Corruption-Robust Offline Reinforcement Learning

Xuezhou Zhang
Princeton University

Yiding Chen
UW Madison

Jerry Zhu
UW Madison

Wen Sun
Cornell University

Abstract

We study the adversarial robustness in offline reinforcement learning. Given a batch dataset consisting of tuples (s, a, r, s'), an adversary is allowed to arbitrarily modify ε fraction of the tuples. From the corrupted dataset the learner aims to robustly identify a near-optimal policy. We first show that a worst-case $\Omega(Hd\varepsilon)$ optimality gap is unavoidable in linear MDP of dimension d, even if the adversary only corrupts the reward element in a tuple. This contrasts with dimension-free results in robust supervised learning and best-known lower-bound in the online RL setting with corruption. Next, we propose robust variants of the Least-Square Value Iteration (LSVI) algorithm utilizing robust supervised learning oracles, which achieve near-matching performances in cases both with and without global data coverage. The algorithm requires the knowledge of ε to design the pessimism bonus in the no-coverage case. Surprisingly, the knowledge of ε is necessary, as we show that being adaptive to unknown ε is impossible. This again contrasts with recent results on corruption-robust online RL and implies that corruption-robust offline RL is a strictly harder problem.

1 Introduction

Offline Reinforcement Learning (RL) (Lange et al., 2012; Levine et al., 2020) has received increasing attention recently due to its appealing property of avoiding online experimentation and making use of offline historical data. In applications such as assistive medical diagnosis and autonomous driving, historical data is abundant and keeps getting generated by high-performing policies (from human doctors/drivers). However, it is often unethical or expensive to allow an online RL algorithm to freely experiment with potentially sub-optimal policies, as often human lives are at stake. Offline RL provides a powerful framework aiming to find a good policy based on historical data alone. Exciting advances have been made in designing stable and high-performing empirical offline RL algorithms (Fujimoto et al., 2019; Laroche et al., 2019; Wu et al., 2019; Kumar et al., 2019, 2020; Agarwal et al., 2020; Kidambi et al., 2020; Siegel et al., 2020; Liu et al., 2020; Yang and Nachum, 2021; Yu et al., 2021). On the theoretical front, recent works have proposed efficient algorithms with theoretical guarantees, based on the principle of *pessimism in face of uncertainty* (Liu et al., 2020; Buckman et al., 2020; Yu et al., 2020; Jin et al., 2020c; Rashidinejad et al., 2021), or variance reduction (Yin et al., 2020, 2021).

In this work, however, we investigate a different aspect of the offline RL framework, namely the statistical robustness in the presence of data corruption. Data corruption is one of the main security threats against modern ML systems: autonomous vehicles can misread traffic signs contaminated by adversarial stickers (Eykholt et al., 2018); chatbots were misguided by tweeter users to make misogynistic and racist remarks (Neff, 2016); recommendation systems are fooled by fake reviews/comments to produce incorrect rankings. Despite the many vulnerabilities, robustness against data corruption has not been extensively studied in RL until recently. To the best of our knowledge, *all* prior works on corruption-robust RL study the online RL setting. As direct extensions to the setting of adversarial bandits, earlier works focus on designing robust algorithms in *fully adversarial* environments, i.e. the reward functions at all rounds are adversarially generated, and show that $O(\sqrt{T})$ regret is achievable (Even-Dar et al., 2009; Neu et al., 2010, 2012; Zimin and Neu, 2013; Rosenberg and Mansour, 2019; Jin et al., 2020a). While such setting might appear certain game-theoretical situations, in most practical scenarios, such as the ones described above, only a small fraction of the data are actually adversarial while the majority of the data are benign.

Proceedings of the 25[th] International Conference on Artificial Intelligence and Statistics (AISTATS) 2022, Valencia, Spain. PMLR: Volume 151.

Recent works start to study the *Huber's contamination* setting (Lykouris et al., 2019; Chen et al., 2021), where both rewards and transitions can be contaminated but only in ε fraction of all episodes. This setting turns out to be significantly harder, and both works can only tolerate at most $\varepsilon \leq O(1/\sqrt{T})$ fraction of corruptions even against oblivious adversaries. Zhang et al. (2021) recently proposes the first online RL algorithm that can be robust against a constant fraction (i.e. $\varepsilon \geq \Omega(1)$) of adaptive corruption on both rewards and transitions while being agnostic to the value of ε, albeit requiring the help of an exploration policy with finite relative condition number.

In this work, we extend the study of robust RL to the offline setting. Following (Lykouris et al., 2019; Chen et al., 2021; Zhang et al., 2021), we study the *Huber's contamination model* in offline reinforcement learning, formally defined in Assumption 2.2. Huber's contamination model is a classic model for studying sparse data contamination, and is widely used in the traditional literature of robust statistics (Huber et al., 1967). We refer interesting readers to a comprehensive survey (Diakonikolas and Kane, 2019) of recent advances along these directions. Motivated by these prior works, in this paper we ask the following question:

Given an offline RL dataset with ε-fraction of corrupted data, what is the information-theoretic limit of robust identification of the optimal policy?

Towards answering this question, we summarize the following contributions of this work:

1. We provide the formal definition of ε-contamination model in offline RL, and establish an information-theoretical lower-bound of $\Omega(Hd\varepsilon)$ in the setting of linear MDP with dimension d.
2. We design a robust variant of the Least-Square Value Iteration (LSVI) algorithm utilizing robust supervised learning oracles with a novel pessimism bonus term, and show that it achieves near-optimal performance in cases with (Theorem 3.2) or without global data coverage (Theorem 3.3).
3. In the without coverage case, we establish a sufficient condition for learning based on the relative condition number with respect to any comparator policy — not necessary the optimal one. When specialized to offline RL without corruption, our partial coverage assumption is much weaker than the full coverage assumption in (Jin et al., 2020c) for linear MDP.
4. In contrast to (Zhang et al., 2021), we show that agnostic learning, i.e. learning without the knowledge of ε, is generally impossible in the offline RL setting, establishing a separation in hardness between online and offline RL in face of data corruption.

While our paper's main contributions are on corruption robust offline RL, it is worth noting when specialized to the clean offline RL setting, i.e., $\varepsilon = 0$, our work also gives two improved results: (1) under the linear MDP setting, we achieve an optimality gap with respect to any comparator policy (not necessarily the optimal one) in the order of $O(d^{3/2}/\sqrt{N})$ with N being the number of offline samples, saving a $\sqrt{d}$ factor over previously best-known results. (2) our analysis works for the setting where offline data only has partial coverage which is formalized using the concept of relative condition number with respect to the comparator policy[1].

2 Preliminaries

To begin with, let us formally introduce the episodic linear MDP setup we will be working with, the data collection and contamination protocol, as well as the robust linear regression oracle.

Environment. We consider an episodic finite-horizon Markov decision process (MDP), $\mathcal{M}(\mathcal{S}, \mathcal{A}, P, R, H, \mu_0)$, where $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action space, $P : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ is the transition function, such that $P(\cdot|s, a)$ gives the distribution over the next state if action a is taken from state s, $R : \mathcal{S} \times \mathcal{A} \to \Delta(\mathbb{R})$ is a stochastic and potentially unbounded reward function, H is the time horizon, and $\mu_0 \in \Delta_{\mathcal{S}}$ is an initial state distribution. The value functions $V_h^\pi : \mathcal{S} \to \mathbb{R}$ is the expected sum of future rewards, starting at time h in state s and executing π, i.e. $V_h^\pi(s) := \mathbb{E}\left[\sum_{t=h}^H R(s_t, a_t)|\pi, s_0 = s\right]$, where the expectation is taken with respect to the randomness of the policy and environment $\mathcal{M}$. Similarly, the *state-action* value function $Q_h^\pi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is defined as $Q_h^\pi(s, a) := \mathbb{E}\left[\sum_{t=h}^H R(s_t, a_t)|\pi, s_0 = s, a_0 = a\right]$. We use π_h^*, Q_h^*, V_h^* to denote the optimal policy, Q-function and value function, respectively. For any function $f : \mathcal{S} \to \mathbb{R}$, we define the Bellman operator as

$$(\mathbb{B}f)(s, a) = \mathbb{E}_{s' \sim P(\cdot|s,a)}[R(s, a) + f(s')]. \quad (1)$$

We then have the Bellman equation

$$V_h^\pi(s) = \langle Q_h^\pi(s, \cdot), \pi_h(\cdot|s)\rangle_{\mathcal{A}},\ Q_h^\pi(s, a) = (\mathbb{B}V_{h+1}^\pi)(s, a)$$

and the Bellman optimality equation

$$V_h^*(s) = \max_a Q_h^*(s, a), \quad Q_h^*(s, a) = (\mathbb{B}V_{h+1}^*)(s, a)$$

We define the averaged state-action distribution d^π of a policy π: $d^\pi(s, a) := \frac{1}{H}\sum_{h=1}^H \mathbb{P}^\pi(s_t = s, a_t =$

[1] Contemporary to ours, Jin et al. (2020c) added a new Corollary 4.5 in the latest arXiv version of their paper that matches with our results.

$a|s_0 \sim \mu_0)$. We aim to learn a policy that maximizes the expected cumulative reward and thus define the performance metric as the suboptimality of the learned policy π compared to a *comparator policy* $\tilde{\pi}$:

$$\text{SubOpt}(\pi, \tilde{\pi}) = \mathbb{E}_{s\sim\mu_0}[V_1^{\tilde{\pi}}(s) - V_1^{\pi}(s)]. \quad (2)$$

Notice that $\tilde{\pi}$ doesn't necessarily have to be the optimal policy π^*, in contrast to most recent results in pessimistic offline RL, such as (Jin et al., 2020c; Buckman et al., 2020).

For the majority of this work, we focus on the linear MDP setting (Yang and Wang, 2019; Jin et al., 2020b).

Assumption 2.1 (Linear MDP)**.** *There exists a known feature map $\phi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$, d unknown signed measures $\mu = (\mu^{(1)}, ..., \mu^{(d)})$ over $\mathcal{S}$ and an unknown vector $\theta \in \mathbb{R}^d$, such that for all $(s, a, s') \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}$,*

$$P(s'|s, a) = \phi(s, a)^\top \mu(s'),\ R(s, a) = \phi(s, a)^\top \theta + \omega$$

where ω is a zero-mean and σ^2-subgaussian distribution. Here we also assume that the parameters are bounded, i.e.$\|\phi(s, a)\| \leq 1$, $\mathbb{E}[R(s, a)] \in [0, 1]$ for all $(s, a) \in \mathcal{S}\times\mathcal{A}$ and $\max(\|\mu(\mathcal{S})\|, \|\theta\|) \leq \sqrt{d}$.

