Appendix

The appendix is organized as follows. In Section A, we describe additional preliminaries required for our technical arguments. In Section B, we analyze our main algorithm, SEVER. In Section C, we specialize our analysis to the important case of Generalized Linear Models (GLMs). In Section D, we describe a variant of our algorithm which performs robust filtering on each iteration of projected gradient descent, and works under more general assumptions. In Section E, we describe concrete applications of SEVER – in particular, how it can be used to robustly optimize in the settings of linear regression, logistic regression, and support vector machines. Finally, in Section F, we provide additional plots from our experimental evaluations.

A Preliminaries

In this section, we formally introduce our setting for robust stochastic optimization.

Notation. For $n \in \mathbb{Z}_+$, we will denote $[n] \stackrel{\text{def}}{=} \{1, \ldots, n\}$. For a vector v, we will let $\|v\|_2$ denote its Euclidean norm. For any $r \geq 0$ and any $x \in \mathbb{R}^d$, let $B(x, r)$ be the ℓ_2 ball of radius r around x. If M is a matrix, we will let $\|M\|_2$ denote its spectral norm and $\|M\|_F$ denote its Frobenius norm. We will write $X \sim_u S$ to denote that X is drawn from the empirical distribution defined by S. We will sometimes use the notation $\mathbb{E}_S$, instead of $\mathbb{E}_{X \sim S}$, for the corresponding expectation. We will also use the same convention for the covariance, i.e. we let Cov_S denote the covariance over the empirical distribution.

Setting. We consider a stochastic optimization setting with outliers. Let $\mathcal{H} \subseteq \mathbb{R}^d$ be a space of parameters. We observe n functions $f_1, \ldots, f_n : \mathcal{H} \to \mathbb{R}$ and we are interested in (approximately) minimizing some target function $\overline{f} : \mathcal{H} \to \mathbb{R}$, related to the f_i's. We will assume for simplicity that the f_i's are differentiable with gradient ∇f_i. (Our results can be easily extended for the case that only a sub-gradient is available.)

In most concrete applications we will consider, there is some true underlying distribution p^* over functions $f : \mathcal{H} \to \mathbb{R}$, and our goal is to find a parameter vector $w^* \in \mathcal{H}$ minimizing $\overline{f}(w) \stackrel{\text{def}}{=} \mathbb{E}_{f \sim p^*}[f(w)]$. Unlike the classical realizable setting, where we assume that $f_1, \ldots, f_n \sim p^*$, we allow for an ε-fraction of the points to be arbitrary outliers. This is captured in the following definition (Definition 2.1) that we restate for convenience:

Definition A.1 (ε-corruption model)**.** Given $\varepsilon > 0$ and a distribution p^* over functions $f : \mathcal{H} \to \mathbb{R}$, data is generated as follows: first, n clean samples $f_1, \ldots, f_n$ are drawn from p^*. Then, an *adversary* is allowed to inspect the samples and replace any εn of them with arbitrary samples. The resulting set of points is then given to the algorithm.

In addition, some of our bounds will make use of the following quantities:

- The ℓ_2-radius r of the domain $\mathcal{H}$: $r = \max_{w \in \mathcal{H}} \|w\|_2$.
- The strong convexity parameter ξ of $\overline{f}$, if it exists. This is the maximal ξ such that $\overline{f}(w) \geq \overline{f}(w_0) + \langle w - w_0, \nabla \overline{f}(w_0) \rangle + \frac{\xi}{2}\|w - w_0\|_2^2$ for all $w, w_0 \in \mathcal{H}$.
- The strong smoothness parameter β of $\overline{f}$, if it exists. This is the minimal β such that $\overline{f}(w) \leq \overline{f}(w_0) + \langle w - w_0, \nabla \overline{f}(w_0) \rangle + \frac{\beta}{2}\|w - w_0\|_2^2$ for all $w, w_0 \in \mathcal{H}$.
- The Lipschitz constant L of $\overline{f}$, if it exists. This is the minimal L such that $\overline{f}(w) - \overline{f}(w_0) \leq L\|w - w_0\|_2$ for all $w, w_0 \in \mathcal{H}$.

B General Analysis of Sever

This section is dedicated to the analysis of Algorithm 1, where we do not make convexity assumptions about the underlying functions $f_1, \ldots, f_n$. In this case, we can show that our algorithm finds an approximate critical point of $\overline{f}$. When we specialize to convex functions, this immediately implies that we find an approximate minimal point of $\overline{f}$.

Our proof proceeds in two parts. First, we define a set of deterministic conditions under which our algorithm finds an approximate minimal point of $\overline{f}$. We then show that, under mild assumptions on our functions, this set of deterministic conditions holds with high probability after polynomially many samples.

For completeness, we recall the definitions of a γ-approximate critical point and a γ-approximate learner:

Definition 2.2 (γ-approximate critical point)**.** Given a function $f : \mathcal{H} \to \mathbb{R}$, a γ-approximate critical point of f, is a point $w \in \mathcal{H}$ so that for all unit vectors v where $w + \delta v \in \mathcal{H}$ for arbitrarily small positive δ, we have that $v \cdot \nabla f(w) \geq -\gamma$.

Definition 2.3 (γ-approximate learner)**.** A learning algorithm $\mathcal{L}$ is called γ-*approximate* if, for any functions $f_1, \ldots, f_n : \mathcal{H} \to \mathbb{R}$ each bounded below on a closed domain $\mathcal{H}$, the output $w = \mathcal{L}(f_{1:n})$ of $\mathcal{L}$ is a γ-approximate critical point of $f(x) := \frac{1}{n}\sum_{i=1}^{n} f_i(x)$.

Deterministic Regularity Conditions We first explicitly demonstrate a set of deterministic conditions on the (uncorrupted) data points. Our deterministic regularity conditions are as follows:

Assumption B.1. Fix $0 < \varepsilon < 1/2$. There exists an unknown set $I_{\text{good}} \subseteq [n]$ with $|I_{\text{good}}| \geq (1-\varepsilon)n$ of "good" functions $\{f_i\}_{i \in I_{\text{good}}}$ and parameters $\sigma_0, \sigma_1 \in \mathbb{R}_+$ such that:

$$\left\| \mathbb{E}_{I_{\text{good}}} \left[\left(\nabla f_i(w) - \nabla \overline{f}(w)\right) \left(\nabla f_i(w) - \nabla \overline{f}(w)\right)^T \right] \right\|_2 \leq (\sigma_0 + \sigma_1 \|w^* - w\|_2)^2, \text{ for all } w \in \mathcal{H} \,, \tag{1}$$

and

$$\|\nabla \hat{f}(w) - \nabla \overline{f}(w)\|_2 \leq (\sigma_0 + \sigma_1 \|w^* - w\|_2)\sqrt{\varepsilon}, \text{ for all } w \in \mathcal{H}, \text{ where } \hat{f} \stackrel{\text{def}}{=} \frac{1}{|I_{\text{good}}|} \sum_{i \in I_{\text{good}}} f_i \,. \tag{2}$$

In Section B.1, we prove the following theorem, which shows that under Assumption B.1 our algorithm succeeds:

Theorem B.2. *Suppose that the functions $f_1, \ldots, f_n, \overline{f} : \mathcal{H} \to \mathbb{R}$ are bounded below, and that Assumption B.1 is satisfied, where $\sigma \stackrel{def}{=} \sigma_0 + \sigma_1 \|w^* - w\|_2$. Then* Sever *applied to $f_1, \ldots, f_n, \sigma$ returns a point $w \in \mathcal{H}$ that, with probability at least $9/10$, is a $(\gamma + O(\sigma\sqrt{\varepsilon}))$-approximate critical point of $\overline{f}$.*

Observe that the above theorem holds quite generally; in particular, it holds for non-convex functions. As a corollary of this theorem, in Section B.2 we show that this immediately implies that Sever robustly minimizes convex functions, if Assumption B.1 holds:

Corollary B.3. *For functions $f_1, \ldots, f_n : \mathcal{H} \to \mathbb{R}$, suppose that Assumption B.1 holds and that $\mathcal{H}$ is convex. Then, with probability at least $9/10$, for some universal constant ε_0, if $\varepsilon < \varepsilon_0$, the output of* Sever *satisfies the following:*

(i) If $\overline{f}$ is convex, the algorithm finds a $w \in \mathcal{H}$ such that $\overline{f}(w) - \overline{f}(w^*) = O((\sigma_0 r + \sigma_1 r^2)\sqrt{\varepsilon} + \gamma r)$.

(ii) If $\overline{f}$ is ξ-strongly convex, the algorithm finds a $w \in \mathcal{H}$ such that

$$\overline{f}(w) - \overline{f}(w^*) = O\left(\frac{\varepsilon}{\xi}(\sigma_0 + \sigma_1 r)^2 + \frac{\gamma^2}{\xi}\right) \,.$$

In the strongly convex case and when $\sigma_1 > 0$, we can remove the dependence on σ_1 and r in the above by repeatedly applying Sever with decreasing r:

Corollary B.4. *For functions $f_1, \ldots, f_n : \mathcal{H} \to \mathbb{R}$, suppose that Assumption B.1 holds, that $\mathcal{H}$ is convex and that $\overline{f}$ is ξ-strongly convex for $\xi \geq C\sigma_1\sqrt{\varepsilon}$ for some absolute constant C. Then, with probability at least $9/10$, for some universal constant ε_0, if $\varepsilon < \varepsilon_0$, we can find a $\widehat{w}$ with*

$$\overline{f}(\widehat{w}) - \overline{f}(w^*) = O\left(\frac{\varepsilon\sigma_0^2 + \gamma^2}{\xi}\right) .$$

and

$$\|\widehat{w} - w^*\|_2 = O\left(\frac{\sqrt{\varepsilon}\sigma_0 + \gamma}{\xi}\right)$$

using at most $O(\log(r\xi/(\gamma + \sigma_0\sqrt{\varepsilon})))$ calls to SEVER.

To concretely use Theorem B.2, Corollary B.3, and Corollary B.4, in Section B.4 we show that the Assumption B.1 is satisfied with high probability under mild conditions on the distribution over the functions, after drawing polynomially many samples:

Proposition B.5. *Let $\mathcal{H} \subset \mathbb{R}^d$ be a closed bounded set with diameter at most r. Let p^* be a distribution over functions $f : \mathcal{H} \to \mathbb{R}$ with $\overline{f} = \mathbb{E}_{f\sim p^*}[f]$ so that $f - \overline{f}$ is L-Lipschitz and β-smooth almost surely. Assume furthermore that for each $w \in \mathcal{H}$ and unit vector v that $\mathbb{E}_{f\sim p^*}[(v \cdot (\nabla f(w) - \overline{f}(w)))^2] \leq \sigma^2/2$. Then for*

$$n = \Omega\left(\frac{dL^2 \log(r\beta L/\sigma^2\varepsilon)}{\sigma^2\varepsilon}\right) ,$$

an ε-corrupted set of points $f_1, \ldots, f_n$ with high probability satisfy Assumption B.1.

The remaining subsections are dedicated to the proofs of Theorem B.2, Corollary B.3, Corollary B.4, and Proposition B.5.

B.1 Proof of Theorem B.2

Throughout this proof we let I_{good} be as in Assumption B.1. We require the following two lemmata. Roughly speaking, the first states that on average, we remove more corrupted points than uncorrupted points, and the second 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 B.6. *If the samples satisfy (1) of Assumption B.1, and if $|S| \geq 2n/3$ then if S' is the output of* FILTER(S, τ, σ)*, we have that*

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

Lemma B.7. *If the samples satisfy Assumption B.1,* FILTER$(S, \tau, \sigma) = S$*, and $n - |S| \leq 11\varepsilon n$, then*

$$\left\| \nabla\overline{f}(w) - \frac{1}{|I_{\text{good}}|} \sum_{i\in S} \nabla f_i(w) \right\|_2 \leq O(\sigma\sqrt{\varepsilon})$$

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

Proof of Theorem B.2 assuming Lemma B.6 and Lemma B.7. 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 B.6 says that each iteration will on average throw out as many elements not in I_{good} from S as elements in I_{good}. In particular, this means 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 $9/10$, it never exceeds $10\varepsilon n$, and therefore at the end of the algorithm, we must have that $n - |S| \leq \varepsilon n + |I_{\text{good}} \backslash S| \leq 11\varepsilon n$. This will allow us to apply Lemma B.7 to complete the proof, using the fact that w is a γ-approximate critical point of $\frac{1}{|I_{\text{good}}|} \sum_{i\in S} \nabla f_i(w)$. □

Thus it suffices to prove these two lemmata. We first prove Lemma B.6:

Proof of Lemma B.6. 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 $\textsc{Filter}(S, \tau, \sigma) = S$. Thus, we can assume that $\mathbb{E}_{i \in S}[\tau_i] \geq 12\sigma$.

