

The Central Role of the Loss Function in Reinforcement Learning

Kaiwen Wang, Nathan Kallus and Wen Sun

Abstract. This paper illustrates the central role of loss functions in data-driven decision making, providing a comprehensive survey on their influence in cost-sensitive classification (CSC) and reinforcement learning (RL). We demonstrate how different regression loss functions affect the sample efficiency and adaptivity of value-based decision making algorithms. Across multiple settings, we prove that algorithms using the binary cross-entropy loss achieve first-order bounds scaling with the optimal policy's cost and are much more efficient than the commonly used squared loss. Moreover, we prove that distributional algorithms using the maximum likelihood loss achieve second-order bounds scaling with the policy variance and are even sharper than first-order bounds. This in particular proves the benefits of distributional RL. We hope that this paper serves as a guide analyzing decision making algorithms with varying loss functions, and can inspire the reader to seek out better loss functions to improve any decision making algorithm.

Key words and phrases: First-Order (Small-Loss) and Second-Order (Variance-Dependent) Bounds, RL with Function Approximation, Distributional RL.

1. INTRODUCTION

The value-based approach to reinforcement learning (RL) reduces the decision making problem to regression: first predict the expected rewards to go under the optimal policy, given state and action, and then one can simply choose the action that maximizes the prediction at every state. This regression, called Q -learning [70], combined with recent decades' advances in deep learning, plays a central role in the empirical successes of deep RL. Groundbreaking examples are DeepMind's use of deep Q -networks to play Atari with no feature engineering [51] and OpenAI's use of deep reward models to align large language models with human preferences via RL fine-tuning [54].

In prediction, we often say a good model is one with low mean-squared error out of sample. Correspondingly, regression is usually done by minimizing the average squared loss between predictions and targets in the training data. However, low mean-squared error may translate loosely to high-quality, downstream decision making. Thus, the natural question arises: is squared loss always the best choice for learning Q -functions?

In this survey article, we highlight that the answer to this question is a resounding “no!” We focus on the the-

TABLE 1

The decision-making regret of value-based RL per loss function, where n is the number of samples. In this paper, we'll see that squared loss cannot adapt to small-cost or small-variance settings while binary-cross-entropy (bce) and maximum likelihood estimation (mle) can. We remark that mle is used with distributional regression and thus requires stronger realizability or completeness conditions. Moreover, all three losses use slightly different eluder dimensions in the online setting. We summarize these nuances in Table 2.

Loss \ Setting	Worst-case	Small cost	Small variance
ℓ_{sq}	$\Theta(1/\sqrt{n})$	$\Theta(1/\sqrt{n})$	$\Theta(1/\sqrt{n})$
ℓ_{bce}	$\Theta(1/\sqrt{n})$	$\mathcal{O}(1/n)$	$\Theta(1/\sqrt{n})$
ℓ_{mle}	$\Theta(1/\sqrt{n})$	$\mathcal{O}(1/n)$	$\mathcal{O}(1/n)$

oretical question: when and how do alternative loss functions attain better guarantees for decision making? We start with the simple setting of cost-sensitive classification (CSC) and show that the binary cross-entropy (bce) loss leads to an improved convergence guarantee that adapts to problem instances with low optimal cost; such a result is called a first-order bound. This result is due to Foster and Krishnamurthy [24], who first observed that the bce loss can have benefits over square loss when outcomes are heteroskedastic and used this observation to derive first-order regret bounds for CSC and contextual bandits (CB). Then, we prove that the maximum likelihood (mle) loss yields an even better convergence guarantee that additionally adapts to problem instances with low variance;

The authors are from Cornell University. Correspondence to Kaiwen Wang (<https://kaiwenw.github.io>).

such a result is called a second-order bound and is strictly stronger than first-order. This result is due to Wang et al. [65], who first observed that the mle loss can have benefits over the bce loss when outcomes have low variance and used this observation to derive second-order bounds for CB and RL. We also provide lower bounds to show that the separation between these loss functions is indeed real, with one result new to this paper separating bce and mle in their ability to attain second-order bounds. The decision making performance that each loss function attains in each type of problem instance is outlined in [Table 1](#).

We then turn to extending these observations and intuitions to RL, following the results of Wang et al. [64, 65], Ayoub et al. [7]. We systematically derive bounds for both online and offline RL, revealing that the trends in [Table 1](#) continue to hold for the more challenging RL settings. Finally, we discuss issues surrounding computational complexity and provide a solution in the hybrid RL setting [58]. In summary, we provide a detailed survey on the decision performance of different loss functions, with the aim of elucidating their central role in RL and data-driven decision making generally. The technical material is largely based on Foster and Krishnamurthy [24], Wang et al. [64, 65], Ayoub et al. [7] with a couple new results along the way.

2. COST-SENSITIVE CLASSIFICATION

To best illuminate the phenomenon, we start with the simplest setting of contextual decision making: cost-sensitive classification (CSC), where learning is done offline, decisions have no impact on future contexts, and full feedback is given for all actions. To make it the simplest CSC setting, we even assume that the action space is finite (an assumption we shed in later sections). An instance of the CSC problem is then characterized by a context space \mathcal{X} , a finite number of actions A , and a distribution d on $\mathcal{X} \times [0, 1]^A$. The value of a policy $\pi : \mathcal{X} \rightarrow \{1, \dots, A\}$ is its average cost under this distribution: $V(\pi) = \mathbb{E}[c(\pi(x))]$, where $x, c(1), \dots, c(A) \sim d$. The optimal value is $V^* = \min_{\pi: \mathcal{X} \rightarrow \{1, \dots, A\}} V(\pi)$. We are given n draws of $x_i, c_i(1), \dots, c_i(A) \sim d$, sampled independently and identically distributed (i.i.d.), based on which we output a policy $\hat{\pi}$ with the aim of it having low $V(\hat{\pi})$.

Let $C : \mathcal{X} \times \mathcal{A} \rightarrow \Delta([0, 1])$ map x, a to the conditional distribution of $c(a)$ given x under d . Here $\Delta([0, 1])$ denotes the set of distributions on $[0, 1]$ that are absolutely continuous with respect to (w.r.t.) a base measure λ , such as Lesbesgue measure for continuous distributions or a counting measure for discrete distributions. We identify such distributions by its density function w.r.t. λ and we write $C(y | x, a)$ for the density of $C(x, a)$ at y . We assume that λ is common across x, a and is known. We can then write value as an expectation w.r.t. x alone:

$$V(\pi) = \mathbb{E}[\bar{C}(x, \pi(x))],$$

where the bar notation on a distribution denotes the mean: $\bar{p} = \int_y yp(y)d\lambda(y)$ for any $p \in \Delta([0, 1])$.

2.1 Solving CSC with Squared-Loss Regression

A value-based approach to CSC is to learn a cost prediction $f(x, a) \approx \bar{C}(x, a)$ by regressing costs on contexts and then use an induced greedy policy: $\pi_f(x) \in \arg \min_a f(x, a)$. A standard way to learn such a cost prediction is to minimize the squared error [5, 45, 30].

Define the squared loss and the excess squared-loss risk for a prediction function f as:

$$\ell_{\text{sq}}(\hat{y}, y) := (\hat{y} - y)^2,$$

$$\mathcal{E}_{\text{sq}}(f) := \sum_a \mathbb{E}[\ell_{\text{sq}}(f(x, a), c(a)) - \ell_{\text{sq}}(\bar{C}(x, a), c(a))].$$

This can be used to expediently bound the sub-optimality of the policy induced by f :

$$\begin{aligned} V(\pi_f) - V^* &= \mathbb{E}[\bar{C}(x, \pi_f(x)) - \bar{C}(x, \pi^*(x))] \\ &\leq \mathbb{E}[\bar{C}(x, \pi_f(x)) - f(x, \pi_f(x))] \\ &\quad + f(x, \pi^*(x)) - \bar{C}(x, \pi^*(x)) \\ (1) \quad &\lesssim \left(\sum_a \mathbb{E}(f(x, a) - \bar{C}(x, a))^2 \right)^{1/2} \\ &= (\mathcal{E}_{\text{sq}}(f))^{1/2}, \end{aligned}$$

where \lesssim means \leq up to a universal constant factor (e.g., above in [Eq. \(1\)](#), it is 2).

How do we learn a predictor with low excess squared-loss risk? We minimize the empirical squared-loss risk over a hypothesis class \mathcal{F} of functions $\mathcal{X} \times \mathcal{A} \rightarrow [0, 1]$:

$$\hat{f}_{\mathcal{F}}^{\text{sq}} \in \arg \min_{f \in \mathcal{F}} \sum_{i=1}^n \sum_{a=1}^A \ell_{\text{sq}}(f(x_i, a), c_i(a)).$$

This procedure is termed nonparametric least squares (since \mathcal{F} is general), and standard results control the excess risk of $\hat{f}_{\mathcal{F}}^{\text{sq}}$. Here we give a version for finite hypothesis classes, while for infinite classes the excess risk convergence depends on their complexity, such as given by the critical radius [61].

ASSUMPTION 1 (Realizability). $\bar{C} \in \mathcal{F}$.

Under [Asm. 1](#), for any $\delta \in (0, 1)$, with probability at least (w.p.a.l.) $1 - \delta$,

$$\mathcal{E}_{\text{sq}}(\hat{f}_{\mathcal{F}}^{\text{sq}}) \lesssim A \log(A|\mathcal{F}|/\delta)/n.$$

Together with [Eq. \(1\)](#), we obtain the following probably approximately correct (PAC) bound:

THEOREM 1. *Under [Asm. 1](#), for any $\delta \in (0, 1)$, w.p.a.l. $1 - \delta$, plug-in squared loss regression enjoys*

$$V(\pi_{\hat{f}_{\mathcal{F}}^{\text{sq}}}) - V^* \lesssim \sqrt{A \log(A|\mathcal{F}|/\delta)/n}.$$

This PAC bound shrinks at a nice parametric rate of $\mathcal{O}(n^{-1/2})$ as the number of samples n grows, but can we do better? The step in Eq. (1), which translates error in predicted means to excess risk in the squared loss, was rather loose. Foster and Krishnamurthy [24] thus investigated the bce loss and showed that it can achieve a first-order bound, which we present next after introducing an important second-order lemma.

2.2 The Second-Order Lemma

We know that estimating a mean of a random variable is easier when the random variable has smaller variance. Our next result recovers this intuition *as a completely deterministic statement about comparing bounded scalars*:

LEMMA 1 (Second-Order Mean Comparison). *Let p, q be two densities on $[0, 1]$ with respect to a common measure λ' . Then*

$$|\bar{p} - \bar{q}| \leq 6\sigma(p)h(p, q) + 8h^2(p, q),$$

where the variance and the squared Hellinger distance are defined as

$$\sigma^2(p) = \int_y y^2 p(y) d\lambda'(y) - \bar{p}^2$$

$$h^2(p, q) = \frac{1}{2} \int_y (\sqrt{p(y)} - \sqrt{q(y)})^2 d\lambda'(y).$$

Here, $h^2(p, q)$ is the squared Hellinger distance, which is an f -divergence, and it is bounded in $[0, 1]$. This lemma is equivalent to Lemma 4.3 of [65] and we provide a simplified proof in Sec. 2.6. Interpreting the inequality, which is a completely deterministic statement, in terms of estimating means, it says that estimation error can be bounded by two terms: one involves the standard deviation times a discrepancy and the other is a *squared* discrepancy. As variance shrinks, the first term vanishes and the second term dominates, which as a squared term we expect to decay quickly. We note that [26, 27] obtain related first-moment and second-moment bounds, which are generally looser and do not directly imply the variance bound in Lem. 1.

2.3 Regression with the Binary-Cross-Entropy Loss: First-Order PAC Bounds for CSC

One way to instantiate Lem. 1 is to let λ' be the counting measure on $\{0, 1\}$ and, given any $f, g \in [0, 1]$, set p, q as the Bernoulli distributions with means f, g , respectively. Bounding $f(1-f) \leq f$, this leads to

$$(2) \quad |f - g| \leq 8\sqrt{f}h_{\text{Ber}}(f, g) + 20h_{\text{Ber}}^2(f, g),$$

where $h_{\text{Ber}}^2(f, g) = \frac{1}{2}(\sqrt{f} - \sqrt{g})^2 + \frac{1}{2}(\sqrt{1-f} - \sqrt{1-g})^2$ is the squared Hellinger distance between Bernoullis with means f and g . This recovers the key inequalities in [24, 64, 7].

Replacing the bound in Eq. (1) with Eq. (2) and using Cauchy-Schwartz, we obtain

$$V(\pi_f) - V^* \lesssim \sqrt{(V(\pi_f) + V^*) \cdot \delta_{\text{Ber}}(f)} + \delta_{\text{Ber}}(f),$$

$$\text{where } \delta_{\text{Ber}}(f) := \sum_a \mathbb{E}[h_{\text{Ber}}^2(\bar{C}(x, a), f(x, a))].$$

Applying the inequality of arithmetic and geometric means (AM-GM), we see that this implies $V(\pi_f) \lesssim V^* + \delta_{\text{Ber}}(f)$. Plugging this implicit inequality back into the above, we have that

$$(3) \quad V(\pi_f) - V^* \lesssim \sqrt{V^* \cdot \delta_{\text{Ber}}(f)} + \delta_{\text{Ber}}(f).$$

Since $V^* \leq 1$, Eq. (3) also implies $V(\pi_f) - V^* \lesssim \sqrt{\delta_{\text{Ber}}(f)}$. That is, if we learn a predictor with low $\sqrt{\delta_{\text{Ber}}(f)}$, then its induced policy has correspondingly low suboptimality. However, Eq. (3) also crucially involves V^* . Thus, if the optimal policy incurs little expected costs so that the first term in Eq. (3) is negligible, we get to *square* the rate of convergence.

How do we learn a predictor with low $\delta_{\text{Ber}}(f)$? Since $\delta_{\text{Ber}}(f)$ is an average divergence between Bernoulli distributions, we could try to fit Bernoullis to the costs. Define the binary-cross-entropy (bce) loss as

$$\ell_{\text{bce}}(\hat{y}, y) := -y \ln \hat{y} - (1-y) \ln(1-\hat{y}).$$

We adopt the convention that $0 \ln 0 = 0$. Then, $\delta_{\text{Ber}}(f)$ is bounded by an exponentiated excess bce-loss risk [24].

LEMMA 2. *For any $f : \mathcal{X} \times \mathcal{A} \rightarrow [0, 1]$,*

$$\delta_{\text{Ber}}(f) \leq \mathcal{E}_{\text{bce}}(f),$$

where $\mathcal{E}_{\text{bce}}(f) := -\sum_{a=1}^A \ln \mathbb{E}[\exp(\frac{1}{2}\ell_{\text{bce}}(\bar{C}(x, a), c(a)) - \frac{1}{2}\ell_{\text{bce}}(f(x, a), c(a)))]$.

PROOF. For each a , let $z \sim \text{Ber}(c(a))$,

$$\begin{aligned} & -\ln \mathbb{E}[\exp(\frac{1}{2}\ell_{\text{bce}}(\bar{C}(x, a), c(a)) - \frac{1}{2}\ell_{\text{bce}}(f(x, a), c(a)))] \\ &= -\ln \mathbb{E}[\exp(\frac{1}{2}(c(a) \ln \frac{f(x,a)}{\bar{C}(x,a)} + (1-c(a)) \ln \frac{1-f(x,a)}{1-\bar{C}(x,a)}))] \\ &\stackrel{(i)}{\geq} -\ln \mathbb{E}[\exp(\frac{1}{2}(z \ln \frac{f(x,a)}{\bar{C}(x,a)} + (1-z) \ln \frac{1-f(x,a)}{1-\bar{C}(x,a)}))] \\ &= -\ln \mathbb{E}[\sqrt{f(x,a)\bar{C}(x,a)} + \sqrt{(1-f(x,a))(1-\bar{C}(x,a))}] \\ &\stackrel{(ii)}{\geq} 1 - \mathbb{E}[\sqrt{f(x,a)\bar{C}(x,a)} + \sqrt{(1-f(x,a))(1-\bar{C}(x,a))}] \\ &\stackrel{(iii)}{=} h_{\text{Ber}}^2(\bar{C}(x, a), f(x, a)). \end{aligned}$$

where (i) is by Jensen's inequality, (ii) is by $-\ln x \geq 1 - x$, (iii) is by completing the square. \square

To learn a predictor with low \mathcal{E}_{bce} , we may consider minimizing the empirical bce-loss risk, simply replacing ℓ_{sq} by ℓ_{bce} in nonparametric least squares:

$$\hat{f}_{\mathcal{F}}^{\text{bce}} \in \arg \min_{f \in \mathcal{F}} \sum_{i=1}^n \sum_{a=1}^A \ell_{\text{bce}}(f(x_i, a), c_i(a)).$$

The bce loss $\ell_{\text{bce}}(\hat{y}, y)$ is exactly the negative log-likelihood of observing y from a Bernoulli distribution with mean \hat{y} . Nevertheless, even if y is *not binary*, it can be used as a general-purpose surrogate loss for regression (sometimes under the moniker "log loss" [24, 7]). In particular, for any density $p \in \Delta([0, 1])$, the mean \bar{p} minimizes expected bce loss:

$$\mathbb{E}_{y \sim p}[\ell_{\text{bce}}(f, y) - \ell_{\text{bce}}(\bar{p}, y)] \geq 2(f - \bar{p})^2.$$

This inequality also means that we could use the excess bce-loss risk to bound Eq. (1) in place of excess squared-loss risk. The point of using bce loss, however, is to do better than Eq. (1) via Eq. (2).

For the final part of the proof, we need to show that minimizing the empirical bce-loss risk gives good control on $\mathcal{E}_{\text{bce}}(\hat{f}_{\mathcal{F}}^{\text{bce}})$. This is implied by the following tail bound.

LEMMA 3. *Let Z_1, \dots, Z_n denote n i.i.d. random variables. For any $\delta \in (0, 1)$, w.p.a.l. $1 - \delta$,*

$$-n \ln \mathbb{E}[\exp(-Z_1)] \leq \sum_{i=1}^n Z_i + \ln(1/\delta).$$

PROOF. We note $\mathbb{E}[\exp(\sum_{i=1}^n Z_i)] = (\mathbb{E}[\exp(Z_1)])^n$. By Chernoff's method, $\Pr(\sum_{i=1}^n Z_i - n \ln \mathbb{E}_{Z_1} \exp(Z_1) \geq t) \leq \exp(-t)$ for all $t > 0$. Finally, set $t = \ln(1/\delta)$. \square

Then, applying Lem. 3 with $Z_i = \frac{1}{2}\ell_{\text{bce}}(f(x_i, a), c_i(a)) - \frac{1}{2}\ell_{\text{bce}}(\bar{C}(x_i, a), c_i(a))$ and a union bound over a and f , we have w.p.a.l. $1 - \delta$, for all $f \in \mathcal{F}$

$$\begin{aligned} n\mathcal{E}_{\text{bce}}(f) &\leq \frac{1}{2} \sum_{i=1}^n \sum_{a=1}^A \ell_{\text{bce}}(f(x_i, a), c_i(a)) \\ &\quad - \ell_{\text{bce}}(\bar{C}(x_i, a), c_i(a)) + A \ln(2A|\mathcal{F}|/\delta). \end{aligned}$$

With Asm. 1, the empirical minimizer $\hat{f}_{\mathcal{F}}^{\text{bce}}$ enjoys

$$\mathcal{E}_{\text{bce}}(\hat{f}_{\mathcal{F}}^{\text{bce}}) \leq \frac{A \ln(A|\mathcal{F}|/\delta)}{n}.$$

Thus, together with Eq. (3) and Lem. 2, we have shown a first-order PAC bound for bce-loss regression:

THEOREM 2. *Under Asm. 1, for any $\delta \in (0, 1)$, w.p.a.l. $1 - \delta$, plug-in bce loss regression enjoys*

$$V(\pi_{\hat{f}_{\mathcal{F}}^{\text{bce}}}) - V^* \lesssim \sqrt{V^* \cdot \frac{A \ln(A|\mathcal{F}|/\delta)}{n}} + \frac{A \ln(A|\mathcal{F}|/\delta)}{n}.$$

This result was first observed in Theorem 3 of [24]. Notably, the bound is *adaptive* to the optimal expected costs V^* and converges at a fast n^{-1} rate when $V^* \lesssim 1/n$. Under the cost minimization setup, first-order bounds are also called 'small-cost' bounds since they converge at a fast rate when the optimal cost V^* is small.

REMARK 1. *A refinement of Eq. (2) keeps the first term as $\sqrt{f(1-f)}h_{\text{Ber}}$ instead of $\sqrt{f}h_{\text{Ber}}$. This would imply a more refined first-order bound that scales as*

$\tilde{\mathcal{O}}(\sqrt{V^*(1 - V^*) \cdot \frac{1}{n}} + \frac{1}{n})$, where the leading term vanishes also if $V^* \approx 1$. Bounds scaling with $1 - V^*$ are sometimes called 'small-reward' bounds [6], which are generally easier to prove than 'small-cost' bounds [46, 64]. For example, if the bound scales with the minimum achievable reward R^* , the case where $R^* = 0$ is actually trivial since all policies are optimal. However, if the bound scales with the minimum achievable cost L^* , the case where $L^* = 0$ is interesting because sub-optimal policies may still have large cost. Thus, following Lykouris, Sridharan and Tardos [46], Wang et al. [64], we focus on the cost minimizing setting in this paper.

2.4 Maximum Likelihood Estimation: Second-Order PAC Bounds for CSC

Can we do even better than a first-order PAC bound with the bce loss? In this section we show that a second-order, variance-adaptive bound is possible if we learn the conditional cost distribution instead of only regressing the mean. To learn the distribution, we use a hypothesis class \mathcal{P} of conditional distributions $\mathcal{X} \times \mathcal{A} \rightarrow \Delta([0, 1])$ and minimize the negative-log likelihood loss from maximum likelihood estimation (mle): for a density $\hat{p} \in \Delta([0, 1])$ and target $y \in [0, 1]$, define

$$\ell_{\text{mle}}(\hat{p}, y) := -\ln \hat{p}(y).$$

Unlike the previous sections where the loss measured the discrepancy of a point prediction, the mle loss measures the discrepancy of a distributional prediction. Indeed, if $\hat{p} = \text{Ber}(\hat{y})$ and $p = \text{Ber}(y)$, then $\mathbb{E}_{y \sim p}[\ell_{\text{mle}}(\hat{p}, y)] = \ell_{\text{bce}}(\hat{y}, y)$ so the bce loss can be viewed as a Bernoulli specialization of the general mle loss. This generality allows us to directly apply Lem. 1 in place of Eq. (1) to obtain for any $p \in \mathcal{P}$:

(4)

$$V(\pi_{\bar{p}}) - V^* \lesssim \sqrt{(\sigma^2(\pi_{\bar{p}}) + \sigma^2(\pi^*)) \delta_{\text{dis}}(p)} + \delta_{\text{dis}}(p),$$

where $\delta_{\text{dis}}(p) := \sum_a \mathbb{E}[h^2(C(x, a), p(x, a))]$ and $\sigma^2(\pi) := \sigma^2(\bar{C}(x, \pi(x)))$. As in the bce section, we then upper bound $\delta_{\text{dis}}(f)$ by an exponentiated excess mle-loss risk.

