Robust Policy Gradient against Strong Data Corruption

Xuezhou Zhang [1] Yiding Chen [1] Xiaojin Zhu [1] Wen Sun [2]

Abstract

We study the problem of robust reinforcement learning under adversarial corruption on both rewards and transitions. Our attack model assumes an *adaptive* adversary who can arbitrarily corrupt the reward and transition at every step within an episode, for at most ε-fraction of the learning episodes. Our attack model is strictly stronger than those considered in prior works. Our first result shows that no algorithm can find a better than $O(\varepsilon)$-optimal policy under our attack model. Next, we show that surprisingly the natural policy gradient (NPG) method retains a natural robustness property if the reward corruption is bounded, and can find an $O(\sqrt{\varepsilon})$-optimal policy. Consequently, we develop a Filtered Policy Gradient (FPG) algorithm that can tolerate even unbounded reward corruption and can find an $O(\varepsilon^{1/4})$-optimal policy. We emphasize that FPG is the first that can achieve a meaningful learning guarantee when a constant fraction of episodes are corrupted. Complimentary to the theoretical results, we show that a neural implementation of FPG achieves strong robust learning performance on the MuJoCo continuous control benchmarks.

1. Introduction

Policy gradient methods are a popular class of Reinforcement Learning (RL) methods among practitioners, as they are amenable to parametric policy classes (Schulman et al., 2015b; 2017), resilient to modeling assumption mismatches (Agarwal et al., 2019; 2020a), and they directly optimizing the cost function of interest. However, one current drawback of these methods and most existing RL algorithms is the lack of robustness to data corruption, which severely limits their applications to high-stack decision-making domains with highly noisy data, such as autonomous driving, quantitative trading, or medical diagnosis.

In fact, data corruption can be a larger threat in the RL paradigm than in traditional supervised learning, because supervised learning is often applied in a controlled environment where data are collected and cleaned by highly-skilled data scientists and domain experts, whereas RL agents are developed to learn in the wild using raw feedbacks from the environment. While the increasing autonomy and less supervision mark a step closer to the goal of general artificial intelligence, they also make the learning system more susceptible to data corruption: autonomous vehicles can misread traffic signs when the signs are contaminated by adversarial stickers (Eykholt et al., 2018); chatbot can be mistaught by a small group of tweeter users to make misogynistic and racist remarks (Neff & Nagy, 2016); recommendation systems can be fooled by a small number of fake clicks/reviews/comments to rank products higher than they should be. Despite the many vulnerabilities, *robustness* against data corruption in RL has not been extensively studied only until recently.

The existing works on *robust* RL are mostly theoretical and can be viewed as a successor of the adversarial bandit literature. However, several drawbacks of this line of approach make them insufficient to modern real-world threats faced by RL agents. We elaborate them below:

1. **Reward vs. transition contamination**: The majority of prior works on adversarial RL focus on reward contamination (Even-Dar et al., 2009; Neu et al., 2010; 2012; Zimin & Neu, 2013; Rosenberg & Mansour, 2019; Jin et al., 2020a), while in reality the adversary often has stronger control during the adversarial interactions. For example, when a chatbot interacts with an adversarial user, the user has full control over both the rewards and transitions during that conversation episode.
2. **Density of contamination**: The existing works that do handle adversarial/time-varying transitions can only tolerate *sublinear* number of interactions being corrupted (Lykouris et al., 2019; Cheung et al., 2019; Ornik & Topcu, 2019; Ortner et al., 2019). These methods would fail when the adversary's attack budget also grows linearly with time, which is often the case in practice.
3. **Practicability**: The majority of these work focuses on the setting of tabular MDPs and cannot be applied to

[1] University of Wisconsin–Madison [2] Cornell University. Correspondence to: Xuezhou Zhang <xzhang784@wisc.edu>, Wen Sun <ws455@cornell.com>.

real-world RL problems that have large state and action spaces and require function approximations.

In this work, we address the above shortcomings by developing a variant of natural policy gradient (NPG) methods that, under the linear value function assumption, are provably robust against strongly adaptive adversaries, who can **arbitrarily contaminate** both rewards and transitions in ε fraction of all learning episodes. Our algorithm does not need to know ε, and is adaptive to the contamination level. Specifically, it guarantees to find an $\tilde{O}(\varepsilon^{1/4})$-optimal policy in a polynomial number of steps. Complementarily, we also present a corresponding lower-bound, showing that no algorithm can consistently find a better than $\Omega(\varepsilon)$ optimal policy, even with infinite data. In addition to the theoretical results, we also develop a neural network implementation of our algorithm which is shown to achieve strong robustness performance on the MuJoCo continuous control benchmarks (Todorov et al., 2012), proving that our algorithm can be applied to real-world, high-dimensional RL problems.

2. Related Work

RL in standard MDPs. Learning MDPs with stochastic rewards and transitions is relatively well-studied for the tabular case (that is, a finite number of states and actions). For example, in the episodic setting, the UCRL2 algorithm (Auer et al., 2009) achieves $O(\sqrt{H^4S^2AT})$ regret, where H is the episode length, S is the state space size, A is the action space size, and T is the total number of steps. Later the UCBVI algorithm (Azar et al., 2017; Dann et al., 2017) achieves the optimal $O(\sqrt{H^2SAT})$ regret matching the lower-bound (Osband & Van Roy, 2016; Dann & Brunskill, 2015). Recent work extends the analysis to various linear setting (Jin et al., 2020b; Yang & Wang, 2019b;a; Zanette et al., 2020; Ayoub et al., 2020; Zhou et al., 2020; Cai et al., 2019; Du et al., 2019; Kakade et al., 2020) with known linear feature. For unknown feature, (Agarwal et al., 2020b) proposes a sample efficient algorithm that explicitly learns feature representation under the assumption that the transition matrix is low rank. Beyond the linear settings, there are works assuming the function class has low Eluder dimension which so far is known to be small only for linear functions and generalized linear models (Osband & Van Roy, 2014). For more general function approximation, (Jiang et al., 2017; Sun et al., 2019) showed that polynomial sample complexity is achievable as long as the MDP and the given function class together induce low Bellman rank and Witness rank, which include almost all prior models such as tabular MDP, linear MDPs (Yang & Wang, 2019b; Jin et al., 2020b), Kernelized nonlinear regulators (Kakade et al., 2020), low rank MDP (Agarwal et al., 2020b), and Bellman completion under linear functions (Zanette et al., 2020).

Policy Gradient and Policy Optimization Policy Gradient (Williams, 1992; Sutton et al., 1999) and Policy optimization methods are widely used in practice (Kakade & Langford, 2002; Schulman et al., 2015b; 2017) and have demonstrated amazing performance on challenging applications (Berner et al., 2019; Akkaya et al., 2019). Unlike model-based approach or Bellman-backup based approaches, PG methods directly optimize the objective function and are often more robust to model-misspecification (Agarwal et al., 2020a). In addition to being robust to model-misspecification, we show in this work that vanilla NPG is also robust to constant fraction and bounded adversarial corruption on both rewards and transitions.

RL with adversarial rewards. Almost all prior works on adversarial RL study the setting where the reward functions can be adversarial but the transitions are still stochastic and remain unchanged throughout the learning process. Specifically, at the beginning of each episode, the adversary must decide on a reward function for this episode, and can not change it for the rest of the episode. Also, the majority of these works focus on tabular MDPs. Early works on adversarial MDPs assume a known transition function and full-information feedback. For example, (Even-Dar et al., 2009) proposes the algorithm MDP-E and proves a regret bound of $\tilde{O}(\tau\sqrt{T\log A})$ in the non-episodic setting, where τ is the mixing time of the MDP; Later, (Zimin & Neu, 2013) consider the episodic setting and propose the O-REPS algorithm which applies Online Mirror Descent over the space of occupancy measures, a key component adopted by (Rosenberg & Mansour, 2019) and (Jin et al., 2020a). O-REPS achieves the optimal regret $\tilde{O}(\sqrt{H^2T\log(SA)})$ in this setting. Several works consider the harder bandit feedback model while still assuming known transitions. The work (Neu et al., 2010) achieves regret $\tilde{O}(\sqrt{H^3AT}/\alpha)$ assuming that all states are reachable with some probability α under all policies. Later, (Neu et al., 2010) eliminates the dependence on α but only achieves $O(T^{2/3})$ regret. The O-REPS algorithm of (Zimin & Neu, 2013) again achieves the optimal regret $\tilde{O}(\sqrt{H^3SAT})$. To deal with unknown transitions, (Neu et al., 2012) proposes the Follow the Perturbed Optimistic Policy algorithm and achieves $\tilde{O}(\sqrt{H^2S^2A^2T})$ regret given full-information feedback. Combining the idea of confidence sets and Online Mirror Descent, the UC-O-REPS algorithm of (Rosenberg & Mansour, 2019) improves the regret to $\tilde{O}(\sqrt{H^2S^2AT})$. A few recent works start to consider the hardest setting assuming unknown transition as well as bandit feedback. (Rosenberg & Mansour, 2019) achieves $O(T^{3/4})$ regret, which is improved by (Jin et al., 2020a) to $\tilde{O}(\sqrt{H^2S^2AT})$, matching the regret of UC-O-REPS in the full information setting. Also, note that the lower bound of $\Omega(\sqrt{H^2SAT})$ (Jin et al., 2018) still applies. In summary, it is found that on tabular MDPs with oblivious reward contamination, an $O(\sqrt{T})$ regret can still be achieved. Recent improvements include best-of-both-worlds algorithms (Jin & Luo, 2020), data-dependent bound (Lee et al., 2020) and extension to linear function approximation (Neu & Olkhovskaya, 2020).

RL with adversarial transitions and rewards. Very few prior works study the problem of both adversarial transitions and adversarial rewards, in fact, only one that we are aware of (Lykouris et al., 2019). They study a setting where only a constant C number of episodes can be corrupted by the adversary, and most of their technical effort dedicate to designing an algorithm that is agnostic to C, i.e. the algorithm doesn't need to know the contamination level ahead of time. As a result, their algorithm takes a multi-layer structure and cannot be easily implemented in practice. Their algorithm achieves a regret of $O(C\sqrt{T})$ for tabular MDPs and $O(C^2\sqrt{T})$ for linear MDPs, which unfortunately becomes vacuous when $C \geq \Omega(\sqrt{T})$ and $C \geq \Omega(T^{1/4})$, respectively. Note that the contamination ratio C/T approaches zero when T increases, and hence their algorithm cannot handle constant fraction contamination. Notably, in all of the above works, the adversary can *partially adapt* to the learner's behavior, in the sense that the adversary can pick an adversary MDP $\mathcal{M}_k$ or reward function r_k at the start of episode k based on the history of interactions so far. However, it can no longer adapt its strategy after the episode starts, and therefore, the learner can still use a randomization strategy to trick the adversary.

A separate line of work studies the *online MDP* setting, where the MDP is not adversarial but *slowly* change over time, and the amount of change is bounded under a total-variation metric (Cheung et al., 2019; Ornik & Topcu, 2019; Ortner et al., 2019; Domingues et al., 2020). Due to the slow-changing nature of the environment, algorithms in these works typically uses a sliding window approach where the algorithm keeps throwing away old data and only learns a policy from recent data, assuming that most of them come from the MDP that the agent is currently experiencing. These methods typically achieve a regret in the form of $O(\Delta^c K^{1-c})$, where Δ is the total variation bound. It is worth noting that all of these regrets become vacuous when the amount of variation is linear in time, i.e. $\Delta \geq \Omega(T)$. Separately, it is shown that when both the transitions and the rewards are adversarial in every episode, the problem is at least as hard as stochastic parity problem, for which no computationally efficient algorithm exists (Yadkori et al., 2013).

Learning robust controller. A different type of robustness has also been considered in RL (Pinto et al., 2017; Derman et al., 2020) and robust control (Zhou & Doyle, 1998; Petersen et al., 2012), where the goal is to learn a control policy that is robust to potential misalignment between the training and deployment environment. Such approaches are often conservative, i.e. the learned polices are sub-optimal even if there is no corruption. In comparison, our approach can learn as effectively as standard RL algorithms without corruption. Interestingly, parallel to our work, a line of concurrent work in the robust control literature (Zhang et al., 2020a;b; 2021) has also found that policy optimization method enjoys some implicit regularization/robustness property that can automatically converge to robust control policies. An interesting future direction could be to understand the connection between these two kind of robustness.

Robust statistics. One of the most important discoveries in modern robust statistics is that there exists computationally efficient and robust estimator that can learn near-optimally even under the strongest adaptive adversary. For example, in the classic problem of Gaussian mean estimation, the recent works (Diakonikolas et al., 2016; Lai et al., 2016) present the first computational and sample-efficient algorithms. The algorithm in (Diakonikolas et al., 2016) can generate a robust mean estimate $\hat{\mu}$, such that $\|\hat{\mu}-\mu\|_2 \leq O(\varepsilon\sqrt{\log{(1/\varepsilon)}})$ under ε corruption. Crucially, the error bound does not scale with the dimension d of the problem, suggesting that the estimator remains robust even in high dimensional problems. Similar results have since been developed for robust mean estimation under weaker assumptions (Diakonikolas et al., 2017), and for supervised learning and unsupervised learning tasks (Charikar et al., 2017; Diakonikolas et al., 2019). We refer readers to (Diakonikolas & Kane, 2019) for a more thorough survey of recent advances in high-dimensional robust statistics.

3. Problem Definitions

A Markov Decision Process (MDP) $\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, r, \gamma, \mu_0)$ is specified by a state space $\mathcal{S}$, an action space $\mathcal{A}$, a transition model $P : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ (where $\Delta(\mathcal{S})$ denotes a distribution over $\mathcal{S}$), a (stochastic and possibly unbounded) reward function $r : \mathcal{S} \times \mathcal{A} \to \Delta(\mathbb{R})$, a discounting factor $\gamma \in [0, 1)$, and an initial state distribution $\mu_0 \in \Delta(\mathcal{S})$, i.e. $s_0 \sim \mu_0$. In this paper, we assume that $\mathcal{A}$ is a small and finite set, and denote $A = |\mathcal{A}|$. A policy $\pi : \mathcal{S} \to \Delta(\mathcal{A})$ specifies a decision-making strategy in which the agent chooses actions based on the current state, i.e., $a \sim \pi(\cdot|s)$.

The value function $V^\pi : \mathcal{S} \to \mathbb{R}$ is defined as the expected discounted sum of future rewards, starting at state s and executing π, i.e. $V^\pi(s) := \mathbb{E}\left[\sum_{t=0}^\infty \gamma^t r(s_t, a_t)|\pi, s_0 = s\right]$, where the expectation is taken with respect to the randomness of the policy and environment $\mathcal{M}$. Similarly, the *state-action* value function $Q^\pi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is defined as $Q^\pi(s, a) := \mathbb{E}\left[\sum_{t=0}^\infty \gamma^t r(s_t, a_t)|\pi, s_0 = s, a_0 = a\right]$.

We define the discounted state-action distribution d_s^π of a policy π: $d^{\pi}_{s'}(s,a) := (1-\gamma)\sum_{t=0}^{\infty}\gamma^t \mathbb{P}^{\pi}(s_t = s, a_t = a|s_0 = s')$, where $\mathbb{P}^{\pi}(s_t = s, a_t = a|s_0 = s')$ is the probability that $s_t = s$ and $a_t = a$, after we execute π from $t = 0$ onwards starting at state s' in model $\mathcal{M}$. Similarly, we define $d^{\pi}_{s',a'}(s,a)$ as: $d^{\pi}_{s',a'}(s,a) := (1-\gamma)\sum_{t=0}^{\infty}\gamma^t \mathbb{P}^{\pi}(s_t = s, a_t = s|s_0 = s', a_0 = a')$. For any state-action distribution ν, we write $d^{\pi}_{\nu}(s,a) := \sum_{(s',a')\in\mathcal{S}\times\mathcal{A}}\nu(s',a')d^{\pi}_{s',a'}(s,a)$. For ease of presentation, we assume that the agent can reset to $s_0 \sim \mu_0$ at any point in the trajectory. We denote $d^{\pi}_{\nu}(s) = \sum_a d^{\pi}_{\nu}(s,a)$.

The goal of the agent is to find a policy π that maximizes the expected value from the starting state s_0, i.e. the optimization problem is: $\max_{\pi} V^{\pi}(\mu_0) := \mathbb{E}_{s\sim\mu_0}V^{\pi}(s)$, where the $\max$ is over some policy class.

For completeness, we specify a d^{π}_{ν}-sampler and an unbiased estimator of $Q^{\pi}(s,a)$ in Algorithm 1, which are standard in discounted MDPs (Agarwal et al., 2019; 2020a). The d^{π}_{ν} sampler samples (s,a) i.i.d from d^{π}_{ν}, and the Q^{π} sampler returns an unbiased estimate of $Q^{\pi}(s,a)$ for a given pair (s,a) by a single roll-out from (s,a). Later, when we define the contamination model and the sample complexity of learning, we treat each call of d^{π}_{ν}-sampler (optionally followed by a $Q^{\pi}(s,a)$-estimator) as a *single episode*, as in practice both of these procedures can be achieved in a single roll-out from μ_0.

Assumption 3.1 (Linear Q function). *For the theoretical analysis, we focus on the setting of linear value function approximation. In particular, we assume that there exists a feature map $\phi : \mathcal{S}\times\mathcal{A}\to\mathbb{R}^d$, such that for any $(s,a)\in\mathcal{S}\times\mathcal{A}$ and any policy $\pi : \mathcal{S}\to\Delta_{\mathcal{A}}$, we have*

$$Q^{\pi}(s,a) = \phi(s,a)^{\top}w^{\pi}, \text{for some } \|w^{\pi}\| \le W \tag{1}$$

We also assume that the feature is bounded, i.e. $\max_{s,a}\|\phi(s,a)\|_2 \le 1$, and the reward function has bounded first and second moments, i.e. $\mathbb{E}[r(s,a)]\in[0,1]$ and Var$(r(s,a)) \le \sigma^2$ for all (s,a).

Remark 3.1. Assumption 3.1 is satisfied, for example, in tabular MDPs and linear MDPs of (Jin et al., 2020b) or (Yang & Wang, 2019a). Unlike most theoretical RL literature, we allow the reward to be stochastic and unbounded. Such a setup aligns better with applications with a low signal-to-noise ratio and motivates the requirement for non-trivial robust learning techniques.

Notation. When clear from context, we write $d^{\pi}(s,a)$ and $d^{\pi}(s)$ to denote $d^{\pi}_{\mu_0}(s,a)$ and $d^{\pi}_{\mu_0}(s)$ respectively. For iterative algorithms which obtain policies at each episode, we let V^i,Q^i and A^i denote the corresponding quantities associated with episode i. For a vector v, we denote $\|v\|_2 = \sqrt{\sum_i v_i^2}$, $\|v\|_1 = \sum_i|v_i|$, and $\|v\|_{\infty} = \max_i|v_i|$. We use Uniform$(\mathcal{A})$ (in short Unif$_{\mathcal{A}}$) to represent a uniform distribution over the set $\mathcal{A}$.

3.1. The Contamination Model

In this paper, we study the robustness of policy gradient methods under the *ε-contamination model*, a widely studied adversarial model in the robust statistics literature, e.g. see (Diakonikolas et al., 2016). In the classic robust mean estimation problem, given a dataset D and a learning algorithm f, the ε-contamination model assumes that the adversary has full knowledge of the dataset D and the learning algorithm f, and can arbitrarily change ε-fraction of the data in the dataset and then send the contaminated data to the learner. The goal of the learner is to identify an $O(\text{poly}(\varepsilon))$-optimal estimator of the mean despite the ε-contamination.

