(Δ_1) of Wang et al. (2023b) holds unconditionally, while our lemma requires $D_\triangle(f \parallel g) \leq \frac{1}{2}$ which is absorbed in the lower order term of the next lemma. This lower order term is a key reason we need the bounded eluder dimension assumption.

Lemma 4.3. *For any $f, g \in L^2(\lambda)$, we have*

$$\left|\bar{f} - \bar{g}\right| \leq 4\sqrt{\mathrm{Var}(f)D_\triangle(f \parallel g)} + 5D_\triangle(f \parallel g). \quad (2)$$

We now bound the regret in a standard way with optimism, *i.e.*, w.h.p. $\bar{f}_k(x_k, a_k) \leq \min_a \bar{C}(x_k, a)$, which is ensured by optimizing the confidence set. Let $\delta_k(x, a) := D_\triangle(f_k(x, a) \parallel C(x, a))$. Then,

$$\begin{aligned}
&\sum_{k=1}^{K} \bar{C}(x_k, a_k) - \min_a \bar{C}(x_k, a) \\
\leq &\sum_{k=1}^{K} \bar{C}(x_k, a_k) - \bar{f}_k(x_k, a_k) && \text{(optimism)} \\
\leq &\sum_{k=1}^{K} 4\sqrt{\mathrm{Var}(C(x_k, a_k))\delta_k(x_k, a_k)} + 5\delta_k(x_k, a_k) && \text{(Eq. (2))} \\
\leq &4\sqrt{\sum_{k=1}^{K} \mathrm{Var}(C(x_k, a_k))\Delta} + 5\Delta, && \text{(Cauchy-Schwarz)}
\end{aligned}$$

where $\Delta = \sum_{k=1}^{K} \delta_k(x_k, a_k)$. Finally, using MLE generalization bound and the fact that $f_k \in \mathcal{F}_{k-1}$, with probability at least $1 - \delta$, we have for all $k \in [K]$: $\sum_{i=1}^{k-1} \delta_k(x_i, a_i) \leq \log(|\mathcal{F}|K/\delta)$ (Wang et al., 2023b, Lemma E.3). Thus, applying pigeon-hole argument of eluder dimension gives $\Delta \leq 4d_{\mathsf{CB}}(1/K)\log(|\mathcal{F}|K/\delta)\log(K)$ (Liu et al., 2022, Proposition 21). This concludes the proof.

4.2. First and Second-Order Gap-Dependent Bounds

While it is known that UCB attains gap-dependent bounds, here we prove *first and second-order gap-dependent bounds* which are novel to the best of our knowledge. Recall that the gap at context x and action a is defined as $\mathrm{Gap}(x, a) := \bar{C}(x, a) - \min_{a^\star \in \mathcal{A}} \bar{C}(x, a^\star)$. We define our novel first and second-order min-gaps as follows:

$$\mathrm{Gap}_{C^\star} = \min_{x \in \mathcal{X}} \min_{\substack{a \in \mathcal{A}: \mathrm{Gap}(x,a) > 0 \\ \wedge \min_{a^\star} \bar{C}(x, a^\star) > 0}} \frac{\mathrm{Gap}(x, a)}{\min_{a^\star} \bar{C}(x, a^\star)},$$

$$\mathrm{Gap}_{\mathsf{Var}} = \min_{x \in \mathcal{X}} \min_{\substack{a \in \mathcal{A}: \mathrm{Gap}(x,a) > 0 \\ \wedge \mathrm{Var}(C(x,a)) > 0}} \frac{\mathrm{Gap}(x, a)}{\sqrt{\mathrm{Var}(C(x, a))}}.$$

The inner min is taken to be ∞ if the condition is empty.

Theorem 4.4. *Assume the premise of Theorem 4.1. If $\max(\mathrm{Gap}_{\mathsf{Var}}, \mathrm{Gap}_{C^\star}) \geq \frac{1}{\sqrt{K}}$, then*

$$\mathrm{Regret}_{\mathsf{CB}}(K) \leq \widetilde{\mathcal{O}}\big(d_{\mathsf{CB}}\beta + d_{\mathsf{CB}}\beta \min\{\mathrm{Gap}_{\mathsf{Var}}^{-1}, \mathrm{Gap}_{C^\star}^{-1}\}\big).$$

As usual, we have a Gap^{-1}-type bound that implies $\mathcal{O}(d_{\mathsf{CB}} \log K)$ regret when the gap is large. Our key innovation lies in the definition of $\mathrm{Gap}_{C^\star}$ and $\mathrm{Gap}_{\mathsf{Var}}$, which are inversely weighted by the optimal mean cost or variance of each context. Our weighted min-gaps are always larger than the standard min-gap (since $\bar{C}(x, a), \mathrm{Var}(C(x, a)) \leq 1$) but they can be much larger in small-loss or near-deterministic regimes. We note that DistUCB's regret is simultaneously bounded by *both* Theorem 4.4 and Theorem 4.1 under the same hyperparameters.

5. Second-Order Bounds for Online DistRL

In this section, we show that the optimistic DistRL algorithm of Wang et al. (2023b) actually enjoys second-order regret and PAC guarantees, which are strictly tighter than the previously known first-order bounds. We first recall the MLE-confidence set for DistRL which generalizes $\mathrm{CS}_{\mathsf{CB}}$ from the warmup. Let $\mathcal{F}$ be a set of conditional distributions, *i.e.*, $(f_1, \ldots, f_H) \in \mathcal{F}$ where $f_h : \mathcal{X} \times \mathcal{A} \to \Delta([0, 1])$, which are candidate functions to fit $Z^\star$ or Z^π (depending on the type of Bellman operator used) with MLE. Given a dataset of state, action, cost, next state tuples, $D = \{x_{h,i}, a_{h,i}, c_{h,i}, x'_{h,i}\}_{h \in [H], i \in [N]}$, and a distributional Bellman operator $\mathcal{T}^D$, the MLE-confidence set is defined as

$$\mathrm{CS}_{\mathsf{RL}}(D; \mathcal{T}^D) = \Big\{f \in \mathcal{F} : \forall h \in [H], \\ \mathcal{L}_{\mathsf{RL}}(f, D) \geq \max_{g \in \mathcal{F}_h} \mathcal{L}_{\mathsf{RL}}(g, D) - \beta\Big\},$$

where $\mathcal{L}_{\mathsf{RL}}(f, D) := \sum_{i=1}^{N} \log f_h(z^f_{h,i} \mid x_{h,i}, a_{h,i})$ and $z^f_{h,i} \sim \mathcal{T}^D_h f_{h+1}(x_{h,i}, a_{h,i})$. In words, $\mathrm{CS}_{\mathsf{RL}}(D; \mathcal{T}^D)$ contains all functions $f \in \mathcal{F}$ such that *for all* $h \in [H]$, f is β-near-optimal w.r.t. the MLE loss for solving $f_h \approx \mathcal{T}^D_h f_{h+1}$. Since this construction happens in a TD fashion, a standard

Algorithm 2 O-DISCO (Wang et al., 2023b)

Input: no. episodes K, distribution class $\mathcal{F}$, UAE flag.
Init $\mathcal{D}_{h,0} \leftarrow \emptyset$ for all $h \in [H]$ and $\mathcal{F}_0 \leftarrow \mathcal{F}$.
for episode $k = 1, 2, \dots, K$ **do**
  Observe init state $x_{1,k}$.
  Set $f^{(k)} \leftarrow \arg\min_{f \in \mathcal{F}_{k-1}} \min_a \bar{f}_1(x_{1,k}, a)$.
  For each h, set $\pi_h^k(x) = \arg\min_a \bar{f}_h^{(k)}(x, a)$.
  if not UAE **then**
    Run π^k from $x_{1,k}$ and get trajectory $x_{1,k}, a_{1,k}, c_{1,k}, .., x_{H,k}, a_{H,k}, c_{H,k}$. Then, $\forall h$, $\mathcal{D}_{h,k} = \mathcal{D}_{h,k-1} \cup \{(x_{h,k}, a_{h,k}, c_{h,k}, x_{h+1,k})\}$.
  else
    For each $h \in [H]$, roll in π^k from $x_{1,k}$ for h steps and take a random action, *i.e.*, $x_{h,k} \sim d_h^{\pi^k}$, $a_{h,k} \sim \mathrm{Unif}(\mathcal{A})$, $c_{h,k} \sim C_h(x_{h,k}, a_{h,k})$, $x'_{h,k} \sim P_h(x_{h,k}, a_{h,k})$. Then, $\mathcal{D}_{h,k} = \mathcal{D}_{h,k-1} \cup \left\{(x_{h,k}, a_{h,k}, c_{h,k}, x'_{h,k})\right\}$.
  end if
  Update $\mathcal{F}_k \leftarrow \mathsf{CS}_{\mathsf{RL}}((\mathcal{D}_{h,k})_{h\in[H]}; \mathcal{T}^{\star,D})$.
end for
Output: $\bar{\pi} = \mathrm{unif}(\pi^{1:K})$.

condition called distributional Bellman Completeness (BC) is needed to guarantee that MLE succeeds for all $h \in [H]$ (Wu et al., 2023; Wang et al., 2023b).

Assumption 5.1 (Bellman Completeness)**.** For all $\pi, h \in [H]$, $f_{h+1} \in \mathcal{F}_{h+1} \implies \mathcal{T}_h^{\pi,D} f_{h+1} \in \mathcal{F}_h$.

BC is a standard assumption in model-free online and offline RL; without it, TD and fitted-Q can diverge or converge to bad fixed points (Tsitsiklis & Van Roy, 1996; Munos & Szepesvári, 2008; Kolter, 2011). As discussed in (Jin et al., 2021a; Wang et al., 2023b), the BC condition can be relaxed to "generalized completeness", *i.e.*, there exist function classes $\mathcal{G}_h$ such that $f_{h+1} \in \mathcal{F}_{h+1} \implies \mathcal{T}_h^{\pi,D} f_{h+1} \in \mathcal{G}_h$.

Then, the O-DISCO algorithm of Wang et al. (2023b) proceeds by selecting the optimistic $f^{(k)}$ in the confidence set $\mathcal{F}_k$ at each round and playing the greedy policy π^k w.r.t. f, where the "playing" can be done with uniform action exploration (UAE). If UAE=True, then for each h, π^k is rolled in for h timesteps and takes a uniform action before the transition tuple is added to the dataset. Note that this requires H rollouts per round but is necessary to capture general MDPs such as low-rank MDPs (Agarwal et al., 2020).

Finally, we adopt the ℓ_1-distributional eluder dimension ($\dim_{\ell_1\mathsf{DE}}$) defined as follows (Wang et al., 2023b).

Definition 5.2 (ℓ_p-distributional eluder dimension)**.** Let $\mathcal{S}$ be any set, Ψ be a set of functions of type $\mathcal{S} \to \mathbb{R}$, and $\mathfrak{D}$ is a set of distributions over $\mathcal{S}$. For any $\varepsilon_0 \in \mathbb{R}_+$, the ℓ_p-distributional eluder dimension ($\dim_{\ell_p,\mathsf{DE}}(\Psi, \mathfrak{D}, \varepsilon_0)$) is the length L of the longest sequence $d^{(1)}, .., d^{(L)} \subset \mathfrak{D}$ s.t. $\exists \varepsilon \geq \varepsilon_0, \forall t \in [L], \exists f \in \Psi$ where $|\mathbb{E}_{d^{(t)}} f| > \varepsilon$ and also $\sum_{i=1}^{t-1} |\mathbb{E}_{d^{(i)}} f|^p \leq \varepsilon^p$.

We work with the same eluder dimensions for RL as in Wang et al. (2023b) which employs the following:

$$\Psi_h = \{(x,a) \mapsto D_\triangle(f_h(x,a) \parallel \mathcal{T}_h^{\star,D} f_{h+1}(x,a)), f \in \mathcal{F}\},$$
$$\mathfrak{D}_h = \{(x,a) \mapsto d_h^\pi(x,a) : \pi \in \Pi\}.$$

Then, the Q-type RL dimension is

$$d_{\mathsf{RL}}(\varepsilon) := \max_h \dim_{\ell_1 \mathsf{DE}}(\Psi_h, \mathfrak{D}_h, \varepsilon).$$

The V-type dimension $d_{\mathsf{RL,V}}$ is analogous with $\Psi_{\mathsf{V},h} = \{x \mapsto \mathbb{E}_{a \sim \mathrm{Unif}(\mathcal{A})}[D_\triangle(f_h(x,a) \parallel \mathcal{T}_h^{\star,D} f_{h+1}(x,a))] : f \in \mathcal{F}\}$. As with d_{CB} (from the CB warmup), the threshold ε is taken as $1/K$ if none is provided. We are now ready to state our online RL result.