LEMMA 4. *For any $f : \mathcal{X} \times \mathcal{A} \rightarrow \Delta([0, 1])$,*

$$\delta_{\text{dis}}(p) \leq \mathcal{E}_{\text{mle}}(p),$$

where $\mathcal{E}_{\text{mle}}(p) := -\sum_{a=1}^A \ln \mathbb{E}[\exp(\frac{1}{2}\ell_{\text{mle}}(C(x, a), c(a)) - \frac{1}{2}\ell_{\text{mle}}(p(x, a), c(a)))]$.

The proof is almost identical to that of Lem. 2, and is even simpler since the inequality marked "(i)" in the proof is not needed. To learn a predictor with low \mathcal{E}_{mle} , we minimize the empirical negative log-likelihood risk:

$$\hat{p}_{\mathcal{P}}^{\text{mle}} \in \arg \min_{p \in \mathcal{P}} L_{\text{mle}}(p),$$

$$\text{where } L_{\text{mle}}(p) := \sum_{i=1}^n \sum_{a=1}^A \ell_{\text{mle}}(p(x_i, a), c_i(a)).$$

We also posit realizability in the distribution class.

ASSUMPTION 2 (Distributional Realizability). $C \in \mathcal{P}$.

While this assumption is more stringent than mean-realizability ([Asm. 1](#) from before), it is required to prove that MLE succeeds [20, 3]. Alternative algorithms such as Scheffé tournament can be used to learn distributions under misspecification, but they are computationally hard to implement [2].

Finally, we apply [Lem. 3](#) with $Z_i = \frac{1}{2}\ell_{\text{mle}}(p(x_i, a), c_i(a)) - \frac{1}{2}\ell_{\text{mle}}(C(x_i, a), c_i(a))$ with union bound over \mathcal{P} , to deduce that w.p.a.l. $1 - \delta$, for all $p \in \mathcal{P}$:

$$(5) \quad n\mathcal{E}_{\text{mle}}(p) \leq \frac{1}{2}L_{\text{mle}}(p) - \frac{1}{2}L_{\text{mle}}(C) + A \ln(A|\mathcal{P}|/\delta).$$

Together with [Asm. 2](#), we have that

$$\mathcal{E}_{\text{mle}}(\hat{p}_{\mathcal{P}}^{\text{mle}}) \leq \frac{A \ln(A|\mathcal{P}|/\delta)}{n}.$$

Thus we have proven a second-order PAC bound for the greedy policy $\hat{\pi}^{\text{mle}} := \pi_{\hat{p}_{\mathcal{P}}^{\text{mle}}}$.

THEOREM 3. *Under [Asm. 2](#), for any $\delta \in (0, 1)$, w.p.a.l. $1 - \delta$, plug-in mle enjoys*

$$V(\hat{\pi}^{\text{mle}}) - V^* \lesssim \sqrt{(\sigma^2(\hat{\pi}^{\text{mle}}) + \sigma^2(\pi^*)) \cdot \frac{A \ln(A|\mathcal{P}|/\delta)}{n}} + \frac{A \ln(A|\mathcal{P}|/\delta)}{n}.$$

Since costs are bounded in $[0, 1]$, we observe that $\sigma^2(\pi) \leq V(\pi)$, and hence a second-order bound is tighter than a first-order bound. However, our approach is distributional with bounds depending on $\ln |\mathcal{P}|$ which can be larger than $\ln |\mathcal{F}|$, and as mentioned earlier, realizability in distribution space is stricter than realizability in the conditional mean. We remark that distributional approaches still achieve superior performance [10, 11, 64, 65] in practice, suggesting that MLE with modern function approximators can well-approximate complex distributions despite the stronger assumption in theory.

2.5 Improved Second-Order PAC Bounds for CSC with Pessimistic MLE

We can derive even tighter bounds if the distribution is learned in a pessimistic manner – that is, the mean of the learned distribution *upper bounds* the true optimal mean V^* with high probability.¹ In this section, we introduce how to achieve pessimism by optimizing over a subset of the function class defined by empirical losses, an approach that is often termed ‘version space’ [25]. This is also important warmup for the optimistic and pessimistic RL algorithms that we consider in the sequel.

¹Here, we say the learned mean is pessimistic if it upper bounds V^* since we’re in the cost minimization setting. Under the reward maximization setting, pessimism would be to lower bound V^* .

We start by defining a subclass of near-optimal distributions w.r.t. the empirical mle loss

$$\mathcal{P}_n := \{p \in \mathcal{P} : L_{\text{mle}}(p) - L_{\text{mle}}(\hat{p}_{\mathcal{P}}^{\text{mle}}) \leq \beta\},$$

where β is a parameter that will be set appropriately. Then, a pessimistic distribution is learnt by selecting the element with the lowest value. The following lemma defines this formally and proves that the learned distribution (a) has low excess risk and (b) is nearly pessimistic.

LEMMA 5. *Under [Asm. 2](#), for any $\delta \in (0, 1)$, set $\beta = 2A \ln(A|\mathcal{P}|/\delta)$ and define,*

$$(6) \quad \hat{p}^{\text{pes}} \in \arg \max_{p \in \mathcal{P}_n} \sum_{i=1}^n \min_a \bar{p}(x_i, a).$$

Then, w.p.a.l. $1 - \delta$, (a) $\mathcal{E}_{\text{mle}}(\hat{p}^{\text{pes}}) \lesssim \frac{A \ln(A|\mathcal{P}|/\delta)}{n}$, and (b) $V(\hat{\pi}^{\text{pes}}) - \mathbb{E}[\min_a \bar{p}^{\text{pes}}(x, a)] \lesssim \frac{\ln(A|\mathcal{P}|/\delta)}{n}$.

PROOF OF LEM. 5. For both claims, we condition on [Eq. \(5\)](#) which holds w.p.a.l. $1 - \delta$. For Claim (a): for any $p \in \mathcal{P}_n$ (which includes \hat{p}^{pes}), we have

$$\begin{aligned} n\mathcal{E}_{\text{mle}}(p) &\leq \frac{1}{2}L_{\text{mle}}(p) - \frac{1}{2}L_{\text{mle}}(\hat{p}_{\mathcal{P}}^{\text{mle}}) + A \ln(A|\mathcal{P}|/\delta) \\ &\leq \frac{1}{2}\beta + A \ln(A|\mathcal{P}|/\delta) \leq 2A \ln(A|\mathcal{P}|/\delta), \end{aligned}$$

where the first inequality is by [Eq. \(5\)](#) and the fact that $\hat{p}_{\mathcal{P}}^{\text{mle}}$ minimizes the empirical risk; and the second inequality is by the definition of \mathcal{P}_n . To prove Claim (b), we first show that $C \in \mathcal{P}_n$: by [Eq. \(5\)](#) and the non-negativity of \mathcal{E}_{mle} , we have $L_{\text{mle}}(C) - L_{\text{mle}}(p) \leq 2A \ln(A|\mathcal{P}|/\delta) = \beta$ for all $p \in \mathcal{P}$ (which includes $\hat{p}_{\mathcal{P}}^{\text{mle}}$). Thus, this shows that C satisfies the \mathcal{P}_n condition, implying its membership in the set. To conclude Claim (b), we have $\sum_{i=1}^n \bar{p}^{\text{pes}}(x_i, \pi^*(x_i)) \geq \sum_{i=1}^n \min_a \bar{p}^{\text{pes}}(x_i, a) \geq \sum_{i=1}^n \min_a \bar{C}(x_i, a)$. Claim (b) then follows by multiplicative Chernoff [75, Theorem 13.5]. \square

With pessimism, the induced policy $\hat{\pi}^{\text{pes}} := \pi_{\hat{p}^{\text{pes}}}$ only suffers one of the terms before [Eq. \(1\)](#), and so

$$\begin{aligned} (7) \quad V(\hat{\pi}^{\text{pes}}) - V^* &\leq \mathbb{E}[\bar{p}^{\text{pes}}(x, \pi^*(x)) - \bar{C}(x, \pi^*(x))] \\ &\lesssim \sqrt{\sigma^2(\pi^*) \cdot \delta_{\text{dis}}(\hat{p}^{\text{pes}})} + \delta_{\text{dis}}(\hat{p}^{\text{pes}}) \\ &\lesssim \sqrt{\sigma^2(\pi^*) \cdot \frac{A \ln(A|\mathcal{P}|/\delta)}{n}} + \frac{A \ln(A|\mathcal{P}|/\delta)}{n}. \end{aligned}$$

Thus, we have proven an *improved* second-order PAC bound for pessimistic mle.

THEOREM 4. *Under [Asm. 2](#), for any $\delta \in (0, 1)$, w.p.a.l. $1 - \delta$, pessimistic mle enjoys*

$$V(\hat{\pi}^{\text{pes}}) - V^* \lesssim \sqrt{\sigma^2(\pi^*) \cdot \frac{A \ln(A|\mathcal{P}|/\delta)}{n}} + \frac{A \ln(A|\mathcal{P}|/\delta)}{n}.$$

Notably, Eq. (7) is an improvement to Eq. (4) since it only contains the variance of the optimal policy π^* , which is a fixed quantity, and not that of the learned policy, which is a random algorithm-dependent quantity. We remark that while pessimism is typically used to solve problems with poor coverage, e.g., offline RL, we see it also plays a crucial role in obtaining finer second-order PAC bounds in CSC, which has full coverage due to complete feedback.

Pessimism could have also been applied with the bce-loss, but there would have been no improvement to the first-order bound. This is because $V(\hat{\pi}^{\text{bce}}) - V^* \leq \sqrt{(V(\hat{\pi}^{\text{bce}}) + V^*) \cdot \frac{C}{n}} + \frac{C}{n}$ already implies $V(\hat{\pi}^{\text{bce}}) - V^* \leq \sqrt{V^* \cdot \frac{C'}{n}} + \frac{C'}{n}$ where $\frac{C'}{C}$ is a universal constant, due to the AM-GM inequality as noted in the text preceding Eq. (3). However, this implicit inequality does not hold for variance-based inequalities, and so pessimism is crucial for removing the dependence on the learned policy's variance.

Finally, we note that pessimistic mle requires more computation than plug-in mle, since we have the extra step of optimizing over \mathcal{P}_n . For one-step settings like CSC or contextual bandits, this can be feasibly implemented with binary search [25] or width computation [41, 22]. However, in multi-step settings like RL as we will soon see, this optimization problem is NP-hard [18].

REMARK 2 (Another improved bound via optimism). We could also consider optimistic mle where $\hat{p}^{\text{op}} \in \arg \min_{p \in \mathcal{P}_n} \sum_{i=1}^n \min_a \bar{p}(x_i, a)$ and $\hat{\pi}^{\text{op}} = \pi_{\hat{p}^{\text{op}}}$. Then, the decomposition of Eq. (7) would look like:

$$\begin{aligned} V(\hat{\pi}^{\text{op}}) - V^* &\leq \mathbb{E}[\bar{C}(x, \hat{\pi}^{\text{op}}(x)) - \overline{\hat{p}^{\text{op}}}(x, \hat{\pi}^{\text{op}}(x))] \\ &\lesssim \sqrt{\sigma^2(\hat{\pi}^{\text{op}}) \cdot \delta_{\text{dis}}(\hat{p}^{\text{op}})} + \delta_{\text{dis}}(\hat{p}^{\text{op}}) \\ &\lesssim \sqrt{\sigma^2(\hat{\pi}^{\text{op}}) \cdot \frac{A \ln(A|\mathcal{P}|/\delta)}{n}} + \frac{A \ln(A|\mathcal{P}|/\delta)}{n}. \end{aligned}$$

This is also an improved second-order PAC bound, which depends only on the variance of the learned policy and not that of π^* . The bound in Thm. 4 may be preferred since $\sigma(\pi^*)$ is a fixed quantity; however, we note that $\sigma^2(\hat{\pi})$ and $\sigma^2(\pi^*)$ are not comparable in general, so neither bound dominates the other.

2.6 Proof of the Second-Order Lemma

The goal of this subsection is to prove the second-order lemma (Lem. 1), a key tool to derive first- and second-order PAC bounds. We prove the result in terms of another divergence called the triangular discrimination: for any densities p, q on $[0, 1]$ w.r.t. a common measure λ' , the triangular discrimination is defined as

$$(8) \quad \Delta(p, q) := \int_y \frac{(p(y) - q(y))^2}{p(y) + q(y)} d\lambda'(y).$$

$\Delta(\cdot)$ is a symmetric f -divergence and is equivalent to the squared Hellinger distance up to universal constants:

$$(9) \quad 2h^2(p, q) \leq \Delta(p, q) \leq 4h^2(p, q),$$

which is a simple consequence of Cauchy-Schwartz [64, Lemma A.1]. Thus, we first prove the second-order lemma using $\Delta(\cdot)$, which is more natural, and then convert the bounds to $h^2(\cdot)$ using Eq. (9).

LEMMA 6. Let p, q be densities on $[0, 1]$ and let q be the one with smaller variance. Then,

$$(10) \quad \sigma^2(p) - \sigma^2(q) \leq 2\sqrt{\sigma^2(q) \cdot \Delta(p, q)} + \Delta(p, q),$$

$$(11) \quad |\bar{p} - \bar{q}| \leq 3\sqrt{\sigma^2(q) \cdot \Delta(p, q)} + 2\Delta(p, q).$$

PROOF OF LEM. 6. We first prove Eq. (10):

$$\begin{aligned} &\sigma^2(p) - \sigma^2(q) \\ &\stackrel{i}{\leq} \int_y (y - \bar{q})^2 (p(y) - q(y)) d\lambda'(y) \\ &\stackrel{ii}{\leq} \sqrt{\int_y (y - \bar{q})^4 (p(y) + q(y)) d\lambda'(y) \cdot \Delta(p, q)} \\ &\stackrel{iii}{\leq} \sqrt{\int_y (y - \bar{q})^2 (p(y) + q(y)) d\lambda'(y) \cdot \Delta(p, q)} \\ &\stackrel{iv}{\leq} \frac{1}{2} \left(\int_y (y - \bar{q})^2 p(y) d\lambda'(y) + \sigma^2(q) \right) + \frac{\Delta(p, q)}{2}, \end{aligned}$$

where (i) is by the definition of variance, (ii) is by Cauchy-Schwarz, (iii) is by the premise that p, q are densities on $[0, 1]$, and (iv) is by AM-GM. Rearranging the inequality that (i) \leq (iv), we get

$$\int_y (y - \bar{q})^2 p(y) d\lambda'(y) \leq 3\sigma^2(q) + \Delta(p, q).$$

Finally, plugging this back into (iii) implies Eq. (10).

Now we prove Eq. (11). Set $c = \frac{\bar{p} + \bar{q}}{2}$. First, consider the case that $D_\Delta(p, q) \leq 1$:

$$\begin{aligned} |\bar{p} - \bar{q}|^2 &= \left| \int_y (p(y) - q(y))(y - c) d\lambda'(y) \right|^2 \\ &\stackrel{i}{\leq} \int_y (p(y) + q(y))(y - c)^2 d\lambda'(y) \cdot \Delta(p, q) \\ &\stackrel{ii}{=} \left(\sigma^2(p) + \sigma^2(q) + 2\left(\frac{\bar{p} - \bar{q}}{2}\right)^2 \right) \Delta(p, q) \\ &\stackrel{iii}{\leq} (\sigma^2(p) + \sigma^2(q)) \Delta(p, q) + \frac{(\bar{p} - \bar{q})^2}{2} \end{aligned}$$

where (i) is by Cauchy-Schwarz, (ii) is by expanding the variance $\sigma^2(f) = \int_y (f(y) - c)^2 d\lambda'(y) - (\bar{f} - c)^2$ which holds for any $c \in \mathbb{R}$, and (iii) is by $\Delta(p, q) \leq 1$. Rearranging terms and using Eq. (10), we get

$$\begin{aligned} |\bar{p} - \bar{q}| &\leq \sqrt{2(\sigma^2(p) + \sigma^2(q)) \Delta(p, q)} \\ &\leq \sqrt{2(3\sigma^2(q) + 2\Delta(p, q)) \Delta(p, q)} \\ &\leq 3\sqrt{\sigma^2(q) \Delta(p, q)} + 2\Delta(p, q). \end{aligned}$$

This finishes the case of $\Delta(p, q) \leq 1$. Otherwise, we simply have $|\bar{p} - \bar{q}| \leq 1 < \Delta(p, q)$. \square

3. LOWER BOUNDS FOR CSC

So far, we have seen that plug-in regression with the squared loss, bce loss and mle loss have progressively more adaptive PAC bounds for CSC. A natural question is if the previous bounds were tight: is it necessary to change the loss function if we want to achieve these sharper and more adaptive bounds? In this section, we answer this in the affirmative by exhibiting counterexamples.

3.1 Plug-in Squared Loss Cannot Achieve First-Order

First, we show that the policy induced by squared loss regression cannot achieve first-order bounds. The following counterexample is due to [24], where we have simplified the presentation and improved constants.

THEOREM 5. *For all $n > 400$, there exists a CSC problem with $|\mathcal{A}| = |\mathcal{X}| = 2$ and a realizable function class with $|\mathcal{F}| = 2$ such that: (a) $V^* \leq \frac{1}{n}$, but (b) $V(\pi_{\hat{f}_{\mathcal{F}}^{\text{sq}}}) - V^* \geq \frac{1}{32\sqrt{n}}$ w.p.a.l. 0.1.*

The intuition is that squared loss regression does not adapt to context-dependent variance, *a.k.a.* heteroskedasticity; so the convergence of squared loss regression is dominated by the worst context's variance. In this counterexample, the second context x^2 occurs with tiny probability n^{-1} but has high variance; however, the empirical squared loss is dominated by this unlikely context.

PROOF OF THM. 5. The structure of the proof is the following: for any $n > 400$, we first construct the CSC problem and realizable function class, and then show that indeed $V^* \leq \mathcal{O}(\frac{1}{n})$. Next, we show that under a bad event which occurs with probability at least 0.1, the function with the lowest empirical squared risk induces a policy that suffers regret which is lower bounded by $\Omega(\frac{1}{\sqrt{n}})$.

Fix any $n > 400$. We begin by setting up the CSC problem: label the two states as x^1, x^2 and the two actions as a^1, a^2 . Set the data generating distribution d as follows: $d(x^1) = 1 - n^{-1}$, $d(x^2) = n^{-1}$ and

$$\begin{aligned} c(a^1) | x^1 &\sim \text{Ber}(\mu_n), & c(a^2) | x^1 &= \nu_n, \\ c(a^1) | x^2 &\sim \text{Ber}(\frac{1}{2}), & c(a^2) | x^2 &= \frac{1}{2}, \end{aligned}$$

where $\mu_n = \frac{1}{8n}$, $\nu_n = \frac{1}{8\sqrt{n}}$. Our realizable function class \mathcal{F} contains two elements: the true $f^*(x, a) = \mathbb{E}[c(a) | x]$ and another function \tilde{f} defined as

$$\begin{aligned} \tilde{f}(x^1, a^1) &= \varepsilon_n, & \tilde{f}(x^1, a^2) &= \nu_n, \\ \tilde{f}(x^2, a^1) &= 0, & \tilde{f}(x^2, a^2) &= \frac{1}{2}, \end{aligned}$$

where $\varepsilon_n = \frac{1}{4\sqrt{n}}$. Note that $\mu_n < \nu_n$ so $\pi^*(x^1) = a^1$ but $\varepsilon_n > \nu_n$ so $\pi_{\tilde{f}}(x^1) = a^2$, *i.e.*, $\pi_{\tilde{f}}$ makes a mistake on x^1 . Also, $V^* = (1 - n^{-1})\mu_n + \frac{1}{2n} \leq \frac{1}{n}$.

Now, we compute the empirical squared-loss risk and show that $\hat{f}_{\mathcal{F}}^{\text{sq}} = \tilde{f}$ under a bad event. The empirical risk can be simplified by shedding shared terms to be:

$$\hat{L}_{\text{sq}}(f) = \sum_{i \in [2]} \frac{n(x^i)}{n} (f(x^i, a^1) - \hat{\mu}(x^i, a^1))^2,$$

where $n(x)$ denotes the number of times x occurs in the dataset and $\hat{\mu}(x, a) = \frac{1}{n(x)} \sum_{i: x_i=x} c_i(a)$ is the empirical conditional mean. We split the bad event into two parts: (\mathfrak{E}_1) x^2 appears only once in the dataset (*i.e.*, $n(x^2) = 1$) and its observed cost at a^1 is 0 (*i.e.*, $\hat{\mu}(x^2, a^1) = 0$); and (\mathfrak{E}_2) $\hat{\mu}(x^1, a^1) \leq 2\mu_n + \frac{3}{n-1}$. We lower bound $\Pr(\mathfrak{E}_1 \cap \mathfrak{E}_2) \geq 0.1$ at the end.

