Appendices

A. Additional 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).

B. Proof for lower bound result

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

$$P_1(s_2|s_1,a_1) = \frac{1-\varepsilon}{2}, P_1(s_3|s_1,a_1) = \frac{1+\varepsilon}{2}, P_1(s_3|s_1,a_2) = P_1(s_3|s_1,a_2) = \frac{1}{2} \tag{8}$$

$$P_2(s_2|s_1,a_1) = \frac{1+\varepsilon}{2}, P_2(s_3|s_1,a_1) = \frac{1-\varepsilon}{2}, P_2(s_3|s_1,a_2) = P_2(s_3|s_1,a_2) = \frac{1}{2} \tag{9}$$

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

$$P\left(\sum_{i=1}^{T}\mathbb{1}_{\{a_1 \text{ is taken at } s_1\}}\mathbb{1}_{\{X_i\neq Y_i\}}\geq \varepsilon T\right)\leq P(\sum_{i=1}^{T}\mathbb{1}_{\{X_i\neq Y_i\}}\geq T\varepsilon)\leq \frac{1}{2} \tag{10}$$

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. ■

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

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

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

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:

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

for $i\neq j$.

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

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

$$
\begin{aligned}
&\leq \frac{\sigma^2}{1-\gamma}+\sum_{t=1}^{\infty}\gamma^t\mathbb{E}\left[\bar{r}(s_t,a_t)^2\right]+2\sum_{i=1}^{\infty}\sum_{j=i+1}^{\infty}\gamma^j\mathbb{E}\left[\bar{r}(s_i,a_i)\bar{r}(s_j,a_j)\right]-\left(\mathbb{E}\left[\hat{Q}^{\pi}(s_0,a_0)\right]-\bar{r}(s_0,a_0)\right)^2\\
&\leq \frac{\sigma^2}{1-\gamma}+\sum_{t=1}^{\infty}\gamma^t\mathbb{E}\left[\bar{r}(s_t,a_t)\right]+2\sum_{i=1}^{\infty}\sum_{j=i+1}^{\infty}\gamma^j\mathbb{E}\left[\bar{r}(s_i,a_i)\right]-\left(\mathbb{E}\left[\hat{Q}^{\pi}(s_0,a_0)\right]-\bar{r}(s_0,a_0)\right)^2\\
&= \frac{\sigma^2}{1-\gamma}+\sum_{t=1}^{\infty}\gamma^t\mathbb{E}\left[\bar{r}(s_t,a_t)\right]+2\sum_{i=1}^{\infty}\frac{\gamma^{i+1}}{1-\gamma}\mathbb{E}\left[\bar{r}(s_i,a_i)\right]-\left(\mathbb{E}\left[\hat{Q}^{\pi}(s_0,a_0)\right]-\bar{r}(s_0,a_0)\right)^2\\
&= \frac{\sigma^2}{1-\gamma}+\frac{1+\gamma}{1-\gamma}\sum_{t=1}^{\infty}\gamma^t\mathbb{E}\left[\bar{r}(s_t,a_t)\right]-\left(\sum_{t=1}^{\infty}\gamma^t\mathbb{E}\left[\bar{r}(s_t,a_t)\right]\right)^2\\
&= -\left(\sum_{t=1}^{\infty}\gamma^t\mathbb{E}\left[\bar{r}(s_t,a_t)\right]-\frac{1+\gamma}{2(1-\gamma)}\right)^2+\frac{(1+\gamma)^2}{4(1-\gamma)^2}+\frac{\sigma^2}{1-\gamma}\\
&\leq -\left(\sum_{t=1}^{\infty}\gamma^t-\frac{1+\gamma}{2(1-\gamma)}\right)^2+\frac{(1+\gamma)^2}{4(1-\gamma)^2}+\frac{\sigma^2}{1-\gamma}=\frac{\gamma}{(1-\gamma)^2}+\frac{\sigma^2}{1-\gamma}
\end{aligned}
$$

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. ■

D. Proofs for Section 4.

Lemma D.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* $1-\delta$*,*

$$
\begin{aligned}
&\mathbb{E}_{s,a\sim d^{(t)}}\left[\left(Q^{\pi^{(t)}}(s,a)-\phi(s,a)^{\top}w^{(t)}\right)^2\right] \qquad (11)\\
&\qquad\leq 4\left(W^2+WH\right)\left(\varepsilon^{(t)}+\sqrt{\frac{8}{M}\log\frac{4d}{\delta}}\right).
\end{aligned}
$$

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

Proof of Lemma D.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.

$$
\begin{aligned}
\hat{f} &= \frac{1}{M}\sum_{i=1}^{M}(y_i-x_i^{\top}w)^2 &(12)\\
f^{\dagger} &= \frac{1}{M}\left[\sum_{i\in C}(y_i^{\dagger}-x_i^{\dagger\top}w)^2+\sum_{i\notin C}(y_i-x_i^{\top}w)^2\right] &(13)\\
f &= \mathbb{E}(y_i-x_i^{\top}w)^2 &(14)
\end{aligned}
$$