Theorem 5.3 (Second-order bounds for Online RL)**.** *Under Assumption 5.1, for any $\delta \in (0,1)$, w.p. at least $1-\delta$, running* O-DISCO *with $\beta = \log(HK|\mathcal{F}|/\delta)$ has $\mathrm{Reg}_{\mathsf{RL}}(K)$ at most,*

$$\widetilde{\mathcal{O}}\left(H\sqrt{\sum_{k=1}^K \mathrm{Var}(Z^{\pi^k}(x_{1,k})) \cdot d_{\mathsf{RL}}\beta} + H^{2.5} d_{\mathsf{RL}}\beta\right).$$

If UAE=True, *then the learned mixture policy $\bar{\pi}$ enjoys the PAC bound: w.p. at least $1-\delta$, $K(V^{\bar{\pi}} - V^\star)$ is at most,*

$$\widetilde{\mathcal{O}}\left(H\sqrt{A\sum_{k=1}^K \mathrm{Var}(Z^{\pi^k}(x_{1,k})) d_{\mathsf{RL,V}}\beta} + AH^{2.5} d_{\mathsf{RL,V}}\beta\right).$$

Compared to prior worst-case bounds for GOLF (Jin et al., 2021a) and small-loss bounds for O-DISCO (Wang et al., 2023b), our new bound has one key improvement: the leading $\sqrt{K}$ terms are replaced by the square root of the sum of return variances $\sum_k \mathrm{Var}(Z^{\pi^k}(x_{1,k}))$. The function class complexity measure $\log|\mathcal{F}|$ can be generalized to bracketing entropy as in Wang et al. (2023b). As Theorem 2.1 shows, our second-order bounds are more general than the first-order bounds of Wang et al. (2023b). For example, in deterministic MDPs where variance is zero, our second-order bound converges at a fast $\widetilde{\mathcal{O}}(1/K)$ rate which is tight up to $\log K$ factors (Wen & Van Roy, 2017). In contrast, $V^\star$ may be non-zero in which case the first-order bound converges at a slow $\widetilde{\Omega}(1/\sqrt{K})$ rate.

It may be surprising that DistRL actually helps for near-deterministic systems. This is because the agent does not *a priori* know that the system is deterministic but a DistRL agent can quickly learn and adapt to this fact, while standard squared loss agents learn to adapt at a slower rate. We highlight that our second-order bound comes easily from $D_\triangle$ generalization bounds of MLE; we do not need any variance weighted regression which almost all prior works to obtain second-order bounds and is hard to extend beyond linear function approximation.

Compared to variance weighted regression, one drawback of our DistRL approach (and other TD-style DistRL algorithms (Wu et al., 2023)) is the requirement of a stronger, distributional completeness assumption (Assumption 5.1), as well as a higher statistical complexity of $\mathcal{F}$ (it is a class of conditional distributions rather than functions). Nevertheless, the empirical success of DistRL suggest these stronger conditions are likely satisfied in practice and the faster second-order rates may indeed offset the increased function class complexity.

5.1. On low-rank MDPs.

Low-rank MDPs (Agarwal et al., 2020) are the standard model for non-linear representation learning in RL (Uehara et al., 2021; Zhang et al., 2022; Ren et al., 2023; Chang et al., 2022), and are defined as follow.

Definition 5.4 (Low-Rank MDP)**.** An MDP is has rank d if each step's transition has a low-rank decomposition $P(x' \mid x, a) = \phi_h^\star(x,a)^\top \mu_h^\star(x')$ where $\phi_h^\star(x,a), \mu_h^\star(x') \in \mathbb{R}^d$ are unknown features that satisfy $\sup_{x,a} \|\phi_h^\star(x,a)\|_2 \le 1$ and $\| \int g \mathrm{d}\mu_h^\star(s') \| \le \|g\|_\infty \sqrt{d}$ for all $g : \mathcal{X} \to \mathbb{R}$.

Our Theorem 5.3 (with UAE) applies to low-rank MDPs the same way as Wang et al. (2023b, Theorem 5.5). In particular, Wang et al. (2023b) showed three important facts for rank-d MDPs: (i) the V-type eluder is controlled $d_{\mathsf{RL,V}}(\varepsilon) \le \mathcal{O}(d \log(d/\varepsilon))$, (ii) given a realizable Φ class, the linear function class $\mathcal{F}^{\text{lin}} = \prod_h \mathcal{F}_h^{\text{lin}}$ defined as

$$\mathcal{F}_h^{\text{lin}} = \Big\{ f(z \mid x,a) = \phi(x,a)^\top w(z) : \phi \in \Phi\,, \\ w : [0,1] \to \mathbb{R}^d,\ \text{s.t.}, \max_z \|w(z)\|_2 \le \sqrt{d} \Big\}$$

satisfies distributional BC (Assumption 5.1), and (iii) if costs are discrete in a uniform grid of M points, the bracketing entropy of $\mathcal{F}^{\text{lin}}$ is $\widetilde{\mathcal{O}}(dM + \log|\Phi|)$. Combining these facts with Theorem 5.3 implies a second-order PAC bound for low-rank MDPs:

Corollary 5.5 (Second-Order PAC Bound for Low-Rank MDPs)**.** *Suppose the MDP has rank d, assume $\phi^\star \in \Phi$ and costs are discrete in a uniform grid of M points, then, w.h.p.,* O-DISCO *with* UAE, $\mathcal{F} = \mathcal{F}^{lin}$ *and* $\beta = dM + \log(|\Phi|/\delta)$ *outputs a policy $\bar\pi$ that satisfies,*

$$V^{\bar\pi} - V^\star \le \widetilde{\mathcal{O}}\left(H\sqrt{\frac{\overline{\mathrm{Var}_{1:K}} \cdot Ad\beta}{K}} + \frac{AdH^{2.5}\beta}{K} \right),$$

where $\overline{\mathrm{Var}_{1:K}} = \frac{1}{K}\sum_{k=1}^K \mathrm{Var}(Z^{\pi^k}(x_{1,k}))$.

To the best of our knowledge, this is the first variance-dependent bound in RL beyond linear function approximation, which is a significant statistical benefit of DistRL.

5.2. Proof Sketch for Theorem 5.3

The new RL tool we'll employ is the following change-of-measure lemma for variance.

Lemma 5.6 (Change of Variance)**.** *For any $f : \mathcal{X} \times \mathcal{A} \to \Delta([0,1])$, π and x_1, we have*

$$\mathbb{E}_{\pi,x_1}[\mathrm{Var}(f_h(x_h,a_h))] \le 2e\,\mathrm{Var}(Z^\pi(x_1)) + \\ 12H^2\mathbb{E}_{\pi,x_1}\Big[\textstyle\sum_{t \ge h} D_\triangle(f_t(x_t,a_t) \parallel \mathcal{T}_t^{\pi,D} f_{t+1}(x_t,a_t))\Big]. \quad (3)$$

For each episode k, by optimism of $\bar f_1^{(k)}$, performance difference lemma and the fact $\mathcal{T}_h^{\pi^k} \bar f_{h+1}^{(k)}(x_h,a_h) = \overline{\mathcal{T}_h^{\pi^k} f_{h+1}^{(k)}}(x_h,a_h)$, we have

$$V^{\pi^k}(x_{1,k}) - V^\star(x_{1,k}) \le V^{\pi^k}(x_{1,k}) - \min_a \bar f_1(x_{1,k},a) \\ = \sum_{h=1}^H \mathbb{E}_{\pi^k,x_{1,k}}\Big[\overline{\mathcal{T}_h^{\pi^k} f_{h+1}^{(k)}}(x_h,a_h) - \bar f_h^{(k)}(x_h,a_h)\Big].$$

Let $\delta_{h,k}(x,a) := D_\triangle(f_h^{(k)}(x,a) \parallel \mathcal{T}_h^{\star,D} f_{h+1}^{(k)}(x,a))$.

$$\sum_{h=1}^H \mathbb{E}_{\pi^k,x_{1,k}}\Big[\overline{\mathcal{T}_h^{\pi^k} f_{h+1}^{(k)}}(x_h,a_h) - \bar f_h^{(k)}(x_h,a_h)\Big]$$
$$\le \sum_{h=1}^H 4\sqrt{\mathbb{E}_{\pi^k,x_{1,k}}[\mathrm{Var}(f_h^{(k)}(x_h,a_h))] \cdot \mathbb{E}_{\pi^k,x_{1,k}}[\delta_{h,k}(x_h,a_h)]} \\ + 5\mathbb{E}_{\pi^k,x_{1,k}}[\delta_{h,k}(x_h,a_h)] \qquad \text{(Eq. (2))}$$
$$\le \sum_{h=1}^H 4\sqrt{(2e\,\mathrm{Var}(Z^\pi(x_{1,k})) + 12H^2\Delta_k) \cdot \mathbb{E}_{\pi^k,x_{1,k}}[\delta_{h,k}(x_h,a_h)]} \\ + 5\mathbb{E}_{\pi^k,x_{1,k}}[\delta_{h,k}(x_h,a_h)] \qquad \text{(Eq. (3))}$$
$$\le 4\sqrt{(2e\,\mathrm{Var}(Z^\pi(x_{1,k})) + 12H^2\Delta_k) \cdot H\Delta_k} + 5H\Delta_k, \\ \text{(Cauchy-Schwarz)}$$

where $\Delta_k := \sum_{h=1}^H \mathbb{E}_{\pi^k,x_{1,k}}[\delta_{h,k}(x_h,a_h)]$. Finally, we can sum over all episodes and use the fact that $\sum_k \Delta_k \le Hd\log K$ w.p. $1-\delta$, where d is the appropriate distributional eluder dimension depending on UAE. This last step is true due to MLE's generalization bound and standard eluder-type arguments from Wang et al. (2023b).

6. Second-Order Bounds for Offline DistRL

We now turn to offline RL and prove that pessimism in the face of uncertainty with MLE-confidence sets enjoys second-order PAC bounds under single-policy coverage. The algorithm we study is P-DISCO (Wang et al., 2023b), which adapts the pessimism-over-confidence-set approach from BCP (Xie et al., 2021) with the DistRL confidence set. As

Algorithm 3 P-DISCO (Wang et al., 2023b)

1: **Input:** datasets $\mathcal{D}_1, \dots, \mathcal{D}_H$, distribution class $\mathcal{F}$, policy class Π.
2: $\forall \pi \in \Pi$, set $\mathcal{F}_\pi \leftarrow \mathsf{CS}_{\mathsf{RL}}((\mathcal{D}_h)_{h\in[H]}; \mathcal{T}^{\pi,D})$.
3: $\forall \pi \in \Pi$, set $f^\pi \leftarrow \arg\max_{f\in\mathcal{F}_\pi} \mathbb{E}_{x_1\sim d_1}[\bar{f}_1(x_1,\pi)]$.
4: **Output:** $\widehat{\pi} = \arg\min_{\pi\in\Pi} \mathbb{E}_{x_1\sim d_1}[\bar{f}_1^\pi(x_1,\pi)]$.

shown in Algorithm 3, P-DISCO returns the best policy with respect to its pessimistic value estimate, induced by the distributional confidence set constructed with the given data.

Following recent advancements in offline RL (Xie et al., 2021; Uehara & Sun, 2022; Jin et al., 2021b), we prove best-effort guarantees that aim to compete with any covered comparator policy $\widetilde{\pi}$ and that only requires weak single-policy coverage. In particular, we do not suffer the strong all-policy coverage condition used in (Chen & Jiang, 2019). Recall the single-policy concentrability w.r.t. the comparator policy $\widetilde{\pi}$ is defined as $C^{\widetilde{\pi}} := \max_h \|\mathrm{d}d_h^{\widetilde{\pi}}/\mathrm{d}\nu_h\|_\infty$. We now state our main result for offline RL.

Theorem 6.1 (Second-order bounds for Offline RL)**.** *Under Assumption 5.1, for any $\delta \in (0,1)$, w.p. at least $1-\delta$, running* P-DISCO *with $\beta = \log(H|\Pi||\mathcal{F}|/\delta)$ learns a policy $\widehat{\pi}$ that enjoys the following bound: for any comparator $\widetilde{\pi} \in \Pi$ (not necessarily the optimal $\pi^\star$), we have*

$$V^{\widehat{\pi}} - V^{\widetilde{\pi}} \le \mathcal{O}\left(H\sqrt{\frac{\mathrm{Var}(Z^{\widetilde{\pi}})C^{\widetilde{\pi}}\beta}{N}} + \frac{H^{2.5}C^{\widetilde{\pi}}\beta}{N} \right).$$

Here, the leading term scales with the variance of the *comparator policy's* returns $\mathrm{Var}(Z^{\widetilde{\pi}})$. Since the variance is bounded by the first moment, this bound immediately improves the small-loss PAC bound of Wang et al. (2023b). In near-deterministic settings, our second-order bound guarantees a fast $1/N$ rate and is tight up to log factors, which is not necessarily the case for small-loss bounds. In particular, our result shows that DistRL is even more robust to poor coverage than as shown in Wang et al. (2023b); that is, P-DISCO can strongly compete with a comparator policy $\widetilde{\pi}$ if one of the following is true: (i) ν has good coverage over $\widetilde{\pi}$, so the $\sqrt{1/N}$ term has a small constant; or (ii) ν has bad (but finite) coverage and $\widetilde{\pi}$ has small *variance*, in which case we can still obtain a fast $1/N$ rate (with constant scaling with coverage). To the best of our knowledge, this is the first second-order bound for offline RL.