We now show that $\hat{f}_{\mathcal{F}}^{\text{sq}} = \tilde{f}$ under $\mathfrak{E}_1 \cap \mathfrak{E}_2$. Under \mathfrak{E}_1 , we lower bound $\hat{L}_{\text{sq}}(f^*)$ by:

$$\frac{n(x^2)}{n} (f^*(x^2, a^1) - \hat{\mu}(x^2, a^1))^2 = \frac{1}{4n}.$$

Under $\mathfrak{E}_1 \cap \mathfrak{E}_2$, the x^2 term of $\hat{L}_{\text{sq}}(\tilde{f})$ vanishes since $\tilde{f}(x^2, a^1) = \hat{\mu}(x^2, a^1)$, and so $\hat{L}_{\text{sq}}(\tilde{f})$ can be bounded by:

$$(\tilde{f}(x^1, a^1) - \hat{\mu}(x^1, a^1))^2 \leq 2\varepsilon_n^2 + 2(2\mu_n + \frac{3}{n-1})^2 < \frac{1}{4n},$$

where the last inequality holds due to $n > 400$. Thus, squared loss regression selects $\hat{f}_{\mathcal{F}}^{\text{sq}} = \tilde{f}$ and the regret of the induced policy can be lower bounded by:

$$V(\pi_{\hat{f}_{\mathcal{F}}^{\text{sq}}}) - V^* = \frac{n-1}{n} (\nu_n - \mu_n) \geq \frac{1}{16} (\frac{1}{\sqrt{n}} - \frac{1}{n}) \geq \frac{1}{32\sqrt{n}}.$$

Probability of the bad event. For \mathfrak{E}_1 , since $n(x^2) \sim \text{Bin}(n, n^{-1})$, thus $\Pr(n(x^2) = 1) = (1 - n^{-1})^{n-1} \geq e^{-1}$. Hence, $\Pr(\mathfrak{E}_1) \geq (2e)^{-1}$. For \mathfrak{E}_2 , we apply the multiplicative Chernoff bound [75, Theorem 13.5], which implies $\hat{\mu}(x^1, a^1) < 2\mu_n + \frac{3}{n-1}$ w.p.a.l. $1 - e^{-3}$. Thus, $\Pr(\mathfrak{E}_1 \cap \mathfrak{E}_2) \geq 1 - (1 - (2e)^{-1}) - e^{-3} \geq 0.1$. \square

3.2 Plug-in BCE Loss Cannot Achieve Second-Order

Next, we show that the bce-loss induced policy cannot achieve second-order bounds. This is a new result that did not appear before.

THEOREM 6. *For all odd $n \in \mathbb{N}$, there exists a CSC problem where $|\mathcal{A}| = 2, |\mathcal{X}| = 1$ and a realizable function class with $|\mathcal{F}| = 2$ such that: (a) $\sigma^2(\pi^*) = 0$, but (b) $V(\pi_{\hat{f}_{\mathcal{F}}^{\text{bce}}}) - V^* \geq \frac{1}{8\sqrt{n}}$ w.p.a.l. $\frac{1}{4}$.*

PROOF OF THM. 6. The proof structure is similar as before: for any odd n , we construct the CSC problem and a realizable function class. We show that $\sigma^2(\pi^*) = 0$ which is the second-order regime; we also sanity check that V^* is bounded away from 0 and 1, to ensure that we're not in the first-order regime. Next, we show that under a bad event which occurs with constant probability, the function with the lowest empirical bce risk induces a policy that suffers regret which is lower bounded by $\Omega(\frac{1}{\sqrt{n}})$.

Fix any odd $n \in \mathbb{N}$. We first construct the CSC problem: label the two actions as a^1, a^2 and drop the context notation since there is one context. Set the data generating distribution d such that: $c(a^1) \sim \text{Ber}(\frac{1}{2} + \varepsilon_n)$ and $c(a^2) = \frac{1}{2}$ w.p. 1. The true conditional means are $f^*(a^1) = \frac{1}{2} + \varepsilon_n$ and $f^*(a^2) = \frac{1}{2}$. In addition to f^* , the function class \mathcal{F} only contains one other function \tilde{f} defined as $\tilde{f}(a^1) = \tilde{f}(a^2) = \frac{1}{2}$. Note that the optimal action is $a^* = a^2$ and the regret of a^1 is $f^*(a^1) - f^*(a^2) = \varepsilon_n$. We also check that $V^* = \Theta(1)$, and so this is not the first-order regime.

Now, we compute the empirical bce-loss risk and show that $\hat{f}_{\mathcal{F}}^{\text{bce}} = \tilde{f}$ under a bad event. Since all elements of \mathcal{F} have the same prediction for a^2 , the empirical bce-loss risk can be simplified to

$$\begin{aligned}\hat{L}_{\text{bce}}(f) &= p \cdot \ell_{\text{bce}}(f(a^1), 0) + (1 - p) \cdot \ell_{\text{bce}}(f(a^1), 1) \\ &= \ell_{\text{bce}}(f(a^1), 1 - p),\end{aligned}$$

where p is the fraction of times that $c(a^1) = 0$ in the dataset. The above loss is convex and its minimizer is $1 - p$. The bad event we consider is that $p > \frac{1}{2}$, under which we have $1 - p < \tilde{f}(a^1) < f^*(a^1)$; since the loss is convex, \tilde{f} indeed achieves lower loss than f^* . Thus, we have that $\hat{f}_{\mathcal{F}}^{\text{bce}} = \tilde{f}$ and the regret of the induced policy is $V(\pi_{\hat{f}_{\mathcal{F}}^{\text{bce}}}) - V^* = f^*(a^1) - f^*(a^2) = \frac{1}{8\sqrt{n}}$.

Probability of the bad event. Kontorovich [40] proved tight lower and upper bounds for binomial small deviations and we will make use of the following result: for all $n \geq 1$ and $\gamma \in [0, \frac{1}{\sqrt{n}}]$, let $\Pr(\text{Bin}(n, \frac{1}{2} - \frac{\gamma}{2}) \leq \lfloor \frac{n}{2} \rfloor) - \Pr(\text{Bin}(n, \frac{1}{2} + \frac{\gamma}{2}) \leq \lfloor \frac{n}{2} \rfloor) \leq \sqrt{n}\gamma$. If n is odd, we have that $\frac{1}{2} = \Pr(\text{Bin}(n, \frac{1}{2}) \leq \lfloor \frac{n}{2} \rfloor) < \Pr(\text{Bin}(n, \frac{1}{2} - \frac{\gamma}{2}) \leq \lfloor \frac{n}{2} \rfloor)$. Thus, $\Pr(\text{Bin}(n, \frac{1}{2} + \frac{\gamma}{2}) < \frac{n}{2}) \geq \frac{1}{2} - \sqrt{n}\gamma$. Setting $\gamma = \frac{1}{4\sqrt{n}}$ (corresponding to $\varepsilon_n = \frac{1}{8\sqrt{n}}$), we have shown that the bad event occurs with probability at least $\frac{1}{4}$. \square

In the above proof, we used two key properties of the empirical bce risk: (1) its minimizer is the empirical mean, and (2) it is convex w.r.t. the prediction (*i.e.*, the first argument). Since squared loss also has these properties, the above result also applies to squared regression. However, since the mle loss learns a distribution rather than just the mean, the counterexample does not apply. Finally, since CSC is the most basic decision making setting, the counterexamples in this section also apply to reinforcement learning via the online-to-batch conversion.

4. SETUP FOR REINFORCEMENT LEARNING

In the preceding sections, we saw how the loss function plays a central role in the sample efficiency of algorithms for CSC, the simplest decision making problem. The commonly used squared loss results in slow $\Theta(1/\sqrt{n})$ rates in benign problem instances where the optimal policy has

small cost (*i.e.*, first-order) or has small variance (*i.e.*, second-order), while the bce or mle losses, respectively, can be used to achieve fast $\mathcal{O}(1/n)$ rates.

In the rest of this paper, we will see that these observations and insights generally transfer to more complex decision making setups, in particular reinforcement learning (RL). Compared to the CSC setting, two new challenges of RL are that (1) the learner receives feedback only for the chosen action (*a.k.a.*, partial or bandit feedback) and (2) the learner sequentially interacts with the environment over multiple time steps. As before, we focus on value-based algorithms with function approximation and prove bounds for problems with high-dimensional observations, *i.e.*, beyond the finite tabular setting.

4.1 Problem Setup

We formalize the RL environment as a Markov Decision Process (MDP) which consists of an observation space \mathcal{X} , action space \mathcal{A} , horizon H , transition kernels $\{P_h : \mathcal{X} \times \mathcal{A} \rightarrow \Delta(\mathcal{X})\}_{h \in [H]}$ and conditional cost distributions $\{C_h : \mathcal{X} \times \mathcal{A} \rightarrow \Delta([0, 1])\}_{h \in [H]}$. We formalize the policy as a tuple of mappings $\pi = \{\pi_h : \mathcal{X} \rightarrow \Delta(\mathcal{A})\}_{h \in [H]}$ that interacts (*a.k.a.* rolls-in) with the MDP as follows: start from an initial state x_1 and at each step $h = 1, 2, \dots, H$, sample an action $a_h \sim \pi_h(x_h)$, collect a cost $c_h \sim C_h(x_h, a_h)$ and transit to the next state $x_{h+1} \sim P_h(x_h, a_h)$. We use $Z^\pi = \sum_{h=1}^H c_h$ to denote the cumulative cost, a random variable, from rolling in π ; we consider the general setup where Z^π is normalized between $[0, 1]$ almost surely which allows for sparse rewards [34]. We use $Z_h^\pi(x_h, a_h) = \sum_{t=h}^H c_t$ to denote the cumulative cost of rolling in π from x_h, a_h at step h . We use $Q_h^\pi(x_h, a_h) = \mathbb{E}[Z_h^\pi(x_h, a_h)]$ and $V_h^\pi(x_h) = Q_h^\pi(x_h, \pi)$ to denote the expected cumulative costs, where we use the shorthand $f(x, \pi) = \mathbb{E}_{a \sim \pi(x)} f(x, a)$ for any f . For simplicity, we assume the initial state x_1 is fixed and known, and we let $V^\pi := V_1^\pi(x_1)$ denote the initial state value of π . Our results can be extended to the case when x_1 is stochastic from an unknown distribution, or, in the online setting, the initial state at round k may even be chosen by an adaptive adversary.

Online RL. The learner aims to compete against the optimal policy denoted as $\pi^* = \arg \min_\pi V_1^\pi(x_1)$. We use Z^*, V^*, Q^* to denote $Z^{\pi^*}, V^{\pi^*}, Q^{\pi^*}$, respectively. The online RL problem iterates over K rounds: for each round $k = 1, 2, \dots, K$, the learner selects a policy π^k to roll-in and collect data, and the goal is to minimize regret,

$$(12) \quad \text{Reg}_{\text{RL}}(K) = \sum_{k=1}^K V^{\pi^k} - V^*.$$

We also consider PAC bounds where the learner outputs π^k at each round but may roll-in with other exploratory policies to better collect data.

Offline RL. The learner is given a dataset of prior interactions with the MDP and, unlike online RL, cannot

TABLE 2

Summary of main RL results in this paper. Since the algorithms presented all use temporal difference learning, all results posit some variant of Bellman Completeness. In online RL, each bound uses a slightly different eluder dimension based on the loss; in offline RL, all bounds use the single policy coverage $C^{\tilde{\pi}}$ (defined in Eq. (21)). Finally, we remark that the online and offline RL algorithms employ optimism and pessimism over a version space, so they are computationally inefficient [18]. For computational efficiency, we study the hybrid RL setting (Sec. 8) where we show that fitted-Q iteration with access to online and offline data achieves a bound that is the sum of online and offline bounds.

Setting	Loss	Function Class	Bellman Completeness	Regret / PAC Bound
Online RL	ℓ_{sq}	$\mathcal{F} : \mathcal{X} \times \mathcal{A} \mapsto [0, 1]$	\mathcal{T}^* (Asm. 3)	$\mathcal{O}\left(H \sqrt{K d_{\text{sq}} \ln(H \mathcal{F} /\delta)}\right)$ (Thm. 7)
	ℓ_{bce}	$\mathcal{F} : \mathcal{X} \times \mathcal{A} \mapsto [0, 1]$	\mathcal{T}^* (Asm. 3)	$\mathcal{O}\left(H \sqrt{V^* \cdot K d_{\text{bce}} \ln(H \mathcal{F} /\delta)} + H^2 d_{\text{bce}} \ln(H \mathcal{F} /\delta)\right)$ (Thm. 8)
	ℓ_{mle}	$\mathcal{P} : \mathcal{X} \times \mathcal{A} \mapsto \Delta([0, 1])$	$\mathcal{T}^{D,*}$ (Asm. 5)	$\mathcal{O}\left(H \sqrt{\sum_{k=1}^K \sigma^2(\pi^k) \cdot d_{\text{mle}} \ln(H \mathcal{P} /\delta)} + H^{2.5} d_{\text{mle}} \ln(H \mathcal{P} /\delta)\right)$ (Thm. 9)
Offline RL	ℓ_{sq}	$\mathcal{F} : \mathcal{X} \times \mathcal{A} \mapsto [0, 1]$	$\mathcal{T}^\pi, \forall \pi \in \Pi$ (Asm. 6)	$\mathcal{O}\left(H \sqrt{\frac{C^{\tilde{\pi}} \ln(H \mathcal{F} /\delta)}{n}}\right)$ (Thm. 10)
	ℓ_{bce}	$\mathcal{F} : \mathcal{X} \times \mathcal{A} \mapsto [0, 1]$	$\mathcal{T}^\pi, \forall \pi \in \Pi$ (Asm. 6)	$\mathcal{O}\left(H \sqrt{V^{\tilde{\pi}} \cdot \frac{C^{\tilde{\pi}} \ln(H \mathcal{F} /\delta)}{n}} + H^2 \frac{C^{\tilde{\pi}} \ln(H \mathcal{F} /\delta)}{n}\right)$ (Thm. 10)
	ℓ_{mle}	$\mathcal{P} : \mathcal{X} \times \mathcal{A} \mapsto \Delta([0, 1])$	$\mathcal{T}^{D,\pi}, \forall \pi \in \Pi$ (Asm. 7)	$\mathcal{O}\left(H \sqrt{\sigma^2(\tilde{\pi}) \cdot \frac{C^{\tilde{\pi}} \ln(H \mathcal{P} /\delta)}{n}} + H^{2.5} \frac{C^{\tilde{\pi}} \ln(H \mathcal{P} /\delta)}{n}\right)$ (Thm. 11)

gather more data by interacting with the environment. The dataset takes the form $\mathcal{D} = (\mathcal{D}_1, \mathcal{D}_2, \dots, \mathcal{D}_H)$ where each \mathcal{D}_h contains n i.i.d. samples $(x_{h,i}, a_{h,i}, c_{h,i}, x'_{h,i})$ where $(x_{h,i}, a_{h,i}) \sim \nu_h$, $c_{h,i} \sim C_h(x_{h,i}, a_{h,i})$ and $x'_{h,i} \sim P_h(x_{h,i}, a_{h,i})$. We note that ν_h is simply the marginal distribution over (x_h, a_h) induced by the data generating process, e.g., mixture of policies. We also recall the (single-policy) coverage coefficient: given a comparator policy $\tilde{\pi}$, define $C^{\tilde{\pi}} = \max_{h \in [H]} \|dd_h^{\tilde{\pi}}/d\nu_h\|_\infty$ [73, 60]. The goal is to learn a policy $\hat{\pi}$ with a PAC guarantee against any comparator policy $\tilde{\pi}$ such that $C^{\tilde{\pi}} < \infty$.

Hybrid RL. We also consider the hybrid setting where the learner is given a dataset as in offline RL, and can also gather more data by interacting with the environment as in online RL [58, 8]. By combining the analyses from both online and offline settings, we prove that fitted Q -iteration (FQI) [53], a computationally efficient algorithm that does not induce optimism or pessimism, can achieve first- and second-order regret and PAC bounds.

Bellman equations. Define the Bellman operator \mathcal{T}^π by $\mathcal{T}_h^\pi f(x, a) = \mathbb{E}_{c \sim C_h(x, a), x' \sim P_h(x, a)}[c + f(x', \pi_{h+1})]$ for any function f and policy π . The Bellman equations are $f_h = \mathcal{T}_h^\pi f_{h+1}$ for all h , where $f_h = Q_h^\pi$ is the unique solution. Also, the Bellman optimality operator \mathcal{T}^* is defined by $\mathcal{T}_h^* f(x, a) = \mathbb{E}_{c \sim C_h(x, a), x' \sim P_h(x, a)}[c + \min_{a'} f(x', a')]$. The Bellman optimality equations are $f_h = \mathcal{T}_h^* f_{h+1}$ for all h , where $f_h = Q_h^*$ is the unique solution.

Distributional Bellman equations. There are also distributional analogs to the above [11]. Let $\mathcal{T}^{D,\pi}$ denote the distributional Bellman operator for policy π , defined by $\mathcal{T}_h^{D,\pi} p(x, a) \stackrel{D}{=} c + p(x', a')$ where $c \sim C_h(x, a)$, $x' \sim P_h(x, a)$, $a' \sim \pi_{h+1}(x')$ for any conditional distribution

p . Here $\stackrel{D}{=}$ denotes equality in distribution. The distributional Bellman equations are $p_h \stackrel{D}{=} \mathcal{T}_h^{D,\pi} p_{h+1}$ for all h , where Z_h^π is a solution. The distributional Bellman optimality operator $\mathcal{T}^{D,*}$ is defined by $\mathcal{T}_h^{D,*} p(x, a) \stackrel{D}{=} c + p(x', a')$ where $c \sim C_h(x, a)$, $x' \sim P_h(x, a)$, $a' = \arg \min_{a'} \bar{p}(x', a')$. The distributional Bellman optimality equations are $p_h \stackrel{D}{=} \mathcal{T}_h^{D,*} p_{h+1}$ for all h , where Z_h^* is a solution.

5. ONLINE RL WITH SQUARED-LOSS REGRESSION

We begin our discussion of RL by solving online RL with optimistic temporal-difference (TD) learning with the squared loss for regression [35, 74], which can be viewed as an abstraction for deep RL algorithms such as DQN [51]. The algorithm is value-based, meaning that it aims to learn the optimal Q -function Q^* , which then induces the optimal policy via greedy action selection $\pi_h^*(x) = \arg \min_a Q_h^*(x, a)$. To learn the Q -function, it uses a function class \mathcal{F} that consists of function tuples $f = (f_1, f_2, \dots, f_H) \in \mathcal{F}$ where $f_h : \mathcal{X} \times \mathcal{A} \rightarrow [0, 1]$ and we use the convention that $f_{H+1} = 0$ for all functions f .

In the sequential RL setting, TD learning is a powerful idea for regressing Q -functions where the function at step h is regressed on the current cost plus a *learned* prediction at the next step $h+1$. This process is also known as bootstrapping. One can view this as an approximation to the Bellman equations $Q_h = \mathcal{T}_h Q_{h+1}$ where \mathcal{T} is a Bellman operator. For online RL, we use the Bellman optimality operator \mathcal{T}^* to learn the optimal Q^* , while in offline RL we use the policy-specific Bellman operator \mathcal{T}^π to learn Q^π for all policies π .

Algorithm 1 Policy Data Collection

```

1: Input: policy  $\pi$ , uniform exploration (UA) flag.
2: if UA flag is True then
3:   for step  $h \in [H]$  do
4:     Roll-in  $\pi$  for  $h$  steps to arrive at  $x_h$ .
5:     Then, randomly act  $a_h \sim \text{Unif}(\mathcal{A})$  and observe  $c_h, x'_h$ .
6:   end for
7: else
8:   Roll-in  $\pi$  for  $H$  steps and collect  $x_1, a_1, c_1, \dots, x_H, a_H, c_H$ .
9:   Label  $x'_h = x_{h+1}$  for all  $h \in [H]$ .
10: end if
11: Output: dataset  $\{(x_h, a_h, c_h, x'_h)\}_{h \in [H]}$ .

```

Algorithm 2 Optimistic Online RL

```

1: Input: number of rounds  $K$ , function class  $\mathcal{F}$ , loss function
 $\ell(\hat{y}, y)$ , threshold  $\beta$ , uniform exploration (UA) flag
2: for round  $k = 1, 2, \dots, K$  do
3:   Denote  $\mathcal{F}_k = \mathcal{C}_\beta^\ell(\mathcal{D}_{<k})$  as the version space defined by:

$$\mathcal{C}_\beta^\ell(\mathcal{D}) = \{f \in \mathcal{F} : \forall h \in [H], L_h^\ell(f_h, f_{h+1}, \mathcal{D}_h)$$


$$(13) \quad - \min_{g_h \in \mathcal{F}_h} L_h^\ell(g_h, f_{h+1}, \mathcal{D}_h) \leq \beta\},$$

   where

$$L_h^\ell(f_h, g, \mathcal{D}_h) = \sum_{i=1}^{|\mathcal{D}_h|} \ell(f_h(x_{h,i}, a_{h,i}), \tau^*(g, c_{h,i}, x'_{h,i}))$$

   and  $\tau^*(g, c, x') = c + \min_{a'} g(x', a')$  is the regression target. In
   the proofs, we use  $L^{\text{sq}}$  if  $\ell = \ell_{\text{sq}}$  and  $L^{\text{bce}}$  if  $\ell = \ell_{\text{bce}}$ .
4:   Get optimistic  $f^k \leftarrow \arg \min_{f \in \mathcal{F}_k} \min_a f_1(x_1, a)$ .
5:   Let  $\pi^k$  be greedy w.r.t.  $f^k$ :  $\pi_h^k(x) = \arg \min_a f_h^k(x, a), \forall h$ .
6:   Gather data  $\mathcal{D}_k \leftarrow \text{Alg. 1}(\pi^k, \text{UA flag})$ .
7: end for

```

To formalize TD learning, let (x_h, a_h, c_h, x'_h) be a transition tuple where c_h, x'_h are sampled conditional on x_h, a_h . For a predictor f_{h+1} at step $h+1$, the regression targets at step h are:

$$\begin{aligned} \tau^*(f_{h+1}, c, x') &= c + \min_{a'} f_{h+1}(x', a'), \\ \tau^\pi(f_{h+1}, c, x') &= c + f_{h+1}(x', \pi_{h+1}), \end{aligned}$$

where τ^* is the target for learning Q^* which we use for online RL, and τ^π is the target for learning Q^π which we use for offline RL. The targets are indeed unbiased estimates of the Bellman backup since $\mathcal{T}_h f_{h+1}(x, a) = \mathbb{E}[\tau(f_{h+1}, c, x')]$. Then, we regress f_h by minimizing the loss $\ell(f_h(x, a), \tau(f_{h+1}, c, x'))$ averaged over the data, where the loss function $\ell(\hat{y}, y)$ captures the discrepancy between the prediction \hat{y} and target y . Note this takes the same form as the regression loss from the CSC warmup.