Unfortunately, the original ε-contamination model is defined for the offline learning setting and does not directly generalize to the online setting, because it doesn't specify the availability of knowledge and the order of actions between the adversary and the learner in the time dimension. In this paper, we define the ε-contamination model for online learning as follows:

Definition 3.1 (ε-contamination model for Reinforcement Learning). Given ε and the clean MDP $\mathcal{M}$, an ε-contamination adversary operates as follows:

1. The adversary has full knowledge of the MDP $\mathcal{M}$ and the learning algorithm, and observes all the historical interactions.I
2. At any time step t, the adversary observes the current state-action pair (s_t, a_t), as well as the reward and next state returned by the environment, (r_t, s_{t+1}). He then can decide whether to replace (r_t, s_{t+1}) with an arbitrary reward and next state $(r^{\dagger}_t, s^{\dagger}_{t+1})\in\mathbb{R}\times\mathcal{S}$.
3. The only constraint on the adversary is that if the learning process terminates after K episodes, he can contaminate in at most εK episodes.

Compared to the standard adversarial models studied in online learning (Shalev-Shwartz et al., 2011), adversarial bandits (Bubeck & Cesa-Bianchi, 2012; Lykouris et al., 2018; Gupta et al., 2019) and adversarial RL (Lykouris et al., 2019; Jin et al., 2020a), the ε-contamination model in Definition 3.1 is stronger in several ways: (1) The adversary can adaptively attack after observing the action of the learner as well as the feedback from the clean environments; (2) the adversary can perturb the data arbitrarily (any real-valued reward and any next state from the state space) rather than sampling it from a pre-specified bounded adversarial reward function or adversarial MDP.

Given the contamination model, our first result is a lower-bound, showing that under the ε-contamination model, one can only hope to find an $O(\varepsilon)$-optimal policy. Exact optimal policy identification is not possible even with infinite data.

Theorem 3.1 (lower bound). *For any algorithm, there exists* *an MDP such that the algorithm fails to find an $\left(\frac{\varepsilon}{2(1-\gamma)}\right)$-optimal policy under the ε-contamination model with a probability of at least $1/4$.*

The high-level idea is that we can construct two MDPs, M and M', with the following properties: 1. No policy can be $O(\varepsilon/(1-\gamma))$ optimal on both MDP simultaneously. 2. An ε-contamination adversary can with large probability mimic one MDP via contamination in the other, regardless of the learner's behavior. Therefore, under contamination, the learner will not be able to distinguish M and M' and must suffer $\Omega(\varepsilon/(1-\gamma))$ gap on at least one of them.

3.2. Background on NPG

Given a differentiable parameterized policy $\pi_\theta : \mathcal{S} \to \Delta(\mathcal{A})$, NPG can be written in the following actor-critc style update form. With the dataset $\{s_i, a_i, \widehat{Q}^{\pi_\theta}(s_i, a_i)\}_{i=1}^N$ where $s_i, a_i \sim d_\nu^{\pi_\theta}$, and $\widehat{Q}^{\pi_\theta}(s_i, a_i)$ is unbiased estimate of $Q^{\pi_\theta}(s, a)$ (e.g., via Q^π-estimator), we have

$$\widehat{w} \in \underset{w:\|w\|_2 \le W}{\arg\min} \sum_{i=1}^N \left(w^\top \nabla \log \pi_\theta(a_i|s_i) - \widehat{Q}^{\pi_\theta}(s_i, a_i) \right)^2$$
$$\theta' = \theta + \eta \widehat{w}. \tag{2}$$

In theoretical part of this work, we focus on softmax linear policy, i.e., $\pi_\theta(a|s) \propto \exp(\theta^\top \phi(s, a))$. In this case, note that $\nabla \log \pi_\theta(a|s) = \phi(s, a)$, and it is not hard to verify that the policy update procedure is equivalent to:

$$\pi_{\theta'}(a|s) \propto \pi_\theta(a|s) \exp\left(\eta \widehat{w}^\top \phi(s, a)\right), \quad \forall s, a,$$

which is equivalent to running Mirror Descent on each state with a reward vector $\widehat{w}^\top \phi(s, \cdot) \in \mathbb{R}^{|\mathcal{A}|}$. We refer readers to (Agarwal et al., 2019) for more detailed explanation of NPG and the equivalence between the form in Eq. (2) and the classic form that uses Fisher information matrix. Similar to (Agarwal et al., 2019), we make the following assumption of having access to an exploratory reset distribution, under which it has been shown that NPG can converge to the optimal policy without contamination.

Assumption 3.2 (Relative condition number)**.** *With respect to any state-action distribution υ, define:*

$$\Sigma_\upsilon = \mathbb{E}_{s,a\sim\upsilon}\left[\phi_{s,a}\phi_{s,a}^\top\right],$$

and define

$$\sup_{w\in\mathbb{R}^d} \frac{w^\top \Sigma_{d^\star} w}{w^\top \Sigma_\nu w} = \kappa, \text{ where } d^*(s, a) = d_{\mu_0}^{\pi^*}(s) \circ \mathit{Unif}_{\mathcal{A}}(a)$$

We assume κ is finite and small w.r.t. a reset distribution ν available to the learner at training time.

4. The Natural Robustness of NPG Against Bounded corruption

Our first result shows that, surprisingly, NPG can already be robust against ε-contamination, if the adversary can only generate small and bounded rewards. In particular, we assume that the adversarial rewards is bounded in $[0, 1]$ (the feature $\phi(s, a)$ is already bounded).

Theorem 4.1 (Natural robustness of NPG)**.** *Under assumptions 3.1 and 3.2, given a desired optimality gap α, there exists a set of hyperparameters agnostic to the contamination level ε, such that Algorithm 2 guarantees with a poly$(1/\alpha, 1/(1-\gamma), |\mathcal{A}|, W, \sigma, \kappa)$ sample complexity that under ε-contamination with adversarial rewards bounded in $[0, 1]$, we have*

$$\mathbb{E}\left[V^*(\mu_0) - V^{\hat{\pi}}(\mu_0)\right] \le \tilde{O}\left(\max\left[\alpha, W\sqrt{\frac{|\mathcal{A}|\kappa\varepsilon}{(1-\gamma)^3}}\,\right]\right)$$

where $\hat{\pi}$ is the uniform mixture of $\pi^{(1)}$ through $\pi^{(T)}$.

A few remarks are in order.

Remark 4.1 (Agnostic to the contamination level ε)**.** It is worth emphasizing that to achieve the above bound, the hyperparameters of NPG are agnostic to the value of ε, and so the algorithm can be applied in the more realistic setting where the agent does not have knowledge of the contamination level ε, similar to what's achieved in (Lykouris et al., 2019) with a complicated nested structure. The same property is also achieved by the FPG algorithm in the next section.

Remark 4.2 (Dimension-independent robustness guarantee)**.** Theorem 4.1 guarantees that NPG can find an $O(\varepsilon^{1/2})$-optimal policy after polynomial number of episodes, provided that $|\mathcal{A}|$ and κ are small. Conceptually, the relative condition number κ indicates how well-aligned the initial state distribution is to the occupancy distribution of the optimal policy. A good initial distribution can have a κ as small as 1, and so κ is independent of d. Interested readers can refer to (Agarwal et al., 2019) (Remark 6.3) for additional discussion on the relative condition number. Here, importantly, the optimality gap does not directly scale with d, and so the guarantee will not blow up on high-dimensional problems. This is an important attribute of robust learning algorithms heavily emphasized in the traditional robust statistics literature.

The proof of Theorem 4.1 relies on the following NPG regret lemma, first developed by (Even-Dar et al., 2009) for the MDP-Expert algorithm and later extend to NPG by (Agarwal et al., 2019; 2020a):

Lemma 4.1 (NPG Regret Lemma)**.** *Suppose Assumption 3.1 and 3.2 hold and Algorithm 2 starts with $\theta^{(0)} = 0$,*

$\eta = \sqrt{2\log|\mathcal{A}|/(W^2T)}$. *Suppose in addition that the (random) sequence of iterates satisfies the assumption that*

$$\mathbb{E}\left[\mathbb{E}_{s,a\sim d^{(t)}}\left[\left(Q^{\pi^{(t)}}(s,a)-\phi(s,a)^\top w^{(t)}\right)^2\right]\right] \le \varepsilon_{stat}^{(t)}.$$

Then, we have that

$$\mathbb{E}\left[\sum_{t=1}^{T}\{V^*(\mu_0)-V^{(t)}(\mu_0)\}\right] \le \frac{W}{1-\gamma}\sqrt{2\log|\mathcal{A}|T}+\sum_{t=1}^{T}\sqrt{\frac{4|\mathcal{A}|\kappa\varepsilon_{stat}^{(t)}}{(1-\gamma)^3}}. \tag{3}$$

Intuitively, Lemma 4.1 decompose the regret of NPG into two terms. The first term corresponds to the regret of standard mirror descent procedure, which scales with $\sqrt{T}$. The second term corresponds to the estimation error on the Q value, which acts as the reward signal for mirror descent. When not under attack, estimation error $\varepsilon_{stat}^{(t)}$ goes to zero as the number of samples M gets larger, which in turn implies the global convergence of NPG. However, when under bounded attack, the generalization error $\varepsilon_{stat}^{(t)}$ will not go to zero even with infinite data. Nevertheless, we can show that it is bounded by $O(\varepsilon^{(t)})$ when the sample size M is large enough, where $\varepsilon^{(t)}$ denotes the fraction of episodes being corrupted in iteration t. Note that by definition, we have $\sum_t \varepsilon^{(t)} \le \varepsilon T$.

Lemma 4.2 (Robustness of linear regression under bounded contamination)**.** *Suppose the adversarial rewards are bounded in* $[0,1]$*, and in a particular iteration t, the adversary contaminates* $\varepsilon^{(t)}$ *fraction of the episodes, then given M episodes, it is guaranteed that with probability at least* $1-\delta$*,*

$$\mathbb{E}_{s,a\sim d^{(t)}}\left[\left(Q^{\pi^{(t)}}(s,a)-\phi(s,a)^\top w^{(t)}\right)^2\right] \le 4\left(W^2+WH\right)\left(\varepsilon^{(t)}+\sqrt{\frac{8}{M}\log\frac{4d}{\delta}}\right). \tag{4}$$

where $H = (\log\delta - \log M)/\log\gamma$ *is the effective horizon.*

This along with the NPG regret lemma guarantees that the expected regret of NPG is bounded by $O(\sqrt{T}+M^{-1/4}+\sqrt{\varepsilon}T)$ which in turn guarantees to identify an $O(\sqrt{\varepsilon})$-optimal policy.

In the special case of tabular MDPs, $\phi(s,a)$ will all be one-hot vectors and W will in general by on the order of $O(\sqrt{SA})$, which means that the bound given by Theorem 4.1 still scales with the size of the state space. In the following corollary, we show that this dependency can be removed through a tighter analysis.

Algorithm 1 d_ν^π sampler and Q^π estimator

```
function $d_\nu^\pi$-SAMPLER
    Input: A reset distribution $\nu \in \Delta(\mathcal{S}\times\mathcal{A})$.
    Sample $s_0, a_0 \sim \nu$.
    Execute $\pi$ from $s_0, a_0$; at any step $t$ with $(s_t, a_t)$, return $(s_t, a_t)$ with probability $1-\gamma$.
function $Q^\pi$-ESTIMATOR
    Input: current state-action $(s,a)$, a policy $\pi$.
    Execute $\pi$ from $(s_0, a_0) = (s,a)$; at step $t$ with $(s_t, a_t)$, terminate with probability $1-\gamma$.
    Return: $\widehat{Q}^\pi(s,a) = \sum_{i=0}^{t} r(s_i, a_i)$.
```

[In an adversarial episode, the adversary can hijack the d_ν^π sampler to return any (s,a) pair and the Q^π-estimator to return any $\widehat{Q}^\pi(s,a) \in \mathbb{R}$.]

Algorithm 2 Natural Policy Gradient (NPG)

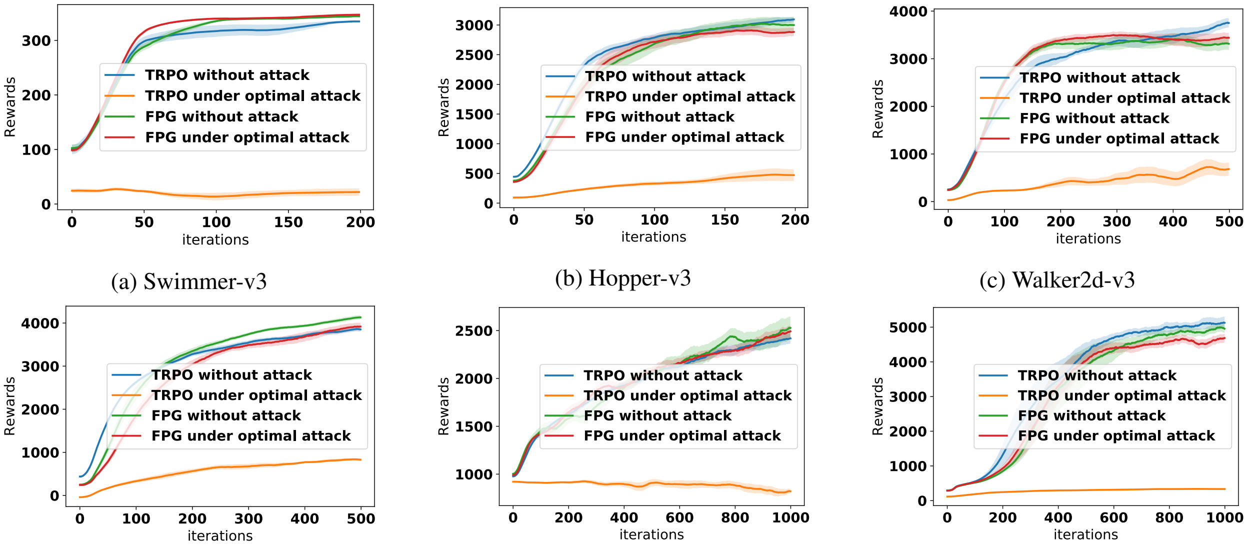
```
Require: Learning rate $\eta$; number of episodes per iteration $M$
Initialize $\theta^{(0)} = 0$.
for $t = 0, 1, \ldots, T-1$ do
    Call Algorithm 1 $M$ times with $\pi^{(t)}$ to obtain a dataset that consist of $s_i, a_i \sim d_\nu^{(t)}$ and $\widehat{Q}^{(t)}(s_i, a_i)$, $i \in [M]$.
    Solve the linear regression problem
        $w^{(t)} = \arg\min_{\|w\|_2 \le W} \sum_{i=1}^{M} \left(\widehat{Q}^{(t)}(s_i,a_i) - w^\top \nabla_\theta \phi(s_i,a_i)\right)^2$
    Update $\theta^{(t+1)} = \theta^{(t)} + \eta w^{(t)}$.
```

Corollary 4.1 (Dimension-free Robustness of NPG in tabular MDPs)**.** *Given a tabular MDP and assumption 3.2, given a desired optimality gap* α*, there exists a set of hyperparameters agnostic to the contamination level* ε*, such that Algorithm 2 guarantees with a* $poly(1/\alpha, 1/(1-\gamma), |\mathcal{A}|, W, \sigma, \kappa)$ *sample complexity that under* ε*-contamination with adversarial rewards bounded in* $[0,1]$*, we have*

$$\mathbb{E}\left[V^*(\mu_0)-V^{\hat{\pi}}(\mu_0)\right] \le \tilde{O}\left(\max\left[\alpha, \sqrt{\frac{|\mathcal{A}|\kappa\varepsilon}{(1-\gamma)^5}}\right]\right)$$

where $\hat{\pi}$ *is the uniform mixture of* $\pi^{(1)}$ *through* $\pi^{(T)}$*.*

In the more general case of linear MDP, W will not necessarily scale with d in an obvious way and thus we leave Theorem 4.1 untouched.



(a) Swimmer-v3

(b) Hopper-v3

(c) Walker2d-v3

(d) HalfCheetah-v3

(e) Ant-v3

(f) Humanoid-v3

Figure 1. Experiment Results on the 6 MuJoTo benchmarks.

Algorithm 3 Robust Linear Regression via SEVER

```
Input: Dataset $\{(x_i, y_i)\}_{i=1:M}$, a standard linear regression solver $\mathcal{L}$, and parameter $\sigma' \in \mathbb{R}_+$.
Initialize $S \leftarrow \{1, \dots, M\}$, $f_i(w) = \|y_i - w^\top x_i\|^2$.
repeat
    $w \leftarrow \mathcal{L}(\{(x_i, y_i)\}_{i \in S})$. ▷ Run learner on $S$.
    Let $\widehat{\nabla} = \frac{1}{|S|}\sum_{i \in S} \nabla f_i(w)$.
    Let $G = [\nabla f_i(w) - \widehat{\nabla}]_{i \in S}$ be the $|S| \times d$ matrix of centered gradients.
    Let $v$ be the top right singular vector of $G$.
    Compute the vector $\tau$ of outlier scores defined via $\tau_i = \left((\nabla f_i(w) - \widehat{\nabla}) \cdot v\right)^2$.
    $S' \leftarrow S$
    if $\frac{1}{|S|}\sum_{i \in S} \tau_i \le c_0 \cdot \sigma'^2$, for some constant $c_0 > 1$ then
        $S = S'$ ▷ We only filter out points if the variance is larger than an appropriately chosen threshold.
    else
        Draw $T$ from Uniform$[0, \max_i \tau_i]$.
        $S = \{i \in S : \tau_i < T\}$.
until $S = S'$.
Return $w$.
```

5. FPG: Robust NPG Against Unbounded Corruption

Our second result is the Filtered Policy Gradient (FPG) algorithm, a robust variant of the NPG algorithm (Kakade, 2001; Agarwal et al., 2019) that can be robust against arbitrary and *potentially unbounded* data corruption. Specifically, FPG replace the standard linear regression solver in NPG with a statistically robust alternative. In this work, we use the SEVER algorithm (Diakonikolas et al., 2019). In practice, one can substitute it with any computationally efficient robust linear regression solver. We show that FPG can find an $O(\varepsilon^{1/4})$-optimal policy under ε-contamination with a polynomial number of samples.

Theorem 5.1. *Under assumptions 3.1 and 3.2, given a desired optimality gap α, there exists a set of hyperparameters agnostic to the contamination level ε, such that Algorithm 2, using Algorithm 3 as the linear regression solver, guarantees with a poly$(1/\alpha, 1/(1-\gamma), |\mathcal{A}|, W, \sigma, \kappa)$ sample complexity that under ε-contamination, we have*

$$\mathbb{E}\left[V^*(\mu_0) - V^{\hat{\pi}}(\mu_0)\right] \le \tilde{O}\left(\max\left[\alpha, \sqrt{\frac{|\mathcal{A}|\kappa\,(W^2 + \sigma W)}{(1-\gamma)^4}}\varepsilon^{1/4}\right]\right). \tag{5}$$

where $\hat{\pi}$ is the uniform mixture of $\pi^{(1)}$ through $\pi^{(T)}$.

The proof of Theorem 5.1 relies on a similar result to Lemma 4.2, which shows that if we use Algorithm 3 as the linear regression subroutine, then $\varepsilon_{stat}^{(t)}$ can be bounded by $O(\sqrt{\varepsilon^{(t)}})$ when the sample size M is large enough, even under unbounded ε-contamination.

