Robust Linear Regression for General Feature Distribution

Tom Norman **Nir Weinberger** **Kfir Y. Levy**
Department of Electrical and Computer Engineering
Technion

Abstract

We investigate robust linear regression where data may be contaminated by an oblivious adversary, i.e., an adversary that knows the data distribution but is otherwise oblivious to the realization of the data samples. This model has been previously analyzed under strong assumptions. Concretely, **(i)** all previous works assume that the covariance matrix of the features is positive definite; **(ii)** most of them assume that the features are centered. Additionally, all previous works make additional restrictive assumptions, e.g., assuming Gaussianity of the features or symmetric distribution of the corruptions.

In this work, we investigate robust regression under a more general set of assumptions: **(i)** the covariance matrix may be either positive definite or positive semi definite, **(ii)** features may not be centered, **(iii)** no assumptions beyond boundedness (or sub-Gaussianity) of the features and the measurement noise. Under these assumptions we analyze a sequential algorithm, namely, a natural SGD variant for this problem, and show that it enjoys a fast convergence rate when the covariance matrix is positive definite. In the positive semi definite case we show that there are two regimes: if the features are centered, we can obtain a standard convergence rate; Otherwise, the adversary can cause any learner to fail arbitrarily.

1 INTRODUCTION

The remarkable recent success of Machine Learning (ML) models has lead to their wide adoption in numerous fields. However, a central challenge that is still standing in deploying ML models in real world scenarios is their *robust* design. That is, the designed model should be immune to data contamination, which may arise due to adversarial corruptions, extreme events, malfunctioning sensors, and other causes. Even beyond ML, designing robust models has proven to be crucial in various applications, including Economics (Zaman et al., 2001), Computer Vision (Gustafsson et al., 2020), Biology (Yeung et al., 2002), and Healthcare (Davies et al., 2004).

In this paper, we investigate the fundamental ML task of *robust linear regression*, and concretely, consider the case that a fraction α of the observations were contaminated by an adversary. In this context, it is well known that standard regression methods are highly sensitive to outliers, and might break down even in the presence of a single contaminated data point. This is illustrated in Appendixes A and B.1.

Past research on robust linear regression roughly falls into one of two categories, according to the power of the adversary: **(i)** An *adaptive adversary*, which is allowed to contaminate the data *after* observing the data samples. This setting was explored, e.g., in Candes and Tao (2005); Charikar et al. (2017); Klivans et al. (2018); Liu et al. (2019); Dalalyan and Thompson (2019); Diakonikolas et al. (2019a,b). It is well known that adaptive adversaries may cause any learner to incur a non vanishing error, depending on the fraction of contaminated samples, irrespective of the number of data points. Conversely, **(ii)** an *oblivious adversary*, which is not allowed to observe the samples, but may know the true statistical properties of the data. This setting was explored, e.g., in Tsakonas et al. (2014); Suggala et al. (2019); Pesme and Flammarion (2020); Sun et al. (2020); D'Orsi et al. (2021). These works have remarkably shown that vanishing error is possible in this model with increased number of data samples, for *any* fraction of contamination.

In this paper, we focus on robust linear regression with an oblivious adversary, with the goal of relaxing some of the limiting assumptions made in previous works. Concretely: **(i)** all previous works assume that the covariance matrix of the features Σ is (strictly) positive definite; and **(ii)** most works assume that the features are centered (i.e., have zero mean). As we show in Appendix B.2 a naive attempt to circumvent the centering assumption by taking pairwise differences of samples fails. In addition, previous works were

Proceedings of the 26th International Conference on Artificial Intelligence and Statistics (AISTATS) 2023, Valencia, Spain. PMLR: Volume 206.

restricted to Gaussian features, or symmetrically distributed corruptions.

The results of our work apply to general feature and noise distributions with bounded (or just sub-Gaussian distributions, and our main contributions are as follows (see Table 1):

- $\Sigma \succeq 0$: For *centered* features, we provide an SGD (Stochastic Gradient Descent) algorithm that ensures a *prediction* error rate of $\mathcal{O}\left(1/(1-\alpha)\sqrt{T}\right)$ after observing T samples. We note that Pesme and Flammarion (2020) also obtain guarantees for this case, nevertheless their work heavily relies on the Gaussian assumption for features and noise. We also show that any algorithm may completely fail for *non-centered* features.
- $\Sigma \succ 0$ (strictly positive definite): We provide an SGD algorithm that ensures an *estimation* error of $\tilde{\mathcal{O}}\left(1/((1-\alpha)\rho)^2 T\right)$, without any assumption on feature centering.

We mention that we make no further assumptions regarding the adversary/data, and that the adversary may inject unbounded perturbations to the contaminated measurements. Our main tool is the use of the Huber loss (Huber, 1964), instead of the standard ℓ_2 loss.

Related Work

Robust statistics dates back to the works of Tukey and Huber (Tukey, 1960; Huber, 1964), with classical works focusing on asymptotic performance (Huber, 1973; Bassett Jr and Koenker, 1978; Pollard, 1991; Van der Vaart, 2000; McMahan et al., 2013), most which are not computable in polynomial time (Rousseeuw, 1984, 1985). A popular approach to robust regression with oblivious adversaries relies on replacing the ℓ_2 loss with a more robust loss function, predominantly either the ℓ_1 loss or the Huber loss (Huber, 1964) (which are convex), as well other non-convex robust losses (Tukey, 1960).

Finite Time Guarantees for Robust regression Non-asymptotic guarantees for robust regression were recently explored in several works; All of them rely on employing a convex robust loss function (either ℓ_1 or Huber). Moreover, all previous works assume strictly positive definite covariance matrix, i.e. $\Sigma \succeq \rho \cdot I$ for $\rho > 0$, and centered features, in addition to other restrictive assumptions detailed below. The conditions in the various works is summarized in Table 1 below. With more detail, Tsakonas et al. (2014), assume that the features x and measurement noise ϵ have zero-mean Gaussian distribution. Their algorithmic approach is to apply ERM (Empirical Risk Minimization) while utilizing the Huber loss. The work of Suggala et al. (2019) similarly assume that x is Gaussian, yet allows the measurement noise be sub-Gaussian. They suggest an algorithm, AdaCRR, which makes several passes over the dataset while thresholding suspicious points. The work of Pesme and Flammarion (2020) makes the same Gaussianity assumptions as Tsakonas et al. (2014), and is the first to provide guarantees for an efficient online algorithm, namely SGD with ℓ_1 loss.

The recent work of D'Orsi et al. (2021) has significantly improved the theoretical understanding by relaxing the Gaussian assumptions of previous works, and, similarly to Tsakonas et al. (2014) provides guarantees to ERM over the Huber loss. Nevertheless, the results have two limiting assumptions. First, it is assumed that both the measurement noise and adversarial perturbations are *symmetrically distributed around zero*, which highly weakens the adversary. Second, an assumption on the features related to their *spreadness* is made (though the features are allowed to be non-centered). We remark that the work of D'Orsi et al. (2021) also shows that one cannot obtain guarantees when the *spreadness* assumption is violated, which at a first glance seems to contradict our results. Nevertheless, as we show in Appendix H, the result of D'Orsi et al. (2021) only holds when the norm of the optimal solution is unbounded, and so the impossibility result does not hold in the natural case for which the norm of the optimal solution is bounded. This settles the result of this paper with that of D'Orsi et al. (2021).

2 PROBLEM FORMULATION

We consider a linear model with contamination by an oblivious adversary.

Model 2.1. $$y = \langle w^*, x \rangle + \epsilon + b \,,$$

where $x \in \mathbb{R}^d$ is a feature vector, $\epsilon \in \mathbb{R}$ is a zero-mean additive noise that is statistically independent of x, and $\langle \cdot, \cdot \rangle$ denotes the standard inner product. The corruption $b \in \mathbb{R}$ is chosen by an adversary that knows w^* as well as the probability distributions of x and ϵ, but is otherwise *oblivious to their realizations*. The probability distribution of the corruption b is constrained to satisfy $\mathbb{P}(b \neq 0) = \alpha$, where $\alpha \in [0, 1)$ is the nominal fraction of samples the adversary is allowed to corrupt.

