Exact Policy Recovery in Offline RL with Both Heavy-Tailed Rewards and Data Corruption

Yiding Chen, [1] **Xuezhou Zhang,** [2] **Qiaomin Xie,** [1] **Xiaojin Zhu,** [1]

[1]UW-Madison, [2]Boston University

Abstract

We study offline reinforcement learning (RL) with heavy-tailed reward distribution and data corruption: (i) Moving beyond subGaussian reward distribution, we require only a bounded $(1+\gamma)$-th moment for $\gamma \in (0, 1]$; (ii) We allow corruptions where an attacker can arbitrarily modify ϵ-fraction of the rewards and transitions in the dataset. We first derive a sufficient optimality condition for generalized Pessimistic Value Iteration (PEVI), which allows various estimators with proper confidence bounds and can be applied to multiple learning settings. In order to handle the data corruption and heavy-tailed reward setting, we prove that the trimmed-mean estimation achieves a minimax optimal error rate $O\big(\sigma\epsilon^{\frac{\gamma}{1+\gamma}}\big)$ for robust mean estimation under heavy-tailed distributions. In the PEVI algorithm, we plug in the trimmed mean estimation and the confidence bound to solve the robust offline RL problem. Standard analysis reveals that data corruption induces a bias term $O\big(H\sigma\epsilon^{\frac{\gamma}{1+\gamma}} + \epsilon H\big)$ in the suboptimality gap, which gives the false impression that any data corruption prevents optimal policy learning. By using the optimality condition for the generalized PEVI, we show that as long as the bias term is less than the "action gap", the policy returned by PEVI achieves the optimal value given sufficient data.

1 Introduction

Reinforcement learning (RL) studies sequential decision-making in a potentially unknown environment (Sutton and Barto 2018). The success of RL requires sufficient interactions with the environment. Unlike RL with online interactions, offline RLs (Fujimoto, Meger, and Precup 2019; Laroche, Trichelair, and Des Combes 2019) utilize batch datasets without further interactions with the environment, which is preferred when there are abundant data generated by high-performing policies. However, offline RL becomes more challenging under data corruption (Eykholt et al. 2018; Neff 2016; Ma et al. 2019; Zhang et al. 2020) and heavy-tailed reward distributions (Bubeck, Cesa-Bianchi, and Lugosi 2013; Dubey et al. 2020), which is the topic of this paper.

Previous studies have primarily focused on problems under certain concentration assumptions, typically requiring that the rewards are bounded or follow distributions with subGaussian tails (Lattimore and Szepesvári 2020). However, there is growing evidence indicating that the subGaussianity assumption may not hold for many real-world scenarios (Arnold 2014; Liebeherr, Burchard, and Ciucu 2012; Borak, Härdle, and Weron 2005), challenging the applicability of algorithms designed solely for sub-Gaussian settings.

In terms of data corruption in RL, prior work (Zhang et al. 2022; Chen et al. 2022) showed that one can apply pessimistic value iteration (PEVI) with robust mean estimation to partially handle data corruption in offline RL, resulting in a policy $\hat{\pi}$ with suboptimality upper bound $\mathrm{SubOpt}(\hat{\pi}) \leq \tilde{O}\Big(\frac{\mathrm{poly}(H,\sigma)}{\sqrt{N}}\Big) + O(H\sigma\epsilon)$. Such an upper bound involves a term diminishing with sample size N and an irreducible bias term involves the corruption level ϵ. This implies that PEVI returns a suboptimal policy even with infinite data.

In this paper, we address the challenge of policy recovery in the presence of both heavy-tailed reward distributions and data corruption. We establish that Trimmed-mean estimation achieves the optimal error rate of $O\Big(\sigma\epsilon^{\frac{\gamma}{1+\gamma}} + \sigma N^{-\frac{\gamma}{1+\gamma}}\Big)$ for the robust mean estimation problem when confronted data corruption and heavy-tailed distribution. When using Trimmed-mean estimation as a subroutine, PEVI generates a nearly optimal policy. In particular, by utilizing the property of action gap, we show that $O(H\sigma\epsilon^{\frac{\gamma}{1+\gamma}} + \epsilon H) < \Delta^{\mathcal{A}}_{\min}$ is sufficient for the policy to achieve the optimal value even under corruption. We summarize our contributions as follows:

1. We show that a modified version of Trimmed-Mean estimation achieves *minimax-optimal* error guarantee for robust mean estimation problems with heavy-tailed distribution and data corruption. Importantly, we only require the distribution to have *bounded* $(1 + \gamma)$*-th centered moment* and allow the variance of the distribution to be infinite. Unlike the truncated empirical mean estimation in (Bubeck, Cesa-Bianchi, and Lugosi 2013), the trimmed mean estimator considered in our paper is both translation-invariant and robust to data corruption. As a result, we show that reward distribution with bounded $(1 + \gamma)$-th moment is sufficient to ensure the success of policy learning, which is a much weaker concentration assumption than the subGaussian or bounded variance assumption typically used in the literature.

2. We present a generalized PEVI and derive an optimality condition based on the action gap. In the offline learning setting with *heavy-tailed reward* and *data corruption*, we plug in the trimmed mean estimation for reward estimation. We show that given sufficient samples, $O(H\sigma\epsilon^{\frac{\gamma}{1+\gamma}} + \epsilon H) < \Delta^{\mathcal{A}}_{\min}$ ensures that the learner takes an optimal action in each state visited by some optimal policy and thus achieves the *optimal value*.

2 Related Work

RL and adversarial attack against RL: Reinforcement learning aims to find the optimal strategy in a Markov Decision Process (MDP) (Sutton and Barto 2018). In online RL, (Azar, Osband, and Munos 2017; Dann, Lattimore, and Brunskill 2017) show that the UCB-style algorithm achieves minimax regret bound. In offline RL, (Jin, Yang, and Wang 2021; Rashidinejad et al. 2021; Xie et al. 2021) use the pessimistic principle to design algorithms for offline policy learning. There are lines of work studying gap-dependent online (Simchowitz and Jamieson 2019; Xu, Ma, and Du 2021; Dann et al. 2021; Jonsson et al. 2020; Wagenmaker, Simchowitz, and Jamieson 2022) and offline (Wang, Cui, and Du 2022; Hu, Kallus, and Uehara 2021) RL. Our paper is closely related to the work on offline gap-dependent RL. However, our main objective is to characterize sufficient conditions for optimality under data corruption instead of optimal sample complexity.

Heavy-tailed bandits: There is a significant body of research dedicated to studying bandit problems under weak moment assumptions. For instance, (Bubeck, Cesa-Bianchi, and Lugosi 2013) focused on the mean multi-armed bandit (MAB) problem with heavy-tailed rewards and utilized robust mean estimation to develop a UCB algorithm that achieves logarithmic regret. The pure-exploration problem for MAB with heavy-tailed distributions was investigated by (Yu et al. 2018). Furthermore, (Medina and Yang 2016; Shao et al. 2018) explored the linear bandit problem with heavy-tailed noise distributions and proposed algorithms with nearly-optimal regret guarantees. (Dubey et al. 2020) examined this problem in the context of cooperative multi-agent settings.

Robust statistics: Robust statistics studies estimation with corrupted data (Huber 1992; Tukey 1960). Recent advances (Diakonikolas et al. 2019a; Lai, Rao, and Vempala 2016) design efficient algorithms for high-dimensional robust statistics. These techniques are applied to more general machine learning tasks, including linear regression (Diakonikolas, Kong, and Stewart 2019), supervised learning (Diakonikolas et al. 2019b; Prasad et al. 2018) and RL (Zhang et al. 2022, 2021). Our work utilizes robust mean estimation to defend data corruption in offline RLs.