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

$$\begin{aligned}\text{Cov}_{i \in S_{\text{good}}}[v \cdot \nabla f_i(w)] &\overset{(a)}{\leq} \frac{3}{2} \text{Cov}_{i \in I_{\text{good}}}[v \cdot \nabla f_i(w)] \\ &= \frac{3}{2} \cdot v^\top \text{Cov}_{i \in I_{\text{good}}}[\nabla f_i(w)] v \leq 2\sigma^2 \,,\end{aligned}$$

where (a) follows since $|S_{\text{good}}| \geq \frac{3}{2} I_{\text{good}}$.

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

$$\mathbb{E}_{i \in S_{\text{good}}}[\tau_i] = \text{Cov}_{i \in S_{\text{good}}}[v \cdot \nabla f_i(w)] + (\mu - \mu_{\text{good}})^2 \leq 2\sigma + (\mu - \mu_{\text{good}})^2 \,.$$

We now split into two cases.

Firstly, if $(\mu - \mu_{\text{good}})^2 \geq 4\sigma^2$, we let $\mu_{\text{bad}} = \mathbb{E}_{i \in S_{\text{bad}}}[v \cdot \nabla f_i(w)]$, and note that $|\mu - \mu_{\text{bad}}||S_{\text{bad}}| = |\mu - \mu_{\text{good}}||S_{\text{good}}|$. We then have that

$$\begin{aligned}\mathbb{E}_{i \in S_{\text{bad}}}[\tau_i] &\geq (\mu - \mu_{\text{bad}})^2 \\ &\geq (\mu - \mu_{\text{good}})^2 \left(\frac{|S_{\text{good}}|}{|S_{\text{bad}}|}\right)^2 \\ &\geq 2\left(\frac{|S_{\text{good}}|}{|S_{\text{bad}}|}\right)(\mu - \mu_{\text{good}})^2 \\ &\geq \left(\frac{|S_{\text{good}}|}{|S_{\text{bad}}|}\right) \mathbb{E}_{i \in S_{\text{good}}}[\tau_i].\end{aligned}$$

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

On the other hand, if $(\mu - \mu_{\text{good}})^2 \leq 4\sigma^2$, then $\mathbb{E}_{i \in S_{\text{good}}}[\tau_i] \leq 6\sigma^2 \leq \mathbb{E}_{i \in S}[\tau_i]/2$. Therefore $\sum_{i \in S_{\text{bad}}} \tau_i \geq \sum_{i \in S_{\text{good}}} \tau_i$ once again. This completes our proof. □

We now prove Lemma B.7.

Proof of Lemma B.7. We need to show that

$$\delta := \left\| \sum_{i \in S} (\nabla f_i(w) - \nabla \overline{f}(w)) \right\|_2 = O(n\sigma\sqrt{\varepsilon}).$$

We note that

$$\begin{aligned}&\left\| \sum_{i \in S} (\nabla f_i(w) - \nabla \overline{f}(w)) \right\|_2 \\ \leq &\left\| \sum_{i \in I_{\text{good}}} (\nabla f_i(w) - \nabla \overline{f}(w)) \right\|_2 + \left\| \sum_{i \in (I_{\text{good}} \backslash S)} (\nabla f_i(w) - \nabla \overline{f}(w)) \right\|_2 + \left\| \sum_{i \in (S \backslash I_{\text{good}})} (\nabla f_i(w) - \nabla \overline{f}(w)) \right\|_2 \\ = &\left\| \sum_{i \in (I_{\text{good}} \backslash S)} (\nabla f_i(w) - \nabla \overline{f}(w)) \right\|_2 + \left\| \sum_{i \in (S \backslash I_{\text{good}})} (\nabla f_i(w) - \nabla \overline{f}(w)) \right\|_2 + O(n\sqrt{\sigma^2 \varepsilon}).\end{aligned}$$

First we analyze

$$\left\| \sum_{i \in (I_{\text{good}} \setminus S)} (\nabla f_i(w) - \nabla \overline{f}(w)) \right\|_2 .$$

This is the supremum over unit vectors v of

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

However, we note that

$$\sum_{i \in I_{\text{good}}} (v \cdot (\nabla f_i(w) - \nabla \overline{f}(w)))^2 = O(n\sigma^2).$$

Since $|I_{\text{good}} \setminus S| = O(n\varepsilon)$, we have by Cauchy-Schwarz that

$$\sum_{i \in (I_{\text{good}} \setminus S)} v \cdot (\nabla f_i(w) - \nabla \overline{f}(w)) = O(\sqrt{(n\sigma^2)(n\varepsilon)}) = O(n\sqrt{\sigma^2 \varepsilon}),$$

as desired.

We note that since for any such v that

$$\sum_{i \in S} (v \cdot (\nabla f_i(w) - \nabla \overline{f}(w)))^2 = \sum_{i \in S} (v \cdot (\nabla f_i(w) - \nabla \hat{f}(w)))^2 + \delta^2 = O(n\sigma^2) + \delta^2$$

(or otherwise our filter would have removed elements) and since $|S \setminus I_{\text{good}}| = O(n\varepsilon)$, and so we have similarly that

$$\left\| \sum_{i \in (S \setminus I_{\text{good}})} \nabla f_i(w) - \nabla \overline{f}(w) \right\|_2 = O(n\sigma\sqrt{\varepsilon} + \delta\sqrt{n\varepsilon}).$$

Combining with the above we have that

$$\delta = O(\sigma\sqrt{\varepsilon} + \delta\sqrt{\varepsilon/n}),$$

and therefore, $\delta = O(\sigma\sqrt{\varepsilon})$ as desired. □

B.2 Proof of Corollary B.3

In this section, we show that the SEVER algorithm finds an approximate global optimum for convex optimization in various settings, under Assumption B.1. We do so by simply applying the guarantees of Theorem B.2 in a fairly black box manner.

Before we proceed with the proof of Corollary B.3, we record a simple lemma that allows us to translate an approximate critical point guarantee to an approximate global optimum guarantee:

Lemma B.8. *Let $f : \mathcal{H} \to \mathbb{R}$ be a convex function and let $x \neq y \in \mathcal{H}$. Let $v = y - x/\|y - x\|_2$ be the unit vector in the direction of $y - x$. Suppose that for some δ that $v \cdot (\nabla f(x)) \geq -\delta$ and $-v \cdot (\nabla f(y)) \geq -\delta$. Then we have that:*

1. *$|f(x) - f(y)| \leq \|x - y\|_2 \delta$.*
2. *If f is ξ-strongly convex, then $|f(x) - f(y)| \leq 2\delta^2/\xi$ and $\|x - y\|_2 \leq 2\delta/\xi$.*

Proof. Let $r = \|x - y\|_2 > 0$ and $g(t) = f(x + tv)$. We have that $g(0) = f(x), g(r) = f(y)$ and that g is convex (or ξ-strongly convex) with $g'(0) \geq -\delta$ and $g'(r) \leq \delta$. By convexity, the derivative of g is increasing on $[0, r]$ and therefore $|g'(t)| \leq \delta$ for all $t \in [0, r]$. This implies that

$$|f(x) - f(y)| = |g(r) - g(0)| = \left| \int_0^r g'(t) dt \right| \leq r\delta .$$

To show the second part of the lemma, we note that if g is ξ-strongly convex that $g''(t) \geq \xi$ for all t. This implies that $g'(r) > g'(0) + \xi r$. Since $g'(r) - g'(0) \leq 2\delta$, we obtain that $r \leq 2\delta/\xi$, from which the second statement follows. □

Proof of Corollary B.3. By applying the algorithm of Theorem B.2, we can find a point w that is a $\gamma' \stackrel{\text{def}}{=} (\gamma + O(\sigma\sqrt{\varepsilon}))$-approximate critical point of $\overline{f}$, where $\sigma \stackrel{\text{def}}{=} \sigma_0 + \sigma_1\|w^* - w\|_2$. That is, for any unit vector v pointing towards the interior of $\mathcal{H}$, we have that $v \cdot \nabla\overline{f}(w) \geq -\gamma'$.

To prove (i), we apply Lemma B.8 to $\overline{f}$ at w which gives that

$$|\overline{f}(w) - \overline{f}(w^*)| \leq r \cdot \gamma'.$$

To prove (ii), we apply Lemma B.8 to $\overline{f}$ at w which gives that

$$|\overline{f}(w) - \overline{f}(w^*)| \leq 2{\gamma'}^2/\xi.$$

Plugging in parameters appropriately then immediately gives the desired bound. □

B.3 Proof of Corollary B.4

We apply Sever iteratively starting with a domain $\mathcal{H}_1 = \mathcal{H}$ and radius $r_1 = r$. After each iteration, we know the resulting point is close to w^* will be able to reduce the search radius.

At step i, we have a domain of radius r_i. As in the proof of Corollary B.3 above, we apply algorithm of Theorem B.2, we can find a point w_i that is a $\gamma_i' \stackrel{\text{def}}{=} (\gamma + O(\sigma_i'\sqrt{\varepsilon}))$-approximate critical point of $\overline{f}$, where $\sigma_i' \stackrel{\text{def}}{=} \sigma_0 + \sigma_1 r_i$. Then using Lemma B.8, we obtain that $\|w_i - w^*\|_2 \leq 2\gamma_i'/\xi$.

Now we can define $\mathcal{H}_{i+1}$ as the intersection of $\mathcal{H}$ and the ball of radius $r_{i+1} = 2\gamma_i'/\xi$ around w_i and repeat using this domain. We have that $r_{i+1} = 2\gamma_i'/\xi = 2\gamma/\xi + O(\sigma_0\sqrt{\varepsilon}/\xi + \sigma_1\sqrt{\varepsilon}r_i/\xi)$. Now if we choose the constant C such that the constant in this $O()$ is $C/4$, then using our assumption that $\xi \geq 2\sigma_1\sqrt{\varepsilon}$, we obtain that

$$r_{i+1} \leq 2\gamma/\xi + C\sigma_0\sqrt{\varepsilon}/4\xi + C\sigma_1\sqrt{\varepsilon}r_i/4\xi \leq 2\gamma/\xi + C\sigma_0\sqrt{\varepsilon}/4 + r_i/4$$

Now if $r_i \geq 8\gamma/\xi + 2C\sigma_0\sqrt{\varepsilon}/\xi$, then we have $r_{i+1} \leq r_i/2$ and if $r_i \leq 8\gamma/\xi + 2C\sigma_0\sqrt{\varepsilon}/\xi$ then we also have $r_{i+1} \leq 8\gamma/\xi + 2C\sigma_0\sqrt{\varepsilon}/\xi$. When r_i is smaller than this we stop and output w_i. Thus we stop in at most $O(\log(r) - \log(8\gamma/\xi + 2C\sigma_0\sqrt{\varepsilon}/\xi)) = O(\log(r\xi/(\gamma + \sigma_0\sqrt{\varepsilon}))$ iterations and have $r_i = O(\gamma/\xi + C\sigma_0\sqrt{\varepsilon})$. But then $\gamma_i' = \gamma + O(\sigma_i'\sqrt{\varepsilon})) \leq \gamma + C(\sigma_0 + \sigma_1 r_i')\sqrt{\varepsilon}/8 = O(\gamma + \sigma_0\sqrt{\varepsilon})$. Using Lemma B.8 we obtain that

$$|\overline{f}(w_i) - \overline{f}(w^*)| \leq 2\gamma_i'^2/\xi = O(\gamma^2/\xi + \sigma_0^2\varepsilon/\xi).$$

as required. The bound on $\|\widehat{w} - w^*\|_2$ follows similarly.