We note that if $\mathcal{P}$ is the distribution of the non-corrupted data samples (x, y) (i.e., $b = 0$), and $\mathcal{Q}$ is the distribution over corrupted samples (for which $b \neq 0$), then the contaminated samples in Model 2.1 are drawn from the mixture

$$(x, y) \sim (1 - \alpha)\mathcal{P} + \alpha\mathcal{Q} \,.$$

This model is known as the α-Huber contamination model (Huber, 1964).

Robust Linear Regression problem setting A learner is given T independent samples $\{(x_i, y_i)\}_{i=1}^T$ from Model 2.1,

Table 1: Assumptions and Rates of Related Work.

Paper	Features	Noise & Adversary	Rates for $\Sigma \succ 0$	Rates for $\Sigma \succeq 0$
Tsakonas et al. (2014)	$x \sim \mathcal{N}(0, I_d)$	$\epsilon \sim \mathcal{N}(0, \sigma^2)$	$\tilde{\mathcal{O}}\left(1/(1-\alpha)^2 T\right)$	N /A
Suggala et al. (2019)	$x \sim \mathcal{N}(0, \Sigma)$	$\epsilon \sim subG(\sigma^2)$	$\tilde{\mathcal{O}}\left(1/(1-\alpha)^2 T\right)$	N /A
Pesme and Flammarion (2020)	$x \sim \mathcal{N}(0, \Sigma)$	$\epsilon \sim \mathcal{N}(0, \sigma^2)$	$\tilde{\mathcal{O}}\left(1/(1-\alpha)^2 T\right)$	$\mathcal{O}\left(1/(1-\alpha)\sqrt{T}\right)$
D'Orsi et al. (2021)	Spreadness of the design matrix X	$\epsilon + b$ has a symmetric distribution around 0	$\tilde{\mathcal{O}}\left(1/(1-\alpha)^2 T\right)$	N /A
This paper	$x \sim subG(\kappa^2)$	$\epsilon \sim subG(\sigma^2)$	$\tilde{\mathcal{O}}\left(1/(1-\alpha)^2 T\right)$	$\mathcal{O}\left(1/(1-\alpha)\sqrt{T}\right)$

and is required to learn a parameter vector $w \in \mathbb{R}^d$ for either the *prediction* problem with respect to the *non-corrupted* data

$$\min_w F(w) := \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{P}} \left[\frac{1}{2}\left(\langle w, x\rangle - y\right)^2\right] , \qquad (1)$$

or the *estimation* problem ($\|\cdot\|$ denotes the ℓ_2 norm)

$$\arg\min_w \|w - w^*\|^2 .$$

In the ordinary prediction or estimation problems, when there is no adversary ($\alpha = 0$), a standard solution method is *least squares* (Gauss, 1809), in which the vector w which minimizes an empirical version of either the prediction error or estimation error. However, this method is known to be fragile for $\alpha \neq 0$; See Appendix A for a simple example.

Given parameter radius D and noise standard-deviation σ, we make the following assumptions throughout most of the paper:

Assumption 2.2 (Bounded Parameter Vector)**.** $w^* \in \mathcal{B} := \left\{w \in \mathbb{R}^d : \|w\| \leq D\right\}$.

Assumption 2.3 (Bounded Zero-Mean Noise)**.** $|\epsilon| \leq \sigma$ *with probability* 1 *and* $\mathbb{E}[\epsilon] = 0$.

Assumption 2.4 (Bounded Feature Vector)**.** $\|x\| \leq 1$ *with probability* 1.

One may surmise that given Assumption 2.2 and Assumption 2.3, the prior knowledge of D and σ can be used by an estimator to *filter* samples which have excessively large magnitude, and thus essentially remove contaminated samples from the dataset. However, as we show in Appendix B.1, an application of this method fails even for a rather simple example (at least a vanilla application of this method). We also mention that in Section 6.1, we relax the assumptions to sub-Gaussian features and measurement noise, rather than strictly bounded.

We denote by Σ the covariance matrix of the feature vector, $\Sigma := \mathbb{E}[(x - \mathbb{E}[x])(x - \mathbb{E}[x])^T]$. In later sections, in order to obtain guarantees on the estimation error we would also assume the following:

Assumption 2.5 (Strictly Positive Definite Covariance Matrix of the Feature Vector)**.** $\Sigma \succeq \rho \cdot I$*, where* $\rho > 0$ *and* I *is the identity matrix.*

We remark that due to Assumption 2.4 ($\|x\| \leq 1$) and the dimension d of the problem, it must hold that $\rho \leq \frac{1}{d}$, but it can also be much smaller. Our bounds will be stated in terms of ρ, which is more general than setting $\rho = \frac{1}{d}$, as adequate for a vector of i.i.d. features.

We will use Assumption 2.5 to show that the expected loss is a strongly convex function in $\mathcal{B}$, and this property will be extensively used. We thus remind the reader the following properties:

Property 2.6. Let f be twice continuously differentiable. f is λ-strongly convex in $\mathcal{B}$ if and only if $\forall w \in \mathcal{B} : \nabla^2 f(w) \succeq \lambda \cdot I$.

Property 2.7. Let f be λ-strongly convex over $\mathcal{B}$. Denote $w_{opt} := \arg\min_{w\in\mathcal{B}} f(w)$. Then, $\forall w \in \mathcal{B} : f(w) - f(w_{opt}) \geq \frac{\lambda}{2}\|w - w_{opt}\|^2$.

Finally, we denote the orthogonal projection onto $\mathcal{B}$ by $\Pi_{\mathcal{B}}(\cdot)$, i.e. $\Pi_{\mathcal{B}}(u) := \arg\min_{w\in\mathcal{B}} \|w - u\|$.

3 HUBER LOSS FOR ROBUST REGRESSION

The ℓ_2 loss function is sensitive to outliers, and so a naive approach to learning in the model Model 2.1 that minimize the ℓ_2 loss while using the contaminated data may completely fail. Simple examples, such as the one we outline in Appendix A, show this phenomena. We thus consider minimizing a *robust* loss function, and concretely, the Huber loss (Huber, 1964). This loss function behaves similarly to the ℓ_2 loss in a region around the origin, yet far from the origin it behaves similarly to the ℓ_1 loss, and hence its gradients are bounded over $\mathbb{R}$. Using the Huber loss is a popular technique in robust regression (e.g. Tsakonas et al. (2014); D'Orsi et al. (2021)), however in contrast to previous works, our analysis is done under a less restrictive set of assumptions, and our algorithm is a simple variant of standard SGD.

The Huber loss $h_R : \mathbb{R} \mapsto \mathbb{R}$ is a convex function, parameterized by a radius parameter R, defined as

$$h_R(s) := \begin{cases} \frac{1}{2}s^2 & \text{, if } |s| \leq R \\ R(|s| - \frac{1}{2}R) & \text{, otherwise} \end{cases} .$$

We will also denote $\phi_R(s) := \min\{R, \max\{s, -R\}\}$, so that $\nabla_w h_R(\langle w, x\rangle) = \phi_R(\langle w, x\rangle) \cdot x$ and $|\phi_R(s)| \leq R$ holds.