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

$$
||\nabla_w f^{\dagger}-\nabla_w f||\leq||\nabla_w f^{\dagger}-\nabla_w\hat{f}||+||\nabla_w\hat{f}-\nabla_w f||. \qquad (15)
$$

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

$$\begin{aligned} & ||\nabla_w \hat{f} - \nabla_w f|| && (19)\\ \leq & \; 2 \left\| \left(\mathbb{E}[xx^\top] - \frac{1}{M} \sum_{i=1}^M x_i x_i^\top \right) w - \left(\mathbb{E}[yx] - \frac{1}{M} \sum_{i=1}^M y_i x_i \right) \right\| && (20)\\ \leq & \; 2 \left(\frac{2}{3M} \log \frac{4d}{\delta} + \sqrt{\frac{2}{M} \log \frac{4d}{\delta}} \right) W + 2 \sqrt{\frac{2}{M} \log \frac{4d}{\delta}} \cdot 2H && (21)\\ \leq & \; 4 \sqrt{\frac{8}{M} \log \frac{4d}{\delta}}\,(W+H)\,, \text{for } M \geq 2 \log \frac{4d}{\delta}. && (22) \end{aligned}$$

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

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

■

Theorem D.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 D.1. First note that $\varepsilon_{stat} = \mathbb{E}_{s,a \sim d^{(t)}}[\left(\phi(s,a)^\top (w^{(t)} - w^*)\right)^2] \leq 4W^2$, 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:

$$\begin{aligned} & \mathbb{E}\left[\mathbb{E}_{s,a \sim d^{(t)}} \left[\left(Q^{\pi^{(t)}}(s,a) - \phi(s,a)^\top w^{(t)} \right)^2 \right] \right] && (25)\\ \leq & \; 4\left(W^2 + HW\right) \left(\varepsilon^{(t)} + \sqrt{\frac{8}{M} \log \frac{4d}{\delta}} \right) + 8\delta W^2 && (26) \end{aligned}$$

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

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

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

$$T = \frac{2W^2\log|\mathcal{A}|}{\alpha^2(1-\gamma)^2} \tag{27}$$

$$\delta = \frac{\alpha^2(1-\gamma)^3}{32W^2|\mathcal{A}|\kappa} \tag{28}$$

$$M = \frac{512|\mathcal{A}|^2W^2(W+H)^2\kappa^2}{\alpha^4(1-\gamma)^6}\log\frac{4d}{\delta}, \tag{29}$$

we get

$$\mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T}\{V^*(\mu_0)-V^{(t)}(\mu_0)\}\right]\leq 3\alpha+\sqrt{\frac{16|\mathcal{A}|\kappa\left(W^2+HW\right)\varepsilon}{(1-\gamma)^3}}. \tag{30}$$

with sample complexity

$$TM=\frac{1024|\mathcal{A}|^2\log|\mathcal{A}|W^4(W+H)^2\kappa^2}{\alpha^6(1-\gamma)^8}\log\frac{128W^2|\mathcal{A}|\kappa d}{\alpha^2(1-\gamma)^3}. \tag{31}$$

■

E. 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

$$\bar{f}(w)=\mathbb{E}_{(X,Y)\sim D_{xy}}[(Y-w^\top X)^2] \tag{32}$$

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

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

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.

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

$$v^\top \mathrm{Cov}_{(X_i,Y_i)\sim D_{xy}}[\nabla f_i(w)]v \le 4s^2\xi + 4Cs^4\|w^*-w\|^2. \tag{33}$$

Proof of Lemma E.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 $\mathbb{E}[g(X)h(e)] = \mathbb{E}_X\left[\mathbb{E}_{h(e)|X}[g(X)h(e)|X]\right] = \mathbb{E}_X\left[g(X)\mathbb{E}_{h(e)|X}[h(e)|X]\right]$. Then

$$\begin{aligned}
&\mathbb{E}\left[\nabla f_i(w)\nabla f_i(w)^\top\right] = 4\mathbb{E}\left[((w^*-w)^\top X_i + e_i)^2 X_i X_i^\top\right] && (34)\\
=\ & 4\mathbb{E}\left[((w^*-w)^\top X_i)^2 X_i X_i^\top\right] + 4\mathbb{E}\left[e_i^2 X_i X_i^\top\right] + 4\mathbb{E}\left[2(w^*-w)^\top X_i e_i X_i X_i^\top\right] && (35)\\
=\ & 4\mathbb{E}\left[((w^*-w)^\top X_i)^2 X_i X_i^\top\right] + 4\mathbb{E}\left[X_i X_i^\top \mathbb{E}\left[e_i^2\middle|X_i\right]\right] && (36)
\end{aligned}$$