Remark B.1. While we don't give explicit bounds on the number of calls to the approximate learner needed by Sever, such bounds can be straightforwardly obtained under appropriate assumptions on the f_i (see, e.g., the following subsection). Two remarks are in order. First, in this case we cannot take advantage of assumptions that only hold at $\overline{f}$ but might not on the corrupted average f. Second, our algorithm can take advantage of a closed form for the minimum. For example, for the case of linear regression considered in Section E, f_i is not Lipschitz with a small constant if x_i is far from the mean, but there is a simple closed form for the minimum of the least squares loss.

B.4 Proof of Proposition B.5

We let I_{good} be the set of uncorrupted functions f_i. It is then the case that $|I_{\text{good}}| \geq (1-\varepsilon)n$. We need to show that for each $w \in \mathcal{H}$ that

$$\text{Cov}_{i \in I_{\text{good}}}[\nabla f_i(w)] \leq 3\sigma^2 I/4 \tag{3}$$

and

$$\left\| \nabla \overline{f}(w) - \frac{1}{|I_{\text{good}}|} \sum_{i \in I_{\text{good}}} \nabla f_i(w) \right\|_2 \leq O(\sigma^2 \sqrt{\varepsilon}). \tag{4}$$

We will proceed by a cover argument. First we claim that for each $w \in \mathcal{H}$ that (3) and (4) hold with high probability. For Equation (3), it suffices to show that for each unit vector v in a cover $\mathcal{N}$ of size $2^{O(d)}$ of the sphere that

$$\mathbb{E}_{i \in I_{\text{good}}}[(v \cdot (\nabla f_i(w) - \overline{f}))^2] \leq 2\sigma^2/3. \tag{5}$$

However, we note that

$$\mathbb{E}_{p^*}[(v \cdot (\nabla f(w) - \overline{f}))^2] \leq \sigma^2/2.$$

Since $|v \cdot (\nabla f(w) - \overline{f})|$ is always bounded by L, Equation (5) holds for each v, w with probability at least $1 - \exp(-\Omega(n\sigma^2/L^2))$ by a Chernoff bound (noting that the removal of an ε-fraction of points cannot increase this by much). Similarly, to show Equation 4, it suffices to show that for each such v that

$$\mathbb{E}_{i \in I_{\text{good}}}[(v \cdot (\nabla f_i(w) - \overline{f}))] \leq O(\sigma \sqrt{\varepsilon}). \tag{6}$$

Noting that

$$\mathbb{E}_{p^*}[(v \cdot (\nabla f(w) - \overline{f}))] = 0$$

A Chernoff bound implies that with probability $1 - \exp(-\Omega(n\sigma^2\varepsilon/L^2))$ that the average over our original set of f's of $(v \cdot (\nabla f(w) - \overline{f}))$ is $O(\sigma\sqrt{\varepsilon})$. Assuming that Equation (5) holds, removing an ε-fraction of these f's cannot change this value by more than $O(\sigma\sqrt{\varepsilon})$. By union bounding over $\mathcal{N}$ and standard net arguments, this implies that Equations (3) and (4) hold with probability $1 - \exp(\Omega(d - n\sigma^2\varepsilon/L^2))$ for any given w.

To show that our conditions hold for all $w \in \mathcal{H}$, we note that by β-smoothness, if Equation (4) holds for some w, it holds for all other w' in a ball of radius $\sqrt{\sigma^2\varepsilon}/\beta$ (up to a constant multiplicative loss). Similarly, if Equation (3) holds at some w, it holds with bound $\sigma^2 I$ for all w' in a ball of radius $\sigma^2/(2L\beta)$. Therefore, if Equations (3) and (4) hold for all w in a $\min(\sqrt{\sigma^2\varepsilon}/\beta, \sigma/(2L\beta))$-cover of $\mathcal{H}$, the assumptions of Theorem B.2 will hold everywhere. Since we have such covers of size $\exp(O(d \log(r\beta L/(\sigma^2\varepsilon))))$, by a union bound, this holds with high probability if

$$n = \Omega\left(\frac{dL^2 \log(r\beta L/\sigma^2\varepsilon)}{\sigma^2\varepsilon}\right),$$

as claimed.

C Analysis of Sever for GLMs

A case of particular interest is that of Generalized Linear Models (GLMs):

Definition C.1. Let $\mathcal{H} \subseteq \mathbb{R}^d$ and $\mathcal{Y}$ be an arbitrary set. Let D_{xy} be a distribution over $\mathcal{H} \times \mathcal{Y}$. For each $Y \in \mathcal{Y}$, let $\sigma_Y : \mathbb{R} \to \mathbb{R}$ be a convex function. The *generalized linear model* (GLM) over $\mathcal{H} \times \mathcal{Y}$ with *distribution* D_{xy} and *link functions* σ_Y is the function $\overline{f} : \mathbb{R}^d \to \mathbb{R}$ defined by $\overline{f}(w) = \mathbb{E}_{X,Y}[f_{X,Y}(w)]$, where

$$f_{X,Y}(w) := \sigma_Y(w \cdot X)\ .$$

A *sample* from this GLM is given by $f_{X,Y}(w)$ where $(X, Y) \sim D_{xy}$.

Our goal, as usual, is to approximately minimize $\overline{f}$ given ε-corrupted samples from D_{xy}. Throughout this section we assume that $\mathcal{H}$ is contained in the ball of radius r around 0, i.e. $\mathcal{H} \subseteq B(0, r)$. Moreover, we will let $w^* = \arg\min_{w \in \mathcal{H}} \overline{f}(w)$ be a minimizer of $\overline{f}$ in $\mathcal{H}$.

This case covers a number of interesting applications, including SVMs and logistic regression. Unfortunately, the tools developed in Appendix B do not seem to be able to cover this case in a simple manner.

In particular, it is unclear how to demonstrate that Assumption B.1 holds after taking polynomially many samples from a GLM. To rectify this, in this section, we demonstrate a different deterministic regularity condition under which we show SEVER succeeds, and we show that this condition holds after polynomially many samples from a GLM. Specifically, we will show that SEVER succeeds under the following deterministic condition:

Assumption C.1. Fix $0 < \varepsilon < 1/2$. There exists an unknown set $I_{\text{good}} \subseteq [n]$ with $|I_{\text{good}}| \geq (1-\varepsilon)n$ of "good" functions $\{f_i\}_{i \in I_{\text{good}}}$ and parameters $\sigma_0, \sigma_2 \in \mathbb{R}_+$ such that such that the following conditions simultanously hold:

- Equation (1) holds with $\sigma_1 = 0$ and the same σ_0, and
- The following equations hold:

$$\|\nabla \hat{f}(w^*) - \nabla \overline{f}(w^*)\|_2 \leq \sigma_0 \sqrt{\varepsilon} \ \text{, and} \tag{7}$$

$$|\hat{f}(w) - \overline{f}(w)| \leq \sigma_2 \sqrt{\varepsilon}, \ \text{for all } w \in \mathcal{H} \ , \tag{8}$$

  where $\hat{f} \stackrel{\text{def}}{=} \frac{1}{|I_{\text{good}}|} \sum_{i \in I_{\text{good}}} f_i$.

In this section, we will show the following two statements. The first demonstrates that Assumption C.1 implies that SEVER succeeds, and the second shows that Assumption C.1 holds after polynomially many samples from a GLM. Formally:

Theorem C.2. *For functions $f_1, \ldots, f_n : \mathcal{H} \to \mathbb{R}$, suppose that Assumption C.1 holds and that $\mathcal{H}$ is convex. Then, for some universal constant ε_0, if $\varepsilon < \varepsilon_0$, there is an algorithm which, with probability at least $9/10$, finds a $w \in \mathcal{H}$ such that*

$$\overline{f}(w) - \overline{f}(w^*) = r(\gamma + O(\sigma_0 \sqrt{\varepsilon})) + O(\sigma_2 \sqrt{\varepsilon}) \ .$$

If the link functions are ξ-strongly convex, the algorithm finds a $w \in \mathcal{H}$ such that

$$\overline{f}(w) - \overline{f}(w^*) = 2\frac{(\gamma + O(\sigma_0 \sqrt{\varepsilon}))^2}{\xi} + O(\sigma_2 \sqrt{\varepsilon}) \ .$$

Proposition C.3. *Let $\mathcal{H} \subseteq \mathbb{R}^d$ and let $\mathcal{Y}$ be an arbitrary set. Let $f_1, \ldots, f_n$ be obtained by picking f_i i.i.d. at random from a GLM $\overline{f}$ over $\mathcal{H} \times \mathcal{Y}$ with distribution D_{xy} and link functions σ_Y, where*

$$n = \Omega\left(\frac{d \log(dr/\varepsilon)}{\varepsilon}\right) \ .$$

Suppose moreover that the following conditions all hold:

1. *$E_{X \sim D_{xy}}[XX^T] \preceq I$,*
2. *$|\sigma_Y'(t)| \leq 1$ for all $Y \in \mathcal{Y}$ and $t \in \mathbb{R}$, and*
3. *$|\sigma_Y(0)| \leq 1$ for all $Y \in \mathcal{Y}$.*

Then with probability at least $9/10$ over the original set of samples, there is a set of $(1-\varepsilon)n$ of the f_i that satisfy Assumption C.1 on $\mathcal{H}$ with $\sigma_0 = 2$, $\sigma_1 = 0$ and $\sigma_2 = 1 + r$.and $\sigma_2 = 1 + r$.

C.1 Proof of Theorem C.2

As before, since SEVER either terminates or throws away at least one sample, clearly it cannot run for more than n iterations. Thus the runtime bound is simple, and it suffices to show correctness.

We first prove the following lemma:

Lemma C.4. *Let $f_1, \dots, f_n$ satisfy Assumption C.1. Then with probability at least 9/10, SEVER applied to $f_1, \dots, f_n, \sigma_0$ returns a point $w \in \mathcal{H}$ which is a $(\gamma + O(\sigma_0\sqrt{\varepsilon}))$-approximate critical point of $\hat{f}$.*

Proof. We claim that the empirical distribution over $f_1, \dots, f_n$ satisfies Assumption B.1 for the function $\hat{f}$ with σ_0 as stated and $\sigma_1 = 0$, with the I_{good} in Assumption B.1 being the same as in the definition of Assumption C.1. Clearly these functions satisfy (2) (since the LHS is zero), so it suffices to show that they satisfy (1) Indeed, we have that for all $w \in \mathcal{H}$,

$$\mathbb{E}_{I_{\text{good}}}[(\nabla f_i(w) - \nabla\hat{f}(w))(\nabla f_i(w) - \nabla\hat{f}(w))^\top] \preceq \mathbb{E}_{I_{\text{good}}}[(\nabla f_i(w) - \nabla\overline{f}(w))(\nabla f_i(w) - \nabla\overline{f}(w))^\top] ,$$

so they satisfy (1), since the RHS is bounded by Assumption C.1. Thus this lemma follows from an application of Theorem B.2. □

With this critical lemma in place, we can now prove Theorem C.2:

Proof of Theorem C.2. Condition on the event that Lemma C.4 holds, and let $w \in \mathcal{H}$ be the output of SEVER. By Assumption C.1, we know that $\hat{f}(w^*) \geq \overline{f}(w^*) - \sigma_2\sqrt{\varepsilon}$, and moreover, w^* is a $\gamma + \sigma_0\sqrt{\varepsilon}$-approximate critical point of $\hat{f}$.

Since each link function is convex, so is $\hat{f}$. Hence, by Lemma B.8, since w is a $(\gamma + O(\sigma_0\sqrt{\varepsilon}))$-approximate critical point of $\hat{f}$, we have $\hat{f}(w) - \hat{f}(w^*) \leq r(\gamma + O(\sigma_0\sqrt{\varepsilon}))$. By Assumption B.1, this immediately implies that $\overline{f}(w) - \overline{f}(w^*) \leq r(\gamma + O(\sigma_0\sqrt{\varepsilon})) + O(\sigma_2\sqrt{\varepsilon})$, as claimed.

The bound for strongly convex functions follows from the exact argument, except using the statement in Lemma B.8 pertaining to strongly convex functions. □

C.2 Proof of Proposition C.3

Proof. We first note that $\nabla f_{X,Y}(w) = X\sigma'_Y(w \cdot X)$. Thus, under Assumption C.1, we have for any v that

$$\mathbb{E}_i[(v \cdot (\nabla f_i(w) - \nabla\overline{f}(w)))^2] \ll \mathbb{E}_i[(v \cdot \nabla f_i(w))^2] + 1 \ll \mathbb{E}_i[(v \cdot X_i)^2] + 1 .$$

In particular, since this last expression is independent of w, we only need to check this single matrix bound.