Variance of $Z(\pi^k)$ vs. $Z(\pi^\star)$. In online RL, Theorem 5.3 and Corollary 5.5 has the average variance of the played policies $Z(\pi^k)$, while in offline RL, Theorem 6.1 has the variance of the optimal policy $Z(\pi^\star)$ (if comparing with optimal policy). From a technical perspective, this dichotomy arises from the fact that in offline RL, single-policy concentrability allows us to change measure to $\pi^\star$, while in online RL, we cannot perform the switch and instead rely on eluder-type arguments. The variances of $Z(\pi^k)$ and $Z(\pi^\star)$ are in general incomparable. Nonetheless, both statements are sharper than the small-loss bound as shown by Theorem D.2. Both are also tight in deterministic settings.

Computational Efficiency. Both O-DISCO and P-DISCO optimize over the confidence set to ensure optimism and pessimism, respectively, but this step is known to be computationally hard even in tabular MDPs (Dann et al., 2018). This is also an issue for other version space algorithms: OLIVE (Jiang et al., 2017), GOLF (Jin et al., 2021a), and BCP (Xie et al., 2021). However, the confidence set is needed for the purpose of deep exploration and can be replaced by myopic strategies such as ε-greedy that are computationally cheap (Dann et al., 2022). Finally, in the sequel, we show that in the case of CBs ($H = 1$), O-DISCO can be efficiently implemented with neural nets via disagreement computation (Feng et al., 2021).

7. Contextual Bandit Experiments

We empirically validate our stronger theory in the contextual bandit setting where our algorithm DISTUCB can be efficiently implemented. We demonstrate that learning the cost distribution (as in DISTUCB) consistently improves performance of the baseline algorithm RegCB (Foster et al., 2018) which uses the squared loss instead of log-likelihood. It's worth noting that cost distribution learning has been shown to be effective in inverse-gap weighted (IGW) algorithms (Wang et al., 2023b); however, our focus here is on optimistic algorithms such as DISTUCB and RegCB. We now describe our efficient implementation with neural networks as function approximators via computing width with the log-likelihood loss.

Efficient Implementation by Computing Width. We group incoming contexts into batches $\mathcal{B}_k \subset \mathcal{X}$ to use GPU parallelism for neural networks. Let $\mathcal{D}_{k-1}$ denote the history so far. Then, recall that DISTUCB aims to compute optimistic actions $a_k = \arg\min_a \min_{f\in\mathcal{F}_{k-1}} \bar{f}(x_k,a)$ for each context $x_k \in \mathcal{B}_k$, where $\mathcal{F}_{k-1}$ is the subset of β-optimal functions w.r.t. the log-likelihood on the history $\mathcal{L}_{\mathsf{CB}}(f,\mathcal{D}_{k-1})$, where β is a hyperparameter. We consider inducing optimism by subtracting the width of $\mathcal{F}_{k-1}$, defined as

$$w_k(x,a) = \max_{f,f'\in\mathcal{F}} \left\{ \bar{f}(x,a) - \bar{f}'(x,a) \right\} \text{ s.t. } f,f' \in \mathcal{F}_{k-1}.$$

Then, given the MLE $g_k = \arg\max_{g\in\mathcal{F}} \mathcal{L}_{\mathsf{CB}}(g,\mathcal{D}_{k-1})$ we can set $f_k := (\bar{g}_k - w_k)$ which satisfies optimism, *i.e.*, $f_k(x_k,a) \le \bar{C}(x_k,a)$, for all a. Thus, the goal now is to compute $w_k(x_k,a)$ for each $x_k \in \mathcal{B}_k$ and $a \in \mathcal{A}$.

We modify the width computation strategy of Feng et al. (2021) to deal with the log-likelihood loss. In particular, given the current MLE g_k parameterized by a neural net, we create a copy g' and train g' for a few steps of gradient ascent on the disagreement objective (g_k is fixed):

$$\sum_{a \in \mathcal{A}} \sum_{x_k \in \mathcal{B}_k} \lambda(\bar{g}'(x_k, a) - \bar{g}_k(x_k, a))^2 / |\mathcal{B}_k| \\ - \sum_{(x,a) \in \mathcal{D}_{k-1}} (\bar{g}'(x, a) - \bar{g}_k(x, a))^2 / |\mathcal{D}_{k-1}| \\ - \sum_{a \in \mathcal{A}} \sum_{x_k \in \mathcal{B}_k} \lambda_1(\bar{g}'(x_k, a) - \bar{g}_k(x_k, a)) / |\mathcal{B}_k|$$

where, the last term of the maximization objective is to avoid a zero gradient when $g_k = g'$. Due to memory constraints, we approximate the second term with a subset of the history. Then, we denote $\widehat{w}_k(x, a) = |\bar{g}_k(x, a) - \bar{g}'(x, a)|$ and set the bonus to be the normalized width $\lambda_2 \cdot \frac{\widehat{w}_k(s,a)}{\max_{a \in \mathcal{A}, x \in \mathcal{B}_k} \widehat{w}_k(x,a)}$. $\lambda, \lambda_1, \lambda_2$ are hyperparameters.

We note that an alternative poly-time algorithm is to binary search for a Lagrange multiplier as in RegCB (Foster et al., 2018), which we also tried. However, the binary search approach requires an optimization oracle at every binary search depth, for every action, whereas disagreement computation only needs one optimization oracle per batch of contexts. Binary searching is thus much more computationally costly and we did not observe any improvement in performance to justify the increased computation. Hence, we use disagreement-based width computation for inducing optimism for all DistUCB and RegCB experiments.

CB Tasks. We now compare DistUCB and RegCB on the three real-world CB tasks: King County Housing (Vanschoren et al., 2013), Prudential Life Insurance (Montoya et al., 2015), and CIFAR-100 (Krizhevsky, 2009). The Housing and Prudential tasks are derived from risk prediction tasks, where a fixed max cost is incurred for over-predicting risk and a low cost is incurred for under-predicting risk (Farsang et al., 2022). The CIFAR-100 task is derived from the image classification task, where 0 cost is given for the correct label, 0.5 cost is given for an almost correct label (*i.e.*, correct superclass), and 1 cost is given otherwise (for wrong superclass). All tasks were rolled out for 5000 steps in batches of 32 examples.

Function Approximators. We use neural networks for squared loss regression in RegCB and maximum likelihood estimation in DistUCB. For the King County Housing dataset and the Prudential Life Insurance dataset, we used 2 hidden-layer MLPs, while for CIFAR-100, we used ResNet-18 (He et al., 2016). This is the same setup as in Wang et al. (2023b, Appendix K).

Algorithm:	RegCB	DistUCB (Ours)
King County Housing (Vanschoren et al., 2013)		
All episodes	.708 (.051)	**.683** (.057)
Last 100 ep.	.676 (.038)	**.640** (.037)
Prudential Life Insurance (Montoya et al., 2015)		
All episodes	.287 (.058)	**.248** (.061)
Last 100 ep.	.278 (.055)	**.236** (.054)
CIFAR-100 (Krizhevsky, 2009)		
All episodes	.890 (.053)	**.862** (.058)
Last 100 ep.	.854 (.053)	**.823** (.060)

Table 1. Average cost over all episodes and last 100 episodes (lower is better). We report 'mean (sem)' over 3 seeds.

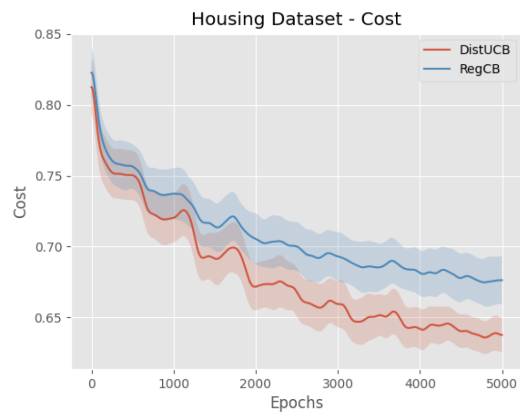

Figure 1. Cost curves for the Housing task (lower is better).

Results. Table 7 shows that cost distribution learning in DistUCB consistently improves the costs and regret compared to the baseline squared loss method RegCB. Also, Fig. 1 shows that DistUCB converges to a smaller cost much faster than RegCB. This reinforces that our stronger theory for MLE-based distribution learning indeed translates to more effective algorithms than standard squared loss regression. We note that in the Housing and Prudential tasks, our costs are actually lower and better than the previously reported numbers by IGW algorithms (Wang et al., 2023b). However, it is worth noting that optimistic algorithms based on width computation is still more computationally costly than IGW algorithms, and a carefully tuned IGW can likely perform just as well in practice.

8. Conclusion

We proved that MLE-based DistRL attains second-order bounds in both online and offline RL, significantly sharpening the previous results of Wang et al. (2023b) and further showing the finite-sample statistical benefits of DistRL. In the CB case, we also proved a novel first and second-order gap-dependent bound and implemented the algorithm, showing it outperforms the previous squared loss method. An interesting direction is to show whether DistRL can obtain even higher-order bounds than second-order.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

Agarwal, A., Kakade, S., Krishnamurthy, A., and Sun, W. Flambe: Structural complexity and representation learning of low rank mdps. *Advances in neural information processing systems*, 33:20095–20107, 2020.

Bellemare, M. G., Dabney, W., and Munos, R. A distributional perspective on reinforcement learning. In *International conference on machine learning*, pp. 449–458. PMLR, 2017.

Bellemare, M. G., Candido, S., Castro, P. S., Gong, J., Machado, M. C., Moitra, S., Ponda, S. S., and Wang, Z. Autonomous navigation of stratospheric balloons using reinforcement learning. *Nature*, 588(7836):77–82, 2020.

Bodnar, C., Li, A., Hausman, K., Pastor, P., and Kalakrishnan, M. Quantile qt-opt for risk-aware vision-based robotic grasping. *Robotics: Science and Systems*, 2020.

Chang, J., Wang, K., Kallus, N., and Sun, W. Learning bellman complete representations for offline policy evaluation. In *International Conference on Machine Learning*, pp. 2938–2971. PMLR, 2022.

Chen, J. and Jiang, N. Information-theoretic considerations in batch reinforcement learning. In *International Conference on Machine Learning*, pp. 1042–1051. PMLR, 2019.

Dabney, W., Ostrovski, G., Silver, D., and Munos, R. Implicit quantile networks for distributional reinforcement learning. In *International conference on machine learning*, pp. 1096–1105. PMLR, 2018a.

Dabney, W., Rowland, M., Bellemare, M., and Munos, R. Distributional reinforcement learning with quantile regression. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018b.

Dann, C., Jiang, N., Krishnamurthy, A., Agarwal, A., Langford, J., and Schapire, R. E. On oracle-efficient pac rl with rich observations. *Advances in neural information processing systems*, 31, 2018.

Dann, C., Mansour, Y., Mohri, M., Sekhari, A., and Sridharan, K. Guarantees for epsilon-greedy reinforcement learning with function approximation. In *International Conference on Machine Learning*, pp. 4666–4689. PMLR, 2022.

Farsang, M., Mineiro, P., and Zhang, W. Conditionally risk-averse contextual bandits. *arXiv preprint arXiv:2210.13573*, 2022.

Fawzi, A., Balog, M., Huang, A., Hubert, T., Romera-Paredes, B., Barekatain, M., Novikov, A., R Ruiz, F. J., Schrittwieser, J., Swirszcz, G., et al. Discovering faster matrix multiplication algorithms with reinforcement learning. *Nature*, 610(7930):47–53, 2022.

Feng, F., Yin, W., Agarwal, A., and Yang, L. Provably correct optimization and exploration with non-linear policies. In *International Conference on Machine Learning*, pp. 3263–3273. PMLR, 2021.

Foster, D. and Rakhlin, A. Beyond ucb: Optimal and efficient contextual bandits with regression oracles. In *International Conference on Machine Learning*, pp. 3199–3210. PMLR, 2020.

Foster, D., Agarwal, A., Dudík, M., Luo, H., and Schapire, R. Practical contextual bandits with regression oracles. In *International Conference on Machine Learning*, pp. 1539–1548. PMLR, 2018.

Foster, D. J. and Krishnamurthy, A. Efficient first-order contextual bandits: Prediction, allocation, and triangular discrimination. *CoRR*, abs/2107.02237, 2021. URL https://arxiv.org/abs/2107.02237.

Haarnoja, T., Zhou, A., Abbeel, P., and Levine, S. 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.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016. doi: 10.1109/CVPR.2016.90.

Hessel, M., Modayil, J., Van Hasselt, H., Schaul, T., Ostrovski, G., Dabney, W., Horgan, D., Piot, B., Azar, M., and Silver, D. Rainbow: Combining improvements in deep reinforcement learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

Ito, S., Hirahara, S., Soma, T., and Yoshida, Y. Tight first- and second-order regret bounds for adversarial linear bandits. *Advances in Neural Information Processing Systems*, 33:2028–2038, 2020.

Jiang, N. and Agarwal, A. Open problem: The dependence of sample complexity lower bounds on planning horizon. In *Conference On Learning Theory*, pp. 3395–3398. PMLR, 2018.