Choice of radius R The radius parameter R can be adjusted by the learner, and involves a trade-off between unnecessarily clipping clean gradients vs. limiting corrupted gradients. Our choice of R is made such that the loss of non-corrupted observations is in the quadratic regime (so its gradient is not affected), while the loss for highly corrupted observations is in the linear regime (so its gradient has bounded norm). Next we detail on how to choose R. Specifically, for a non-corrupted sample i ($b_i = 0$) and $w \in \mathcal{B}$,

$$\begin{aligned}|\langle w, x_i\rangle - y_i| &\leq |\langle w, x_i\rangle| + |\langle w^*, x_i\rangle| + |\epsilon_i| \\ &\leq \|x_i\| \cdot (\|w\| + \|w^*\|) + |\epsilon_i| \\ &\leq \|w\| + \|w^*\| + \sigma\,,\end{aligned}$$

and we limit our search to $\mathcal{B}$ since we know that $w^* \in \mathcal{B}$. Thus, taking $R := 6D + \sigma$ assures that the gradient of non-corrupted samples is as for the ℓ_2 loss (The choice of possibly unnecessarily large constant 6 is made here for consistency with later derivations in Section 5). With this choice, for any $w \in \mathcal{B}$ and non-corrupted sample (x_i, y_i), we have with probability 1 that

$$h_R(\langle w, x_i\rangle - y_i) = \tfrac{1}{2}(\langle w, x_i\rangle - y_i)^2\,. \tag{2}$$

Robust Objective Function The objective function we choose is the *expected Huber loss* with respect to the contaminated data, given by

$$L_R(w) := \mathop{\mathbb{E}}_{x,\epsilon,b}[h_R(\langle w, x\rangle - y)]\,.$$

Since $L_R(\cdot)$ is defined with respect to the contaminated data distribution we can efficiently compute unbiased estimates of its gradients and apply SGD. We will use $L_R(\cdot)$ as a proxy to the true expected loss $F(\cdot)$ appearing in Equation (1). The next lemma explicitly relates $L_R(\cdot)$ to $F(\cdot)$, as follows:

Lemma 3.1. *Let $R := 6D + \sigma$. Then, $\forall w \in \mathcal{B}$: $L_R(w) = (1-\alpha)F(w) + \alpha H(w)$, where $H(w) := \mathbb{E}_{x,\epsilon,b}[h_R(\langle w - w^*, x\rangle - \epsilon - b)\,|b \neq 0]$ is the expected Huber loss of corrupted samples.*

The proof is based on the law of total expectation on $L_R(\cdot)$ with respect to b, and can be found in Appendix C.

4 THE GENERAL FEATURE COVARIANCE MATRIX CASE

In this section, we make no assumptions on Σ, that is, allow it to have vanishing eigenvalues. We show that there is a stark difference between centered and non-centered features, i.e., whether $\mathbb{E}[x] \neq 0$ or not. For non-centered features, we provide an example showing that obtaining a low prediction error is generally impossible, even for a known $\mathbb{E}[x]$. For centered features ($\mathbb{E}[x] = 0$), we show that an efficient predictor can be learned (Theorem 4.2). Our proposed algorithm does not require the knowledge of the corruptions fraction α.

Low Prediction Error is Generally Impossible Let $\alpha = \frac{1}{2}$ and consider two one-dimensional models with parameter vectors w_1^*, w_2^*, where $w_2^* = -w_1^* = 1$. For both models we assume $\epsilon = 0, x = 1$ with probability 1; thus $\mathbb{E}[x]$ is known and equals 1. The adversary chooses his corruptions for every model $m \in [1, 2]$ so $b^{(m)} = -2 \cdot w_m^*$ with probability $\alpha = \frac{1}{2}$ and 0 otherwise. Then $y^{(m)} = w_m^* + b^{(m)}$ and both $y^{(1)}$ and $y^{(2)}$ have the same probability distribution of $y^{(m)} = \pm 1$ with probability $\alpha = \frac{1}{2}$. Since the contaminated data has the same distribution in both cases, then a learner cannot distinguish between these models. Thus, this adversary can cause any learner to incur a fixed (non-decreasing) prediction error, irrespective of the number of available samples. In Appendix K we provide a generalization for any $\alpha \in [0, 1]$.

A Prediction Error Bound for Centered Features We next show that if $\mathbb{E}[x] = 0$, then a low prediction error can be achieved even if Σ has vanishing eigenvalues. The following is the key lemma which shows that the expected ℓ_2 error can be bounded by expected Huber loss error.

Lemma 4.1. *For any $w \in \mathcal{B}$: $F(w) - F(w^*) \leq \frac{1}{1-\alpha}(L_R(w) - L_R(w^*))$.*

Proof. We first show that $H(w) \geq H(w^*)$ for all $w \in \mathbb{R}^d$. Indeed, for any $w \in \mathbb{R}^d$:

$$\begin{aligned}H(w) &:= \mathop{\mathbb{E}}_{x,\epsilon,b}\left[h_R(\langle w - w^*, x\rangle - \epsilon - b)\,\middle|\, b \neq 0\right] \\ &\overset{(a)}{\geq} \mathop{\mathbb{E}}_{\epsilon,b}\left[h_R\left(\mathop{\mathbb{E}}_{x}[\langle w - w^*, x\rangle - \epsilon - b]\right)\,\middle|\, b \neq 0\right] \\ &\overset{(b)}{=} \mathop{\mathbb{E}}_{\epsilon,b}\left[h_R\left(\left\langle w - w^*, \mathbb{E}[x]\right\rangle - \epsilon - b\right)\,\middle|\, b \neq 0\right] \\ &\overset{(c)}{=} \mathop{\mathbb{E}}_{\epsilon,b}\left[h_R(-\epsilon - b)\,\middle|\, b \neq 0\right] \\ &= \mathop{\mathbb{E}}_{x,\epsilon,b}\left[h_R(\langle w^* - w^*, x\rangle - \epsilon - b)\,\middle|\, b \neq 0\right] \\ &= H(w^*)\,,\end{aligned}$$

where (a) follows from Jensen's inequality applied to the convex function $h_R(\cdot)$, as well as from the independence assumptions, (b) follows from the linearity of the inner product, (c) follows from the assumption $\mathbb{E}[x] = 0$.

Using $H(w) \geq H(w^*)$ together with Lemma 3.1, gives

Algorithm 1 Huber SGD

Input: $D > 0, R > 0, T \in \mathbb{N}_+$
$w_1 := 0$
$\eta := \frac{D}{R\sqrt{T}}$
for $t = 1$ **to** T **do**
  Draw (x_t, y_t) from (the contaminated) Model 2.1
  $g_t := \phi_R\left(\langle w_t, x_t\rangle - y_t\right) \cdot x_t$
  $w_{t+1} := \Pi_{\mathcal{B}}\left(w_t - \eta \cdot g_t\right)$
end for
Return: $\bar{w} := \frac{1}{T}\sum_{t=1}^{T} w_t$

$\forall w \in \mathcal{B}$,

$$\begin{aligned} F(w) - F(w^*) &= \frac{1}{1-\alpha}\left(L_R(w) - L_R(w^*)\right) \\ &\quad - \frac{\alpha}{1-\alpha}\left(H(w) - H(w^*)\right) \\ &\leq \frac{1}{1-\alpha}\left(L_R(w) - L_R(w^*)\right) . \end{aligned}$$

□

Lemma 4.1 implies that by applying SGD to the expected Huber loss $L_R(\cdot)$ (while using the contaminated data), leads to guarantees on the true expected prediction error. Algorithm 1 formalizes this observation, and a theoretical guarantee on its error is formalized in the next theorem.