We let our good set be the set of samples with $|X| \leq 80\sqrt{d/\varepsilon}$ that were not corrupted. We use Lemma A.18 of [DKK+17]. This shows that with 90% probability that the non-good samples make up at most an $\varepsilon/2 + \varepsilon/160$-fraction of the original samples, and that $\mathbb{E}[XX^T]$ over the good samples is at most $2I$. This proves that the spectral bound holds everywhere. Applying it to the $\nabla f_{X,Y}(w^*)$, we find also with 90% probability that the expectation over all samples of $\nabla f_{X,Y}(w^*)$ is within $\sqrt{\varepsilon}/3$ of $\nabla\overline{f}(w^*)$. Additionally, throwing away the samples with $|\nabla f_{X,Y}(w^*) - \nabla\overline{f}(w^*)| > 80\sqrt{d/\varepsilon}$ changes this by at most $\sqrt{\varepsilon}/2$. Finally, it also implies that the variance of $\nabla f_{X,Y}(w^*)$ is at most $3/2I$, and therefore, throwing away any other ε-fraction of the samples changes it by at most an additional $\sqrt{3\varepsilon/2}$.

We only need to show that $|\mathbb{E}_{i\ \text{good}}[f_i(w)] - \mathbb{E}_X[f_X(w)]| \leq \sqrt{\varepsilon}$ for all $w \in \mathcal{H}$. For this we note that since the f_X and f_i are all 1-Lipschitz, it suffices to show that $|\mathbb{E}_{i\ \text{good}}[f_i(w)] - \mathbb{E}_X[f_X(w)]| \leq (1 + |w|)\sqrt{\varepsilon}/2$ on an $\varepsilon/2$-cover of $\mathcal{H}$. For this it suffices to show that the bound will hold pointwise except with probability $\exp(-\Omega(d\log(r/\varepsilon)))$. We will want to bound this using pointwise concentration and union bounds, but this runs into technical problems since very large values of $X \cdot w$ can lead to large values of f, so we will need to make use of the condition above that the average of $X_iX_i^T$ over our good samples is bounded by $2I$. In particular, this implies that the contribution to the average of $f_i(w)$ over the good i coming from samples where $|X_i \cdot w| \geq 10|w|/\sqrt{\varepsilon}$ is at most $\sqrt{\varepsilon}(1 + |w|)/10$. We consider the average of $f_i(w)$ over the remaining i. Note that these values are uniform random samples from $f_X(w)$ conditioned on $|X| \leq 80\sqrt{d/\varepsilon}$ and

$|X_i \cdot w| < 10|w|/\sqrt{\varepsilon}$. It will suffices to show that taking n samples from this distribution has average within $(1+|w|)\sqrt{\varepsilon}/2$ of the mean with high probability. However, since $|f_X(w)| \leq O(1+|X \cdot w|)$, we have that over this distribution $|f_X(w)|$ is always $O(1+|w|)/\sqrt{\varepsilon}$, and has variance at most $O(1+|w|)^2$. Therefore, by Bernstein's Inequality, the probability that n random samples from $f_X(w)$ (with the above conditions on X) differ from their mean by more than $(1+|w|)\sqrt{\varepsilon}/2$ is

$$\exp(-\Omega(n^2(1+|w|)^2\varepsilon/((1+|w|)^2+n(1+|w|)^2))) = \exp(-\Omega(n\varepsilon)).$$

Thus, for n at least a sufficiently large multiple of $d\log(dr/\varepsilon)/\varepsilon$, this holds for all w in our cover of $\mathcal{H}$ with high probability. This completes the proof. □

D An Alternative Algorithm: Robust Filtering in Each Iteration

In this section, we describe another algorithm for robust stochastic optimization. This algorithm uses standard robust mean estimation techniques to compute approximate gradients pointwise, which it then feeds into a standard projective gradient descent algorithm. This algorithm in practice turns out to be somewhat slower than the one employed in the rest of this paper, because it employs a filtering algorithm at every step of the projective gradient descent, and does not remember which points were filtered between iterations. On the other hand, we present this algorithm for two reasons. Firstly, because it is a conceptually simpler interpretation of the main ideas of this paper, and secondly, because the algorithm works under somewhat more general assumptions. In particular, this algorithm only requires that for each $w \in \mathcal{H}$ that there is a corresponding good set of functions, rather than that there exists a single good set that works simultaneously for all w.

In particular, we can make do with the following somewhat weaker assumption:

Assumption D.1. Fix $0 < \varepsilon < 1/2$ and parameter $\sigma \in \mathbb{R}_+$. For each $w \in \mathcal{H}$, there exists an unknown set $I_{\text{good}} = I_{\text{good}}(w) \subseteq [n]$ with $|I_{\text{good}}| \geq (1-\varepsilon)n$ of "good" functions $\{f_i\}_{i \in I_{\text{good}}}$ such that:

$$\left\| \mathbb{E}_{I_{\text{good}}} \left[\left(\nabla f_i(w) - \nabla \overline{f}(w)\right) \left(\nabla f_i(w) - \nabla \overline{f}(w)\right)^T \right] \right\|_2 \leq (\sigma)^2 , \tag{9}$$

and

$$\|\nabla \hat{f}(w) - \nabla \overline{f}(w)\|_2 \leq \sigma\sqrt{\varepsilon}, \text{ where } \hat{f} \stackrel{\text{def}}{=} \frac{1}{|I_{\text{good}}|} \sum_{i \in I_{\text{good}}} f_i . \tag{10}$$

We make essential use of the following result, which appears in both [DKK+17, SCV18]:

Theorem D.2. *Let $\mu \in \mathbb{R}^d$ and a collection of points $x_i \in \mathbb{R}^d$, $i \in [n]$ and $\sigma > 0$. Suppose that there exists $I_{\text{good}} \subseteq [n]$ with $|I_{\text{good}}| \geq (1-\varepsilon)n$ satisfying the following:*

$$\frac{1}{|I_{\text{good}}|} \sum_{i \in I_{\text{good}}} (x_i - \mu)(x_i - \mu)^\top \preceq \sigma^2 I \text{ and } \left\| \frac{1}{|I_{\text{good}}|} \sum_{i \in I_{\text{good}}} (x_i - \mu) \right\|_2 \leq \sigma\sqrt{\varepsilon}. \tag{11}$$

Then, if $\varepsilon < \varepsilon_0$ for some universal constant ε_0, there is an efficient algorithm, Algorithm $\mathcal{A}$, which outputs an estimate $\hat{\mu} \in \mathbb{R}^d$ such that $\|\hat{\mu} - \mu\|_2 = O(\sigma\sqrt{\varepsilon})$.

Our general robust algorithm for stochastic optimization will make calls to Algorithm $\mathcal{A}$ in a black-box manner, as well as to the projection operator onto $\mathcal{H}$. We will measure the cost of our algorithm by the total number of such calls.

Remark D.1. While it is not needed for the theoretical results established in this subsection, we note that the robust mean estimation algorithm of [DKK+17] relies on an iterative outlier removal method only requiring basic eigenvalue computations (SVD), while the [SCV18] algorithm employs semidefinite programming. In our experiments, we use the algorithm in [DKK+17] and variants thereof.

Using the above black-box, together with known results on convex optimization with errors, we obtain the following meta-theorem:

Theorem D.3. *For functions $f_1, \ldots, f_n : \mathcal{H} \to \mathbb{R}$, bounded below on a closed domain $\mathcal{H}$, suppose that either Assumption D.1 is satisfied with some parameters $\varepsilon, \sigma > 0$. Then there exists an efficient algorithm that finds an $O(\sigma\sqrt{\varepsilon})$-approximate critical point of $\overline{f}$.*

Proof. We note that by applying Algorithm A on $\{\nabla f_i(w)\}$, we can find an approximation to $\nabla\overline{f}(w)$ with error $O(\sigma\sqrt{\varepsilon})$. We note that standard projective gradient descent algorithms can be made to run efficiently even if the gradients given are only approximate, and this can be used to find our $O(\sigma\sqrt{\varepsilon})$-approximate critical point. □

E Applications of the General Algorithm

In this section, we present three concrete applications of our general robust algorithm. In particular, we describe how to robustly optimize models for linear regression, support vector machines, and logistic regression, in Sections E.1, E.2, E.3, respectively.

E.1 Linear Regression

In this section, we demonstrate how our results apply to linear regression. We are given pairs $(X_i, Y_i) \in \mathbb{R}^d \times \mathbb{R}$ for $i \in [n]$. The X_i's are drawn i.i.d. from a distribution D_x, and $Y_i = \langle w^*, X_i\rangle + e_i$, for some unknown $w^* \in \mathbb{R}^d$ and the noise random variables e_i's are drawn i.i.d. from some distribution D_e. Given $(X_i, Y_i) \sim D_{xy}$, the joint distribution induced by this process, let $f_i(w) = (Y_i - \langle w, X_i\rangle)^2$. The goal is then to find a $\widehat{w}$ approximately minimizing the objective function

$$\overline{f}(w) = \mathbb{E}_{(X,Y)\sim D_{xy}}[(Y - \langle w, X\rangle)^2] \,.$$

We work with the following assumptions:

Assumption E.1. Given the model for linear regression described above, assume the following conditions for D_e and D_x:

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

Our main result for linear regression is the following:

Theorem E.2. *Let $\varepsilon > 0$, and let D_{xy} be a distribution over pairs (X, Y) which satisfies the conditions of Assumption E.1. Suppose we are given $O\left(\frac{d^5}{\varepsilon^2}\right)$ ε-noisy samples from D_{xy}. Then in either of the following two cases, there exists an algorithm that, with probability at least $9/10$, produces a $\widehat{w}$ with the following guarantees:*

1. *If $\mathbb{E}_{X\sim D_x}[XX^T] \succeq \gamma I$ for $\gamma = \Omega(\sqrt{\varepsilon})$, then $\overline{f}(\widehat{w}) \le \overline{f}(w^*) + O\left(\frac{(\xi+\varepsilon)\varepsilon}{\gamma}\right)$ and $\|\widehat{w} - w^*\|_2 = O\left(\frac{\sqrt{\xi\varepsilon}+\varepsilon}{\gamma}\right)$.*
2. *If $\|w^*\|_2 \le r$, then $\overline{f}(\widehat{w}) \le \overline{f}(w^*) + O(((\sqrt{\xi} + \sqrt{\varepsilon})r + \sqrt{C}r^2)\sqrt{\varepsilon})$.*

The proof will follow from two lemmas (proved in Section E.1.1 and E.1.2, respectively). First, we will bound the covariance of the gradient, in Lemma E.3:

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

$$v^\top \operatorname{Cov}_{(X,Y)\sim D_{xy}} \left[\nabla f_i(w,(X,Y))\right] v \le 4\sigma^2\xi + 4C\sigma^4\|w^* - w\|_2^2 \,.$$

With this in hand, we can prove Lemma E.4, giving us a polynomial sample complexity which is sufficient to satisfy the conditions of Assumption B.1.

Lemma E.4. *Suppose D_{xy} satisfies the conditions of Assumption E.1. Given $O(d^5/\varepsilon^2)$ ε-noisy samples from D_{xy}, then with probability at last 9/10, they satisfy Assumption B.1 with parameters $\sigma_0 = 30\sqrt{\xi} + \sqrt{\varepsilon}$ and $\sigma_1 = 18\sqrt{C+1}$.*

The proof concludes by applying Corollary B.4 or case (i) of Corollary B.3 for the first and second cases respectively.

E.1.1 Proof of Lemma E.3

Note that for this setting we have that $f(w,z) = f(w,x,y) = (y - \langle w, x\rangle)^2$. We then have that $\nabla_w f(w,z) = -2(\langle w^* - w, x\rangle + e)x$. Our main claim is the following:

Claim E.5. We have that $\operatorname{Cov}[\nabla_w f(w,z)] = 4\,\mathbb{E}_{X\sim D}\left[\langle w^* - w, x\rangle^2 (xx^T)\right] + 4\operatorname{Var}[E]\Sigma - 4\Sigma(w^* - w)(w^* - w)^T\Sigma$.