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

$$\begin{aligned}
v^\top\mathbb{E}\left[((w^*-w)^\top X_i)^2 X_i X_i^\top\right] v &= \mathbb{E}\left[((w^*-w)^\top X_i)^2 (v^\top X_i)^2\right] && (37)\\
&\le \sqrt{\mathbb{E}\left[((w^*-w)^\top X_i)^4\right]\mathbb{E}\left[(v^\top X_i)^4\right]} && (38)\\
&\le Cs^4\|w^*-w\|^2 && (39)
\end{aligned}$$

and

$$v^\top\mathbb{E}\left[X_i X_i^\top \mathbb{E}\left[e_i^2\middle|X_i\right]\right] v \le \xi v^\top \mathbb{E}\left[X_i X_i^\top\right] v \le s^2\xi \tag{40}$$

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

$$v^\top \mathrm{Cov}_{(X_i,Y_i)\sim D_{xy}}[\nabla f_i(w)]v \le v^\top\mathbb{E}\left[\nabla f_i(w)\nabla f_i(w)^\top\right] v \le 4s^2\xi + 4Cs^4\|w^*-w\|^2. \tag{41}$$

■

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

Theorem E.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 $\varepsilon' = \tilde{C}\,(\log(1/\tau)/n + \varepsilon) = O(1)$, 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 E.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 $w \in \mathcal{H}$:*

$$\left\| \frac{1}{|I_{good}|}\sum_{i\in I_{good}} \nabla f_i(w) - \nabla \bar{f}(w) \right\| \leq \sigma\alpha(\varepsilon,n,\tau) \tag{42}$$

and

$$\left\| \frac{1}{|I_{good}|}(\nabla f_i(w) - \nabla \bar{f}(w))(\nabla f_i(w) - \nabla \bar{f}(w))^\top \right\| \leq \sigma^2\beta(\varepsilon,n,\tau) \tag{43}$$

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

Proposition E.1. Given a linear regression model $f_i(w) = (Y_i - w^\top X_i)^2$ satisfying Assumption E.1, $X_i \sim D_x$, $D_e \sim D_e$, with probability at least $1-\tau$, $\{f_i\}_{i\in[n]}$ satisfies Assumption E.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 E.1. By Theorem E.1 and Lemma E.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.

$$\begin{aligned}
&\left\| \frac{1}{|I_{good}|}(\nabla f_i(w) - \nabla \bar{f}(w))(\nabla f_i(w) - \nabla \bar{f}(w))^\top \right\| && (44)\\
\leq\ &\left\| \frac{1}{|I_{good}|}(\nabla f_i(w) - \nabla \bar{f}(w))(\nabla f_i(w) - \nabla \bar{f}(w))^\top - \|\mathrm{Cov}_{f\in p^*}[\nabla f]\|\, I \right\| + \|\mathrm{Cov}_{f\in p^*}[\nabla f]\| && (45)\\
\leq\ &\left(4s^2\xi + 4Cs^4\|w^* - w\|^2\right) O\left(\frac{d\log d}{\log(1/\tau)+n\varepsilon} + 1\right) && (46)\\
\leq\ &\left(2s\sqrt{\xi} + 2\sqrt{C}s^2\|w^* - w\|\right)^2 O\left(\frac{d\log d}{\log(1/\tau)+n\varepsilon} + 1\right) =: \sigma^2\beta(\varepsilon,n,\tau). && (47)
\end{aligned}$$

$$\|\nabla \hat{f}(w) - \nabla \bar{f}(w)\| \leq \sigma O\left(\sqrt{(d\log d)/n} + \sqrt{\varepsilon} + \sqrt{\log(1/\tau)/n}\right) =: \sigma\alpha(\varepsilon,n,\tau). \tag{48}$$

■

E.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 E.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 $f_1,\ldots,f_n,\bar{f} : \mathcal{H} \to \mathbb{R}$ are bounded below, and that Assumption E.2 is satisfied. Then SEVER applied to $f_1,\ldots,f_n$ 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 E.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 $f_1, \ldots, f_n : \mathcal{H} \to \mathbb{R}$, suppose that Assumption E.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 E.1, we can prove the following expected optimality gap:

Theorem E.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 E.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*

$$\mathbb{E}\left[\bar{f}(w) - \bar{f}(w^*)\right] = O\left(\left(sr\sqrt{\xi} + s^2r^2\right)\left(\tau + \sqrt{(d\log d)/n} + \sqrt{\varepsilon} + \sqrt{\log(1/\tau)/n}\right)\right). \tag{49}$$

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

Proof of Theorem E.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:

$$\begin{aligned}
\bar{f}(w) - \bar{f}(w^*) &= \mathbb{E}_{X,Y}\left[(Y - w^\top X)^2 - (Y - {w^*}^\top X)^2\right] && (50)\\
&= \mathbb{E}_{X,e}\left[(w^* - w)^\top X + e)^2 - e^2\right] && (51)\\
&= (w^* - w)^\top \mathbb{E}_X[XX^\top](w^* - w) \leq s^2(w - w^*)^2 \leq 4s^2r^2. && (52)
\end{aligned}$$