Theorem 4.2. *Given Assumptions 2.2, 2.3 and 2.4, let $\bar{w}$ be the output of Algorithm 1. Then, $\mathbb{E}\left[F(\bar{w})\right] - F(w^*) \leq \frac{RD}{(1-\alpha)\sqrt{T}}$.*

Proof. Note that $L_R(\cdot)$ is convex and $\mathcal{B}$ has a finite diameter D. Also, note that for a given $w_t \in \mathcal{B}$ then $g_t := \phi_R\left(\langle w_t, x_t\rangle - y_t\right) \cdot x_t$ is an unbiased estimate for the gradient of $L_R(\cdot)$ at w_t. Moreover, since $\phi_R(\cdot)$ is bounded by R and the features are bounded by 1, then the norm of g_t is bounded by R. Thus, applying projected SGD as is done in Algorithm 1 ensures that (see e.g. Shalev-Shwartz (2012, Theorem 14.8)),

$$\mathbb{E}\left[L_R(\bar{w})\right] - L_R(w^*) \leq \frac{RD}{\sqrt{T}} .$$

Combining this with Lemma 4.1 concludes the proof. □

5 THE STRICTLY POSITIVE DEFINITE FEATURE COVARIANCE MATRIX CASE

In this section, we assume that $\Sigma \succeq \rho \cdot I$ for some known, strictly positive, $\rho > 0$, and show that under this stronger condition (compared to the general case discussed in Section 4), it is possible to obtain guarantees even if $\mathbb{E}[x] \neq 0$. We also establish faster convergence rates compared to the general case.

As a first step, we consider the case in which the expectation of the features $\mathbb{E}[x]$ is known to the learner, so it can perfectly center the features. We show that feeding these centered samples to an appropriate variant of SGD leads to an accurate estimation of w^* with an error rate of $\mathcal{O}\left(1/((1-\alpha)\rho)^2 T\right)$ (Theorem 5.4). Afterwards, we consider the case of unknown features expectation, and show that an algorithm that is based on an initial phase of expectation estimation is effective, and achieves the same estimation rate, up to logarithmic factors in T (Theorem 5.7).

5.1 The Known Expectation Case

The main difficulty in the analysis of this setting is that when the features are not centered, the minimizer of $L_R(\cdot)$ might be different from w^*. A natural way to avoid it is to center the features, i.e. $x_i - \mathbb{E}[x]$. In fact, Model 2.1 can be written with centered feature vectors, as follows,

Model 5.1. $y := \langle w^*, x - \mathbb{E}[x]\rangle + \langle w^*, \mathbb{E}[x]\rangle + \epsilon + b$.

This is still a linear model, albeit it has two differences compared to the original Model 2.1. First, the norm of the centered features might be larger than 1. To resolve this we recall that the radius parameter of the Huber loss was set to $R := 6D + \sigma$, and so for any non-corrupted sample in Model 5.1

$$\begin{aligned} \left|\langle w, x - \mathbb{E}[x]\rangle - y\right| &\leq D\left(\|x\| + \|\mathbb{E}[x]\|\right) + |y| \\ &\leq 2D + D + \sigma \leq R , \end{aligned}$$

where the first inequality follows from the triangle inequality and the second one follows from Assumption 2.4. Thus, Equation (2) still applies with probability 1 for any $w \in \mathcal{B}$ and non-corrupted sample (x_i, y_i).

Second, Model 5.1 has an additional unknown quantity, to wit $\langle w^*, \mathbb{E}[x]\rangle$. As we next show, this additional quantity does not alter the properties of $L_R(w)$ which are essential to the analysis of the algorithm. To this end, we define, With slight abuse of notation,

$$L_R(w) = \mathop{\mathbb{E}}_{x,\epsilon,b}\left[h_R\left(\left\langle w, x - \mathbb{E}[x]\right\rangle - y\right)\right] . \quad (3)$$

This expected Huber loss of centered features is also minimized by w^*:

Lemma 5.2. *$L_R(w)$ is $(1-\alpha)\rho$-strongly convex in $\mathcal{B}$ and $w^* = \arg\min_{w\in\mathcal{B}} L_R(w)$.*

The full proof can be found in Appendix D, which establishes Lemma D.1, a more general version of this lemma.

Huber Loss SGD with Centered Features (Algorithm 2) Lemma 5.2 is instrumental in achieving fast convergence

Algorithm 2 Huber SGD for known expectation

Input: $R > 0, \lambda > 0, T \in \mathbb{N}_+, r \in \mathbb{R}^d$
$w_1 := 0$
for $t = 1$ **to** T **do**
  Draw (x_t, y_t) from Model 2.1
  $\eta_t := 1/\lambda t$
  $g_t := \phi_R\left(\langle w_t, x_t - r\rangle - y_t\right) \cdot (x_t - r)$
  $w_{t+1} := \Pi_{\mathcal{B}}\left(w_t - \eta_t \cdot g_t\right)$
end for
Return: $\bar{w} := \frac{2}{T}\sum_{t=1+T/2}^{T} w_t$

of the Huber loss, and, in turn, for the estimation error $\|\bar{w} - w^*\|^2$. As the lemma implies, $L_R(\cdot)$ is strongly convex and maintains the same global optimum w^* as the true expected loss $F(\cdot)$. Thus, it is natural to apply an appropriate version of SGD for strongly-convex functions to $L_R(\cdot)$ while using the contaminated data with *centered* features, in order to obtain guarantees to $\|\bar{w} - w^*\|^2$. This is formulated in Algorithm 2. Concretely, Algorithm 2 utilizes a SGD with $\frac{1}{2}$-suffix averaging, which has the following guarantees:

Lemma 5.3 (Rakhlin et al. (2012, Theorem 5))**.** *Consider SGD with $\frac{1}{2}$-suffix averaging and with step size $\eta_t := 1/\lambda t$. Suppose f is λ-strongly convex, and that $\mathbb{E}[\|g_t\|^2] \leq G^2$ for all t. Then for any T, it holds that $\forall w \in \mathcal{B} : \mathbb{E}\left[f(\bar{w})\right] - f(w) \leq \frac{9G^2}{\lambda T}$.*

Now, by using Lemma 5.2 we can bound the estimation error of Algorithm 2 as follows,

Theorem 5.4. *Let $\bar{w}$ be the output of Algorithm 2 with input $(R, (1-\alpha)\rho, T, \mathbb{E}\left[x\right])$. Then,*

$$\mathbb{E}\left[\|\bar{w} - w^*\|^2\right] \leq \frac{72R^2}{((1-\alpha)\rho)^2 \cdot T} .$$

The full proof can be found in Appendix E. At high level, we use the fact that g_t is an unbiased gradient estimate for $\nabla L_R(w_t)$ and is bounded with probability 1. Lemma 5.2 and Lemma 5.3 then lead to the desired result.

5.2 The Unknown Expectation Case

In this section, we present Algorithm 3, which generalizes the algorithm developed in the previous section to the unknown $\mathbb{E}\left[x\right]$ case. The algorithm is based on sample splitting, and so we assume, for simplicity, that $2T$ samples are provided to the algorithm. Similarly to the previous section, the learning algorithm is based on centering the features before feeding them to an SGD with $\frac{1}{2}$-suffix averaging. The only difference is that in Algorithm 3 we center the features using the empirical mean based on first T samples, $\mu := \frac{1}{T}\sum_{t=1}^{T} z_t$.

Algorithm 3 Huber SGD for unknown expectation

Input: $R > 0, \lambda > 0, T \in \mathbb{N}_+$
Phase 1: Compute μ
Draw T samples $\{(z_i, y_i)\}_{i=1}^{T}$ from Model 2.1, and estimate the mean, $\mu := 1/T \sum_{i=1}^{T} z_i$
Phase 2: SGD with $\frac{1}{2}$-suffix averaging given μ
Return: $\bar{w}$:= output of Algorithm 2 with input (R, λ, T, μ)

Similarly to the previous section, with this definition, Model 2.1 can be written as follows,

Model 5.5. $$y = \langle w^*, x - \mu\rangle + \langle w^*, \mu\rangle + \epsilon + b\,.$$

Our proposed Algorithm 3 has two phases, one for each of its sub-samples. In the first phase, μ is estimated using T samples, and in the second phase, T steps of SGD are performed on fresh samples that are centered using μ. Since the samples used in the SGD algorithm are independent of the estimator μ (which is a function of different samples), the analysis of the SGD algorithm can be made conditionally on μ, without affecting the distribution of the samples.