Lemma 5.1 (Robustness of SEVER under unbounded contamination)**.** *Suppose the adversarial rewards are unbounded, and in a particular iteration t, the adversarial contaminate $\varepsilon^{(t)}$ fraction of the episodes, then given M episodes, it is guaranteed that if $\varepsilon^{(t)} \le c$, for some absolute* *constant c, and any constant $\tau \in [0, 1]$, we have*

$$\mathbb{E}\left[\mathbb{E}_{s,a\sim d^{(t)}}\left[\left(Q^{\pi^{(t)}}(s,a) - \phi(s,a)^\top w^{(t)}\right)^2\right]\right] \tag{6}$$
$$\leq O\left(\left(W^2 + \frac{\sigma W}{1-\gamma}\right)\left(\sqrt{\varepsilon^{(t)}} + f(d,\tau)M^{-\frac{1}{2}} + \tau\right)\right).$$

where $f(d, \tau) = \sqrt{d \log d} + \sqrt{\log(1/\tau)}$.

In Lemma 5.1, c is the break point of SEVER and is an absolute constant that does not depend on the data, and $(1-\tau)$ is the probability that the clean data satisfies a certain stability condition which suffices for robust learning.

6. Robust NPG with Exploration via Policy Cover

The Policy Cover-Policy Gradient (PC-PG) algorithm, defined in Algorithm 4, is an exploratory policy gradient methods recently developed by (Agarwal et al., 2020a). Intuitively, PC-PG is a spiritually inheritor of the RMax algorithm (Brafman & Tennenholtz, 2002), and encourages exploration by adding reward bonuses in directions of the feature space that past polices (stored in the policy cover) haven't visited sufficiently. Similar to the NPG algorithm, we show that PC-PG enjoys a (weaker) natural robustness against bounded data corruption. This gives us the following robustness guarantee:

Theorem 6.1 (Best hyperparameters, assuming known ε)**.** *There exists a set of hyperparameters, such that Algorithm 4 guarantees with probability at least $1 - \delta$*

$$\mathbb{E}\left[V^*(\mu_0) - V^{\hat{\pi}}(\mu_0)\right] \leq \tilde{O}\left(d^2\varepsilon^{1/7}\right) \tag{7}$$

with $\mathit{poly}\,(d, W, \sigma, \kappa, |\mathcal{A}|, 1/(1-\gamma), 1/\alpha)$ number of episodes.

Remark 6.1 (The scaling with dimension d)**.** Compared to the guarantee of vanilla NPG, PC-PG alleviate the requirement of a good initial distribution with small relative conditional number. However, this process introduce a dependency on d. In particular, the gap in Theorem 6.1 is on the order of $\tilde{O}\left(d^2\varepsilon^{1/7}\right)$. This implies that for any fixed ε, the bound becomes vacuous for high dimensional problems where $d \geq \Omega(\varepsilon^{-2/3})$. Intuitively, the dependency on d is introduced because PC-PG is trying to find a initial state-action distribution with good coverage, i.e. a distribution whose covariance matrix has a lower-bounded smallest eigenvalue. Under the assumption that $\|\phi(s,a)\|_2 \leq 1$, such a distribution will have a covariance matrix whose eigenvalues are all on the order of $O(1/d)$. and so the value of κ will be on the order of $O(d)$, which by Theorem 5.1 will similarly introduce a d dependency. We expect that for a robust RL algorithm to avoid the d dependency, it must gradually find a state-action distribution approaching d^*. How to design such an algorithm is left as an open problem.

Algorithm 4 Robust Policy Cover-Policy Gradient (PC-PG)

```
Input: iterations $N$, threshold $\beta$, regularizer $\lambda$
Initialize $\pi^0(a|s)$ to be uniform.
for episode $n = 0, \ldots N-1$ do
    Define the policy cover's state-action distribution $\rho^n_{\text{cov}}$ as
        $$\rho^n_{\text{cov}}(s,a) = \sum_{i=0}^{n} d^i(s,a)/(n+1)$$
    Sample $\{s_i, a_i\}_{i=1}^K \sim \rho^n_{\text{cov}}(s,a)$ and estimate the covariance of $\pi^n$ as
        $$\widehat{\Sigma}^n = (n+1)\left(\sum_{i=1}^{K}\phi(s_i,a_i)\phi(s_i,a_i)^\top/K\right) + \lambda I$$
    Set the exploration bonus $b^n$ to reward infrequently visited state-action under $\rho^n_{\text{cov}}$
        $$b^n(s,a) = \frac{\mathbf{1}\{(s,a) \,:\, \phi(s,a)^\top(\widehat{\Sigma}^n_{\text{cov}})^{-1}\phi(s,a) \geq \beta\}}{1-\gamma}.$$
    Update $\pi^{n+1}$ = Robust-NPG-Update$(\rho^n_{\text{cov}}, b^n)$ [Alg. 7 in the appendix, similar to Alg. 2].
return $\hat{\pi} :=$ Uniform$\{\pi^0, ..., \pi^{N-1}\}$.
```

7. Experiments

In the theoretical analysis, we rely on the assumption of linear Q function, finite action space and exploratory initial state distribution to prove the robustness guarantees for NPG and FPG. In this section, we present a practical implementation of FPG, based on the *Trusted Region Policy Optimization* (TRPO) algorithm (Schulman et al., 2015a), in which the conjugate gradient step (equivalent to the linear regression step in Alg. 2) is robustified with SEVER. The pseudo-code and implementation details are discussed in appendix G[1]. In this section, we demonstrate its empirical performance on the MuJoCo benchmarks (Todorov et al., 2012), a set of high-dimensional continuous control domains where both assumptions no longer holds, and show that FPG can still consistently performs near-optimally with and without attack.

Attack mechanism: While designing and calculating the *optimal* attack strategy against a deep RL algorithm is still a challenging problem and active area of research (Ma et al., 2019; Zhang et al., 2020c), here we describe the poisoning

[1] A Pytorch Implementation of FPG-TRPO can be found at https://github.com/zhangxz1123/FilteredPolicyGradient

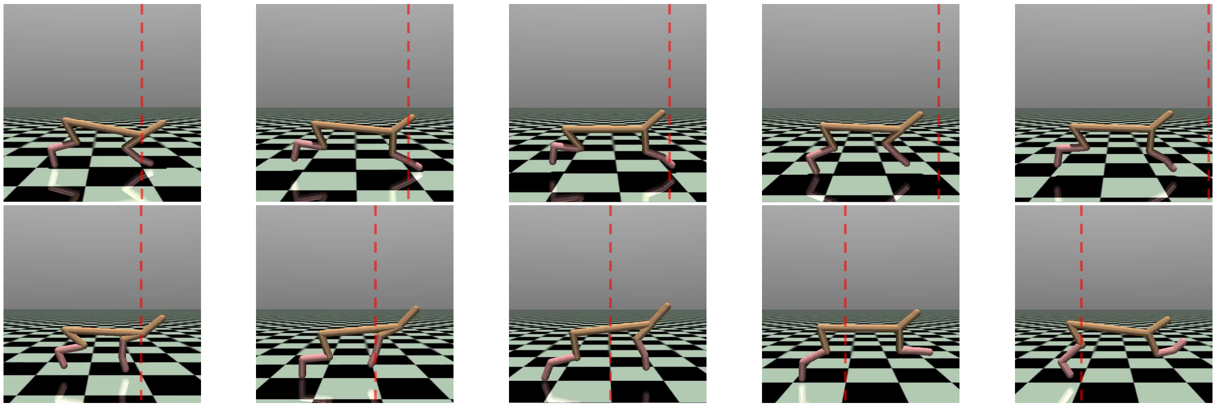
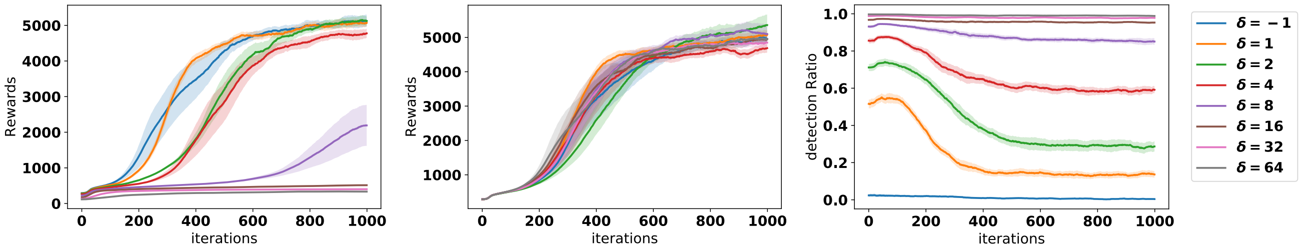


Figure 2. Consecutive Frames of Half-Cheetah trained with TRPO (top row) and FPG (bottom row) respectively under $\delta = 100$ attack. The dashed red line serves as a stationary reference object. TRPO was fooled to learn a "running backward" policy, contrasted with the normal "running forward" policy learned by FPG.



(a) TRPO Rewards

(b) FPG Rewards

(c) FPG Detection Ratio

Figure 3. Detailed Results on Humanoid-v3.

strategy used in our empirical evaluation, which, despite being simple, can fool non-robust RL algorithms with ease. Conceptually, policy gradient methods can be viewed as a stochastic gradient ascent method, where each iteration can be simplified as:

$$\theta^{(t+1)} = \theta^{(t)} + g^{(t)} \tag{8}$$

where $g^{(t)}$ is a gradient step that ideally points in the direction of fastest policy improvement. Assuming that $g^{(t)}$ is a good estimate of the gradient direction, then a simple attack strategy is to try to perturb $g^{(t)}$ to point in the $-g^{(t)}$ direction, in which case the policy, rather than improving, will deteriorate as learning proceed. A straightforward way to achieve this is to flip the rewards and multiply them by a big constant δ in the adversarial episodes. In the linear regression subproblem of Alg. 2, this would result in a set of (x, y) pairs whose y becomes $-\delta y$. This in expectation will make the best linear regressor w point to the opposite direction, which is precisely what we want.

This attack strategy is therefore parameterized by a single parameter δ, which guides the magnitude of the attack, and is **adaptively tuned** against each learning algorithm in the experiments: Throughout the experiment, we set the contamination level $\varepsilon = 0.01$, and tune δ among the values of $[1, 2, 4, 8, 16, 32, 64]$ to find the most effective magnitude against each learning algorithm. All experiments are repeated with 3 random seeds and the mean and standard deviations are plotted in the figures.

Results: The experiment results are shown in Figure 1. Consistent patterns can be observed across all environments: vanilla `TRPO` performs well without attack but fails completely under the adaptive attack (which choose $\delta = 64$ in all environments). `FPG`, on the other hand, matches the performance of vanilla `TRPO` with or without attack. Figure 2 showcase two half-cheetah control policies learned by `TRPO` and `FPG` under attack with $\delta = 100$. Interestingly, due to the large negative adversarial rewards, TRPO actually learns the "running backward" policy, showing that our attack strategy indeed achieves what it's designed for. In contrast, `FPG` is still able to learn the "running forward" policy despite the attack.

Figure 3 shows the detailed performances of `TRPO` and `FPG` across different δ's on the hardest *Humanoid* environment. One can observe that `TRPO` actually learns robustly under attacks of small magnitude ($\delta = 1, 2, 4$) and achieves similar performances to itself in clean environments, verifying our theoretical result in Theorem 4.1. In contrast, `FPG` remains robust across all values of δ's. Figure 3c shows the proportion of adversary data detected and removed by `FPG`'s filtering subroutine throughout the learning process. One can observe that as the attack norm δ increases, the filtering algorithm also does a better job detecting the adversarial data and thus protect the algorithm from getting inaccurate gradient estimates. Similar patterns can be observed in all the other environments, and we defer the additional figures to the appendix.

8. Discussions

To summarize, in this work we present a robust policy gradient algorithm `FPG`, and show theoretically and empirically that it can learn in the presence of strong data corruption. Despite our results, many open questions remain unclear and are interesting directions to pursue further:

1. `FPG` does not handle exploration and relies on an exploratory initial distribution. Can we design algorithms that achieve the same *dimension-free* robustness guarantee without such assumptions?
2. Our $O(\varepsilon^{1/4})$ upper-bound and $O(\varepsilon)$ lower-bound are not tight. Information theoretically, what is the best robustness guarantee one can achieve under ε-contamination?
3. The `SEVER` algorithm requires computing the top eigenvalue of an $n \times d$ matrix, which is memory and time consuming when using large neural networks (large d). More computationally efficient robust learning method will be extremely valuable to make FPG truly scale.
4. In the experiment, we focus on TRPO as the closest variant of NPG. Can other policy gradient algorithm, such as PPO and SAC, be robustified in similar fashions and achieve strong empirical performance?

We believe that answering these questions will be important steps towards more robust reinforcement learning.

9. Acknowledgements

We would like to thank Ankit Pensia and Ilias Diakonikolas for valuable discussions on SEVER and other robust statistics techniques. Xiaojin Zhu acknowledges NSF grants 1545481, 1704117, 1836978, 2041428, 2023239 and MADLab AF CoE FA9550-18-1-0166. Xuezhou Zhang is supported in part by NSF Award DMS-2023239.

References

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

Agarwal, A., Henaff, M., Kakade, S., and Sun, W. Pc-pg: Policy cover directed exploration for provable policy gradient learning. *arXiv preprint arXiv:2007.08459*, 2020a.

Agarwal, A., Kakade, S., Krishnamurthy, A., and Sun, W. Flambe: Structural complexity and representation learning of low rank mdps. *Advances in Neural Information Processing Systems*, 33, 2020b.

Akkaya, I., Andrychowicz, M., Chociej, M., Litwin, M., McGrew, B., Petron, A., Paino, A., Plappert, M., Powell, G., Ribas, R., et al. Solving rubik's cube with a robot hand. *arXiv preprint arXiv:1910.07113*, 2019.

Auer, P., Jaksch, T., and Ortner, R. Near-optimal regret bounds for reinforcement learning. In *Advances in neural information processing systems*, pp. 89–96, 2009.

Ayoub, A., Jia, Z., Szepesvari, C., Wang, M., and Yang, L. F. Model-based reinforcement learning with value-targeted regression. *arXiv preprint arXiv:2006.01107*, 2020.

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

Berner, C., Brockman, G., Chan, B., Cheung, V., Dębiak, P., Dennison, C., Farhi, D., Fischer, Q., Hashme, S., Hesse, C., et al. Dota 2 with large scale deep reinforcement learning. *arXiv preprint arXiv:1912.06680*, 2019.

Brafman, R. I. and Tennenholtz, M. R-max-a general polynomial time algorithm for near-optimal reinforcement learning. *Journal of Machine Learning Research*, 3(Oct): 213–231, 2002.

Bubeck, S. and Cesa-Bianchi, N. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *arXiv preprint arXiv:1204.5721*, 2012.

Cai, Q., Yang, Z., Jin, C., and Wang, Z. Provably efficient exploration in policy optimization. *arXiv preprint arXiv:1912.05830*, 2019.

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

Cheung, W. C., Simchi-Levi, D., and Zhu, R. Non-stationary reinforcement learning: The blessing of (more) optimism. *Available at SSRN 3397818*, 2019.

Dann, C. and Brunskill, E. Sample complexity of episodic fixed-horizon reinforcement learning. In *Advances in Neural Information Processing Systems*, pp. 2818–2826, 2015.

Dann, C., Lattimore, T., and Brunskill, E. Unifying pac and regret: Uniform pac bounds for episodic reinforcement learning. In *Advances in Neural Information Processing Systems*, pp. 5713–5723, 2017.

Derman, E., Mankowitz, D., Mann, T., and Mannor, S. A bayesian approach to robust reinforcement learning. In *Uncertainty in Artificial Intelligence*, pp. 648–658. PMLR, 2020.

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

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

Diakonikolas, I., Kamath, G., Kane, D. M., Li, J., Moitra, A., and Stewart, A. Being robust (in high dimensions) can be practical. *arXiv preprint arXiv:1703.00893*, 2017.

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

Diakonikolas, I., Kane, D. M., and Pensia, A. Outlier robust mean estimation with subgaussian rates via stability. *Advances in Neural Information Processing Systems*, 33, 2020.

Domingues, O. D., Ménard, P., Pirotta, M., Kaufmann, E., and Valko, M. A kernel-based approach to non-stationary reinforcement learning in metric spaces. *arXiv preprint arXiv:2007.05078*, 2020.

Du, S. S., Luo, Y., Wang, R., and Zhang, H. Provably efficient q-learning with function approximation via distribution shift error checking oracle. In *Advances in Neural Information Processing Systems*, pp. 8060–8070, 2019.

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

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

Gupta, A., Koren, T., and Talwar, K. Better algorithms for stochastic bandits with adversarial corruptions. *arXiv preprint arXiv:1902.08647*, 2019.

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

Jin, C., Allen-Zhu, Z., Bubeck, S., and Jordan, M. I. Is q-learning provably efficient? In *Advances in Neural Information Processing Systems*, pp. 4863–4873, 2018.

Jin, C., Netrapalli, P., Ge, R., Kakade, S. M., and Jordan, M. I. A short note on concentration inequalities for random vectors with subgaussian norm. *arXiv preprint arXiv:1902.03736*, 2019.

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

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

Jin, T. and Luo, H. Simultaneously learning stochastic and adversarial episodic mdps with known transition. *arXiv preprint arXiv:2006.05606*, 2020.

Kakade, S. and Langford, J. Approximately optimal approximate reinforcement learning. In *ICML*, volume 2, pp. 267–274, 2002.

Kakade, S., Krishnamurthy, A., Lowrey, K., Ohnishi, M., and Sun, W. Information theoretic regret bounds for online nonlinear control. *Advances in Neural Information Processing Systems*, 33, 2020.

Kakade, S. M. A natural policy gradient. *Advances in neural information processing systems*, 14:1531–1538, 2001.

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

Lee, C.-W., Luo, H., Wei, C.-Y., and Zhang, M. Bias no more: high-probability data-dependent regret bounds for adversarial bandits and mdps. *Advances in Neural Information Processing Systems*, 33, 2020.

Lykouris, T., Mirrokni, V., and Paes Leme, R. Stochastic bandits robust to adversarial corruptions. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*, pp. 114–122, 2018.

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

Ma, Y., Zhang, X., Sun, W., and Zhu, J. Policy poisoning in batch reinforcement learning and control. In *Advances in Neural Information Processing Systems*, pp. 14570–14580, 2019.

Neff, G. and Nagy, P. Automation, algorithms, and politicsl talking to bots: Symbiotic agency and the case of tay. *International Journal of Communication*, 10:17, 2016.

Neu, G. and Olkhovskaya, J. Online learning in mdps with linear function approximation and bandit feedback. *arXiv preprint arXiv:2007.01612*, 2020.

Neu, G., György, A., and Szepesvári, C. The online loop-free stochastic shortest-path problem. In *COLT*, volume 2010, pp. 231–243. Citeseer, 2010.

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

Ornik, M. and Topcu, U. Learning and planning for time-varying mdps using maximum likelihood estimation. *arXiv preprint arXiv:1911.12976*, 2019.

Ortner, R., Gajane, P., and Auer, P. Variational regret bounds for reinforcement learning. In *UAI*, pp. 16, 2019.

Osband, I. and Van Roy, B. Model-based reinforcement learning and the eluder dimension. *Advances in Neural Information Processing Systems*, 27:1466–1474, 2014.

Osband, I. and Van Roy, B. On lower bounds for regret in reinforcement learning. *arXiv preprint arXiv:1608.02732*, 2016.

Petersen, I. R., Ugrinovskii, V. A., and Savkin, A. V. *Robust Control Design Using H-∞ Methods*. Springer Science & Business Media, 2012.