In online RL, the algorithm we consider (Alg. 2) performs TD learning optimistically by maintaining a version space constructed with the TD loss. Specifically, given a dataset $\mathcal{D} = (\mathcal{D}_1, \dots, \mathcal{D}_H)$ where each $\mathcal{D}_h = \{x_{h,i}, a_{h,i}, c_{h,i}, x'_{h,i}\}_{i \in [n]}$ is a set of transition tuples, the version space $\mathcal{C}_\beta^\ell(\mathcal{D})$ is defined in Eq. (13) of Alg. 2. Intuitively, the version space contains all functions $f \in \mathcal{F}$

which nearly minimize the empirical TD risk measured by loss function ℓ , for all time steps h . This construction is useful since it satisfies two properties with high probability. First, any function in $\mathcal{C}_\beta^\ell(\mathcal{D})$ has small population TD risk (a.k.a. Bellman error) w.r.t. ℓ , so we can be assured that choosing any function from the version space is a good estimate of the desired Q^* . Second, we have that Q^* is an element of the version space, which provides a means to achieve optimism (or pessimism) by optimizing over the version space. Indeed, by selecting the function in the version space with the minimum initial state value, we are guaranteed to select a function that lower bounds the optimal policy's cost V^* .

We now summarize the online RL algorithm (Alg. 2), which proceeds iteratively. At each round $k = 1, 2, \dots, K$, the learner selects an optimistic function f^k from the version space defined by previously collected data: $f^k \leftarrow \arg \min_{f \in \mathcal{F}_k} \min_a f_1(x_1, a)$ where $\mathcal{F}_k = \mathcal{C}_\beta^\ell(\mathcal{D}_{<k})$ and $\mathcal{D}_{<k}$ denotes the previously collected data. Then, let π^k be the greedy policy w.r.t. f^k : $\pi^k(x) = \arg \min_a f^k(x, a)$. Finally, roll-in with π^k to collect data, as per Alg. 1.

The roll-in procedure (Alg. 1) has two variants depending on the uniform action (UA) flag. If UA is enabled, we roll-in H times with a slightly modified policy: for each $h \in [H]$, we collect a datapoint from $\pi^k \circ_h \text{unif}(\mathcal{A})$, which denotes the policy that executes π^k for $h-1$ steps and switches to uniform actions at step h . If UA is disabled, we roll-in π^k once and collect trajectory $x_{1,k}, a_{1,k}, c_{1,k}, \dots, x_{H,k}, a_{H,k}, c_{H,k}$. While UA requires H roll-ins per round, this more exploratory data collection is useful for proving bounds with non-linear MDPs. The collected data is then used to define the confidence set at the next round.

As a historical remark, this algorithm was first proposed with the squared loss ℓ_{sq} under the name GOLF by [35] and then extended with the mle loss ℓ_{mle} under the name O-DISCO by [64]. In this section, we focus on the squared loss case, recovering the results of [35]. In the subsequent sections, we propose a new variant with the bce loss ℓ_{bce} , and then finally discuss application of the mle loss, recovering the results of [65].

We now state the Bellman Completeness (BC) assumption needed to ensure that Alg. 2 succeeds [15, 35, 73, 14].

ASSUMPTION 3 (\mathcal{T}^* **-BC).** $\mathcal{T}_h^* f_{h+1} \in \mathcal{F}_h$ for all $h \in [H]$ and $f_{h+1} \in \mathcal{F}_{h+1}$.

BC ensures that the TD-style regression which bootstraps on the next prediction is realizable, playing the same role as realizability (Asm. 1) in the CSC setting. In fact, BC implies realizability in Q^* : $Q_h^* \in \mathcal{F}_h$ for all h , which can be verified by using the Bellman optimality equations and induction from $h = H \rightarrow 1$. While appealing, Q^* -realizability is not sufficient for sample efficient

RL [62, 28, 71] and TD learning can diverge or converge to bad points with realizability alone [59, 53, 39]. We note that Q^* -realizability becomes sufficient when combined with low coverability and generative access to the MDP [48], where the learner can reset to any previously observed states. We believe that the techniques in this paper can lead to first- and second-order bounds with realizability plus generative access for example. However, we do not pursue this direction here since exchanging BC for other conditions is orthogonal to our study of loss functions.

We also define the eluder dimension,² a flexible structural measure that quantifies the complexity of exploration and representation learning [35].

DEFINITION 1 (Eluder Dimension). *Fix any set \mathcal{S} , function class $\Psi = \{\psi : \mathcal{S} \rightarrow \mathbb{R}\}$, distribution class $\mathcal{M} = \{\mu : \Delta(\mathcal{S})\}$, threshold ε_0 , and number $q \in \mathbb{N}$. The ℓ_q -eluder dimension $\text{EluDim}_q(\Psi, \mathcal{M}, \varepsilon_0)$ is defined as the length of the longest sequence $\mu^{(1)}, \dots, \mu^{(L)} \subset \mathcal{M}$ s.t. $\exists \varepsilon \geq \varepsilon_0, \forall t \in [L], \exists \psi \in \Psi$ s.t. $|\mathbb{E}_{\mu^{(t)}}[\psi]| > \varepsilon$ but $\sum_{i < t} |\mathbb{E}_{\mu^{(i)}}[\psi]|^q \leq \varepsilon^q$.*

Taking $\mathcal{S} = \mathcal{X} \times \mathcal{A}$, we will instantiate the Ψ class to be a set of TD errors measured by the regression loss function. For example with squared loss, we set $\Psi_h^{\text{sq}} = \{\mathcal{E}_h^{\text{sq}}(\cdot; f) : f \in \mathcal{F}\}$ where

$$\mathcal{E}_h^{\text{sq}}(x, a; f) := (f_h(x, a) - \mathcal{T}_h^* f_{h+1}(x, a))^2.$$

The distribution class \mathcal{M} will be the set of all visitation distributions by any policy, i.e., $\mathcal{M}_h = \{x, a \mapsto d_h^\pi(x, a) : \pi \in \Pi\}$ where $d_h^\pi(x, a)$ is the state-action visitation distribution of π at time step h . If UA is enabled, then we will have $\mathcal{S} = \mathcal{X}$ and the “V-type” distribution class is $\mathcal{M}_h^V = \{x \mapsto d_h^\pi(x) : \pi \in \Pi\}$ where $d_h^\pi(x)$ is the state visitation distribution of π at time step h . Moreover we let $\Psi_h^{\text{sq}, V} = \{\mathbb{E}_{a \sim \text{unif}(\mathcal{A})}[\psi(x, a)] : \psi \in \Psi_h^{\text{sq}}\}$ denote the “V-type” function class, where “V-type” refers to the fact that the functions only depend on state and not action. Thus, define the eluder dimension for squared loss:

$$d_{\text{sq}} = \max_{h \in [H]} \text{EluDim}_2(\Psi_h^{\text{sq}}, \mathcal{M}_h, 1/K),$$

$$d_{\text{sq}}^V = \max_{h \in [H]} \text{EluDim}_2(\Psi_h^{\text{sq}, V}, \mathcal{M}_h^V, 1/K).$$

We now state the guarantees for [Alg. 2](#) with the squared loss ℓ_{sq} , which recovers the main results of Jin, Liu and Miryoosefi [35].

²[Def. 1](#) is often called the *distributional* eluder dimension to distinguish it from the classic eluder dimension of [57]. To not confuse with distributional RL, we simply refer to it as the eluder dimension.

THEOREM 7. *Under [Asm. 3](#), for any $\delta \in (0, 1)$, w.p.a.l. $1 - \delta$, [Alg. 2](#) with the squared loss ℓ_{sq} and $\beta = 2 \ln(H|\mathcal{F}|/\delta)$ enjoys the following:*

$$\sum_{k=1}^K V^{\pi^k} - V^* \leq \tilde{\mathcal{O}}(H\sqrt{K \cdot d\beta}),$$

where $d = d_{\text{sq}}$ if UA is false and $d = Ad_{\text{sq}}^V$ if UA is true, where A is the number of actions.

This shows that [Alg. 2](#) with the squared loss is guaranteed to learn a policy that converges to the optimal policy at a $\Theta(K^{-1/2})$ rate, which is the minimax-optimal rate. In [Sec. 5.1](#) we show that the V-type dimension can be bounded by the rank of the transition kernel in a low-rank MDP [3, 4], a canonical model for RL with non-linear function approximation.

Computational complexity of version space algorithms. The algorithms we present for online and offline RL optimize over version spaces to establish optimism and pessimism, and this optimization over version space is computationally inefficient in general [18]. The computational hardness comes from the non-convex optimization, not necessarily from any loss function itself. However, we note that the version space optimization is oracle-efficient in the one-step $H = 1$ setting, a.k.a. contextual bandits [25, 22, 65]. In the RL setting, there are also approaches to mitigate the computational hardness of optimism. One common approach in practice is to use myopic exploration strategies such as epsilon-greedy [51, 10], which also has bounded regret when exploration is easy [19, 76]. Another approach which we later present is to work in the hybrid RL setting, where the learner can both interact with the MDP in an online manner and has access to offline data with good coverage [58]. In [Sec. 8](#), we will see that our results for all loss functions naturally carry over to the hybrid setting, giving us both computational and statistical efficiency.

We now prove the main theorem for squared loss [Thm. 7](#). We prove the nested lemmas in the Appendix.

PROOF OF THM. 7. We define the excess squared-loss risk for $f \in \mathcal{F}$ under the visitation distribution of π as

$$\mathcal{E}_h^{\text{sq}}(\pi; f) := \mathbb{E}_\pi[\mathcal{E}_h^{\text{sq}}(x_h, a_h; f)],$$

and also set $\mathcal{E}_{\text{sq}}^{\text{RL}} = \sum_{h=1}^H \mathcal{E}_h^{\text{sq}}$. We first establish an optimism lemma for f^k and prove it in the appendix.

LEMMA 7. *Let $\ell = \ell_{\text{sq}}$ and \mathcal{D}_h be a dataset where the i -th datapoint is collected from π^i , and denote $\mathcal{D} = (\mathcal{D}_1, \dots, \mathcal{D}_H)$. Then under BC ([Asm. 3](#)), for any $\delta \in (0, 1)$, let $\beta = 2 \ln(H|\mathcal{F}|/\delta)$ and define*

$$(14) \quad \hat{f}^{\text{op}} \in \arg \min_{f \in \mathcal{C}_\beta^{\text{sq}}(\mathcal{D})} \min_a f_1(x_1, a).$$

W.p.a.l. $1 - \delta$, we have (a) $\sum_{i=1}^n \mathcal{E}_{\text{sq}}^{\text{RL}}(\pi^i; \hat{f}^{\text{op}}) \leq 2H\beta$, and (b) $\min_a \hat{f}_1^{\text{op}}(x_1, a) \leq V^$.*

By [Lem. 7](#), we have $\sum_{k=1}^K V^{\pi^k} - V^* \leq \sum_{k=1}^K V^{\pi^k} - \min_a f_1^k(x_1, a)$, which can be further decomposed by the performance difference lemma (PDL) [[2, 37](#)].

LEMMA 8 (PDL). $\forall f = (f_1, f_2, \dots, f_H)$ and π , we have $V^\pi - f_1(x_1, \pi) = \sum_{h=1}^H \mathbb{E}_\pi[(\mathcal{T}_h^\pi f_{h+1} - f_h)(x_h, a_h)]$.

By PDL and Cauchy-Schwarz, we have

$$\begin{aligned} (15) \quad & \sum_{k=1}^K V^{\pi^k} - f_1^k(x_1, \pi^k(x_1)) \\ &= \sum_{k=1}^K \sum_{h=1}^H \mathbb{E}_{\pi^k}[\mathcal{T}_h f_{h+1}^k(x_h, a_h) - f_h^k(x_h, a_h)] \\ &\leq \sum_{k=1}^K \sum_{h=1}^H \sqrt{\mathcal{E}_h^{\text{sq}}(f^k, \pi^k)} \leq \sum_{k=1}^K \sqrt{H \mathcal{E}_h^{\text{RL}}(f^k, \pi^k)} \\ &\leq \sqrt{HK \sum_{k=1}^K \mathcal{E}_h^{\text{RL}}(f^k, \pi^k)}. \end{aligned}$$

The final step is to bound $\sum_{k=1}^K \mathcal{E}_h^{\text{RL}}(f^k, \pi^k)$. By [Lem. 7](#), we have that $\sum_{i < k} \mathcal{E}_h^{\text{RL}}(f^k, \pi^i) \lesssim H\beta$ for all k , which is very similar except that the expectations are taken under previous policies $\pi^{< k}$ instead of π^k . It turns out that the eluder dimension can establish a link between the two, via the following ‘‘pigeonhole principle’’ lemma:

LEMMA 9 (Pigeonhole). Fix a number $N \in \mathbb{N}$, a sequence of functions $\psi^{(1)}, \dots, \psi^{(N)} \in \Psi$, and distributions $\mu^{(1)}, \dots, \mu^{(N)} \in \mathcal{M}$. Suppose $\sum_{i < j} |\mathbb{E}_{\mu^{(i)}}[\psi^{(j)}]|^q \leq \beta^q$ for all $j \in [N]$. Then we have $\sum_{j=1}^N |\mathbb{E}_{\mu^{(j)}}[\psi^{(j)}]| \leq 2 \text{EluDim}_q(\Psi, \mathcal{M}, N^{-1}) \cdot (E + \beta^q \ln(EN))$, where $E := \sup_{\mu \in \mathcal{M}, \psi \in \Psi} |\mathbb{E}_\mu[\psi]|$ is the envelope.

If we interpret $\psi^{(i)}$ as the regression error at round i , then [Lem. 9](#) essentially states that ratio of (online) out-of-distribution errors (i.e., $\psi^{(i)}$ measured under $p^{(i)}$) to the (offline) in-distribution errors (i.e., $\psi^{(i)}$ measured under $p^{(1)}, \dots, p^{(i-1)}$) is bounded by the eluder dimension. This generalizes similar results from [[57, 35, 44, 64, 74](#)].

Finally, going back to the regret decomposition in [Eq. \(15\)](#), applying the pigeonhole lemma implies that $\sum_{k=1}^K \mathcal{E}_h^{\text{RL}}(f^k, \pi^k) \leq \tilde{\mathcal{O}}(d_{\text{sq}} H \beta)$. Thus, we have shown that $\sum_{k=1}^K V^{\pi^k} - f_1^k(x_1, \pi^k(x_1)) \leq \tilde{\mathcal{O}}(H \sqrt{K d_{\text{sq}} \beta})$, which proves the desired regret bound.

Moreover, if the uniform action (UA) flag is set, we also perform a change of measure so that: $\sum_{k=1}^K \mathcal{E}_h^{\text{RL}}(f^k, \pi^k) \leq A \sum_{k=1}^K \sum_{h=1}^H \mathbb{E}_{\pi^k \circ_h \text{unif}(\mathcal{A})}((f^h - \mathcal{T}_h f_{h+1}^k)(x_h, a_h))^2 \lesssim A d_{\text{sq}}^V H \beta$. Plugging into [Eq. \(15\)](#) proves the desired PAC bound of $\tilde{\mathcal{O}}(H \sqrt{AK d_{\text{sq}}^V \beta})$. This finishes the proof. \square

5.1 Verifying Assumptions for Low-Rank MDPs

In this subsection, we show that the assumptions in [Thm. 7](#) (as well as subsequent theorems with other loss

functions) are satisfied in low-rank MDPs [[3](#)], a class of rich-observation MDPs where the transition kernel has an unknown low-rank decomposition.

DEFINITION 2 (Low-Rank MDP). An MDP has rank d if its transition kernel has a low-rank decomposition: $P_h(x' | x, a) = \phi_h^*(x, a)^\top \mu_h^*(x')$ where $\phi_h^*, \mu_h^* \in \mathbb{R}^d$ are unknown feature maps that satisfy $\|\phi_h^*(x, a)\|_2 \leq 1$ and $\|\int g d\mu_h^*(x')\|_2 \leq \|g\|_\infty \sqrt{d}$ for all x, a, x' and $g : \mathcal{X} \rightarrow \mathbb{R}$. We also require that the expected cost is linear in the features: $\bar{C}_h(x, a) = \phi_h^*(x, a)^\top v_h^*$ for some unknown vectors $v_h^* \in \mathbb{R}^d$ that satisfy $\|v_h^*\|_2 \leq \sqrt{d}$.

This model captures non-linear representation learning since ϕ^* and μ^* are *unknown* and can be non-linear. The low-rank MDP model also generalizes many other models such as linear MDPs (where ϕ^* is known) [[36](#)], block MDPs [[50](#)] and latent variable models [[52](#)].

To perform representation learning, we posit a feature class $\Phi = \Phi_1 \times \dots \times \Phi_H$ where each $\phi_h : \mathcal{X} \times \mathcal{A} \rightarrow \mathbb{R}^d \in \Phi_h$ is a candidate for the ground truth features ϕ^* .

ASSUMPTION 4 (ϕ^* -realizability). $\phi_h^* \in \Phi_h$ for all h .

Then, the following class of linear functions in Φ satisfies all the assumptions needed in [Thm. 7](#) and [Thm. 8](#), a subsequent result with the bce loss.

$\mathcal{F}_h^{\text{lin}} := \{\text{clip}(\langle \phi_h(\cdot), w \rangle, 0, 1) : w \in \mathbb{R}^d \text{ s.t. } \|w\|_2 \leq 2\sqrt{d}\}$, where $\text{clip}(y, l, h) := \max(\min(y, h), l)$. This function class is sensible because Bellman backups of any function are linear in ϕ_h^* ; thus Q -functions are Bellman backups via the Bellman equations, they are linear in ϕ^* . The clipping is to ensure that the functions are bounded in $[0, 1]$ which is true for the desired Q^* .

We now show that \mathcal{F}^{lin} satisfies BC ([Asm. 3](#)).

LEMMA 10. In a low-rank MDP, under [Asm. 4](#), \mathcal{F}^{lin} satisfies Bellman Completeness ([Asms. 3 and 6](#)).

PROOF OF LEM. 10. Fix any $f_{h+1} \in \mathcal{F}_{h+1}^{\text{lin}}$ and π . We want to show $\mathcal{T}_h^\pi f_{h+1} \in \mathcal{F}_h^{\text{lin}}$. First, we note $\mathcal{T}_h^\pi f_{h+1}(x, a)$ is equal to

$$(16) \quad \phi_h^*(x, a)^\top (v_h^* + \int_{x'} f_{h+1}(x', \pi(x')) d\mu_h^*(x')).$$

Setting $w = v_h^* + \int_{x'} f_{h+1}(x', \pi(x')) d\mu_h^*(x')$, we indeed have that $\|w\|_2 \leq \sqrt{d} + \sqrt{d} \|f_{h+1}\|_\infty \leq 2\sqrt{d}$, which implies $\mathcal{T}_h^\pi f_{h+1}(x, a) \in \mathcal{F}_h^{\text{lin}}$. \square

Moreover, we can also show that the V-type eluder dimension is bounded by the rank d of the low-rank MDP, as defined in [Def. 2](#). We note this applies to both ℓ_1 and ℓ_2 eluder dimensions.

LEMMA 11. *In a low-rank MDP with rank d , we have $\text{EluDim}_1(\Psi_h^V, \mathcal{M}_h^V, \varepsilon) \leq \text{EluDim}_2(\Psi_h^V, \mathcal{M}_h^V, \varepsilon) \leq \mathcal{O}(d \ln(d/\varepsilon))$ for all steps $h \in [H]$ and function classes $\Psi_h^V \subset \mathcal{X} \rightarrow \mathbb{R}$.*

PROOF. $\text{EluDim}_2(\Psi_h^V, \mathcal{M}_h^V, \varepsilon) \leq \mathcal{O}(d \ln(d/\varepsilon))$ can be proved by applying a standard elliptical potential argument to the decomposition in Eq. (16); for example, see [35] or Theorem G.4 of [64]. Then, $\text{EluDim}_1(\Psi_h^V, \mathcal{M}_h^V, \varepsilon) \leq \text{EluDim}_2(\Psi_h^V, \mathcal{M}_h^V, \varepsilon)$ is a simple consequence of the fact that $\sqrt{\sum_i x_i^2} \leq \sum_i |x_i|$ [64, Lemma 5.4]. \square

Since the above lemma holds for all values of Ψ_h^V , this implies that d_{sq}^V (and $d_{\text{bce}}^V, d_{\text{mle}}^V$ to be defined in future theorems) are all bounded by $\tilde{\mathcal{O}}(d)$ in low-rank MDPs. Finally, one can also show that the bracketing entropy of $\mathcal{F}_h^{\text{lin}}$ is $\tilde{\mathcal{O}}(d + \log |\Phi|)$. We note that our PAC bounds can all be extended to allow for infinite classes such as \mathcal{F}^{lin} via a standard bracketing argument, e.g., see [35, 64] for detailed extensions. Thus, we have established that our bounds hold in low-rank MDPs when the algorithm uses the linear function class \mathcal{F}^{lin} .

6. FIRST/SECOND-ORDER BOUNDS FOR ONLINE RL

As we learned from the CSC warmup, algorithms with the squared loss can be sub-optimal in small-cost or small-variance problems. The intuition is that squared loss regression bounds do not capture the underlying heteroskedastic variance; indeed, minimizing squared loss implicitly assumes that the underlying distribution is a homoskedastic Gaussian. We also learned that simply swapping the loss function for the bce loss or mle loss can yield first-order or second-order bounds that are more sample efficient in small-loss or small-variance settings. We now show that this observation smoothly extends to RL as well. In particular, we will sharpen the Cauchy-Schwarz step in the proof of Thm. 7 by leveraging Eq. (2) from the CSC warmup.

6.1 First-Order Bounds with BCE Regression

In this subsection, we analyze Alg. 2 with the bce loss ℓ_{bce} and derive improved first-order bounds. Before stating guarantees with the bce loss, we first define the eluder dimension which measures discrepancy with the Bernoulli squared hellinger distance. Let $\Psi_h^{\text{bce}} = \{\delta_h^{\text{Ber}}(\cdot; f) : f \in \mathcal{F}\}$ where

$$\delta_h^{\text{Ber}}(x, a; f) := h_{\text{Ber}}^2(f_h(x, a), \mathcal{T}_h^* f_{h+1}(x, a))^2,$$

and $\Psi_h^{\text{bce}, V} = \{\mathbb{E}_{a \sim \text{unif}(\mathcal{A})}[\psi(x, a)] : \psi \in \Psi_h^{\text{bce}}\}$. Then define the eluder dimension for bce loss:

$$d_{\text{bce}} = \max_{h \in [H]} \text{EluDim}_1(\Psi_h^{\text{bce}}, \mathcal{M}_h, 1/K),$$

$$d_{\text{bce}}^V = \max_{h \in [H]} \text{EluDim}_1(\Psi_h^{\text{bce}, V}, \mathcal{M}_h^V, 1/K).$$

The following guarantees for Alg. 2 with bce loss is new.