Adversarial RL and robust RL: RL is vulnerable to adversarial attacks (Ma et al. 2019; Zhang et al. 2020; Huang et al. 2017; Sun et al. 2020; Behzadan and Munir 2017). Corruption robust RL performs policy learning under data corruption (Lykouris et al. 2021; Wei, Dann, and Zimmert 2022; Zhang et al. 2021, 2022; Chen et al. 2022), which usually results in a bias term in the performance guarantee due to the data corruption. (Niss and Tewari 2020; Kapoor, Patel, and Kar 2019) study multi-armed bandits under data corruption using robust statistics. They show that if the corruption level is not high enough to make the robust reward estimation of a suboptimal to be larger than that of an optimal arm, then the learner suffers only sublinear regret, which captures an optimal arm. We use this intuition to study offline RL under data corruption. There is a separate line of works studying distributionally robust RL problem (Shi and Chi 2022; Panaganti et al. 2022) where the state transition is specified by some uncertainty sets. Our setting is significantly different from this line of works.

3 Preliminary

MDP formulation: We consider a finite horizon episodic tabular Markov Decision Process (MDP) $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, H, p_0)$ with finite state space $|\mathcal{S}| = S$, finite action space $|\mathcal{A}| = A$, transition matrices $\mathcal{P} = \{P_h\}_{h=1}^H$, reward distributions $\mathcal{R} = \{\mathcal{R}_h\}_{h=1}^H$, and initial state distribution p_0. We assume the rewards are scholastic and the expectations of reward distributions are bounded in $[0,1]$, i.e. for all $(s,a,h) \in \mathcal{S} \times \mathcal{A} \times [H]$, $r_h(s,a) := \mathbb{E}_{R_h(s,a)\sim\mathcal{R}_h(s,a)}[R_h(s,a)] \in [0,1]$. Later on, we will study MDPs with different concentration assumptions on the reward distributions.

Policy and value function: A policy $\pi = \{\pi_h\}_{h=1}^H$ from a deterministic policy class Π is a sequence of deterministic functions that map from state to action: $\pi_h : \mathcal{S} \mapsto \mathcal{A}, \forall h$. The state value function of π is defined as $V_h^\pi(s) := \mathbb{E}\Big[\sum_{t=h}^H R_t(s_t, \pi_t(s_t)) \Big| s_t = s\Big]$. We similarly define the state-action value function: $Q_h^\pi(s,a) := \mathbb{E}[R_h(s,a)] + \mathbb{E}_{s_{h+1}\sim P_h(\cdot|s,a)}\big[V_{h+1}^\pi(s_{h+1})\big]$. The value of a policy is the expectation of $V_1^\pi(s)$ over the initial state distribution: $V_{p_0}^\pi := \mathbb{E}_{s_1\sim p_0}[V_1^\pi(s_1)]$. An *optimal policy* is one that simultaneously maximizes $V_h^\pi(s)$ for all h and s. We use $\Pi^* \subseteq \Pi$ to denote the set of all deterministic optimal policies. And we use $V_h^*(\cdot)$, $Q_h^*(\cdot,\cdot)$, $V_{p_0}^*$ to denote the state value function, state-action value function, and value of the optimal policies. We use $d_h^\pi(s) := \mathbb{E}_\pi[\mathbb{I}\{s_h = s\}]$, $d_h^\pi(s,a) := \mathbb{E}_\pi[\mathbb{I}\{(s_h,a_h) = (s,a)\}]$ to denote the state occupancy distribution and state-action occupancy distribution under policy π.

Performance measure: In this paper, we mainly focus on the offline setting and use the suboptimality gap as the performance measure for a policy: $\mathrm{SubOpt}(\pi) := V_{p_0}^* - V_{p_0}^\pi$. Our goal is to find a policy with a small suboptimality gap.

Policy gap and action gap: Among policies that fail to achieve the optimal value, the best one has the smallest suboptimality gap. We call this gap the *policy gap*: $\Delta^\Pi_{\min} := \min_{\pi\in\Pi: V_{p_0}^\pi < V_{p_0}^*} \mathrm{SubOpt}(\pi)$. In contrast, we define a more fine-grained *action gap* by $\Delta^{\mathcal{A}}_{\min} := \min_{(h,s,a):\Delta_h(s,a)>0, s\in\mathcal{S}_h} \Delta_h(s,a)$, where $\Delta_h(s,a) := V_h^*(s) - Q_h^*(s,a)$ and $\mathcal{S}_h := \big\{s \in \mathcal{S} : \exists \pi^* \in \Pi^*, \text{ s.t. } d_h^{\pi^*}(s) > 0\big\}$. For notation convenience we assume there is at least one (s,a,h) tuple s.t. $s \in \mathcal{S}_h$ and $\Delta_h(s,a) > 0$ to exclude trivial MDPs. A simi-

lar notion of $\Delta_{\min}^{\mathcal{A}}$ has been introduced in (Simchowitz and Jamieson 2019; Wang, Cui, and Du 2022), Our notation of $\Delta_{\min}^{\mathcal{A}}$ is a refinement over theirs where the minimum is over *only the* (s,h) *pairs covered by at least an optimal policy*. We can show that our action gap is always no less than policy gap, and the difference can be large:

Proposition 1. *For any MDP* $\mathcal{M}$*, there exists* $(\pi^*, s', h') \in \Pi^* \times \mathcal{S} \times [H]$*, s.t.* $d_{h'}^{\pi^*}(s') > 0$*, and* $\Delta_{\min}^{\Pi} \le d_{h'}^{\pi^*}(s')\Delta_{\min}^{\mathcal{A}} \le \Delta_{\min}^{\mathcal{A}}$.

Intuitively, by definition of $\Delta_{\min}^{\mathcal{A}}$, there exists a (s', a', h') tuples and an optimal policy π^* s.t. $\Delta_{h'}(s', a') = \Delta_{\min}^{\mathcal{A}}$ and $d_{h'}^{\pi^*}(s') > 0$. We can design a suboptimal policy $\tilde{\pi}$ by choosing the suboptimal action a' at state s' and step h' and follow π^* in all other states or steps. The suboptimality of $\tilde{\pi}$, $d_{h'}^{\pi^*}(s')\Delta_{\min}^{\mathcal{A}}$, depends on the state occupancy measure $d_{h'}^{\pi^*}(s')$. Because $\Delta_{\min}^{\Pi}$ is a lower bound on the suboptimality of all suboptimal policies, we conclude that $\Delta_{\min}^{\Pi} \le d_{h'}^{\pi^*}(s')\Delta_{\min}^{\mathcal{A}}$. $d_{h'}^{\pi^*}(s')$ can be very close to 0 in some MDPs, thus $\Delta_{\min}^{\Pi}$ can be much smaller than $\Delta_{\min}^{\mathcal{A}}$.

4 Sufficient Condition for Exact Optimal Policy Recovery in Offline RL

In this section, we provide a sufficient condition for exact optimal policy recovery in offline RL. Our characterization is based on the well-known PEVI algorithm (Jin, Yang, and Wang 2021), we slightly generalize it in Algorithm 1 to decouple RL from the estimators on mean rewards and transitions. This enables us to plug in different estimators later based on specific data assumptions, such as when the data is drawn from heavy-tailed distributions or adversarially corrupted. We then achieve different exact optimal policy recovery guarantees accordingly. Concretely, Algorithm 1 calls a REWARDESTIMATOR f to obtain a confidence interval $\hat{r}_h(s,a) \pm b_h^1(s,a)$ for the reward $r_h(s,a)$, and a TRANSITIONESTIMATOR g to obtain a confidence interval $\widehat{\mathrm{PV}}_{h,s,a} \pm b_h^2(s,a)$ for the expectation of a vector $\underline{V}_h$ under the transition multinomial $P_{h,s,a}$. These estimators will be instantiated differently in Section 5 based on different data assumptions. The notation $\mathcal{D}_{r|hsa}$ stands for the set of reward values observed at stage h in state s under action a in the offline dataset; similarly for the set of next states $\mathcal{D}_{s'|hsa}$.