Jiang, N., Krishnamurthy, A., Agarwal, A., Langford, J., and Schapire, R. E. Contextual decision processes with low bellman rank are pac-learnable. In *International Conference on Machine Learning*, pp. 1704–1713. PMLR, 2017.

Jin, C., Yang, Z., Wang, Z., and Jordan, M. I. Provably efficient reinforcement learning with linear function approximation. In *Conference on Learning Theory*, pp. 2137–2143. PMLR, 2020.

Jin, C., Liu, Q., and Miryoosefi, S. Bellman eluder dimension: New rich classes of rl problems, and sample-efficient algorithms. *Advances in neural information processing systems*, 34:13406–13418, 2021a.

Jin, Y., Yang, Z., and Wang, Z. Is pessimism provably efficient for offline rl? In *International Conference on Machine Learning*, pp. 5084–5096. PMLR, 2021b.

Kolter, J. The fixed points of off-policy td. *Advances in Neural Information Processing Systems*, 24, 2011.

Krizhevsky, A. *Learning Multiple Layers of Features from Tiny Images*. 2009.

Liu, Q., Chung, A., Szepesvári, C., and Jin, C. When is partially observable reinforcement learning not scary? In *Conference on Learning Theory*, pp. 5175–5220. PMLR, 2022.

Liu, Y., Swaminathan, A., Agarwal, A., and Brunskill, E. Provably good batch off-policy reinforcement learning without great exploration. *Advances in neural information processing systems*, 33:1264–1274, 2020.

Lyle, C., Bellemare, M. G., and Castro, P. S. A comparative analysis of expected and distributional reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 4504–4511, 2019.

Ma, Y., Jayaraman, D., and Bastani, O. Conservative offline distributional reinforcement learning. *Advances in Neural Information Processing Systems*, 34:19235–19247, 2021.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., et al. Human-level control through deep reinforcement learning. *nature*, 518(7540): 529–533, 2015.

Montoya, A., BigJek14, Bull, denisedunleavy, egrad, FleetwoodHack, Imbayoh, PadraicS, Pru_Admin, tpitman, and Cukierski, W. Prudential life insurance assessment, 2015. URL https://kaggle.com/competitions/prudential-life-insurance-assessment.

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

Olkhovskaya, J., Mayo, J., van Erven, T., Neu, G., and Wei, C.-Y. First- and second-order bounds for adversarial linear contextual bandits. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=NTSbj2otOA.

Polyanskiy, Y. and Wu, Y. *Information Theory: From Coding to Learning*. 2023. https://people.lids.mit.edu/yp/homepage/data/itbook-export.pdf.

Ren, T., Zhang, T., Lee, L., Gonzalez, J. E., Schuurmans, D., and Dai, B. Spectral decomposition representation for reinforcement learning. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=FBMLeaXpZN.

Rowland, M., Bellemare, M., Dabney, W., Munos, R., and Teh, Y. W. An analysis of categorical distributional reinforcement learning. In *International Conference on Artificial Intelligence and Statistics*, pp. 29–37. PMLR, 2018.

Rowland, M., Munos, R., Azar, M. G., Tang, Y., Ostrovski, G., Harutyunyan, A., Tuyls, K., Bellemare, M. G., and Dabney, W. An analysis of quantile temporal-difference learning. *arXiv preprint arXiv:2301.04462*, 2023a.

Rowland, M., Tang, Y., Lyle, C., Munos, R., Bellemare, M. G., and Dabney, W. The statistical benefits of quantile temporal-difference learning for value estimation. *arXiv preprint arXiv:2305.18388*, 2023b.

Topsoe, F. Some inequalities for information divergence and related measures of discrimination. *IEEE Transactions on information theory*, 46(4):1602–1609, 2000.

Tsitsiklis, J. and Van Roy, B. Analysis of temporal-diffference learning with function approximation. *Advances in neural information processing systems*, 9, 1996.

Uehara, M. and Sun, W. Pessimistic model-based offline reinforcement learning under partial coverage. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=tyrJsbKAe6.

Uehara, M., Zhang, X., and Sun, W. Representation learning for online and offline rl in low-rank mdps. In *International Conference on Learning Representations*, 2021.

Vanschoren, J., van Rijn, J. N., Bischl, B., and Torgo, L. Openml: Networked science in machine learning. *SIGKDD Explorations*, 15(2):49–60, 2013. doi: 10.1145/2641190.2641198. URL http://doi.acm.org/10.1145/2641190.2641198.

Wang, K., Kallus, N., and Sun, W. Near-minimax-optimal risk-sensitive reinforcement learning with cvar. *International Conference on Machine Learning*, 2023a.

Wang, K., Zhou, K., Wu, R., Kallus, N., and Sun, W. The benefits of being distributional: Small-loss bounds for reinforcement learning. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023b. URL https://openreview.net/forum?id=S3Y0VvegGm.

Wen, Z. and Van Roy, B. Efficient reinforcement learning in deterministic systems with value function generalization. *Mathematics of Operations Research*, 42(3): 762–782, 2017.

Wu, R., Uehara, M., and Sun, W. Distributional offline policy evaluation with predictive error guarantees. *International Conference of Machine Learning*, 2023.

Xie, T., Cheng, C.-A., Jiang, N., Mineiro, P., and Agarwal, A. Bellman-consistent pessimism for offline reinforcement learning. *Advances in neural information processing systems*, 34:6683–6694, 2021.

Yang, D., Zhao, L., Lin, Z., Qin, T., Bian, J., and Liu, T.-Y. Fully parameterized quantile function for distributional reinforcement learning. *Advances in neural information processing systems*, 32, 2019.

Zanette, A. and Brunskill, E. Tighter problem-dependent regret bounds in reinforcement learning without domain knowledge using value function bounds. In *International Conference on Machine Learning*, pp. 7304–7312. PMLR, 2019.

Zhang, T., Ren, T., Yang, M., Gonzalez, J., Schuurmans, D., and Dai, B. Making linear mdps practical via contrastive representation learning. In *International Conference on Machine Learning*, pp. 26447–26466. PMLR, 2022.

Zhang, Z., Chen, Y., Lee, J. D., and Du, S. S. Settling the sample complexity of online reinforcement learning. *arXiv preprint arXiv:2307.13586*, 2023.

Zhao, H., He, J., Zhou, D., Zhang, T., and Gu, Q. Variance-dependent regret bounds for linear bandits and reinforcement learning: Adaptivity and computational efficiency. In Neu, G. and Rosasco, L. (eds.), *Proceedings of Thirty Sixth Conference on Learning Theory*, volume 195 of *Proceedings of Machine Learning Research*, pp. 4977–5020. PMLR, 12–15 Jul 2023. URL https://proceedings.mlr.press/v195/zhao23a.html.

Zhou, R., Zihan, Z., and Du, S. S. Sharp variance-dependent bounds in reinforcement learning: Best of both worlds in stochastic and deterministic environments. In *International Conference on Machine Learning*, pp. 42878–42914. PMLR, 2023.

Appendices

A. Notations

Table 2. List of Notations

$\mathcal{S}, \mathcal{A}, A$	State and action spaces, and $A = \lvert\mathcal{A}\rvert$.
$\Delta(S)$	The set of distributions supported by set S.
$\bar{d}$	The expectation of any real-valued distribution d, *i.e.*, $\bar{d} = \int y d(y) \mathrm{d}\lambda(y)$.
$\mathrm{Var}(d)$	The variance of any real-valued distribution d, *i.e.*, $\mathrm{Var}(d) = \int (y - \bar{d})^2 d(y) \mathrm{d}\lambda(y)$.
$[N]$	$\{1, 2, \dots, N\}$ for any $N \in \mathbb{N}$.
$Z_h^\pi(x, a)$	Distribution of $\sum_{t=h}^H c_t$ given $x_h = x, a_h = a$ rolling in from π.
$Q_h^\pi(x, a), V_h^\pi(x)$	$Q_h^\pi(x, a) = \bar{Z}_h^\pi(x, a)$ and $V_h^\pi = \mathbb{E}_{a \sim \pi(x)}[Q_h^\pi(x, a)]$.
$Z_h^\star, Q_h^\star, V_h^\star$	$Z_h^\pi, Q_h^\pi, V_h^\pi$ with $\pi = \pi^\star$, the optimal policy.
$\mathcal{T}_h^\pi, \mathcal{T}_h^\star$	The Bellman operators that act on functions.
$\mathcal{T}_h^{\pi,D}, \mathcal{T}_h^{\star,D}$	The distributional Bellman operators that act on conditional distributions.
$V^\pi, Z^\pi, V^\star, Z^\star$	$V^\pi = V_1^\pi(x_1)$, $Z^\pi = Z_1^\pi(x_1)$. $V^\star, Z^\star$ are defined similarly with $\pi^\star$.
$d_h^\pi(x, a)$	The probability of π visiting (x, a) at time h.
$C^{\widetilde{\pi}}$	Coverage coefficient $\max_h \left\lVert \mathrm{d}d_h^{\widetilde{\pi}} / \mathrm{d}\nu_h \right\rVert_\infty$.
$D_\triangle(f \parallel g)$	Triangular discrimination between f, g.
$H(f \parallel g)$	Hellinger distance between f, g.

A.1. Statistical Distances

Let f, g be distributions over $\mathcal{Y}$. Then,

$$D_\triangle(f \parallel g) = \sum_y \frac{(f(y) - g(y))^2}{f(y) + g(y)},$$
$$H^2(f \parallel g) = \frac{1}{2} \sum_y \left(\sqrt{f(y)} - \sqrt{g(y)} \right)^2.$$

Lemma A.1. *For any distributions f, g, we have $2H^2(f \parallel g) \le D_\triangle(f \parallel g) \le 4H^2(f \parallel g)$.*

Proof. Recall that $D_\triangle(f \parallel g) = \int_y \left(\frac{f(y) - g(y)}{\sqrt{f(y) + g(y)}} \right)^2$. Apply $\frac{1}{\sqrt{f(y)} + \sqrt{g(y)}} \le \frac{1}{\sqrt{f(y) + g(y)}} \le \frac{\sqrt{2}}{\sqrt{f(y)} + \sqrt{g(y)}}$. $\square$

B. Proofs for CB Lemmas

Lemma 4.2. *For $f, g \in L^2(\lambda)$ s.t. $D_\triangle(f \parallel g) \le \frac{1}{2}$,*

$$\left|\bar{f} - \bar{g}\right| \le 2\sqrt{(\mathrm{Var}(f) + \mathrm{Var}(g)) D_\triangle(f \parallel g)}. \tag{1}$$

Proof. For any constant c and random variable X, recall that $\mathbb{E}(X - c)^2 = \mathrm{Var}(X) + (\mathbb{E}X - c)^2$. Thus,

$$\begin{aligned}
\bar{f} - \bar{g} &= \sum_z (z - c)(f(z) - g(z)) \\
&\le \sqrt{\sum_z (z - c)^2 (f(z) + g(z))} \cdot \sqrt{\sum_z \frac{(f(z) - g(z))^2}{f(z) + g(z)}} && \text{(Cauchy-Schwartz)} \\
&= \sqrt{\mathrm{Var}(f) + \mathrm{Var}(g) + (\bar{f} - c)^2 + (\bar{g} - c)^2} \cdot \sqrt{D_\triangle(f \parallel g)}.
\end{aligned}$$

To minimize the bound, set $c = \frac{\bar{f} + \bar{g}}{2}$ to get,

$$\begin{aligned}
\left|\bar{f} - \bar{g}\right| &\le \sqrt{\mathrm{Var}(f) + \mathrm{Var}(g) + (\bar{f} - \bar{g})^2/2} \cdot \sqrt{D_\triangle(f \parallel g)} \\
&\le \sqrt{(\mathrm{Var}(f) + \mathrm{Var}(g)) D_\triangle(f \parallel g)} + \left|\bar{f} - \bar{g}\right| \sqrt{D_\triangle(f \parallel g)/2}.
\end{aligned}$$

Rearranging and using the fact that $D_\triangle(f \parallel g) < 2$,

$$\left|\bar{f} - \bar{g}\right| \le \frac{1}{1 - \sqrt{D_\triangle(f \parallel g)/2}} \sqrt{(\mathrm{Var}(f) + \mathrm{Var}(g)) D_\triangle(f \parallel g)}.$$

Finally, use the facts that $\frac{1}{1-\varepsilon} \le 2$ for $\varepsilon \in [0, \frac{1}{2}]$ and $\sqrt{D_\triangle(f \parallel g)/2} \le \frac{1}{2}$ by the premise. $\square$

Lemma B.1. *For any $f, g \in \Delta([0, 1])$, we have*

$$|\mathrm{Var}(f) - \mathrm{Var}(g)| \le 4\sqrt{(\mathrm{Var}(f) + D_\triangle(f \parallel g)) D_\triangle(f \parallel g)} \tag{4}$$

Proof. Recall that $\mathrm{Var}(f) = \frac{1}{2}\mathbb{E}_{z,z' \sim f \otimes f}[(z - z')^2]$. So if f' is the distribution of $\frac{1}{2}(z - z')^2$ where $z, z' \sim f$, then $\mathrm{Var}(f) = \bar{f}'$. Since $(z - z')^2 \in [0, 1]$, we can use Eq.(Δ_2) of (Wang et al., 2023b) to get $\left|\bar{f} - \bar{g}\right| \le \sqrt{(4\bar{f} + D_\triangle(f \parallel g)) D_\triangle(f \parallel g)}$. Thus, $|\mathrm{Var}\, f - \mathrm{Var}\, g| \le \sqrt{(4 \mathrm{Var}_f + D_\triangle(f' \parallel g')) D_\triangle(f' \parallel g')}$.