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 E.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, $\forall 0 \leq p < \sqrt{\varepsilon}$

$$P(A_2, A_3(p) \mid A_1) = 0. \tag{53}$$

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

$$P(A_2, A_3(p) \mid A_1) \leq p. \tag{54}$$

By Proposition E.1,

$$P(A_1) \geq 1 - \tau. \tag{55}$$

By Lemma E.3,

$$P(A_2 \mid A_1) \geq 1 - \sqrt{\varepsilon}, \tag{56}$$

and thus

$$1 - P(A_1, A_2) = 1 - P(A_2 \mid A_1)P(A_1) \leq \tau + \sqrt{\varepsilon}. \tag{57}$$

Then, we have:

$$\begin{aligned}
&P\left(\bar{f}(w) - \bar{f}(w^*) > x_1/\sqrt{p} \mid A_1, A_2\right) && (58)\\
\leq &P\left(A_3(p) \mid A_1, A_2\right) = P(A_2, A_3(p) \mid A_1)/P(A_2 \mid A_1) && (59)\\
\leq &\begin{cases} 0 & 0 \leq p < \sqrt{\varepsilon} \\ \frac{p}{1-\sqrt{\varepsilon}} & \sqrt{\varepsilon} \leq p \leq 1 \end{cases}. && (60)
\end{aligned}$$

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

$$P\left(\bar{f}(w) - \bar{f}(w^*) > x \middle| A_1, A_2\right) \leq \begin{cases} 0 & x \geq x_1\varepsilon^{-1/4} \\ \frac{1}{1-\sqrt{\varepsilon}}\frac{x_1^2}{x^2} & x_1 \leq x < x_1\varepsilon^{-1/4} \\ 1 & 0 \leq x < x_1 \end{cases}. \tag{61}$$

Algorithm 4 SEVER$(f_{1:n}, \mathcal{L}, \sigma)$

Input: Sample functions $f_1, \ldots, f_n : \mathcal{H} \to \mathbb{R}$, bounded below on a closed domain $\mathcal{H}$, γ-approximate learner $\mathcal{L}$, and parameter $\sigma \in \mathbb{R}_+$.
Initialize $S \leftarrow \{1, \ldots, n\}$.
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 τ 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', τ, σ) ▷ Remove some i's with the largest scores τ_i from S; see Algorithm 5.
until $S = S'$.
Return w.

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

$$\begin{aligned}
\mathbb{E}\left[\bar{f}(w) - \bar{f}(w^*)\right] &\leq \mathbb{E}\left[\bar{f}(w) - \bar{f}(w^*) \middle| A_1, A_2\right] P(A_1, A_2) + 4s^2 r^2 (1 - P(A_1, A_2)) && (62)\\
&\leq \int_0^\infty P\left(\bar{f}(w) - \bar{f}(w^*) > x \middle| A_1, A_2\right) dx + 4s^2 r^2 (\tau + \sqrt{\varepsilon}) && (63)\\
&= \int_0^{x_1} 1 dx + \frac{1}{1 - \sqrt{\varepsilon}} \int_{x_1}^{x_1 \varepsilon^{-1/4}} \frac{x_1^2}{x^2} dx + 4s^2 r^2 (\tau + \sqrt{\varepsilon}) && (64)\\
&\leq 2x_1 + 4s^2 r^2 (\tau + \sqrt{\varepsilon}) && (65)\\
&= 2Mr\sigma \left(\alpha / \sqrt{\varepsilon} + \sqrt{\alpha^2 + \beta} \right) \sqrt{\varepsilon} + 4s^2 r^2 (\tau + \sqrt{\varepsilon}) && (66)\\
&= O\left(\left(sr\sqrt{\xi} + s^2 r^2 \right) \left(\tau + \sqrt{(d \log d)/n} + \sqrt{\varepsilon} + \sqrt{\log(1/\tau)/n} \right) \right) && (67)
\end{aligned}$$

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

E.3. Proof of Theorem E.2

In this proof, we mainly follow the steps in (Diakonikolas et al., 2019) but use our notion of stability in Assumption E.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 4 and Algorithm 5. Throughout this proof we let I_{good} be as in

Algorithm 5 FILTER(S, τ, σ)

Input: Set $S \subseteq [n]$, vector τ of outlier scores, and parameter $\sigma \in \mathbb{R}_+$.
If $\frac{1}{|S|} \sum_{i \in S} \tau_i \leq c_0 \cdot \sigma^2$, 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 $\{i \in S : \tau_i < T\}$.