If the sum of confidence bound $b_h^1(s,a) + b_h^2(s,a)$ is uniformly bounded on (s,a,h) tuples that are covered by the optimal policies, we can get a clean suboptimality guarantee for Algorithm 1:

Theorem 1 (Bound on suboptimality)**.** *Suppose for all* $(s,a,h) \in \mathcal{S} \times \mathcal{A} \times [H]$*, with probability at least* $1 - \frac{\delta}{SAH}$*, we have:*

$$|\hat{r}_h(s,a) - r_h(s,a)| \le b_h^1(s,a)$$
$$\left|\widehat{\mathrm{PV}}_{h,s,a} - P_{h,s,a}^{\top}\underline{V}_{h+1}\right| \le b_h^2(s,a).$$

If $\forall (s,a,h) \in \big\{(s,a,h) : \exists \pi^* \in \Pi^*,\ s.t.\ d_h^{\pi^*}(s,a) > 0\big\}$*, we have* $b_h^1(s,a) + b_h^2(s,a) \le b$*, then with probability at least* $1-\delta$*,* $\hat{\pi}$ *returned by Algorithm 1 satisfies*

$$\mathrm{SubOpt}(\hat{\pi}) \le 2Hb. \tag{1}$$

Algorithm 1: Generalized PEVI

Input: dataset $\mathcal{D} = \bigcup_{h=1}^{H}\left\{\left(s_{h,i}, a_{h,i}, r_{h,i}, s'_{h,i}\right)\right\}_{i=1}^{N}$. confidence level δ.
Set $\underline{Q}_{H+1}(s,a) = 0$, $\underline{V}_{H+1}(s) = 0$ for all (s,a)
for $h = H, \ldots, 1$ **do**
  for $s \in \mathcal{S}$, $a \in \mathcal{A}$ **do**
    $\big(\hat{r}_h(s,a), b_h^1(s,a)\big) \leftarrow f(\mathcal{D}_{r|hsa}, \frac{\delta}{2SAH})$
    $\left(\widehat{\mathrm{PV}}_{h,s,a}, b_h^2(s,a)\right) \leftarrow g(\mathcal{D}_{s'|hsa}, \underline{V}_{h+1}, \frac{\delta}{2SAH})$
    $\underline{Q}_h(s,a) = \max(0, \hat{r}_h(s,a) - b_h^1(s,a) + \widehat{\mathrm{PV}}_{h,s,a} - b_h^2(s,a))$
  end for
  for $s \in \mathcal{S}$ **do**
    $\underline{V}_h(s) = \max_{a \in \mathcal{A}} \underline{Q}_h(s,a)$
    $\hat{\pi}_h(s) = \mathrm{argmax}_{a \in \mathcal{A}} \underline{Q}_h(s,a)$
  end for
end for
Return: $\hat{\pi}$.

When the confidence bounds are small enough, the estimation for value function in Algorithm 1 will be accurate and $\hat{\pi}$ will choose the optimal action in each state with positive occupancy measure. With this intuition, we get a sufficient condition for optimality:

Theorem 2 (Optimality condition)**.** *Under the conditions in Theorem 1, if* $2Hb < \Delta_{\min}^{\mathcal{A}}$*, then* $\mathrm{SubOpt}(\hat{\pi}) = 0$ *with probability at least* $1 - \delta$.

Theorem 2 provides a general condition for optimal policy identification, which results in different guarantees given different estimators and corresponding confidence bounds. One can also derive an optimality condition using policy gap $\Delta_{\min}^{\Pi}$: because the set of deterministic optimal policies Π^* is discrete, when (1) is less than $\Delta_{\min}^{\Pi}$, $\mathrm{SubOpt}(\hat{\pi}) = 0$. However, this argument usually results in an overly conservative optimality condition. We defer the detailed discussion to Section 6.

In the case of learning with i.i.d. offline dataset, Theorem 4.1 of (Wang, Cui, and Du 2022) provides a dedicated sample complexity guarantee for offline optimal policy identification when the rewards are deterministic and known. Under a similar i.i.d. learning setting but with subGaussian rewards, we show, in Section 5.1, that when the reward and transition estimators f and g are specified to be empirical mean estimators with Hoeffding-style confidence bound, Theorem 2 provides a similar sample complexity bound. However, our main focus is to use Algorithm 1 to study the robust offline learning setting in Section 5.2, which is much more challenging.

5 Case Studies

The meta-algorithm Algorithm 1 and its theoretical guarantee in Section 4 can be applied to various data generative models and reward distributions given estimators with proper confidence bounds. In this section, we present two

case studies. We start with a standard learning setting in Section 5.1 as a warm-up where the dataset consists of i.i.d. samples and the reward distributions are subGaussian; we then present our main result in Section 5.2 with a harder learning setting where the dataset can be corrupted and reward distributions are heavy-tailed. In both case studies, we provide sufficient conditions for optimality derived using Theorem 2.

5.1 Warm-up: i.i.d. dataset with subGaussian rewards

We first consider the standard offline learning setting with an i.i.d. dataset and a subGaussian rewards distribution. The exact policy recovery condition is known (Wang, Cui, and Du 2022), but our purpose here is to illustrate how one can instantiate Theorem 2 with f, g, in anticipation of our main result in the next section. We assume the reward distributions are subGausian:

Assumption 1 (SubGaussian rewards)**.** *For all $(s,a,h) \in \mathcal{S}\times\mathcal{A}\times[H]$, $\mathcal{R}_h(s,a)$ is subGaussian with mean $r_h(s,a) := \mathbb{E}_{X\sim\mathcal{R}_h(s,a)}[X] \in [0,1]$ and parameter σ^2, $\sigma > 0$, i.e. $\mathbb{E}_{X\sim\mathcal{R}_h(s,a)}[\exp(s(X - r_h(s,a)))] \le \exp\big(\sigma^2 s^2/2\big)$, for all $s \in \mathbb{R}$.*

In our offline learning setting, we consider the data generative model similar to (Wang, Foster, and Kakade 2020), where the learning agent has access to an offline dataset drawn from some data distribution but cannot have further interaction with the MDP. The i.i.d. dataset is generated as a set of transition tuples instead of trajectories. Specifically,

Definition 1 (Offline dataset)**.** *An* offline dataset $\mathcal{D}$ *of size N collected with data distributions $\mu = \{\mu_h\}_{h\in[H]}$ is a multiset consisting of N transition tuples sampled at each time step:*

$$\mathcal{D} = \bigcup_{h=1}^{H} \big\{\big(s_{h,i}, a_{h,i}, r_{h,i}, s'_{h,i}\big)\big\}_{i=1}^{N}$$

where $(s_{h,i}, a_{h,i}) \sim \mu_h$, $r_{h,i} \sim \mathcal{R}_h(s_{h,i}, a_{h,i})$ and $s'_{h,i} \sim P_h(\cdot \mid s_h^i, a_h^i)$.

We assume the data distribution μ has *uniform coverage on all optimal policies*:

Assumption 2 (Uniform optimal policy coverage)**.** *There exists $P > 0$, s.t. $\mu_h(s,a) \ge P$, for all $(s,a,h) \in \big\{(s,a,h) : \exists \pi^* \in \Pi^*,\ s.t.\ d_h^{\pi^*}(s,a) > 0\big\}$.*

As shown in Section D of (Wang, Cui, and Du 2022), this assumption is necessary for optimal policy recovery.