With slight abuse of notation define

$$L_R(w) = \mathop{\mathbb{E}}_{x,\epsilon,b}\left[h_R\left(\langle w - w^*, x - \mu\rangle - \langle w^*, \mu\rangle - \epsilon - b\right)\right] .$$

The key lemma for the analysis is as follows,

Lemma 5.6. *$\forall \mu \in \mathbb{R}^d :$ $L_R(w)$ is $(1-\alpha)\rho$-strongly convex in $\mathcal{B}$.*

The proof follows directly from the more general Lemma D.1 (by setting with $v = q = \mu$), and utilizes the fact that μ is constant in the second phase (conditioned on the randomization of the first phase).

We next state error bound guarantees for Algorithm 3. Essentially, similarly to the case of known expectation, Algorithm 3 applies a strongly-convex variant of SGD to the above defined $L_R(\cdot)$. However, we establish this fast error bound despite the fact that the optimum of $L_R(\cdot)$ in $\mathcal{B}$ *is not necessarily* w^* (in contrast to the simpler case of known expectation).

Theorem 5.7. *Let $\bar{w}$ be the output of Algorithm 3 with input $(R, (1-\alpha)\rho, T)$. Then,*

$$\mathbb{E}\left[\|\bar{w} - w^*\|^2\right] \leq \frac{16R^2 \cdot (36\log T + 13)}{((1-\alpha)\rho)^2 \cdot T} .$$

A glimpse of the Proof. The main challenge is that now, as previously mentioned, w^* is not necessarily the optimum of $L_R(\cdot)$ in $\mathcal{B}$. To circumvent this, we analyze an auxiliary expected loss function which we define as follows,

$$\tilde{L}_R(w) := \mathop{\mathbb{E}}_{x,\epsilon,b}\left[h_R\left(\left\langle w - w^*, x - \mathbb{E}\left[x\right]\right\rangle - \langle w^*, \mu\rangle - \epsilon - b\right)\right] .$$

This function is not the one minimized by the algorithm, however, as we show in the appendix it is $(1-\alpha)\rho$-strongly convex and minimized by w^*. We then give the following upper bound:

$$\mathbb{E}\left[\tilde{L}_R(\bar{w}) - L_R(\bar{w})\right] \leq \mathcal{O}\left(\mathbb{E}\left[\|\bar{w} - w^*\|^2\right] + \frac{\log T}{T}\right). \tag{4}$$

The proof proceeds by showing that we can bound

$$\|\bar{w} - w^*\|^2 \leq \mathcal{O}\left(\tilde{L}_R(\bar{w}) - L_R(\bar{w}) + L_R(\bar{w}) - L_R(w^*)\right).$$

Then we utilize Eq. (4) together with the SGD guarantees to establish the proof. The full proof appears in Appendix F. □

6 EXTENSIONS

We present three extensions for the case where $\Sigma \succeq \rho \cdot I$: In Section 6.1 we go beyond the assumptions of bounded features and noise, and extend our results to the sub-Gaussian norm case. In Section 6.2, we point out to an algorithm that does not require knowledge of the fraction of contaminated samples α, but rather implicitly adapts to it. This improves over Algorithms 2 and 3 that require α. Finally, in Section 6.3, we describe an extension that does not require prior knowledge of $\|w^*\|$, but rather implicitly adapts to it.

6.1 Sub-Gaussian Norm Noise and Feature Vectors

We next relax the assumptions that the noise and the norm of the feature vectors are bounded with probability 1, and replace them with the following sub-Gaussian norm assumptions.

Assumption 6.1 (Sub-Gaussian Norm Feature Vector)**.** *The feature vector x has a sub-Gaussian norm distribution with variance proxy κ^2, that is $\mathbb{P}(\|x\| > u) \leq 2e^{-\frac{u^2}{2\kappa^2}}$ for all $u \in \mathbb{R}$.*

Assumption 6.2 (Sub-Gaussian Norm Noise)**.** *The noise ϵ has a sub-Gaussian norm distribution variance proxy σ^2, that is $\mathbb{P}(|\epsilon| > u) \leq 2e^{-\frac{u^2}{2\sigma^2}}$ for all $u \in \mathbb{R}$.*

Assumption 6.1 replaces Assumption 2.4 and Assumption 6.2 replaces Assumption 2.3. We show the following (The full proof appears in Appendix G).

Theorem 6.3. *Under Assumptions 6.1 and 6.2, define, $R = C \cdot (\kappa D + \sigma)\sqrt{\log T}$, where $C > 0$ is an explicit constant, which depends logarithmically on κ, ρ, σ. Denote $\bar{w}$ as the output of Algorithm 3 with input $(R, (1-\alpha)\rho, \mathcal{B}, T)$, then,*

$$\mathbb{E}\left[\|\bar{w} - w^*\|^2\right] \leq \left(D^2 + \frac{64R^2 \cdot (36\log T + 13)}{((1-\alpha)\rho)^2}\right) \cdot \frac{1}{T}.$$

The core idea of the proof is standard, and is based on reducing the sub-Gaussian feature and noise case to the bounded case previously analyzed. This is possible since it holds with a sufficiently high probability under the sub-Gaussian assumption that the feature and noise are bounded at all steps of the algorithm. Thus, the convergence analysis of the algorithm can condition on this high probability event. In turn, it is also shown that the low probability event that either the noise or the features have excessively large norm has a negligible effect on the total error. However, beyond the various technical details involved (see Appendix G), the conditioning on the high probability event may alter two properties that hold without this conditioning (regarding the minimal eigenvalue of the covariance matrix of the features, and the expected value of the noise). The proof of Theorem 6.3 addresses these delicate differences.

6.2 Adaptivity to Contamination Fraction α

In Section 5 we have assumed that α is known in order to find an accurate estimation of w^*. Our derivation shows that the expected Huber loss $L_R(\cdot)$ is $(1-\alpha)\rho$-strongly-convex, and we encode this information into the learning rate of SGD with $\frac{1}{2}$-suffix averaging that we employ. Indeed, Algorithms 2 and 3 require the strong-convexity parameter in order to ensure fast convergence.

In practice, it is unrealistic to assume that the fraction of contamination α is known. Fortunately, Cutkosky and Orabona (2018) have derived a novel and practical first order algorithm for stochastic convex optimization, that enables to implicitly adapt to the strong-convexity of the problem at hand. So, if we apply this algorithm instead of SGD with $\frac{1}{2}$-suffix averaging, then we immediately obtain adaptivity to both α and ρ. We elaborate on this approach in Appendix M.

6.3 Adaptivity to the norm of w^*

In Section 5 we have assumed that $\|w^*\| \leq D$ is known a priori to hold, and that knowledge of D was used to determine the Huber loss radius as $R = 6D + \sigma$. This assured strong convexity of the expected Huber loss $L_R(\cdot)$, even at the origin, and allowed for simple initialization of the algorithm at $w_1 = 0$. In this section we describe a proper algorithm for the case in which a tight bound on $\|w^*\|$ is not known a priori. The algorithm implicitly adapts to the unknown norm $\|w^*\|$, and to achieve an estimation error of ϵ it requires $T = \tilde{\Omega}\left(\frac{1}{((1-\alpha)\rho)^2}\left(\|w^*\|^2 + \frac{1}{\epsilon}\right)\right)$ samples and gradient computations. Next we describe our approach.

Our adaptive variant employs the Huber loss with $R := 6 + \sigma$. While this does not ensure strong-convexity around the origin, it does ensure the strong-convexity of $L_R(\cdot)$ at a ball of radius 1 around w^*. Then, at a high level, our algorithm is based on two SGD phases: **(i)** the first phase is initialized at the origin, and utilizes a recently developed adaptive

SGD algorithm Carmon and Hinder (2022) applied to the expected Huber loss. This algorithm essentially zooms in on the true norm of w^* by performing a line-search, and at the end of its run it is assured to output a solution $\bar{w}_1$ that lies at the strongly convex region of $L_R(\cdot)$, with high probability. **(ii)** At the second phase, a version of standard SGD for strongly convex function is initialized at $\bar{w}_1$. This assures fast convergence to w^*, and concretely the estimation error enjoys a convergence rate of $\tilde{O}(1/T)$. The analysis of this algorithm is based on *centered features assumption*, and appears in Appendix I.