Pinto, L., Davidson, J., Sukthankar, R., and Gupta, A. Robust adversarial reinforcement learning. In *International Conference on Machine Learning*, pp. 2817–2826. PMLR, 2017.

Rosenberg, A. and Mansour, Y. Online convex optimization in adversarial markov decision processes. *arXiv preprint arXiv:1905.07773*, 2019.

Schulman, J., Levine, S., Abbeel, P., Jordan, M., and Moritz, P. Trust region policy optimization. In *International conference on machine learning*, pp. 1889–1897, 2015a.

Schulman, J., Moritz, P., Levine, S., Jordan, M., and Abbeel, P. High-dimensional continuous control using generalized advantage estimation. *arXiv preprint arXiv:1506.02438*, 2015b.

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., and Klimov, O. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Shalev-Shwartz, S. et al. Online learning and online convex optimization. *Foundations and trends in Machine Learning*, 4(2):107–194, 2011.

Sun, W., Jiang, N., Krishnamurthy, A., Agarwal, A., and Langford, J. Model-based rl in contextual decision processes: Pac bounds and exponential improvements over model-free approaches. In *Conference on Learning Theory*, pp. 2898–2933. PMLR, 2019.

Sutton, R. S., McAllester, D. A., Singh, S. P., and Mansour, Y. Policy gradient methods for reinforcement learning with function approximation. In *Advances in Neural Information Processing Systems*, volume 99, pp. 1057–1063, 1999.

Todorov, E., Erez, T., and Tassa, Y. Mujoco: A physics engine for model-based control. In *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 5026–5033. IEEE, 2012.

Tropp, J. A. An introduction to matrix concentration inequalities. *arXiv preprint arXiv:1501.01571*, 2015.

Williams, R. J. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3-4):229–256, 1992.

Yadkori, Y. A., Bartlett, P. L., Kanade, V., Seldin, Y., and Szepesvári, C. Online learning in markov decision processes with adversarially chosen transition probability distributions. In *Advances in neural information processing systems*, pp. 2508–2516, 2013.

Yang, L. and Wang, M. Sample-optimal parametric q-learning using linearly additive features. In *International Conference on Machine Learning*, pp. 6995–7004. PMLR, 2019a.

Yang, L. F. and Wang, M. Reinforcement learning in feature space: Matrix bandit, kernels, and regret bound. *arXiv preprint arXiv:1905.10389*, 2019b.

Zanette, A., Brandfonbrener, D., Brunskill, E., Pirotta, M., and Lazaric, A. Frequentist regret bounds for randomized least-squares value iteration. In *International Conference on Artificial Intelligence and Statistics*, pp. 1954–1964, 2020.

Zhang, K., Hu, B., and Basar, T. Policy optimization for h-2 linear control with h-∞ robustness guarantee: Implicit regularization and global convergence. In *Learning for Dynamics and Control*, pp. 179–190. PMLR, 2020a.

Zhang, K., Hu, B., and Basar, T. On the stability and convergence of robust adversarial reinforcement learning: A case study on linear quadratic systems. *Advances in Neural Information Processing Systems*, 33, 2020b.

Zhang, K., Zhang, X., Hu, B., and Başar, T. Derivative-free policy optimization for risk-sensitive and robust control design: Implicit regularization and sample complexity. *arXiv preprint arXiv:2101.01041*, 2021.

Zhang, X., Ma, Y., Singla, A., and Zhu, X. Adaptive reward-poisoning attacks against reinforcement learning. *arXiv preprint arXiv:2003.12613*, 2020c.

Zhou, D., He, J., and Gu, Q. Provably efficient reinforcement learning for discounted mdps with feature mapping. *arXiv preprint arXiv:2006.13165*, 2020.

Zhou, K. and Doyle, J. C. *Essentials of robust control*, volume 104. Prentice hall Upper Saddle River, NJ, 1998.

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

Zinkevich, M. Online convex programming and generalized infinitesimal gradient ascent. In *Proceedings of the 20th international conference on machine learning (icml-03)*, pp. 928–936, 2003.

Appendices

Table of Contents

A. Proof for the lower-bound result

Theorem A.1 (Theorem 3.1)**.** *For any algorithm, there exists an MDP such that the algorithm fails to find an* $\left(\frac{\varepsilon}{2(1-\gamma)}\right)$*-optimal policy under the* ε*-contamination model with a probability of at least* $1/4$*.*

Proof of Theorem A.1. Consider two MDPs M_1, M_2, both with 3 states and 2 actions, defined as

$$P_1(s_2|s_1,a_1) = \frac{1-\varepsilon}{2}, P_1(s_3|s_1,a_1) = \frac{1+\varepsilon}{2}, P_1(s_3|s_1,a_2) = P_1(s_3|s_1,a_2) = \frac{1}{2} \tag{9}$$

$$P_2(s_2|s_1,a_1) = \frac{1+\varepsilon}{2}, P_2(s_3|s_1,a_1) = \frac{1-\varepsilon}{2}, P_2(s_3|s_1,a_2) = P_2(s_3|s_1,a_2) = \frac{1}{2} \tag{10}$$

and for both MDPs s_2, s_3 are absorbing states with constant reward 1 and 0, respectively. So for M_1, the optimal policy is $\pi_1^*(s_1) = a_2$, and for M_2, the optimal policy is $\pi_2^*(s_1) = a_1$. In both cases, choosing the alternative action in s_1 will incur a suboptimality gap of $\frac{\varepsilon}{2(1-\gamma)}$.

Let $N(\cdot)$ be the probability function of Bernoulli distribution on $\{s_2, s_3\}$: $N(x) = \begin{cases} 1 & \text{if } x = s_2 \\ 0 & \text{if } x = s_3 \end{cases}$. First of all, notice that an 2ε-*oblivious adversary* can make the two MDPs M_1, M_2 indistinguishable by changing $P_1(\cdot \mid s_1, a_1)$ to be $(1 - \frac{2\varepsilon}{1+\varepsilon})P_1(\cdot \mid s_1, a_1) + \frac{2\varepsilon}{1+\varepsilon}N(\cdot)$, which is exactly $P_2(\cdot \mid s_1, a_1)$. Note that $\frac{2\varepsilon}{1+\varepsilon} \leq 2\varepsilon$ and thus can be achieved by a 2ε-oblivious adversary.

When the two MDPs are indistinguishable, any rollout has the same probability under both MDP, and thus conditioned on any roll-out, the learner can at best obtain an $\frac{\varepsilon}{2(1-\gamma)}$-optimal policy with probability $1/2$ on both MDP.

What remains to be shown is that with high probability, the ε-contamination adversary can simulate the oblivious adversary.

Let X_i, Y_i be Bernoulli random variables s.t. $X_i = \begin{cases} s_2 & U \leq \frac{1-\varepsilon}{2} \\ s_3 & \text{o.w.} \end{cases}$, $Y_i = \begin{cases} s_2 & U \leq \frac{1+\varepsilon}{2} \\ s_3 & \text{o.w.} \end{cases}$, where U is picked uniformly random in $[0, 1]$. Then (X_i, Y_i) is a coupling with law: $P((X_i, Y_i) = (s_2, s_2)) = \frac{1-\varepsilon}{2}$, $P((X_i, Y_i) = (s_2, s_3)) = 0$, $P((X_i, Y_i) = (s_3, s_2)) = \varepsilon$, $P((X_i, Y_i) = (s_3, s_3)) = \frac{1-\varepsilon}{2}$, X_i and Y_i can be thought as the outcome of $P_1(\cdot \mid s_1, a_1)$, $P_2(\cdot \mid s_1, a_1)$ respectively. The ε-contamination adversary can simulate the oblivious adversary by changing X_i to Y_i when $X_1 \neq Y_i$, which has probability ε. This is possible when there are at most ε fraction of index i s.t. $X_i \neq Y_i$. Suppose there are T episodes, then

$$P\left(\sum_{i=1}^{T} \mathbb{1}_{\{a_1 \text{ is taken at } s_1\}} \mathbb{1}_{\{X_i \neq Y_i\}} \geq \varepsilon T\right) \leq P(\sum_{i=1}^{T} \mathbb{1}_{\{X_i \neq Y_i\}} \geq T\varepsilon) \leq \frac{1}{2} \tag{11}$$

because the median of Binomial(n, p) is at most $\lceil np \rceil$. Therefore, the probability that the adaptive adversary can simulate the oblivious adversary throughout T episodes is at least $1/2$. Assuming that when the adversary fails to simulate, the learner automatically succeed in finding the optimal policy, then we've established that the learner will still fail to find an $\left(\frac{\varepsilon}{2(1-\gamma)}\right)$-optimal policy with probability $1/4$ on both MDPs. ■

B. Property of $\hat{Q}(s,a)$ sampled from Algorithm 1

To prepare for the analysis that follows, we first show that the $\hat{Q}(s,a)$ sampled from Algorithm 1 is unbiased and has bounded variance.

Lemma B.1. $\mathbb{E}\left[\hat{Q}^\pi(s,a)\right] = Q^\pi(s,a)$, *Var*$(\hat{Q}^\pi(s,a)) \le \frac{\gamma}{(1-\gamma)^2} + \frac{\sigma^2}{1-\gamma}$. *The bound for variance is tight.*

Proof of Lemma B.1. In the following, we treat (s_0, a_0) as deterministic.

$$\begin{aligned}
\mathbb{E}\left[\hat{Q}^\pi(s_0,a_0)\right] =& \sum_{k=0}^{\infty} \mathbb{E}\left[\sum_{t=0}^{T} r(s_t,a_t) \middle| T=k\right] P(T=k) \quad \text{(by law of total expectation)}\\
=& \sum_{k=0}^{\infty} \mathbb{E}\left[\sum_{t=0}^{k} r(s_t,a_t)\right](1-\gamma)\gamma^k \quad \text{(each } r(s,a) \text{ is independent of } T\text{)}\\
=&(1-\gamma)\sum_{k=0}^{\infty}\frac{\gamma^k}{1-\gamma}\mathbb{E}\left[r(a_k,s_k)\right]\\
=&Q^\pi(s_0,a_0)
\end{aligned}$$

Now, we upperbound the variance. Let $\bar{r}(s,a) := r(s,a) - e(s,a)$ be the expected reward over the zero-mean noise. Because the zero-mean noise is independent of state transition, we observe that:

$$\begin{aligned}
\mathbb{E}\left[r(s,a)\right] =& \mathbb{E}\left[\bar{r}(s,a)\right]\\
\mathbb{E}\left[r(s,a)^2\right] =& \mathbb{E}\left[(\bar{r}(s,a)+e(s,a))^2\right] = \mathbb{E}\left[\bar{r}(s,a)^2\right] + \mathbb{E}\left[e(s,a)^2\right] \le \mathbb{E}\left[\bar{r}(s,a)^2\right] + \sigma^2\\
\mathbb{E}\left[r(s_i,a_i)r(s_j,a_j)\right] =& \mathbb{E}\left[(\bar{r}(s_i,a_i)+e(s_i,a_i))(\bar{r}(s_j,a_j)+e(s_j,a_j))\right] = \mathbb{E}\left[\bar{r}(s_i,a_i)\bar{r}(s_j,a_j)\right],
\end{aligned}$$

for $i \neq j$.

Given the above observations, we can bound the variance as follows

$$\begin{aligned}
& \text{Var}(\hat{Q}^\pi(s_0,a_0))\\
\le\;& \sigma^2 + \mathbb{E}\left[(\hat{Q}^\pi(s_0,a_0) - \bar{r}(s_0,a_0))^2\right] - \left(\mathbb{E}\left[\hat{Q}^\pi(s_0,a_0)\right] - \bar{r}(s_0,a_0)\right)^2 \quad \text{(separate the variance of } r(s_0,a_0)\text{)}\\
=\;& \sigma^2 + \sum_{k=1}^{\infty}(1-\gamma)\gamma^k\mathbb{E}\left[\left(\sum_{t=1}^{k} r(s_t,a_t)\right)^2\right] - \left(\mathbb{E}\left[\hat{Q}^\pi(s_0,a_0)\right] - \bar{r}(s_0,a_0)\right)^2\\
=\;& \sigma^2 + \sum_{k=1}^{\infty}(1-\gamma)\gamma^k\left(\sum_{t=1}^{k}\mathbb{E}\left[r(s_t,a_t)^2\right] + 2\sum_{i=1}^{k}\sum_{j=i+1}^{k}\mathbb{E}\left[r(s_i,a_i)r(s_j,a_j)\right]\right) - \left(\mathbb{E}\left[\hat{Q}^\pi(s_0,a_0)\right] - \bar{r}(s_0,a_0)\right)^2\\
=\;& \sigma^2 + \sum_{t=1}^{\infty}\gamma^t\mathbb{E}\left[r(s_t,a_t)^2\right] + 2\sum_{i=1}^{\infty}\sum_{j=i+1}^{\infty}\gamma^j\mathbb{E}\left[r(s_i,a_i)r(s_j,a_j)\right] - \left(\mathbb{E}\left[\hat{Q}^\pi(s_0,a_0)\right] - \bar{r}(s_0,a_0)\right)^2\\
\le\;& \frac{\sigma^2}{1-\gamma} + \sum_{t=1}^{\infty}\gamma^t\mathbb{E}\left[\bar{r}(s_t,a_t)^2\right] + 2\sum_{i=1}^{\infty}\sum_{j=i+1}^{\infty}\gamma^j\mathbb{E}\left[\bar{r}(s_i,a_i)\bar{r}(s_j,a_j)\right] - \left(\mathbb{E}\left[\hat{Q}^\pi(s_0,a_0)\right] - \bar{r}(s_0,a_0)\right)^2\\
\le\;& \frac{\sigma^2}{1-\gamma} + \sum_{t=1}^{\infty}\gamma^t\mathbb{E}\left[\bar{r}(s_t,a_t)\right] + 2\sum_{i=1}^{\infty}\sum_{j=i+1}^{\infty}\gamma^j\mathbb{E}\left[\bar{r}(s_i,a_i)\right] - \left(\mathbb{E}\left[\hat{Q}^\pi(s_0,a_0)\right] - \bar{r}(s_0,a_0)\right)^2\\
=\;& \frac{\sigma^2}{1-\gamma} + \sum_{t=1}^{\infty}\gamma^t\mathbb{E}\left[\bar{r}(s_t,a_t)\right] + 2\sum_{i=1}^{\infty}\frac{\gamma^{i+1}}{1-\gamma}\mathbb{E}\left[\bar{r}(s_i,a_i)\right] - \left(\mathbb{E}\left[\hat{Q}^\pi(s_0,a_0)\right] - \bar{r}(s_0,a_0)\right)^2
\end{aligned}$$

$$
\begin{aligned}
&= \frac{\sigma^2}{1-\gamma} + \frac{1+\gamma}{1-\gamma}\sum_{t=1}^{\infty}\gamma^t\mathbb{E}\left[\bar{r}(s_t,a_t)\right] - \left(\sum_{t=1}^{\infty}\gamma^t\mathbb{E}\left[\bar{r}(s_t,a_t)\right]\right)^2 \\
&= -\left(\sum_{t=1}^{\infty}\gamma^t\mathbb{E}\left[\bar{r}(s_t,a_t)\right] - \frac{1+\gamma}{2(1-\gamma)}\right)^2 + \frac{(1+\gamma)^2}{4(1-\gamma)^2} + \frac{\sigma^2}{1-\gamma} \\
&\leq -\left(\sum_{t=1}^{\infty}\gamma^t - \frac{1+\gamma}{2(1-\gamma)}\right)^2 + \frac{(1+\gamma)^2}{4(1-\gamma)^2} + \frac{\sigma^2}{1-\gamma} = \frac{\gamma}{(1-\gamma)^2} + \frac{\sigma^2}{1-\gamma}
\end{aligned}
$$

The last line is because:

$$\sum_{t=1}^{\infty}\gamma^t\mathbb{E}\left[\bar{r}(s_t,a_t)\right] \leq \sum_{t=1}^{\infty}\gamma^t = \frac{\gamma}{1-\gamma} \leq \frac{1+\gamma}{2(1-\gamma)}.$$

The equality can be reached by the following reward setting: let $P(1 = \bar{r}(s_1,a_1) = \cdots = \bar{r}(s_t,a_t) = \cdots) = 1$ and therefore is tight. ■

C. Proofs for Section 4.

Lemma C.1 (Lemma 4.2). *Suppose the adversarial rewards are bounded in* $[0,1]$, *and in a particular iteration* t, *the adversary contaminates* $\varepsilon^{(t)}$ *fraction of the episodes, then given M episodes, it is guaranteed that with probability at least* $1-\delta$,

$$\mathbb{E}_{s,a\sim d^{(t)}}\left[\left(Q^{\pi^{(t)}}(s,a) - \phi(s,a)^\top w^{(t)}\right)^2\right] \leq 4\left(W^2 + WH\right)\left(\varepsilon^{(t)} + \sqrt{\frac{8}{M}\log\frac{4d}{\delta}}\right).$$

where $H = (\log\delta - \log M)/\log\gamma$ *is the effective horizon.*

Proof of Lemma C.1. First of all, observe that since the adversarial reward is bounded in $[0,1]$, with probability $1-\delta$, the $\hat{Q}(s,a)$ estimates collected in the adversarial episodes are bounded by $H := (\log\delta - \log M)/\log\gamma$.

Conditioned on the above event, consider three loss functions $\hat{f}$, $f^\dagger$ and f, representing the loss w.r.t. clean data, corrupted data and underlying distribution respectively, i.e.

$$\hat{f} = \frac{1}{M}\sum_{i=1}^{M}(y_i - x_i^\top w)^2 \tag{12}$$

$$f^\dagger = \frac{1}{M}\left[\sum_{i\in C}(y_i^\dagger - x_i^{\dagger\top} w)^2 + \sum_{i\notin C}(y_i - x_i^\top w)^2\right] \tag{13}$$

$$f = \mathbb{E}(y_i - x_i^\top w)^2 \tag{14}$$

Then, for all w, we can make the following decomposition

$$||\nabla_w f^\dagger - \nabla_w f|| \leq ||\nabla_w f^\dagger - \nabla_w \hat{f}|| + ||\nabla_w \hat{f} - \nabla_w f||. \tag{15}$$

We next bound each of the two terms in equation 15. For the first term,

$$
\begin{aligned}
&||\nabla_w f^\dagger - \nabla_w \hat{f}|| && (16)\\
=\ &\left\|\frac{2}{M}\sum_{i\in C}\left[(x_i^\dagger x_i^{\dagger\top} - x_i x_i^\top)w + (y_i^\dagger x_i^\dagger - y_i x_i)\right]\right\| && (17)\\
\leq\ &4\,(W+H)\,\varepsilon^{(t)} && (18)
\end{aligned}
$$

where the last step uses the fact that $|C|/M \leq \varepsilon^{(t)}$, and $||x|| \leq 1$, $|y^\dagger| \leq H$ and $||w|| \leq W$. For the second term

$$||\nabla_w \hat{f} - \nabla_w f|| \tag{19}$$

$$\leq \quad 2\left\|\left(\mathbb{E}[xx^\top] - \frac{1}{M}\sum_{i=1}^{M} x_i x_i^\top\right) w - \left(\mathbb{E}[yx] - \frac{1}{M}\sum_{i=1}^{M} y_i x_i\right)\right\| \tag{20}$$

$$\leq \quad 2\left(\frac{2}{3M}\log\frac{4d}{\delta} + \sqrt{\frac{2}{M}\log\frac{4d}{\delta}}\right) W + 2\sqrt{\frac{2}{M}\log\frac{4d}{\delta}}\cdot 2H \tag{21}$$

$$\leq \quad 4\sqrt{\frac{8}{M}\log\frac{4d}{\delta}}\,(W+H)\,\text{, for } M \geq 2\log\frac{4d}{\delta}. \tag{22}$$

where in step (21) we apply Matrix Bernstein inequality (Tropp, 2015) on the first term and vector Hoeffding's inequality (Jin et al., 2019) on the second term. The constant in Corollary 7 of (Jin et al., 2019) is instantiated to be $c = 1$, because boundedness means we always have condition 2 in Lemma 2 of (Jin et al., 2019). This condition is all we need throughout the proof for the vector Hoeffding.