Under this standard offline learning setting, it is sufficient to use empirical mean estimator in both the reward estimator and transition estimator:

$$\hat{r}^{\text{emp}}_{h,s,a} = \frac{1}{N_h(s,a)} \sum_{r\in\mathcal{D}_{r|hsa}} r \tag{2}$$

$$\widehat{\text{PV}}^{\text{emp}}_{h,s,a} = \frac{1}{N_h(s,a)} \sum_{s'\in\mathcal{D}_{s'|hsa}} \underline{V}_{h+1}(s'), \tag{3}$$

where $N_h(s,a) = \big|\mathcal{D}_{r|hsa}\big| = \big|\mathcal{D}_{s'|hsa}\big|$. We use the convention that $0/0 = 0$. The confidence bounds are given by the following lemma:

Proposition 2 (Confidence bound)**.** *If Assumption 1 holds, then for all $(s,a,h) \in \mathcal{S}\times\mathcal{A}\times[H]$, with probability at least $1 - \frac{\delta}{2SAH}$:*

$$\Big|\hat{r}^{\text{emp}}_{h,s,a} - r_h(s,a)\Big| \le b^{1,\text{emp}}_{h,s,a},$$
$$\Big|\widehat{\text{PV}}^{\text{emp}}_{h,s,a} - P^{\top}_{h,s,a}\underline{V}_{h+1}\Big| \le b^{2,\text{emp}}_{h,s,a}.$$

where

$$b^{1,\text{emp}}_{h,s,a} = \sigma\sqrt{\frac{2\log\frac{8SAH}{\delta}}{N_h(s,a)}}, b^{2,\text{emp}}_{h,s,a} = H\sqrt{\frac{\log\frac{8SAH}{\delta}}{2N_h(s,a)}}$$

In this case study, the reward and transition estimators are defined to be:

$$f_{\text{emp}}\left(\mathcal{D}_{r|hsa}, \frac{\delta}{2SAH}\right) := \left(\hat{r}^{\text{emp}}_{h,s,a}, b^{1,\text{emp}}_{h,s,a}\right)$$
$$g_{\text{emp}}\left(\mathcal{D}_{s'|hsa}, \underline{V}_{h+1}, \frac{\delta}{2SAH}\right) := \left(\widehat{\text{PV}}^{\text{emp}}_{h,s,a}, b^{2,\text{emp}}_{h,s,a}\right)$$

Given the reward estimator and transition estimator, we can get the following optimality condition by applying Theorem 2:

Proposition 3 (Optimality condition)**.** *Suppose Assumption 1, 2 holds. We specify the reward and transition estimators in Algorithm 1 to be f_{emp} and g_{emp}. Let $\hat{\pi}$ be the policy returned by Algorithm 1 given an offline dataset $\mathcal{D}$ generated according to Definition 1. If $4H(2\sigma + H)\frac{\log\frac{8SAH}{\delta}}{\sqrt{NP}} < \Delta^{\mathcal{A}}_{\min}$, then $\text{SubOpt}(\hat{\pi}) = 0$ with probability at least $1-\delta$.*

Proposition 3 translates Theorem 2 to a sample complexity bound by using empirical mean estimation with Hoeffding-style confidence bound. This result is similar to Theorem 4.1 of (Wang, Cui, and Du 2022) but with a slightly worse dependence on H. We are now ready to present our main results in the robust offline learning setting.

5.2 Main results: corrupted dataset and heavy-tailed reward distributions

When (i) the reward distributions have weaker concentrations, and (ii) the dataset is corrupted, the learning problem becomes more challenging. Nonetheless, Algorithm 1 can be adapted to this setting by using powerful robust estimators. We first provide a novel analysis that allows an existing robust estimator to handle *unbounded variance* and *data corruption*, then instantiate the exact policy recovery condition under this estimator.

Formally, we first relax the SubGaussian reward assumption in Assumption 1 by only assuming the reward distributions to have bounded $(1+\gamma)$-th centered moment:

Assumption 3 (Heavy-tailed reward distributions)**.** *There exsits $\gamma \in (0,1]$ and $\sigma > 0$, s.t. for all $(s,a,h) \in \mathcal{S}\times\mathcal{A}\times[H]$, $\mathbb{E}_{X\sim\mathcal{R}_h(s,a)}\Big[(X - r_h(s,a))^{1+\gamma}\Big] \le \sigma^{1+\gamma}$, where $r_h(s,a) = \mathbb{E}_{X\sim\mathcal{R}_h(s,a)}[X] \in [0,1]$.*

(Bubeck, Cesa-Bianchi, and Lugosi 2013) first studies this reward distribution in multi-armed bandits. The reward

distributions may not have finite variance, making the reward estimation itself a hard problem, even given clean data without data corruption. (Bubeck, Cesa-Bianchi, and Lugosi 2013) shows that empirical mean estimator results in a significantly wider confidence interval, which is not satisfactory. In this section, we study offline RL with a corrupted dataset, on top of this heavy-tailed reward model. Specifically, we consider an ϵ-corruption model on the offline dataset where **both rewards and transitions** can be corrupted, which is much more challenging than the learning problem in Definition 1:

Definition 2 (ϵ-corruption model)**.** *Let $\epsilon \geq 0$. An ϵ-corrupted offline dataset $\mathcal{D}$ is a multiset generated by the following procedure: a clean offline dataset $\tilde{\mathcal{D}} = \bigcup_{h=1}^{H}\left\{\left(s_{h,i}, a_{h,i}, \tilde{r}_{h,i}, \tilde{s}'_{h,i}\right)\right\}_{i=1}^{N}$ is generated according to Definition 1; an adversary is allowed to inspect the whole dataset $\tilde{\mathcal{D}}$ and replace up to ϵ fraction of the reward entries and transition entries with something arbitrary for each (s,a,h) tuple. We denote the corrupted dataset as $\mathcal{D} = \bigcup_{h=1}^{H}\left\{\left(s_{h,i}, a_{h,i}, r_{h,i}, s'_{h,i}\right)\right\}_{i=1}^{N}$. In other words, we require $\frac{\sum_{i=1}^{N}\mathbb{I}\{(s_{h,i},a_{h,i})=(s,a), r_{h,i}\neq\tilde{r}_{h,i}\}}{N_h(s,a)} \leq \epsilon$ and $\frac{\sum_{i=1}^{N}\mathbb{I}\{(s_{h,i},a_{h,i})=(s,a), s'_{h,i}\neq\tilde{s}'_{h,i}\}}{N_h(s,a)} \leq \epsilon$ for all (s,a,h).*

In the robust learning setting defined in Definition 2, the corrupted rewards can be unbounded. And importantly, the learning agent has no access to the clean dataset $\tilde{\mathcal{D}}$ and can only learn from the corrupted dataset $\mathcal{D}$.