7 EXPERIMENTS

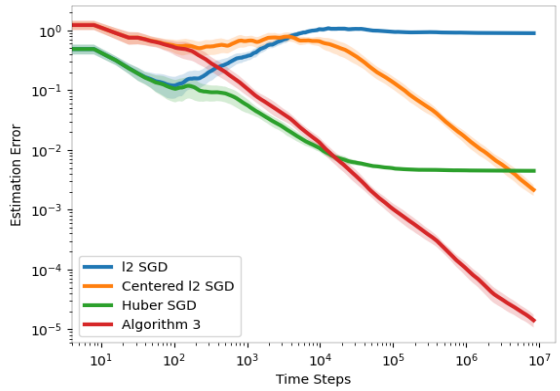

Figure 1: Results for $\alpha = 0.01$.

We perform two experiments under Model 5.5 in $d = 5$ dimensions. The optimum w^* is sampled from the unit ball and is the same for both experiments. We also let $\epsilon \sim \text{Uniform}[-0.1, 0.1]$, and $x \sim \text{Uniform}[-1/\sqrt{d}, 0]^d$ such that $\|x\| \leq 1$ w.p. 1. The adversary picks $b = 10^5$ with probability α and 0 otherwise. We test two cases $\alpha \in \{0.01, 0.7\}$ for four algorithms: Centered + Non-Centered Projected SGD over ℓ_2 loss, Huber SGD[1] and Centered Huber SGD (Algorithm 3) with radius parameter $R := 6D + \sigma$ for both. All of the methods use the same learning rate $\eta_t := 1/((1-\alpha)\rho t)$, and are given the same samples and observations for the gradient computation. We compute the estimate of $\mathbb{E}[x]$ on the fly rather than using extra samples (i.e., we use $\mu_t := 1/t \sum_{i=1}^{t-1} x_t$ instead of $\mu = 1/T \sum_{t=1}^{T} z_t$). We repeat each experiment 25 times and add confidence intervals. The experiments, shown in Figures 1 and 2, clearly demonstrate the benefit of our approach compared to the baselines. A comparison with additional algorithms can be found in Appendix L.

[1] Huber SGD is related to the application of Algorithm 2 directly to the Huber loss without first centering the features.

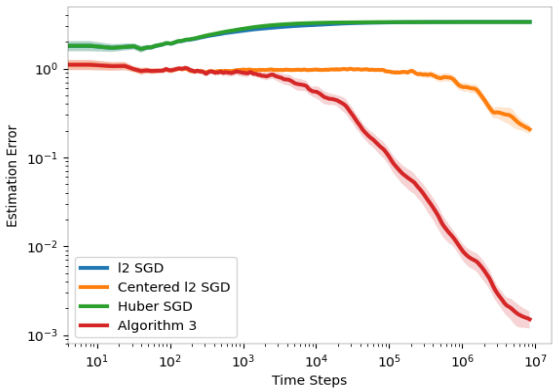

Figure 2: Results for $\alpha = 0.7$.

8 CONCLUSION

In this paper, we have analyzed robust linear regression under general assumptions on the feature vectors, noise and the oblivious adversary. We have shown that low prediction error requires either centered features or strict positivity of the features covariance matrix, and provided efficient SGD-style algorithms and established error rate convergence bounds. Finally, we have provided SGD variants that do not require prior knowledge on the fraction of contamination. Our work highlights that a combination of two basic techniques, namely using the Huber loss, together with feature centering is very powerful, and may find use even for more complex ML models. While the linear regression model is both basic and important, an important avenue for future work is to generalize the algorithms and the error bounds to more elaborated ML models.

9 ACKNOWLEDGMENTS

The research of K.Y. Levy was supported by the Israel Science Foundation, grant No. 447/20. The research of N. Weinberger was supported by the Israel Science Foundation, grant No. 1782/22.

References

G. Bassett Jr and R. Koenker. Asymptotic theory of least absolute error regression. *Journal of the American Statistical Association*, 73(363):618–622, 1978. URL https://www.jstor.org/stable/pdf/2286611.pdf.

E. J. Candes and T. Tao. Decoding by linear programming. *IEEE transactions on information theory*, 51(12):4203–4215, 2005. URL https://ieeexplore.ieee.org/iel5/18/32943/01542412.pdf.

Y. Carmon and O. Hinder. Making sgd parameter-free, 2022. URL https://arxiv.org/pdf/2205.02160.

N. Cesa-Bianchi and G. Lugosi. *Prediction, learning, and games*. Cambridge university press, Cambridge New York, 2006.

M. Charikar, J. Steinhardt, and G. Valiant. Learning from untrusted data. In *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing*, pages 47–60, 2017. URL https://dl.acm.org/doi/pdf/10.1145/3055399.3055491.

A. Cutkosky and F. Orabona. Black-box reductions for parameter-free online learning in banach spaces. In *Conference On Learning Theory*, pages 1493–1529. PMLR, 2018. URL http://proceedings.mlr.press/v75/cutkosky18a/cutkosky18a.pdf.

A. S. Dalalyan and P. Thompson. Outlier-robust estimation of a sparse linear model using ell_1-penalized huber's m-estimator. *arXiv preprint arXiv:1904.06288*, 2019. URL https://arxiv.org/pdf/1904.06288.

P. L. Davies, R. Fried, and U. Gather. Robust signal extraction for on-line monitoring data. *Journal of Statistical Planning and Inference*, 122(1-2):65–78, 2004. URL https://www.econstor.eu/bitstream/10419/77367/2/2002-02.pdf.

I. Diakonikolas, G. Kamath, D. Kane, J. Li, A. Moitra, and A. Stewart. Robust estimators in high-dimensions without the computational intractability. *SIAM Journal on Computing*, 48(2):742–864, 2019a. URL https://epubs.siam.org/doi/pdf/10.1137/17M1126680.

I. Diakonikolas, W. Kong, and A. Stewart. Efficient algorithms and lower bounds for robust linear regression. In *Proceedings of the Thirtieth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 2745–2754. SIAM, 2019b. URL https://epubs.siam.org/doi/pdf/10.1137/1.9781611975482.170.

T. D'Orsi, G. Novikov, and D. Steurer. Consistent regression when oblivious outliers overwhelm. In M. Meila and T. Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 2297–2306. PMLR, 18–24 Jul 2021. URL https://arxiv.org/pdf/2009.14774 and https://proceedings.mlr.press/v139/d-orsi21a.html.

C. F. Gauss. Least squares, 1809.

F. K. Gustafsson, M. Danelljan, and T. B. Schon. Evaluating scalable bayesian deep learning methods for robust computer vision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 318–319, 2020. URL https://openaccess.thecvf.com/content_CVPRW_2020/papers/w20/Gustafsson_Evaluating_Scalable_Bayesian_Deep_Learning_Methods_for_Robust_Computer_Vision_CVPRW_2020_paper.pdf.

E. Hazan. Introduction to online convex optimization, 2021. URL https://arxiv.org/pdf/1909.05207.

P. J. Huber. Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, 35(1):73–101, Mar. 1964. doi: 10.1214/aoms/1177703732. URL https://doi.org/10.1214/aoms/1177703732.

P. J. Huber. Robust regression: asymptotics, conjectures and monte carlo. *The annals of statistics*, pages 799–821, 1973. URL https://www.jstor.org/stable/pdf/2958283.pdf.

S. Kakade. Lecture notes in multivariate analysis, dimensionality reduction, and spectral methods, April 2010. URL https://homes.cs.washington.edu/~sham/courses/stat991_mult/lectures/MatrixConcen.pdf.