Now it suffices to bound $D_\triangle(f' \parallel g')$ by $4D_\triangle(f \parallel g)$, which we do by data processing inequality and tensorization of Hellinger. In particular, the tensorization of H^2 is given by $H^2(f \otimes f \parallel g \otimes g) = 2 - 2(1 - H^2(f \parallel g)/2)^2$ (Polyanskiy & Wu, 2023, Eqn. 7.26) and using $1 - (1 - x/2)^2 \le x$ implies that $H^2(f \otimes f \parallel g \otimes g) \le 2H^2(f \parallel g)$. Thus,

$$\begin{aligned}
D_\triangle(f' \parallel g') &\le D_\triangle(f \otimes f \parallel g \otimes g) && \text{(data processing ineq.)} \\
&\le 4H^2(f \otimes f \parallel g \otimes g) && (D_\triangle \le 4H^2) \\
&\le 8H^2(f \parallel g) && \text{(tensorization of } H^2) \\
&\le 4D_\triangle(f \parallel g). && (2H^2 \le D_\triangle)
\end{aligned}$$

$\square$

Lemma 4.3. *For any $f, g \in L^2(\lambda)$, we have*

$$\left|\bar{f} - \bar{g}\right| \le 4\sqrt{\mathrm{Var}(f) D_\triangle(f \parallel g)} + 5D_\triangle(f \parallel g). \tag{2}$$

Proof. If $D_\triangle(f \parallel g) > \frac{1}{2}$, then we trivially have $\left|\bar{f} - \bar{g}\right| \leq 1 \leq 2D_\triangle(f \parallel g)$ since $\bar{f}, \bar{g} \in [0,1]$. Thus, we can assume $D_\triangle(f \parallel g) \leq \frac{1}{2}$. Starting from Eq. (1), we can bound the sum of two variances as follows,

$$\begin{aligned}
\mathrm{Var}(f) + \mathrm{Var}(g) &= 2\,\mathrm{Var}(f) + \mathrm{Var}(g) - \mathrm{Var}(f) \\
&\leq 2\,\mathrm{Var}(f) + 4\sqrt{(\mathrm{Var}(f) + D_\triangle(f \parallel g))D_\triangle(f \parallel g)} && \text{(Eq. (4))} \\
&\leq 2\,\mathrm{Var}(f) + 4\sqrt{\mathrm{Var}(f) D_\triangle(f \parallel g)} + 4D_\triangle(f \parallel g) \\
&\leq 4\,\mathrm{Var}(f) + 6D_\triangle(f \parallel g). && \text{(AM-GM)}
\end{aligned}$$

Hence, we have

$$\begin{aligned}
\left|\bar{f} - \bar{g}\right| &\leq 2\sqrt{(\mathrm{Var}(f) + \mathrm{Var}(g))D_\triangle(f \parallel g)} \\
&\leq 2\sqrt{(4\,\mathrm{Var}(f) + 6D_\triangle(f \parallel g))D_\triangle(f \parallel g)} && \text{(above inequality)} \\
&\leq 4\sqrt{\mathrm{Var}(f) D_\triangle(f \parallel g)} + 5D_\triangle(f \parallel g).
\end{aligned}$$

This finishes the proof. □

C. Proof for Gap-dependent Bounds for CB

Define $d_{\mathsf{CB}}(\varepsilon) = \dim_{\ell_1 \mathsf{E}}(\{(x,a) \mapsto D_\triangle(f(x,a) \parallel C(x,a)) : f \in \mathcal{F}\}, \varepsilon)$ is the ℓ_1-eluder dimension at threshold ε (Liu et al., 2022).

Theorem 4.4. *Assume the premise of Theorem 4.1. If* $\max(\mathrm{Gap}_{\mathsf{Var}}, \mathrm{Gap}_{C^\star}) \geq \frac{1}{\sqrt{K}}$, *then*

$$\mathrm{Regret}_{\mathsf{CB}}(K) \leq \widetilde{\mathcal{O}}\big(d_{\mathsf{CB}}\beta + d_{\mathsf{CB}}\beta \min\{\mathrm{Gap}_{\mathsf{Var}}^{-1}, \mathrm{Gap}_{C^\star}^{-1}\}\big).$$

Proof of Theorem 4.4. Define $\delta_k(x,a) := D_\triangle(f_k(x,a) \parallel C(x,a))$ and $\Delta = \sum_k \delta_k(x_k,a_k)$, the same notation as in Section 4.1. We partition episodes into burn-in and stable episodes, where stable episodes are those that satisfy: $\delta_k(x_k,a_k) \leq \mathrm{Var}(C(x_k,a_k))$. Let $\mathcal{E}$ denote the set of stable episodes and $\neg\mathcal{E}$ are the burn-in episodes.

Step 1: burn-in episodes have $\mathcal{O}(\Delta)$ regret.

$$\begin{aligned}
\sum_{k \in \mathcal{E}_1 \cap \mathcal{E}_2^C} \bar{C}(x_k,a_k) - \min_a \bar{C}(x_k,a) &\leq \sum_{k \in \mathcal{E}_1 \cap \mathcal{E}_2^C} \bar{C}(x_k,a_k) - \bar{f}_k(x_k,a_k) && \text{(optimism)} \\
&\leq \sum_{k \in \mathcal{E}_1 \cap \mathcal{E}_2^C} 4\sqrt{\mathrm{Var}(C(x_k,a_k))\delta_k(x_k,a_k)} + 5\delta_k(x_k,a_k) && \text{(Eq. (2))} \\
&\leq \sum_{k \in \mathcal{E}_1 \cap \mathcal{E}_2^C} 4\delta_k(x_k,a_k) + 5\delta_k(x_k,a_k) && (\neg\mathcal{E}) \\
&\leq \sum_{k=1}^K 9\delta_k(x_k,a_k) = 9\Delta.
\end{aligned}$$

This implies that $\sum_{k \notin \mathcal{E}} \bar{C}(x_k,a_k) - \min_a \bar{C}(x_k,a) \leq 9\Delta$.

Step 2: stable episodes have gap-dependent regret. We now argue those episodes in $\mathcal{E}$ have large gap. For each k, optimism implies that $\bar{f}_k(x_k,a_k) \leq \min_a \bar{C}(x_k,a) = \bar{C}(x_k,a_k) - \mathrm{Gap}(x_k,a_k)$. This implies that $\mathrm{Gap}(x_k,a_k) \leq \bar{C}(x_k,a_k) - \bar{f}_k(x_k,a_k)$. By $\mathcal{E}$, we have $4\sqrt{\mathrm{Var}(C(x_k,a_k))\delta_k(x_k,a_k)} + 5\delta_k(x_k,a_k) \leq 9\sqrt{\mathrm{Var}(C(x_k,a_k))\delta_k(x_k,a_k)}$, and hence the previous display implies

$$\bar{C}(x_k,a_k) - \bar{f}_k(x_k,a_k) \leq 9\sqrt{\mathrm{Var}(C(x_k,a_k))\delta_k(x_k,a_k)}.$$

If this is zero, then the regret for the episode is zero. If this is non-zero, we have $\mathrm{Gap}(x_k, a_k) > 0$ and $\mathrm{Var}(C(x_k, a_k)) > 0$, which implies that

$$9\sqrt{\delta_k(x_k, a_k)} \geq \frac{\mathrm{Gap}(x_k, a_k)}{\sqrt{\mathrm{Var}(C(x_k, a_k))}} \geq \mathrm{Gap}_{\mathsf{Var}}.$$

Now we will invoke the standard peeling technique (Lemma C.2) on $9\sqrt{\delta_k(x_k, a_k)}$. For any $\zeta > 0$, we have

$$\sum_{k=1}^{K} \mathbb{I}\left[\delta_k(x_k, a_k) \geq \zeta\right] \leq 4 d_{\mathsf{CB}}(\zeta)\beta \log(K)\zeta^{-1}, \tag{5}$$

because $\mathbb{I}\left[\delta_k(x_k, a_k) \geq \zeta\right] \leq \zeta^{-1}\delta_k(x_k, a_k)$ and the summation of $\delta_k(x_k, a_k)$ is bounded by the eluder dimension with log factors (Wang et al., 2023b, Theorem 5.3). This indeed satisfies the assumption of Lemma C.2 with $C = 4d_{\mathsf{CB}}(\mathrm{Gap}_{\mathsf{Var}}^2)\beta\log(K)$. Thus, we can bound the stable episode regret as follows:

$$\begin{aligned}
&\sum_{k \in \mathcal{E}_1 \cap \mathcal{E}_2} \bar{C}(x_k, a_k) - \min_a \bar{C}(x_k, a) \\
&\leq \sum_{k \in \mathcal{E}_1 \cap \mathcal{E}_2} 9\sqrt{\mathrm{Var}(C(x_k, a_k))\delta_k(x_k, a_k)} && \text{(same steps as before and } \mathcal{E}) \\
&\leq \sum_{k \in \mathcal{E}_1 \cap \mathcal{E}_2} 9\sqrt{\delta_k(x_k, a_k)} && (C(x_k, a_k) \in [0, 1]) \\
&\leq 18 \cdot 16 d_{\mathsf{CB}}(\mathrm{Gap}_{\mathsf{Var}}^2)\beta\log(K)\,\mathrm{Gap}_{\mathsf{Var}}^{-1}. && (9\sqrt{\delta_k(x_k, a_k)} \leq 18 \text{ and Lemma C.2})
\end{aligned}$$

In the last inequality, note that we invoke Lemma C.2 directly on $\sqrt{\delta_k}$. Thus, we have shown the $\mathrm{Gap}_{\mathsf{Var}}$-dependent regret:

$$\mathrm{Regret}_{\mathsf{CB}}(K) \leq 11 \cdot 4 d_{\mathsf{CB}}(K^{-1})\beta\log(K) + 288\frac{d_{\mathsf{CB}}(\mathrm{Gap}_{\mathsf{Var}}^2)\beta\log(K)}{\mathrm{Gap}_{\mathsf{Var}}}.$$

Following the same steps, and using Lemma C.1, we can prove the same result for $\mathrm{Gap}_{C^\star}$. Therefore, we have shown that

$$\mathrm{Regret}_{\mathsf{CB}}(K) \leq \widetilde{\mathcal{O}}\left(d_{\mathsf{CB}} + \min\left\{\frac{d_{\mathsf{CB}}(\mathrm{Gap}_{\mathsf{Var}}^2)}{\mathrm{Gap}_{\mathsf{Var}}}, \frac{d_{\mathsf{CB}}(\mathrm{Gap}_{C^\star}^2)}{\mathrm{Gap}_{C^\star}}\right\}\right).$$

Finally, notice that if $\mathrm{Gap}_{\mathsf{Var}} \geq \frac{1}{\sqrt{K}}$, $d_{\mathsf{CB}}(\mathrm{Gap}_{\mathsf{Var}}^2) \leq d_{\mathsf{CB}}(1/K) = d_{\mathsf{CB}}$ by monotonicity of the eluder dimension. If $\mathrm{Gap}_{\mathsf{Var}} < \frac{1}{\sqrt{K}}$ then $1/\mathrm{Gap}_{\mathsf{Var}} \geq \sqrt{K}$ anyways, and so this small-gap regime results in a $\mathcal{O}(\sqrt{K})$ bound; in this case, we already have a better second-order bound in Theorem 4.1. This finishes the proof for Theorem 4.4. □

Lemma C.1. *For each episode k, we have*

$$\bar{C}(x_k, a_k) - \min_a \bar{C}(x_k, a) \leq 3\sqrt{\min_a \bar{C}(x_k, a) \cdot \delta_k(x_k, a_k)} + 6\delta_k(x_k, a_k).$$

Proof. By optimism and Wang et al. (2023b, Equation Δ_2), we have

$$\bar{C}(x_k, a_k) - \min_a \bar{C}(x_k, a) \leq \bar{C}(x_k, a_k) - \bar{f}_k(x_k, a_k) \leq 2\sqrt{\bar{C}(x_k, a_k)\delta_k(x_k, a_k)} + \delta_k(x_k, a_k).$$

Using AM-GM, this can be further bounded by $\frac{1}{2}\bar{C}(x_k, a_k) + 3\delta_k(x_k, a_k)$. Rearranging, this implies $\bar{C}(x_k, a_k) \leq 2\min_a \bar{C}(x_k, a) + 6\delta_k(x_k, a_k)$. Therefore, plugging this back into above,

$$\begin{aligned}
\bar{C}(x_k, a_k) - \min_a \bar{C}(x_k, a) &\leq 2\sqrt{(2\min_a \bar{C}(x_k, a) + 6\delta_k(x_k, a_k))\delta_k(x_k, a_k)} + \delta_k(x_k, a_k) \\
&\leq 3\sqrt{\min_a \bar{C}_k(x_k, a) \cdot \delta_k(x_k, a_k)} + 6\delta_k(x_k, a_k).
\end{aligned}$$