Similar to Section 5.1, our first step is to design RewardEstimator f and TransitionEstimator g with proper confidence bound for Algorithm 1. We first formally define the robust mean estimation problem, which captures the hardness of the reward estimation problem:

Definition 3 (Robust mean estimation with heavy-tailed distribution)**.** *Let $\gamma \in (0,1]$, $\sigma \geq 0$, $\epsilon \in (0,1)$. Let $\mathcal{P}$ be a heavy-tailed distribution in $\mathbb{R}$ with bounded $(1+\gamma)$-th centered moment: $\mathbb{E}_{X\sim\mathcal{P}}\left[|X-\mu|^{1+\gamma}\right] \leq \sigma^{1+\gamma}$, where $\mu := \mathbb{E}_{X\sim\mathcal{P}}[X]$. Given an i.i.d. dataset $\tilde{X}_1, \ldots, \tilde{X}_N$ drawn from $\mathcal{P}$, an adversary can inspect the dataset and replace an ϵ-faction of the data points with arbitrary values. The corrupted dataset $X_1, \ldots, X_N$ is revealed to the learning algorithm, which attempts to estimate μ, the mean of $\mathcal{P}$.*

Trimmed Mean estimation is a well-studied estimator in robust statistics (Lugosi and Mendelson 2021, 2019). However, most prior work are limited to distributions with *sub-Gaussian* distribution or at most distribution with *bounded variance*. Surprisingly, we show that the Trimmed Mean estimator in (Lugosi and Mendelson 2021) can be directly applied to robust mean estimation in Definition 3 and resolves both difficulties simultaneously. For completeness, we present the Trimmed Mean estimator: Trimmed-Mean in Algorithm 2 in Appendix A.

Theorem 3 (Trimmed-Mean for heavy-tailed distribution)**.** *Suppose $\gamma \in (0,1]$, $\epsilon < \frac{1}{32}$, $\delta \in (0,1)$ and $N > 96\log\frac{4}{\delta}$. Given N samples generated by the ϵ-corruption model in Definition 3, Algorithm 2 outputs a $\hat{\mu}$, s.t. with probability at least $1-\delta$, $|\hat{\mu}-\mu| \leq C_{1,\gamma}\sigma\epsilon^{\frac{\gamma}{1+\gamma}} + C_{2,\gamma}\sigma\left(\frac{1}{N}\log\frac{8}{\delta}\right)^{\frac{\gamma}{1+\gamma}}$, where $C_{1,\gamma} = 128A_\gamma$, $C_{2,\gamma} = 768A_\gamma$ and A_γ is the smallest value s.t. $A_\gamma((1+x)\log(1+x)-x) \geq x^{\frac{\gamma+1}{\gamma}}/\left(1+x^{\frac{1}{\gamma}}\right)$ for all $x>0$.*

The error bound in Theorem 3 involves a bias term $O\left(\sigma\epsilon^{\frac{\gamma}{1+\gamma}}\right)$ and a statistical error term $\tilde{O}\left(\sigma N^{-\frac{\gamma}{1+\gamma}}\right)$. The bias is caused by data corruption why the statistical error term is due to finite sample. Importantly, both bias and statistical error term *meets the information-theoretic lower bound* (up to constants). Our new analysis is based on a variant of *Bernstein inequality under weak moment assumption*. We defer the details and more discussion about Theorem 3 to the end of this section.

We use the Trimmed-Mean estimator in Algorithm 2 and its confidence bound for reward estimation to handle the corrupted reward. The estimated reward is set to be: for all $(s,a,h) \in \mathcal{S}\times\mathcal{A}\times[H]$,

$$\hat{r}^{\mathrm{TM}}_{h,s,a} = \textsc{Trimmed-Mean}\left(\mathcal{D}_{r|hsa}, \epsilon, \frac{\delta}{4SAH}\right), \quad (4)$$

recall that $\mathcal{D}_{r|hsa}$ is the set of all rewards received in (s,a) visitations at step h. We use the same empirical mean estimator in (3) but with modified confidence bound to account for the effect of data corruption on the state transition. Formally, we have:

Proposition 4 (Confidence bound)**.** *If Assumption 3 holds, then for all $(s,a,h) \in \mathcal{S}\times\mathcal{A}\times[H]$, with probability at least $1-\frac{\delta}{2SAH}$:*

$$\begin{aligned}\left|\hat{r}^{\mathrm{TM}}_{h,s,a} - r_h(s,a)\right| \leq & b^{1,\mathrm{TM}}_{h,s,a}\\ \left|\widehat{\mathrm{PV}}^{\mathrm{emp}}_{h,s,a} - P^{\top}_{h,s,a}\underline{V}_{h+1}\right| \leq & b^{2,\mathrm{robust}}_{h,s,a}.\end{aligned}$$

where $\widehat{\mathrm{PV}}^{\mathrm{emp}}_{h,s,a}$ is defined in (3) *and*

$$b^{1,\mathrm{TM}}_{h,s,a} = \begin{cases}\infty & \text{if } N_h(s,a) \leq 96\log\frac{8SAH}{\delta}\\ C_{1,\gamma}\sigma\epsilon^{\frac{\gamma}{1+\gamma}} + C_{2,\gamma}\sigma\left(\frac{\log\frac{32SAH}{\delta}}{N_h(s,a)}\right)^{\frac{\gamma}{1+\gamma}} & \text{o.w.}\end{cases} \quad (5)$$

$$b^{2,\mathrm{robust}}_{h,s,a} = \epsilon H + H\sqrt{\frac{\log\frac{8SAH}{\delta}}{2N_h(s,a)}}, \quad (6)$$

where $C_{1,\gamma}$ and $C_{2,\gamma}$ are specified in Theorem 3.

$b^{1,\mathrm{TM}}_{h,s,a}$ is the confidence bound for the Trimmed-Mean estimator when applied to reward estimation. The success of the Trimmed-Mean estimation requires a minimum number of samples. So we simply set $b^{1,\mathrm{TM}}_{h,s,a}$ to ∞ when $N_h(s,a)$ is less than the threshold. Setting $b^{1,\mathrm{TM}}_{h,s,a}$ to ∞ looks excessive at the first glance. However, by Theorem 1, we can see that the suboptimality of $\hat{\pi}$ only depends on the bonus for (s,a,h) tuples covered by some optimal policy. By Assumption 2, the sample size requirement of Trimmed-Mean is met with high probability for any (s,a,h) tuples covered

by some optimal policy when N, the number of samples, is large enough.

In this case study, the reward and transition estimators are defined to be:

$$f_{\text{robust}}\left(\mathcal{D}_{r|hsa}, \frac{\delta}{2SAH}\right) := \left(\hat{r}^{\text{TM}}_{h,s,a}, b^{1,\text{TM}}_{h,s,a}\right)$$

$$g_{\text{robust}}\left(\mathcal{D}_{s'|hsa}, \underline{V}_{h+1}, \frac{\delta}{2SAH}\right) := \left(\widehat{\text{PV}}^{\text{emp}}_{h,s,a}, b^{2,\text{robust}}_{h,s,a}\right)$$

By applying Theorem 2, we get the following optimality condition:

Theorem 4 (Optimality condition)**.** *Suppose Assumption 3, 2 holds and $\epsilon < \frac{1}{32}$, $N > \frac{768}{P}\left(\log\frac{8SA}{\delta}\right)^2$. We specify the reward and transition estimators in Algorithm 1 to be f_{robust} and g_{robust}. Let $\hat{\pi}$ be the policy returned by Algorithm 1 given an offline dataset $\mathcal{D}$, where $\mathcal{D}$ is generated according to Definition 2. If $2H\left(C_{1,\gamma}\sigma\epsilon^{\frac{\gamma}{1+\gamma}} + \epsilon H\right) + 4H\left(\frac{\sqrt{2}C_{2,\gamma}\sigma}{(NP)^{\frac{\gamma}{1+\gamma}}} + \frac{H}{\sqrt{NP}}\right)\log\frac{32SAH}{\delta} < \Delta^{\mathcal{A}}_{\min}$, then $\text{SubOpt}(\hat{\pi}) = 0$ with probability at least $1-\delta$.*