Clean Data Collection. We consider the offline setting, where a clean dataset $\tilde{D} = \{(\tilde{s}_i, \tilde{a}_i, \tilde{r}_i, \tilde{s}'_i)\}_{i=1:N}$ of transitions is collected a priori by an unknown experimenter. In this work, we assume the stochasticity of the clean data collecting process, i.e. there exists an offline state-action distribution $\nu \in \Delta(\mathcal{S} \times \mathcal{A})$, s.t. $(\tilde{s}_i, \tilde{a}_i) \sim \nu(s, a)$, $\tilde{r}_i \sim R(\tilde{s}_i, \tilde{a}_i)$ and $\tilde{s}'_i \sim P(\tilde{s}_i, \tilde{a}_i)$. When there is no corruption, $\tilde{D}$ will be observed by the learner. However, in this work, we study the setting where the data is contaminated by an adversary before revealed to the learner.

Contamination model. We define an adversarial model that can be viewed as a direct extension to the ε-*contamination model* studied in the traditional robust statistics literature.

Assumption 2.2 (ε-Contamination in offline RL)**.** *Given $\varepsilon \in [0, 1]$ and a set of clean tuples $\tilde{D} = \{(\tilde{s}_i, \tilde{a}_i, \tilde{r}_i, \tilde{s}'_i)\}_{i=1:N}$, the adversary is allowed to inspect the tuples and replace any εN of them with arbitrary transition tuples $(s, a, r, s') \in \mathcal{S}\times\mathcal{A}\times\mathbb{R}\times\mathcal{S}$. The resulting set of transitions is then revealed to the learner. We will call such a set of samples ε-corrupted, and denote the contaminated dataset as $D = \{(s_i, a_i, r_i, s'_i)\}_{i=1:N}$. In other words, there are at most εN number of indices i, on which $(\tilde{s}_i, \tilde{a}_i, \tilde{r}_i, \tilde{s}'_i) \neq (s_i, a_i, r_i, s'_i)$.*

Under ε-contamination, we assume access to a robust linear regression oracle.

Assumption 2.3 (Robust least-square oracle (RLS))**.** *Given a set of ε-contaminated samples $S = \{(x_i, y_i)\}_{1:N}$, where the clean data is generated as: $\tilde{x}_i \sim \nu$, $P(\|x\| \leq 1) = 1$, $\tilde{y}_i = \tilde{x}_i\top w^* + \gamma_i$, where γ_i's are subgaussian noise with zero-mean and γ^2-variance. Then, a robust least-square oracle returns an estimator $\hat{w}$, such that*

1. *If $\mathbb{E}_\nu[xx^\top] \succeq \xi$, then with probability at least $1 - \delta$, $\|\hat{w} - w^*\|_2 \leq c_1(\delta) \cdot \left(\sqrt{\frac{\gamma^2 poly(d)}{\xi^2 N}} + \frac{\gamma}{\xi}\varepsilon\right)$*
2. *With probability at least $1 - \delta$, $\mathbb{E}_\nu\left(\|x^\top(\hat{w} - w^*)\|_2^2\right) \leq c_2(\delta) \cdot \left(\frac{\gamma^2 poly(d)}{N} + \gamma^2\varepsilon\right)$*

where c_1 and c_2 hide absolute constants and polylog$(1/\delta)$.

Such guarantees are common in the robust statistics literature, see e.g. (Bakshi and Prasad, 2020; Pensia et al., 2020; Klivans et al., 2018). In particular, in the simpler setting of bounded reward, i.e. $r_i \in [0, 1]$ for all i, Regular Least Square (RLS) already satisfies Assumption 2.3 with poly$d = O(d)$, see e.g. Appendix F of (Lykouris et al., 2019). We note that while we focus on oracles with such guarantees, our algorithm and analysis are modular and allow one to easily plug in oracles with stronger or weaker guarantees.

3 Algorithms and Main Results

In this work, we focus on a Robust variant of Least-Squares Value Iteration (LSVI)-style algorithms (Jin et al., 2020c), which directly calls a robust least-square oracle to estimate the Bellman operator $\hat{\mathbb{B}}\hat{V}_h(s, a)$. Optionally, it may also subtract a pessimistic bonus $\Gamma_h(s, a)$ during the Bellman update. A template of such an algorithm is defined in Algorithm 1. In section 3.2 and 3.3, we present two variants of the LSVI algorithm designed for two different settings, depending on whether the data has full coverage over the whole state-action space or not. However, before that, we first present an algorithm-independent minimax lower-bound that illustrates the hardness of the robust learning problem in offline RL, in contrast to classic results in statistical estimation and supervised learning.

3.1 Minimax Lower-bound

Theorem 3.1 (Minimax Lower bound)**.** *Under assumptions 2.1 (linear MDP) and 2.2 (ε-contamination), for any fixed data-collecting distribution ν, no algorithm $L : (\mathcal{S}\times\mathcal{A}\times\mathbb{R}\times\mathcal{A})^N \to \Pi$ can find a better than $O(dH\varepsilon)$-optimal policy with probability more than $1/4$ on all MDPs. Specifically,*

$$\min_{L,\nu} \max_{\mathcal{M}, f_c} SubOpt(\hat{\pi}, \pi^*) = \Omega\left(dH\varepsilon\right) \quad (3)$$

where f_c denotes an ε-contamination strategy that corrupts the data based on the MDP $\mathcal{M}$ and clean data $\tilde{D}$ and returns a contaminated dataset, and L denotes an algorithm that takes the contaminated dataset and return a policy $\hat{\pi}$, i.e. $\hat{\pi} = L(f_c(\mathcal{M}, \tilde{D}))$.

The detailed proof is presented in appendix B, but the high-level idea is simple. Consider the tabular MDP setting which is a special case of linear MDP with $d = SA$. For any data generating distribution ν, by the pigeonhole principle, there must exists a least-sampled (s, a) pair, for which $\nu(s, a) \leq 1/SA$. If the adversary *concentrate* all its attack budget on this least sampled (s, a) pair, it can perturb the empirical reward on this (s, a) pair to be as much as $\hat{r}(s, a) = r(s, a) + SA\varepsilon$. Further more, assume that there exists another (s^*, a^*) such that $r(s^*, a^*) = r(s, a) + SA\varepsilon/2$. Then, the learner has no way to tell if truly $r(s, a) > r(s^*, a^*)$ (i.e., the learner believes what she observes and believes there is no contamination) or if the data is contaminated and in fact $r(s, a) < r(s^*, a^*)$. Either could be true and whichever alternative the learner chooses to believe, it will suffer at least $SAH\varepsilon/2$ optimality gap in one of the two scenarios.

Remark 3.1 (dimension scaling)**.** Theorem 3.1 says that even if the algorithm has control over the data collecting distribution ν (without knowing $\mathcal{M}$ a priori), it can still do no better than $\Omega(dH\varepsilon)$ in the worst-case, which implies that robustness is fundamentally impossible in high-dimensional problems where $d \gtrsim 1/\varepsilon$. This is in sharp contrast to the classic results in the robust statistics literature, where estimation errors are found to not scale with the problem dimension, in settings such as robust mean estimation (Diakonikolas et al., 2016; Lai et al., 2016) and robust supervised learning (Charikar et al., 2017; Diakonikolas et al., 2019). From the construction we can see that the dimension scaling appears fundamentally due to a *multi-task learning* effect: the learner must perform SA separate reward mean estimation problems for each (s, a) pair, while the data is provided as a mixture for all these tasks. As a result, the adversary can concentrate on one particular task, raising the contamination level to effectively $d\varepsilon$.

Remark 3.2 (Offline vs. Online RL)**.** We note that the construction in Theorem 3.1 remains valid even if the adversary only contaminates the rewards, and if the adversary is oblivious and perform the contamination based only on the data generating distribution ν rather than the instantiated dataset $\tilde{D}$. In contrast, the best-known lower-bound for robust online RL is $\Omega(H\varepsilon)$ (Zhang et al., 2021). It remains unknown whether $\Omega(H\varepsilon)$ is tight, as no algorithm yet can achieve a matching upper-bound without additional information. We will come back to this discussion in section 3.3.

In what follows, we show that the above lower-bound is tight in both d and ε, by presenting two upper-bound results nearly matching the lower-bound.

Algorithm 1 Robust Least-Square Value Iteration (R-LSVI)

1: Input: Dataset $D = \{(s_i, a_i, r_i, s_i')\}_{1:N}$; pessimism bonus $\Gamma_h(s, a) \geq 0$, robust least-squares Oracle: $RLS(\cdot)$.
2: Split the dataset randomly into H subset: $D_h = \{(s_i^h, a_i^h, r_i^h, s_i'^h)\}_{1:(N/H)}$, for $h \in [H]$.
3: Initialization: Set $\hat{V}_{H+1}(s) \leftarrow 0$.
4: **for** step $h = H, H-1, \ldots, 1$ **do**
5: Set $\hat{w}_h \leftarrow RLS\left(\left\{(\phi(s_i^h, a_i^h), y_i^h)\right\}_{i \in D_h}\right)$, where $y_i^h = r_i^h + \hat{V}_{h+1}(s_i'^h)$.
6: Set $\hat{Q}_h(s, a) \leftarrow \phi(s, a)^\top \hat{w}_h - \Gamma_h(s, a)$, clipped within $[0, H - h + 1]$.
7: Set $\hat{\pi}_h(a|s) \leftarrow \arg\max_a \hat{Q}_h(s, a)$ and $\hat{V}_h(s) \leftarrow \max_a \hat{Q}_h(s, a)$.
8: Output: $\{\hat{\pi}_h\}_{h=1}^H$.

3.2 Robust Learning with Data Coverage

To begin with, we study the simple setting where the offline data has sufficient coverage over the whole state-action distribution. This is often considered as a strong assumption. However, results in this setting will establish meaningful comparison to the above lower-bound and the no-coverage results later. In the context of linear MDP, we say that a data generating distribution has coverage if it satisfies the following assumption.

Assumption 3.1 (Uniform Coverage)**.** *Under assumption 2.1, define $\Sigma_\nu := \mathbb{E}_\nu[\phi(s, a)\phi(s, a)^\top]$ as the covariance matrix of ν. We say that the data generating distribution ξ-covers the state-action space for $\xi > 0$, if $\Sigma_\nu \succeq \xi I$ i.e. the smallest eigenvalue of Σ_ν is strictly positive and at least ξ.*

Under such an assumption, we show that the R-LSVI without pessimism bonus can already be robust to data contamination.