□

Lemma C.2 (Peeling Lemma)**.** *Suppose $g_1, g_2, \dots, g_K : \mathcal{Z} \to [0,1]$ and $z_1, z_2, \dots, z_K \in \mathcal{Z}$ satisfy $g_k(z_k) \geq \mathrm{Gap}$ for all k. Moreover, suppose there exists $C > 0$ such that for any $\zeta \geq \mathrm{Gap}$, we have $\sum_k \mathbb{I}\left[g_k(z_k) \geq \zeta\right] \leq C\zeta^{-2}$. Then,*

$$\sum_{k=1}^{K} g_k(z_k) \leq 4C\,\mathrm{Gap}^{-1}.$$

Proof. Divide $[\mathrm{Gap}, 1]$ into $N = \lceil \log(1/\mathrm{Gap}) \rceil$ intervals, where the $i \in [N]$-th interval is $[2^{i-1}\,\mathrm{Gap}, 2^i\,\mathrm{Gap})$. Then, we bound the sum via a standard peeling argument: note that $g_k(z_k)\mathbb{I}\left[g_k(z_k) \in [2^{i-1}\,\mathrm{Gap}, 2^i\,\mathrm{Gap})\right] \leq 2^i\,\mathrm{Gap}\,\mathbb{I}\left[g_k(z_k) \geq 2^{i-1}\,\mathrm{Gap}\right]$. Therefore,

$$\begin{aligned}
\sum_k g_k(z_k) &= \sum_k \sum_{i=1}^{N} g_k(z_k)\mathbb{I}\left[g_k(z_k) \in [2^{i-1}\,\mathrm{Gap}, 2^i\,\mathrm{Gap})\right] \\
&\leq \sum_k \sum_{i=1}^{N} 2^i\,\mathrm{Gap}\,\mathbb{I}\left[g_k(z_k) \geq 2^{i-1}\,\mathrm{Gap}\right] \\
&\leq \sum_{i=1}^{N} 2^i\,\mathrm{Gap} \cdot C 2^{-2i+2}\,\mathrm{Gap}^{-2} && \text{(premise)} \\
&= 4C\,\mathrm{Gap}^{-1} \sum_{i=1}^{N} 2^{-i} \leq 4C\,\mathrm{Gap}^{-1}.
\end{aligned}$$

□

D. RL Lemmas

Lemma D.1 (Performance Difference)**.** *For any $f : (\mathcal{X} \times \mathcal{A} \to \mathbb{R})^H$, policy π and x_1, we have*

$$V^\pi(x_1) - f_1(x_1, \pi(x_1)) = \sum_{h=1}^{H} \mathbb{E}_{\pi, x_1}[(\mathcal{T}_h^\pi f_{h+1} - f_h)(x_h, a_h)].$$

Proof. See Wang et al. (2023b, Lemma H.2). □

Theorem D.2 (Second-order implies Small-loss)**.** *For online RL, suppose we have a second-order bound: $\sum_{k=1}^K V^{\pi^k}(x_{1,k}) - V^\star(x_{1,k}) \leq \sqrt{c \sum_{k=1}^K \mathrm{Var}(Z^{\pi^k}(x_{1,k}))} + c$, for some $c \in \mathbb{R}_+$. Then, we also have a small-loss (first-order) bound: $\sum_{k=1}^K V^{\pi^k}(x_{1,k}) - V^\star(x_{1,k}) \leq \sqrt{2c \sum_{k=1}^K V^\star(x_{1,k})} + 3c$.*

For offline RL, suppose we have a second-order bound w.r.t. comparator policy π_{comp}: $V^{\widehat{\pi}} - V^{\pi_{\text{comp}}} \leq \sqrt{\frac{c' \mathrm{Var}(Z(\pi_{\text{comp}}))}{N}} + \frac{c'}{N}$. Then, we also have a small-loss (first-order) bound: $V^{\widehat{\pi}} - V^{\pi_{\text{comp}}} \leq \sqrt{\frac{c' V^{\pi_{\text{comp}}}}{N}} + \frac{c'}{N}$.

Proof. The offline RL claim follows from $\mathrm{Var}(Z(\pi_{\text{comp}})) \leq V^{\pi_{\text{comp}}}$ because returns are bounded between $[0,1]$ and variance is bounded by second moment, which is bounded by first moment. So, we will focus on the online RL claim for the remainder of the proof.

$$\begin{aligned}
\sum_{k=1}^K V^{\pi^k}(x_{1,k}) - V^\star(x_{1,k}) &\leq \sqrt{c \sum_{k=1}^K \mathrm{Var}(Z^{\pi^k}(x_{1,k})) + c} && \text{(premise)} \\
&\leq \sqrt{c \sum_{k=1}^K V^{\pi^k}(x_{1,k}) + c} && (6) \\
&\leq \frac{1}{2}c + \frac{1}{2}\sum_{k=1}^K V^{\pi^k}(x_{1,k}) + c, && \text{(AM-GM)}
\end{aligned}$$

which implies

$$\sum_{k=1}^{K} V^{\pi^k}(x_{1,k}) \le 2\sum_{k=1}^{K} V^\star(x_{1,k}) + 3c.$$

Plugging this back into Eq. (6) gives

$$\sum_{k=1}^{K} V^{\pi^k}(x_{1,k}) - V^\star(x_{1,k}) \le \sqrt{2c\sum_{k=1}^{K} V^\star(x_{1,k}) + 3c^2} + c,$$

which finishes the proof. □

D.1. Variance Change of Measure

Lemma 5.6 (Change of Variance). *For any $f : \mathcal{X}\times\mathcal{A}\to\Delta([0,1])$, π and x_1, we have*

$$\begin{aligned}&\mathbb{E}_{\pi,x_1}[\mathrm{Var}(f_h(x_h,a_h))] \le 2e\,\mathrm{Var}(Z^\pi(x_1)) + \\ &12H^2\mathbb{E}_{\pi,x_1}\Big[\textstyle\sum_{t\ge h} D_\triangle(f_t(x_t,a_t) \parallel \mathcal{T}_t^{\pi,D} f_{t+1}(x_t,a_t))\Big].\end{aligned} \tag{3}$$

Proof. Apply law of total variance to the variance term of Theorem D.3, *i.e.*,

$$\begin{aligned}\mathrm{Var}(Z_1^\pi(x_1)) &= \mathbb{E}_{\pi,x_1}[\mathrm{Var}(f_h(x_h,a_h)\mid x_h,a_h,x_1)\mid x_1] + \mathrm{Var}_{\pi,x_1}(\mathbb{E}[f_h(x_h,a_h)\mid x_h,a_h,x_1]\mid x_1)\\ &\ge \mathbb{E}_{\pi,x_1}[\mathrm{Var}(f_h(x_h,a_h)\mid x_h,a_h,x_1)\mid x_1].\end{aligned}$$

□

Theorem D.3. *Fix any $f : \mathcal{X}\times\mathcal{A}\to\Delta([0,1])$ and any policy π. Define $\delta_h(x,a) := D_\triangle(f_h(x,a) \parallel \mathcal{T}_h^{\pi,D} f_{h+1}(x,a))$ and $\Delta_h(x_h,a_h) := \sum_{t=h}^{H}\mathbb{E}_{\pi,x_h,a_h}[\delta_t(x_t,a_t)]$. Then, for all $h\in[H], x_h, a_h$, we have*

$$\mathrm{Var}(f_h(x_h,a_h)) \le 2e\,\mathrm{Var}(Z_h^\pi(x_h,a_h)) + 12H(H-h+1)\Delta_h(x_h,a_h). \tag{7}$$

Therefore, for any x_1,

$$\mathbb{E}_{\pi,x_1}[\mathrm{Var}(f_h(x_h,a_h))] \le 2e\,\mathrm{Var}(Z_1^\pi(x_1)) + 12H^2\mathbb{E}_{\pi,x_1}[\Delta_h(x_h,a_h)]. \tag{8}$$

Proof. The main technical lemma is Lemma D.4, which is proven with induction. Given this lemma, use the fact that $(1+H^{-1})^H \le e$ to get

$$\mathrm{Var}(f_h(x_h,a_h)) \le \sum_{t=h}^{H} e\big(\mathbb{E}_{\pi,x_h,a_h}\big[2\,\mathrm{Var}_{c_t,x_{t+1}}(c_t + V_{t+1}^\pi(x_{t+1})\mid x_t,a_t) + 12H\Delta_t(x_t,a_t)\big]\big).$$

Recall that $\mathrm{Var}(Z_h^\pi(x_h,a_h)) = \sum_{t=h}^{H}\mathbb{E}_{\pi,x_h,a_h}\big[\mathrm{Var}_{c_t,x_{t+1}}(c_t+V_{t+1}^\pi(x_{t+1})\mid x_t,a_t)\big]$, by the law of total variance. Also for any $t\ge h$, we have $\mathbb{E}_{\pi,x_h,a_h}\Delta_t(x_t,a_t) \le \Delta_h(x_h,a_h)$. Thus,

$$\mathrm{Var}(f_h(x_h,a_h)) \le 2e\,\mathrm{Var}(Z_h^\pi(x_h,a_h)) + 12H(H-h+1)\Delta_h(x_h,a_h),$$

which proves the claim. □

Lemma D.4. *For all $h\in[H], x_h, a_h$, we have*

$$\mathrm{Var}(f_h(x_h,a_h)) \le \sum_{t=h}^{H}(1+H^{-1})^{t-h+1}\big(\mathbb{E}_{\pi,x_h,a_h}\big[2\,\mathrm{Var}_{c_t,x_{t+1}}(c_t+V_{t+1}^\pi(x_{t+1})\mid x_t,a_t) + 12H\Delta_t(x_t,a_t)\big]\big). \tag{9}$$

Proof. First observe that

$$\operatorname{Var}(f_h(x_h, a_h)) \le (1 + H^{-1}) \operatorname{Var}(\mathcal{T}_h^{\pi,D} f_{h+1}(x_h, a_h)) + 12H\delta_h(x_h, a_h), \tag{10}$$

because by Eq. (4) and AM-GM, we have

$$\begin{aligned}
\operatorname{Var}(f_h(x_h, a_h)) - \operatorname{Var}(\mathcal{T}_h^{\pi,D} f_{h+1}(x_h, a_h)) &\le 4\sqrt{(\operatorname{Var}(\mathcal{T}_h^{\pi,D} f_{h+1}(x_h, a_h)) + \delta_h(x_h, a_h))\delta_h(x_h, a_h)} \\
&\le 4\sqrt{\operatorname{Var}(\mathcal{T}_h^{\pi,D} f_{h+1}(x_h, a_h))\delta_h(x_h, a_h)} + 4\delta_h(x_h, a_h) \\
&\le H^{-1} \operatorname{Var}(\mathcal{T}_h^{\pi,D} f_{h+1}(x_h, a_h)) + 8H\delta_h(x_h, a_h) + 4\delta_h(x_h, a_h).
\end{aligned}$$

We now proceed to show Eq. (9) by induction. The base case $h = H$ is true since $\operatorname{Var}(\mathcal{T}_H^{\pi,D} f_{H+1}(x_H, a_H)) = \operatorname{Var}(C_H(x_H, a_H)) = \operatorname{Var}(c_H + V_{H+1}^\pi(x_{H+1}) \mid x_H, a_H)$.