THEOREM 8. *Under Asm. 3, for any $\delta \in (0, 1)$, w.p.a.l. $1 - \delta$, Alg. 2 with the bce loss ℓ_{bce} and $\beta = 2 \ln(H|\mathcal{F}|/\delta)$ enjoys the following:*

$$\sum_{k=1}^K (V^{\pi^k} - V^*) \leq \tilde{\mathcal{O}}(H \sqrt{V^* K \cdot d\beta} + H^2 d\beta),$$

where $d = d_{\text{bce}}$ if UA is false and $d = Ad_{\text{bce}}^V$ if UA is true.

Compared to the non-adaptive bounds of squared loss (Thm. 7), the above bce loss bounds are first-order and shrinks with the optimal policy's cost V^* . This adaptive scaling with V^* gives the bound a small-cost property: if $V^* \leq \mathcal{O}(1/K)$ (i.e., if the optimal policy achieves low cost), then the leading term vanishes and the bound enjoys logarithmic-in- K regret, i.e., $\sum_{k=1}^K V^{\pi^k} - V^* \leq \tilde{\mathcal{O}}(H^2 d\beta)$. In other words, by dividing both sides by K , the sub-optimality gap of the best learned policy shrinks at a fast $\tilde{\mathcal{O}}(1/K)$ rate. Moreover, since $V^* \leq 1$, Thm. 8 is never worse than the $\tilde{\mathcal{O}}(\sqrt{K})$ rate from Thm. 7, and so these two bounds match in the worst-case but bce loss is strictly better in the small-cost regime.

We remark that Thm. 8 uses the same completeness assumption as Thm. 7, although its bound contains a slightly different eluder dimension d_{bce} instead of d_{sq} . The eluder dimension for bce is different because the bce loss naturally measures Bellman error via the Bernoulli squared hellinger while the squared loss uses squared distance. The change to ℓ_1 eluder is due to the generalization bound analysis for log-losses, although it is actually sharper than ℓ_2 eluder [44, 64]. However, this is not a significant change for our MDP of interest: Lem. 11 ensures that both eluder dimensions are bounded by $\tilde{\mathcal{O}}(d)$ in low-rank MDPs with rank d . Indeed, this implies that the first-order bound in Thm. 8 can be specialized for low-rank MDPs by the same argument as before, yielding the first small-loss bound for low-rank MDPs in online RL without distributional RL [64]. Characterizing the exact differences between these eluder dimension variants is an interesting question for future work.

In terms of related works, several other works also propose to use different losses than squared loss to evaluate the Bellman error. For example, Bas-Serrano et al. [9] argued that the logistic loss is advantageous since it is convex, whereas the squared loss is not convex in the Q -function due to the max operator. Farebrother et al. [21] is a more applied paper which shows that classification losses such as cross-entropy scale much better with deep networks than squared loss. Overall these other works provide other reasons for why squared loss is sub-optimal, while we focus on the improved sample efficiency aspect of employing alternative losses. We now prove Thm. 8.

PROOF OF THM. 8. For the bce loss, we measure the Bellman error of $f \in \mathcal{F}$ under π using the squared Hellinger distance of Bernoullis (as defined in Eq. (2)):

$$\delta_h^{\text{Ber}}(\pi; f) := \mathbb{E}_\pi[\delta_h^{\text{Ber}}(x_h, a_h; f)],$$

We define the bce excess risk $\mathcal{E}_h^{\text{bce}}(\pi; f)$ as:

$$-\ln \mathbb{E}_{\pi}[\exp(\frac{1}{2}\ell_{\text{bce}}(\mathcal{T}_h f_{h+1}(x_h, a_h), \tau^*(f_{h+1}, c_h, x_{h+1})) - \frac{1}{2}\ell_{\text{bce}}(f_h(x_h, a_h), \tau^*(f_{h+1}, c_h, x_{h+1}))].$$

Note that $\mathcal{T}_h^* f_{h+1}(x_h, a_h) = \mathbb{E}[\tau^*(f_{h+1}, c_h, x_{h+1}) | x_h, a_h]$, which is realizable by BC. Recall that $\delta_h^{\text{Ber}} \leq \mathcal{E}_h^{\text{bce}}$ by

Lem. 2. We also write $\delta_{\text{Ber}}^{\text{RL}} = \sum_{h=1}^H \delta_h^{\text{Ber}}$ and $\mathcal{E}_{\text{bce}}^{\text{RL}} = \sum_{h=1}^H \mathcal{E}_h^{\text{bce}}$. The following lemma establishes optimism and is analogous to **Lem. 7**.

LEMMA 12. Let $\ell = \ell_{\text{bce}}$. Under the same setup as **Lem. 7** with \hat{f}^{op} selected from $\mathcal{C}_{\beta}^{\text{bce}}$ instead of $\mathcal{C}_{\beta}^{\text{sq}}$, w.p.a.l. $1 - \delta$, we have (a) $\sum_{i=1}^n \mathcal{E}_{\text{bce}}^{\text{RL}}(\hat{f}^{\text{op}}, \pi^i) \leq 2H\beta$, and (b) $\min_a \hat{f}_1^{\text{op}}(x_1, a) \leq V^*$.

By **Lem. 12**, we have $\sum_{k=1}^K (V^{\pi^k} - V^*) \leq \sum_{k=1}^K (V^{\pi^k} - \min_a f_1^k(x_1, a))$. Then, the proof follows similarly as the squared loss case from before, except that we apply the finer Eq. (2) in place of Cauchy-Schwarz:

$$\begin{aligned} (17) \quad & \sum_{k=1}^K (V^{\pi^k} - f_1^k(x_1, \pi^k(x_1))) \\ &= \sum_{k=1}^K \sum_{h=1}^H \mathbb{E}_{\pi^k} [\mathcal{T}_h^* f_{h+1}^k(x_h, a_h) - f_h^k(x_h, a_h)] \\ &\leq \sum_{k=1}^K \sum_{h=1}^H \sqrt{\mathbb{E}_{\pi^k}[f_h^k(x_h, a_h)] \cdot \delta_h^{\text{Ber}}(f^k, \pi^k)} \\ &\quad + \delta_h^{\text{Ber}}(f^k, \pi^k). \\ &\leq \sum_{k=1}^K \sqrt{\sum_{h=1}^H \mathbb{E}_{\pi^k}[f_h^k(x_h, a_h)] \cdot \delta_{\text{Ber}}^{\text{RL}}(f^k, \pi^k)} \\ &\quad + \delta_{\text{Ber}}^{\text{RL}}(f^k, \pi^k). \end{aligned}$$

Now, we bound $\sum_{h=1}^H \mathbb{E}_{\pi^k}[f_h^k(x_h, a_h)]$ by $H V^{\pi^k}$ plus some lower-order error terms, which we achieve with a ‘self-bounding’ lemma:

LEMMA 13. Define δ_h^{Ber} that uses \mathcal{T}^{π} instead of \mathcal{T}^* :

$$\delta_h^{\text{Ber}}(f, \pi, x_h, a_h) := h_{\text{Ber}}^2(f_h(x_h, a_h), \mathcal{T}_h^{\pi} f_{h+1}(x_h, a_h)).$$

Then, for any f, π, x_h, a_h ,

$$f_h(x_h, a_h) \leq eQ_h^{\pi}(x_h, a_h) + 77H\delta_{\text{Ber}}^{\text{RL}}(f, \pi).$$

This implies the corollary:

$$\mathbb{E}_{\pi}[f_h(x_h, a_h)] \lesssim V^{\pi} + H\delta_{\text{Ber}}^{\text{RL}}(f, \pi).$$

By **Lem. 13**, we can bound Eq. (17) by

$$\begin{aligned} &\lesssim \sum_{k=1}^K \sqrt{HV^{\pi^k} \delta_{\text{Ber}}^{\text{RL}}(f^k, \pi^k)} + H\delta_{\text{Ber}}^{\text{RL}}(f^k, \pi^k) \\ &\leq \sqrt{H \sum_{k=1}^K V^{\pi^k} \cdot \sum_{k=1}^K \delta_{\text{Ber}}^{\text{RL}}(f^k, \pi^k)} \\ &\quad + H \sum_{k=1}^K \delta_{\text{Ber}}^{\text{RL}}(f^k, \pi^k). \end{aligned}$$

Algorithm 3 Optimistic Online Distributional RL

- 1: **Input:** number of rounds K , conditional distribution class \mathcal{P} , threshold β , uniform exploration (UA) flag
 - 2: **for** round $k = 1, 2, \dots, K$ **do**
 - 3: Define confidence set $\mathcal{P}_k = \mathcal{C}_{\beta}^{\text{mle}}(\mathcal{D}_{<k})$ where we define:

$$\mathcal{C}_{\beta}^{\text{mle}}(\mathcal{D}) = \{p \in \mathcal{P} : \forall h \in [H], L_h^{\text{mle}}(p_h, p_{h+1}, \mathcal{D}_h)$$

$$(18) \quad - \min_{g_h \in \mathcal{P}_h} L_h^{\text{mle}}(g_h, p_{h+1}, \mathcal{D}_h) \leq \beta\},$$
where

$$L_h^{\text{mle}}(p_h, g, \mathcal{D}_h) = \sum_{i=1}^{|\mathcal{D}_h|} \ell_{\text{mle}}(p_h(x_{h,i}, a_{h,i}), \tau^{\text{D}, *}(g, c_{h,i}, x'_{h,i}))$$
and $\tau^{\text{D}, *}(g, c, x') = c + Z$, $Z \sim g(x', \pi_{\bar{g}}(x'))$ be the mle target.
Note that if c, x' are sampled conditional on x, a , then the target a sample of the random variable $\mathcal{T}_h^{\text{D}, *} g(x, a)$.
 - 4: Get optimistic $p^k \leftarrow \arg \min_{p \in \mathcal{P}_k} \min_a \bar{p}_1(x_1, a)$.
 - 5: Let π^k be greedy w.r.t. p^k : $\pi_h^k(x) = \arg \min_a \bar{p}_h^k(x, a), \forall h$.
 - 6: Gather data $\mathcal{D}_k \leftarrow \text{Alg. 1}(\pi^k, \text{UA flag}).$
 - 7: **end for**
-

By **Lem. 12** and the pigeonhole principle, the error terms $\mathcal{E}_{\text{bce}}^{\text{RL}}$ can be bounded similarly as in the squared loss proof: we can bound $\sum_{k=1}^K \delta_{\text{Ber}}^{\text{RL}}(f^k, \pi^k)$ by $\tilde{\mathcal{O}}(d_{\text{bce}} H \beta)$ if UA is false, and by $\tilde{\mathcal{O}}(Ad_{\text{bce}}^V H \beta)$ if UA is true.

Thus, setting $d = d_{\text{bce}}$ if UA is false, and $d = Ad_{\text{bce}}^V$ if UA is true, we have proven

$$\sum_{k=1}^K (V^{\pi^k} - V^*) \leq H \sqrt{\sum_{k=1}^K V^{\pi^k} \cdot d \beta} + H^2 d \beta.$$

Finally, we observe an implicit inequality where we can replace $\sum_{k=1}^K V^{\pi^k}$ by KV^* by collecting a factor of 3, as shown by the following lemma.

LEMMA 14. If $\sum_{k=1}^K (V^{\pi^k} - V^*) \leq c \sqrt{\sum_{k=1}^K V^{\pi^k}} + c^2$, then $\sum_{k=1}^K V^{\pi^k} - V^* \leq c \sqrt{2KV^*} + 3c^2$.

This concludes the proof of **Thm. 8**. □

6.2 Second-Order Bounds with MLE

A natural question is how can we achieve second-order bounds in RL? In this section, we consider a distributional variant of the online RL algorithm that uses the mle loss to learn the cost-to-go distributions Z^* . RL algorithms that learn the cost-to-go distributions are often referred to as *distributional RL* (DistRL) [11] and have resulted in a plethora of empirical success [10, 17, 31, 12, 32, 21]. Distributional losses, such as the mle loss and quantile regression loss, were initially motivated by improve representation learning and multi-task learning, but a theoretically rigorous explanation was an open question. Recently, [64, 65] provided an answer to this mystery by proving that DistRL automatically yields first- and second-order bounds in RL, thus establishing the benefits of DistRL.

In this section, we review the results of [65], a refinement of [64] that introduced the mle-loss variant of the optimistic online RL algorithm. To learn the optimal policy's cost-to-go distributions Z^* , we posit a conditional distribution class \mathcal{P} that consists of conditional distribution tuples $p = (p_1, p_2, \dots, p_H) \in \mathcal{P}$ where $p_h : \mathcal{X} \times \mathcal{A} \rightarrow \Delta([0, 1])$. We use the convention that p_{H+1} is deterministic point-mass at 0 for all conditional distributions p . Then, [Alg. 3](#) takes exactly the same structure as [Alg. 2](#) except that it performs a distributional variant of TD to solve the distributional Bellman equation $Z_h^* \stackrel{D}{=} \mathcal{T}_h^D Z_{h+1}^*$. It uses mle to learn the cost-to-go distributions and acts greedily with respect to the learned distribution's mean.

To ensure that distributional TD learning succeeds, we assume distributional BC (DistBC) [72, 64].

ASSUMPTION 5 ($\mathcal{T}^{D,*}$ -DistBC). $\mathcal{T}_h^{D,*} p_{h+1} \in \mathcal{P}_h$ for all $h \in [H]$ and $p_{h+1} \in \mathcal{P}_{h+1}$.

Distributional BC posits that the distribution class is closed under the distributional Bellman operator and is a stronger condition than the standard completeness condition ([Asm. 3](#)). Nevertheless, in low-rank MDPs with discrete cost distributions, [Asm. 5](#) can be satisfied with a linear distribution class [65, Section 5.1], which also has bounded bracketing entropy of $\tilde{\mathcal{O}}(dM + \log |\Phi|)$ where M is the number of discretizations.

Next, we define the eluder dimension for mle loss. Let $\Psi_h^{\text{mle}} = \{\delta_h^{\text{dis}}(\cdot; p) : p \in \mathcal{P}\}$ where

$$\delta_h^{\text{dis}}(x, a; p) := h^2(p_h(x, a), \mathcal{T}^{D,*} p_h(x, a))$$

and $\Psi_h^{\text{mle}, V} = \{\mathbb{E}_{a \sim \text{unif}(\mathcal{A})}[\psi(x, a)] : \psi \in \Psi_h^{\text{mle}}\}$. Define:

$$d_{\text{mle}} = \max_{h \in [H]} \text{EluDim}_1(\Psi_h^{\text{mle}}, \mathcal{M}_h, 1/K),$$

$$d_{\text{mle}}^V = \max_{h \in [H]} \text{EluDim}_1(\Psi_h^{\text{mle}, V}, \mathcal{M}_h^V, 1/K).$$

The following is the main online RL result from [65].

THEOREM 9. Under [Asm. 5](#), for any $\delta \in (0, 1)$, w.p.a.l. $1 - \delta$, [Alg. 3](#) with the mle loss ℓ_{mle} and $\beta = 2 \ln(H|\mathcal{P}|/\delta)$ enjoys the following:

$$\sum_{k=1}^K (V^{\pi^k} - V^*) \leq \tilde{\mathcal{O}}(H \sqrt{\sum_{k=1}^K \sigma^2(\pi^k) \cdot d\beta} + H^{2.5} d\beta),$$

where $d = d_{\text{mle}}$ if UA is false and $d = Ad_{\text{mle}}^V$ if UA is true.

The above mle loss bounds scales with the variances of the policies selected by the algorithm, and are thus are called second-order (a.k.a. variance dependent) bounds. As we saw in the CSC setting, a second-order bound is actually strictly sharper than the first-order bound and this is also true in RL [65, Theorem 2.1]. The variance bound can be much tighter in near-deterministic settings where the optimal policy's cost is far from zero.

One drawback of the DistRL approach is that it requires modeling the entire conditional distributions which are more complex than the conditional mean, i.e., \mathcal{P} is generally larger and more complex than \mathcal{F} from [Thms. 7](#) and [8](#). Also, DistRL requires completeness w.r.t. the distributional Bellman operator, which is stronger than standard BC. However, it is worth noting that in practice DistRL is often much more performant than non-distributional approaches, which suggests that these modeling conditions (e.g., distributional completeness) are not so restrictive [10, 17, 21]. Closing this gap between theory and practice is an interesting future direction.

We also highlight two recent works on second-order bounds for contextual bandits, the one-step special case of RL where DistBC simplifies to distributional realizability ([Asm. 2](#)). In this setting, Pacchiano [55] proved a second-order bound for contextual bandits with only mean realizability, which is weaker than distributional realizability, by using thresholded least squares. Concurrently, Jia et al. [33] obtained a similar bound as [55] for the strong adversary setting and a complementary bound for the weak adversary setting. Jia et al. [33] also furnished a lower bound proving that the eluder-based second-order bounds in [Thm. 9](#) and [65] are tight and unimprovable even if the number of actions is less than the eluder dimension.

Computational considerations for bce and mle. We remark that switching from squared loss to bce does not incur any computational overhead and is a single line of code in practice. The mle loss is used for distribution fitting and thus the function approximator should output a distribution instead of a scalar. In practice, the distribution can be modeled with histograms [10, 31, 21] or quantiles [17] and is only a constant factor more to compute and maintain than the non-distributional losses. Thus, the mle loss and distributional RL more generally also do not incur any notable computational overhead in practice.

PROOF OF THM. 9. We measure the distributional Bellman error of $p \in \mathcal{P}$ under π with the squared Hellinger distance:

$$\delta_h^{\text{dis}}(\pi; p) := \mathbb{E}_\pi[\delta_h^{\text{dis}}(x_h, a_h; p)],$$

We define the mle excess risk $\mathcal{E}_h^{\text{mle}}(p, \pi)$ as:

$$-\ln \mathbb{E}_\pi[\exp(\frac{1}{2} \ell_{\text{bce}}(\mathcal{T}_h^{D,*} p_{h+1}(x_h, a_h), \tau^{D,*}(p, c_h, x_{h+1})) - \frac{1}{2} \ell_{\text{bce}}(p_h(x_h, a_h), \tau^{D,*}(p, c_h, x_{h+1})))],$$

Recall we have that $\delta_h^{\text{dis}} \leq \mathcal{E}_h^{\text{mle}}$ by [Lem. 4](#). We also write $\delta_{\text{dis}}^{\text{RL}} = \sum_{h=1}^H \delta_h^{\text{dis}}$ and $\mathcal{E}_{\text{mle}}^{\text{RL}} = \sum_{h=1}^H \mathcal{E}_h^{\text{mle}}$. We now establish the optimism lemma, which is analogous to [Lem. 7](#) and [Lem. 12](#).

LEMMA 15. Let $\ell = \ell_{\text{mle}}$ and \mathcal{D}_h be the same as in Lems. 7 and 12. Then, under Asm. 5, for any $\delta \in (0, 1)$ let $\beta = 2 \ln(H|\mathcal{P}|/\delta)$ and define

$$\hat{p}^{\text{op}} \in \arg \min_{p \in \mathcal{C}_{\beta}^{\text{mle}}(\mathcal{D})} \min_a \bar{p}_1(x_1, a)$$

W.p.a.l. $1 - \delta$, we have (a) $\sum_{i=1}^n \mathcal{E}_{\text{mle}}^{\text{RL}}(\hat{p}^{\text{op}}, \pi^i) \leq 2H\beta$ and (b) $\min_a \bar{p}_1^{\text{op}}(x_1, a) \leq V^*$.

By Lem. 15, we have $\sum_{k=1}^K V^{\pi^k} - V^* \leq \sum_{k=1}^K V^{\pi^k} - \min_a \bar{p}_1^k(x_1, a)$. Now we apply the second-order lemma:

$$\begin{aligned} & \sum_{k=1}^K (V^{\pi^k} - \bar{p}_1^k(x_1, \pi^k(x_1))) \\ &= \sum_{k=1}^K \sum_{h=1}^H \mathbb{E}_{\pi^k} [\mathcal{T}_h^{\star} \bar{p}_{h+1}^k(x_{h+1}) - \bar{p}_h^k(x_h, a_h)] \\ &= \sum_{k=1}^K \sum_{h=1}^H \mathbb{E}_{\pi^k} [\overline{\mathcal{T}_h^{\text{D}, \star} p_{h+1}^k(x_{h+1})} - \bar{p}_h^k(x_h, a_h)] \\ &\leq \sum_{k=1}^K \sqrt{\sum_{h=1}^H \mathbb{E}_{\pi^k} [\sigma^2(p_h^k(x_h, a_h))] \cdot \delta_{\text{dis}}^{\text{RL}}(p^k, \pi^k)} \\ &\quad + \delta_{\text{dis}}^{\text{RL}}(p^k, \pi^k). \end{aligned}$$

Now, we bound the variance term by $\sum_{h=1}^H \mathbb{E}_{\pi^k} [\sigma^2(c_h + V_{h+1}^{\pi^k}(x_{h+1}))]$ plus some lower-order error terms. We achieve this with the following lemma, which can be viewed as a variance analog of Lem. 13.