Theorem 3.2 (Robust Learning under ξ-Coverage)**.** *Under assumption 2.1, 2.2 and 3.1, for any $\xi, \varepsilon > 0$, given a dataset of size N, Algorithm 1 with bonus $\Gamma_h(s, a) = 0$ achieves*

$$SubOpt(\hat{\pi}, \pi^*) \leq \tilde{O}\left(\sqrt{\frac{(\sigma + H)^2 H^3 poly(d)}{\xi^2 N}} + \frac{(\sigma + H)H^2}{\xi}\varepsilon\right) \quad (4)$$

with probability at least $1 - \delta$.

The proof of Theorem 3.2 follows readily from the standard analysis of approximated value iterations and rely on the following classic result connecting the Bellman error to the suboptimality of the learned policy, see e.g. Section 2.3 of (Jiang, 2020).

Lemma 3.1 (Optimality gap of VI)**.** *Under assumption 2.1, Algorithm 1 with $\Gamma_h(s,a) = 0$ satisfies*

$$\begin{aligned} SubOpt(\hat{\pi}, \pi^*) &\leq 2H \max_{s,a,h} |\hat{Q}_h(s,a) - (\mathbb{B}_h \hat{V}_{h+1})(s,a)| \\ &\leq 2H \max_{s,a,h} \|\phi(s,a)\|_2 \cdot \|\hat{w}_h - w_h^*\|_2 \quad (5) \end{aligned}$$

where $w_h^* := \theta + \int_{\mathcal{S}} \hat{V}_{h+1}(s')\mu_h(s')ds'$ is the best linear predictor.

The result then follows immediately using property 1 of the robust least-square oracle and the fact that $\mathbb{E}[((r(s,a) + \hat{V}(s')) - (\mathbb{B}_h\hat{V})(s,a))^2|s,a] \leq (\sigma + H)^2$ (Lemma A.2).

Remark 3.3 (Data Splitting and tighter d-dependency)**.** The data splitting in step 2 of Algorithm 1 is mainly for the sake of theoretical analysis and is not required for practical implementations. Nevertheless, it directly contributes to our tighter bounds. Specifically, the data splitting makes $\hat{V}_{h+1}$, which is learned based on D_{h+1}, independent from D_h, at the cost of an additional H multiplicative factor. In contrast, the typical covering argument used in online RL will introduce another $O(d^{1/2})$ multiplicative factor, and naively applying it to the offline RL setting will make the finally sample complexity scales as $O(d^{3/2})$, see e.g. Corollary 4.5 of (Jin et al., 2020c). Our result above, when specialized to offline RL without corruption (i.e., $\varepsilon = 0$), achieves the following results.

Corollary 3.1 (Uncorrupted Learning under ξ-Coverage)**.** *Under assumption 2.1 and 3.1, for any $\xi > 0$, given a clean dataset of size N, with bonus $\Gamma_h(s,a) = 0$ and ridge regression with regularizer coefficient $\lambda = 1$ as the RLS solver, Algorithm 1 achieves with probability at least $1-\delta$*

$$SubOpt(\hat{\pi}, \pi^*) \leq \tilde{O}\left(\frac{H^3 d}{\xi\sqrt{N}}\right). \quad (6)$$

Remark 3.4 (Tolerable ε)**.** Notice that Theorem 3.2 requires $\varepsilon \leq \xi$ to provide a non-vacuous bound. This is because if $\varepsilon > \xi$, then similar to the lower-bound construction in Theorem 3.1, the adversary can corrupt all the data along the eigenvector direction corresponding to the smallest eigenvalue, in which case the empically estimated reward along that direction can be arbitrarily far away from the true reward even with a robust mean estimator, and thus the estimation error becomes vacuous.

Remark 3.5 (Unimprovable gap)**.** Notice in contrast to classic RL results, Theorem 3.2 implies that in the presence of data contamination, there exists an unimprovable optimality gap $(\sigma+H)H^2\varepsilon/\xi$ for the proposed algorithm, even if the learner has access to infinite data. Also note that because $\|\phi(s,a)\| \leq 1$, ξ is at most $1/d$. This implies that asymptotically, $V^* - V^{\hat{\pi}} \leq O(H^3 d\varepsilon)$ when ξ is on the order of $1/d$, matching the lower-bound upto H factors.

Remark 3.6 (Agnosticity to problem parameters)**.** It is worth noting that in theorem 3.2, the algorithm does not require the knowledge of ε or ξ, and thus works in the agnostic setting where these parameters are not available to the learner (given that the robust least-square oracle is agnostic). In other words, the algorithm and the bound are adaptive to both ε and ξ. This point will be revisited in the next section.

3.3 Robust Learning without Coverage

Next, we consider the harder setting where assumption 3.1 does not hold, as often in practice, the offline data will not cover the whole state-action space. Instead, we provide a much weaker sufficient condition under which offline RL is possible.

Assumption 3.2 (relative condition number)**.** *For any given comparator policy $\tilde{\pi}$, under assumption 2.1 and 2.2, define the relative condition number as*

$$\kappa = \sup_w \frac{w^\top \tilde{\Sigma} w}{w^\top \Sigma_\nu w} \quad (7)$$

where $\tilde{\Sigma}$ denotes $\Sigma_{d^{\tilde{\pi}}}$ and we take the convention that $\frac{0}{0} = 0$. We assume that $\kappa < \infty$.

The relative condition number is recently introduced in the policy gradient literature (Agarwal et al., 2019; Zhang et al., 2021). Intuitively, the relative condition number measures the worst-case density ratio between the occupancy distribution of comparator policy and the data generating distribution. For example, in a tabular MDP, $\kappa = \max_{s,a} \frac{d^{\tilde{\pi}}(s,a)}{\nu(s,a)}$. Here, we show that a finite relative condition number with respect to an *arbitrary* comparator policy is already sufficient for offline RL, for both clean and contaminated setting.

Without data coverage, we now rely on pessimism to retain reasonable behavior. However, the challenge, in this case, is to design a valid confidence bonus using only the corrupted data. We now present our constructed pessimism bonus that allows Algorithm 1 to handle ε-corruption, albeit requiring the knowledge of ε.

Theorem 3.3 (Robust Learning without Coverage)**.** *Under assumption 2.1, 2.2 and 3.2, with $\varepsilon > 0$, given*

any comparator policy $\tilde{\pi}$ with $\kappa < \infty$, define the ε-robust empirical covariance as

$$\Lambda_h = \frac{3}{5}\left(\frac{H}{N}\sum_{i\in\mathbb{D}_h}\phi(s_i^h,a_i^h)\phi(s_i^h,a_i^h)^\top + (\varepsilon+\lambda)\cdot I\right), \quad \lambda = c'\cdot dH\log(N/\delta)/N \tag{8}$$

where D_h denotes the data for step h and c' is an absolute constant. Then, Algorithm 1 with pessimism bonus

$$\Gamma_h(s,a) = \left(\frac{(\sigma+H)\sqrt{H}poly(d)}{\sqrt{N}} + ((\sigma+H)+2H\sqrt{d})\sqrt{\varepsilon} + \sqrt{d\lambda}\right)\sqrt{c_2(\delta/H)}\|\phi(s,a)\|_{\Lambda_h^{-1}} \tag{9}$$

will with probability at least $1-\delta$ achieve

$$SubOpt(\hat{\pi},\tilde{\pi}) \le \tilde{O}\left(\frac{(\sigma+H)\sqrt{H^3\kappa}poly(d)}{\sqrt{N}} + ((\sigma+H)H + H^2\sqrt{d})\sqrt{d\kappa\varepsilon}\right) \tag{10}$$

Remark 3.7 (Arbitrary comparator policy)**.** Notice that in comparison to Theorem 4.2 of (Jin et al., 2020c), Lemma 3.2 allows the comparator policy to be arbitrary, and the implication is profound. Specifically, Lemma 3.2 indicates that a pessimism-style algorithm *always* retains reasonable behavior, in the sense that, given enough data, it will eventually find the best policy among all the policies covered by the data generating distribution, i.e. $\arg\max_\pi V^\pi(\mu)$, s.t. $\kappa(\pi) < \infty$. Similar to the ξ-coverage, when specialized to standard offline RL, our analysis provides a tighter bound.

Corollary 3.2 (Uncorrupted Learning without Coverage)**.** *Under assumption 2.1 and 3.2, given any comparator policy $\tilde{\pi}$ with $\kappa < \infty$, define the empirical covariance as*

$$\Lambda_h = \frac{H}{N}\sum_{i=1}^{N/H}\phi(s_i^h,a_i^h)\phi(s_i^h,a_i^h)^\top + \lambda\cdot I, \quad \lambda = c'\cdot dH\log(N/\delta)/N \tag{11}$$

where c' is an absolute constant. Then, with pessimism bonus

$$\Gamma_h(s,a) = H\left(\sqrt{d\cdot\lambda} + \sqrt{\frac{Hd\log(N/\delta\lambda)}{N}}\right)\cdot\|\phi(s,a)\|_{\Lambda_h^{-1}}$$

and ridge regression with regularizer coefficient λ as the RLS solver, Algorithm 1 will with probability at least $1-\delta$ achieve

$$SubOpt(\hat{\pi},\tilde{\pi}) \le \tilde{O}\left(\left(H^2d + H^{2.5}\sqrt{d}\right)\sqrt{\frac{d\kappa}{N}}\right) \tag{12}$$

We note that the leading term (first term) $O(d^{3/2})$ is directly due to the assumption that the linear MDP parameter $\max(\|\mu(\mathcal{S})\|,\|\theta\|) \le \sqrt{d}$. If instead $\max(\|\mu(\mathcal{S})\|,\|\theta\|) \le \rho$ for some ρ indepdent of d, then the above bound will become linear in d. In contrast, the covering-number style analysis will generate $d^{3/2}$ regardless of the parameter norm, since its second term will become $O(d^{3/2})$ and dominate (as one needs to perform a covering argument to cover the quadratic penalty term $\Gamma_h(s,a)$).

The proof of Theorem 3.3 is technical but largely follows the analysis framework of pessimism-based offline RL and consists of two main steps. The first step establishes $\Gamma_h(s,a)$ as a valid bonus by showing

$$|\hat{Q}_h(s,a) - (\mathbb{B}_h\hat{V}_{h+1})(s,a)| \le \Gamma_h(s,a), \text{ w.p. } 1-\delta/H. \tag{13}$$

The second step applies the following Lemma connectingthe optimality gap with the expectation of $\Gamma_h(s,a)$ under visitation distribution of the comparator policy.