There are two terms on the LHS of the optimality condition in Theorem 4: the first term $2H\left(C_{1,\gamma}\sigma\epsilon^{\frac{\gamma}{1+\gamma}} + \epsilon H\right)$ involves the corruption level ϵ, which characterizes the bias caused by data corruption; the second term $4H\left(\frac{\sqrt{2}C_{2,\gamma}\sigma}{(NP)^{\frac{\gamma}{1+\gamma}}} + \frac{H}{\sqrt{NP}}\right)\log\frac{32SAH}{\delta}$ involves N, the size of the dataset, which characterizes the statistical error. If

$$2H\left(C_{1,\gamma}\sigma\epsilon^{\frac{\gamma}{1+\gamma}} + \epsilon H\right) < \Delta^{\mathcal{A}}_{\min} \tag{7}$$

then for N large enough, the optimality condition holds with high probability. This implies a key difference between robust RL and robust mean estimation: in robust mean estimation, it is never possible to learn the true mean even regardless of sample size due to the data corruption (Lai, Rao, and Vempala 2016); however, in robust RL, $\Delta^{\mathcal{A}}_{\min}$ creates a quantization effect, enabling the exact identification of a policy with the optimal value despite minor corruption. This is reassuring because we can still aim to find a policy with the optimal value as long as (7) holds.

More discussion on Theorem 3 and the minimax optimality (Bubeck, Cesa-Bianchi, and Lugosi 2013) provides a Median-of-Means estimator and a truncated empirical mean estimator for the mean estimation problem under heavy-tailed distribution, both are designed without the consideration of data corruption. The Median-of-Means estimator achieves the same rate as Theorem 3 for $\epsilon = 0$. Their truncated empirical mean estimator requires the uncentered moment $\mathbb{E}_{X\sim\mathcal{P}}\left[|X|^{1+\gamma}\right]$ to be bounded by some constant u, which increases as μ moves away from 0. However, this assumption leads to their error bound blowing up as u increases. In contrast, our algorithm handles data corruption and the error bound in Theorem 3 is translation invariant w.r.t. μ, which makes it significantly stronger.

Importantly, Algorithm 2 is minimax optimal up-to some constant:

Theorem 5 (Error lower bound of the learning problem in Theorem 3)**.** *Given any learning algorithm $\mathscr{A}$, $\sigma > 0, \epsilon > 0$ and sufficiently large $N \in \mathbb{Z}_+$, there exists a distribution $\mathcal{P}$ with bounded $(1+\gamma)$-th centered moment and an adversary satisfying the constraints in Definition 3, s.t. any learning algorithm, given N data points from $\mathcal{P}$ with ϵ-fraction of corruption, will suffer an error at least $\Omega\left(\sigma\epsilon^{\frac{\gamma}{1+\gamma}} + \sigma N^{-\frac{\gamma}{1+\gamma}}\right)$ with at least constant probability.*

When $\epsilon = 0$, Theorem 5 implies the following error lower bound for mean estimation problem with i.i.d. data from a distribution with bounded $(1+\gamma)$-th centered moment:

Corollary 1. *Given any σ and sufficiently large N, there exists a distribution $\mathcal{D}$ with bounded $(1+\gamma)$-th centered moment, s.t. given N i.i.d. samples from the distribution, any learning algorithm will suffer an error at least $\Omega\left(\sigma N^{-\frac{\gamma}{1+\gamma}}\right)$ with at least constant probability.*

(Lugosi and Mendelson 2021) guarantees an error $\tilde{O}(\sigma\sqrt{\epsilon} + \sigma/\sqrt{N})$ for the case when $\gamma = 1$, which is captured by Theorem 3. When $\gamma < 1$, our Theorem 3 provides a larger bias term of $O\left(\sigma\epsilon^{\frac{\gamma}{1+\gamma}}\right)$ and a slower convergence rate of $O(\sigma N^{-\frac{\gamma}{1+\gamma}})$. As shown in Theorem 5, these discrepancies are consequences of the inherent difficulty of the learning problem. The weaker moment assumption makes the estimation more challenging, leading to a larger error.

Proof sketch of Theorem 3 Algorithm 2 chooses $\tilde{\epsilon} = \tilde{O}(\epsilon + 1/N)$ as the trimming portion. It first splits the sample into two batches: D_1 and D_2. The trimming threshold α, β are set to be the $\tilde{\epsilon}$ and $(1-\tilde{\epsilon})$-quantile of D_1. The algorithm use α, β to define a clipping function $\phi_{\alpha,\beta}(\cdot)$, s.t. $\phi_{\alpha,\beta}(x) = \beta$ if $x > \beta$; $\phi_{\alpha,\beta}(x) = x$ if $\alpha \le x \le \beta$; $\phi_{\alpha,\beta}(x) = \alpha$ if $x < \alpha$. The algorithm simply returns the truncated mean of D_2: $\hat{\mu} = \frac{1}{|D_2|}\sum_{x\in D_2}\phi_{\alpha,\beta}(x)$.

In the proof of Theorem 3, we derive a novel Bernstein's inequality under weak moment assumption as a key lemma and conduct a refined analysis on the quantile of the heavy-tailed distribution. The remaining parts of the proof of Theorem 3 follow the main steps in Proof of Theorem 1 in (Lugosi and Mendelson 2021). We first present the variant of Bernstein's inequality below:

Lemma 1 (Bernstein's inequality under weak moment assumption)**.** *Suppose X_j, $j = 1, \ldots, n$ is a sequence of independent zero-mean random variable bounded by $|X_j| \le M$ and there exists $\gamma \in (0, 1]$, s.t.*

$$\mathbb{E}\,|X_j|^{1+\gamma} \le \sigma^{1+\gamma}, \text{ for all } j = 1, \ldots, n.$$

then there exists $A_\gamma \ge 1$ (depending only on γ) s.t.:

$$\mathbb{P}\left(\frac{1}{n}\sum_{j=1}^{n} X_j > t\right) \le \exp\left\{-\frac{n}{A_\gamma}\frac{t^{\frac{\gamma+1}{\gamma}}}{\sigma^{\frac{1+\gamma}{\gamma}} + Mt^{\frac{1}{\gamma}}}\right\}.$$

Let $\tilde{D}_1 \cup \tilde{D}_2$ be the uncorrupted dataset. The estimation

error of $\hat{\mu}$ can be decomposed as:

$$\begin{aligned}|\hat{\mu}-\mu| \leq & \left|\frac{1}{|D_2|}\sum_{x\in D_2}\phi_{\alpha,\beta}(x)-\frac{1}{\left|\tilde{D}_2\right|}\sum_{x\in\tilde{D}_2}\phi_{\alpha,\beta}(x)\right| \\ & +\left|\frac{1}{\left|\tilde{D}_2\right|}\sum_{x\in\tilde{D}_2}\phi_{\alpha,\beta}(x)-\mathbb{E}_{X\sim\mathcal{P}}[\phi_{\alpha,\beta}(X)]\right| \\ & +\left|\mathbb{E}_{X\sim\mathcal{P}}[\phi_{\alpha,\beta}(X)]-\mu\right| \\ =: & B_1+B_2+B_3\end{aligned}$$

Because $\tilde{D}_2$ and D_2 differ by at most $2\epsilon\left|D_2\right|$ entries,

$$B_1 \leq 2\epsilon \max_{x,y\in\mathbb{R}}|\phi_{\alpha,\beta}(x)-\phi_{\alpha,\beta}(y)| = 2\epsilon(\beta-\alpha).$$

Because $\left\{\phi_{\alpha,\beta}(x) : x\in\tilde{D}_2\right\}$ consists i.i.d samples from a distribution with bounded $(1+\gamma)$-th centered moment and support bounded between $[\alpha,\beta]$, by Lemma 4:

$$B_2 \leq \tilde{O}\left(\frac{\sigma}{N^{\frac{\gamma}{1+\gamma}}}+\frac{|\beta-\alpha|}{N}\right)$$

By concentration of Bernoulli random variables, α and β are close to the $\tilde{\epsilon}$ and $(1-\tilde{\epsilon})$-quantile of distribution $\mathcal{P}$. Furthermore, we can show that the truncated random variable $\phi_{\alpha,\beta}(X)$, where $X\sim\mathcal{P}$, has a mean close to the original random variable:

$$B_3 \leq O\left(\sigma\tilde{\epsilon}^{\frac{\gamma}{1+\gamma}}\right).$$

We finish the proof by combining these together.