Proof. Let us use the notation $A = \nabla_w f(w,z)$ and $\mu = \mathbb{E}[A]$. By definition, we have that $\operatorname{Cov}[A] = \mathbb{E}[AA^T] - \mu\mu^T$.

Note that $\mu = \mathbb{E}_z[\nabla_w f(w,z)] = \mathbb{E}_z[(-2\langle w^* - w, x\rangle + e)x] = -2\Sigma(w^* - w)$, where we use the fact that $\mathbb{E}_z[e] = 0$ and e is independent of x. Therefore, $\mu\mu^T = 4\Sigma(w^* - w)(w^* - w)^T\Sigma$.

To calculate $\mathbb{E}[AA^T]$, note that $A = \nabla_w f(w,z) = -2(\langle w^* - w, x\rangle + e)x$, and $A^T = -2(\langle w^* - w, x\rangle + e)x^T$. Therefore, $AA^T = 4(\langle w^* - w, x\rangle^2 + e^2 + 2\langle w^* - w, x\rangle e)(xx^T)$ and

$$\mathbb{E}_z[AA^T] = 4\,\mathbb{E}_x[\langle w^* - w, x\rangle^2 (xx^T)] + 4\operatorname{Var}[e]\Sigma + 0 \,,$$

where we again used the fact that the noise e is independent of x and its expectation is zero.

By gathering terms, we get that

$$\operatorname{Cov}[\nabla_w f(w,z)] = 4\,\mathbb{E}_x[\langle w^* - w, x\rangle^2 (xx^T)] + 4\operatorname{Var}[e]\Sigma - 4\Sigma(w^* - w)(w^* - w)^T\Sigma \,.$$

This completes the proof. □

Given the above claim, we can bound from above the spectral norm of the covariance matrix of the gradients as follows: Specifically, for a unit vector v, the quantity $v^T \operatorname{Cov}[\nabla_w f(w,z)]v$ is bounded from above by a constant times the following quantities:

- The first term is $v^T\,\mathbb{E}_x[\langle w^* - w, x\rangle^2 (xx^T)]v = \mathbb{E}_x[\langle w^* - w, x\rangle^2 \cdot \langle v, x\rangle^2)]$. By Cauchy-Schwarz and our 4th moment bound, this is at most $C\sigma^4\|w^* - w\|_2^2$, where $\Sigma \preceq \sigma^2 I$.
- The second term is at most the upper bound of the variance of the noise ξ times σ^2.
- The third term is at most $v^T\Sigma(w^* - w)(w^* - w)^T\Sigma v$, which by our bounded covariance assumption is at most $\sigma^4\|w^* - w\|_2^2$.

This gives the parameters in the meta-theorem.

E.1.2 Proof of Lemma E.4

Let S be the set of uncorrupted samples and I be the subset of S with $\|X\|_2 \le 2\sqrt{d}/\varepsilon^{1/4}$. We will take I_{good} to be the subset of I that are not corrupted.

Firstly, we show that with probability at least $39/40$, at most an $\varepsilon/2$-fraction of points in S have $\|X\|_2 > 2\sqrt{d}/\varepsilon^{1/4}$, and so $|I_{\text{good}}| \ge (1-\varepsilon)|S|$. Note that $\mathbb{E}_D[\|X\|_2^4] = \mathbb{E}_D[(\sum_{j=1}^d X_j^2)^2] \le \sum_{j=1}^d \sum_{k=1}^d \sqrt{\mathbb{E}_D[X_j^2]\,\mathbb{E}[X_k^2]} \le Cd^2$, since $\mathbb{E}_{D_x}[XX^T] \preceq I$. Thus, by Markov's inequality, $\Pr_D[\|X\|_2 > 2\sqrt{d}(C/\varepsilon)^{1/4}] = \Pr_D[\|X\|_2^4 > 16d^2/\varepsilon] \le \varepsilon/16$. By a Chernoff bound, since $n \ge 10\varepsilon^2$ this probability is at most $\varepsilon/2$ for the uncorrupted samples with probability at least $39/40$.

Next, we show that (1) holds with probability at least $39/40$. To do this, we will apply Lemma E.3 to I_{good}. Since S consists of independent samples, the variance over the randomness of S of $|S|\,\mathbb{E}_S[e^2]$ is at most $|S|\xi$. By Chebyshev's inequality, except with probability $1/99$, we have that $\mathbb{E}_S[e^2] \le 99\xi$ and since $I_{\text{good}} \subset S$, $\mathbb{E}_{I_{\text{good}}}[e^2] \le |S|\,\mathbb{E}_S[e^2]/|I| \le 100\xi$. This is condition (i) of Lemma E.3.

We note that I consists of $\Omega(d^5/\varepsilon^2)$ independent samples from D conditioned on $\|X\|_2 < 2\sqrt{d}/\varepsilon^{1/4}$, a distribution that we will call D'. Since the VC-dimension of all halfspaces in $\mathbb{R}^d$ is $d+1$, by the VC inequality, we have that, except with probability $1/80$, for any unit vector v and $T \in \mathbb{R}$ that $|\Pr_I[v\cdot X > T] - \Pr_{D'}[v\cdot X > T]| \le \varepsilon/d^2$. Note that for unit vector v and positive integer m, $\mathbb{E}[(v.X)^m] = \int_0^\infty m(v\cdot X)^{m-1}\Pr[v\cdot X > T]dT$. Thus we have that

$$\begin{aligned}
\mathbb{E}_I[(v.X)^m] &= \int_0^\infty m(v\cdot X)^{m-1}\Pr_I[v\cdot X > T]dT \\
&\le \int_0^{2d^{1/2}(C/\varepsilon)^{1/4}} m(v\cdot X)^{m-1}(\Pr_{D'}[v\cdot X > T] + \varepsilon/d^2)dT \\
&= \mathbb{E}_{D'}[(v.X)^m] + (2d^{1/2}(C/\varepsilon)^{1/4})^m(\varepsilon/d^2) \\
&\le (1+\varepsilon)\,\mathbb{E}_D[(v.X)^m] + 2^m C^{m/4}(\varepsilon/d^2)^{1-m/4}\,.
\end{aligned}$$

Applying this for $m=2$ gives $\mathbb{E}_I[XX^T] \preceq (1+\varepsilon+4\sqrt{C}\varepsilon/d^2)I \preceq 2I$ and with $m=4$ gives $\mathbb{E}_I[(v.X)^4] \le (1+\varepsilon)C + 16C$. Similar bounds apply to I_{good}, with an additional $1+\varepsilon$ factor.

Thus, with probability at least $39/40$, I_{good} satisfies the conditions of Lemma E.3 with $\xi := 100\xi$, $\sigma^2 := 2$ and $C := 5C$. Hence, it satisfies (1) with $\sigma_0 = 20\sqrt{\xi}$ and $\sigma_1 = 18\sqrt{C+1}$.

For (2), note that $\nabla_w f_i(w) = (w\cdot x_i - y_i)x_i = ((w-w^*)\cdot x_i)x_i - e_i x_i$. We will separately bound $\|\mathbb{E}_{I_{\text{good}}}[((w-w^*)\cdot X)X] - \mathbb{E}_D[((w-w^*)\cdot X)X]\|_2$ and $\|\mathbb{E}_{I_{\text{good}}}[eX] - \mathbb{E}_D[eX]\|_2$.

We will repeatedly make use of the following, which bounds how much removing points or probability mass affects an expectation in terms of its variance:

Claim E.6. For a mixture of distributions $P = (1-\delta)Q + \delta R$ for distributions P, Q, R and a real valued function f, we have that $|\mathbb{E}_{X\sim P}[f(X)] - \mathbb{E}_{X\sim Q}[f(X)]| \le 2\sqrt{\delta\,\mathbb{E}_{X\sim P}[f(X)^2]}/(1-\delta)$

Proof. By Cauchy-Schwarz $|\mathbb{E}_{X\sim R}[f(X)]| \le \sqrt{\mathbb{E}_{X\sim R}[f(X)^2]} \le \sqrt{\mathbb{E}_{X\sim P}[f(X)^2]/\delta}$. Since $\mathbb{E}_{X\sim P}[f(X)] = (1-\delta)\,\mathbb{E}_{X\sim Q}[f(X)] + \delta\,\mathbb{E}_{X\sim R}[f(X)]$, this implies that $|\mathbb{E}_{X\sim P}[f(X)]/(1-\delta) - \mathbb{E}_{X\sim Q}[f(X)]| \le \sqrt{\delta\,\mathbb{E}_{X\sim P}[f(X)^2]}/(1-\delta)$. However $|\mathbb{E}_{X\sim P}[f(X)]/(1-\delta) - \mathbb{E}_{X\sim P}[f(X)]| = (\delta/(1-\delta))|\mathbb{E}_{X\sim P}[f(X)]| \le \sqrt{\delta\,\mathbb{E}_{X\sim P}[f(X)^2]}/(1-\delta)$ and the triangle inequality gives the result. □

We can apply this to $P = I$ and $Q = I_{\text{good}}$ with $\delta = \varepsilon/2$ and also to $P = D$ and $Q = D'$ with $\delta = \varepsilon/16$, with error $2\sqrt{\delta}/(1-\delta) \le 2\sqrt{\varepsilon}$ in either case.

For the first of term we wanted to bound, we have $\|\mathbb{E}_{I_{\text{good}}}[((w-w^*)\cdot X)X] - \mathbb{E}_D[((w-w^*)\cdot X)X]\|_2 = \|(w-w^*)^T\left(\mathbb{E}_{I_{\text{good}}}[XX^T] - \mathbb{E}_D[XX^T]\right)\|_2 \le \|w-w^*\|_2 \|\mathbb{E}_{I_{\text{good}}}[XX^T] - \mathbb{E}_D[XX^T]\|_2$. For any unit vector v, the VC dimension argument above gave that $|\mathbb{E}_I[(v\cdot X)^2] - \mathbb{E}_{D'}[(v\cdot X)^2)]| \le 4\sqrt{C}\varepsilon/d^2$ and Claim E.6 both gives that $|\mathbb{E}_I[(v\cdot X)^2] - \mathbb{E}_{I_{\text{good}}}[(v\cdot X)^2)]| \le 2\sqrt{\varepsilon\,\mathbb{E}_I[(v.X)^4]} \le 10\sqrt{C\varepsilon}$ and that $|\mathbb{E}_D[(v\cdot X)^2] - \mathbb{E}_{D'}[(v\cdot X)^2)]| \le 2\sqrt{\varepsilon\,\mathbb{E}_D[(v.X)^4]} \le 2\sqrt{C\varepsilon}$. By the triangle inequality, we have that $|\mathbb{E}_D[(v\cdot X)^2] - \mathbb{E}_{I_{\text{good}}}[(v\cdot X)^2)]| \le 16\sqrt{C\varepsilon}$. Since this holds for all unit v and the matrices involved are symmetric, we have that $\|\mathbb{E}_{I_{\text{good}}}[XX^T] -$

$\mathbb{E}_D[XX^T]\|_2 \le 16\sqrt{C\varepsilon}$. The overall first term is bounded by $\|\mathbb{E}_{I_{\text{good}}}[((w-w^*)\cdot X)X] - \mathbb{E}_D[((w-w^*)\cdot X)X]\|_2 \le 16\sqrt{C\varepsilon}\|w-w^*\|_2$.

Now we want to bound the second term, $\|\mathbb{E}_{I_{\text{good}}}[eX] - \mathbb{E}_D[eX]\|_2$. Note that $\mathbb{E}_D[eX] = \mathbb{E}_D[e]\,\mathbb{E}_D[X] = 0$. So we need to bound $\mathbb{E}_{I_{\text{good}}}[eX]$. First we bound the expectation and variance on D' using Claim E.6. It yields that, for any unit vector v, $|\mathbb{E}_{D'}[e(v\cdot X)]| \le 2\sqrt{\varepsilon\,\mathbb{E}_D[e^2(v\cdot X)^2]} \le 2\sqrt{\varepsilon\xi}$.