Lemma 3.2 (Suboptimality for Pessimistic Value Iteration)**.** *Under assumption 2.1, and under the event $\mathcal{E}$ that the $\Gamma_h(s,a)$ satisfies the required property of bounding the Bellman error, i.e. $|\hat{Q}_h(s,a)-(\mathbb{B}_h\hat{V}_{h+1})(s,a)| \le \Gamma_h(s,a), \forall h\in[H]$, then against any comparator policy $\tilde{\pi}$, it achieves*

$$SubOpt(\hat{\pi},\tilde{\pi}) \le 2\sum_{h=1}^{H}\mathbb{E}_{d^{\tilde{\pi}}}[\Gamma_h(s,a)] \tag{14}$$

We then further upper-bound the expectation through the following inequality, which bounds the distribution shift effect using the relative condition number κ:

$$\mathbb{E}_{d^{\tilde{\pi}}}\left[\sqrt{\phi(s,a)^\top\Lambda^{-1}\phi(s,a)}\right] \le \sqrt{5d\kappa} \tag{15}$$

The detailed proof can be found in Appendix C. Note that the prior work (Jin et al., 2020c) only establishes results in terms of the suboptimality comparing with the optimal policy, and when specializes to linear MDPs, they assume the offline data has global full coverage. We replace these redundant assumptions with a single assumption of partial coverage with respect to any comparator policy, in the form of a finite relative condition number.

Remark 3.8 (Novel bonus term)**.** One of our main algorithmic contributions is the new bonus term that upper-bound the effect of data contamination on the Bellman error. Ignoring ε-independent additive terms and absolute constants, our bonus term has the form

$$H\sqrt{\varepsilon}\cdot\sqrt{\phi(s,a)^\top\Lambda^{-1}\phi(s,a)}. \tag{16}$$

In comparison, below is the one used in (Lykouris et al., 2019) for online corruption-robust RL:

$$H\varepsilon\cdot\sqrt{\phi(s,a)^\top\Lambda^{-2}\phi(s,a)}. \tag{17}$$

In the tabular case, (17) evaluates to $H\varepsilon/\nu(s,a)$ and (16) evaluates to $H\sqrt{\varepsilon/\nu(s,a)}$, and thus (17) is actually tighter than (16) for $\nu(s,a)\geq\varepsilon$. However, in the linear MDP case, the relation between the two is less obvious. As we shall see, when offline distribution has good coverage, i.e. Λ is well-conditioned, (17) appears to be tighter. However, as the smallest eigenvalue of Λ goes to zero, a.k.a. lack of coverage, (17) actually blows up rapidly, whereas both (16) and the actual achievable gap remain bounded.

We demonstrate these behaviors with a numerical simulation, shown in Figure 1. In the simulation, we compare the size of three terms

$$\text{maximum possible gap} = \max_{\|y\|_\infty\leq 2H,\|y\|_0\leq\varepsilon N}\phi(s,a)^\top\Lambda^{-1}\left(\frac{1}{N}\sum_{i=1}^{N}\phi(s_i,a_i)\cdot y_i\right)$$

$$\text{bonus 1} = H\varepsilon\cdot\sqrt{\phi(s,a)^\top\Lambda^{-2}\phi(s,a)}$$

$$\text{bonus 2} = H\sqrt{\varepsilon}\cdot\sqrt{\phi(s,a)^\top\Lambda^{-1}\phi(s,a)}$$

The maximum possible gap is defined as above since for any (s,a) pair and in any step h, the bias introduced to its Bellman update due to corruption takes the form of

$$\phi(s,a)^\top\Lambda^{-1}\left(\frac{1}{N}\sum_{i=1}^{N}\phi_i\left(\tilde{y}_i-y_i\right)\right) \tag{18}$$

where $\tilde{y}_i=\tilde{r}_i+\hat{V}_{h+1}(\tilde{s}'_i)$ and $y_i=r_i+\hat{V}_{h+1}(s'_i)$, in which $\tilde{r}_i$ and $\tilde{x}'_i$ are the clean reward and transitions. For the sake of clarity, here we assume that the adversary only contaminates the reward and transitions in a bounded fashion while keeping the current (s,a)-pairs unchanged. (18) can then be upper-bounded by (**??**), because there are at most εN tuples on which $\tilde{r}_i\neq r_i$ or $\tilde{s}'_i\neq s'_i$, and for any such tuple $(\tilde{r}_i+\hat{V}_{h+1}(\tilde{s}'_i))-(r_i+\hat{V}_{h+1}(s'_i))\leq 2H$.

In the simulation, we set $H=1$ to ignore the scaling on time horizon and let $\lambda=1$; We let both the test data $\phi(s,a)$ and the training data $\phi(s_i,a_i)$ to be sampled from a truncated standard Gaussian distribution in $\mathbb{R}^3$, denoted by ν, with mean 0, and covariance eigenvalues $1,1,\lambda_{\min}$. We set the training data size set to $N=10^6$ and contamination level set to $\varepsilon=0.01$. The x-axis tracks $-\log(\lambda_{\min})$, while the y-axis tracks $\mathbb{E}_{s,a\sim\nu}\text{bonus}(s,a)$, with expectation being approximated by 1000 test samples from ν. It can be seen that bonus 1 starts off closely upper-bounding the maximum possible gap when the data has good coverage, but increases rapidly as $\lambda_{\min}$ decreases. Note that for a fixed N, bonus 1 will eventually plateau at $HN\varepsilon/\lambda$, but this term scales with N, so the error blows up as the number of samples grows, which certainly is not desirable. Bonus 2, on the other hand, is not as tight as bonus 1 when there is good data coverage, but remains intact regardless of the value of $\lambda_{\min}$, which is essential for the more challenging setting with poor data coverage.

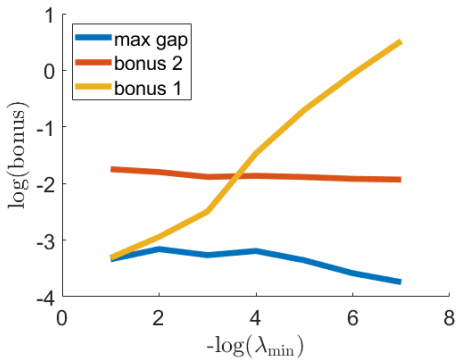

Figure 1: bonus size simulation

This new bonus term can be of independent interest in other robust RL contexts. For example, in the online corruption-robust RL problem, as a result of using the looser bonus term (16), the algorithm in (Lykouris et al., 2019) can only handle $\varepsilon=T^{-3/4}$ amount of corruptions in the linear MDP setting, while being able to handle $\varepsilon=T^{-1/2}$ amount of corruptions in the tabular setting, due to the tabular bonus being tighter. Our bonus term can be directly plugged into their algorithm, allowing it to handle up to $\varepsilon=T^{-1/2}$ amount of corruption even in the linear MDP setting, achieving an immediate improvement over previous results.[2]

Note that our algorithm and theorem are adaptive to the unknown relative coverage κ, but is not adaptive to the level of contamination ε (i.e., algorithm requires knowing ε or a tight upper bound of ε). One may ask whether there exists an agnostic result, similar to Theorem 3.2, where an algorithm can be adaptive simultaneously to unknown values of ε and coverage parameter κ. Our last result shows that this is unfortunately not pos-

[2]Though our bound improve their result, the tolerable corruption amount is still sublinear, which is due to the multi-layer scheduling procedure used in their algorithm.

sible without full data coverage. In particular, we show that no algorithm can achieve a best-of-both-worlds guarantee in both clean and ε-corrupted environments. More specifically, in this setting, κ is still unknown to the learner, and the adversary either corrupt ε amount of tuples (ε is known) or does not corrupt at all—but the learner does not know which situation it is.

Theorem 3.4 (Agnostic learning is impossible without full coverage)**.** *Under assumption 2.1 and 3.2, for any algorithm* $L : (\mathcal{S} \times \mathcal{A} \times \mathbb{R} \times \mathcal{A})^N \to \Pi$ *that achieves diminishing suboptimality in clean environment, i.e., for any clean dataset* $\tilde{\mathcal{D}}$ *it achieves* $SubOpt(L(\tilde{\mathcal{D}})) = g(N)$ *for some positive function* g *such that* $\lim_{N\to\infty} g(N) = 0$*, we have that for any* $\varepsilon \in (0, 1/2]$*, there exists an MDP* $\mathcal{M}^\dagger$ *such that with probability at least* $1/4$*,* $\max_{f_c} SubOpt(\hat{\pi}, \tilde{\pi}) \geq 1/2$.

Intuitively, the logic behind this result is that to achieve vanishing errors in the clean environment, the learner has no choice but to *trust* all data as clean. However, it is also possible that the same dataset could be generated under some adversarial corruption from another MDP with a very different optimal policy—thus the learner cannot be robust to corruption under that MDP.

Specifically, consider a 2-arm bandit problem. The learner observes a dataset of N data points of arm-reward pairs, of which p fraction is arm a_1 and $(1-p)$ fraction is arm a_2. For simplicity, we assume that N is large enough such that the empirical distribution converges to the underlying sampling distribution. Assume further that the average reward observed for a_1 is $\hat{r}_1 = \frac{1}{2} + \frac{\varepsilon}{2p}$, for some $\varepsilon \leq p$, and the average reward observed for a_2 is $\frac{1}{2}$. Given such a dataset, two data generating processes can generate such a dataset with equal likelihood and thus indistinguishable based only on the data:

1. There is no contamination. The MDP has a reward setting where a_1 indeed has reward $r_1 = \text{Bernoulli}(\frac{1}{2} + \frac{\varepsilon}{2p})$ and a_2 has $r_2 = \text{Bernoulli}(\frac{1}{2})$. Since there is no corruption, $\kappa = 1/p$ in this MDP.

2. The data is ε-corrupted. In particular, in this MDP, the actual reward of a_1 is $r_1 = \text{Bernoulli}(\frac{1}{2} - \frac{\varepsilon}{2p})$, and the adversary is able to increase empirical mean by ε/p via changing εN number of data points from $(a_1, 0)$ to $(a_1, 1)$. One can show that this can be achieved by the adversary with probability at least $1/2$ (which is where the probability $1/2$ in the theorem statement comes from). In this MDP, we have $\kappa = 1/(1-p)$.