We now prove the induction step: suppose the Eq. (9) is true for $h+1$; we want to show the h case is true. By the law of total conditional variance, we have that $\operatorname{Var}(\mathcal{T}_h^{\pi,D} f_{h+1}(x_h, a_h))$ is equal to:

$$\begin{aligned}
&\mathbb{E}[\operatorname{Var}(c_h + f_{h+1}(x_{h+1}, \pi(x_{h+1})) \mid x_{h+1}, c_h, x_h, a_h) \mid x_h, a_h] \\
&+ \operatorname{Var}(\mathbb{E}[c_h + f_{h+1}(x_{h+1}, \pi(x_{h+1})) \mid x_{h+1}, c_h, x_h, a_h] \mid x_h, a_h) \\
&= \mathbb{E}[\operatorname{Var}(f_{h+1}(x_{h+1}, \pi(x_{h+1})) \mid x_{h+1}) \mid x_h, a_h] + \operatorname{Var}_{c_h, x_{h+1} \sim C_h, P_h(x_h, a_h)}\big(c_h + \bar{f}_{h+1}(x_{h+1}, \pi(x_{h+1}))\big).
\end{aligned}$$

The first term is controlled by the induction hypothesis. The second term is handled by Lemma D.5. Therefore,

$$\begin{aligned}
&\operatorname{Var}(\mathcal{T}_h^{\pi,D} f_{h+1}(x_h, a_h)) \\
&\le \mathbb{E}_{\pi, x_h, a_h} \sum_{t=h+1}^{H} (1 + H^{-1})^{t-h} \big(2\mathbb{E}_{\pi, x_{h+1}, a_{h+1}}\big[\operatorname{Var}_{c_t, x_{t+1}}\big(c_t + V_{t+1}^\pi(x_{t+1}) \mid x_t, a_t\big) + 12H\Delta_t(x_t, a_t)\big]\big) \\
&+ 2\operatorname{Var}_{c_h, x_{h+1} \sim C_h, P_h(x_h, a_h)}\big(c_h + V_{h+1}^\pi(x_{h+1})\big) + 4H\mathbb{E}_{\pi, x_h, a_h}\Delta_{h+1}(x_{h+1}, a_{h+1}).
\end{aligned}$$

Thus, by Eq. (10), we have

$$\begin{aligned}
&\operatorname{Var}(f_h(x_h, a_h)) \\
&\le \mathbb{E}_{\pi, x_h, a_h} \sum_{t=h+1}^{H} (1 + H^{-1})^{t-h+1} \big(2\mathbb{E}_{\pi, x_{h+1}, a_{h+1}}\big[\operatorname{Var}_{c_t, x_{t+1}}\big(c_t + V_{t+1}^\pi(x_{t+1}) \mid x_t, a_t\big) + 12H\Delta_t(x_t, a_t)\big]\big) \\
&+ (1 + H^{-1})\big(2\operatorname{Var}_{c_h, x_{h+1} \sim C_h, P_h(x_h, a_h)}\big(c_h + V_{h+1}^\pi(x_{h+1})\big) + 4H\mathbb{E}_{\pi, x_h, a_h}\Delta_{h+1}(x_{h+1}, a_{h+1})\big) + 12H\delta_h(x_h, a_h) \\
&\le \sum_{t=h+1}^{H} (1 + H^{-1})^{t-h+1} \big(2\mathbb{E}_{\pi, x_h, a_h}\big[\operatorname{Var}_{c_t, x_{t+1}}\big(c_t + V_{t+1}^\pi(x_{t+1}) \mid x_t, a_t\big) + 12H\Delta_t(x_t, a_t)\big]\big) \\
&+ (1 + H^{-1})\big(2\operatorname{Var}_{c_h, x_{h+1} \sim C_h, P_h(x_h, a_h)}\big(c_h + V_{h+1}^\pi(x_{h+1})\big) + 12H\Delta_h(x_h, a_h)\big) \\
&= \sum_{t=h}^{H} (1 + H^{-1})^{t-h+1} \big(2\mathbb{E}_{\pi, x_h, a_h}\big[\operatorname{Var}_{c_t, x_{t+1}}\big(c_t + V_{t+1}^\pi(x_{t+1}) \mid x_t, a_t\big) + 12H\Delta_t(x_t, a_t)\big]\big),
\end{aligned}$$

which finishes the induction. ◻

Lemma D.5.

$$\begin{aligned}
&\operatorname{Var}_{c_h, x_{h+1} \sim C_h, P_h(x_h, a_h)}\big(c_h + \bar{f}_{h+1}(x_{h+1}, \pi(x_{h+1}))\big) \\
&\le 2\operatorname{Var}_{c_h, x_{h+1} \sim C_h, P_h(x_h, a_h)}\big(c_h + V_{h+1}^\pi(x_{h+1})\big) + 4(H - h)\mathbb{E}_{\pi, x_h, a_h}\Delta_{h+1}(x_{h+1}, a_{h+1}).
\end{aligned}$$

Proof. Recall that $\operatorname{Var}(X + Y) \le 2\operatorname{Var}(X) + 2\operatorname{Var}(Y)$ and hence,

$$\begin{aligned}
&\operatorname{Var}_{c_h, x_{h+1} \sim C_h, P_h(x_h, a_h)}\big(c_h + \bar{f}_{h+1}(x_{h+1}, \pi(x_{h+1}))\big) \\
&\le 2\operatorname{Var}_{c_h, x_{h+1} \sim C_h, P_h(x_h, a_h)}\big(c_h + V_{h+1}^\pi(x_{h+1})\big) + 2\operatorname{Var}_{x_{h+1} \sim P_h(x_h, a_h)}\big(\bar{f}_{h+1}(x_{h+1}, \pi(x_{h+1})) - V_{h+1}^\pi(x_{h+1})\big)
\end{aligned}$$

For the second term, we first bound the envelope of $\bar{f}_{h+1}(x_{h+1}, \pi(x_{h+1})) - V^{\pi}_{h+1}(x_{h+1})$ as follows:

$$\left|\bar{f}_{h+1}(x_{h+1}, \pi(x_{h+1})) - V^{\pi}_{h+1}(x_{h+1})\right| \leq \sum_{t=h+1}^{H} \mathbb{E}_{\pi, x_{h+1}}\left[\left|\bar{f}_t(x_t, a_t) - \mathcal{T}^{\pi}_t \bar{f}_{t+1}(x_t, a_t)\right|\right] \quad \text{(PDL)}$$

$$\leq \sum_{t=h+1}^{H} \mathbb{E}_{\pi, x_{h+1}}\left[\sqrt{2\delta_t(x_t, a_t)}\right] \quad \text{(Eq.}(\Delta_1)\text{ of Wang et al. (2023b))}$$

This enables us to bound the variance,

$$\begin{aligned}
&\mathrm{Var}_{x_{h+1} \sim P_h(x_h, a_h)}\left(\bar{f}_{h+1}(x_{h+1}, \pi(x_{h+1})) - V^{\pi}_{h+1}(x_{h+1})\right) \\
&\leq \mathbb{E}_{x_{h+1} \sim P_h(x_h, a_h)}\left[\left(\bar{f}_{h+1}(x_{h+1}, \pi(x_{h+1})) - V^{\pi}_{h+1}(x_{h+1})\right)^2\right] \\
&\leq \mathbb{E}_{x_{h+1} \sim P_h(x_h, a_h)}\left[\left(\sum_{t=h+1}^{H} \mathbb{E}_{\pi, x_{h+1}}\left[\sqrt{2\delta_t(x_t, a_t)}\right]\right)^2\right] \\
&\leq (H-h)\mathbb{E}_{x_{h+1} \sim P_h(x_h, a_h)}\left[\sum_{t=h+1}^{H}\left(\mathbb{E}_{\pi, x_{h+1}}\left[\sqrt{2\delta_t(x_t, a_t)}\right]\right)^2\right] \\
&\leq (H-h)\mathbb{E}_{x_{h+1} \sim P_h(x_h, a_h)}\left[\sum_{t=h+1}^{H} \mathbb{E}_{\pi, x_{h+1}}[2\delta_t(x_t, a_t)]\right],
\end{aligned}$$

as desired. □

E. Proofs for Online RL

Theorem 5.3 (Second-order bounds for Online RL)**.** *Under Assumption 5.1, for any $\delta \in (0,1)$, w.p. at least $1-\delta$, running* O-DISCO *with $\beta = \log(HK|\mathcal{F}|/\delta)$ has $\mathrm{Reg}_{\mathsf{RL}}(K)$ at most,*

$$\widetilde{\mathcal{O}}\left(H\sqrt{\sum_{k=1}^{K} \mathrm{Var}(Z^{\pi^k}(x_{1,k})) \cdot d_{\mathsf{RL}}\beta} + H^{2.5} d_{\mathsf{RL}}\beta\right).$$

If UAE=TRUE*, then the learned mixture policy $\bar{\pi}$ enjoys the PAC bound: w.p. at least $1-\delta$, $K(V^{\bar{\pi}} - V^{\star})$ is at most,*

$$\widetilde{\mathcal{O}}\left(H\sqrt{A\sum_{k=1}^{K} \mathrm{Var}(Z^{\pi^k}(x_{1,k})) d_{\mathsf{RL,V}}\beta} + AH^{2.5} d_{\mathsf{RL,V}}\beta\right).$$

Proof of Theorem 5.3. As noted by (Wang et al., 2023b, Proof of Theorem 5.5), the confidence set construction guarantees two facts w.p. $1-\delta$: **for all** $k \in [K]$,

(i) Optimism: $\min_a \bar{f}^{(k)}_1(x_{1,k}, a) \leq V^{\star}(x_{1,k})$ (since $Z^{\pi}(x_{1,k}) \in \mathcal{F}_k$); and

(ii) Small-generalization error: for all h, we have

If UAE=FALSE. $\sum_{i<k} \mathbb{E}_{\pi^i}[\delta_{h,k}(s_h, a_h)] \leq c\beta$;
If UAE=TRUE. $\sum_{i<k} \mathbb{E}_{\pi^i}\left[\mathbb{E}_{a' \sim \mathrm{unif}(\mathcal{A})}[\delta_{h,k}(s_h, a_h)]\right] \leq c\beta$,

for some universal constant c.

Let $\delta_{h,k}(x,a) := D_{\triangle}(f^{(k)}_h(x,a) \parallel \mathcal{T}^{\star,D}_h f^{(k)}_{h+1}(x,a))$ and $\Delta_k := \sum_{h=1}^{H} \mathbb{E}_{\pi^k, x_{1,k}}[\delta_{h,k}(x_h, a_h)]$. We now decompose the regret into two parts.

$$
\begin{aligned}
&\sum_k V^{\pi^k}(x_{1,k}) - V^\star(x_{1,k}) \\
&\leq \sum_k V^{\pi^k}(x_{1,k}) - \min_a \bar{f}_1^{(k)}((x_{1,k}), a) && \text{(Optimism)} \\
&= \sum_k \sum_{h=1}^{H} \mathbb{E}_{\pi^k, x_{1,k}}\left[\mathcal{T}_h^{\pi^k} \bar{f}_{h+1}^{(k)}(x_h, a_h) - \bar{f}_h^{(k)}(x_h, a_h)\right] && \text{(PDL)} \\
&= \sum_k \sum_{h=1}^{H} \mathbb{E}_{\pi^k, x_{1,k}}\left[\overline{\mathcal{T}_h^{\pi^k} f_{h+1}^{(k)}}(x_h, a_h) - \bar{f}_h^{(k)}(x_h, a_h)\right] \\
&\leq \sum_{h,k} 4\sqrt{\mathbb{E}_{\pi^k, x_{1,k}}[\mathrm{Var}(f_h^{(k)}(x_h, a_h))] \cdot \mathbb{E}_{\pi^k, x_{1,k}}[\delta_{h,k}(x_h, a_h)]} + 5\mathbb{E}_{\pi^k, x_{1,k}}[\delta_{h,k}(x_h, a_h)] && \text{(Eq. (2))} \\
&\leq \sum_{h,k} 4\sqrt{\left(\mathrm{Var}(Z^{\pi^k}(x_{1,k})) + 12H^2\Delta_k\right) \cdot \mathbb{E}_{\pi^k, x_{1,k}}[\delta_{h,k}(x_h, a_h)]} + 5\mathbb{E}_{\pi^k, x_{1,k}}[\delta_{h,k}(x_h, a_h)] && \text{(Eq. (3))} \\
&\leq \sum_k 4\sqrt{\left(2e\,\mathrm{Var}(Z^{\pi^k}(x_{1,k})) + 12H^2\Delta_k\right) \cdot H\Delta_k} + 5\Delta_k && \text{(Cauchy-Schwarz)} \\
&\leq \sum_k 4\sqrt{2e\,\mathrm{Var}(Z^{\pi^k}(x_{1,k}))H\Delta_k} + (4\sqrt{12}+5)H^{1.5}\Delta_k \\
&\leq 4\sqrt{2e\sum_k \mathrm{Var}(Z^{\pi^k}(x_{1,k}))H\sum_k \Delta_k} + (4\sqrt{12}+5)H^{1.5}\sum_k \Delta_k.
\end{aligned}
$$

The final step is to bound $\sum_k \Delta_k$, which is the same as in (Wang et al., 2023b). In particular, if UAE=FALSE, then $\sum_k \Delta_k \leq cH \dim_{\ell_1,DE}(1/K)\beta \log(K)$. If UAE=TRUE, then $\sum_k \Delta_k \leq cAH \dim_{\ell_1,DE}(1/K)\beta \log(K)$. This concludes the proof. □

E.1. Bounding Q-type distributional Eluder in Linear MDPs

Recall the Linear MDP definition (Jin et al., 2020).

Definition E.1 (Linear and Low-Rank MDP)**.** A transition model $P_h : \mathcal{X} \times \mathcal{A} \to \Delta(\mathcal{X})$ has rank d if there exist features $\phi_h^\star : \mathcal{X} \times \mathcal{A} \to \mathbb{R}^d, \mu_h^\star : \mathcal{X} \to \mathbb{R}^d$ such that $P_h(x' \mid x, a) = \phi_h^\star(x,a)^\top \mu_h^\star(x')$ for all x, a, x'. Also, assume $\max_{x,a} \|\phi_h^\star(x,a)\|_2 \leq 1$ and $\|\int g \mathrm{d}\mu_h^\star\|_2 \leq \|g\|_\infty \sqrt{d}$ for all functions $g : \mathcal{X} \to \mathbb{R}$. The MDP is called low-rank if P_h is low-rank for all $h \in [H]$. The MDP is called linear if $\{\phi_h^\star\}_{h\in[H]}$ is known.