Now, let M be sufficiently large, and instantiate w to be w^t, i.e. the constrained linear regression solution w.r.t $f^\dagger$, then our result above implies that for any vector v such that $||w+v|| \leq W$, we have $\nabla_w f^\dagger(w^t)^\top v/||v|| \geq 0$, and thus

$$\nabla_w f(w^t)^\top v/||v|| \geq -4\,(W+H)\left(\varepsilon^{(t)} + \sqrt{\frac{8}{M}\log\frac{4d}{\delta}}\right) \tag{23}$$

which by Lemma B.8 of (Diakonikolas et al., 2019) implies that

$$\varepsilon^{(t)}_{stat} \leq 4\left(W^2 + HW\right)\left(\varepsilon^{(t)} + \sqrt{\frac{8}{M}\log\frac{4d}{\delta}}\right)\text{, w.p. } 1-2\delta. \tag{24}$$

■

Theorem C.1 (Theorem 4.1)**.** *Under assumptions 3.1 (linear Q function) and 3.2 (reset distribution with small κ), given a desired optimality gap α, there exists a set of hyperparameters agnostic to the contamination level ε, such that Algorithm 2 guarantees with a $poly(1/\alpha, 1/(1-\gamma), |\mathcal{A}|, W, \sigma, \kappa)$ sample complexity that under ε-contamination with adversarial rewards bounded in $[0,1]$, we have*

$$\mathbb{E}\left[V^*(\mu_0) - V^{\hat{\pi}}(\mu_0)\right] \leq \tilde{O}\left(\max\left[\alpha, W\sqrt{\frac{|\mathcal{A}|\kappa\varepsilon}{(1-\gamma)^3}}\right]\right)$$

where $\hat{\pi}$ is the uniform mixture of $\pi^{(1)}$ through $\pi^{(T)}$.

Proof of Theorem C.1. First note that $\varepsilon_{stat} = \mathbb{E}_{s,a\sim d^{(t)}}[(\phi(s,a)^\top(w^{(t)} - w^*))^2] \leq 4W^2$, because $\|\phi(s,a)\| \leq 1$ and $\|w^{(t)}\|, \|w^*\| \leq W$. As a result, the high probability bound in Lemma 4.2 can be ready translate into an expected bound:

$$\mathbb{E}\left[\mathbb{E}_{s,a\sim d^{(t)}}\left[\left(Q^{\pi^{(t)}}(s,a) - \phi(s,a)^\top w^{(t)}\right)^2\right]\right] \leq 4\left(W^2 + HW\right)\left(\varepsilon^{(t)} + \sqrt{\frac{8}{M}\log\frac{4d}{\delta}}\right) + 8\delta W^2 \tag{25}$$

where δ becomes a free parameter. Plugging this into Lemma 4.1, we get

$$\begin{aligned}
&\mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T}\{V^*(\mu_0) - V^{(t)}(\mu_0)\}\right] \\
\leq\;& \frac{W}{1-\gamma}\sqrt{\frac{2\log|\mathcal{A}|}{T}} + \frac{1}{T}\sum_{t=1}^{T}\sqrt{\frac{4|\mathcal{A}|\kappa\varepsilon^{(t)}_{stat}}{(1-\gamma)^3}} \\
\leq\;& \frac{W}{1-\gamma}\sqrt{\frac{2\log|\mathcal{A}|}{T}} + \frac{1}{T}\sum_{t=1}^{T}\sqrt{\frac{16|\mathcal{A}|\kappa\left((W^2+HW)\left(\varepsilon^{(t)} + \sqrt{\frac{8}{M}\log\frac{4d}{\delta}}\right) + 2\delta W^2\right)}{(1-\gamma)^3}}
\end{aligned}$$

$$
\begin{aligned}
&\leq \frac{W}{1-\gamma}\sqrt{\frac{2\log|\mathcal{A}|}{T}} + \frac{1}{T}\sum_{t=1}^{T}\sqrt{\frac{16|\mathcal{A}|\kappa\left((W^2+HW)\sqrt{\frac{8}{M}\log\frac{4d}{\delta}}+2\delta W^2\right)}{(1-\gamma)^3}} + \frac{1}{T}\sum_{t=1}^{T}\sqrt{\frac{16|\mathcal{A}|\kappa\left(W^2+HW\right)\varepsilon^{(t)}}{(1-\gamma)^3}} \\
&\leq \frac{W}{1-\gamma}\sqrt{\frac{2\log|\mathcal{A}|}{T}} + \sqrt{\frac{16|\mathcal{A}|\kappa\left((W^2+HW)\sqrt{\frac{8}{M}\log\frac{4d}{\delta}}+2\delta W^2\right)}{(1-\gamma)^3}} + \sqrt{\frac{16|\mathcal{A}|\kappa\left(W^2+HW\right)\varepsilon}{(1-\gamma)^3}}
\end{aligned}
$$

where the last step is by Cauchy Schwarz and the fact that the attacker only has ε budget to distribute, which implies that $\sum_{t=1}^{T}\varepsilon^{(t)} = T\varepsilon$. Setting

$$
T = \frac{2W^2\log|\mathcal{A}|}{\alpha^2(1-\gamma)^2} \tag{26}
$$

$$
\delta = \frac{\alpha^2(1-\gamma)^3}{32W^2|\mathcal{A}|\kappa} \tag{27}
$$

$$
M = \frac{512|\mathcal{A}|^2W^2(W+H)^2\kappa^2}{\alpha^4(1-\gamma)^6}\log\frac{4d}{\delta}, \tag{28}
$$

we get

$$
\mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T}\{V^*(\mu_0)-V^{(t)}(\mu_0)\}\right] \leq 3\alpha + \sqrt{\frac{16|\mathcal{A}|\kappa\left(W^2+HW\right)\varepsilon}{(1-\gamma)^3}}. \tag{29}
$$

with sample complexity

$$
TM = \frac{1024|\mathcal{A}|^2\log|\mathcal{A}|W^4(W+H)^2\kappa^2}{\alpha^6(1-\gamma)^8}\log\frac{128W^2|\mathcal{A}|\kappa d}{\alpha^2(1-\gamma)^3}. \tag{30}
$$

■

Next, we prove this tighter version of Theorem 4.1 in the special case of tabular MDPs.

Corollary C.1 (Corollary 4.1)**.** *Given a tabular MDP and assumption 3.2, given a desired optimality gap α, there exists a set of hyperparameters agnostic to the contamination level ε, such that Algorithm 2 guarantees with a poly$(1/\alpha, 1/(1-\gamma), |\mathcal{A}|, W, \sigma, \kappa)$ sample complexity that under ε-contamination with adversarial rewards bounded in $[0,1]$, we have*

$$
\mathbb{E}\left[V^*(\mu_0)-V^{\hat{\pi}}(\mu_0)\right] \leq \tilde{O}\left(\max\left[\alpha, \sqrt{\frac{|\mathcal{A}|\kappa\varepsilon}{(1-\gamma)^5}}\right]\right) \tag{31}
$$

where $\hat{\pi}$ is the uniform mixture of $\pi^{(1)}$ through $\pi^{(T)}$.

The proof follows the exact same structure as the proof of Theorem C.1, but with a tighter robustness bound of linear regression.

Lemma C.2. *Assume a tabular MDP and the adversarial rewards are bounded in $[0,1]$, and in a particular iteration t, the adversary contaminates $\varepsilon^{(t)}$ fraction of the episodes, then given M episodes, it is guaranteed that with probability at least $1-\delta$,*

$$
\mathbb{E}_{s,a\sim d^{(t)}}\left[\left(Q^{\pi^{(t)}}(s,a)-\phi(s,a)^\top w^{(t)}\right)^2\right] \leq H^2\varepsilon^{(t)} + 3\left(W^2+WH\right)\sqrt{\frac{\log 1/\delta}{M}}. \tag{32}
$$

where $H = (\log\delta - \log M)/\log\gamma$ is the effective horizon.

Proof of Lemma C.2. The proof is largely based on Lemma G.1 of (Agarwal et al., 2020a). We assumed that the constrained linear regression problem is solved using Projected Online Gradient Descent (Zinkevich, 2003) on the sequence of loss functions $(w^\top\phi_i - \hat{Q}_i)^2$, i.e.

$$
w_{i+1} = \text{Proj}_{\|w\|\leq W}\left(w_i - \eta_i(w_i^\top\phi_i - \hat{Q}_i)\phi_i\right), \text{ for all } i\in[M], \tag{33}
$$

where $\eta_i = W^2/((W+H)\sqrt{N})$ and we set $w^{(t)} = \frac{1}{M}\sum_{i=1}^{M} w_i$.

Using the projected online gradient descent regret guarantee, we have that:

$$\sum_{i\in C}(w_i^\top \phi_i^\dagger - \hat{Q}_i^\dagger)^2 + \sum_{i\notin C}(w_i^\top \phi_i - \hat{Q}_i)^2 \le \sum_{i\in C}(w^{\star\top} \phi_i^\dagger - \hat{Q}_i^\dagger)^2 + \sum_{i\notin C}(w^{\star\top} \phi_i - \hat{Q}_i)^2 + \underbrace{W(W+H)}_{:=Q}\sqrt{M}. \tag{34}$$

which implies

$$\begin{aligned} &\sum_{i\in[M]}(w_i^\top \phi_i - \hat{Q}_i)^2 - \sum_{i\in[M]}(w^{\star\top} \phi_i - \hat{Q}_i)^2 & (35)\\ \le &\sum_{i\in C}\left[(w^{\star\top} \phi_i^\dagger - \hat{Q}_i^\dagger)^2 - (w^{\star\top} \phi_i - \hat{Q}_i)^2\right] - \sum_{i\in C}\left[(w_i^\top \phi_i^\dagger - \hat{Q}_i^\dagger)^2 - (w_i^\top \phi_i - \hat{Q}_i)^2\right] + Q\sqrt{M}. & (36)\end{aligned}$$

We now want to show by induction that $w_i^\top \phi \in [0,H]$ for any i and ϕ. $w_0 = 0$ which satisfies $w_0^\top \phi \in [0,H]$. Now, assume that $w_i^\top \phi \in [0,H]$, we want to show $w_{i+1}^\top \phi \in [0,H]$. In a tabular MDP, ϕ is an one-hot vector, and thus for $\phi \ne \phi_i$, $w_{i+1}^\top \phi = w_i^\top \phi \in [0,H]$. If $\phi = \phi_i$, then

$$w_{i+1}^\top \phi = \Big(w_i - \eta_i(w_i^\top \phi_i - \hat{Q}_i)\phi_i\Big)^\top \phi_i \le (1-\eta_i) w_i^\top \phi_i + \eta \hat{Q}_i \in [0,H] \tag{37}$$

because both $w_i \top \phi_i$ (by induction hypothesis) and $\hat{Q}_i$ (by assumption on bounded attack) are in $[0,H]$. Therefore, we have shown that $w_i^\top \phi \in [0,H]$ for any i and ϕ. Then, (36) implies that

$$\sum_{i\in[M]}(w_i^\top \phi_i - \hat{Q}_i)^2 \le \sum_{i\in[M]}(w^{\star\top} \phi_i - \hat{Q}_i)^2 + 2H^2\varepsilon^{(t)}M + Q\sqrt{M}. \tag{38}$$

Denote random variable $z_i = (\theta_i \cdot x_i - y_i)^2 - (\theta^\star \cdot x_i - y_i)^2$. Denote $\mathbb{E}_i$ as the expectation taken over the randomness at step i conditioned on all history $t = 1$ to $i-1$. Note that for $\mathbb{E}_i[z_i]$, we have:

$$\begin{aligned} &\mathbb{E}_i\left[(\theta_i \cdot x - y)^2 - (\theta^\star \cdot x - y)^2\right] & (39)\\ &= \mathbb{E}_i\left[(\theta_i \cdot x - \mathbb{E}[y|x])^2\right] & (40)\\ &\quad - \mathbb{E}_i\left[2(\theta_i \cdot x - \mathbb{E}[y|x])(\mathbb{E}[y|x] - y) - (\theta^\star \cdot x - \mathbb{E}[y|x])^2 + 2(\theta^\star \cdot x - \mathbb{E}[y|x])(\mathbb{E}[y|x] - y))\right] & (41)\\ &= \mathbb{E}_i\left[(\theta_i \cdot x - \mathbb{E}[y|x])^2 - (\theta^\star \cdot x - \mathbb{E}[y|x])^2\right], & (42)\end{aligned}$$

where we use $\mathbb{E}[\mathbb{E}[y|x] - y] = 0$. Also for $|z_i|$, we can show that for $|z_i|$ we have:

$$|z_i| = |(\theta_i \cdot x_i - \theta^\star \cdot x_i)(\theta_i \cdot x_i + \theta^\star \cdot x_i - 2y_i)| \le W(2W + 2H) = 2W(W+H). \tag{43}$$

Note that z_i forms a Martingale difference sequence. Using Azuma-Hoeffding's inequality, we have that with probability at least $1-\delta$:

$$\left|\sum_{i=1}^{M} z_i - \sum_{i=1}^{M}\mathbb{E}_i\left[(\theta_i \cdot x - \mathbb{E}[y|x])^2 - (\theta^\star \cdot x - \mathbb{E}[y|x])^2\right]\right| \le 2W(W+H)\sqrt{\ln(1/\delta)M}, \tag{44}$$

which implies that:

$$\begin{aligned} &\sum_{i=1}^{M}\mathbb{E}_i\left[(\theta_i \cdot x - \mathbb{E}[y|x])^2 - (\theta^\star \cdot x - \mathbb{E}[y|x])^2\right] \le \sum_{i=1}^{M} z_i + 2W(W+H)\sqrt{\ln(1/\delta)M} & (45)\\ &\le 2W(W+H)\sqrt{\ln(1/\delta)M} + 2H^2 M\varepsilon^{(t)} + Q\sqrt{M}. & (46)\end{aligned}$$

Apply Jensen's inequality on the LHS of the above inequality, we have that:

$$\mathbb{E}\left(\hat{\theta} \cdot x - \mathbb{E}[y|x]\right)^2 \le \mathbb{E}\left(\theta^\star \cdot x - \mathbb{E}[y|x]\right)^2 + 2H^2\varepsilon^{(t)} + (Q + 2W(W+H))\sqrt{\frac{\ln(1/\delta)}{M}}. \tag{47}$$

∎

D. A modified analysis for SEVER

In this section, we will derive an expected error bound for SEVER (Diakonikolas et al., 2019) when applied to a linear regression problem. The high level idea is to use the results of (Diakonikolas et al., 2020) to show the existence of a stable set and change the probabilistic argument in (Diakonikolas et al., 2019) to an expectation argument. We note that the original result in (Diakonikolas et al., 2019) works only with probability $9/10$, and there is no direct way of translating it into either a high-probability argument or an expectation argument.

In the following, we consider a robust linear regression problem. We observe pairs $(X_i, Y_i) \in \mathbb{R}^d \times \mathbb{R}$ for $i \in [n]$, where X_i's are drawn i.i.d. from a distribution D_x and $Y_i = w^{*\top} X_i + e_i$ for some unknown $w^* \in \mathbb{R}^d$. e_i's are i.i.d, noise from some distribution $D_{e|x}$. Note that here e_i and X_i may not be independent. We let D_{xy} be the joint distribution of (X, Y). Let $f_i(w) = (Y_i - w^\top X_i)^2$. Given a multiset of observations $\{(X_i, Y_i)\}_{i=1}^n$, our goal is to minimize the objective function

$$\bar{f}(w) = \mathbb{E}_{(X,Y)\sim D_{xy}}[(Y - w^\top X)^2] \tag{48}$$

on a convex feasible set $\mathcal{H}$. Let $r := \max_{w\in\mathcal{H}} \|w\|$ be the ℓ_2-radius of $\mathcal{H}$. In the following, we use $\|\cdot\|$ to denote the spectral norm of a matrix and the 2-norm of a vector. We use Cov to denote the covariance matrix of a random vector: $\mathrm{Cov}[X] = \mathbb{E}\left[(X - \mathbb{E}X)(X - \mathbb{E}X)^\top\right]$. When S is a set, we use $\mathbb{E}_S$ and Cov_S to denote the expectation and covariance over the empirical distribution on S. We allow for an ε-fraction of the observations to be arbitrary outliers. The ε-corruption model is defined in more detail in the Appendix A of (Diakonikolas et al., 2019).

Due to our application, we make assumptions on the linear regression model that is slight different from Assumption E.1 in (Diakonikolas et al., 2019):

Assumption D.1. *Given the model for linear regression described above, assume the following conditions for $D_{e|x}$ and D_x:*

- $\mathbb{E}\left[e|X\right] = 0$*;*
- $\mathbb{E}\left[e^2\middle|X\right] \leq \xi$*;*
- $\mathbb{E}_{X\sim D_x}[XX^\top] \preceq s^2 I$ *for some* $s > 0$*;*
- *There is a constant* $C > 0$*, such that for all unit vectors* v*,* $\mathbb{E}_{X\sim D_x}[\langle v, X\rangle^4] \leq Cs^4$.

In (Diakonikolas et al., 2019), the noise term e and X are independent. We weaken the assumption on e and bound its first and second moments conditional on X.