Now, since the algorithm achieves a diminishing suboptimal gap in all clean environments, it must return a_1 with high probability given such a dataset, due to the possibility of the learner facing the data generation process 1. However, committing to action a_1 will incur $\varepsilon/2p$ suboptimal gap in the second MDP with the data generation process 2. On the other hand, note that the relative condition number in the second MDP is bounded, i.e. $\frac{1}{1-p} \leq 2$ for $\varepsilon \leq p \leq 1/2$. Therefore, for any $\varepsilon \in (0, 1/2]$, one can construct such an instance with $p = \varepsilon$, such that the relative condition number for the second MDP is $\frac{1}{1-p} \leq 2$ and the relative condition number for the first MDP is $\frac{1}{\varepsilon} < \infty$, while the learner would always suffer $\varepsilon/2p = 1/2$ suboptimality gap in the second MDP if she had to commit to a_1 under the first MDP where data is clean.

Remark 3.9 (Offline vs. Online RL: Agnostic Learning)**.** Theorem 3.4 shows that no algorithm can simultaneously achieve good performance in both clean and corrupted environments without knowing which one it is currently experiencing. This is in sharp contrast to the recent result in (Zhang et al., 2021), which shows that in the online RL setting, natural policy gradient (NPG) algorithm can find an $O(\sqrt{\kappa\varepsilon})$-optimal policy for any unknown contamination level ε with the help of an exploration policy with finite relative condition number. Without such a helper policy, however, robust RL is much harder, and the best-known result (Lykouris et al., 2019) can only handle $\varepsilon \leq O(1/\sqrt{T})$ corruption, but still does not require the knowledge of ε. Intuitively, such adaptivity is lost in the offline setting, because the learner is no longer able to evaluate the current policy by collecting on-policy data. In the online setting, the construction in Theorem 3.4 will not work. Our construction heavily relies on the fact that ν has ε probability of sampling a_1, which allows adversary in the second MDP to concentrate its corruption budget all on a_1. In the online setting, one can simply uniform randomly try a_1 and a_2 to significantly increase the probability of sampling a_1 which in turn makes the estimation of r_1 accurate (up to $O(\varepsilon)$ in the corrupted data generation process).

4 Discussions and Conclusion

In this paper, we studied corruption-robust RL in the offline setting. We provided an information-theoretical lower bound and two near-matching upper-bounds for cases with or without full data coverage, respectively. We also establish an impossibility result, showing that an agnostic algorithm is impossible in corruption-robust offline RL and distincting the offline setting from the online counterpart. Finally, when specialized to the uncorrupted setting, our algorithm and analysis also obtained tighter bounds than prior works.

Acknowledgement

Zhu acknowledges NSF grants 1545481, 1704117, 1836978, 2041428, 2023239, ARO MURI W911NF2110317 and MAD- Lab AF CoE FA9550-18-1-0166.

References

Abbasi-Yadkori, Y., Pál, D., and Szepesvári, C. (2011). Improved algorithms for linear stochastic bandits. In *NIPS*, volume 11, pages 2312–2320.

Agarwal, A., Kakade, S. M., Lee, J. D., and Mahajan, G. (2019). On the theory of policy gradient methods: Optimality, approximation, and distribution shift. *arXiv preprint arXiv:1908.00261*.

Agarwal, R., Schuurmans, D., and Norouzi, M. (2020). An optimistic perspective on offline reinforcement learning. In *International Conference on Machine Learning*, pages 104–114. PMLR.

Bakshi, A. and Prasad, A. (2020). Robust linear regression: Optimal rates in polynomial time. *arXiv preprint arXiv:2007.01394*.

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

Cai, Q., Yang, Z., Jin, C., and Wang, Z. (2020). Provably efficient exploration in policy optimization. In *International Conference on Machine Learning*, pages 1283–1294. PMLR.

Charikar, M., Steinhardt, J., and Valiant, G. (2017). Learning from untrusted data. In *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing*, pages 47–60.

Chen, Y., Du, S. S., and Jamieson, K. (2021). Improved corruption robust algorithms for episodic reinforcement learning. *arXiv preprint arXiv:2102.06875*.

Diakonikolas, I., Kamath, G., Kane, D., Li, J., Moitra, A., and Stewart, A. (2016). Robust estimators in high dimensions without the computational intractability. In *2016 IEEE 57th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 655–664.

Diakonikolas, I., Kamath, G., Kane, D., Li, J., Steinhardt, J., and Stewart, A. (2019). Sever: A robust meta-algorithm for stochastic optimization. In *International Conference on Machine Learning*, pages 1596–1606.

Diakonikolas, I. and Kane, D. M. (2019). Recent advances in algorithmic high-dimensional robust statistics. *arXiv preprint arXiv:1911.05911*.

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

Eykholt, K., Evtimov, I., Fernandes, E., Li, B., Rahmati, A., Xiao, C., Prakash, A., Kohno, T., and Song, D. (2018). Robust physical-world attacks on deep learning visual classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1625–1634.

Fujimoto, S., Meger, D., and Precup, D. (2019). Off-policy deep reinforcement learning without exploration. In *International Conference on Machine Learning*, pages 2052–2062. PMLR.

Huber, P. J. et al. (1967). The behavior of maximum likelihood estimates under nonstandard conditions. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 221–233. University of California Press.

Jiang, N. (2020). Notes on tabular methods.

Jin, C., Jin, T., Luo, H., Sra, S., and Yu, T. (2020a). Learning adversarial markov decision processes with bandit feedback and unknown transition. In *International Conference on Machine Learning*, pages 4860–4869. PMLR.

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

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

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

Klivans, A., Kothari, P. K., and Meka, R. (2018). Efficient algorithms for outlier-robust regression. In *Conference On Learning Theory*, pages 1420–1430. PMLR.

Kumar, A., Fu, J., Tucker, G., and Levine, S. (2019). Stabilizing off-policy q-learning via bootstrapping error reduction. *arXiv preprint arXiv:1906.00949*.

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

Lai, K. A., Rao, A. B., and Vempala, S. (2016). Agnostic estimation of mean and covariance. In *2016 IEEE 57th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 665–674. IEEE.

Lange, S., Gabel, T., and Riedmiller, M. (2012). Batch reinforcement learning. In *Reinforcement learning*, pages 45–73. Springer.

Laroche, R., Trichelair, P., and Des Combes, R. T. (2019). Safe policy improvement with baseline bootstrapping. In *International Conference on Machine Learning*, pages 3652–3661. PMLR.

Levine, S., Kumar, A., Tucker, G., and Fu, J. (2020). Offline reinforcement learning: Tutorial, review, and perspectives on open problems. *arXiv preprint arXiv:2005.01643*.

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

Lykouris, T., Simchowitz, M., Slivkins, A., and Sun, W. (2019). Corruption robust exploration in episodic reinforcement learning. *arXiv preprint arXiv:1911.08689*.

Neff, G. (2016). Talking to bots: Symbiotic agency and the case of tay. *International Journal of Communication*.

Neu, G., Antos, A., György, A., and Szepesvári, C. (2010). Online markov decision processes under bandit feedback. In *Advances in Neural Information Processing Systems*, pages 1804–1812.

Neu, G., Gyorgy, A., and Szepesvári, C. (2012). The adversarial stochastic shortest path problem with unknown transition probabilities. In *Artificial Intelligence and Statistics*, pages 805–813.

Pensia, A., Jog, V., and Loh, P.-L. (2020). Robust regression with covariate filtering: Heavy tails and adversarial contamination. *arXiv preprint arXiv:2009.12976*.

Rashidinejad, P., Zhu, B., Ma, C., Jiao, J., and Russell, S. (2021). Bridging offline reinforcement learning and imitation learning: A tale of pessimism. *arXiv preprint arXiv:2103.12021*.

Rosenberg, A. and Mansour, Y. (2019). Online stochastic shortest path with bandit feedback and unknown transition function. In *Advances in Neural Information Processing Systems*, pages 2212–2221.

Siegel, N. Y., Springenberg, J. T., Berkenkamp, F., Abdolmaleki, A., Neunert, M., Lampe, T., Hafner, R., Heess, N., and Riedmiller, M. (2020). Keep doing what worked: Behavioral modelling priors for offline reinforcement learning. *arXiv preprint arXiv:2002.08396*.

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

Yang, L. F. and Wang, M. (2019). Sample-optimal parametric q-learning using linearly additive features. *arXiv preprint arXiv:1902.04779*.

Yang, M. and Nachum, O. (2021). Representation matters: Offline pretraining for sequential decision making. *arXiv preprint arXiv:2102.05815*.

Yin, M., Bai, Y., and Wang, Y.-X. (2020). Near optimal provable uniform convergence in off-policy evaluation for reinforcement learning. *arXiv preprint arXiv:2007.03760*.

Yin, M., Bai, Y., and Wang, Y.-X. (2021). Near-optimal offline reinforcement learning via double variance reduction. *arXiv preprint arXiv:2102.01748*.

Yu, T., Kumar, A., Rafailov, R., Rajeswaran, A., Levine, S., and Finn, C. (2021). Combo: Conservative offline model-based policy optimization. *arXiv preprint arXiv:2102.08363*.

Yu, T., Thomas, G., Yu, L., Ermon, S., Zou, J., Levine, S., Finn, C., and Ma, T. (2020). Mopo: Model-based offline policy optimization. *arXiv preprint arXiv:2005.13239*.

Zanette, A., Cheng, C.-A., and Agarwal, A. (2021). Cautiously optimistic policy optimization and exploration with linear function approximation. *arXiv preprint arXiv:2103.12923*.

Zhang, X., Chen, Y., Zhu, X., and Sun, W. (2021). Robust policy gradient against strong data corruption. *arXiv preprint arXiv:2102.05800*.

Zimin, A. and Neu, G. (2013). Online learning in episodic markovian decision processes by relative entropy policy search. In *Advances in neural information processing systems*, pages 1583–1591.

Appendices

A Basics

Lemma A.1. $\|w_h^*\| \leq H\sqrt{d}$ *for all* h.