Next we bound the expectation on I. Since I consists of independent samples from D', the covariance matrix over the randomness on I of $|I|\,\mathbb{E}_I[eX - \mathbb{E}_{D'}[eX]]$ is $|I|\,\mathbb{E}_{D'}[(eX-\mathbb{E}_{D'}[eX])(eX-\mathbb{E}_{D'}[eX])^T] \le |I|\,\mathbb{E}_{D'}[XX^T] \le |I|(1+\varepsilon)I$ and its expectation is 0. Thus the expectation over the randomness of I of $(|I|^2\|\mathbb{E}_I[eX] - \mathbb{E}_{D'}[eX]\|_2)^2$ is $\mathrm{Tr}(|I|\,\mathbb{E}_{D'}[(eX-\mathbb{E}_{D'}[eX])(eX-\mathbb{E}_{D'}[eX])^T]) \le |I|(1+\varepsilon+4\xi\varepsilon|I|)d$. By Markov's inequality, except with probability $1/40$, $\Pr[\|\mathbb{E}_I[eX]\|_2 \ge 2\sqrt{\xi\varepsilon}+\varepsilon] \le d/|I|\varepsilon^2$. Since $|I| \ge 40d/\varepsilon^2$. This happens with probability at least $1/40$.

Next we bound the expectation on I_{good} which follows by a slight variation of Claim E.6. Let $J = I - I_{\text{good}}$. Then, for any v, $\mathbb{E}_J[e(v\cdot X)] \le \sqrt{\mathbb{E}_J[e^2]\,\mathbb{E}_J[(v\cdot X)^2]} \le \sqrt{\mathbb{E}_S[e^2]\,\mathbb{E}_I[(v\cdot X)^2]}|J|/\sqrt{|S||I|} \le \sqrt{100\xi(1+\varepsilon+4\sqrt{C}\varepsilon/d^2)}|J|/\sqrt{|S||I|} \le 20|J|\sqrt{\xi/|S||I|}$ by bounds we obtained earlier. Now $\|\mathbb{E}_{I_{\text{good}}}[eX]\|_2 = \|(|I|/|I_{\text{good}}|)\,\mathbb{E}_I[eX] - (|J|/|I_{\text{good}}|)\,\mathbb{E}_J[eX]\|_2 \le 20\sqrt{\xi\varepsilon}+\varepsilon+(1+\varepsilon)\sqrt{\xi\varepsilon/16} \le 30\sqrt{\xi\varepsilon}+\varepsilon$.

We can thus take $\sigma_0 = 30\sqrt{\xi}+\sqrt{\varepsilon}$ and $\sigma_1 = 18\sqrt{C+1} \ge 16\sqrt{C}$ to get (2).

To get both (2) and (1) hold with $\sigma_0 = 30\sqrt{\xi}+\sqrt{\varepsilon}$ and $\sigma_1 = 18\sqrt{C+1}$). This happens with probability at least $9/10$ by a union bound on the probabilistic assumptions above.

E.2 Support Vector Machines

In this section, we demonstrate how our results apply to learning support vector machines (i.e., halfspaces under hinge loss). In particular, we describe how SVMs fit into the GLM framework described in Section C.

We are given pairs $(X_i, Y_i) \in \mathbb{R}^d \times \{\pm 1\}$ for $i \in [n]$, which are drawn from some distribution D_{xy}. Let $L(w,(x,y)) = \max\{0, 1-y(w\cdot x)\}$, and $f_i(w) = L(w,(x_i,y_i))$. The goal is to find a $\widehat{w}$ approximately minimizing the objective function

$$\overline{f}(w) = \mathbb{E}_{(X,Y)\sim D_{xy}}[L(w,(X,Y))].$$

One technical point is that f_i does not have a gradient everywhere – instead, we will be concerned with the sub-gradients of the f_i's. All our results which operate on the gradients also work for sub-gradients. To be precise, we will take the sub-gradient to be 0 when the gradient is undefined:

Definition E.1. Let ∇f_i be the *sub-gradient* of $f_i(w)$ with respect to w, where $\nabla f_i = -y_i x_i$ if $y_i(w\cdot x_i) < 1$, and 0 otherwise.

To get a bound on the error of hinge loss, we will need to assume the marginal distribution D_x is anti-concentrated.

Definition E.2. A distribution is δ-*anticoncentrated* if at most an $O(\delta)$-fraction of its probability mass is within Euclidean distance δ of any hyperplane.

We work with the following assumptions:

Assumption E.7. Given the model for SVMs as described above, assume the following conditions for the marginal distribution D_x:

- $\mathbb{E}_{X\sim D_x}[XX^T] \preceq I$;
- D_x is $\varepsilon^{1/4}$-anticoncentrated.

Our main result on SVMs is the following:

Theorem E.8. *Let $\varepsilon > 0$, and let D_{xy} be a distribution over pairs (X, Y), where the marginal distribution D_x satisfies the conditions of Assumption E.7. Then there exists an algorithm that with probability 9/10, given $O(d \log(d/\varepsilon)/\varepsilon)$ ε-noisy samples from D_{xy}, returns a $\widehat{w}$ such that for any w^*,*

$$\mathbb{E}_{(X,Y)\sim D_{xy}}[L(\widehat{w}, (X, Y))] \leq \mathbb{E}_{(X,Y)\sim D_{xy}}[L(w^*, (X, Y))] + O(\varepsilon^{1/4}).$$

Our approach will be to fit this problem into the GLM framework developed in Section C. First, we will restrict our search over w to $\mathcal{H}$, a ball of radius $r = \varepsilon^{-1/4}$. As we argue in Lemma E.9, this restriction comes at a cost of at most $O(\varepsilon^{1/4})$ in our algorithm's loss. With this restriction, we will argue that the problem satisfies the conditions of Proposition C.3. This allows us to argue that, with a polynomial number of samples, we can obtain a set of f_i's satisfying the conditions of Assumption C.1. This will allow us to apply Theorem C.2, concluding the proof.

We start by showing that, due to anticoncentration of D, there is a $w' \in \mathcal{H}$ with loss close to w^*:

Lemma E.9. *Let w' be a rescaling of w^*, such that $\|w'\|_2 \leq \varepsilon^{-1/4}$ (i.e. $w' = \min\{1, \varepsilon^{-1/4}/\|w^*\|_2\}w^*$). Then $\mathbb{E}_{(X,Y)\sim D_{xy}}[L(w', (X, Y))] \leq \mathbb{E}_{(X,Y)\sim D_{xy}}[L(w^*, (X, Y))] + O(\varepsilon^{1/4})$.*

Proof. If $w' = w^*$, then $\mathbb{E}_{(X,Y)\sim D_{xy}}[L(w', (X, Y))] = \mathbb{E}_{(X,Y)\sim D_{xy}}[L(w^*, (X, Y))]$.

Otherwise, we break into case analysis, based on the value of (x, y):

- $|w' \cdot x| > 1$: If $y(w' \cdot x) > 1$, then $L(w', (x, y)) = L(w^*, (x, y)) = 0$. If $y(w' \cdot x) < -1$, then $L(w', (x, y)) = 1 - y(w' \cdot x) \leq 1 - y(w^* \cdot x) = L(w^*, (x, y))$. Both cases use the fact that $\|w'\|_2 < \|w^*\|_2$.
- $|w' \cdot x| \leq 1$: In this case, we have that $L(w', (x, y)) \leq 2$. Since $L(w^*, (x, y)) \geq 0$, we have that $L(w', (x, y)) \leq L(w^*, (x, y)) + 2$.

Note that if $|w' \cdot x| \leq 1$, then x is within $1/\|w'\|_2 = \varepsilon^{1/4}$ of the hyperplane defined by the normal vector w'. Since D_x is $\varepsilon^{1/4}$-anticoncentrated, we have that $\Pr_{X\sim D_x}[|w' \cdot X| \leq 1] \leq \varepsilon^{1/4}$. Thus, we have that $\mathbb{E}_{(X,Y)\sim D_{xy}}[L(w', (X, Y))] \leq \mathbb{E}_{(X,Y)\sim D_{xy}}[L(w^*, (X, Y)) + 2 \cdot \mathbb{1}(|w' \cdot X| \leq 1)] \leq \mathbb{E}_{(X,Y)\sim D_{xy}}[L(w^*, (X, Y))] + O(\varepsilon^{1/4})$. □

Proof of Theorem E.8. We first show that this problem fits into the GLM framework, in particular, satisfying the conditions of Proposition C.3. The link function is $\sigma_y(t) = \max\{0, 1 - yt\}$, giving us the loss function $L(w, (x, y)) = \sigma_y(w \cdot x)$. We let $\mathcal{H}$ be the set $\|w\|_2 \leq \varepsilon^{-1/4}$, giving us the parameter $r = \varepsilon^{-1/4}$. Condition 1 is satisfied by Assumption E.7. For $y \in \{-1, 1\}$, $\sigma'_y(t) = 0$ for $yt \geq 1$ and $\sigma'_y(t) = -y$ for $yt < 1$. Thus we have that $|\sigma'_1(t)| \leq 1$ for all t and y, satisfying Condition 2. Finally, one can observe that $\sigma_y(0) = 1$ for all y, satisfying Condition 3. Thus we can apply Proposition C.3: if we take $O(d \log(dr/\varepsilon)/\varepsilon)$ ε-corrupted samples, then they satisfy Assumption C.1 on $\mathcal{H}$ with $\sigma_0 = 2$, $\sigma_1 = 0$ and $\sigma_2 = 1 + \varepsilon^{-1/4}$, with probability 9/10.

Now we can apply the algorithm of Theorem C.2. Since the loss is convex, we get a vector $\widehat{w}$ with $\overline{f}(\widehat{w}) - \overline{f}(w^{*\prime}) = O((\sigma_0 r + \sigma_1 r^2 + \sigma_2)\sqrt{\varepsilon}) = O((2\varepsilon^{-1/4} + \varepsilon^{-1/4})\sqrt{\varepsilon}) = O(\varepsilon^{1/4})$ where $w^{*\prime}$ is the minimizer of $\overline{f}$ on $\mathcal{H}$.

We thus have that $\overline{f}(\hat{w}) \leq \overline{f}(w^{*\prime}) + O(\varepsilon^{1/4}) \leq \overline{f}(w') + O(\varepsilon^{1/4}) \leq \overline{f}(w^*) + O(\varepsilon^{1/4})$. The second inequality follows because $w^{*\prime}$ is the minimizer of $\overline{f}$ on $\mathcal{H}$, and the third inequality follows from Lemma E.9. □

E.3 Logistic Regression

In this section, we demonstrate how our results apply to logistic regression. In particular, we describe how logistic regression fits into the GLM framework described in Section C.

We are given pairs $(X_i, Y_i) \in \mathbb{R}^d \times \{\pm 1\}$ for $i \in [n]$, which are drawn from some distribution D_{xy}. Let $\phi(t) = \frac{1}{1+\exp(-t)}$. Logistic regression is the model where $y = 1$ with probability $\phi(w \cdot x)$, and $y = -1$ with probability $\phi(-w \cdot x)$. We define the loss function to be the log-likelihood of y given x. More precisely, we let $f_i(w, (x_i, y_i)) = L(w, (x_i, y_i))$, which is defined as follows:

$$L(w, (x, y)) = \frac{1 + y}{2} \ln\left(\frac{1}{\phi(w \cdot x)}\right) + \frac{1 - y}{2} \ln\left(\frac{1}{\phi(-w \cdot x)}\right) = \frac{1}{2}\left(-\ln\left(\phi(w \cdot x)\phi(-w \cdot x)\right) - y(w \cdot x)\right).$$

The gradient of this function is $\nabla L(w,(x,y)) = \frac{1}{2}(\phi(w\cdot x)-\phi(-w\cdot x)-y)x$. The goal is to find a $\widehat{w}$ approximately minimizing the objective function

$$\overline{f}(w) = \mathbb{E}_{(X,Y)\sim D_{xy}}[L(w,(X,Y))].$$

We work with the following assumptions:

Assumption E.10. Given the model for logistic regression as described above, assume the following conditions for the marginal distribution D_x:

- $\mathbb{E}_{X\sim D_x}[XX^T] \preceq I$;
- D_x is $\varepsilon^{1/4}\sqrt{\log(1/\varepsilon)}$-anticoncentrated.