Assumption E.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:

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

$$\frac{2(1 - \varepsilon)\sigma^2}{c_1 - 2\varepsilon} \left(\alpha(\varepsilon, n, \tau)^2 + \beta(\varepsilon, n, \tau) \right) \tag{68}$$

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

$$\mathbb{E}[|I_{\text{good}} \cap (S \backslash S')|] \leq \mathbb{E}[|([n] \backslash I_{\text{good}}) \cap (S \backslash S')|]. \tag{69}$$

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

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

$$\left\| \nabla \bar{f}(w) - \frac{1}{|I_{\text{good}}|} \sum_{i \in S} \nabla f_i(w) \right\|_2 \leq O\left(\sigma \left(\alpha(\varepsilon, n, \tau) + \sqrt{\alpha(\varepsilon, n, \tau)^2 + \beta(\varepsilon, n, \tau)} \sqrt{\varepsilon / p} \right) \right) \tag{70}$$

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

Proof of Theorem E.2 assuming Lemma E.3 and Lemma E.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 E.3 says that with probability at least $1 - p$, SEVER will not remove too many points, this will allow us to apply Lemma E.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 E.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 $\mathbb{E}_{i \in S}[\tau_i] \geq \frac{2(1-\varepsilon)\sigma^2}{c_1 - 2\varepsilon} \left(\alpha(\varepsilon, n, \tau)^2 + \beta(\varepsilon, n, \tau) \right)$.

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

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

$$\begin{aligned}
\text{Cov}_{i \in S_{\text{good}}}[v \cdot \nabla f_i(w)] &\leq \frac{1 - \varepsilon}{c_1 - \varepsilon} \text{Cov}_{i \in I_{\text{good}}}[v \cdot \nabla f_i(w)] \quad (\text{since } |S_{\text{good}}| \geq \tfrac{c_1 - \varepsilon}{1 - \varepsilon} |I_{\text{good}}|) && (71) \\
&= \frac{1 - \varepsilon}{c_1 - \varepsilon} \left(\frac{1}{|I_{\text{good}}|} \sum_{i \in I_{\text{good}}} (v \cdot (\nabla f_i(w) - \bar{f}(w)))^2 - (\bar{f}(w) - \mathbb{E}_{i \in I_{\text{good}}}[v \cdot \nabla f_i(w)])^2 \right) && (72) \\
&\leq \frac{(1 - \varepsilon)\sigma^2}{c_1 - \varepsilon} \left(\alpha(\varepsilon, n, \tau)^2 + \beta(\varepsilon, n, \tau) \right) \quad (\text{By Assumption E.2}), && (73)
\end{aligned}$$

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

$$\mathbb{E}_{i \in S_{\text{good}}}[\tau_i] = \text{Cov}_{i \in S_{\text{good}}}[v \cdot \nabla f_i(w)] + (\mu - \mu_{\text{good}})^2 \leq \frac{(1 - \varepsilon)\sigma^2}{c_1 - \varepsilon} \left(\alpha(\varepsilon, n, \tau)^2 + \beta(\varepsilon, n, \tau) \right) + (\mu - \mu_{\text{good}})^2 . \tag{74}$$

We now split into two cases.

Firstly, if

$$(\mu - \mu_{\text{good}})^2 \geq \frac{\varepsilon}{c_1 - 2\varepsilon} \frac{(1 - \varepsilon)\sigma^2}{c_1 - \varepsilon} \left(\alpha(\varepsilon, n, \tau)^2 + \beta(\varepsilon, n, \tau) \right), \tag{75}$$

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

$$\mathbb{E}_{i \in S_{\text{bad}}}[\tau_i] = \text{Cov}_{i \in S_{\text{bad}}}[v \cdot \nabla f_i(w)] + (\mu - \mu_{\text{bad}})^2 \geq (\mu - \mu_{\text{bad}})^2 \tag{76}$$

$$= (\mu - \mu_{\text{good}})^2 \left(\frac{|S_{\text{good}}|}{|S_{\text{bad}}|}\right)^2 \tag{77}$$

$$\geq \frac{|S_{\text{good}}|}{|S_{\text{bad}}|}\frac{c_1-\varepsilon}{\varepsilon}(\mu - \mu_{\text{good}})^2 \quad \text{(because } |S_{\text{good}}| \geq (c_1-\varepsilon)n \text{ and } |S_{\text{bad}}| \leq \varepsilon n\text{)} \tag{78}$$

$$= \frac{|S_{\text{good}}|}{|S_{\text{bad}}|}\left(\frac{c_1-2\varepsilon}{\varepsilon}(\mu - \mu_{\text{good}})^2 + (\mu - \mu_{\text{good}})^2\right) \tag{79}$$

$$\geq \frac{|S_{\text{good}}|}{|S_{\text{bad}}|}\left(\frac{(1-\varepsilon)\sigma^2}{c_1-\varepsilon}\left(\alpha(\varepsilon,n,\tau)^2 + \beta(\varepsilon,n,\tau)\right) + (\mu - \mu_{\text{good}})^2\right) \quad \text{(by (75))} \tag{80}$$

$$\geq \frac{|S_{\text{good}}|}{|S_{\text{bad}}|}\mathbb{E}_{i\in S_{\text{good}}}[\tau_i] \quad \text{(by (74))}. \tag{81}$$

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 E.3. Define the event

$$A = \{n - |S| \leq (1+1/p)\varepsilon n\}, \tag{82}$$

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 E.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

$$4\sigma^2 \geq \frac{2(1-\varepsilon)\sigma^2}{c_1 - 2\varepsilon} \tag{83}$$

And Lemma E.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 E.4.

