Robust Policy Gradient against Strong Data Corruption

Xuezhou Zhang [1] **Yiding Chen** [1] **Jerry 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

*Equal contribution [1]Department of Computer Sciences, University of Wisconsin–Madison [2]Cornell University. Correspondence to: Xuezhou Zhang <xzhang784@wisc.edu>.

Proceedings of the 38^{th} International Conference on Machine Learning, PMLR 139, 2021.

the setting of tabular MDPs and cannot be applied to 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

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. Additional discussions on other RL algorithms in standard stochastic MDPs can be found in appendix A.

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.

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: $d_{s'}^\pi(s, a) := (1-\gamma)\sum_{t=0}^{\infty} \gamma^t \mathbb{P}^\pi(s_t = s, a_t = a | s_0 = s')$, 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 $d_{s',a'}^\pi(s, a)$ as: $d_{s',a'}^\pi(s, a) := (1-\gamma)\sum_{t=0}^{\infty} \gamma^t \mathbb{P}^\pi(s_t = s, a_t = s | s_0 = s', a_0 = a')$. For any state-action distribution ν, we write $d_\nu^\pi(s, a) := \sum_{(s',a') \in \mathcal{S} \times \mathcal{A}} \nu(s', a') d_{s',a'}^\pi(s, a)$. 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 $d_\nu^\pi(s) = \sum_a d_\nu^\pi(s, a)$.

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_ν^π-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_ν^π sampler samples (s, a) i.i.d from d_ν^π, 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_ν^π-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*

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

We also assume that the feature is bounded, i.e. $\max_{s,a} \|\phi(s, a)\|_2 \leq 1$, and the reward function has bounded first and second moments, i.e. $\mathbb{E}[r(s, a)] \in [0, 1]$ and $Var(r(s, a)) \leq \sigma^2$ 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 $(r_t^\dagger, s_{t+1}^\dagger) \in \mathbb{R} \times \mathcal{S}$.
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

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

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

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

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*

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

Then, we have that

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

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 $\varepsilon^{(t)}_{stat}$ 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 $\varepsilon^{(t)}_{stat}$ 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 $1-\delta$,*

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

Algorithm 1 d_ν^π sampler and Q^π estimator

function d_ν^π-SAMPLER
  Input: A reset distribution $\nu \in \Delta(\mathcal{S} \times \mathcal{A})$.
  Sample $s_0, a_0 \sim \nu$.
  Execute π 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^π-ESTIMATOR
  Input: current state-action (s, a), a policy π.
  Execute π 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 η; 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 $\theta^{(t+1)} = \theta^{(t)} + \eta w^{(t)}$.

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.

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.

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 $S \leftarrow \{1, \ldots, M\}$, $f_i(w) = \|y_i - w^\top x_i\|^2$.
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 τ 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.

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*

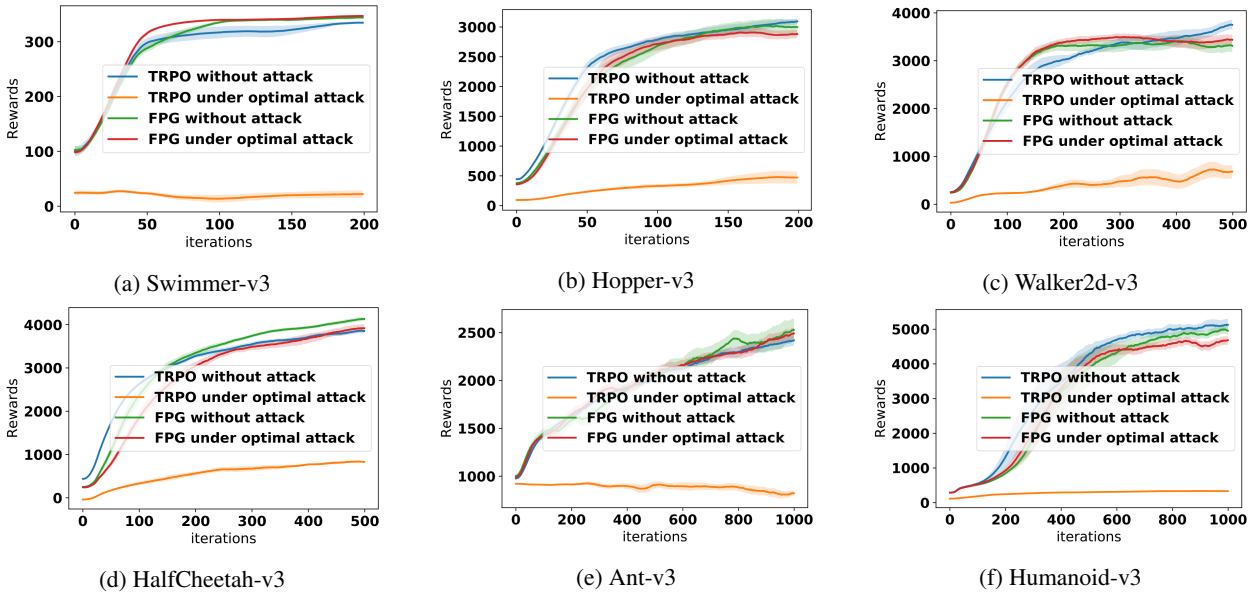
$$\mathbb{E}\left[V^*(\mu_0) - V^{\hat{\pi}}(\mu_0)\right] \le \tilde{O}\left(\max\left[\alpha, \sqrt{\frac{|\mathcal{A}|\kappa\,(W^2 + \sigma W)}{(1-\gamma)^4}} \varepsilon^{1/4} \right] \right). \tag{5}$$