LEMMA 16. Define the state-action analog of δ_h^{dis} :

$$\delta_h^{\text{dis}}(p, \pi, x_h, a_h) := h^2(p_h(x_h, a_h), \mathcal{T}_h^{\text{D}, \pi} p_{h+1}(x_h, a_h)),$$

where $\mathcal{T}_h^{\text{D}, \pi} p(x, a) \stackrel{D}{=} C(x, a) + p(X', \pi(X'))$ is the distributional Bellman backup of p under π . Then, for any p, π, x_h, a_h , we have

$$\sigma^2(p_h(x_h, a_h)) \leq 2e\sigma^2(Z_h^{\pi}(x_h, a_h)) + H\delta_{\text{dis}}^{\text{RL}}(p, \pi),$$

where $\sigma^2(Z_h^{\pi}(x_h, a_h))$ denotes the variance of the random variable $Z_h^{\pi}(x_h, a_h)$. This also implies the corollary:

$$\mathbb{E}[\sigma^2(p_h(x_h, a_h))] \lesssim \sigma^2(Z^{\pi}) + H^2\delta_{\text{dis}}^{\text{RL}}(p, \pi).$$

By Lem. 16, we can bound Eq. (19) by

$$\begin{aligned} & \lesssim \sum_{k=1}^K \sqrt{H\sigma^2(Z^{\pi^k}) \cdot \delta_{\text{dis}}^{\text{RL}}(p^k, \pi^k)} + H^{1.5}\delta_{\text{dis}}^{\text{RL}}(p^k, \pi^k) \\ & \leq \sqrt{H \sum_{k=1}^K \sigma^2(Z^{\pi^k}) \cdot \sum_{k=1}^K \delta_{\text{dis}}^{\text{RL}}(p^k, \pi^k)} \\ & \quad + H^{1.5} \sum_{k=1}^K \delta_{\text{dis}}^{\text{RL}}(p^k, \pi^k). \end{aligned}$$

By Lem. 15 and the pigeonhole principle, the error terms $\mathcal{E}_{\text{mle}}^{\text{RL}}$ can be bounded similarly as before: we can bound $\sum_{k=1}^K \delta_{\text{dis}}^{\text{RL}}(p^k, \pi^k)$ by $\tilde{\mathcal{O}}(d_{\text{mle}} H \beta)$ if UA is false, and by $\tilde{\mathcal{O}}(Ad_{\text{mle}}^V H \beta)$ if UA is true. This finishes the proof of Thm. 9. \square

Algorithm 4 Pessimistic Offline RL

- 1: **Input:** function class \mathcal{F} , offline dataset \mathcal{D} , loss function $\ell(\hat{y}, y)$, threshold β .
 - 2: **for** each policy $\pi \in \Pi$ **do**
 - 3: Denote $\mathcal{F}_{\pi} = \mathcal{C}_{\beta}^{\ell}(\mathcal{D}; \pi)$ as the version space defined by:
 - $$\mathcal{C}_{\beta}^{\ell}(\mathcal{D}; \pi) = \{f \in \mathcal{F} : \forall h \in [H], L_h^{\ell}(f_h, f_{h+1}, \mathcal{D}_h, \pi)$$

$$(20) \quad - \min_{g_h \in \mathcal{F}_h} L_h^{\ell}(g_h, f_{h+1}, \mathcal{D}_h, \pi) \leq \beta\},$$
 - where
 - $$L_h^{\ell}(f_h, g, \mathcal{D}_h, \pi) = \sum_{i=1}^{|\mathcal{D}_h|} \ell(f_h(x_{h,i}, a_{h,i}), \tau^{\pi}(g, c_{h,i}, x'_{h,i}))$$
 - and $\tau^{\pi}(g, c, x') = c + g(x', \pi)$ is the regression target. In the proofs, we use L^{sq} if $\ell = \ell_{\text{sq}}$ and L^{bce} if $\ell = \ell_{\text{bce}}$.
 - 4: Get pessimistic $f^{\pi} \leftarrow \arg \max_{f \in \mathcal{F}_{\pi}} \min_a f_1(x_1, a)$.
 - 5: **end for**
 - 6: **Return:** $\hat{\pi} = \arg \min_{\pi \in \Pi} \min_a f_1^{\pi}(x_1, a)$.
-

7. OFFLINE RL VIA PESSIONISTIC REGRESSION

In offline RL, we are given a dataset \mathcal{D} of size N and the goal is to learn a good policy in a purely offline manner, without any interactions with the environment. Since we cannot explore in offline RL, a natural strategy is to be cautious about any states and actions not covered by the given dataset – that is, we should be conservative or pessimistic about unseen parts of the environment where we may make catastrophic errors [43, 56, 73]. Indeed, it is intuitively clear that we can only hope to learn a good policy on the support of the given data. This will be soon formalized with the single-policy coverage coefficient.

We summarize the offline RL algorithm in Alg. 4. We achieve pessimism by maximizing over the version space defined in Eq. (20), which is an inversion of online RL which minimizes over a similar version space. The only other difference is the regression target:

$$\tau^{\pi}(f_{h+1}, c, x') = c + f_{h+1}(x', \pi_{h+1}),$$

is an unbiased estimate of $\mathcal{T}_h^{\pi} f_{h+1}$ in contrast to the online case where τ^{\star} was used to estimate $\mathcal{T}_h^{\star} f_{h+1}$. Thus, we instead use the policy-wise BC for offline RL:

ASSUMPTION 6 (\mathcal{T}^{π} -BC). $\mathcal{T}_h^{\pi} f_{h+1} \in \mathcal{F}_h$ for all $h \in [H]$, $f_{h+1} \in \mathcal{F}_{h+1}$ and $\pi \in \Pi$.

Since we may take $\pi = \pi_f$, this is technically stronger than Asm. 3. Nevertheless, Asm. 6 is also satisfied in low-rank MDPs by the linear function class \mathcal{F}^{lin} and so changing from Asm. 3 to Asm. 6 does not change any conclusions we make. As a historical remark, Alg. 4 was first proposed with the squared loss ℓ_{sq} under the name BCP by [73] and then extended with the mle loss ℓ_{mle} under the name P-DISCO by [64].

We introduce the single-policy coverage coefficient: for any given comparator policy in the policy class $\tilde{\pi} \in \Pi$, its

coverage coefficient is defined by:

$$(21) \quad C^{\tilde{\pi}} := \max_{h \in [H]} \max_{x,a} \frac{d_h^{\tilde{\pi}}(x,a)}{\nu_h(x,a)}.$$

For simplicity, we set the policy class to all greedy policies induced by our function class $\Pi_{\mathcal{F}} = \{\pi_f : f \in \mathcal{F}\}$.³ In the following theorem, the squared loss case recovers the results of [73] and the bce loss result is new.

THEOREM 10. *Under Asm. 6, for any $\delta \in (0, 1)$, w.p.a.l. $1 - \delta$, Alg. 4 with $\beta = 2 \ln(H|\mathcal{F}|/\delta)$ has the following guarantees each loss function:*

1. If $\ell = \ell_{\text{sq}}$, then for any comparator policy $\tilde{\pi} \in \Pi_{\mathcal{F}}$,

$$V^{\tilde{\pi}} - V^{\tilde{\pi}} \leq \tilde{\mathcal{O}}(H \sqrt{\frac{C^{\tilde{\pi}} \beta}{n}}).$$

2. If $\ell = \ell_{\text{bce}}$, then for any comparator policy $\tilde{\pi} \in \Pi_{\mathcal{F}}$,

$$V^{\tilde{\pi}} - V^{\tilde{\pi}} \leq \tilde{\mathcal{O}}(H \sqrt{V^{\tilde{\pi}} \cdot \frac{C^{\tilde{\pi}} \beta}{n}} + H^2 \frac{C^{\tilde{\pi}} \beta}{n}).$$

We see that the squared loss algorithm always converges at a slow $\tilde{\mathcal{O}}(1/\sqrt{n})$ rate. Simply changing the squared loss to the bce loss yields a first-order bound that converges at a fast $\tilde{\mathcal{O}}(1/n)$ rate in the small-cost regime where $V^{\tilde{\pi}} \lesssim 1/n$, and is never worse than the squared loss bound since $V^{\tilde{\pi}} \leq 1$. Again, the only change needed to achieve the improved bound is to change the loss function from squared loss to bce loss, which mirrors our observations from before. One difference with the first-order online RL bound is that small-cost term here is $V^{\tilde{\pi}}$ instead of V^* . Of course, we can set $\tilde{\pi} = \pi^*$ to recover the same small-cost term. However, this offline RL bound is more general since it can be applied to any comparator policy $\tilde{\pi}$ with bounded coverage coefficient.

PROOF OF THM. 10. We only prove the bce case as the squared loss case follows essentially the same structure. The key difference compared to online RL is that we establish pessimism instead of optimism.

LEMMA 17 (Pessimism). *Let $\ell = \ell_{\text{bce}}$. Under Asm. 3, for any $\delta \in (0, 1)$, setting $\beta = \Theta(\ln(H|\mathcal{F}|/\delta))$. Then, w.p.a.l. $1 - \delta$, for all $\pi \in \Pi$, (a) $\mathcal{E}_{\text{bce}}^{\text{RL}}(\hat{f}^{\pi}, \nu) \leq \frac{2H\beta}{n}$, and (b) $\min_a \hat{f}_1^{\pi}(x_1, a) \geq V^{\pi}$.*

PROOF OF LEM. 17. The proof is essentially identical to that of Lem. 12 where we show that w.p.a.l. $1 - \delta$, (1) all elements of the version space have low excess risk and (2) Q^{π} lies in the version space. The only difference is that \hat{f}^{π} is defined as the argmax rather than argmin, so that we have pessimism (greater than V^{π}) instead of optimism. \square

³The offline RL results can be extended for general, infinite policy classes with log covering numbers [16] or entropy integrals [38].

Algorithm 5 Pessimistic Offline Distributional RL

1: **Input:** conditional distribution class \mathcal{P} , offline dataset \mathcal{D} , threshold β .
2: **for** each policy $\pi \in \Pi$ **do**
3: Denote $\mathcal{P}_{\pi}^{\text{mle}} = \mathcal{C}_{\beta}^{\text{mle}}(\mathcal{D}; \pi)$ as the version space defined by:

$$\mathcal{C}_{\beta}^{\text{mle}}(\mathcal{D}; \pi) = \{p \in \mathcal{P} : \forall h \in [H], L_h^{\text{mle}}(p_h, p_{h+1}, \mathcal{D}_h, \pi) - \min_{g_h \in \mathcal{F}_h} L_h^{\text{mle}}(g_h, p_{h+1}, \mathcal{D}_h, \pi) \leq \beta\},$$
where $L_h^{\text{mle}}(f_h, g, \mathcal{D}_h, \pi)$ is

$$\sum_{i=1}^{|\mathcal{D}_h|} \ell_{\text{mle}}(f_h(x_{h,i}, a_{h,i}), \tau^{\text{D}, \pi}(g, c_{h,i}, x'_{h,i}))$$
and $\tau^{\text{D}, \pi}(g, c, x') = c + Z, Z \sim g(x', \pi(x'))$ is the mle target.
Note that if c, x' are sampled conditional on x, a , then the target is a sample of the random variable $\mathcal{T}_h^{\text{D}, \pi} g(x, a)$.
4: Get pessimistic $p^{\pi} \leftarrow \arg \max_{p \in \mathcal{P}_{\pi}} \min_a \bar{p}_1(x_1, a)$.
5: **end for**
6: **Return:** $\hat{\pi} = \arg \min_{\pi \in \Pi} \min_a \bar{p}_1^{\pi}(x_1, a)$.

By Lem. 17, we have $V^{\hat{\pi}} - V^{\tilde{\pi}} \leq \min_a f_1^{\hat{\pi}}(x_1, a) - V^{\tilde{\pi}}$. Then, by definition of $\hat{\pi}$, we further bound this by $\min_a f_1^{\hat{\pi}}(x_1, a) - V^{\tilde{\pi}}$. Now, we decompose with PDL:

$$\begin{aligned} & \min_a f_1^{\hat{\pi}}(x_1, a) - V^{\tilde{\pi}} \\ &= \sum_{h=1}^H \mathbb{E}_{\tilde{\pi}}[f_h^{\tilde{\pi}}(x_h, a_h) - \mathcal{T}_h^{\tilde{\pi}} f_{h+1}^{\tilde{\pi}}(x_h, a_h)] \\ &\leq \sqrt{\sum_{h=1}^H \mathbb{E}_{\tilde{\pi}}[f_h^{\tilde{\pi}}(x_h, a_h)] \cdot \delta_{\text{Ber}}^{\text{RL}}(f^{\tilde{\pi}}, \tilde{\pi}) + \delta_{\text{Ber}}^{\text{RL}}(f^{\tilde{\pi}}, \tilde{\pi})} \\ &\lesssim \sqrt{HV^{\tilde{\pi}} \cdot \delta_{\text{Ber}}^{\text{RL}}(f^{\tilde{\pi}}, \tilde{\pi})} + H \delta_{\text{Ber}}^{\text{RL}}(f^{\tilde{\pi}}, \tilde{\pi}) \end{aligned}$$

By importance sampling and Lem. 17, the error terms can be bounded by $\tilde{\mathcal{O}}(C^{\tilde{\pi}} \cdot \frac{H\beta}{n})$. This completes the proof of Thm. 10. \square

7.1 Second-Order Bounds via Distributional RL

We now show that DistRL with the mle loss can yield second-order guarantees, recovering the results of [65]. We make a few minor changes to the pessimistic offline RL algorithm. We consider the set of greedy policies w.r.t. the means of the conditional distribution as our policy class $\Pi_{\mathcal{P}} = \{\pi_p : p \in \mathcal{P}\}$. Following Xie et al. [73], Wang et al. [65], in offline RL, we posit the policy-wise distributional BC condition.

ASSUMPTION 7 ($\mathcal{T}^{\text{D}, \pi}$ -DistBC). $\mathcal{T}_h^{\text{D}, \pi} p_{h+1} \in \mathcal{P}_h$ for all $h \in [H], p_{h+1} \in \mathcal{P}_{h+1}$ and $\pi \in \Pi$.

THEOREM 11. *Under Asm. 7, for any $\delta \in (0, 1)$, w.p.a.l. $1 - \delta$, Alg. 5 with $\beta = 2 \ln(H|\mathcal{P}|/\delta)$ has the following guarantee: for any comparator policy $\tilde{\pi} \in \Pi_{\mathcal{P}}$,*

$$V^{\hat{\pi}} - V^{\tilde{\pi}} \leq \tilde{\mathcal{O}}(H \sqrt{\sigma^2(\tilde{\pi}) \cdot \frac{C^{\tilde{\pi}} \beta}{n}} + H^{2.5} \frac{C^{\tilde{\pi}} \beta}{n}).$$

Algorithm 6 Fitted Q -Iteration for Hybrid RL

```

1: Input: number of rounds  $K$ , function class  $\mathcal{F}$ , offline dataset
    $\mathcal{D}^{\text{off}}$ , loss function  $\ell(\hat{y}, y)$ , uniform exploration (UA) flag
2: for episode  $k = 1, 2, \dots, K$  do
3:   for each  $h = H, H-1, \dots, 1$  do
4:     Recall the loss from Alg. 2 (Eq. \(13\)):

$$L_h^\ell(f_h, g, \mathcal{D}_h) = \sum_{i=1}^{|\mathcal{D}_h|} \ell(f_h(x_{h,i}, a_{h,i}), \tau^*(g, c_{h,i}, x'_{h,i}))$$

5:     Set  $f_h^k = \arg \min_{f_h \in \mathcal{F}_h} L_h^\ell(f_h, f_{h+1}^k, \mathcal{D}_h^{\text{off}} \cup \mathcal{D}_{<h}^{\text{on}})$ .
6:   end for
7:   Let  $\pi^k$  be greedy w.r.t.  $f^k$ :  $\pi_h^k(x) = \arg \min_a f_h^k(x, a)$ .
8:   Gather data  $\mathcal{D}_k^{\text{on}} \leftarrow \text{Alg. 1}(\pi^k, \text{UA flag})$ .
9: end for

```

Since $\sigma^2(\tilde{\pi}) \leq V^{\tilde{\pi}}$, this implies a first-order bound as well. This variance bound can be much tighter in near-deterministic settings where the comparator's variance is near zero, but its cost is far from zero. However, as was the case in online RL, DistRL still has the drawbacks of requiring a distributional class and DistBC. While these are more stringent conditions in theory, DistRL has achieved state-of-the-art in many offline RL tasks as well [47], suggesting that the benefits of DistRL can outweigh the stronger modeling assumptions in practice. The proof of [Thm. 11](#), which we omit due to space, follows from the same argument as the proof of [Thm. 10](#), coupled with the variance arguments from [Thm. 9](#). The interested reader may find the full proof in [65].

8. COMPUTATIONAL EFFICIENCY VIA HYBRID RL

While we have exhibited the central role of loss functions, achieving tight variance-adaptive bounds, in both online and offline RL, one issue which we have not yet addressed is computational efficiency. As mentioned earlier, optimizing over the version space is computationally difficult (NP-hard) even in tabular MDPs [18].

In this section, we discuss a solution via the hybrid RL setting where the learner can access an offline dataset \mathcal{D}^{off} with good coverage and also interact with the environment. We show that Fitted-Q Iteration (FQI) [53], a computationally efficient algorithm, can also enjoy first- and second-order guarantees by simply regressing with the bce and mle losses. The FQI algorithm in the hybrid setting was first proposed with the squared loss ℓ_{sq} by [58] and our extensions to the bce and mle losses are novel.

Intuitively, the offline dataset mitigates the need for optimism, while the online interactions mitigate the need for pessimism – together, they obviate the need for maintaining a version space. For the following guarantees, we use $C^{\tilde{\pi}}$ to denote the coverage coefficient of the comparator policy $\tilde{\pi}$ under the data generating distribution of \mathcal{D}^{off} . We also assume the offline dataset to be as large as the number of interactions, *i.e.*, $|\mathcal{D}^{\text{off}}| \geq \Omega(K)$ [58].

THEOREM 12. *Under [Asm. 3](#) and $|\mathcal{D}^{\text{off}}| \geq \Omega(K)$, for any $\delta \in (0, 1)$, w.p.a.l. $1 - \delta$, [Alg. 6](#) has the following guarantees for each loss function:*

1. *If $\ell = \ell_{\text{sq}}$, for any comparator policy $\tilde{\pi} \in \Pi_{\mathcal{F}}$,*

$$\sum_{k=1}^K (V^{\pi^k} - V^{\tilde{\pi}}) \leq \tilde{\mathcal{O}}(H \sqrt{K \cdot (d + C^{\tilde{\pi}})\beta}),$$
where $d = d_{\text{sq}}$ if UA is false, and $d = Ad_{\text{sq}}^V$ if UA is true.
2. *If $\ell = \ell_{\text{bce}}$, for any comparator policy $\tilde{\pi} \in \Pi_{\mathcal{F}}$,*

$$\begin{aligned} \sum_{k=1}^K (V^{\pi^k} - V^{\tilde{\pi}}) &\leq \tilde{\mathcal{O}}(H \sqrt{V^{\tilde{\pi}} K \cdot (d + C^{\tilde{\pi}})\beta} \\ &\quad + H^2(d + C^{\tilde{\pi}})\beta) \end{aligned}$$
where $d = d_{\text{bce}}$ if UA is false, and $d = Ad_{\text{bce}}^V$ if UA is true.

Importantly, we see that simply changing the loss from ℓ_{sq} to ℓ_{bce} again leads to improved first-order bounds, which again supports our earlier observations. Compared with our prior results, the main advantage of [Thm. 12](#) is computational: it bounds the sub-optimality of a computationally efficient algorithm FQI, which much more closely resembles deep RL algorithms such as DQN [51]. From a statistical perspective, the hybrid RL bound is actually worse than either pure online or offline bounds, since it takes the form of:

$$\text{online RL bound} + \text{offline RL bound}.$$

Indeed, the hybrid RL bounds contain both the structural condition such as eluder dimension and the coverage coefficient $V^{\tilde{\pi}}$. This form will be made clear in the proof, which simply combines the prior online and offline RL results. We finally discuss some related works. [7] analyzed FQI with ℓ_{bce} in the pure offline setting and proved a first-order bound that depends on the much larger global coverage coefficient $C^{\Pi} = \max_{\tilde{\pi} \in \Pi} C^{\tilde{\pi}}$, which is needed to analyze FQI in the pure offline setting [15]. Also, [49] is able to achieve computationally efficient learning low-rank MDPs without requiring offline data with good partial coverage; however, their bounds are neither first nor second-order. It would be interesting future work to adapt the techniques in this paper to derive variance dependent bounds for computationally efficient algorithms without requiring good offline data.

PROOF OF [THM. 12](#). For any comparator policy $\tilde{\pi}$, we decompose:

$$\begin{aligned} \sum_{k=1}^K (V^{\pi^k} - V^{\tilde{\pi}}) &= \sum_{k=1}^K \mathbb{E}[V^{\pi^k} - \min_a f_1^k(x_1, a)] \\ &\quad + \mathbb{E}[\min_a f_1^k(x_1, a) - V^{\tilde{\pi}}] \end{aligned}$$

We see that the first term is exactly the same term in the online RL proof after we apply optimism (*e.g.*, [Eq. \(15\)](#)); thus the first term is bounded by the online RL results,

Algorithm 7 Distributional FQI for Hybrid RL

```

1: Input: number of rounds  $K$ , conditional distribution class  $\mathcal{P}$ , offline dataset  $\mathcal{D}^{\text{off}}$ , loss function  $\ell(\hat{y}, y)$ , uniform exploration (UA) flag
2: for episode  $k = 1, 2, \dots, K$  do
3:   for each  $h = H, H - 1, \dots, 1$  do
4:     Recall the loss from Alg. 3 (Eq. \(18\)):

$$L_h^{\text{mle}}(p_h, g, \mathcal{D}_h) = \sum_{i=1}^{|\mathcal{D}_h|} \ell_{\text{mle}}(p_h(x_{h,i}, a_{h,i}), \tau^{\text{D}, *}(g, c_{h,i}, x'_{h,i}))$$

5:     Set  $p_h^k = \arg \min_{p_h \in \mathcal{P}_h} L_h^{\text{mle}}(p_h, p_{h+1}^k, \mathcal{D}_h^{\text{off}} \cup \mathcal{D}_{< k}^{\text{on}})$ .
6:   end for
7:   Let  $\pi^k$  be greedy w.r.t.  $p^k$ :  $\pi_h^k(x) = \arg \min_a p_h^k(x, a)$ .
8:   Gather data  $\mathcal{D}_k^{\text{on}} \leftarrow \text{Alg. 1}(\pi^k, \text{UA flag})$ .
9: end for

```

e.g., [Thms. 7](#) and [8](#). We also see that the second term is exactly the same term in the offline RL proof after applying pessimism. Thus, we can bound the second term by the offline RL results, e.g., [Thm. 10](#). Since we posit the offline dataset has as many samples as the online dataset, the offline bound matches the online one in terms of K . This completes the proof and shows why the bound in hybrid RL is the sum of online and offline RL bounds. \square

Finally, to apply the mle loss to achieve second-order bounds, we naturally extend FQI with DistRL which closely resembles deep DistRL algorithms such as C51 [\[10\]](#). This gives the following new second-order guarantees for hybrid RL.

THEOREM 13. *Under [Asm. 5](#) and $|\mathcal{D}^{\text{off}}| \geq \Omega(K)$, for any $\delta \in (0, 1)$, w.p.a.l. $1 - \delta$, [Alg. 7](#) has the following guarantee: for any comparator policy $\tilde{\pi} \in \Pi_{\mathcal{P}}$,*

$$\begin{aligned} \sum_{k=1}^K (V^{\pi^k} - V^{\tilde{\pi}}) &\leq \tilde{\mathcal{O}}(H^{2.5}(d + C^{\tilde{\pi}})\beta \\ &+ H\sqrt{(\sigma^2(\tilde{\pi})K + \sum_{k=1}^K \sigma^2(\pi^k)) \cdot (d + C^{\tilde{\pi}})\beta}), \end{aligned}$$

where $d = d_{\text{mle}}$ if UA is false, and $d = Ad_{\text{mle}}^V$ if UA is true.