6 Comparison between Different Optimality Conditions

In Section 4. we derive an optimality condition based on the suboptimal gap of actions in Theorem 2. Alternatively, we can get another optimality condition with the following observation: $\hat{\pi}$ is optimal if the suboptimality gap $\mathrm{SubOpt}(\hat{\pi})$ is less than the policy gap $\Delta^{\Pi}_{\min}$. Formally, we can get the following sufficient condition for optimality with Theorem 1:

Proposition 5 (Optimality condition)**.** *Under the conditions in Theorem 1, if* $2Hb < \Delta^{\Pi}_{\min}$*, then* $\mathrm{SubOpt}(\hat{\pi}) = 0$ *with probability at least* $1-\delta$*.*

By Proposition 1, the action gap $\Delta^{\mathcal{A}}_{\min} \leq \Delta^{\Pi}_{\min}$ and the difference can be large. This means the condition in Proposition 5 is usually more conservative and thus less preferable than that in Theorem 2. In the following, we use contextual bandit as an illustrative example to show that why utilizing the action gap idea leads to a better sufficient condition.

When $H=1$, MDP is specialized to contextual bandit. And Algorithm 1 returns a policy $\hat{\pi}$ that chooses the action with the largest lower confidence bound (LCB) in each state. Similar to the discussion above, we can make sure $\hat{\pi}$ is optimal by comparing either the action gap or policy gap. We will show that utilizing the action gap is preferable.

In contextual bandit, the action gap can be written as:

$$\Delta^{\mathcal{A}}_{\min} = \min_{(s,a)\in\mathcal{C}}(r_1(s,\pi^*(s))-r_1(s,a)),$$

where π^* is an optimal policy and

$$\mathcal{C} := \{(s,a) : s\in\mathrm{supp}(p_0), r_1(s,a)\neq r_1(s,\pi^*(s))\}.$$

Because the best *suboptimal* policy should only choose a suboptimal action in one state, we can write the policy gap as:

$$\Delta^{\Pi}_{\min} = \min_{(s,a)\in\mathcal{C}} p_0(s)(r_1(s,\pi^*(s))-r_1(s,a)).$$

Because $p_0(s)$ can be very small for some state s, the policy gap $\Delta^{\Pi}_{\min}$ can be much smaller than the action gap $\Delta^{\mathcal{A}}_{\min}$.

Since there is no state transition in contextual bandits, $b^2_1(\cdot,\cdot)=0$ and the value function estimation in Algorithm 1 can be written as:

$$\underline{Q}_1(s,a) = \max\{0, \hat{r}_1(s,a)-b^1_1(s,a)\} \quad \forall s,a \tag{8}$$

By the definition of $\hat{\pi}$ and the fact that $b_1(\cdot,\cdot)$ is a proper confidence bound, with probability at least $1-\delta/4$, the suboptimality of $\hat{\pi}$ at any s can be bounded by:

$$\begin{aligned}&V^*_1(s)-Q_1(s,\hat{\pi}(s)) = r_1(s,\pi^*(s))-r_1(s,\hat{\pi}(s)) \\ \leq & r_1(s,\pi^*(s))-\underline{Q}_1(s,\hat{\pi}(s)) \leq r_1(s,\pi^*(s))-\underline{Q}_1(s,\pi^*(s)) \\ = & r_1(s,\pi^*(s))-\max\{0,\hat{r}_1(s,\pi^*(s))-b^1_1(s,\pi^*(s))\} \\ \leq & 2b^1_1(s,\pi^*(s)),\end{aligned} \tag{9}$$

where π^* is an optimal policy. Thus under the conditions in Theorem 1, the suboptimality gap of $\hat{\pi}$ can be bounded by:

$$\begin{aligned}\mathrm{SubOpt}(\hat{\pi}) = & \mathbb{E}_{s\sim p_0}[V^*_1(s)-Q_1(s,\hat{\pi}(s))] \\ \leq & \mathbb{E}_{s\sim p_0}\left[2b^1_1(s,\pi^*(s))\right] \leq 2b.\end{aligned} \tag{10}$$

We can ensure the optimality of $\hat{\pi}$ by using either the action gap or policy gap:

- on one hand, by (9), if $2b_1(s,\pi^*(s)) \leq 2b < \Delta^{\mathcal{A}}_{\min}$ for all $s\in\mathcal{S}$, then for all $s\in\mathcal{S}$, $\hat{\pi}$ chooses an optimal action and thus achieves the optimal value;
- on the other hand, by (10), if $2b < \Delta^{\Pi}_{\min}$, then $\hat{\pi}$ achieve the optimal value.

However, the condition $2b < \Delta^{\Pi}_{\min}$ is more conservative than $2b < \Delta^{\mathcal{A}}_{\min}$ because $\Delta^{\Pi}_{\min}$ can be much smaller than $\Delta^{\mathcal{A}}_{\min}$. Similarly, in the more general MDP setting, $\Delta^{\mathcal{A}}_{\min}$ and $\Delta^{\Pi}_{\min}$ differ by at least a factor of state occupancy probability as shown in Proposition 1, thus Theorem 2 provides a more desirable optimality condition than Proposition 5.

7 Conclusion

We provided a new optimality condition for corruption-robust offline RL with heavy-tailed rewards. We show that if $\tilde{O}\left(H\sigma\epsilon^{\frac{\gamma}{1+\gamma}}+\epsilon H^2\right) < \Delta^{\mathcal{A}}_{\min}$, then a modified pessimistic value iteration algorithm can obtain a policy with the optimal value even under data corruption.

Future work should answer the question: what is the **sufficient and necessary** condition for learners to get a policy with optimal value? A less fundamental but equally interesting direction is to strengthen the sample complexity in this paper.

Acknowledgements

We would like to thank Yudong Chen for valuable discussions. Xie is partially supported by NSF grant 1955997. This project is supported in part by NSF grants 1545481, 1704117, 1836978, 2023239, 2041428, 2202457, ARO MURI W911NF2110317, and AF CoE FA9550-18-1-0166.

References

Arnold, B. C. 2014. Pareto distribution. *Wiley StatsRef: Statistics Reference Online*, 1–10.

Azar, M. G.; Osband, I.; and Munos, R. 2017. Minimax regret bounds for reinforcement learning. In *International Conference on Machine Learning*, 263–272. PMLR.

Behzadan, V.; and Munir, A. 2017. Vulnerability of deep reinforcement learning to policy induction attacks. In *Machine Learning and Data Mining in Pattern Recognition: 13th International Conference, MLDM 2017, New York, NY, USA, July 15-20, 2017, Proceedings 13*, 262–275. Springer.

Borak, S.; Härdle, W.; and Weron, R. 2005. Stable distributions. *Statistical tools for finance and insurance*, 1: 21–44.

Bubeck, S.; Cesa-Bianchi, N.; and Lugosi, G. 2013. Bandits with heavy tail. *IEEE Transactions on Information Theory*, 59(11): 7711–7717.