D.1. Stability with subgaussian rate

We first note that the gradient of f_i, $\nabla f_i(w)$ has bounded covariance matrix. We will show this by following the proof of Lemma E.3 in (Diakonikolas et al., 2019), but make minor changes as we do not assume e and X are independent:

Lemma D.1 (A variant of Lemma E.3 in (Diakonikolas et al., 2019))**.** *Suppose D_{xy} satisfies the conditions of Assumption D.1. Then for all unit vectors $v \in \mathbb{R}^d$, we have*

$$v^\top \mathrm{Cov}_{(X_i,Y_i)\sim D_{xy}}[\nabla f_i(w)]v \leq 4s^2\xi + 4Cs^4\|w^* - w\|^2. \tag{49}$$

Proof of Lemma D.1. We first note that $f_i(w) = (Y_i - w^\top X_i)^2$ and $\nabla f_i(w) = -2((w^* - w)^\top X_i + e_i)X_i$. By the property of conditional expectation, for any function $g(\cdot), h(\cdot)$, we have $\mathbb{E}\left[g(X)h(e)\right] = \mathbb{E}_X\left[\mathbb{E}_{h(e)|X}\left[g(X)h(e)|X\right]\right] = \mathbb{E}_X\left[g(X)\mathbb{E}_{h(e)|X}\left[h(e)|X\right]\right]$. Then

$$\begin{aligned}
& \mathbb{E}\left[\nabla f_i(w)\nabla f_i(w)^\top\right] = 4\mathbb{E}\left[((w^* - w)^\top X_i + e_i)^2 X_i X_i^\top\right] && (50)\\
= \;& 4\mathbb{E}\left[((w^* - w)^\top X_i)^2 X_i X_i^\top\right] + 4\mathbb{E}\left[e_i^2 X_i X_i^\top\right] + 4\mathbb{E}\left[2(w^* - w)^\top X_i e_i X_i X_i^\top\right] && (51)\\
= \;& 4\mathbb{E}\left[((w^* - w)^\top X_i)^2 X_i X_i^\top\right] + 4\mathbb{E}\left[X_i X_i^\top \mathbb{E}\left[e_i^2|X_i\right]\right] && (52)
\end{aligned}$$

By Assumption D.1, for all unit vectors $v \in \mathbb{R}^d$, we have

$$v^\top \mathbb{E}\left[((w^* - w)^\top X_i)^2 X_i X_i^\top\right] v \;=\; \mathbb{E}\left[((w^* - w)^\top X_i)^2 (v^\top X_i)^2\right] \tag{53}$$

$$\begin{aligned} &\leq \sqrt{\mathbb{E}\left[((w^* - w)^\top X_i)^4\right]\mathbb{E}\left[(v^\top X_i)^4\right]} && (54)\\ &\leq Cs^4\|w^* - w\|^2 && (55)\end{aligned}$$

and

$$v^\top \mathbb{E}\left[X_i X_i^\top \mathbb{E}\left[e_i^2 | X_i\right]\right] v \leq \xi v^\top \mathbb{E}\left[X_i X_i^\top\right] v \leq s^2 \xi \tag{56}$$

Thus for all unit vectors $v \in \mathbb{R}^d$, we have

$$v^\top \operatorname{Cov}_{(X_i,Y_i)\sim D_{xy}}[\nabla f_i(w)] v \leq v^\top \mathbb{E}\left[\nabla f_i(w) \nabla f_i(w)^\top\right] v \leq 4s^2\xi + 4Cs^4\|w^* - w\|^2. \tag{57}$$

■

We then use the following Theorem D.1 to show that the observations $f_1, \ldots, f_n$ satisfies the Assumption D.2 with high probability:

Theorem D.1 (Theorem 1.4 in (Diakonikolas et al., 2020))**.** *Fix any* $0 < \tau < 1$*. Let* S *be a multiset of* n *i.i.d. samples from a distribution on* $\mathbb{R}^d$ *with mean* μ *and covariance* Σ*. Let* $\varepsilon' = \tilde{C}\left(\log(1/\tau)/n + \varepsilon\right) = O(1)$*, for some constant* $\tilde{C} > 0$*. Then, with probability at least* $1 - \tau$*, there exists a subset* $S' \subseteq S$ *such that* $|S'| \geq (1 - \varepsilon')n$ *and for every* $S'' \subseteq S'$ *with* $|S''| \geq (1 - 2\varepsilon')|S'|$*, the following conditions hold: (i)* $\|\mu_{S''} - \mu\| \leq \sqrt{\|\Sigma\|}\delta$*, and (ii)* $\|\overline{\Sigma}_{S''} - \|\Sigma\| I\| \leq \|\Sigma\|\delta^2/(2\varepsilon')$*, for* $\delta = O\left(\sqrt{(d\log d)/n} + \sqrt{\varepsilon} + \sqrt{\log(1/\tau)/n}\right)$*.*

where $\mu_{S''} = \frac{1}{|S''|}\sum_{x\in S''} x$ and $\overline{\Sigma}_{S''} = \frac{1}{|S''|}\sum_{x\in S''}(x-\mu)(x-\mu)^\top$.

We use a notion of stability similar to that in (Diakonikolas et al., 2019) but allow the parameter to depend on the confidence level and sample size:

Assumption D.2 (A variant of Assumption B.1 in (Diakonikolas et al., 2019))**.** *Fix* $0 < \varepsilon < 1/2$*. With probability at least* $1 - \tau$*, there exists an unknown set* $I_{good} \subseteq [n]$ *with* $|I_{good}| \geq (1 - \varepsilon)n$ *of "good" functions* $\{f_i\}_{i\in I_{good}}$ *and parameters* σ*,* $\alpha(\varepsilon, n, \tau), \beta(\varepsilon, n, \tau) \in \mathbb{R}_+$ *such that for all* $w \in \mathcal{H}$*:*

$$\left\|\frac{1}{|I_{good}|}\sum_{i\in I_{good}} \nabla f_i(w) - \nabla \bar{f}(w)\right\| \leq \sigma\alpha(\varepsilon, n, \tau) \tag{58}$$

and

$$\left\|\frac{1}{|I_{good}|}(\nabla f_i(w) - \nabla \bar{f}(w))(\nabla f_i(w) - \nabla \bar{f}(w))^\top\right\| \leq \sigma^2\beta(\varepsilon, n, \tau) \tag{59}$$

We can then equivalently write Theorem D.1 as the following Proposition:

Proposition D.1. Given a linear regression model $f_i(w) = (Y_i - w^\top X_i)^2$ satisfying Assumption D.1, $X_i \sim D_x$, $D_e \sim D_e$, with probability at least $1 - \tau$, $\{f_i\}_{i\in[n]}$ satisfies Assumption D.2 with $\sigma = 2s\sqrt{\xi} + 2\sqrt{C}s^2\|w^* - w\|$, $\alpha(\varepsilon, n, \tau) = O\left(\sqrt{(d\log d)/n} + \sqrt{\varepsilon} + \sqrt{\log(1/\tau)/n}\right)$ and $\beta(\varepsilon, n, \tau) = \left(\frac{d\log d}{\log(1/\tau)+n\varepsilon} + 1\right)$.

Proof of Proposition D.1. By Theorem D.1 and Lemma D.1, with probability at least $1 - \tau$, there exist an unknown set $I_{good} \subseteq [n]$ with $|I_{good}| \geq (1 - \varepsilon')n$, s.t.

$$\begin{aligned} &\left\|\frac{1}{|I_{good}|}(\nabla f_i(w) - \nabla \bar{f}(w))(\nabla f_i(w) - \nabla \bar{f}(w))^\top\right\| && (60)\\ \leq\ &\left\|\frac{1}{|I_{good}|}(\nabla f_i(w) - \nabla \bar{f}(w))(\nabla f_i(w) - \nabla \bar{f}(w))^\top - \|\operatorname{Cov}_{f\in p^*}[\nabla f]\| I\right\| + \|\operatorname{Cov}_{f\in p^*}[\nabla f]\| && (61)\\ \leq\ &\left(4s^2\xi + 4Cs^4\|w^* - w\|^2\right) O\left(\frac{d\log d}{\log(1/\tau) + n\varepsilon} + 1\right) && (62)\\ \leq\ &\left(2s\sqrt{\xi} + 2\sqrt{C}s^2\|w^* - w\|\right)^2 O\left(\frac{d\log d}{\log(1/\tau) + n\varepsilon} + 1\right) =: \sigma^2\beta(\varepsilon, n, \tau). && (63)\end{aligned}$$

$$\|\nabla \hat{f}(w) - \nabla \bar{f}(w)\| \leq \sigma O\left(\sqrt{(d\log d)/n} + \sqrt{\varepsilon} + \sqrt{\log(1/\tau)/n}\right) =: \sigma\alpha(\varepsilon, n, \tau). \tag{64}$$

■

D.2. The expected optimality gap

In order to prove the expected optimality gap, we first state a slightly modified version of the main theorem in (Diakonikolas et al., 2019) by specifying the probability of success;

Theorem D.2 (Theorem B.2 in (Diakonikolas et al., 2019))**.** *Let the corruption level $\varepsilon \in [0, c]$, for some small enough $c > 0$. Suppose that the functions $f_1, \dots, f_n, \bar{f} : \mathcal{H} \to \mathbb{R}$ are bounded below, and that Assumption D.2 is satisfied. Then SEVER applied to $f_1, \dots, f_n$ returns a point $w \in \mathcal{H}$ that, fix $p \geq \sqrt{\varepsilon}$, with probability at least $1 - p$, is a $O\left(\sigma\left(\alpha(\varepsilon, n, \tau) + \sqrt{\alpha(\varepsilon, n, \tau)^2 + \beta(\varepsilon, n, \tau)}\sqrt{\varepsilon/p}\right)\right)$-approximate critical point of $\bar{f}$, i.e. for all unit vectors v where $w + \lambda v \in \mathcal{H}$ for arbitrarily small positive λ, we have that $v \cdot \nabla f(w) \geq -O\left(\sigma\left(\alpha(\varepsilon, n, \tau) + \sqrt{\alpha(\varepsilon, n, \tau)^2 + \beta(\varepsilon, n, \tau)}\sqrt{\varepsilon/p}\right)\right)$.*

if $\bar{f}$ is convex, we have the following optimality gap. Recall r is the radius of the convex set $\mathcal{H}$ where w^* belongs.

Corollary D.1 (Corollary B.3 in (Diakonikolas et al., 2019))**.** *Let the corruption level $\varepsilon \in [0, c]$, for some small enough $c > 0$. For functions $f_1, \dots, f_n : \mathcal{H} \to \mathbb{R}$, suppose that Assumption D.2 holds and that $\mathcal{H}$ is convex. Then, fix $p \geq \sqrt{\varepsilon}$, with probability at least $1 - p$, the output of SEVER satisfies the following: if $\bar{f}$ is convex, the algorithm finds a $w \in \mathcal{H}$ such that $\bar{f}(w) - \bar{f}(w^*) = O\left(r\sigma\left(\alpha(\varepsilon, n, \tau) + \sqrt{\alpha(\varepsilon, n, \tau)^2 + \beta(\varepsilon, n, \tau)}\sqrt{\varepsilon/p}\right)\right)$*

Given Theorem D.1, we can prove the following expected optimality gap:

Theorem D.3 (expected optimality gap)**.** *Let the corruption level $\varepsilon \in [0, c]$, for some small enough $c > 0$. Let $\mathcal{H}$ be a convex set. Given n samples from a linear regression model $f(w) = (Y - w^\top X)^2$ satisfying Assumption D.1, where $X \sim D_x$, $e \sim D_e$, $Y = w^{*\top} X + e$ for some unknown $w^* \in \mathcal{H}$, SEVER will find a $w \in \mathcal{H}$, such that*

$$\mathbb{E}\left[\bar{f}(w) - \bar{f}(w^*)\right] = O\left(\left(sr\sqrt{\xi} + s^2r^2\right)\left(\tau + \sqrt{(d\log d)/n} + \sqrt{\varepsilon} + \sqrt{\log(1/\tau)/n}\right)\right). \tag{65}$$

where the expectation above is over both the randomness of SEVER and (X_i, Y_i) pairs.

Proof of Theorem D.3. In the following, we use α and β as shorthands of $\alpha(\varepsilon, n, \tau)$ and $\beta(\varepsilon, n, \tau)$. We first show that $\bar{f}(w) - \bar{f}(w^*)$ is upper bounded:

$$\begin{aligned} \bar{f}(w) - \bar{f}(w^*) &= \mathbb{E}_{X,Y}\left[(Y - w^\top X)^2 - (Y - w^{*\top} X)^2\right] &(66)\\ &= \mathbb{E}_{X,e}\left[(w^* - w)^\top X + e)^2 - e^2\right] &(67)\\ &= (w^* - w)^\top \mathbb{E}_X[XX^\top](w^* - w) \leq s^2(w - w^*)^2 \leq 4s^2r^2. &(68)\end{aligned}$$

For some constant $M > 0$, define $x_1 := Mr\sigma\left(\alpha/\sqrt{\varepsilon} + \sqrt{\alpha^2 + \beta}\right)\sqrt{\varepsilon}$. Let A_1 be the event of {Assumption D.2 holds}. Let A_2 be the event of {SEVER removes less than $(1 + 1/\sqrt{\varepsilon})\varepsilon n$ points}. Let $A_3(p)$ be the event of $\left\{\bar{f}(w) - \bar{f}(w^*) > Mr\sigma\left(\alpha + \sqrt{\alpha^2 + \beta}\sqrt{\varepsilon/p}\right)\right\}$. Then, $\forall 0 \leq p < \sqrt{\varepsilon}$

$$P(A_2, A_3(p) \mid A_1) = 0. \tag{69}$$

By Corollary D.1, $\forall\sqrt{\varepsilon} \leq p \leq 1$

$$P(A_2, A_3(p) \mid A_1) \leq p. \tag{70}$$

By Proposition D.1,

$$P(A_1) \geq 1 - \tau. \tag{71}$$

By Lemma D.3,

$$P(A_2 \mid A_1) \geq 1 - \sqrt{\varepsilon}, \tag{72}$$

and thus

$$1 - P(A_1, A_2) = 1 - P(A_2 \mid A_1)P(A_1) \leq \tau + \sqrt{\varepsilon}. \tag{73}$$

Then, we have:

$$P\left(\bar{f}(w) - \bar{f}(w^*) > x_1/\sqrt{p} \mid A_1, A_2\right) \tag{74}$$

Algorithm 5 SEVER$(f_{1:n}, \mathcal{L}, \sigma)$

```
Input: Sample functions $f_1, \dots, f_n : \mathcal{H} \to \mathbb{R}$, bounded below on a closed domain $\mathcal{H}$, $\gamma$-approximate learner $\mathcal{L}$, and parameter $\sigma \in \mathbb{R}_+$.
Initialize $S \leftarrow \{1, \dots, n\}$.
repeat
    $w \leftarrow \mathcal{L}(\{f_i\}_{i \in S})$. ▷ Run approximate learner on points in $S$.
    Let $\widehat{\nabla} = \frac{1}{|S|} \sum_{i \in S} \nabla f_i(w)$.
    Let $G = [\nabla f_i(w) - \widehat{\nabla}]_{i \in S}$ be the $|S| \times d$ matrix of centered gradients.
    Let $v$ be the top right singular vector of $G$.
    Compute the vector $\tau$ of *outlier scores* defined via $\tau_i = \left((\nabla f_i(w) - \widehat{\nabla}) \cdot v\right)^2$.
    $S' \leftarrow S$
    $S \leftarrow$ FILTER$(S', \tau, \sigma)$ ▷ Remove some $i$'s with the largest scores $\tau_i$ from $S$; see Algorithm 6.
until $S = S'$.
Return $w$.
```

$$\le P\left(A_3(p) \mid A_1, A_2\right) = P(A_2, A_3(p) \mid A_1)/P(A_2 \mid A_1) \tag{75}$$

$$\le \begin{cases} 0 & 0 \le p < \sqrt{\varepsilon} \\ \frac{p}{1-\sqrt{\varepsilon}} & \sqrt{\varepsilon} \le p \le 1 \end{cases}. \tag{76}$$

Let $x = x_1/\sqrt{p}$, we have:

$$P\left(\bar{f}(w) - \bar{f}(w^*) > x \middle| A_1, A_2\right) \le \begin{cases} 0 & x \ge x_1 \varepsilon^{-1/4} \\ \frac{1}{1-\sqrt{\varepsilon}} \frac{x_1^2}{x^2} & x_1 \le x < x_1 \varepsilon^{-1/4} \\ 1 & 0 \le x < x_1 \end{cases}. \tag{77}$$

By Proposition D.1 and law of total expectation, we can bound the expected optimality gap by:

$$\begin{aligned}
\mathbb{E}\left[\bar{f}(w) - \bar{f}(w^*)\right] &\le \mathbb{E}\left[\bar{f}(w) - \bar{f}(w^*) \middle| A_1, A_2\right] P(A_1, A_2) + 4s^2r^2(1 - P(A_1, A_2)) && (78)\\
&\le \int_0^\infty P\left(\bar{f}(w) - \bar{f}(w^*) > x \middle| A_1, A_2\right) dx + 4s^2r^2(\tau + \sqrt{\varepsilon}) && (79)\\
&= \int_0^{x_1} 1 dx + \frac{1}{1-\sqrt{\varepsilon}} \int_{x_1}^{x_1\varepsilon^{-1/4}} \frac{x_1^2}{x^2} dx + 4s^2r^2(\tau + \sqrt{\varepsilon}) && (80)\\
&\le 2x_1 + 4s^2r^2(\tau + \sqrt{\varepsilon}) && (81)\\
&= 2Mr\sigma\left(\alpha/\sqrt{\varepsilon} + \sqrt{\alpha^2 + \beta}\right)\sqrt{\varepsilon} + 4s^2r^2(\tau + \sqrt{\varepsilon}) && (82)\\
&= O\left(\left(sr\sqrt{\xi} + s^2r^2\right)\left(\tau + \sqrt{(d\log d)/n} + \sqrt{\varepsilon} + \sqrt{\log(1/\tau)/n}\right)\right) && (83)
\end{aligned}$$

Note that the expectation above is over both the randomness of SEVER and (X_i, Y_i) pairs. ■

D.3. Proof of Theorem D.2

In this proof, we mainly follow the steps in (Diakonikolas et al., 2019) but use our notion of stability in Assumption D.2. We also allow the success probability to vary, so that we can give an expected error bound later on.

We first restate the SEVER algorithm in Algorithm 5 and Algorithm 6. Throughout this proof we let I_{good} be as in Assumption D.2. We require the following three lemmas. Roughly speaking, the first states that with high probability, we will not remove too many points throughtout the process, the second states that on average, we remove more corrupted points than uncorrupted points, and the third states that at termination, and if we have not removed too many points, then we have reached a point at which the empirical gradient is close to the true gradient. Formally:

Algorithm 6 FILTER(S, τ, σ)

```
Input: Set $S \subseteq [n]$, vector $\tau$ of outlier scores, and parameter $\sigma \in \mathbb{R}_+$.
If $\frac{1}{|S|}\sum_{i\in S}\tau_i \le c_0 \cdot \sigma^2$, for some constant $c_0 > 1$, return $S$ ▷ We only filter out points if the variance is larger than an appropriately chosen threshold.
Draw $T$ from the uniform distribution on $[0, \max_i \tau_i]$.
Return $\{i \in S : \tau_i < T\}$.
```

Lemma D.2. *If the samples satisfy Assumption D.2, $|S| \ge c_1 n$, and the filtering threshold is at least*

$$\frac{2(1-\varepsilon)\sigma^2}{c_1 - 2\varepsilon}\left(\alpha(\varepsilon, n, \tau)^2 + \beta(\varepsilon, n, \tau)\right) \tag{84}$$

then if S' is the output of FILTER(S, τ, σ)*, we have that*

$$\mathbb{E}[|I_{\text{good}} \cap (S\backslash S')|] \le \mathbb{E}[|([n]\backslash I_{\text{good}}) \cap (S\backslash S')|]. \tag{85}$$

Lemma D.3 (Revised version of Lemma 6 in (Diakonikolas et al., 2019))**.** *Assume filtering threshold is $4(\alpha(\varepsilon, n, \tau)^2 + \beta(\varepsilon, n, \tau))\sigma^2$, $\varepsilon \le 1/16$, then we have that for any given $p \ge \sqrt{\varepsilon}$, with probability at least $1-p$, $n - |S| \le (1+1/p)\varepsilon n$ when the filtering algorithm terminates.*

Lemma D.4. *If the samples satisfy Assumption D.2,* FILTER$(S, \tau, \sigma) = S$*, and $n - |S| \le (1+1/p)\varepsilon n$, for $p \ge \sqrt{\varepsilon}$, then*

$$\left\|\nabla \bar{f}(w) - \frac{1}{|I_{\text{good}}|}\sum_{i\in S}\nabla f_i(w)\right\|_2 \le O\left(\sigma\left(\alpha(\varepsilon, n, \tau) + \sqrt{\alpha(\varepsilon, n, \tau)^2 + \beta(\varepsilon, n, \tau)}\sqrt{\varepsilon/p}\right)\right) \tag{86}$$

Before we prove these lemmata, we show how together they imply Theorem D.2.

Proof of Theorem D.2 assuming Lemma D.3 and Lemma D.4. First, we note that the algorithm must terminate in at most n iterations. This is easy to see as each iteration of the main loop except for the last must decrease the size of S by at least 1.