The hybrid second-order bound, being the sum of the second-order bounds for online and offline DistRL ([Thms. 9](#) and [11](#)), contains both the variance of the played policies as well as the variance of the comparator policy. Nevertheless, the hybrid second-order bound still implies a hybrid first-order bound by the same AM-GM argument as in CSC. Thus, this again shows that DistRL yields a notable benefit compared to other losses.

9. DISCUSSION AND CONCLUSION

From the one-step CSC to online, offline and hybrid RL, we see time and time again that the loss function plays a central role in the adaptivity and efficiency of decision making algorithms. The classical squared loss always converges at a slow $\tilde{\mathcal{O}}(1/\sqrt{n})$ rate and cannot adapt to

easier problem instances with heteroskedasticity. The bce loss can serve as a drop-in improvement that yields first-order bounds with a much faster $\tilde{\mathcal{O}}(1/n)$ rate when the optimal cost is small. Switching from conditional-mean learning to conditional-distribution learning, the mle loss can tighten the bounds further with a second-order guarantee, that is bounds that converge at a $\tilde{\mathcal{O}}(1/n)$ rate in near-deterministic settings even if the optimal cost is large. Crucially, these gaps in performance are not merely theoretical as they have been observed many times by the deep RL community [\[21, 11, 31, 7, 47\]](#). The theory outlined herein is very general and can be applied to a wide range of settings including imitation learning [\[23\]](#), model-based [\[26, 66\]](#), risk-sensitive RL [\[63, 67\]](#), and robust bandits [\[38\]](#), and RL [\[13\]](#). Moreover, the principles herein can improve algorithms for post-training large language models [\[29, 1, 68, 77\]](#), learning query optimizers [\[42, 69\]](#) and many more real-world applications. We hope to have not only clearly demonstrated that the loss function choice is important in RL, but also to inspire the reader to seek out opportunities for better loss functions to improve their decision-making algorithms.

10. ACKNOWLEDGMENTS

This material is based upon work supported by a Google PhD Fellowship and grants NSF IIS-1846210, NSF IIS-2154711, NSF CAREER 2339395 and DARPA Learning Network Cyberagents (LANCER). We also thank the editor and anonymous reviewers for useful discussions and feedback.

REFERENCES

- [1] ADLER, B., AGARWAL, N., AITHAL, A., ANH, D. H., BHATTACHARYA, P., BRUNDYN, A., CASPER, J., CATANZARO, B., CLAY, S., COHEN, J. et al. (2024). Nemotron-4 340B Technical Report. *arXiv preprint arXiv:2406.11704*.
- [2] AGARWAL, A., JIANG, N., KAKADE, S. M. and SUN, W. (2019). Reinforcement learning: Theory and algorithms. *CS Dept., UW Seattle, Seattle, WA, USA, Tech. Rep* **32** 96.
- [3] AGARWAL, A., KAKADE, S., KRISHNAMURTHY, A. and SUN, W. (2020). Flambe: Structural complexity and representation learning of low rank mdps. *Advances in neural information processing systems* **33** 20095–20107.
- [4] AGARWAL, A., SONG, Y., SUN, W., WANG, K., WANG, M. and ZHANG, X. (2023). Provable benefits of representational transfer in reinforcement learning. In *The Thirty Sixth Annual Conference on Learning Theory* 2114–2187. PMLR.
- [5] AUDIBERT, J.-Y. and TSYBAKOV, A. B. (2007). Fast learning rates for plug-in classifiers.
- [6] AUER, P., CESA-BIANCHI, N., FREUND, Y. and SCHAPIRE, R. E. (2002). The nonstochastic multiarmed bandit problem. *SIAM journal on computing* **32** 48–77.
- [7] AYOUB, A., WANG, K., LIU, V., ROBERTSON, S., MCINERNY, J., LIANG, D., KALLUS, N. and SZEPESVARI, C. (2024). Switching the Loss Reduces the Cost in Batch Reinforcement Learning. In *Forty-first International Conference on Machine Learning*.

- [8] BALL, P. J., SMITH, L., KOSTRIKOV, I. and LEVINE, S. (2023). Efficient online reinforcement learning with offline data. In *International Conference on Machine Learning* 1577–1594. PMLR.
- [9] BAS-SERRANO, J., CURI, S., KRAUSE, A. and NEU, G. (2021). Logistic Q-learning. In *International conference on artificial intelligence and statistics* 3610–3618. PMLR.
- [10] BELLEMARE, M. G., DABNEY, W. and MUNOS, R. (2017). A distributional perspective on reinforcement learning. In *International conference on machine learning* 449–458. PMLR.
- [11] BELLEMARE, M. G., DABNEY, W. and ROWLAND, M. (2023). *Distributional Reinforcement Learning*. MIT Press <http://www.distributional-rl.org>.
- [12] BELLEMARE, M. G., CANDIDO, S., CASTRO, P. S., GONG, J., MACHADO, M. C., MOITRA, S., PONDA, S. S. and WANG, Z. (2020). Autonomous navigation of stratospheric balloons using reinforcement learning. *Nature* **588** 77–82.
- [13] BENNETT, A., KALLUS, N., OPRESCU, M., SUN, W. and WANG, K. (2024). Efficient and Sharp Off-Policy Evaluation in Robust Markov Decision Processes. *Advances in Neural Information Processing Systems*.
- [14] CHANG, J., WANG, K., KALLUS, N. and SUN, W. (2022). Learning bellman complete representations for offline policy evaluation. In *International Conference on Machine Learning* 2938–2971. PMLR.
- [15] CHEN, J. and JIANG, N. (2019). Information-theoretic considerations in batch reinforcement learning. In *International Conference on Machine Learning* 1042–1051. PMLR.
- [16] CHENG, C.-A., XIE, T., JIANG, N. and AGARWAL, A. (2022). Adversarially trained actor critic for offline reinforcement learning. In *International Conference on Machine Learning* 3852–3878. PMLR.
- [17] DABNEY, W., OSTROVSKI, G., SILVER, D. and MUNOS, R. (2018). Implicit quantile networks for distributional reinforcement learning. In *International conference on machine learning* 1096–1105. PMLR.
- [18] DANN, C., JIANG, N., KRISHNAMURTHY, A., AGARWAL, A., LANGFORD, J. and SCHAPIRE, R. E. (2018). On oracle-efficient pac rl with rich observations. *Advances in neural information processing systems* **31**.
- [19] DANN, C., MANSOUR, Y., MOHRI, M., SEKHARI, A. and SRIDHARAN, K. (2022). Guarantees for epsilon-greedy reinforcement learning with function approximation. In *International conference on machine learning* 4666–4689. PMLR.
- [20] DEVROYE, L. and LUGOSI, G. (2001). *Combinatorial methods in density estimation*. Springer Science & Business Media.
- [21] FAREBROTHER, J., ORBAY, J., VUONG, Q., TAIGA, A. A., CHEBOTAR, Y., XIAO, T., IRPAN, A., LEVINE, S., CASTRO, P. S., FAUST, A., KUMAR, A. and AGARWAL, R. (2024). Stop Regressing: Training Value Functions via Classification for Scalable Deep RL. In *Forty-first International Conference on Machine Learning*.
- [22] FENG, F., YIN, W., AGARWAL, A. and YANG, L. (2021). Provably correct optimization and exploration with non-linear policies. In *International Conference on Machine Learning* 3263–3273. PMLR.
- [23] FOSTER, D. J., BLOCK, A. and MISRA, D. (2024). Is Behavior Cloning All You Need? Understanding Horizon in Imitation Learning. *arXiv preprint arXiv:2407.15007*.
- [24] FOSTER, D. J. and KRISHNAMURTHY, A. (2021). Efficient first-order contextual bandits: Prediction, allocation, and triangular discrimination. *Advances in Neural Information Processing Systems* **34** 18907–18919.
- [25] FOSTER, D., AGARWAL, A., DUDÍK, M., LUO, H. and SCHAPIRE, R. (2018). Practical contextual bandits with regression oracles. In *International Conference on Machine Learning* 1539–1548. PMLR.
- [26] FOSTER, D. J., KAKADE, S. M., QIAN, J. and RAKHLIN, A. (2021). The statistical complexity of interactive decision making. *arXiv preprint arXiv:2112.13487*.
- [27] FOSTER, D. J., RAKHLIN, A., SEKHARI, A. and SRIDHARAN, K. (2022a). On the complexity of adversarial decision making. *Advances in Neural Information Processing Systems* **35** 35404–35417.
- [28] FOSTER, D. J., KRISHNAMURTHY, A., SIMCHI-LEVI, D. and XU, Y. (2022b). Offline Reinforcement Learning: Fundamental Barriers for Value Function Approximation. In *Conference on Learning Theory* 3489–3489. PMLR.
- [29] GAO, Z., CHANG, J. D., ZHAN, W., OERTELL, O., SWAMY, G., BRANTLEY, K., JOACHIMS, T., BAGNELL, J. A., LEE, J. D. and SUN, W. (2024). Rebel: Reinforcement learning via regressing relative rewards. *arXiv preprint arXiv:2404.16767*.
- [30] HU, Y., KALLUS, N. and MAO, X. (2022). Fast rates for contextual linear optimization. *Management Science* **68** 4236–4245.
- [31] IMANI, E. and WHITE, M. (2018). Improving regression performance with distributional losses. In *International conference on machine learning* 2157–2166. PMLR.
- [32] JASON MA, Y., JAYARAMAN, D. and BASTANI, O. (2022). Conservative Offline Distributional Reinforcement Learning. *Advances in neural information processing systems*.
- [33] JIA, Z., QIAN, J., RAKHLIN, A. and WEI, C.-Y. (2024). How Does Variance Shape the Regret in Contextual Bandits? *Advances in Neural Information Processing Systems*.
- [34] JIANG, N. and AGARWAL, A. (2018). Open problem: The dependence of sample complexity lower bounds on planning horizon. In *Conference On Learning Theory* 3395–3398. PMLR.
- [35] JIN, C., LIU, Q. and MIRYOOSEFI, S. (2021). Bellman eluder dimension: New rich classes of rl problems, and sample-efficient algorithms. *Advances in neural information processing systems* **34** 13406–13418.
- [36] JIN, C., YANG, Z., WANG, Z. and JORDAN, M. I. (2020). Provably efficient reinforcement learning with linear function approximation. In *Conference on learning theory* 2137–2143. PMLR.
- [37] KAKADE, S. and LANGFORD, J. (2002). Approximately optimal approximate reinforcement learning. In *Proceedings of the Nineteenth International Conference on Machine Learning* 267–274.
- [38] KALLUS, N., MAO, X., WANG, K. and ZHOU, Z. (2022). Doubly robust distributionally robust off-policy evaluation and learning. In *International Conference on Machine Learning* 10598–10632. PMLR.
- [39] KOLTER, J. (2011). The fixed points of off-policy TD. *Advances in neural information processing systems* **24**.
- [40] KONTOROVICH, A. (2024). Binomial small deviations. Tweet.
- [41] KRISHNAMURTHY, A., AGARWAL, A., HUANG, T.-K., DAUMÉ III, H. and LANGFORD, J. (2019). Active learning for cost-sensitive classification. *Journal of Machine Learning Research* **20** 1–50.
- [42] KRISHNAN, S., YANG, Z., GOLDBERG, K., HELLERSTEIN, J. and STOICA, I. (2018). Learning to optimize join queries with deep reinforcement learning. *arXiv preprint arXiv:1808.03196*.
- [43] KUMAR, A., ZHOU, A., TUCKER, G. and LEVINE, S. (2020). Conservative q-learning for offline reinforcement learning. *Advances in Neural Information Processing Systems* **33** 1179–1191.
- [44] LIU, Q., CHUNG, A., SZEPESVÁRI, C. and JIN, C. (2022). When is partially observable reinforcement learning not scary? In *Conference on Learning Theory* 5175–5220. PMLR.

- [45] LUEDTKE, A. and CHAMBAZ, A. (2017). Faster rates for policy learning. *arXiv preprint arXiv:1704.06431*.
- [46] LYKOURIS, T., SRIDHARAN, K. and TARDOS, É. (2018). Small-loss bounds for online learning with partial information. In *Conference on Learning Theory* 979–986. PMLR.
- [47] MA, Y., JAYARAMAN, D. and BASTANI, O. (2021). Conservative offline distributional reinforcement learning. *Advances in neural information processing systems* **34** 19235–19247.
- [48] MHAMMEDI, Z., FOSTER, D. J. and RAKHLIN, A. (2024). The Power of Resets in Online Reinforcement Learning. *arXiv preprint arXiv:2404.15417*.
- [49] MHAMMEDI, Z., BLOCK, A., FOSTER, D. J. and RAKHLIN, A. (2024). Efficient model-free exploration in low-rank mdps. *Advances in Neural Information Processing Systems* **36**.
- [50] MISRA, D., HENAFF, M., KRISHNAMURTHY, A. and LANGFORD, J. (2020). Kinematic state abstraction and provably efficient rich-observation reinforcement learning. In *International conference on machine learning* 6961–6971. PMLR.
- [51] MNIIH, V., KAVUKCUOGLU, K., SILVER, D., RUSU, A. A., VENESS, J., BELLEMARE, M. G., GRAVES, A., RIEDMILLER, M., FIDJELAND, A. K., OSTROVSKI, G. et al. (2015). Human-level control through deep reinforcement learning. *nature* **518** 529–533.
- [52] MODI, A., CHEN, J., KRISHNAMURTHY, A., JIANG, N. and AGARWAL, A. (2024). Model-free representation learning and exploration in low-rank mdps. *Journal of Machine Learning Research* **25** 1–76.
- [53] MUNOS, R. and SZEPESVÁRI, C. (2008). Finite-Time Bounds for Fitted Value Iteration. *Journal of Machine Learning Research* **9**.
- [54] OUYANG, L., WU, J., JIANG, X., ALMEIDA, D., WAINWRIGHT, C., MISHKIN, P., ZHANG, C., AGARWAL, S., SLAMA, K., RAY, A. et al. (2022). Training language models to follow instructions with human feedback. *Advances in neural information processing systems* **35** 27730–27744.
- [55] PACCHIANO, A. (2024). Second Order Bounds for Contextual Bandits with Function Approximation. *arXiv preprint arXiv:2409.16197*.
- [56] RASHIDINEJAD, P., ZHU, B., MA, C., JIAO, J. and RUSSELL, S. (2021). Bridging offline reinforcement learning and imitation learning: A tale of pessimism. *Advances in Neural Information Processing Systems* **34** 11702–11716.
- [57] RUSSO, D. and VAN ROY, B. (2013). Eluder dimension and the sample complexity of optimistic exploration. *Advances in Neural Information Processing Systems* **26**.
- [58] SONG, Y., ZHOU, Y., SEKHARI, A., BAGNELL, D., KRISHNAMURTHY, A. and SUN, W. (2023). Hybrid RL: Using both offline and online data can make RL efficient. In *The Eleventh International Conference on Learning Representations*.
- [59] TSITSIKLIS, J. and VAN ROY, B. (1996). Analysis of temporal-difference learning with function approximation. *Advances in neural information processing systems* **9**.
- [60] UEHARA, M. and SUN, W. (2022). Pessimistic Model-based Offline Reinforcement Learning under Partial Coverage. In *International Conference on Learning Representations*.
- [61] WAINWRIGHT, M. J. (2019). *High-dimensional statistics: A non-asymptotic viewpoint* **48**. Cambridge university press.
- [62] WANG, R., FOSTER, D. and KAKADE, S. M. (2021). What are the Statistical Limits of Offline RL with Linear Function Approximation? In *International Conference on Learning Representations*.
- [63] WANG, K., KALLUS, N. and SUN, W. (2023). Near-minimax-optimal risk-sensitive reinforcement learning with cvar. In *International Conference on Machine Learning* 35864–35907. PMLR.
- [64] WANG, K., ZHOU, K., WU, R., KALLUS, N. and SUN, W. (2023). The benefits of being distributional: Small-loss bounds for reinforcement learning. *Advances in Neural Information Processing Systems* **36**.
- [65] WANG, K., OERTELL, O., AGARWAL, A., KALLUS, N. and SUN, W. (2024a). More Benefits of Being Distributional: Second-Order Bounds for Reinforcement Learning. *International Conference on Machine Learning*.
- [66] WANG, Z., ZHOU, D., LUI, J. and SUN, W. (2024b). Model-based RL as a Minimalist Approach to Horizon-Free and Second-Order Bounds. *arXiv preprint arXiv:2408.08994*.
- [67] WANG, K., LIANG, D., KALLUS, N. and SUN, W. (2024c). A Reductions Approach to Risk-Sensitive Reinforcement Learning with Optimized Certainty Equivalents. *arXiv preprint arXiv:2403.06323*.
- [68] WANG, K., KIDAMBI, R., SULLIVAN, R., AGARWAL, A., DANN, C., MICHI, A., GELMI, M., LI, Y., GUPTA, R., DUBEY, A. et al. (2024d). Conditional Language Policy: A General Framework for Steerable Multi-Objective Finetuning. *Findings of Empirical Methods in Natural Language Processing*.
- [69] WANG, J., WANG, K., LI, Y., KALLUS, N., TRUMMER, I. and SUN, W. (2024e). JoinGym: An Efficient Join Order Selection Environment. *Reinforcement Learning Journal* **1**.
- [70] WATKINS, C. J. and DAYAN, P. (1992). Q-learning. *Machine learning* **8** 279–292.
- [71] WEISZ, G., AMORTILA, P. and SZEPESVÁRI, C. (2021). Exponential lower bounds for planning in mdps with linearly-realizable optimal action-value functions. In *Algorithmic Learning Theory* 1237–1264. PMLR.
- [72] WU, R., UEHARA, M. and SUN, W. (2023). Distributional offline policy evaluation with predictive error guarantees. In *International Conference on Machine Learning* 37685–37712. PMLR.
- [73] XIE, T., CHENG, C.-A., JIANG, N., MINEIRO, P. and AGARWAL, A. (2021). Bellman-consistent pessimism for offline reinforcement learning. *Advances in neural information processing systems* **34** 6683–6694.
- [74] XIE, T., FOSTER, D. J., BAI, Y., JIANG, N. and KAKADE, S. M. (2023). The Role of Coverage in Online Reinforcement Learning. In *The Eleventh International Conference on Learning Representations*.
- [75] ZHANG, T. (2023). *Mathematical Analysis of Machine Learning Algorithms*. Cambridge University Press. <https://doi.org/10.1017/9781009093057>
- [76] ZHANG, S., LI, H., WANG, M., LIU, M., CHEN, P.-Y., LU, S., LIU, S., MURUGESAN, K. and CHAUDHURY, S. (2023). On the Convergence and Sample Complexity Analysis of Deep Q-Networks with ϵ -Greedy Exploration. In *Thirty-seventh Conference on Neural Information Processing Systems*.
- [77] ZHOU, J. P., WANG, K., CHANG, J., GAO, Z., KALLUS, N., WEINBERGER, K. Q., BRANTLEY, K. and SUN, W. (2025). Q#: Provably Optimal Distributional RL for LLM Post-Training. *arXiv preprint arXiv:2502.20548*.

11. APPENDIX

11.1 Proof of Lemmas in Sec. 5

LEMMA 7. Let $\ell = \ell_{\text{sq}}$ and \mathcal{D}_h be a dataset where the i -th datapoint is collected from π^i , and denote $\mathcal{D} = (\mathcal{D}_1, \dots, \mathcal{D}_H)$. Then under BC (Asm. 3), for any $\delta \in (0, 1)$, let $\beta = 2 \ln(H|\mathcal{F}|/\delta)$ and define

$$(14) \quad \hat{f}^{\text{op}} \in \arg \min_{f \in \mathcal{C}_{\beta}^{\text{sq}}(\mathcal{D})} \min_a f_1(x_1, a).$$

W.p.a.l. $1 - \delta$, we have (a) $\sum_{i=1}^n \mathcal{E}_{\text{sq}}^{\text{RL}}(\pi^i; \hat{f}^{\text{op}}) \leq 2H\beta$, and (b) $\min_a \hat{f}_1^{\text{op}}(x_1, a) \leq V^*$.

PROOF OF LEM. 7. By standard martingale concentration via Freedman's inequality, w.p.a.l. $1 - \delta$, for all f, h , we have

$$(23) \quad \begin{aligned} \sum_{i=1}^n \mathcal{E}_h^{\text{sq}}(f, \pi^i) &\leq \ln(H|\mathcal{F}|/\delta) + L_h^{\text{sq}}(f_h, f_{h+1}, \mathcal{D}_h) \\ &\quad - L_h^{\text{sq}}(\mathcal{T}_h f_{h+1}, f_{h+1}, \mathcal{D}_h). \end{aligned}$$

Let $g_h^f \in \arg \min_{g_h \in \mathcal{G}_h} L_h^*(g_h, f_{h+1}, \mathcal{D}_h)$ denote the empirical risk minimizer, as used in the definition of $\mathcal{C}_{\beta}^*(\mathcal{D})$ (Eq. (13)). Under the BC premise,

$$\begin{aligned} \sum_{i=1}^n \mathcal{E}_h^{\text{sq}}(f, \pi^i) &\leq \ln(H|\mathcal{F}|/\delta) + L_h^{\text{sq}}(f_h, f_{h+1}, \mathcal{D}_h) \\ &\quad - L_h^{\text{sq}}(g_h^f, f_{h+1}, \mathcal{D}_h). \end{aligned}$$

Thus, any $f \in \mathcal{C}_{\beta}^{\text{sq}}(\mathcal{D})$ satisfies $\sum_{i=1}^n \mathcal{E}_h^{\text{sq}}(f, \pi^i) \leq 2\beta$, which proves Claim (a). For Claim (b), we prove that $Q^* \in \mathcal{C}_{\beta}^{\text{sq}}(\mathcal{D})$: by Eq. (23) and non-negativity of \mathcal{E}^{sq} , we have $L_h^{\text{sq}}(\mathcal{T}_h f_{h+1}, f_{h+1}, \mathcal{D}_h) - L_h^{\text{sq}}(g_h^f, f_{h+1}, \mathcal{D}_h) \leq \ln(H|\mathcal{F}|/\delta) = \beta$. Then, setting $f = Q^*$ and applying $Q_h^* = \mathcal{T}_h^* Q_{h+1}^*$ shows that Q^* satisfies the version space condition. Thus, $Q^* \in \mathcal{C}_{\beta}^{\text{sq}}(\mathcal{D})$ and Claim (b) follows by definition of \hat{f}^{op} . \square

The following is a proof for a stronger version of the pigeonhole lemma (Lem. 9). In particular, Lem. 9 follows when $\varepsilon_0 = \frac{1}{N}$.