S. M. Kakade and A. Tewari. On the generalization ability of online strongly convex programming algorithms. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, editors, *Advances in Neural Information Processing Systems*, volume 21. Curran Associates, Inc., 2008. URL https://proceedings.neurips.cc/paper/2008/file/f90f2aca5c640289d0a29417bcb63a37-Paper.pdf.

A. Klivans, P. K. Kothari, and R. Meka. Efficient algorithms for outlier-robust regression. In *Conference On Learning Theory*, pages 1420–1430. PMLR, 2018. URL http://proceedings.mlr.press/v75/klivans18a/klivans18a.pdf.

L. Liu, T. Li, and C. Caramanis. High dimensional robust m-estimation: Arbitrary corruption and heavy tails. *arXiv preprint arXiv:1901.08237*, 2019. URL https://arxiv.org/pdf/1901.08237.

H. B. McMahan, G. Holt, D. Sculley, M. Young, D. Ebner, J. Grady, L. Nie, T. Phillips, E. Davydov, D. Golovin, et al. Ad click prediction: a view from the trenches. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1222–1230, 2013. URL https://dl.acm.org/doi/pdf/10.1145/2487575.2488200.

S. Pesme and N. Flammarion. Online robust regression via SGD on the ℓ_1 loss. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 2540–2552. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/1ae6464c6b5d51b363d7d96f97132c75-Paper.pdf.

D. Pollard. Asymptotics for least absolute deviation regression estimators. *Econometric Theory*, 7(2):186–199, 1991. URL http://www.math.pku.edu.cn/teachers/xirb/Courses/QR2013/Pollard91ET.pdf.

A. Rakhlin, O. Shamir, and K. Sridharan. Making gradient descent optimal for strongly convex stochastic optimization, 2012. URL http://arxiv.org/pdf/1109.5647.

P. J. Rousseeuw. Least median of squares regression. *Journal of the American statistical association*, 79(388): 871–880, 1984. URL https://www.jstor.org/stable/pdf/2288718.pdf.

P. J. Rousseeuw. Multivariate estimation with high breakdown point. *Mathematical statistics and applications*, 8(37):283–297, 1985. URL https://wis.kuleuven.be/stat/robust/papers/publications-1985/rousseeuw-multivariateestimationhighbreakdown-1985.pdf.

S. Shalev-Shwartz. *Online Learning and Online Convex Optimization*. now, 2012. ISBN 9781601985477.

A. S. Suggala, K. Bhatia, P. Ravikumar, and P. Jain. Adaptive hard thresholding for near-optimal consistent robust regression. In *Conference on Learning Theory*, pages 2892–2897. PMLR, 2019. URL http://proceedings.mlr.press/v99/suggala19a/suggala19a.pdf.

Q. Sun, W.-X. Zhou, and J. Fan. Adaptive huber regression. *Journal of the American Statistical Association*, 115(529):254–265, 2020. URL https://www.tandfonline.com/doi/pdf/10.1080/01621459.2018.1543124.

E. Tsakonas, J. Jalden, N. D. Sidiropoulos, and B. Ottersten. Convergence of the Huber regression m-estimate in the presence of dense outliers. *IEEE Signal Processing Letters*, 21(10):1211–1214, Oct. 2014. doi: 10.1109/lsp.2014.2329811. URL https://doi.org/10.1109/lsp.2014.2329811.

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

A. W. Van der Vaart. *Asymptotic statistics*, volume 3. Cambridge university press, 2000.

R. Vershynin. *High-dimensional probability : An introduction with applications in data science*. Cambridge University Press, Cambridge, United Kingdom New York, NY, 2018. ISBN 9781108231596.

M. S. Yeung, J. Tegnér, and J. J. Collins. Reverse engineering gene networks using singular value decomposition and robust regression. *Proceedings of the National Academy of Sciences*, 99(9):6163–6168, 2002. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC122920/.

A. Zaman, P. J. Rousseeuw, and M. Orhan. Econometric applications of high-breakdown robust regression techniques. *Economics Letters*, 71(1):1–8, 2001. URL https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2151062.

A Least Squares is Fragile in the Presence of Outliers

As the next simple example demonstrates, for any $\alpha \in (0,1)$ the adversary can make the estimation error of least squares arbitrarily large, even for infinite number of samples.

Example A.1. Let $w^* \in \mathbb{R}$, assume that $\epsilon = 0$ with probability 1, and that x is distributed uniformly over $\{0,2\}$. Also consider the following adversary, for $C > 0$:

$$b_i := \begin{cases} \frac{C}{\alpha} & \text{, with probability } \alpha \\ 0 & \text{, otherwise} \end{cases} \text{, } \forall i \in [1,\dots,T] \,.$$

Then $y_i = x_i \cdot w^* + b_i$ and on the population level (with infinite number of samples) the expected ℓ_2 loss is

$$\mathop{\mathbb{E}}_{x,\epsilon,b}\left[\tfrac{1}{2}\left(\langle w, x\rangle - y\right)^2\right] = \frac{1}{2}(w - w^*)^2 - C \cdot (w - w^*) + \frac{C^2}{2\alpha} \,.$$

Its minimal value is attained for $w_{min} = w^* + C$ and may be arbitrarily far from w^*.

B Other Methods Which Might Be Inefficient

B.1 Naive Filtering

If D, σ are known, then the scale of the measurements y is $\mathcal{O}(D + \sigma)$. Thus, one might think of the following naive filtering approach to robust regression: Set a threshold $\tau > D + \sigma$ (in order not to filter out non-contaminated points), filter out all data points (x_i, y_i) such that $|y_i| \geq \tau$, and apply standard linear regression to the remaining points.

We next demonstrate via a simple example that a vanilla application of this approach to robust regression fails.

Example B.1. Assume the following model $w^* = 1, \epsilon = 0$ with probability 1, and

$$x = \begin{cases} 1 & \text{, with probability } \frac{1}{2} \\ -1 & \text{, with probability } \frac{1}{2} \end{cases} \,,$$

and assume that the learner uses a threshold $\tau > 0$ for the filtering. Further consider an adversary which chooses its contamination as follows,

$$b = \begin{cases} \tau & \text{, with probability } \alpha \\ 0 & \text{, with probability } 1 - \alpha \end{cases} \,.$$

The learner knows that $\mathbb{E}[x] = 0$, and a bound D on w^*, taken w.l.o.g. to be $D = 1$, as well as a bound of 1 on the norm of the feature vector. The learner uses the naive filtering method, which means that he chooses $\tau > D$ and removes any pair (x_i, y_i) such that $|y_i| > \tau$ from the given samples.

It is immediate to show that the filtered data has the following distribution:

$$(x,y) = \begin{cases} (1,1) & \text{, with probability } \frac{1-\alpha}{2-\alpha} \\ (-1,-1) & \text{, with probability } \frac{1-\alpha}{2-\alpha} \\ (-1,\tau-1) & \text{, with probability } \frac{\alpha}{2-\alpha} \end{cases} \,.$$

Hence, an exact solution that minimizes the ℓ_2 loss would approach in the limit of large number of samples to

$$w^*_{\text{filtering}} = \frac{\mathbb{E}[x \cdot y]}{\mathbb{E}[x^2]} = 2 \cdot \frac{1-\alpha}{2-\alpha} \cdot 1 + \frac{\alpha}{2-\alpha} \cdot (\tau - 1) = 1 - \frac{\alpha\tau}{2-\alpha} = w^* - \frac{\alpha\tau}{2-\alpha} \,.$$

Thus, for any $\tau > D = 1$ and $\alpha > 0, w^*_{\text{filtering}} \neq w^*$.

Hence, this simple example demonstrates that naive filtering is inefficient, which reinforces the need to design more complex methods, as the one described in the paper.