Consider the following linear function class:

$$
\mathcal{F}_h^{\text{lin}} = \left\{ f(z \mid x, a) = \langle \phi^\star(x,a), w(z) \rangle \text{ s.t. } w : [0,1] \to \mathbb{R}^d, \max_z \|w(z)\|_2 \leq \alpha\sqrt{d} \text{ and } \max_{x,a,z} \langle \phi^\star(x,a), w(z) \rangle \leq \alpha \right\}, \tag{11}
$$

Wang et al. (2023b) showed two nice facts about $\mathcal{F}^{\text{lin}}$. First, it satisfies Bellman Completeness (Assumption 5.1). Moreover, under the assumption that costs are discretized into a uniform grid of M points, this class's bracketing entropy is $\widetilde{\mathcal{O}}(dM)$. Note that discretization is necessary to bound the statistical complexity of $\mathcal{F}^{\text{lin}}$ and is also common in practice, *e.g.*, C51 (Bellemare et al., 2017) and Rainbow (Hessel et al., 2018) both set $M = 51$, which works well in Atari; also the optimal policy's performance in the discretized MDP can also be bounded by the discretization error (Wang et al., 2023a).

We now show a new fact about $\mathcal{F}^{\text{lin}}$. If we further assume that per-step cost and cost-to-go distributions have minimum mass $\eta_{\min} > 0$ on each element of its support, then we can bound the appropriate Q-type distributional eluder dimension for linear MDPs as $\widetilde{\mathcal{O}}(d\eta_{\min}^{-1}\log(1/\varepsilon))$. This is formalized in the following assumption.

Assumption E.2. For all $f \in \mathcal{F}^{\text{lin}}$ and $h \in [H]$, if $f_h(z \mid x, a) = \mathcal{T}_h^{\star,D} f_{h+1}(z \mid x, a)$, then $f_h(z \mid x, a) + \mathcal{T}_h^{\star,D} f_{h+1}(z \mid x, a) \geq \eta_{\min}$.

If cost-to-go and per-step cost distributions have a minimum mass, then this assumption is satisfied.

Theorem E.3. *Suppose the MDP is a linear MDP and Assumption E.2. Fix any $h \in [H]$ and define*

$$\Psi_h = \left\{(x,a) \mapsto D_\triangle(f_h(x,a) \parallel \mathcal{T}_h^{\star,D} f_{h+1}(x,a)) : f \in \mathcal{F}_h^{\text{lin}}\right\},$$
$$\mathcal{D}_h = \{(x,a) \mapsto d_h^\pi(x,a) : \pi \in \Pi\}.$$

Then, $\dim_{\ell_1 DE}(\Psi_h, \mathcal{D}_h, \varepsilon) \le \mathcal{O}(d\eta_{\min}^{-1} \log(dM/(\eta_{\min}\varepsilon)))$.

Proof. Fix any h. Suppose $(d^{(k)}, f^{(k)})_{k\in[T]}$ is any sequence such that for all $k \in [T]$, $d^{(k)} \in \mathcal{D}_h$, $f^{(k)} \in \Psi_h$ and $(d^{(k)}, f^{(k)})$ is (ε, ℓ_1)-independent of its predecessors. By definition, the largest possible T is the eluder dimension of interest, so we now proceed to bound T.

For any k, since $f^{(k)} \in \Psi_h$, there exists $w^{(k)}, v^{(k)} : [0,1] \to \mathbb{R}^d$ satisfying normalization constraints of Eq. (11) such that $f^{(k)}(x,a) = D_\triangle(z \mapsto \phi_h^\star(x,a)^\top w^{(k)}(z) \parallel z \mapsto \phi_h^\star(x,a)^\top v^{(k)}(z))$. Note that $v^{(k)}$ exists by Bellman completeness of $\mathcal{F}_h^{\text{lin}}$.

Now we simplify the $D_\triangle$ term with the assumption: for any k,

$$\begin{aligned}
\mathbb{E}_{d^{(k)}} D_\triangle(f_h^{(k)}(x,a) \parallel \mathcal{T}_h^{\star,D} f_{h+1}^{(k)}(x,a)) &= \mathbb{E}_{d^{(k)}} \sum_z \frac{(f_h^{(k)}(z \mid x,a) - \mathcal{T}_h^{\star,D} f_{h+1}^{(k)}(z \mid x,a))^2}{f_h^{(k)}(z \mid x,a) + \mathcal{T}_h^{\star,D} f_{h+1}^{(k)}(z \mid x,a)} \\
&\le \eta_{\min}^{-1} \mathbb{E}_{d^{(k)}} \sum_z (\phi_h^\star(x,a)^\top (w^{(k)}(z) - v^{(k)}(z)))^2 && \text{(Assumption E.2)} \\
&\le \eta_{\min}^{-1} \mathbb{E}_{d^{(k)}} \|\phi_h^\star(x,a)\|_{\Sigma_k^{-1}}^2 \cdot \sum_z \|w^{(k)}(z) - v^{(k)}(z)\|_{\Sigma_k}^2, && \text{(CS)}
\end{aligned}$$

where $\Sigma_k := \sum_{i<k} \mathbb{E}_{d^{(i)}}[\phi_h^\star(x_h,a_h)\phi_h^\star(x_h,a_h)^\top] + \lambda I$ and $\lambda > 0$ will be set soon. For the second factor,

$$\begin{aligned}
\sum_z \|w^{(k)}(z) - v^{(k)}(z)\|_{\Sigma_k}^2 &= \sum_z \sum_{i<k} \mathbb{E}_{d^{(i)}} \left(\phi_h^\star(x,a)^\top (w^{(k)}(z) - v^{(k)}(z))\right)^2 + M\lambda d \\
&\le \sum_{i<k} \mathbb{E}_{d^{(i)}} \left(\sum_z \left|\phi_h^\star(x,a)^\top (w^{(k)}(z) - v^{(k)}(z))\right|\right)^2 + M\lambda d \\
&\le \sum_{i<k} \mathbb{E}_{d^{(i)}} D_\triangle(f_h^{(k)}(x,a) \parallel \mathcal{T}_h^{\star,D} f_{h+1}^{(k)}(x,a)) + M\lambda d && (D_{TV}^2 \le D_\triangle) \\
&\le \varepsilon + M\lambda d && ((\varepsilon,\ell_1)\text{-independent sequence}) \\
&= 2\varepsilon. && (\text{set } \lambda = \varepsilon/(dM))
\end{aligned}$$

Thus, we have shown that

$$\begin{aligned}
T\varepsilon &< \sum_k \mathbb{E}_{d^{(k)}} D_\triangle(f_h^{(k)}(x,a) \parallel \mathcal{T}_h^{\star,D} f_{h+1}^{(k)}(x,a)) && ((\varepsilon,\ell_1)\text{-independent sequence}) \\
&\le \eta_{\min}^{-1} \sum_k \mathbb{E}_{d^{(k)}} \|\phi_h^\star(x,a)\|_{\Sigma_k^{-1}}^2 \cdot 2\varepsilon \\
&\le 2\eta_{\min}^{-1} \varepsilon \cdot d \log(1 + TM/\varepsilon^2),
\end{aligned}$$

where we used elliptical potential in the last step (Uehara et al., 2021, Lemma 19 & 20), which is applicable since $\mathbb{E}_{d^{(k)}} \|\phi_h^\star(x,a)\|_{\Sigma_k^{-1}}^2 = \mathbb{E}_{d^{(k)}} \phi_h^\star(x,a)^\top \Sigma_k^{-1} \phi_h^\star(x,a) = \mathrm{Tr}(\mathbb{E}_{d^{(k)}}[\phi_h^\star(x,a)\phi_h^\star(x,a)^\top]\Sigma_k^{-1})$. Thus, (Uehara et al., 2021, Lemma 19 & 20) implies that

$$T < 2\eta_{\min}^{-1} d \log(1 + TM/\varepsilon^2),$$

which finally implies,

$$T \le 12\eta_{\min}^{-1} d \log(1 + 2\eta_{\min}^{-1} dM/\varepsilon^2),$$

by (Wang et al., 2023b, Lemma G.5).

□

F. Proofs for Offline RL

Theorem 6.1 (Second-order bounds for Offline RL)**.** *Under Assumption 5.1, for any $\delta \in (0,1)$, w.p. at least $1-\delta$, running* P-DISCO *with $\beta = \log(H|\Pi||\mathcal{F}|/\delta)$ learns a policy $\widehat{\pi}$ that enjoys the following bound: for any comparator $\widetilde{\pi} \in \Pi$ (not necessarily the optimal $\pi^\star$), we have*

$$V^{\widehat{\pi}} - V^{\widetilde{\pi}} \leq \mathcal{O}\left(H\sqrt{\frac{\operatorname{Var}(Z^{\widetilde{\pi}})C^{\widetilde{\pi}}\beta}{N}} + \frac{H^{2.5}C^{\widetilde{\pi}}\beta}{N} \right).$$

Proof of Theorem 6.1. As noted by (Wang et al., 2023b, Proof of Theorem 6.1), the confidence set construction guarantees two facts w.p. $1-\delta$:

(i) Pessimism: for all π, $V^\pi \leq \bar{f}_1^\pi(x_1,\pi)$ (since $Z^\pi \in \mathcal{F}_\pi$); and

(ii) Small-generalization error: for all π and h, $\mathbb{E}_{\nu_h}[D_\triangle(f_h^\pi(x,a) \parallel \mathcal{T}_h^{\pi,D} f_{h+1}^\pi(x,a))] \leq c\beta N^{-1}$ for some universal constant c.

Let $\delta_h^\pi(x,a) := D_\triangle(f_h^\pi(x,a) \parallel \mathcal{T}_h^{\pi,D} f_{h+1}^\pi(x,a))$ and $\Delta^\pi := \sum_{h=1}^H \mathbb{E}_\pi[\delta_h^\pi(x_h,a_h)]$. We now bound the performance difference between $\widehat{\pi}$ and $\widetilde{\pi}$:

$$\begin{aligned}
V^{\widehat{\pi}} - V^{\widetilde{\pi}} &\leq \bar{f}_1^{\widehat{\pi}}(x_1,\widehat{\pi}) - V^{\widetilde{\pi}} && \text{(Pessimism)} \\
&\leq \bar{f}_1^{\widetilde{\pi}}(x_1,\widetilde{\pi}) - V^{\widetilde{\pi}} && \text{(Defn of } \widehat{\pi}) \\
&= \sum_{h=1}^H \mathbb{E}_{\widetilde{\pi}}\left[\left(\bar{f}_h^{\widetilde{\pi}} - \mathcal{T}_h^{\widetilde{\pi}} \bar{f}_{h+1}^{\widetilde{\pi}}\right)(x_h,a_h)\right] && \text{(PDL Lemma D.1)} \\
&\leq \sum_{h=1}^H 4\sqrt{\mathbb{E}_{\widetilde{\pi}}[\operatorname{Var}(f_h^{\widetilde{\pi}}(x_h,a_h))] \cdot \mathbb{E}_{\widetilde{\pi}}[\delta_h^{\widetilde{\pi}}(x_h,a_h)]} + 5\mathbb{E}_{\widetilde{\pi}}[\delta_h^{\widetilde{\pi}}(x_h,a_h)] && \text{(Eq. (2))} \\
&\leq \sum_{h=1}^H 4\sqrt{(2e\operatorname{Var}(Z^{\widetilde{\pi}}) + 12H^2\Delta^{\widetilde{\pi}}) \cdot \mathbb{E}_{\widetilde{\pi}}[\delta_h^{\widetilde{\pi}}(x_h,a_h)]} + 5\mathbb{E}_{\widetilde{\pi}}[\delta_h^{\widetilde{\pi}}(x_h,a_h)] && \text{(Eq. (3))} \\
&\leq 4\sqrt{(2e\operatorname{Var}(Z^{\widetilde{\pi}}) + 12H^2\Delta^{\widetilde{\pi}}) \cdot H\Delta^{\widetilde{\pi}}} + 5\Delta^{\widetilde{\pi}} && \text{(Cauchy-Schwarz)} \\
&\leq 4\sqrt{2e\operatorname{Var}(Z^{\widetilde{\pi}})H\Delta^{\widetilde{\pi}}} + (4\sqrt{12}+5)H^{1.5}\Delta^{\widetilde{\pi}}.
\end{aligned}$$

Finally, bound $\Delta^{\widetilde{\pi}}$ by change of measure and the generalization bound of MLE (fact (ii)):

$$\Delta^{\widetilde{\pi}} \leq C^{\widetilde{\pi}} \sum_{h=1}^H \mathbb{E}_{\nu_h}[\delta_h^{\widetilde{\pi}}(x_h,a_h)] \leq C^{\widetilde{\pi}} H \cdot c\beta N^{-1}.$$

Therefore,

$$V^{\widehat{\pi}} - V^{\widetilde{\pi}} \leq \mathcal{O}\left(H\sqrt{\frac{C^{\widetilde{\pi}}\operatorname{Var}(Z^{\widetilde{\pi}})\beta}{N}} + \frac{H^{2.5}C^{\widetilde{\pi}}\beta}{N} \right).$$

□