It thus suffices to prove correctness. Note that Lemma D.3 says that with probability at least $1-p$, SEVER will not remove too many points, this will allow us to apply Lemma D.4 to complete the proof, using the fact that w is a critical point of $\frac{1}{|I_{\text{good}}|}\sum_{i\in S}\nabla f_i(w)$. ■

Thus it suffices to prove these three lemmata.

Proof of Lemma D.2. Let $S_{\text{good}} = S \cap I_{\text{good}}$ and $S_{\text{bad}} = S\backslash I_{\text{good}}$. We wish to show that the expected number of elements thrown out of S_{bad} is at least the expected number thrown out of S_{good}. We note that our result holds trivially if FILTER$(S, \tau, \sigma) = S$. Thus, we can assume that $\mathbb{E}_{i\in S}[\tau_i] \ge \frac{2(1-\varepsilon)\sigma^2}{c_1 - 2\varepsilon}\left(\alpha(\varepsilon, n, \tau)^2 + \beta(\varepsilon, n, \tau)\right)$.

It is easy to see that the expected number of elements thrown out of S_{bad} is proportional to $\sum_{i\in S_{\text{bad}}}\tau_i$, while the number removed from S_{good} is proportional to $\sum_{i\in S_{\text{good}}}\tau_i$ (with the same proportionality). Hence, it suffices to show that $\sum_{i\in S_{\text{bad}}}\tau_i \ge \sum_{i\in S_{\text{good}}}\tau_i$.

We first note that since $\text{Cov}_{i\in I_{\text{good}}}[\nabla f_i(w)] \preceq \sigma^2 I$, we have that

$$\begin{aligned}
\text{Cov}_{i\in S_{\text{good}}}[v\cdot\nabla f_i(w)] &\le \frac{1-\varepsilon}{c_1-\varepsilon}\,\text{Cov}_{i\in I_{\text{good}}}[v\cdot\nabla f_i(w)] \quad \left(\text{since } |S_{\text{good}}| \ge \tfrac{c_1-\varepsilon}{1-\varepsilon}|I_{\text{good}}|\right) && (87)\\
&= \frac{1-\varepsilon}{c_1-\varepsilon}\left(\frac{1}{|I_{\text{good}}|}\sum_{i\in I_{\text{good}}}(v\cdot(\nabla f_i(w) - \bar{f}(w)))^2 - (\bar{f}(w) - \mathbb{E}_{i\in I_{\text{good}}}[v\cdot\nabla f_i(w)])^2\right) && (88)\\
&\le \frac{(1-\varepsilon)\sigma^2}{c_1-\varepsilon}\left(\alpha(\varepsilon, n, \tau)^2 + \beta(\varepsilon, n, \tau)\right) \quad (\text{By Assumption D.2}), && (89)
\end{aligned}$$

Let $\mu_{\text{good}} = \mathbb{E}_{i \in S_{\text{good}}}[v \cdot \nabla f_i(w)]$ and $\mu = \mathbb{E}_{i \in S}[v \cdot \nabla f_i(w)]$. Note that

$$\mathbb{E}_{i \in S_{\text{good}}}[\tau_i] = \operatorname{Cov}_{i \in S_{\text{good}}}[v \cdot \nabla f_i(w)] + (\mu - \mu_{\text{good}})^2 \leq \frac{(1-\varepsilon)\sigma^2}{c_1 - \varepsilon} \left(\alpha(\varepsilon, n, \tau)^2 + \beta(\varepsilon, n, \tau)\right) + (\mu - \mu_{\text{good}})^2 \,. \tag{90}$$

We now split into two cases.

Firstly, if

$$(\mu - \mu_{\text{good}})^2 \geq \frac{\varepsilon}{c_1 - 2\varepsilon} \frac{(1-\varepsilon)\sigma^2}{c_1 - \varepsilon} \left(\alpha(\varepsilon, n, \tau)^2 + \beta(\varepsilon, n, \tau)\right), \tag{91}$$

we let $\mu_{\text{bad}} = \mathbb{E}_{i \in S_{\text{bad}}}[v \cdot \nabla f_i(w)]$, and note that $|\mu - \mu_{\text{bad}}||S_{\text{bad}}| = |\mu - \mu_{\text{good}}||S_{\text{good}}|$. We then have that

$$\begin{aligned}
\mathbb{E}_{i \in S_{\text{bad}}}[\tau_i] &= \operatorname{Cov}_{i \in S_{\text{bad}}}[v \cdot \nabla f_i(w)] + (\mu - \mu_{\text{bad}})^2 \geq (\mu - \mu_{\text{bad}})^2 && (92)\\
&= (\mu - \mu_{\text{good}})^2 \left(\frac{|S_{\text{good}}|}{|S_{\text{bad}}|}\right)^2 && (93)\\
&\geq \frac{|S_{\text{good}}|}{|S_{\text{bad}}|} \frac{c_1 - \varepsilon}{\varepsilon} (\mu - \mu_{\text{good}})^2 \quad \text{(because } |S_{\text{good}}| \geq (c_1 - \varepsilon)n \text{ and } |S_{\text{bad}}| \leq \varepsilon n\text{)} && (94)\\
&= \frac{|S_{\text{good}}|}{|S_{\text{bad}}|} \left(\frac{c_1 - 2\varepsilon}{\varepsilon} (\mu - \mu_{\text{good}})^2 + (\mu - \mu_{\text{good}})^2\right) && (95)\\
&\geq \frac{|S_{\text{good}}|}{|S_{\text{bad}}|} \left(\frac{(1-\varepsilon)\sigma^2}{c_1 - \varepsilon} \left(\alpha(\varepsilon, n, \tau)^2 + \beta(\varepsilon, n, \tau)\right) + (\mu - \mu_{\text{good}})^2\right) \quad \text{(by (91))} && (96)\\
&\geq \frac{|S_{\text{good}}|}{|S_{\text{bad}}|} \mathbb{E}_{i \in S_{\text{good}}}[\tau_i] \quad \text{(by (90))}. && (97)
\end{aligned}$$

Hence, $\sum_{i \in S_{\text{bad}}} \tau_i \geq \sum_{i \in S_{\text{good}}} \tau_i$.

On the other hand, if $(\mu - \mu_{\text{good}})^2 \leq \frac{\varepsilon}{c_1 - 2\varepsilon} \frac{(1-\varepsilon)\sigma^2}{c_1 - \varepsilon} \left(\alpha(\varepsilon, n, \tau)^2 + \beta(\varepsilon, n, \tau)\right)$, then $\mathbb{E}_{i \in S_{\text{good}}}[\tau_i] \leq \left(1 + \frac{\varepsilon}{c - 2\varepsilon}\right) \frac{(1-\varepsilon)\sigma^2}{c_1 - \varepsilon} \left(\alpha(\varepsilon, n, \tau)^2 + \beta(\varepsilon, n, \tau)\right) \leq \mathbb{E}_{i \in S}[\tau_i]/2$. Therefore $\sum_{i \in S_{\text{bad}}} \tau_i \geq \sum_{i \in S_{\text{good}}} \tau_i$ once again. This completes our proof. ■

Proof of Lemma D.3. Define the event

$$A = \{n - |S| \leq (1 + 1/p)\varepsilon n\}, \tag{98}$$

and we want to lower-bound $P(A)$. Given that $\varepsilon \leq 1/16$, the threshold is $4(\alpha(\varepsilon, n, \tau)^2 + \beta(\varepsilon, n, \tau))\sigma^2$ and $p \geq \sqrt{\varepsilon}$, and conditioned on the event A, it can be verified that the asusumption of Lemma D.2 is satisfied. In particular, simple calculation shows that given $c_1 = 1 - (1 + 1/p)\varepsilon$, $\varepsilon \leq 1/16$, $p \geq \sqrt{\varepsilon}$, we have

$$4\sigma^2 \geq \frac{2(1-\varepsilon)\sigma^2}{c_1 - 2\varepsilon} \tag{99}$$

And Lemma D.2 implies that $|([n] \backslash I_{\text{good}}) \cap S| + |I_{\text{good}} \backslash S|$ is a supermartingale. Since its initial size is at most εn, with probability at least $1 - p$, it never exceeds $\varepsilon n / p$, and therefore at the end of the algorithm, we must have that $n - |S| \leq \varepsilon n + |I_{\text{good}} \backslash S| \leq (1 + 1/p)\varepsilon n$. ■

We now prove Lemma D.4.

Proof of Lemma D.4. We note that

$$\left\| \sum_{i \in S} (\nabla f_i(w) - \nabla \bar{f}(w)) \right\|_2 \tag{100}$$

$$\leq \left\| \sum_{i \in I_{\text{good}}} (\nabla f_i(w) - \nabla \bar{f}(w)) \right\|_2 + \left\| \sum_{i \in (I_{\text{good}} \backslash S)} (\nabla f_i(w) - \nabla \bar{f}(w)) \right\|_2 + \left\| \sum_{i \in (S \backslash I_{\text{good}})} (\nabla f_i(w) - \nabla \bar{f}(w)) \right\|_2 \tag{101}$$

$$\leq \left\| \sum_{i \in (I_{\text{good}} \backslash S)} (\nabla f_i(w) - \nabla \bar{f}(w)) \right\|_2 + \left\| \sum_{i \in (S \backslash I_{\text{good}})} (\nabla f_i(w) - \nabla \bar{f}(w)) \right\|_2 + n\sigma\alpha(\varepsilon, n, \tau). \tag{102}$$

First we analyze

$$\left\| \sum_{i \in (I_{\text{good}} \backslash S)} (\nabla f_i(w) - \nabla \bar{f}(w)) \right\|_2 . \tag{103}$$

This is the supremum over unit vectors v of

$$\sum_{i \in (I_{\text{good}} \backslash S)} v \cdot (\nabla f_i(w) - \nabla \bar{f}(w)). \tag{104}$$

However, we note that

$$\sum_{i \in I_{\text{good}}} (v \cdot (\nabla f_i(w) - \nabla \bar{f}(w)))^2 \leq n\sigma^2 \beta(\varepsilon, n, \tau). \tag{105}$$

Since $|I_{\text{good}} \backslash S| \leq (1 + 1/p)\varepsilon n$, we have by Cauchy-Schwarz that

$$\sum_{i \in (I_{\text{good}} \backslash S)} v \cdot (\nabla f_i(w) - \nabla \bar{f}(w)) = \sqrt{(n\sigma^2 \beta(\varepsilon, n, \tau))((1 + 1/p)\varepsilon n)} = n\sigma \sqrt{\beta(\varepsilon, n, \tau)(1 + 1/p)\varepsilon}, \tag{106}$$

as desired.

Let

$$\Delta := \left\| \sum_{i \in S} (\nabla f_i(w) - \nabla \bar{f}(w)) \right\|_2 . \tag{107}$$

Because our Filter algorithm terminates with $n - |S| \leq (1+1/p)\varepsilon n$, and the stopping condition is set as $\| \frac{1}{|S|} \sum_{i \in S} (\nabla f_i(w) - \nabla \hat{f}(w))(\nabla f_i(w) - \nabla \hat{f}(w))^\top \| \leq 4(\alpha(\varepsilon, n, \tau)^2 + \beta(\varepsilon, n, \tau))\sigma^2$, we note that since for any such v that

$$\sum_{i \in S} (v \cdot (\nabla f_i(w) - \nabla \bar{f}(w)))^2 = \sum_{i \in S} (v \cdot (\nabla f_i(w) - \nabla \hat{f}(w)))^2 + |S|(v \cdot (\nabla \hat{f}(w) - \nabla \bar{f}(w)))^2 \tag{108}$$

$$\leq \sum_{i \in S} (v \cdot (\nabla f_i(w) - \nabla \hat{f}(w)))^2 + \Delta^2/|S| \leq n4(\alpha(\varepsilon, n, \tau)^2 + \beta(\varepsilon, n, \tau))\sigma^2 + \Delta^2/((1 - (1 + 1/p)\varepsilon)n) \tag{109}$$

and since $|S \backslash I_{\text{good}}| \leq (1 + 1/p)\varepsilon n$, and so we have similarly that

$$\left\| \sum_{i \in (S \backslash I_{\text{good}})} \nabla f_i(w) - \nabla \bar{f}(w) \right\|_2 \leq 2n\sigma \sqrt{\alpha(\varepsilon, n, \tau)^2 + \beta(\varepsilon, n, \tau)} \sqrt{(1 + 1/p)\varepsilon} + \Delta \sqrt{\frac{(1 + 1/p)\varepsilon}{1 - (1 + 1/p)\varepsilon}}. \tag{110}$$

Combining with the above we have that

$$\frac{\Delta}{n} \leq \sigma\alpha(\varepsilon, n, \tau) + \sigma \sqrt{\beta(\varepsilon, n, \tau)(1 + 1/p)\varepsilon} + 2\sigma \sqrt{\alpha(\varepsilon, n, \tau)^2 + \beta(\varepsilon, n, \tau)} \sqrt{(1 + 1/p)\varepsilon} + \frac{\Delta}{n} \sqrt{\frac{(1 + 1/p)\varepsilon}{1 - (1 + 1/p)\varepsilon}}, \tag{111}$$

Thus

$$\frac{\Delta}{n} \leq \frac{1}{1 - \sqrt{\frac{(1+1/p)\varepsilon}{1-(1+1/p)\varepsilon}}} \left(\sigma\alpha(\varepsilon, n, \tau) + 6\sigma \sqrt{\alpha(\varepsilon, n, \tau)^2 + \beta(\varepsilon, n, \tau)} \sqrt{\varepsilon/p} \right) \tag{112}$$

and therefore, $\frac{\Delta}{n} = O\left(\sigma \left(\alpha(\varepsilon, n, \tau) + \sqrt{\alpha(\varepsilon, n, \tau)^2 + \beta(\varepsilon, n, \tau)} \sqrt{\varepsilon/p} \right) \right)$ as desired. ■

E. Proofs for Section 5

Lemma E.1 (Lemma 5.1)**.** *Suppose the adversarial rewards are unbounded, and in a particular iteration t, the adversarial contaminate $\varepsilon^{(t)}$ fraction of the episodes, then given M episodes, it is guaranteed that if $\varepsilon^{(t)} \leq c$, for some absolute constant c, and any constant $\tau \in [0,1]$, we have*

$$\begin{aligned}&\mathbb{E}\left[\mathbb{E}_{s,a\sim d^{(t)}}\left[\left(Q^{\pi^{(t)}}(s,a) - \phi(s,a)^\top w^{(t)}\right)^2\right]\right] \\ &\leq O\left(\left(W^2 + \frac{\sigma W}{1-\gamma}\right)\left(\sqrt{\varepsilon^{(t)}} + f(d,\tau)M^{-\frac{1}{2}} + \tau\right)\right).\end{aligned} \tag{113}$$

where $f(d,\tau) = \sqrt{d\log d} + \sqrt{\log(1/\tau)}$.

Proof of Lemma E.1. The proof of Lemma 5.1 follows by instantiating Theorem D.3 to our specific linear regression problem instance. To specify the constants in Theorem D.3, we make the following observations

1. By Lemma B.1, we have that $\xi = \frac{1}{(1-\gamma)^2} + \frac{\sigma^2}{1-\gamma}$.
2. Since $\|X\| \leq 1$, $\mathbb{E}_{X\sim D_x}\left[XX^\top\right] \leq I$, and thus $s = 1$.
3. $\max_{\|v\|=1}\mathbb{E}\left[(v^\top X)^4\right] \leq \mathbb{E}\left[\|v\|^4\|X\|^4\right] \leq 1$, thus $C = 1$.

Plugging in the above instantiation to Theorem D.3 concludes the proof. ■

Theorem E.1 (Theorem 5.1)**.** *Under assumptions 3.1 and 3.2, given a desired optimality gap α, there exists a set of hyperparameters agnostic to the contamination level ε, such that Algorithm 2, using Algorithm 3 as the linear regression solver, guarantees with a poly$(1/\alpha, 1/(1-\gamma), |\mathcal{A}|, W, \sigma, \kappa)$ sample complexity that under ε-contamination, we have*

$$\begin{aligned}&\mathbb{E}\left[V^*(\mu_0) - V^{\hat{\pi}}(\mu_0)\right] \\ &\leq \tilde{O}\left(\max\left[\alpha, \sqrt{\frac{|\mathcal{A}|\kappa\left(W^2+\sigma W\right)}{(1-\gamma)^4}}\varepsilon^{1/4}\right]\right).\end{aligned} \tag{114}$$

where $\hat{\pi}$ is the uniform mixture of $\pi^{(1)}$ through $\pi^{(T)}$.

Proof of Theorem E.1. Denote $z := 2W$ and again $\varepsilon_{stat} \leq (2W)^2 = z^2$. Denote $\left(W^2 + \frac{\sigma W}{1-\gamma}\right) = b$. Notice that Lemma 5.1 only holds when $\varepsilon^{(t)} \leq c$ for some absolute constant c, and there are at most $\varepsilon T/c$ iterations in which $\varepsilon^{(t)} > c$, which incurs at most $\varepsilon_{stat} \leq z^2$ error. Given this observation we can now plugging Lemma 5.1 into Lemma 4.1, and we get

$$\mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T}\{V^*(\mu_0) - V^{(t)}(\mu_0)\}\right] \tag{115}$$

$$\leq \frac{W}{1-\gamma}\sqrt{\frac{2\log|\mathcal{A}|}{T}} + \frac{1}{T}\sum_{t=1}^{T}\sqrt{\frac{4|\mathcal{A}|\kappa\varepsilon_{stat}^{(t)}}{(1-\gamma)^3}} \tag{116}$$

$$\leq \frac{W}{1-\gamma}\sqrt{\frac{2\log|\mathcal{A}|}{T}} + \frac{z^2}{c}\varepsilon + \frac{1}{T}\sum_{t=1}^{T}\sqrt{\frac{4|\mathcal{A}|\kappa b\left(\sqrt{\varepsilon^{(t)}} + \sqrt{(d\log d)/M} + \sqrt{\log(1/\tau)/M} + \tau\right)}{(1-\gamma)^3}} \tag{117}$$

$$\leq \frac{W}{1-\gamma}\sqrt{\frac{2\log|\mathcal{A}|}{T}} + \frac{z^2}{c}\varepsilon + \sqrt{\frac{4|\mathcal{A}|\kappa b\left(\sqrt{(d\log d)/M} + \sqrt{\log(1/\tau)/M} + \tau\right)}{(1-\gamma)^3}} + \frac{1}{T}\sum_{t=1}^{T}\sqrt{\frac{4|\mathcal{A}|\kappa b\sqrt{\varepsilon^{(t)}}}{(1-\gamma)^3}} \tag{118}$$

$$\leq \frac{W}{1-\gamma}\sqrt{\frac{2\log|\mathcal{A}|}{T}} + \frac{z^2}{c}\varepsilon + \sqrt{\frac{4|\mathcal{A}|\kappa b\left(\sqrt{(d\log d)/M} + \sqrt{\log(1/\tau)/M} + \tau\right)}{(1-\gamma)^3}} + \sqrt{\frac{4|\mathcal{A}|\kappa b}{(1-\gamma)^3}}\varepsilon^{1/4} \tag{119}$$