Proof. By definition, we have

$$w_h^* = \theta + \int_{\mathcal{S}} \hat{V}_{h+1}(s')\mu_h(s')ds' \tag{19}$$

and thus

$$\begin{aligned}
\|w_h^*\| &\leq \|\theta\| + \|\int_{\mathcal{S}} \hat{V}_{h+1}(s')\mu_h(s')ds'\| && (20)\\
&\leq \|\theta\| + \int_{\mathcal{S}} \|\hat{V}_{h+1}(s')\mu_h(s')\|ds' && (21)\\
&\leq \sqrt{d} + (H-h+1)\sqrt{d} && (22)\\
&\leq H\sqrt{d}. && (23)
\end{aligned}$$

■

Lemma A.2. *Note that* $\mathbb{E}[[(r(s,a) + \hat{V}(s')) - (\mathbb{B}_h\hat{V})(s,a)]^2|s,a] \leq \gamma^2 = (\sigma + H/2)^2$

Proof.

$$Var(X+Y) = Var(X) + Var(Y) + 2Cov(X,Y) \leq Var(X) + Var(Y) + 2\sqrt{Var(X)Var(Y)}$$

Because $0 \leq \hat{V}(s') \leq H$,

$$\begin{aligned}
&\mathbb{E}[(\hat{V}(s') - \mathbb{E}[\hat{V}(s')|s,a])^2|s,a] = \mathbb{E}[\hat{V}(s')^2|s,a] - \mathbb{E}[\hat{V}(s')|s,a]^2 && (24)\\
\leq &H\mathbb{E}[\hat{V}(s')|s,a] - \mathbb{E}[\hat{V}(s')|s,a]^2 \leq \frac{H^2}{4}. && (25)
\end{aligned}$$

$$\begin{aligned}
&\mathbb{E}[[(r(s,a) + \hat{V}(s')) - (\mathbb{B}_h\hat{V})(s,a)]^2|s,a] && (26)\\
=&\mathbb{E}[[(r(s,a) + \hat{V}(s')) - \mathbb{E}[r(s,a) + \hat{V}(s')|s,a]]^2|s,a] && (27)\\
=&\mathbb{E}[(r(s,a) - \mathbb{E}[r(s,a)|s,a])^2|s,a] + \mathbb{E}[(\hat{V}(s') - \mathbb{E}[\hat{V}(s')|s,a])^2|s,a] && (28)\\
&+ 2\mathbb{E}[(r(s,a) - \mathbb{E}[r(s,a)|s,a])(\hat{V}(s') - \mathbb{E}[\hat{V}(s')|s,a])|s,a] && (29)\\
\leq&\mathbb{E}[(r(s,a) - \mathbb{E}[r(s,a)|s,a])^2|s,a] + \mathbb{E}[(\hat{V}(s') - \mathbb{E}[\hat{V}(s')|s,a])^2|s,a] && (30)\\
&+ 2\sqrt{\mathbb{E}[(r(s,a) - \mathbb{E}[r(s,a)|s,a])^2|s,a]\mathbb{E}[(\hat{V}(s') - \mathbb{E}[\hat{V}(s')|s,a])^2|s,a]} \quad \text{(By Cauchy's Ineq)} && (31)\\
=&Var(r(s,a) \mid (s,a)) + Var(\hat{V}(s') \mid (s,a)) + 2\sqrt{Var(r(s,a) \mid (s,a))Var(\hat{V}(s') \mid (s,a))} && (32)\\
=&\left(\sqrt{Var(r(s,a) \mid (s,a))} + \sqrt{Var(\hat{V}(s') \mid (s,a))}\right)^2 \leq (\sigma + H/2)^2 && (33)
\end{aligned}$$

■

B Proof of the Minimax Lower-bound

Proof of Theorem 3.1. Given any dimension d, time horizon H, consider a tabular MDP with action space size $A > 2$ and state space size $S \leq \left(\frac{A}{2}\right)^{H/2}$ s.t. $SA = d$. Consider a "tree" with self-loops, which has S nodes and

depth $\lceil \log_{A/2}\left(S\left(\frac{A}{2}-1\right)+1\right)\rceil$. There is 1 node at the first level, $\frac{A}{2}$ nodes at the second level, $\left(\frac{A}{2}\right)^2$ nodes at the third level, ..., $\left(\frac{A}{2}\right)^{\lceil \log_{A/2}\left(S\left(\frac{A}{2}-1\right)+1\right)\rceil-2}$ nodes at the second to last level. The rest nodes are all at the last level. Define the MDP induced by this graph, where each state corresponds to a node, and each action corresponds to an edge. The agent always starts from the first level. For each state at the first $\lceil \log_{A/2}\left(S\left(\frac{A}{2}-1\right)+1\right)\rceil-2$ levels, there are $A/2$ actions that lead to child nodes, and the rest leads back to that state, i.e. self-loops. The leaf states are absorbing state, i.e. all actions lead to self-loops. Denote this transition structure as P. Let's consider two MDPs with the same transition structure and different reward function, i.e. $M=(P,R)$, $M'=(P,R')$.

For MDP M, define $R(s^*,a^*)=\text{Bernoulli}(SA\varepsilon/2)$ on one particular (s^*,a^*) pair, where s^* is a leaf state at the last level, a^* is a self-loop action. Every other (s,a) pair receive reward 0. Let $(s',a')=\arg\min_{(s,a)}\nu(s,a)$ be the state-action pair appears least often in the data collecting distribution. For MCP M', define $R'(s^*,a^*)=\text{Bernoulli}(SA\varepsilon/2)$, $R'(s',a')=\text{Bernoulli}(SA\varepsilon)$ and 0 everywhere else. Then, it can be easily verified that: on M, the expected cumulative reward of the optimal policy is $\left(H-\lceil \log_{A/2}\left(S\left(\frac{A}{2}-1\right)+1\right)\rceil\right)SA\varepsilon/2$; on M', the expected cumulative reward of the optimal policy is at least $\left(H-\lceil \log_{A/2}\left(S\left(\frac{A}{2}-1\right)+1\right)\rceil\right)SA\varepsilon$; no policy can be simultaneously better than $\left(H-\lceil \log_{A/2}\left(S\left(\frac{A}{2}-1\right)+1\right)\rceil\right)SA\varepsilon/4$-optimal on both M and M'. Note that because $S\le\left(\frac{A}{2}\right)^{H/2}$,

$$\left(H-\lceil \log_{A/2}\left(S\left(\frac{A}{2}-1\right)+1\right)\rceil\right)SA\varepsilon/4=\Omega(HSA\varepsilon). \tag{34}$$

With probability at least $1/2$, we have $N(s',a')\le T\nu(s',a')\le T/SA$ by the pigeonhole principle. Conditioning on $N(s',a')\le T/SA$, with probability at least $1/2$, the amount of positive reward $r(s',a')$ will not exceed $SA\varepsilon N(s',a')\le\varepsilon T$, and thus an ε-contamination adversary can perturb all the positive rewards on (s',a') to 0. In other words, with probability $1/4$, the learner will observe a dataset whose likelihood under M and ($M'+\varepsilon$-contamination) are exactly the same, and thus the learner must suffer at least $\Omega(HSA\varepsilon)$ regret on one of the MDPs. ■

C Proof of Upper-bounds

Proof of Lemma 3.2. Applying Lemma F.2 with $\pi=\hat{\pi}$, $\pi'=\tilde{\pi}$, and $\{\hat{Q}_h\}_{h=1}^H$ being the Q-functions constructed by the meta-algorithm, we have

$$\begin{aligned}\hat{V}_1(s)-V_1^{\tilde{\pi}}(s)=&\sum_{h=1}^H\mathbb{E}_{\tilde{\pi}}\left[\langle\hat{Q}_h(s_h,\cdot),\hat{\pi}_h(\cdot|s_h)-\tilde{\pi}_h(\cdot|s_h)\rangle_{\mathcal{A}}|s_1=s\right]\\&+\sum_{h=1}^H\mathbb{E}_{\tilde{\pi}}\left[\hat{Q}_h(s_h,a_h)-(\mathbb{B}_h\hat{V}_{h+1})(s_h,a_h)|s_1=s\right]\end{aligned} \tag{35}$$

Similarly, applying Lemma F.2 with $\pi=\pi'=\hat{\pi}$, we have

$$\hat{V}_1(s)-V_1^{\hat{\pi}}(s)=\sum_{h=1}^H\mathbb{E}_{\hat{\pi}}\left[\hat{Q}_h(s_h,a_h)-(\mathbb{B}_h\hat{V}_{h+1})(s_h,a_h)|s_1=s\right] \tag{36}$$

Then, we have

$$\begin{aligned}
\text{SubOpt}(\hat{\pi}, \tilde{\pi}) =& \left(V_1^{\hat{\pi}}(\mu) - \hat{V}_1(\mu)\right) + \left(\hat{V}_1(\mu) - V_1^{\hat{\pi}}(\mu)\right) && (37)\\
=& -\sum_{h=1}^{H} \mathbb{E}_{\tilde{\pi}}\left[(\mathbb{B}_h \hat{V}_{h+1}) - \hat{Q}_h\right] + \sum_{h=1}^{H} \mathbb{E}_{\tilde{\pi}}\left[(\mathbb{B}_h \hat{V}_{h+1}) - \hat{Q}_h\right] && (38)\\
&+ \sum_{h=1}^{H} \mathbb{E}_{\tilde{\pi}}\left[\langle \hat{Q}_h(s_h, \cdot), \tilde{\pi}_h(\cdot|s_h) - \hat{\pi}_h(\cdot|s_h)\rangle_{\mathcal{A}}\right] && (39)\\
\leq& 0 + 2\sum_{h=1}^{H} \mathbb{E}_{\tilde{\pi}}[\Gamma_h(s,a)] + 0 && (40)\\
=& 2\sum_{h=1}^{H} \mathbb{E}_{\tilde{\pi}}[\Gamma_h(s,a)] && (41)
\end{aligned}$$

as needed. ■

Proof of Theorem 3.3. To simplify the notation, below we use M for the number of data points per time step, i.e. $M := N/H$. We first show that

$$|\hat{Q}_h(s,a) - (\mathbb{B}_h \hat{V}_{h+1})(s,a)| \leq \Gamma(s,a). \tag{42}$$

The robust least-square oracle guarantees

$$\begin{aligned}
\mathbb{E}_\nu\left(\|x^\top(\hat{w} - w^*)\|_2^2\right) \leq c_2(\delta) \cdot \left(\frac{\gamma^2 \text{poly}(d)}{M} + \gamma^2\varepsilon\right) && (43)\\
\implies \|\hat{w}_h - w_h^*\|_\Sigma^2 \leq c_2(\delta) \cdot \left(\frac{\gamma^2 \text{poly}(d)}{M} + \gamma^2\varepsilon\right) && (44)\\
\implies \|\hat{w}_h - w_h^*\|_{\Sigma+(2\varepsilon+\lambda)I}^2 \leq c_2(\delta) \cdot \left(\frac{\gamma^2 \text{poly}(d)}{M} + \gamma^2\varepsilon + (2\varepsilon+\lambda)H^2 d\right) && (45)
\end{aligned}$$