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*

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



(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.

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. 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. 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., 2020), here we describe the poisoning 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:

$$\theta^{(t+1)} = \theta^{(t)} + g^{(t)} \tag{7}$$

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

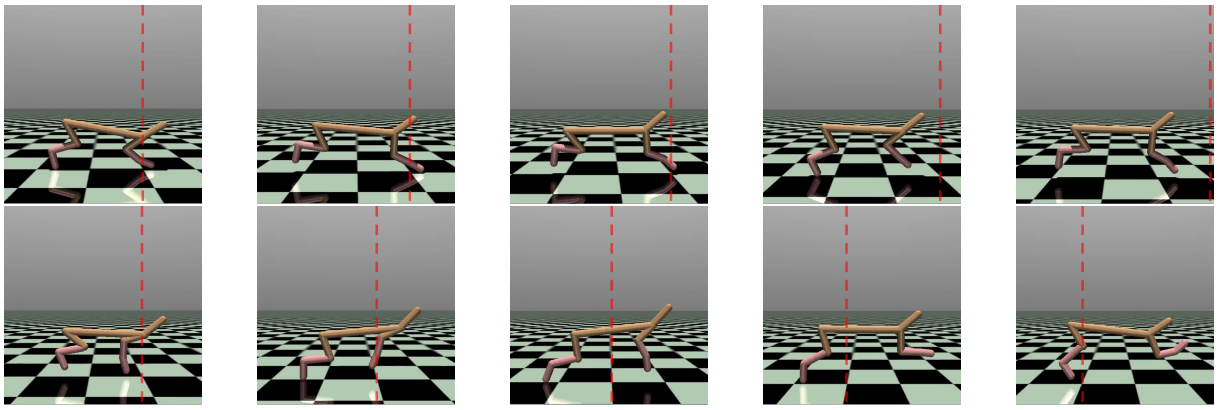
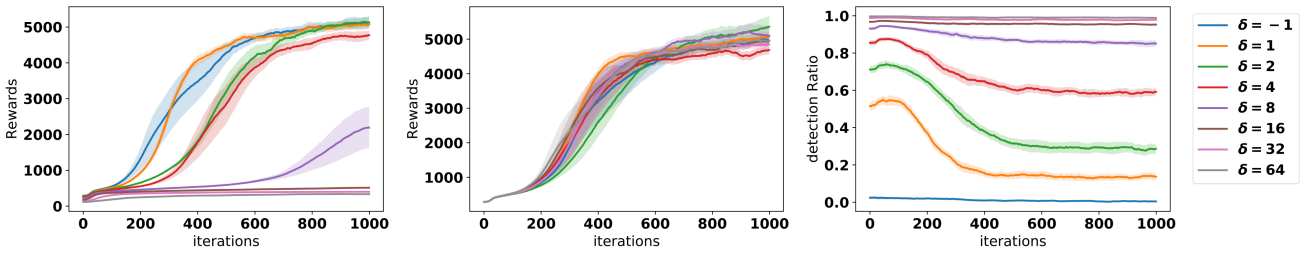


Figure 2. Consecutive Frames of Half-Cheetah trained with TRPO (top row) and FPG (bottom row) respectively under $\delta = 100$ attack. TRPO was fooled to learn a "running backward" policy, contrasted with the normal "running forward" policy learned by FPG.



(a) TRPO Rewards (b) FPG Rewards (c) FPG Detection Ratio

Figure 3. Detailed Results on Humanoid-v3.

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.

7. 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:

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 robust reinforcement learning.

8. 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.

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

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.