Proof of Lemma E.4. We note that

$$\left\|\sum_{i\in S}(\nabla f_i(w) - \nabla \bar{f}(w))\right\|_2 \tag{84}$$

$$\leq \left\|\sum_{i\in I_{\text{good}}}(\nabla f_i(w) - \nabla \bar{f}(w))\right\|_2 + \left\|\sum_{i\in (I_{\text{good}}\backslash S)}(\nabla f_i(w) - \nabla \bar{f}(w))\right\|_2 + \left\|\sum_{i\in (S\backslash I_{\text{good}})}(\nabla f_i(w) - \nabla \bar{f}(w))\right\|_2 \tag{85}$$

$$\leq \left\|\sum_{i\in (I_{\text{good}}\backslash S)}(\nabla f_i(w) - \nabla \bar{f}(w))\right\|_2 + \left\|\sum_{i\in (S\backslash I_{\text{good}})}(\nabla f_i(w) - \nabla \bar{f}(w))\right\|_2 + n\sigma\alpha(\varepsilon,n,\tau). \tag{86}$$

First we analyze

$$\left\|\sum_{i\in (I_{\text{good}}\backslash S)}(\nabla f_i(w) - \nabla \bar{f}(w))\right\|_2. \tag{87}$$

This is the supremum over unit vectors v of

$$\sum_{i\in(I_{\text{good}}\setminus S)} v\cdot(\nabla f_i(w)-\nabla\bar{f}(w)). \tag{88}$$

However, we note that

$$\sum_{i\in I_{\text{good}}}(v\cdot(\nabla f_i(w)-\nabla\bar{f}(w)))^2\le n\sigma^2\beta(\varepsilon,n,\tau). \tag{89}$$

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

$$\sum_{i\in(I_{\text{good}}\setminus S)} v\cdot(\nabla f_i(w)-\nabla\bar{f}(w))=\sqrt{(n\sigma^2\beta(\varepsilon,n,\tau))((1+1/p)\varepsilon n)}=n\sigma\sqrt{\beta(\varepsilon,n,\tau)(1+1/p)\varepsilon}, \tag{90}$$

as desired.

Let

$$\Delta:=\left\|\sum_{i\in S}(\nabla f_i(w)-\nabla\bar{f}(w))\right\|_2. \tag{91}$$

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

$$\sum_{i\in S}(v\cdot(\nabla f_i(w)-\nabla\bar{f}(w)))^2=\sum_{i\in S}(v\cdot(\nabla f_i(w)-\nabla\hat{f}(w)))^2+|S|(v\cdot(\nabla\hat{f}(w)-\nabla\bar{f}(w)))^2 \tag{92}$$

$$\le\sum_{i\in S}(v\cdot(\nabla f_i(w)-\nabla\hat{f}(w)))^2+\Delta^2/|S|\le n4(\alpha(\varepsilon,n,\tau)^2+\beta(\varepsilon,n,\tau))\sigma^2+\Delta^2/((1-(1+1/p)\varepsilon)n) \tag{93}$$

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

$$\left\|\sum_{i\in(S\setminus I_{\text{good}})}\nabla f_i(w)-\nabla\bar{f}(w)\right\|_2\le 2n\sigma\sqrt{\alpha(\varepsilon,n,\tau)^2+\beta(\varepsilon,n,\tau)}\sqrt{(1+1/p)\varepsilon}+\Delta\sqrt{\frac{(1+1/p)\varepsilon}{1-(1+1/p)\varepsilon}}. \tag{94}$$

Combining with the above we have that

$$\frac{\Delta}{n}\le\sigma\alpha(\varepsilon,n,\tau)+\sigma\sqrt{\beta(\varepsilon,n,\tau)(1+1/p)\varepsilon}+2\sigma\sqrt{\alpha(\varepsilon,n,\tau)^2+\beta(\varepsilon,n,\tau)}\sqrt{(1+1/p)\varepsilon}+\frac{\Delta}{n}\sqrt{\frac{(1+1/p)\varepsilon}{1-(1+1/p)\varepsilon}}, \tag{95}$$

Thus

$$\frac{\Delta}{n}\le\frac{1}{1-\sqrt{\frac{(1+1/p)\varepsilon}{1-(1+1/p)\varepsilon}}}\left(\sigma\alpha(\varepsilon,n,\tau)+6\sigma\sqrt{\alpha(\varepsilon,n,\tau)^2+\beta(\varepsilon,n,\tau)}\sqrt{\varepsilon/p}\right) \tag{96}$$