We can get a similar result to that for hinge loss for logistic regression:

Theorem E.11. *Let $\varepsilon > 0$, and let D_{xy} be a distribution over pairs (X,Y), where the marginal distribution D_x satisfies the conditions of Assumption E.10. Then there exists an algorithm that with probability $9/10$, given $O(d\log(d/\varepsilon)/\varepsilon)$ ε-noisy samples from D_{xy}, returns a $\widehat{w}$ such that for any w^*,*

$$\mathbb{E}_{(X,Y)\sim D_{xy}}[L(\widehat{w},(X,Y))] \le \mathbb{E}_{(X,Y)\sim D_{xy}}[L(w^*,(X,Y))] + O(\varepsilon^{1/4}\sqrt{\log(1/\varepsilon)}).$$

The approach is very similar to that of Theorem E.8, which we repeat here for clarity. First, we will restrict our search over w to $\mathcal{H}$, a ball of radius $r = \varepsilon^{-1/4}\sqrt{\log(1/\varepsilon)}$. As we argue in Lemma E.12, this restriction comes at a cost of at most $O(\varepsilon^{1/4}\sqrt{\log(1/\varepsilon)})$ in our algorithm's loss. With this restriction, we will argue that the problem satisfies the conditions of Proposition C.3. This allows us to argue that, with a polynomial number of samples, we can obtain a set of f_i's satisfying the conditions of Assumption C.1. This will allow us to apply Theorem C.2, concluding the proof.

We start by showing that, due to anticoncentration of D, there is a $w' \in \mathcal{H}$ with loss close to w^*:

Lemma E.12. *Let w' be a rescaling of w^*, such that $\|w'\|_2 \le \varepsilon^{-1/4}\sqrt{\ln(1/\varepsilon)}$ (i.e. $w' = \min\{1, \varepsilon^{-1/4}\sqrt{\ln(1/\varepsilon)}/\|w^*\|_2\}w^*$). Then $\mathbb{E}_{(X,Y)\sim D_{xy}}[L(w',(X,Y))] \le \mathbb{E}_{(X,Y)\sim D_{xy}}[L(w^*,(X,Y))] + O(\varepsilon^{1/4}\sqrt{\ln(1/\varepsilon)})$.*

Proof. We need the following claim:

Claim E.13.

$$|t| \le -\ln(\phi(t)\phi(-t)) \le |t| + 3\exp(-|t|)$$

Proof. Recalling that $\phi = 1/(1+\exp(-t))$, we have that $-\ln(\phi(t)\phi(-t)) = \ln(\exp(t)+\exp(-t)+2)$. Since $\exp(t)+\exp(-t)+2 \ge \exp(|t|)$, we have $|t| \le -\ln(\phi(t)\phi(-t))$. On the other hand, $\ln(\exp(t)+\exp(-t)+2) = |t| + \ln(1+2\exp(-|t|)+\exp(-2|t|)) \le |t| + \ln(1+3\exp(-|t|)) \le |t| + 3\exp(-|t|)$. □

For any $x \in \mathbb{R}^d$, we have that:

$$\begin{aligned} -\ln(\phi(w'\cdot x)\phi(-w'\cdot x)) - y(w'\cdot x) - 3\exp(-3|w'\cdot x|) &\le |w'\cdot x| - y(w'\cdot x) \\ &\le |w^*\cdot x| - y(w^*\cdot x) \\ &\le -\ln(\phi(w^*\cdot x)\phi(-w^*\cdot x)) - y(w^*\cdot x) \end{aligned}$$

The first and last inequality hold by Claim E.13. For the second inequality, we do a case analysis on y. When $y = \operatorname{sign}(w'\cdot x) = \operatorname{sign}(w^*\cdot x)$, then both sides of the inequality are 0. When $y = -\operatorname{sign}(w'\cdot x) = -\operatorname{sign}(w^*\cdot x)$, then the inequality becomes $2|w'\cdot x| \le 2|w^*\cdot x|$, which holds since $\|w'\|_2 \le \|w^*\|_2$. We thus have that for any $y \in \{\pm 1\}$, $L(w',(x,y)) \le L(w^*,(x,y)) + \frac{3}{2}\exp(-3|w'\cdot x|)$. If $|w'\cdot x| \le \frac{1}{3}\ln(1/\varepsilon)$, then $L(w',(x,y)) \le L(w^*,(x,y)) + \frac{3}{2}$. If $|w'\cdot x| \ge \frac{1}{3}\ln(1/\varepsilon)$, then $L(w',(x,y)) \le L(w^*,(x,y)) + \frac{3}{2}\varepsilon$. Since $\|w'\|_2 \le \varepsilon^{-1/4}\sqrt{\ln(1/\varepsilon)}$ and D_x is $\varepsilon^{1/4}\sqrt{\ln(1/\varepsilon)}$-anticoncentrated, we have that $\Pr_{D_x}[|w'\cdot x| \le \frac{1}{3}\ln(1/\varepsilon)] \le O(\varepsilon^{1/4}\sqrt{\ln(1/\varepsilon)})$. Thus, $\mathbb{E}_{(X,Y)\sim D_{xy}}[L(w',(X,Y))] \le \mathbb{E}_{(X,Y)\sim D_{xy}}[L(w^*,(X,Y))] + O(\varepsilon^{1/4}\sqrt{\ln(1/\varepsilon)})$, as desired. □

With this in hand, we can conclude with the proof of Theorem E.11.

Proof of Theorem E.11. We first show that this problem fits into the GLM framework, in particular, satisfying the conditions of Proposition C.3. The link function is $\sigma_y(t) = \frac{1}{2}(-\ln(\phi(t)\phi(-t)) - yt)$, giving us the loss function $L(w,(x,y)) = \sigma_y(w \cdot x)$. We let $\mathcal{H}$ be the set $\|w\|_2 \le \varepsilon^{-1/4}\sqrt{\ln(1/\varepsilon)}$, giving us the parameter $r = \varepsilon^{-1/4}\sqrt{\ln(1/\varepsilon)}$. Condition 1 is satisfied by Assumption E.10. For $y \in \{-1,1\}$, $\sigma_y'(t) = \frac{1}{2}(\phi(t) - \phi(-t) - y)$, which gives that $|\sigma_y'(t)| \le 1$ for all t and y, satisfying Condition 2. Finally, $\sigma_y(0) = \ln 2 < 1$ for all y, satisfying Condition 3. Thus we can apply Proposition C.3: if we take $O(d\log(dr/\varepsilon)/\varepsilon)$ ε-corrupted samples, then they satisfy Assumption C.1 on $\mathcal{H}$ with $\sigma_0 = 2$, $\sigma_1 = 0$ and $\sigma_2 = 1 + \varepsilon^{-1/4}\sqrt{\ln(1/\varepsilon)}$, with probability 9/10.

Now we can apply the algorithm of Theorem C.2. Since the loss is convex, we get a vector $\widehat{w}$ with $\overline{f}(\widehat{w}) - \overline{f}(w^{*\prime}) = O((\sigma_0 r + \sigma_1 r^2 + \sigma_2)\sqrt{\varepsilon}) = O((2\varepsilon^{-1/4}\sqrt{\ln(1/\varepsilon)} + \varepsilon^{-1/4}\sqrt{\ln(1/\varepsilon)})\sqrt{\varepsilon}) = O(\varepsilon^{1/4}\sqrt{\ln(1/\varepsilon)})$ where $w^{*\prime}$ is the minimizer of $\overline{f}$ on $\mathcal{H}$.

We thus have that $\overline{f}(\hat{w}) \le \overline{f}(w^{*\prime}) + O(\varepsilon^{1/4}\sqrt{\ln(1/\varepsilon)}) \le \overline{f}(w') + O(\varepsilon^{1/4}\sqrt{\ln(1/\varepsilon)}) \le \overline{f}(w^*) + O(\varepsilon^{1/4}\sqrt{\ln(1/\varepsilon)})$. The second inequality follows because $w^{*\prime}$ is the minimizer of $\overline{f}$ on $\mathcal{H}$, and the third inequality follows from Lemma E.12. □

F Additional Experimental Results

In this section, we provide additional plots of our experimental results, comparing with all baselines considered.

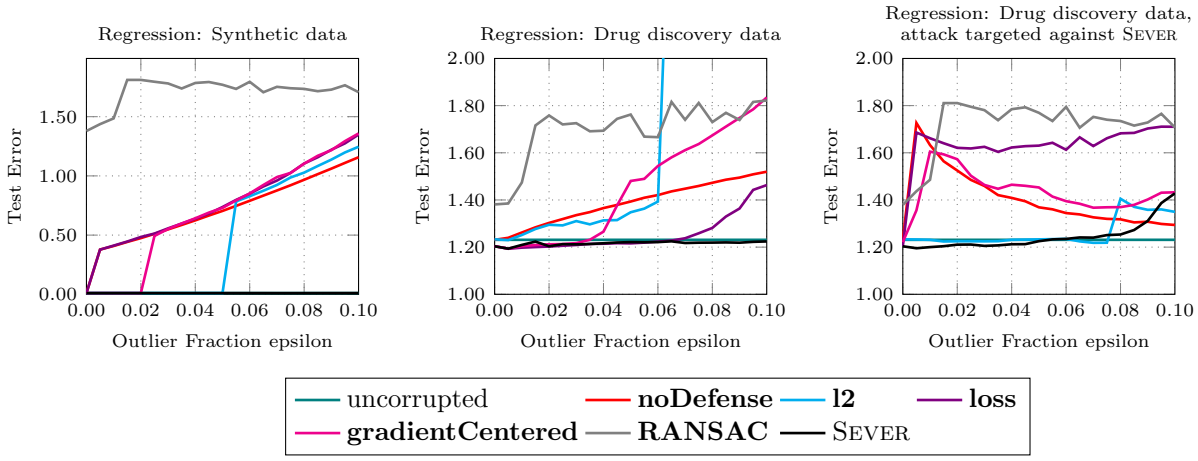

Figure 7: ε vs test error for baselines and SEVER on synthetic data and the drug discovery dataset. The left and middle figures show that SEVER continues to maintain statistical accuracy against our attacks which are able to defeat previous baselines. The right figure shows an attack with parameters chosen to increase the test error SEVER on the drug discovery dataset as much as possible. Despite this, SEVER still has relatively small test error.

References

[ABL14] P. Awasthi, M. F. Balcan, and P. M. Long. The power of localization for efficiently learning linear separators with noise. In *Symposium on Theory of Computing (STOC)*, pages 449–458, 2014.

[BDLS17] S. Balakrishnan, S. S. Du, J. Li, and A. Singh. Computationally efficient robust sparse estimation in high dimensions. In *Proceedings of the 30th Conference on Learning Theory, COLT 2017*, pages 169–212, 2017.

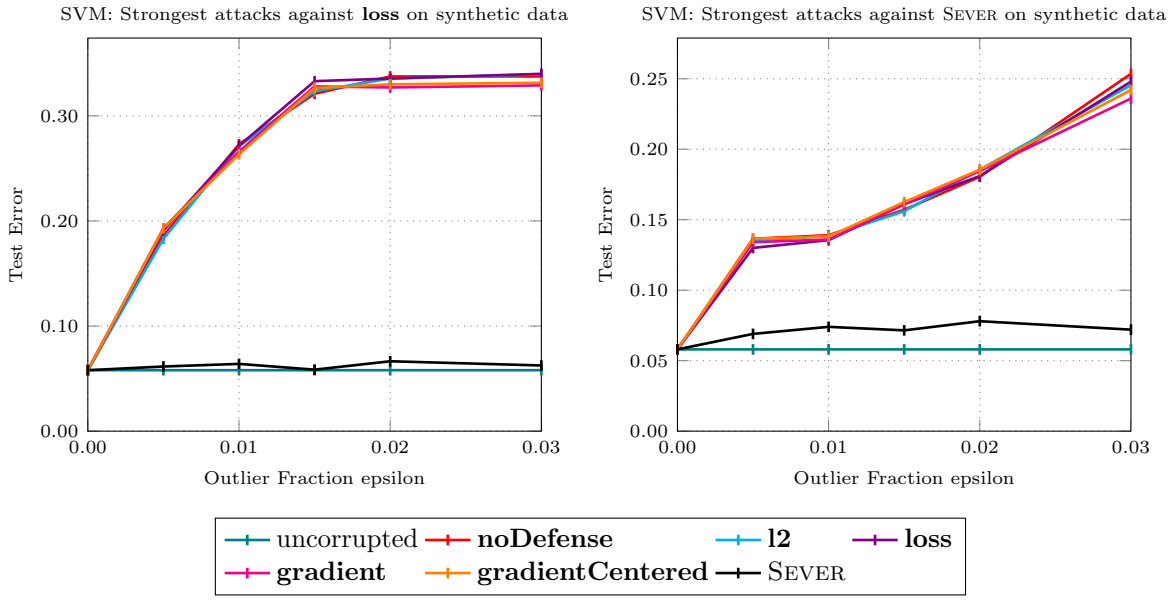

Figure 8: ε vs test error for baselines and SEVER on synthetic data. The left figure demonstrates that SEVER is accurate when outliers manage to defeat previous baselines. The right figure shows the result of attacks which increased the test error the most against SEVER. Even in this case, SEVER performs much better than the baselines.