Then,

$$\begin{aligned}
|\hat{Q}_h(s,a) - (\mathbb{B}_h \hat{V}_{h+1})(s,a)| =& |\phi(s,a)(\hat{w}_h - w_h^*)| && (46)\\
\leq& \|\hat{w}_h - w_h^*\|_{(\Sigma+(2\varepsilon+\lambda)I)} \|\phi(s,a)\|_{(\Sigma+(2\varepsilon+\lambda)I)^{-1}} && (47)\\
\leq& \sqrt{c_2(\delta) \cdot \left(\frac{\gamma^2 \text{poly}(d)}{M} + \gamma^2\varepsilon + (2\varepsilon+\lambda)H^2 d\right)} \|\phi(s,a)\|_{(\Sigma+(2\varepsilon+\lambda)I)^{-1}} && (48)\\
\leq& \sqrt{c_2(\delta)} \cdot \left(\frac{\gamma \text{poly}(d)}{\sqrt{M}} + (\gamma + 2H\sqrt{d})\sqrt{\varepsilon} + H\sqrt{d\lambda})\right) \|\phi(s,a)\|_{\Lambda^{-1}} && (49)
\end{aligned}$$

where the last step are due to $W \leq H\sqrt{d}$ and

$$\begin{aligned}
\Lambda =& \frac{3}{5}\left(\frac{1}{M}\sum_{i=1}^{M} \phi_i \phi_i^\top + (\varepsilon+\lambda) \cdot I\right) && (50)\\
\preceq& \frac{3}{5}\left(\frac{1}{M}\sum_{i=1}^{M} \tilde{\phi}_i \tilde{\phi}_i^\top + (2\varepsilon+\lambda) \cdot I\right) && (51)\\
\preceq& (\Sigma + (2\varepsilon+\lambda) \cdot I) && (52)
\end{aligned}$$

where $\tilde{\phi}$ denotes the clean data and the last step applies Lemma F.3 because $M(2\varepsilon+\lambda) \geq \Omega(d\log(M/\delta))$ due to the definition of λ and $\varepsilon \geq 0$.

Next, we show that Algorithm 1 achieves the desired optimality gap. By Lemma 3.2, we have

$$\text{SubOpt}(\hat{\pi}) \leq 2H\mathbb{E}_{\pi^*}[\Gamma(s,a)] \tag{53}$$

$$\leq \sqrt{c_2(\delta)} \cdot \left(\frac{\gamma H \text{poly}(d)}{\sqrt{N}} + (H\gamma + 2H^2\sqrt{d})\sqrt{\varepsilon} + H^2\sqrt{d\lambda} \right) \mathbb{E}_{\pi^*} \left[\|\phi(s,a)\|_{\Lambda^{-1}} \right] \tag{54}$$

Focusing on the last term, applying Lemma F.3 again, we have

$$\mathbb{E}_{d^*}[\|\phi(s,a)\|_{\Lambda^{-1}}] \leq \mathbb{E}_{d^*}[\|\phi(s,a)\|_{(\frac{1}{5}(\Sigma+\lambda I))^{-1}}] \tag{55}$$

$$= \mathbb{E}_{d^*}\left[\sqrt{\phi^\top (\frac{1}{5}(\Sigma+\lambda I))^{-1}\phi} \right] \tag{56}$$

$$\leq \sqrt{\mathbb{E}_{d^*}[\phi^\top (\frac{1}{5}(\Sigma+\lambda I))^{-1}\phi]} \tag{57}$$

$$\leq \sqrt{tr\left(\Sigma_* (\frac{1}{5}(\Sigma+\lambda I))^{-1} \right)} \tag{58}$$

$$\leq \sqrt{\kappa tr\left(\Sigma (\frac{1}{5}(\Sigma+\lambda I))^{-1} \right)} \tag{59}$$

$$\leq \sqrt{5\kappa \sum_{i=1}^{d} \frac{\sigma_i}{\sigma_i + \lambda}} \tag{60}$$

$$\leq \sqrt{5d\kappa} \tag{61}$$

Combining the two terms give the desired results. ■

D Proof of uncorrupted learning results

In this section, we prove the conclusion in Corollary 3.1 and 3.2. The proof follows closely the classic analysis of Least Squared Value Iteration (LSVI) methods with the only difference being the data splitting, which allows us to ditch the covering argument and obtain a tighter bound. Such a trick is only possible in the offline setting where the data are assumed to be i.i.d. For completeness, we specify the uncorrupted algorithm in Alg. 2.

Algorithm 2 Uncorrupted Least-Square Value Iteration (LSVI)

Input: Dataset $D = \{(s_i, a_i, r_i, s_i')\}_{1:N}$; pessimism bonus $\Gamma_h(s,a) \geq 0$, $\lambda > 0$.
Split the dataset randomly into H subset: $D_h = \{(s_i^h, a_i^h, r_i^h, s_i'^h)\}_{1:(N/H)}$, for $h \in [H]$.
Initialization: Set $\hat{V}_{H+1}(s) \leftarrow 0$.
for step $h = H, H-1, \ldots, 1$ **do**
  Set $\Lambda_h \leftarrow \frac{H}{M} \sum_{i=1}^{N/H} \phi(s_i^h, a_i^h)\phi(s_i^h, a_i^h)^\top + \lambda \cdot I$.
  Set $\hat{w}_h \leftarrow \Lambda_h^{-1}(\frac{H}{N} \sum_{i=1}^{N/H} \phi(s_i^h, a_i^h) \cdot (r_i^h + \hat{V}_{h+1}(s_i^{h+1})))$.
  Set $\hat{Q}_h(s,a) \leftarrow \phi(s,a)^\top \hat{w}_h - \Gamma_h(s,a)$, clipped within $[0, H-h+1]$.
  Set $\hat{\pi}_h(a|s) \leftarrow \arg\max_a \hat{Q}_h(s,a)$ and $\hat{V}_h(s) \leftarrow \max_a \hat{Q}_h(s,a)$.
Output: $\{\hat{\pi}_h\}_{h=1}^H$.

We first prove the following lemma:

Lemma D.1 (Bound on the Bellman Error)**.** *Under assumption 2.1, given a dataset of size N, Algorithm 1 achieves*

$$|(\mathbb{B}_h \hat{V}_{h+1})(s,a) - \hat{Q}_h(s,a)| \leq H \left(\sqrt{d \cdot \lambda} + \sqrt{\frac{Hd \log(N/\delta\lambda)}{N}} \right) \cdot \sqrt{\phi(x,a)^\top \Lambda_h^{-1} \phi(x,a)}$$

for all $(s,a,h) \in \mathcal{S} \times \mathcal{A} \times [H]$, with probability at least $1-\delta$.

Proof. We start by applying the following decomposition

$$
\begin{aligned}
&(\mathbb{B}_h \hat{V}_{h+1})(s,a) - \hat{Q}_h(s,a) && (62)\\
=&(\mathbb{B}_h \hat{V}_{h+1})(s,a) - (\hat{\mathbb{B}}_h \hat{V}_{h+1})(s,a) && (63)\\
=&\underbrace{\phi(s,a)^\top w_h - \phi(s,a)^\top \Lambda_h^{-1} \left(\frac{H}{N} \sum_{i=1}^{N/H} \phi(s_i,a_i) \cdot (\mathbb{B}_h \hat{V}_{h+1})(s_i,a_i) \right)}_{\text{(i)}} - && (64)\\
&\underbrace{\phi(s,a)^\top \Lambda_h^{-1} \left(\frac{H}{N} \sum_{i=1}^{N/H} \phi(s_i,a_i) \cdot \left(r_i + \hat{V}_{h+1}(s_i') - (\mathbb{B}_h \hat{V}_{h+1})(s_i,a_i) \right) \right)}_{\text{(ii)}} && (65)
\end{aligned}
$$

Therefore, by triangle inequality we have

$$
|(\mathbb{B}_h \hat{V}_{h+1})(s,a) - \hat{Q}_h(s,a)| \leq |\text{(i)}| + |\text{(ii)}| \tag{66}
$$

Then, we bound the two terms separately:

$$
\begin{aligned}
|\text{(i)}| &= \left| \phi(s,a)^\top w_h - \phi(s,a)^\top \Lambda_h^{-1} \left(\frac{H}{N} \sum_{i=1}^{N/H} \phi(s_i,a_i) \cdot \phi(s_i,a_i)^\top w_h \right) \right| \\
&= \left| \phi(s,a)^\top w_h - \phi(s,a)^\top \Lambda_h^{-1} (\Lambda_h - \lambda \cdot I) w_h \right| = \lambda \cdot |\phi(s,a)^\top \Lambda_h^{-1} w_h| \\
&\leq \lambda \cdot \|w_h\|_{\Lambda_h^{-1}} \cdot \|\phi(s,a)\|_{\Lambda_h^{-1}} \leq H\sqrt{d \cdot \lambda} \cdot \sqrt{\phi(s,a)^\top \Lambda_h^{-1} \phi(s,a)}.
\end{aligned}
$$

For the second term, define

$$
\varepsilon_i^h(V) = r_i^h + V(s_i^{h\prime}) - (\mathbb{B}_h V)(s_i^h, a_i^h) \tag{67}
$$

Then, we have

$$
\begin{aligned}
|\text{(ii)}| &= \left| \phi(s,a)^\top \Lambda_h^{-1} \left(\frac{H}{N} \sum_{i=1}^{N/H} \phi(s_i,a_i) \cdot \varepsilon_i^h(\hat{V}_{h+1}) \right) \right| \\
&\leq \underbrace{\left\| \frac{H}{N} \sum_{i=1}^{N/H} \phi(s_i,a_i) \cdot \varepsilon_i^h(\hat{V}_{h+1}) \right\|_{\Lambda_h^{-1}}}_{\text{(iii)}} \cdot \sqrt{\phi(x,a)^\top \Lambda_h^{-1} \phi(x,a)}.
\end{aligned} \tag{68}
$$

From here, because of our data splitting, $\hat{V}_{h+1}$ is independent from D_h, and thus we can bypass the covering argument and directly apply matrix concentrations. In particular, by applying Lemma F.1, we have that with probability at least $1-\delta$

$$
\text{(iii)} \leq H\sqrt{\frac{Hd\log(1+N/H\lambda) + 2H\log(1/\delta)}{N}} \tag{69}
$$