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. ■

F. Proofs for Section 5

Lemma F.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 $\varepsilon^{(t)}\le c$, for some absolute constant c, and any constant $\tau\in[0,1]$, we have*

$$\mathbb{E}\left[\mathbb{E}_{s,a\sim d^{(t)}}\left[\left(Q^{\pi^{(t)}}(s,a)-\phi(s,a)^\top w^{(t)}\right)^2\right]\right] \tag{97}$$

$$\le O\left(\left(W^2+\frac{\sigma W}{1-\gamma}\right)\left(\sqrt{\varepsilon^{(t)}}+f(d,\tau)M^{-\frac{1}{2}}+\tau\right)\right).$$

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

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

1. By Lemma C.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 E.3 concludes the proof. ■

Theorem F.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*

$$\begin{aligned}&\mathbb{E}\left[V^*(\mu_0) - V^{\hat{\pi}}(\mu_0)\right] \\ &\qquad \leq \tilde{O}\left(\max\left[\alpha, \sqrt{\frac{|\mathcal{A}|\kappa\left(W^2+\sigma W\right)}{(1-\gamma)^4}}\varepsilon^{1/4}\right]\right).\end{aligned} \tag{98}$$

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

Proof of Theorem F.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

$$\mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T}\{V^*(\mu_0) - V^{(t)}(\mu_0)\}\right] \tag{99}$$

$$\leq \frac{W}{1-\gamma}\sqrt{\frac{2\log|\mathcal{A}|}{T}} + \frac{1}{T}\sum_{t=1}^{T}\sqrt{\frac{4|\mathcal{A}|\kappa\varepsilon_{stat}^{(t)}}{(1-\gamma)^3}} \tag{100}$$

$$\leq \frac{W}{1-\gamma}\sqrt{\frac{2\log|\mathcal{A}|}{T}} + \frac{z^2}{c}\varepsilon + \frac{1}{T}\sum_{t=1}^{T}\sqrt{\frac{4|\mathcal{A}|\kappa b\left(\sqrt{\varepsilon^{(t)}} + \sqrt{(d\log d)/M} + \sqrt{\log(1/\tau)/M} + \tau\right)}{(1-\gamma)^3}} \tag{101}$$

$$\leq \frac{W}{1-\gamma}\sqrt{\frac{2\log|\mathcal{A}|}{T}} + \frac{z^2}{c}\varepsilon + \sqrt{\frac{4|\mathcal{A}|\kappa b\left(\sqrt{(d\log d)/M} + \sqrt{\log(1/\tau)/M} + \tau\right)}{(1-\gamma)^3}} + \frac{1}{T}\sum_{t=1}^{T}\sqrt{\frac{4|\mathcal{A}|\kappa b\sqrt{\varepsilon^{(t)}}}{(1-\gamma)^3}} \tag{102}$$

$$\leq \frac{W}{1-\gamma}\sqrt{\frac{2\log|\mathcal{A}|}{T}} + \frac{z^2}{c}\varepsilon + \sqrt{\frac{4|\mathcal{A}|\kappa b\left(\sqrt{(d\log d)/M} + \sqrt{\log(1/\tau)/M} + \tau\right)}{(1-\gamma)^3}} + \sqrt{\frac{4|\mathcal{A}|\kappa b}{(1-\gamma)^3}}\varepsilon^{1/4} \tag{103}$$

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

$$T = \frac{2W^2\log|\mathcal{A}|}{\alpha^2(1-\gamma)^2} \tag{104}$$

$$\tau = \frac{\alpha^2(1-\gamma)^3}{4|\mathcal{A}|b\kappa} \tag{105}$$

$$M = \frac{16|\mathcal{A}|^2b^2\kappa^2}{\alpha^4(1-\gamma)^6}\max\left[d\log d, \log(1/\tau)\right] \tag{106}$$

we get

$$\mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T}\{V^*(\mu_0) - V^{(t)}(\mu_0)\}\right] \leq O\left(\alpha + \sqrt{\frac{|\mathcal{A}|\kappa b}{(1-\gamma)^3}}\varepsilon^{1/4}\right). \tag{107}$$

Algorithm 6 FPG-TRPO

Input: initial policy parameter θ_0; initial value function parameter ϕ_0.
Hyperparameters: KL-divergence limit δ; backtracking coefficient α; maximum number of backtracking steps K; upper-bound of corruption level ε; 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_{ϕ_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. 7 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

$$\theta_{k+1} = \theta_k + \alpha^j \sqrt{\frac{2\delta}{\hat{x}_k \hat{g}_k}} \hat{x}_k \tag{109}$$

  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:

$$\phi_{k+1} = \arg\min_{\phi} \frac{1}{|S_k|} \sum_{(t,i)\in S_k} \left(V_\phi(s_{t,i}) - \hat{R}_{t,i} \right)^2 \tag{110}$$

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

Algorithm 7 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 2ε 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 τ 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 $(S, \hat{x})$.