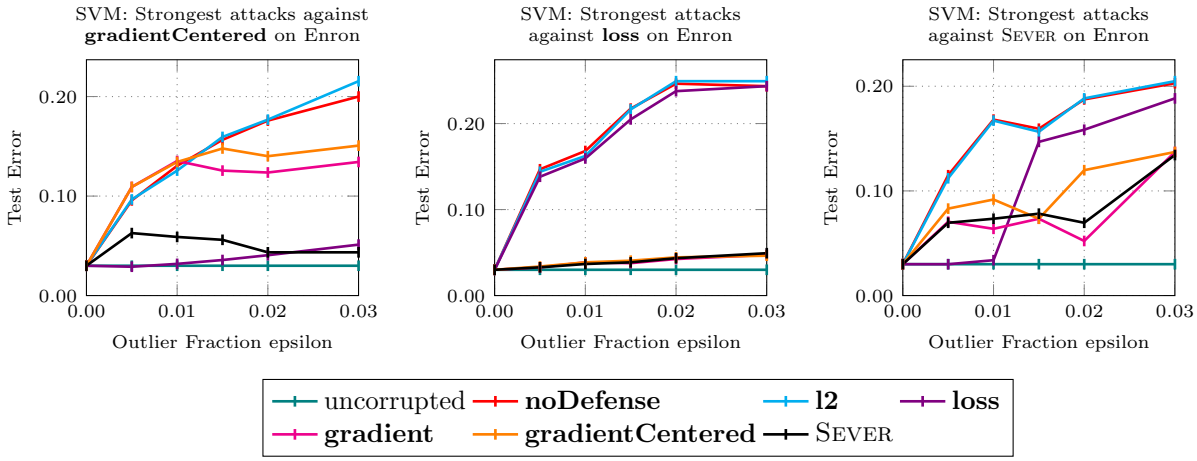

Figure 9: ε versus test error for baselines and SEVER on the Enron spam corpus. The left and middle figures are the attacks which perform best against two baselines, while the right figure performs best against SEVER. Though other baselines may perform well in certain cases, only SEVER is consistently accurate. The exception is for certain attacks at $\varepsilon = 0.03$, which, as shown in Figure 6, require three rounds of outlier removal for any method to obtain reasonable test error – in these plots, our defenses perform only two rounds.

[Ben17] Y. Bengio. The consciousness prior. *arXiv preprint arXiv:1709.08568*, 2017.

[BJK15] K. Bhatia, P. Jain, and P. Kar. Robust regression via hard thresholding. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems*

2015, pages 721–729, 2015.

[BJKK17] K. Bhatia, P. Jain, P. Kamalaruban, and P. Kar. Consistent robust regression. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017*, pages 2107–2116, 2017.

[BKNS00] M. M¿ Breunig, H. Kriegel, R. T. Ng, and J. Sander. Lof: identifying density-based local outliers. In *ACM sigmod record*, volume 29, pages 93–104. ACM, 2000.

[BNJT10] M. Barreno, B. Nelson, A. D. Joseph, and J. D. Tygar. The security of machine learning. *Machine Learning*, 81(2):121–148, 2010.

[BNL12] B. Biggio, B. Nelson, and P. Laskov. Poisoning attacks against support vector machines. In *International Conference on Machine Learning (ICML)*, pages 1467–1474, 2012.

[CLL+17] X. Chen, C. Liu, B. Li, K. Lu, and D. Song. Targeted backdoor attacks on deep learning systems using data poisoning. *arXiv preprint arXiv:1712.05526*, 2017.

[CSV17] M. Charikar, J. Steinhardt, and G. Valiant. Learning from untrusted data. In *Proceedings of STOC 2017*, pages 47–60, 2017.

[DKK+16] I. Diakonikolas, G. Kamath, D. M. Kane, J. Li, A. Moitra, and A. Stewart. Robust estimators in high dimensions without the computational intractability. In *Proceedings of FOCS'16*, pages 655–664, 2016.

[DKK+17] I. Diakonikolas, G. Kamath, D. M. Kane, J. Li, A. Moitra, and A. Stewart. Being robust (in high dimensions) can be practical. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017*, pages 999–1008, 2017. Full version available at https://arxiv.org/abs/1703.00893.

[DKK+18] I. Diakonikolas, G. Kamath, D. M. Kane, J. Li, A. Moitra, and A. Stewart. Robustly learning a gaussian: Getting optimal error, efficiently. In *Proceedings of the Twenty-Ninth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2018*, pages 2683–2702, 2018. Full version available at https://arxiv.org/abs/1704.03866.

[DKS16] I. Diakonikolas, D. M. Kane, and A. Stewart. Robust learning of fixed-structure bayesian networks. *CoRR*, abs/1606.07384, 2016.

[DKS17a] I. Diakonikolas, D. M. Kane, and A. Stewart. Learning geometric concepts with nasty noise. *CoRR*, abs/1707.01242, 2017.

[DKS17b] I. Diakonikolas, D. M. Kane, and A. Stewart. List-decodable robust mean estimation and learning mixtures of spherical gaussians. *CoRR*, abs/1711.07211, 2017.

[DKS17c] I. Diakonikolas, D. M. Kane, and A. Stewart. Statistical query lower bounds for robust estimation of high-dimensional gaussians and gaussian mixtures. In *58th IEEE Annual Symposium on Foundations of Computer Science, FOCS 2017*, pages 73–84, 2017. Full version available at http://arxiv.org/abs/1611.03473.

[DKS19] I. Diakonikolas, W. Kong, and A. Stewart. Efficient algorithms and lower bounds for robust linear regression. In *Proceedings of the Twenty-Ninth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2019*, pages 2745–2754, 2019.

[EEF+18] K. Eykholt, I. Evtimov, E. Fernandes, B. Li, A. Rahmati, C. Xiao, A. Prakash, T. Kohno, and D. Song. Robust physical-world attacks on deep learning visual classification. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[FB81] M. A. Fischler and R. C. Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, 1981.

[Gur16] Gurobi Optimization, Inc. Gurobi optimizer reference manual, 2016.

[HA04] V. Hodge and J. Austin. A survey of outlier detection methodologies. *Artificial intelligence review*, 22(2):85–126, 2004.

[HL17] S. B. Hopkins and J. Li. Mixture models, robustness, and sum of squares proofs. *CoRR*, abs/1711.07454, 2017.

[Hub64] P. J. Huber. Robust estimation of a location parameter. *Ann. Math. Statist.*, 35(1):73–101, 03 1964.

[KL17] P. W. Koh and P. Liang. Understanding black-box predictions via influence functions. In *International Conference on Machine Learning (ICML)*, 2017.

[KLS09] A. R. Klivans, P. M. Long, and R. A. Servedio. Learning halfspaces with malicious noise. *Journal of Machine Learning Research (JMLR)*, 10:2715–2740, 2009.

[KS17a] P. K. Kothari and J. Steinhardt. Better agnostic clustering via relaxed tensor norms. *CoRR*, abs/1711.07465, 2017.

[KS17b] P. K. Kothari and D. Steurer. Outlier-robust moment-estimation via sum-of-squares. *CoRR*, abs/1711.11581, 2017.

[KSL18] P. W. Koh, J. Steinhardt, and P. Liang. Stronger data poisoning attacks break data sanitization defenses. *arXiv preprint arXiv:1811.00741*, 2018.

[LAT+08] J.Z. Li, D.M. Absher, H. Tang, A.M. Southwick, A.M. Casto, S. Ramachandran, H.M. Cann, G.S. Barsh, M. Feldman, L.L. Cavalli-Sforza, and R.M. Myers. Worldwide human relationships inferred from genome-wide patterns of variation. *Science*, 319:1100–1104, 2008.

[Löf04] J. Löfberg. YALMIP: A toolbox for modeling and optimization in MATLAB. In *CACSD*, 2004.

[LRV16] K. A. Lai, A. B. Rao, and S. Vempala. Agnostic estimation of mean and covariance. In *Proceedings of FOCS'16*, 2016.

[LWSV16] B. Li, Y. Wang, A. Singh, and Y. Vorobeychik. Data poisoning attacks on factorization-based collaborative filtering. In *Advances in Neural Information Processing Systems (NIPS)*, 2016.

[MAP06] V. Metsis, I. Androutsopoulos, and G. Paliouras. Spam filtering with naive Bayes – which naive Bayes? In *CEAS*, volume 17, pages 28–69, 2006.

[MV18] M. Meister and G. Valiant. A data prism: Semi-verified learning in the small-alpha regime. In *Proceedings of the 31st Conference On Learning Theory*, volume 75 of *Proceedings of Machine Learning Research*, pages 1530–1546. PMLR, 06–09 Jul 2018.

[NPXNR14] A. Newell, R. Potharaju, L. Xiang, and C. Nita-Rotaru. On the practicality of integrity attacks on document-level sentiment analysis. In *Workshop on Artificial Intelligence and Security (AISec)*, pages 83–93, 2014.

[NT13] N. H. Nguyen and T. D. Tran. Exact recoverability from dense corrupted observations via ℓ_1-minimization. *IEEE Transactions on Information Theory*, 59(4):2017–2035, 2013.

[NTN11] N. M. Nasrabadi, T. D. Tran, and N. Nguyen. Robust lasso with missing and grossly corrupted observations. In *Advances in Neural Information Processing Systems (NIPS)*, 2011.

[OSB+18] I. Olier, N. Sadawi, G. R. Bickerton, J. Vanschoren, C. Grosan, L. Soldatova, and Ross D. King. Meta-qsar: a large-scale application of meta-learning to drug design and discovery. *Machine Learning*, 107(1):285–311, Jan 2018.

[Owe07] A. B. Owen. A robust hybrid of lasso and ridge regression. *Contemporary Mathematics*, 443(7):59–72, 2007.

[PLJD10] P. Paschou, J. Lewis, A. Javed, and P. Drineas. Ancestry informative markers for fine-scale individual assignment to worldwide populations. *Journal of Medical Genetics*, 47:835–847, 2010.

[PSBR18] A. Prasad, A. S. Suggala, S. Balakrishnan, and P. Ravikumar. Robust estimation via robust gradient estimation. *CoRR*, abs/1802.06485, 2018.

[QV18] M. Qiao and G. Valiant. Learning discrete distributions from untrusted batches. In *Innovations in Theoretical Computer Science (ITCS)*, 2018.

[RD99] P. J. Rousseeuw and K. V. Driessen. A fast algorithm for the minimum covariance determinant estimator. *Technometrics*, 41(3):212–223, 1999.

[RPW+02] N. Rosenberg, J. Pritchard, J. Weber, H. Cann, K. Kidd, L.A. Zhivotovsky, and M.W. Feldman. Genetic structure of human populations. *Science*, 298:2381–2385, 2002.

[SCV18] J. Steinhardt, M. Charikar, and G. Valiant. Resilience: A criterion for learning in the presence of arbitrary outliers. In *Innovations in Theoretical Computer Science (ITCS)*, 2018.

[SKL17] J. Steinhardt, P. W. Koh, and P. Liang. Certified defenses for data poisoning attacks. In *Advances in Neural Information Processing Systems (NIPS)*, 2017.

[Ste17] J. Steinhardt. Does robustness imply tractability? A lower bound for planted clique in the semi-random model. *arXiv*, 2017.

[SZS+14] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus. Intriguing properties of neural networks. In *International Conference on Learning Representations (ICLR)*, 2014.

[Tuk60] J. W. Tukey. A survey of sampling from contaminated distributions. *Contributions to probability and statistics*, 2:448–485, 1960.

[Tuk75] J.W. Tukey. Mathematics and picturing of data. In *Proceedings of ICM*, volume 6, pages 523–531, 1975.