LEMMA 18. Let $E := \sup_{\nu \in \mathcal{M}, \psi \in \Psi} |\mathbb{E}_p \psi|$. Fix any $N \in \mathbb{N}$, $\psi^{(1)}, \dots, \psi^{(N)} \in \Psi$, and $\nu^{(1)}, \dots, \nu^{(N)} \in \mathcal{M}$. Let β be a constant s.t. $\sum_{i < j} |\mathbb{E}_{\nu^{(i)}}[\psi^{(j)}]|^q \leq \beta^q$ for all $j \in [N]$. Then, $\sum_{j=1}^N |\mathbb{E}_{\nu^{(j)}} \psi^{(j)}| \leq \inf_{\varepsilon_0 \in (0, 1)} \{N\varepsilon_0 + \text{EluDim}_q(\Psi, \mathcal{M}, \varepsilon_0) \cdot (2E + \beta^q \ln(E\varepsilon_0^{-1}))\}$.

PROOF OF LEM. 18. Fix any $q \in \mathbb{N}$; the proof will be for the ℓ_q eluder dimension. We say a distribution $\nu \in \mathcal{M}$ is ε -independent of a subset $\Gamma \subset \mathcal{M}$ if there exists $\psi \in \Psi$ s.t. $|\mathbb{E}_\nu \psi| > \varepsilon$ but also $\sum_{\mu \in \Gamma} (\mathbb{E}_\mu [\psi])^q \leq \varepsilon^q$. Conversely, we say ν is ε -dependent on Γ if for all $\psi \in \Psi$, we have $|\mathbb{E}_\nu \psi| \leq \varepsilon$ or $\sum_{\mu \in \Gamma} (\mathbb{E}_\mu [\psi])^q > \varepsilon^q$. For any $\Gamma \subset \mathcal{M}$ and $\nu \in \mathcal{M}$, we let $N(\nu, \Gamma, \varepsilon_0)$ denote the largest number of

disjoint subsets of Γ that ν is ε_0 -dependent on. We also use the shorthand $\mu^{(<j)} = \{\mu^{(1)}, \mu^{(2)}, \dots, \mu^{(j-1)}\}$.

Claim 1: If $|\mathbb{E}_{\mu^{(j)}}[\psi^{(j)}]| > \varepsilon$, then $N(\mu^{(j)}, \mu^{(<j)}, \varepsilon) \leq \beta^q \varepsilon^{-1}$. By definition of $N := N(\mu^{(j)}, \mu^{(<j)}, \varepsilon)$, there are disjoint subsets $S^{(1)}, \dots, S^{(N)} \subset \mu^{(<j)}$ s.t. each $S^{(i)}$ satisfies $\sum_{\mu \in S^{(i)}} |\mathbb{E}_\mu[\psi^{(j)}]| > \varepsilon$ since $|\mathbb{E}_{\mu^{(j)}}[\psi^{(j)}]| > \varepsilon$ by premise. Thus, summing over all such subsets gives $N\varepsilon < \sum_{i < j} |\mathbb{E}_{\mu^{(i)}}[\psi^{(j)}]|^q \leq \beta^q$, proving Claim 1.

Claim 2 (Pigeonhole): For any ε_0 and any sequence $\mu^{(1)}, \dots, \mu^{(N)} \in \mathcal{M}$, there exists $j \leq N$ such that $N(\mu^{(j)}, \mu^{(<j)}, \varepsilon_0) \geq \lfloor \frac{(N-1)}{\text{EluDim}_q(\Psi, \mathcal{M}, \varepsilon_0)} \rfloor$. Recall that if $\mu^{(1)}, \dots, \mu^{(L)} \subset \mathcal{M}$ satisfies for all $j \in [L]$, $\mu^{(j)}$ is ε_0 -independent of $p^{(<j)}$, then $L \leq \text{EluDim}_q(\Psi, \mathcal{M}, \varepsilon_0)$ by definition. To prove the claim, we maintain $J := \lfloor \frac{(k-1)}{\text{EluDim}_q(\Psi, \mathcal{M}, \varepsilon_0)} \rfloor$ disjoint sequences $S^{(1)}, \dots, S^{(J)} \subset \mu^{(<k)}$ s.t. each $S^{(i)}$ has the property that each element is ε -independent of its predecessors. We initialize $S^{(1)} = \dots = S^{(J)} = \emptyset$ and iteratively add elements $\mu^{(1)}, \dots, \mu^{(N)}$ until $\mu^{(j)}$ is ε_0 -dependent on all these disjoint subsequences, at which point the claim is proven. If there exists a subsequence which $\mu^{(j)}$ is ε_0 -independent of, we add $\mu^{(j)}$ to that subsequence, which preserves the invariant condition. This process indeed terminates since otherwise one subsequence would have more elements than $\text{EluDim}_q(\Psi, \mathcal{M}, \varepsilon_0)$, a contradiction.

Claim 3: For any ε , $\sum_{j=1}^N \mathbb{I}[|\mathbb{E}_{\mu^{(j)}}[\psi^{(j)}]| > \varepsilon] \leq (\beta^q \varepsilon^{-1} + 1) \text{EluDim}_q(\Psi, \mathcal{M}, \varepsilon) + 1$. Let κ denote the left hand sum and so let i_1, \dots, i_κ be all the indices j s.t. $|\mathbb{E}_{\mu^{(j)}}[\psi^{(j)}]| > \varepsilon$. By Claim 2, there exists $j \leq \kappa$ s.t. $\lfloor \frac{(\kappa-1)}{\text{EluDim}_q(\Psi, \mathcal{M}, \varepsilon)} \rfloor \leq L(\mu^{(i_j)}, \mu^{(<i_j)}, \varepsilon)$. Then by Claim 1, this is further upper bounded by $\beta^q \varepsilon^{-1}$. Rearranging proves the claim.

Concluding the proof. For any ε_0 , we have

$$\begin{aligned} \sum_{j=1}^N |\mathbb{E}_{\mu^{(j)}}[\psi^{(j)}]| &= \sum_{j=1}^N \int_0^E \mathbb{I}[|\mathbb{E}_{\mu^{(j)}}[\psi^{(j)}]| > y] dy \\ &\leq N\varepsilon_0 + \sum_{j=1}^N \int_{\varepsilon_0}^E \mathbb{I}[|\mathbb{E}_{\mu^{(j)}}[\psi^{(j)}]| > y] dy \\ &\stackrel{(i)}{\leq} N\varepsilon_0 + \int_{\varepsilon_0}^E \{(\beta^q y^{-1} + 1) \text{EluDim}_q(\Psi, \mathcal{M}, y) + 1\} dy \\ &\stackrel{(ii)}{\leq} N\varepsilon_0 + \int_{\varepsilon_0}^E \{(\beta^q y^{-1} + 1) \text{EluDim}_q(\Psi, \mathcal{M}, \varepsilon_0) + 1\} dy \\ &\stackrel{(iii)}{\leq} N\varepsilon_0 + \text{EluDim}_q(\Psi, \mathcal{M}, \varepsilon_0)(2E + \beta^q \ln(E\varepsilon_0^{-1})), \end{aligned}$$

where (i) is by Claim 3, (ii) is by monotonicity of the eluder dimension, and (iii) is by $\int_{\varepsilon_0}^E y^{-1} = \ln(E\varepsilon_0^{-1})$. \square

11.2 Proofs for Lemmas in Sec. 6

LEMMA 12. Let $\ell = \ell_{\text{bce}}$. Under the same setup as Lem. 7 with \hat{f}^{op} selected from $\mathcal{C}_{\beta}^{\text{bce}}$ instead of $\mathcal{C}_{\beta}^{\text{sq}}$, w.p.a.l. $1 - \delta$, we have (a) $\sum_{i=1}^n \mathcal{E}_{\text{bce}}^{\text{RL}}(\hat{f}^{\text{op}}, \pi^i) \leq 2H\beta$, and (b) $\min_a \hat{f}_1^{\text{op}}(x_1, a) \leq V^*$.

PROOF OF LEM. 12. By Lem. 3 extended on martingale sequences [3], w.p.a.l. $1 - \delta$, for all $f \in \mathcal{F}, h \in [H]$,

$$\sum_{i=1}^n \mathcal{E}_h^{\text{Ber}}(f, \pi^i) \leq \ln(H|\mathcal{F}|/\delta) + \frac{1}{2}L_h^{\text{bce}}(f_h, f_{h+1}, \mathcal{D}_h)$$

$$(24) \quad - \frac{1}{2}L_h^{\text{bce}}(\mathcal{T}_h^* f_{h+1}, f_{h+1}, \mathcal{D}_h).$$

Let $g_h^f := \arg \min_{g_h \in \mathcal{G}_h} L_h(g_h, f_{h+1}, \mathcal{D}_h)$ denote the empirical risk minimizer. Under the BC premise,

$$\sum_{i=1}^n \mathcal{E}_h^{\text{Ber}}(f, \pi^i) \leq \ln(H|\mathcal{F}|/\delta) + \frac{1}{2}L_h^{\text{bce}}(f_h, f_{h+1}, \mathcal{D}_h)$$

$$- \frac{1}{2}L_h^{\text{bce}}(g_h^f, f_{h+1}, \mathcal{D}_h).$$

Thus, any $f \in \mathcal{C}_\beta(\mathcal{D})$ satisfies $\sum_{i=1}^n \mathcal{E}_h^{\text{Ber}}(f, \pi^i) \leq \frac{1}{2}\beta + \ln(H|\mathcal{F}|/\delta) \leq 2\beta$, which proves Claim (a). For Claim (b), we prove that $Q^* \in \mathcal{C}_\beta(\mathcal{D})$. By Eq. (24) and non-negativity of \mathcal{E}^{Ber} , we have $L_h^{\text{bce}}(\mathcal{T}_h^* f_{h+1}, f_{h+1}, \mathcal{D}_h) - L_h^{\text{bce}}(g_h^f, f_{h+1}, \mathcal{D}_h) \leq 2\ln(H|\mathcal{F}|/\delta) = \beta$. Then, setting $f = Q^*$ and noting that $Q_h^* = \mathcal{T}_h^* Q_{h+1}^*$ shows that Q^* satisfies the confidence set condition. Thus, $Q^* \in \mathcal{C}_\beta(\mathcal{D})$ and Claim (b) follows by definition of \hat{f}^{op} . \square

LEMMA 13. Define δ_h^{Ber} that uses \mathcal{T}^π instead of \mathcal{T}^* :

$$\delta_h^{\text{Ber}}(f, \pi, x_h, a_h) := h^2_{\text{Ber}}(f_h(x_h, a_h), \mathcal{T}_h^\pi f_{h+1}(x_h, a_h)).$$

Then, for any f, π, x_h, a_h ,

$$f_h(x_h, a_h) \leq eQ_h^\pi(x_h, a_h) + 77H\delta_{\text{Ber}}^{\text{RL}}(f, \pi).$$

This implies the corollary:

$$\mathbb{E}_\pi[f_h(x_h, a_h)] \lesssim V^\pi + H\delta_{\text{Ber}}^{\text{RL}}(f, \pi).$$

PROOF OF LEM. 13. Fix any f, π . We use the shorthand $\delta_t(x, a) = \delta_t^{\text{Ber}}(f, \pi, x, a)$ to simplify notation. The corollary follows from the main claim via $\mathbb{E}_\pi[Q_h^\pi(x_h, a_h)] \leq V^\pi$, since costs are non-negative. To prove the main claim, we establish the following claim by induction:

$$f_h(x_h, a_h) \leq \sum_{t=h}^H (1 + \frac{1}{H})^{t-h} \mathbb{E}_\pi[c_t +$$

$$(25) \quad 28H\delta_t(x_t, a_t) \mid x_h, a_h].$$

The base case of $h = H+1$ holds since $f_{H+1} = 0$. For the induction step, fix any $h \in [H]$ and suppose that Eq. (25) is true for $h+1$. By Eq. (2) and AM-GM, we have

$$f_h(x_h, a_h) \leq (1 + \frac{1}{H})\mathcal{T}_h^\pi f_{h+1}(x_h, a_h) + 28H\delta_h(x_h, a_h)$$

By definition, $\mathcal{T}_h^\pi f_{h+1}(x_h, a_h) = \mathbb{E}_\pi[c_h + f_{h+1}(x_{h+1}, a_{h+1}) \mid x_h, a_h]$, so we can apply induction hypothesis to f_{h+1} . This proves the inductive claim Eq. (25). Then, we prove the main claim by using the fact $(1 + \frac{1}{H})^H \leq e$. The corollary then follows by $\mathbb{E}_\pi[Q_h^\pi(x_h, a_h)] \leq V^\pi$ which holds due to the non-negativity of costs. \square

LEMMA 14. If $\sum_{k=1}^K (V^{\pi^k} - V^*) \leq c\sqrt{\sum_{k=1}^K V^{\pi^k}} + c^2$, then $\sum_{k=1}^K V^{\pi^k} - V^* \leq c\sqrt{2KV^*} + 3c^2$.

PROOF OF LEM. 14. By AM-GM, the premise implies $\sum_{k=1}^K (V^{\pi^k} - V^*) \leq \frac{1}{2} \sum_{k=1}^K V^{\pi^k} + \frac{3c^2}{2}$, which simplifies to $\sum_{k=1}^K V^{\pi^k} \leq 2KV^* + 3c^2$. Hence, plugging this back into the premise yields the desired bound. \square

LEMMA 15. Let $\ell = \ell_{\text{mle}}$ and \mathcal{D}_h be the same as in Lems. 7 and 12. Then, under Asm. 5, for any $\delta \in (0, 1)$ let $\beta = 2\ln(H|\mathcal{P}|/\delta)$ and define

$$\hat{p}^{\text{op}} \in \arg \min_{p \in \mathcal{C}_\beta^{\text{mle}}(\mathcal{D})} \min_a \bar{p}_1(x_1, a)$$

W.p.a.l. $1 - \delta$, we have (a) $\sum_{i=1}^n \mathcal{E}_{\text{mle}}^{\text{RL}}(\hat{p}^{\text{op}}, \pi^i) \leq 2H\beta$ and (b) $\min_a \bar{p}_1^{\text{op}}(x_1, a) \leq V^*$.

PROOF OF LEM. 15. The proof follows similarly as the bce case of Lem. 12. By Lem. 3 extended on martingale sequences [3], w.p.a.l. $1 - \delta$, for all $p \in \mathcal{P}, h \in [H]$,

$$\sum_{i=1}^n \mathcal{E}_h^{\text{dis}}(p, \pi^i) \leq \ln(H|\mathcal{P}|/\delta) + \frac{1}{2}L_h^{\text{mle}}(p_h, p_{h+1}, \mathcal{D}_h)$$

$$(26) \quad - \frac{1}{2}L_h^{\text{mle}}(\mathcal{T}_h^{\text{D},*} p_{h+1}, p_{h+1}, \mathcal{D}_h).$$

Let $g_h^p := \arg \min_{g_h \in \mathcal{P}_h} L_h^{\text{mle}}(g_h, p_{h+1}, \mathcal{D}_h)$ denote the empirical maximum likelihood estimate. Under the distributional BC premise, we have

$$\sum_{i=1}^n \mathcal{E}_h^{\text{dis}}(p, \pi^i) \leq \ln(H|\mathcal{P}|/\delta)$$

$$+ \frac{1}{2}L_h^{\text{mle}}(p_h, p_{h+1}, \mathcal{D}_h) - \frac{1}{2}L_h^{\text{mle}}(g_h^p, p_{h+1}, \mathcal{D}_h).$$

Thus, any $p \in \mathcal{C}_\beta^{\text{mle}}(\mathcal{D})$ satisfies $\sum_{i=1}^n \mathcal{E}_h^{\text{dis}}(p, \pi^i) \leq \frac{1}{2}\beta + \ln(H|\mathcal{P}|/\delta) \leq 2\beta$, which proves Claim (a). For Claim (b), we prove that $Z^* \in \mathcal{C}_\beta^{\text{mle}}(\mathcal{D})$. By Eq. (26) and non-negativity of \mathcal{E}^{dis} , we have $L_h^{\text{mle}}(\mathcal{T}_h^{\text{D},*} p_{h+1}, p_{h+1}, \mathcal{D}_h) - L_h^{\text{mle}}(g_h^p, p_{h+1}, \mathcal{D}_h) \leq 2\ln(H|\mathcal{P}|/\delta) = \beta$. Then, setting $p = Z^*$ and noting that $Z_h^* = \mathcal{T}_h^{\text{D},*} Z_{h+1}^*$ shows that Z^* satisfies the confidence set condition. Thus, $Z^* \in \mathcal{C}_\beta^{\text{mle}}(\mathcal{D})$ and Claim (b) follows by definition of \hat{p}^{op} . \square

LEMMA 16. Define the state-action analog of δ_h^{dis} :

$$\delta_h^{\text{dis}}(p, \pi, x_h, a_h) := h^2(p_h(x_h, a_h), \mathcal{T}_h^{\text{D},\pi} p_{h+1}(x_h, a_h)),$$

where $\mathcal{T}_h^{\text{D},\pi} p(x, a) \stackrel{D}{=} C(x, a) + p(X', \pi(X'))$ is the distributional Bellman backup of p under π . Then, for any p, π, x_h, a_h , we have

$$\sigma^2(p_h(x_h, a_h)) \leq 2e\sigma^2(Z_h^\pi(x_h, a_h)) + H\delta_{\text{dis}}^{\text{RL}}(p, \pi),$$

where $\sigma^2(Z_h^\pi(x_h, a_h))$ denotes the variance of the random variable $Z_h^\pi(x_h, a_h)$. This also implies the corollary:

$$\mathbb{E}_\pi[\sigma^2(p_h(x_h, a_h))] \lesssim \sigma^2(Z^\pi) + H^2\delta_{\text{dis}}^{\text{RL}}(p, \pi).$$

PROOF OF LEM. 16. Fix any p, π . We use the shorthand $\delta_t(x, a) = \delta_t^{\text{dis}}(p, \pi, x, a)$ to simplify notation. First, note that the corollary follows from the main claim since

the law of total variance (LTV) implies $\mathbb{E}[\sigma^2(Z_h^\pi(x_h, a_h))] \leq \sigma^2(Z^\pi)$, where recall the LTV states: for any random variable X, Y :

$$\sigma^2(Y) = \mathbb{E}[\sigma^2(Y | X)] + \sigma^2(\mathbb{E}[Y | X]).$$

We now establish the main claim.

Step 1. We first show the following claim by induction: for all h ,

$$\begin{aligned} \sigma^2(p_h(x_h, a_h)) &\leq \sum_{t=h}^H (1 + \frac{1}{H})^{t-h} \mathbb{E}_\pi[8H\delta_t(x_t, a_t) \\ (27) \quad &2\sigma^2(c_t + \bar{p}_{t+1}(x_{t+1}, \pi(x_{t+1}))) | x_h, a_h] \end{aligned}$$

The base case $h = H + 1$ is true since $\sigma^2(p_{H+1}) = 0$. For the induction step, fix any $h \in [H]$ and suppose that the induction hypothesis (IH; Eq. (27)) is true for $h + 1$.

By our second-order lemma for variance (Eq. (10)),

$$\begin{aligned} \sigma^2(p_h(x_h, a_h)) &\leq (1 + \frac{1}{H})\sigma^2(\mathcal{T}_h^{\text{D}, \pi} p_{h+1}(x_h, a_h)) \\ &\quad + 8H\delta_h(x_h, a_h). \end{aligned}$$

Then, we use LTV to condition on c_h, x_{h+1} (*i.e.*, the outer mean/variance are w.r.t. c_h, x_{h+1} , the inner mean/variance are w.r.t. p_{h+1}): $\sigma^2(\mathcal{T}_h^{\text{D}, \pi} p_{h+1}(x_h, a_h))$ is equal to

$$\begin{aligned} &\mathbb{E}[\sigma^2(p_{h+1}(x_{h+1}, \pi(x_{h+1})) | c_h, x_{h+1})] \\ &\quad + \sigma^2(c_h + \bar{p}_{h+1}(x_{h+1}, \pi(x_{h+1}))). \end{aligned}$$

We bound the first term by the IH, which completes the proof for Eq. (27).

Step 2. By the above claim and $(1 + \frac{1}{H})^H \leq e$, we have

$$\begin{aligned} \sigma^2(p_h(x_h, a_h)) &\leq 8H\delta_{\text{dis}}^{\text{RL}}(p, \pi) + \\ &2e \sum_{t=h}^H \mathbb{E}_\pi[\sigma^2(c_t + \bar{p}_{t+1}(x_{t+1}, \pi(x_{t+1}))) | x_h, a_h]. \end{aligned}$$

Step 3. Lastly, it suffices to convert the above variance term to $\sigma^2(c_t + V_{t+1}^\pi(x_{t+1}))$, since $\sigma^2(Z_h^\pi(x_h, a_h)) = \sum_{t=h}^H \mathbb{E}_\pi[\sigma^2(c_t + V_{t+1}^\pi(x_{t+1})) | x_h, a_h]$ by LTV. To perform this switch in variance, observe that:

$$(28) \quad |\bar{p}_h(x_h, \pi(x_h)) - V^\pi(x_h)| \lesssim \sum_{t=h}^H \mathbb{E}_\pi[\sqrt{\delta_t(x_t, a_t)}],$$

by the PDL and the second-order lemma (Lem. 1). Also, recall that $\sigma^2(X) \leq 2\sigma^2(Y) + 2\sigma^2(X - Y)$. Thus, we have

$$\begin{aligned} &\sigma^2(c_t + \bar{p}_{t+1}(x_{t+1}, \pi(x_{t+1}))) \\ &\leq 2\sigma^2(c_t + V_{t+1}^\pi(x_{t+1})) \\ &\quad + 2\sigma^2(\bar{p}_{t+1}(x_{t+1}, \pi(x_{t+1})) - V_{t+1}^\pi(x_{t+1})) \\ &\leq 2\sigma^2(c_t + V_{t+1}^\pi(x_{t+1})) + H \sum_{t=h}^H \mathbb{E}_\pi[\delta_t(x_t, a_t)], \end{aligned}$$

where the last inequality used Eq. (28) and Cauchy-Schwarz. Thus we have shown that

$$\begin{aligned} \sigma^2(p_h(x_h, a_h)) &\lesssim H^2\delta_{\text{dis}}^{\text{RL}}(p, \pi) + \\ &4e \sum_{t=h}^H \mathbb{E}_\pi[\sigma^2(c_t + \bar{p}_{t+1}(x_{t+1}, \pi(x_{t+1}))) | x_h, a_h]. \end{aligned}$$

□