Chen, Y.; Zhang, X.; Zhang, K.; Wang, M.; and Zhu, X. 2022. Byzantine-Robust Online and Offline Distributed Reinforcement Learning. *arXiv preprint arXiv:2206.00165*.

Dann, C.; Lattimore, T.; and Brunskill, E. 2017. Unifying PAC and regret: Uniform PAC bounds for episodic reinforcement learning. *Advances in Neural Information Processing Systems*, 30.

Dann, C.; Marinov, T. V.; Mohri, M.; and Zimmert, J. 2021. Beyond value-function gaps: Improved instance-dependent regret bounds for episodic reinforcement learning. *Advances in Neural Information Processing Systems*, 34: 1–12.

Diakonikolas, I.; Kamath, G.; Kane, D.; Li, J.; Moitra, A.; and Stewart, A. 2019a. Robust estimators in high-dimensions without the computational intractability. *SIAM Journal on Computing*, 48(2): 742–864.

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

Diakonikolas, I.; Kong, W.; and Stewart, A. 2019. Efficient algorithms and lower bounds for robust linear regression. In *Proceedings of the Thirtieth Annual ACM-SIAM Symposium on Discrete Algorithms*, 2745–2754. SIAM.

Dubey, A.; et al. 2020. Cooperative multi-agent bandits with heavy tails. In *International conference on machine learning*, 2730–2739. PMLR.

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*, 1625–1634.

Fujimoto, S.; Meger, D.; and Precup, D. 2019. Off-policy deep reinforcement learning without exploration. In *International conference on machine learning*, 2052–2062. PMLR.

Hu, Y.; Kallus, N.; and Uehara, M. 2021. Fast rates for the regret of offline reinforcement learning. *arXiv preprint arXiv:2102.00479*.

Huang, S.; Papernot, N.; Goodfellow, I.; Duan, Y.; and Abbeel, P. 2017. Adversarial attacks on neural network policies. *arXiv preprint arXiv:1702.02284*.

Huber, P. J. 1992. Robust estimation of a location parameter. *Breakthroughs in statistics: Methodology and distribution*, 492–518.

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

Jonsson, A.; Kaufmann, E.; Ménard, P.; Darwiche Domingues, O.; Leurent, E.; and Valko, M. 2020. Planning in markov decision processes with gap-dependent sample complexity. *Advances in Neural Information Processing Systems*, 33: 1253–1263.

Kapoor, S.; Patel, K. K.; and Kar, P. 2019. Corruption-tolerant bandit learning. *Machine Learning*, 108(4): 687–715.

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)*, 665–674. IEEE.

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

Lattimore, T.; and Szepesvári, C. 2020. *Bandit algorithms*. Cambridge University Press.

Liebeherr, J.; Burchard, A.; and Ciucu, F. 2012. Delay bounds in communication networks with heavy-tailed and self-similar traffic. *IEEE Transactions on Information Theory*, 58(2): 1010–1024.

Lugosi, G.; and Mendelson, S. 2019. Mean estimation and regression under heavy-tailed distributions: A survey. *Foundations of Computational Mathematics*, 19(5): 1145–1190.

Lugosi, G.; and Mendelson, S. 2021. Robust multivariate mean estimation: the optimality of trimmed mean. *The Annals of Statistics*, 49(1): 393–410.

Lykouris, T.; Simchowitz, M.; Slivkins, A.; and Sun, W. 2021. Corruption-robust exploration in episodic reinforcement learning. In *Conference on Learning Theory*, 3242–3245. PMLR.

Ma, Y.; Zhang, X.; Sun, W.; and Zhu, J. 2019. Policy poisoning in batch reinforcement learning and control. *Advances in Neural Information Processing Systems*, 32.

Medina, A. M.; and Yang, S. 2016. No-regret algorithms for heavy-tailed linear bandits. In *International Conference on Machine Learning*, 1642–1650. PMLR.

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

Niss, L.; and Tewari, A. 2020. What You See May Not Be What You Get: UCB Bandit Algorithms Robust to ε-Contamination. In *Conference on Uncertainty in Artificial Intelligence*, 450–459. PMLR.

Panaganti, K.; Xu, Z.; Kalathil, D.; and Ghavamzadeh, M. 2022. Robust reinforcement learning using offline data. *arXiv preprint arXiv:2208.05129*.

Prasad, A.; Suggala, A. S.; Balakrishnan, S.; and Ravikumar, P. 2018. Robust estimation via robust gradient estimation. *arXiv preprint arXiv:1802.06485*.

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.

Shao, H.; Yu, X.; King, I.; and Lyu, M. R. 2018. Almost optimal algorithms for linear stochastic bandits with heavy-tailed payoffs. *Advances in Neural Information Processing Systems*, 31.

Shi, L.; and Chi, Y. 2022. Distributionally robust model-based offline reinforcement learning with near-optimal sample complexity. *arXiv preprint arXiv:2208.05767*.

Simchowitz, M.; and Jamieson, K. G. 2019. Non-asymptotic gap-dependent regret bounds for tabular mdps. *Advances in Neural Information Processing Systems*, 32.

Sun, J.; Zhang, T.; Xie, X.; Ma, L.; Zheng, Y.; Chen, K.; and Liu, Y. 2020. Stealthy and efficient adversarial attacks against deep reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 5883–5891.

Sutton, R. S.; and Barto, A. G. 2018. *Reinforcement learning: An introduction*. MIT press.

Tukey, J. W. 1960. A survey of sampling from contaminated distributions. *Contributions to probability and statistics*, 448–485.

Wagenmaker, A. J.; Simchowitz, M.; and Jamieson, K. 2022. Beyond no regret: Instance-dependent pac reinforcement learning. In *Conference on Learning Theory*, 358–418. PMLR.

Wang, R.; Foster, D.; and Kakade, S. M. 2020. What are the Statistical Limits of Offline RL with Linear Function Approximation? In *International Conference on Learning Representations*.

Wang, X.; Cui, Q.; and Du, S. S. 2022. On gap-dependent bounds for offline reinforcement learning. *Advances in Neural Information Processing Systems*, 35: 14865–14877.

Wei, C.-Y.; Dann, C.; and Zimmert, J. 2022. A model selection approach for corruption robust reinforcement learning. In *International Conference on Algorithmic Learning Theory*, 1043–1096. PMLR.

Xie, T.; Jiang, N.; Wang, H.; Xiong, C.; and Bai, Y. 2021. Policy finetuning: Bridging sample-efficient offline and online reinforcement learning. *Advances in neural information processing systems*, 34: 27395–27407.

Xu, H.; Ma, T.; and Du, S. 2021. Fine-grained gap-dependent bounds for tabular mdps via adaptive multi-step bootstrap. In *Conference on Learning Theory*, 4438–4472. PMLR.

Yu, X.; Shao, H.; Lyu, M. R.; and King, I. 2018. Pure Exploration of Multi-Armed Bandits with Heavy-Tailed Payoffs. In *UAI*, 937–946.

Zhang, X.; Chen, Y.; Zhu, X.; and Sun, W. 2021. Robust policy gradient against strong data corruption. In *International Conference on Machine Learning*, 12391–12401. PMLR.

Zhang, X.; Chen, Y.; Zhu, X.; and Sun, W. 2022. Corruption-robust offline reinforcement learning. In *International Conference on Artificial Intelligence and Statistics*, 5757–5773. PMLR.

Zhang, X.; Ma, Y.; Singla, A.; and Zhu, X. 2020. Adaptive reward-poisoning attacks against reinforcement learning. In *International Conference on Machine Learning*, 11225–11234. PMLR.