Algorithm 7 Robust NPG Update

```
Input $\rho^n_{\text{cov}}$, $b^n$, learning rate $\eta$, sample size $M$ for critic fitting, iterations $T$
Define $\mathcal{K}^n = \{s : \forall a \in \mathcal{A}, b^n(s,a) = 0\}$
Initialize policy $\pi^0 : \mathcal{S} \to \Delta(\mathcal{A})$, such that
    $\pi^0(\cdot|s) = \begin{cases} \text{Uniform}(\mathcal{A}) & s \in \mathcal{K}^n \\ \text{Uniform}(\{a \in \mathcal{A} : b^n(s,a) > 0\}) & s \notin \mathcal{K}^n. \end{cases}$
for $t = 0 \to T-1$ do
    Draw $M$ i.i.d samples $\left\{s_i, a_i, \widehat{Q}^{\pi^t}(s_i, a_i; r + b^n)\right\}_{i=1}^{M}$ with $s_i, a_i \sim \rho^n_{\text{cov}}$ (see Alg 1)
    Critic fit: Call Algorithm 3 to solve for the robust linear regression problem
        $\theta^t = \arg\min_{\|\theta\| \le W} \sum_{i=1}^{M} \left(\theta \cdot \phi(s_i, a_i) - \left(\widehat{Q}^{\pi^t}(s_i, a_i; r + b^n) - b^n(s_i, a_i)\right)\right)^2$
    Actor update
        $\pi^{t+1}(\cdot|s) \propto \pi^t(\cdot|s) \exp\left(\eta\left(b^n(s,\cdot) + \theta^t \cdot \phi(s,\cdot)\right) \mathbf{1}\{s \in \mathcal{K}^n\}\right)$   (125)
return $\pi^n := \text{Uniform}\{\pi^0, ..., \pi^{T-1}\}$.
```

where the last two steps are by Cauchy Schwarz and the fact that the attacker only has ε budget to distribute, which implies that $\sum_{t=1}^{T} \varepsilon^{(t)} = T\varepsilon$. Setting

$$T = \frac{2W^2 \log|\mathcal{A}|}{\alpha^2(1-\gamma)^2} \tag{120}$$

$$\tau = \frac{\alpha^2(1-\gamma)^3}{4|\mathcal{A}|b\kappa} \tag{121}$$

$$M = \frac{16|\mathcal{A}|^2 b^2 \kappa^2}{\alpha^4(1-\gamma)^6} \max\left[d \log d, \log(1/\tau)\right] \tag{122}$$

we get

$$\mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T}\{V^*(\mu_0) - V^{(t)}(\mu_0)\}\right] \le O\left(\alpha + \sqrt{\frac{|\mathcal{A}|\kappa b}{(1-\gamma)^3}}\varepsilon^{1/4}\right). \tag{123}$$

with sample complexity

$$TM = \frac{32W^2|\mathcal{A}|^2 \log|\mathcal{A}| b^2 \kappa^2}{\alpha^6(1-\gamma)^8} \max\left[d \log d, \log(1/\tau)\right]. \tag{124}$$

■

F. Proof of Theorem 6.1

In PC-PG, aside from the robust linear regression step in Algorithm 7, in step 4 of Algorithm 4, we also needs to robustly estimate the covariance matrix under ε-contamination. Luckily, by assumption, $\phi(s,a)$ is bounded, and thus the current empirical mean estimation is already robust to adversarial contamination:

Lemma F.1 (Robust variant of Lemma G.3 in (Agarwal et al., 2020a)). *Given $\nu \in \Delta(S \times A)$ and K ε-contaminated samples from ν. Denote $\Sigma = \mathbb{E}_{(s,a)\sim\nu}\left[\phi(s,a)\phi(s,a)^\top\right]$. Then, with probability at least $1-\delta$, we have that under ε-corruption*

$$\max_{\|x\| \le 1}\left|x^\top\left(\sum_{i=1}^{K}\phi(s_i,a_i)\phi(s_i,a_i)^\top / K - \Sigma\right)x\right| \le \sqrt{\frac{8\log(8d/\delta)}{K}} + 2\varepsilon. \tag{126}$$

Proof. Without contamination, Lemma G.3 in (Agarwal et al., 2020a) shows that

$$\max_{\|x\|\le 1}\left|x^\top\left(\sum_{i=1}^{K}\phi(s_i,a_i)\phi(s_i,a_i)^\top/K-\Sigma\right)x\right|\le\frac{2\log(8d/\delta)}{3K}+\sqrt{\frac{2\log(8d/\delta)}{K}}. \tag{127}$$

Since both x and $\phi(s,a)$ has norm bounded by 1, the ε fraction of contaminated samples can only bias the estimate by at most 2ε, i.e. with ε-contamination

$$\max_{\|x\|\le 1}\left|x^\top\left(\sum_{i=1}^{K}\phi(s_i,a_i)\phi(s_i,a_i)^\top/K-\Sigma\right)x\right|\le\sqrt{\frac{8\log(8d/\delta)}{K}}+2\varepsilon. \tag{128}$$

■

Lemma F.2 (Lemma G.4 in (Agarwal et al., 2020a))**.** *Denote* $\eta(K)=\sqrt{\frac{8\log(8d/\delta)}{K}}+2\varepsilon$*. Then, under* ε*-contamination,* $\phi(s,a)^\top(\Sigma^n_{cov})^{-1}\phi(s,a)\le\beta$ *is guaranteed with probability* $1-\delta$*, if* $N\eta(K)\le\lambda/2$*.*

Lemma F.3 (variant of Lemma C.2 in (Agarwal et al., 2020a))**.** *Assuming that for all iterations* n *but* m *of them, we have* $\phi(s,a)^\top(\Sigma^n_{cov})^{-1}\phi(s,a)\le\beta$ *for* $(s,a)\in\mathcal{K}^n$*, then*

$$V^*-V^{\hat\pi}\le\frac{1}{1-\gamma}\left(2W\sqrt{\frac{\log A}{T}}+2\sqrt{\beta\lambda W^2}+\frac{1}{NT}\sum_{n=0}^{N-1}\sum_{t=0}^{T-1}2\sqrt{\beta N\varepsilon_{stat}^{(n,t)}}+\frac{2I_N(\lambda)}{\beta N}+2Hm\right) \tag{129}$$

Proof. The proof follows exactly the proof of Lemma C.2 in (Agarwal et al., 2020a), except that we note that for iterations in which the assumption is not satisfied, the worst-case loss is bounded:

$$\frac{1}{T}\sum_{t=0}^{T-1}\left(\mathbb{E}_{(s,a)\sim\tilde d_{\mathcal{M}^n}}\left(A^t_{b_n}(s,a)-\hat A^t_{b_n}(s,a)\right)\mathbf{1}\{s\in\mathcal{K}^n\}\right)\le 2H \tag{130}$$

■

Proof of Theorem 6.1. First of all, we need to upper-bound m. The condition in Lemma F.2 is satisfied as long as $2\varepsilon^{(n)}\le\frac{\lambda}{4N}$ and $K\ge\frac{128N^2\log(8\tilde d/\delta)}{\lambda^2}$. Also note that $\sum_{n=0}^{N-1}\varepsilon^{(n)}\le N\varepsilon$, and thus m is at most $\frac{8N^2\varepsilon}{\lambda}$.

Also, by Lemma C.1,

$$\varepsilon_{stat}^{(n,t)}\le 4\left(W^2+WH\right)\left(\varepsilon^{(n,t)}+\sqrt{\frac{8}{M}\log\frac{4d}{\delta}}\right). \tag{131}$$

Plugging both into Lemma F.3, we get

$$V^*-V^{\hat\pi}\le\frac{1}{1-\gamma}\left(2W\sqrt{\frac{\log A}{T}}+2\sqrt{\beta\lambda W^2}+2\sqrt{4\left(W^2+WH\right)\beta N\left(\varepsilon+\sqrt{\frac{8}{M}\log\frac{4d}{\delta}}\right)}+\frac{2I_N(\lambda)}{\beta N}+\frac{16HN^2\varepsilon}{\lambda}\right) \tag{132}$$

Let

$$T=\frac{4W^2\log A}{(1-\gamma)^2\alpha^2},\qquad\lambda=1,\qquad\beta=\frac{\alpha^2(1-\gamma)^2}{4W^2},\qquad N=\frac{4W^2d\log(N+1)}{\alpha^3(1-\gamma)^3} \tag{133}$$

$$M=\frac{2d^2\log^2(N+1)(W^2+WH)^2\log(\frac{4d}{\delta})}{\alpha^6(1-\gamma)^6},\qquad K=128N^2\log(8\tilde d/\delta) \tag{134}$$

Then, (132) gives

$$V^*-V^{\hat\pi}\le 4\alpha+\sqrt{\frac{16(W^2WH)d\log(N+1)}{\alpha(1-\gamma)^3}\varepsilon}+\frac{256HW^4d^2\log^2(N+1)}{\alpha^6(1-\gamma)^6}\varepsilon \tag{135}$$

Algorithm 8 FPG-TRPO

```
Input: initial policy parameter $\theta_0$; initial value function parameter $\phi_0$.
Hyperparameters: KL-divergence limit $\delta$; backtracking coefficient $\alpha$; maximum number of backtracking steps $K$; upper-bound of corruption level $\varepsilon$; episode length $H$; batch size $M$.
for $k = 0, 1, \ldots$ do
    Collect set of $M$ trajectories $D_k = \{\tau_i\}_{1:M}$ by running policy $\pi_k = \pi(\theta_k)$ in the environment.
    Compute rewards-to-go $\hat{R}_{t,i} = \sum_{h=t}^{H} \gamma^{h-t} r_{h,i}$.
    Using GAE to compute advantage estimate $\hat{A}_{t,i}$ based on the current value function $V_{\phi_k}$.
    Compute and save $\hat{g}_{t,i} = \nabla_\theta \log \pi_\theta(a_{t,i}, s_{t,i})|_{\theta_k}$ for all $t = 1:H$ and $i = 1:M$.
    Call the filtered conjugate gradient algorithm in Alg. 9 to get $S_k \subset [M] \times [H], \hat{x}_k = FCG(\hat{g}_{t,i}, \hat{A}_{t,i})$.
    Compute policy gradient estimate $\hat{g}_k = \frac{1}{|S_k|} \sum_{(t,i)\in S_k} \hat{g}_{t,i}\hat{A}_{t,i}$.
    Update the policy by backtracking line search with

        $\theta_{k+1} = \theta_k + \alpha^j \sqrt{\frac{2\delta}{\hat{x}_k \hat{g}_k}} \hat{x}_k$        (138)

    where $j \in \{0, 1, 2, ..., K\}$ is the smallest value which improves the sample loss and satisfies the sample KL-divergence constraint.
    Fit the value function by regression on mean-squared error on the filtered trajectories $S_k$:

        $\phi_{k+1} = \arg\min_{\phi} \frac{1}{|S_k|} \sum_{(t,i)\in S_k} \left(V_\phi(s_{t,i}) - \hat{R}_{t,i}\right)^2$        (139)

    In practice, one often only take a few gradient steps in each iteration $k$, instead of optimizing to convergence.
```

Algorithm 9 Filtered Conjugate Gradient (FCG)

```
Input: $\hat{g}_{t,i}, \hat{A}_{t,i}$
Hyperparameters: Number of iterations $r$ (default $r = 4$), fraction of data filtered in each iteration $p$ (default $p = \varepsilon/2$, i.e. filter out $2\varepsilon$ data in total).
Initialize $S = \{1, 2, \ldots, M\}$.
for $k = 1, \ldots, r$ do
    Call standard CG to solve for $\hat{x} = \hat{F}^{-1}\hat{g}$, where $\hat{F} = \frac{1}{S}\sum_{(t,i)\in S} \hat{g}_{t,i}\hat{g}_{t,i}^\top$ and $\hat{g} = \frac{1}{S}\sum_{(t,i)\in S} \hat{g}_{t,i}\hat{A}_{t,i}$.
    Compute the residues $r_{t,i} = \hat{g}_{t,i}\hat{g}_{t,i}^\top \hat{x} - \hat{g}_{t,i}\hat{A}_{t,i}$ for $(t,i) \in S$ and save in a matrix $G$ of size $d \times |S|$.
    Let $v$ be the top right singular vector of $G$.
    Compute the vector $\tau$ of outlier scores defined via $\tau_{t,i} = \left(r_{t,i}^\top v\right)^2$.
    Remove $(HMp)$ number of $(t,i)$ pair with the largest outlier scores from $S$.
Call standard CG one more time and return $(S, \hat{x})$.
```

Let $\alpha = \varepsilon^{1/7}$, then

$$
\begin{aligned}
V^* - V^{\hat{\pi}} &\leq 4\varepsilon^{1/7} + \sqrt{\frac{16(W^2WH)d\log(N+1)}{(1-\gamma)^3}}\varepsilon^{3/7} + \frac{256HW^4d^2\log^2(N+1)}{(1-\gamma)^6}\varepsilon^{1/7} && (136) \\
&\leq \tilde{O}(d^2\varepsilon^{1/7}) && (137)
\end{aligned}
$$

This concludes the proof. ■

G. Implementation Details of FPG-TRPO

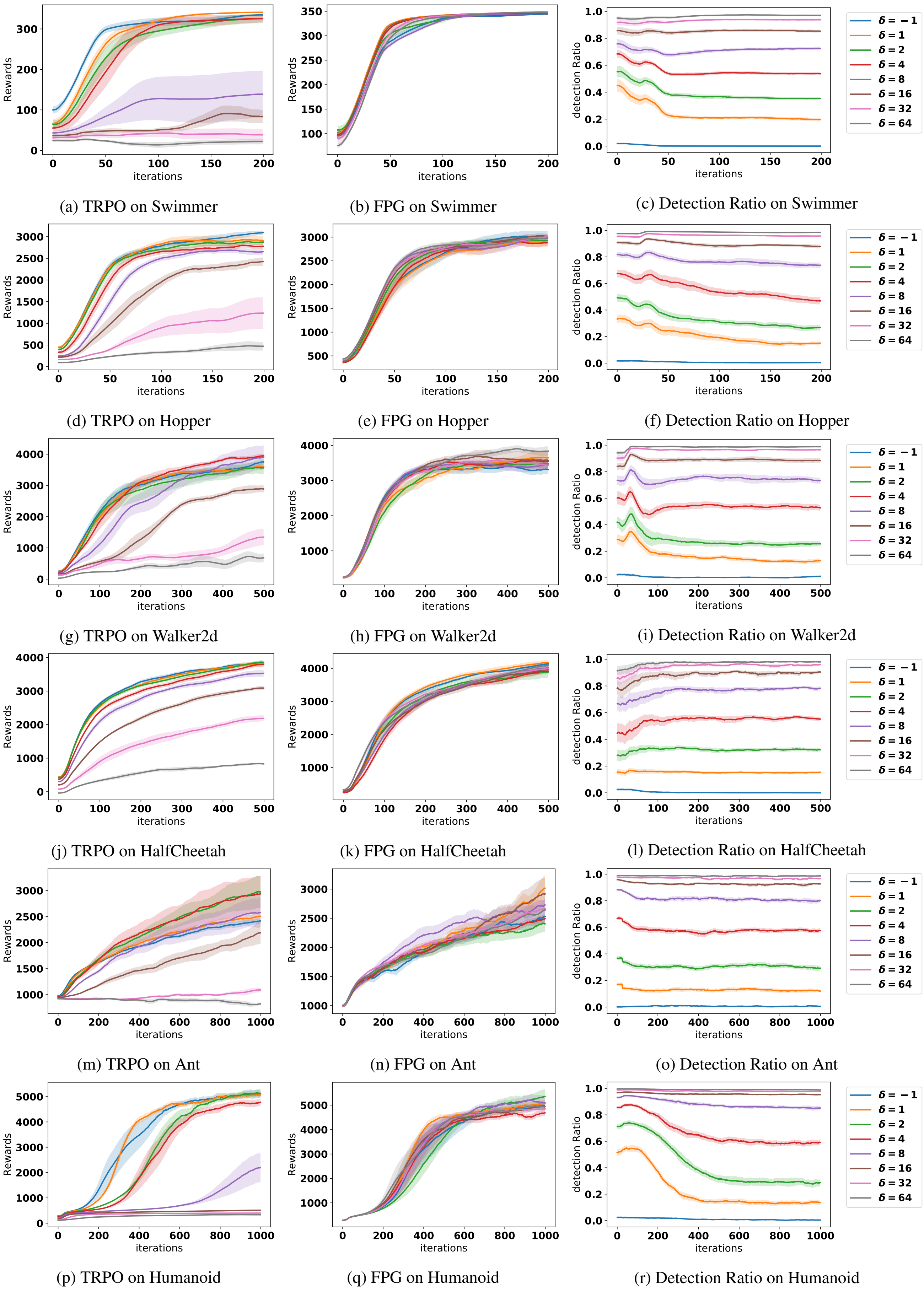
In the experiment, we use a TRPO variant of FPG implementation, which differs from Alg. 2 in several ways:

Parameters	Values	Description
γ	0.995	discounting factor.
λ	0.97	GAE parameter (Schulman et al., 2015b).
l2-reg	0.001	L2 regularization weight in value loss.
δ	0.01	KL constraint in TRPO.
damping	0.1	damping factor in conjugate gradient.
batch-size	25000	number of time steps per policy gradient iteration.
α	0.5	backtracking coefficient.
K	10	maximum number of backtracking steps.

Table 1. Hyperparameters for FPG-TRPO.

1. Most existing TRPO implementation uses the conjugate gradient (CG) method instead of linear regression to solve for the matrix inverse vector product problem. We follow this convention and design FPG-TRPO to use a filtered conjugate gradient (FCG) subroutine to replace the standard CG produce. The FPG procedure is detailed in Alg. 9. At a high level FCG performs a filtering algorithm (a.k.a. outlier removal) on the residues of CG with respect to each data point.
2. Again following existing TRPO implementations, FPG-TRPO builds another network to estimate the value function for the purpose of variance reduction, effectively resulting in an actor-critic algorithm. Instead of performing robust learning procedure on both policy and value function learning, we perform the main filtering algorithm on the policy learning procedure (the CG step discussed above), which also returns a filtered subset of data as a by-product. We then use this filtered subset of data to perform the rest of the learning procedure, including value function update and the sample loss estimation in backtracking line search. This allows us to perform the robust learning procedure only once per PG iteration.
3. FPG-TRPO uses a deterministic variant of the filtering algorithm suggested in (Diakonikolas et al., 2019), which empirically performs better and is simpler to implement than the stochastic variant used for theoretical analysis. Specifically, the filtering algorithm will simply remove a fixed fraction of points with the largest deviation along the top singular value direction (step 9 of Alg. 9).

The pseudo-code of `FPG-TRPO` can be found in Alg. 8. Similar to the NPG variant of FPG, the only difference between Alg. 8 and a standard TRPO implementation is the replacement of the CG subroutine with the FCG subroutine. This modular implementation allows one to easily replace Alg. 9 with any state-of-the-art robust CG procedure in the future. Table 1 lists all the hyper-parameters we used in our experiments, which are taken from open-source implementations of TRPO tuned for the MuJoCo environments. Our code to reproduce the experiment result is included in the supplementary material and will be open-sourced. Finally, Figure 4 presents the detailed results on all experiments, completing the partial results shown in Figure 3.



(a) TRPO on Swimmer

(b) FPG on Swimmer

(c) Detection Ratio on Swimmer

(d) TRPO on Hopper

(e) FPG on Hopper

(f) Detection Ratio on Hopper

(g) TRPO on Walker2d

(h) FPG on Walker2d

(i) Detection Ratio on Walker2d

(j) TRPO on HalfCheetah

(k) FPG on HalfCheetah

(l) Detection Ratio on HalfCheetah

(m) TRPO on Ant

(n) FPG on Ant

(o) Detection Ratio on Ant

(p) TRPO on Humanoid

(q) FPG on Humanoid

(r) Detection Ratio on Humanoid

Figure 4. Detailed Results on the MuJoCo benchmarks.