Combining the two terms gives

$$
|(\mathbb{B}_h \hat{V}_{h+1})(s,a) - \hat{Q}_h(s,a)| \leq H \left(\sqrt{d \cdot \lambda} + \sqrt{\frac{Hd\log(N/\delta\lambda)}{N}} \right) \cdot \sqrt{\phi(x,a)^\top \Lambda_h^{-1} \phi(x,a)} \tag{70}
$$

■

Now, given Lemma D.1, applying Lemma 3.2, we have

$$
\text{SubOpt}(\hat{\pi}, \tilde{\pi}) \leq 2\sum_{h=1}^{H} \mathbb{E}_{d^{\tilde{\pi}}}[\Gamma_h(s,a)] \leq 2H^2 \left(\sqrt{d \cdot \lambda} + \sqrt{\frac{Hd\log(N/\delta\lambda)}{N}} \right) \cdot \mathbb{E}_{d^{\tilde{\pi}}}[\sqrt{\phi(x,a)^\top \Lambda_h^{-1} \phi(x,a)}] \tag{71}
$$

The last step would be to bound $\mathbb{E}_{d^{\tilde{\pi}}}[\sqrt{\phi(x,a)^\top \Lambda_h^{-1}\phi(x,a)}]$, similar to the last section. In particular, applying Lemma F.3, we have

$$\begin{aligned}\mathbb{E}_{d^{\tilde{\pi}}}\left[\sqrt{\phi(x,a)^\top \Lambda_h^{-1}\phi(x,a)}\right] \leq & \mathbb{E}_{d^{\tilde{\pi}}}\left[\sqrt{3\phi(x,a)^\top(\Sigma+\lambda I)\phi(x,a)}\right] && (72)\\ \leq & \sqrt{3\mathbb{E}_{d^{\tilde{\pi}}}\left[\phi(x,a)^\top(\Sigma+\lambda\cdot I)\phi(x,a)\right]} && (73)\\ \leq & \sqrt{3d\kappa} && (74)\end{aligned}$$

where step 72 requires $\lambda \geq H\Omega(d\log(N/\delta))/N$. Thus,

$$\begin{aligned}\mathrm{SubOpt}(\hat{\pi},\tilde{\pi}) \leq & 2H^2\left(\sqrt{d\cdot\lambda}+\sqrt{Hd\log(N/\delta\lambda)}\right)\sqrt{\frac{3d\kappa}{N}} && (75)\\ \leq & \tilde{O}\left(H^2\left(d\sqrt{\log(N/\delta)}+\sqrt{Hd\log(N/(d\delta))}\right)\sqrt{\frac{3d\kappa}{N}}\right) && (76)\end{aligned}$$

E Lower-bound on best-of-both-world results

Proof of Theorem 3.4. Consider two instances of the offline RL problem, with two MDPs, M and M', both of which are actually simple two-arm bandit problems, along with their data generating distribution ν and ν', defined below.

1. Instance 1: Bandit M has $r_1 = \text{Bernoulli}(\frac{1}{2}+\frac{\varepsilon}{2p})$ and $r_2 = \text{Bernoulli}(\frac{1}{2})$. The data generating distribution is $\nu(a_1) = p$ and $\nu(a_2) = 1-p$. The relative condition number is $1/p$.
2. Instance 2: Bandit M has $r_1 = \text{Bernoulli}(\frac{1}{2}-\frac{\varepsilon}{2p})$ and $r_2 = \text{Bernoulli}(\frac{1}{2})$. The data generating distribution is $\nu(a_1) = p$ and $\nu(a_2) = 1-p$, same as instance 1. The relative condition number is $1/(1-p)$.

Let D and D' be i.i.d. datasets of size N generated by instances 1 and 2, respectively, generated by the following *coupling* process. First, the actions are sampled from ν and shared across instances, e.g. $N_D(a_1) = N_{D'}(a_1)$ and $N_D(a_2) = N_{D'}(a_2)$. Then, the rewards of a_2 are sampled from $\text{Bernoulli}(\frac{1}{2})$ and shared across tasks, e.g. $N_D(a_2,0) = N_{D'}(a_2,0)$ and $N_D(a_2,1) = N_{D'}(a_2,1)$.

Finally, let X_i, Y_i be Bernoulli random variables s.t. $X_i = \begin{cases}0 & U \leq \frac{1}{2}-\frac{\varepsilon}{2p}\\ 1 & \text{o.w.}\end{cases}$, $Y_i = \begin{cases}0 & U \leq \frac{1}{2}+\frac{\varepsilon}{2p}\\ 1 & \text{o.w.}\end{cases}$, where U is picked uniformly random in $[0,1]$. Then (X_i, Y_i) is a coupling with law: $P((X_i,Y_i)=(0,0)) = \frac{1}{2}-\frac{\varepsilon}{2p}$, $P((X_i,Y_i)=(1,0)) = 0$, $P((X_i,Y_i)=(0,1)) = \frac{\varepsilon}{2p}$, $P((X_i,Y_i)=(s_3,s_3)) = \frac{1}{2}-\frac{\varepsilon}{2p}$, X_i and Y_i can be thought as the outcome of $\text{Bernoulli}(\frac{1}{2}+\frac{\varepsilon}{2p})$, $\text{Bernoulli}(\frac{1}{2}+\frac{\varepsilon}{2p})$ respectively. Then, let the rewards of a_1 of the two instances be generated by Y_i and X_i respectively. We then have

$$P\Big(\sum_{i=1}^{N(a_1)}\mathbb{1}\left[X_i \neq Y_i\right]) \geq P(N(a_1)\leq pN)\cdot P(\sum_{i=1}^{pN}\mathbb{1}\left[X_i\neq Y_i\right]) \geq \frac{1}{2}\cdot\frac{1}{2} = \frac{1}{4} \quad (77)$$

In other words, with probability at least $\frac{1}{4}$, instance 1 and 2 are indistinguishable under ε-contamination, in particular the adversary can replace at most εN of $(a_1,0)$ with $(a_1,1)$ in D' to replicate D. Therefore, instance 1 and (instance 2 + ε-contamination) are with probability at least $1/4$ indistinguishable. Now, if an algorithm wants to achieve best of both world guarantee, it must return a_1 as the optimal arm with high probability when observing a dataset generated as above, in which case it will suffer a suboptimality of $\frac{\varepsilon}{2p}$ if the data is generated by (instance 2 + ε-contamination). As $p \geq \varepsilon \geq 0$ goes to 0, this gap blows up, while the relative condition number $1/(1-p)$ remains bounded, thus contradiction.

■

F Technical Lemmas

Lemma F.1 (Concentration of Self-Normalized Processes (Abbasi-Yadkori et al., 2011)). *Let $\{\varepsilon_t\}_{t=1}^{\infty}$ be a real-valued stochastic process that is adaptive to a filtration $\{\mathcal{F}_t\}_{t=0}^{\infty}$. That is, ε_t is $\mathcal{F}_t$-measurable for all $t \geq 1$. Moreover, we assume that, for any $t \geq 1$, conditioning on $\mathcal{F}_{t-1}$, ε_t is a zero-mean and σ-subGaussian random variable such that*

$$\mathbb{E}[\varepsilon_t|\mathcal{F}_{t-1}] = 0 \qquad \text{and} \qquad \mathbb{E}[\exp(\lambda\varepsilon_t)|\mathcal{F}_{t-1}] \leq \exp(\lambda^2\sigma^2/2), \quad \forall\lambda \in \mathbb{R}. \tag{78}$$

Besides, let $\{\phi_t\}_{t=1}^{\infty}$ be an $\mathbb{R}^d$-valued stochastic process such that ϕ_t is $\mathcal{F}_{t-1}$-measurable for all $t \geq 1$. Let $M_0 \in \mathcal{R}^{d\times d}$ be a deterministic and positive-definite matrix, and we define $M_t = M_0 + \sum_{s=1}^{t} \phi_s\phi_s^\top$ for all $t \geq 1$. Then for any $\delta > 0$, with probability at least $1-\delta$, we have for all $t \geq 1$ that

$$\Big\|\sum_{s=1}^{t} \phi_s \cdot \varepsilon_s\Big\|^2_{M_t^{-1}} \leq 2\sigma^2 \cdot \log\left(\frac{\det(M_t)^{1/2}\det(M_0)^{-1/2}}{\delta}\right).$$

Lemma F.2 (Extended Value Difference (Cai et al., 2020)). *Let $\pi = \{\pi_h\}_{h=1}^H$ and $\pi' = \{\pi'_h\}_{h=1}^H$ be two arbitrary policies and let $\{\hat{Q}_h\}_{h=1}^H$ be any given Q-functions. For any $h \in [H]$, we define a value function $\hat{V}_h : \mathcal{S} \to \mathbb{R}$ by letting $\hat{V}_h(x) = \langle \hat{Q}_h(x,\cdot), \pi_h(\cdot|x)\rangle_{\mathcal{A}}$ for all $s \in \mathcal{S}$. Then for all $s \in \mathcal{S}$, we have*

$$\hat{V}_1(s) - V_1^{\pi'}(s) = \sum_{h=1}^{H} \mathbb{E}_{\pi'}\left[\langle \hat{Q}_h(s_h,\cdot), \pi_h(\cdot|s_h) - \pi'_h(\cdot|s_h)\rangle_{\mathcal{A}} | s_1 = s\right] \tag{79}$$

$$+ \sum_{h=1}^{H} \mathbb{E}_{\pi'}\left[\hat{Q}_h(s_h,a_h) - (\mathbb{B}_h\hat{V}_{h+1})(s_h,a_h) | s_1 = s\right], \tag{80}$$

where the expectation $\mathbb{E}_{\pi'}$ is taken with respect to the trajectory generated by π', and $\mathbb{B}_h$ is the Bellman operator.

Lemma F.3 (Concentration of Covariances (Zanette et al., 2021)). *Let $\{\phi_i\}_{1:N} \subset \mathbb{R}^d$ be i.i.d. samples from an underlying bounded distribution ν, with $\|\phi_i\|_i \leq 1$ and covariance Σ. Define*

$$\Lambda = \sum_{i=1}^{N} \phi_i\phi_i^\top + \lambda \cdot I \tag{81}$$

for some $\lambda \geq \Omega(d\log(N/\delta))$. Then, we have that with probability at least $(1-\delta)$,

$$\frac{1}{3}(N\Sigma + \lambda I) \preceq \Lambda \preceq \frac{5}{3}(N\Sigma + \lambda I) \tag{82}$$

Proof. See (Zanette et al., 2021) Lemma 39 for a detailed proof. ■