with sample complexity

$$TM = \frac{32W^2|\mathcal{A}|^2 \log|\mathcal{A}| b^2 \kappa^2}{\alpha^6 (1-\gamma)^8} \max\left[d \log d, \log(1/\tau)\right]. \tag{108}$$

■

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:

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

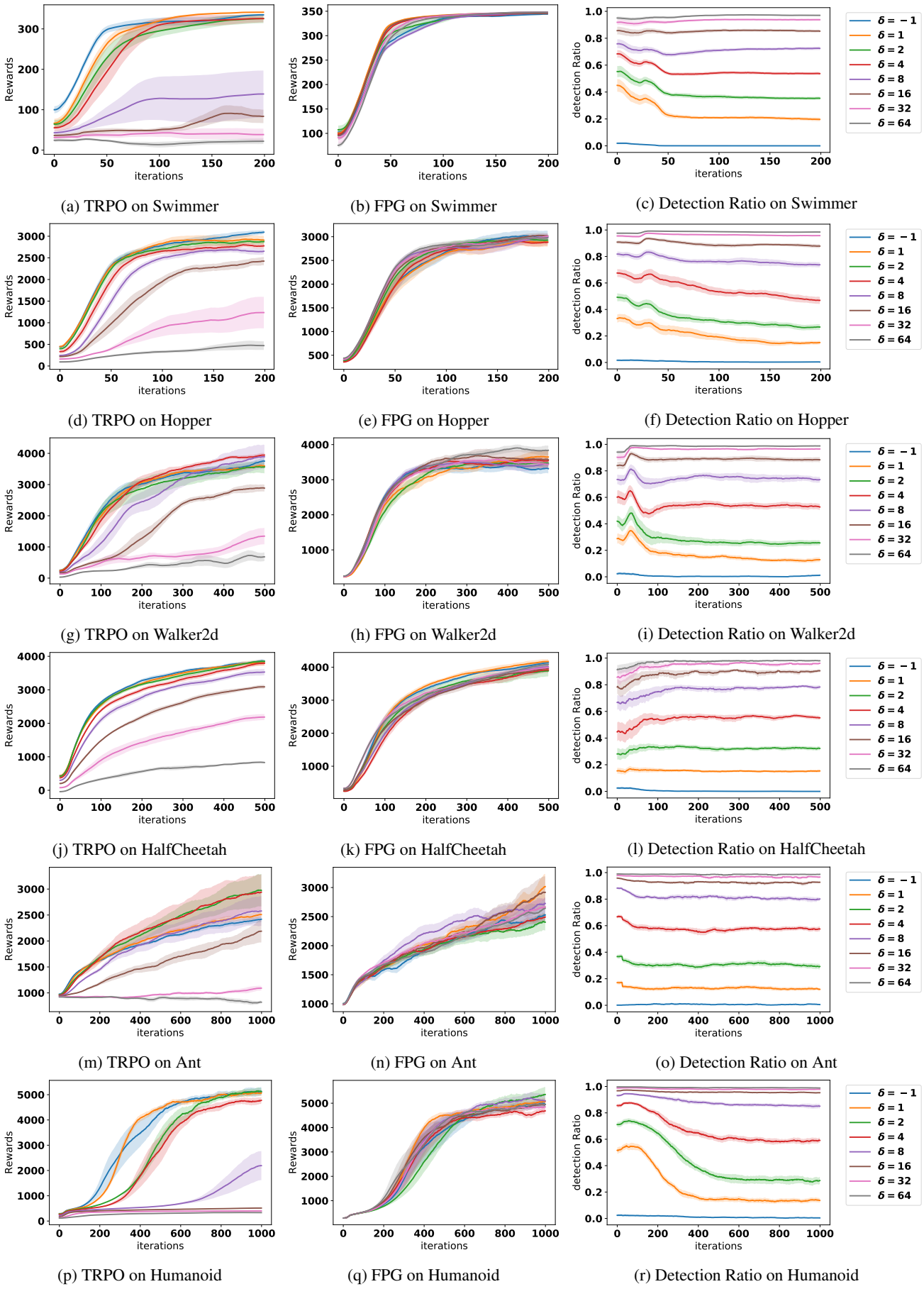
Parameters	Values	Description
γ	0.995	discounting factor.
λ	0.97	GAE parameter (Schulman et al., 2015b).
l2-reg	0.001	L2 regularization weight in value loss.
δ	0.01	KL constraint in TRPO.
damping	0.1	damping factor in conjugate gradient.
batch-size	25000	number of time steps per policy gradient iteration.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

Table 1. Hyperparameters for FPG-TRPO.

gradient (FCG) subroutine to replace the standard CG produce. The FPG procedure is detailed in Alg. 7. 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. 7).

The pseudo-code of `FPG-TRPO` can be found in Alg. 6. Similar to the NPG variant of FPG, the only difference between Alg. 6 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. 7 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.