B.2 Centering with Pairwise Differences

An assumption prevailing in other works on this problem is that the features are centered (i.e., have zero mean). In this paper we remove this assumption by an empirical centering of the features, to wit, reducing the empirical mean from the samples. An alternative approach, which is seemingly simpler, is to reduce the original model to a model with centered feature vectors by computing differences of pairs of adjacent samples. Specifically, such an estimator subtracts the first sample from the second one, the third from the fourth and so on. Formally, $\forall i \in [1, 2, \ldots, T/2]$ (and even T) we let

- $z_i := x_{2i} - x_{2i-1}$ be the equivalent feature vector;
- $e_i := \epsilon_{2i} - \epsilon_{2i-1}$ be the equivalent noise;
- $r_i := b_{2i} - b_{2i-1}$ be the equivalent contamination.

The resulting effective model is $y_{2i} - y_{2i-1} = \langle w^*, z_i \rangle + e_i$. The effective feature vector z_i has zero mean, and so this is a centered model. An additional benefit of this method is that the distribution of the effective total noise and contamination, $e_i + r_i$, is centered too, and is even symmetric around the origin because of the i.i.d. assumption. Nonetheless, this method for centering has a major drawback: The contamination probability α is altered by the pairwise difference operation. Specifically, let us assume that conditioned on the event $b_i \neq 0$, the corruptions b_i have a continuous distribution, and that $\alpha \in (0, 1)$. Then, the probability that the ith sample in the new model is not corrupted is

$$\begin{aligned}
\mathbb{P}(r_i \neq 0) &= 1 - \mathbb{P}(b_{2i} = 0 \cap b_{2i-1} = 0) + \mathbb{P}(b_{2i} \neq b_{2i-1} \cap b_{2i} \neq 0 \cap b_{2i-1} \neq 0) \\
&= 1 - (1-\alpha)^2 + 0 \\
&= 2\alpha - \alpha^2 \\
&> \alpha \,,
\end{aligned}$$

where the first inequality follows from the continuity of the distribution of the corruptions. The estimation error bounds derived in this paper (as well as all the other ones we mention in the introduction) attempt to establish an accurate dependence of the error on the contamination probability α. In this respect, the pairwise difference technique leads to strictly sub-optimal dependence of the convergence rate of the estimation error with respect to α. Indeed, instead of an error dependence of multiplicative factor $\frac{1}{(1-\alpha)^2}$ it will obtain a factor of $\frac{1}{(1-2\alpha+\alpha^2)^2}$. As α tends to 1 the ratio between these two factors becomes unbounded. On the other hand, if, say, $\alpha < \frac{1}{2}$, then this ratio is bounded by 4. Hence, up to this constant factor, the pairwise difference approach leads to an efficient centering technique, and consequently, the noise can be assumed symmetric. However, the centering technique is just a method to simplify the original problem, and there is still the need to solve the resulting problem. As said in the introduction, the algorithms proposed in previous works, require assumptions like spreadness, Gaussianity or positive definite covariance matrix. Our algorithm removes these assumptions.

C Proof of Lemma 3.1

Lemma C.1 (Lemma 3.1)**.** *Let $R := 6D + \sigma$. Then, $L_R(w) = (1-\alpha)F(w) + \alpha H(w)$, for all $w \in \mathcal{B}$ where*

$$H(w) := \mathop{\mathbb{E}}_{x,\epsilon,b}\left[h_R\left(\langle w - w^*, x\rangle - \epsilon - b\right) \middle| b \neq 0\right]$$

is the expected Huber loss of corrupted samples.

Proof. Take $w \in \mathcal{B}$. We use the law of total expectation with respect to b on $L_R(w)$. We start with conditioning on $b = 0$: For R as stated in the lemma,

$$\begin{aligned}
&\mathop{\mathbb{E}}_{x,\epsilon,b}\left[h_R\left(\langle w - w^*, x\rangle - \epsilon\right) \middle| b = 0\right] \\
&= \mathop{\mathbb{E}}_{x,\epsilon,b}\left[\frac{1}{2}\left(\langle w - w^*, x\rangle - \epsilon\right)^2 \middle| b = 0\right] \\
&= \mathop{\mathbb{E}}_{x,\epsilon}\left[\frac{1}{2}\left(\langle w - w^*, x\rangle - \epsilon\right)^2\right] \\
&= F(w) \,,
\end{aligned}$$

where the first equality follows from our choice of R and Equation (2). The second equality follows since b, ϵ and x are statistically independent and last equality follows from the definition of $F(w)$.
Then, by the law of total expectation

$$\begin{aligned} L_R(w) &= (1-\alpha) \mathop{\mathbb{E}}_{x,\epsilon,b} \left[h_R \left(\langle w - w^*, x \rangle - \epsilon \right) \Big| b = 0 \right] \\ &\quad + \alpha \mathop{\mathbb{E}}_{x,\epsilon,b} \left[h_R \left(\langle w - w^*, x \rangle - \epsilon - b \right) \Big| b \neq 0 \right] \\ &= (1-\alpha) F(w) + \alpha H(w) \, . \end{aligned}$$

□

D A Generalization of Lemma 5.2, Lemma 5.6 and Lemma F.3

Lemma D.1. *Given $v, q \in \mathbb{R}^d$ such that $\|v\|, \|q\| \leq 1$ and $R = 6D + \sigma$, let*

$$G_R(w) := \mathop{\mathbb{E}}_{x,\epsilon,b} \left[h_R \left(\langle w - w^*, x - v \rangle - \langle w^*, q \rangle - \epsilon - b \right) \right] \, .$$

Then, $G_R(\cdot)$ is $(1-\alpha)\rho$-strongly convex. Furthermore, if $v = \mathbb{E}[x]$, then $w^* = \arg\min_{w \in \mathcal{B}} G_R(w)$.

Proof. We show that $G_R(\cdot)$ is a sum of a $(1-\alpha)\rho$-strongly function and a convex function. As such, it is $(1-\alpha)\rho$-strongly convex.

Define

$$H(w) := \mathop{\mathbb{E}}_{x,\epsilon,b} \left[h_R \left(\langle w - w^*, x - v \rangle - \langle w^*, q \rangle - \epsilon - b \right) \Big| b \neq 0 \right] \, ,$$

which is a convex function as an average of convex functions. Also define

$$\begin{aligned} F(w) &:= \mathop{\mathbb{E}}_{x,\epsilon,b} \left[h_R \left(\langle w - w^*, x - v \rangle - \langle w^*, q \rangle - \epsilon - b \right) \Big| b = 0 \right] \\ &\overset{(a)}{=} \mathop{\mathbb{E}}_{x,\epsilon,b} \left[\frac{1}{2} \left(\langle w - w^*, x - v \rangle - \langle w^*, q \rangle - \epsilon \right)^2 \Big| b = 0 \right] \\ &\overset{(b)}{=} \mathop{\mathbb{E}}_{x,\epsilon} \left[\frac{1}{2} \left(\langle w - w^*, x - v \rangle - \langle w^*, q \rangle - \epsilon \right)^2 \right] \\ &= \mathop{\mathbb{E}}_{x,\epsilon} \left[\frac{1}{2} \langle w - w^*, x - v \rangle^2 \right] \\ &\quad + \mathop{\mathbb{E}}_{x,\epsilon} \left[\epsilon \cdot \langle w - w^*, x - v \rangle + \langle w^*, q \rangle \cdot \langle w - w^*, x - v \rangle \right] + S \, , \end{aligned}$$

where S is a constant independent of w, and where (a) follows from the boundness assumptions on x, v, q, ϵ and $R := 6D + \sigma$, and (b) follows from the assumption that x, ϵ, b are statistically independent.

$F(\cdot)$ is a polynomial and hence twice continuously differentiable. We take the second derivative and show it is positive definite,

$$\begin{aligned} \nabla^2 F(w) &= \mathop{\mathbb{E}}_{x} \left[(x - v)(x - v)^T \right] \\ &= \mathop{\mathbb{E}}_{x} \left[(x - \mathbb{E}[x] + \mathbb{E}[x] - v)(x - \mathbb{E}[x] + \mathbb{E}[x] - v)^T \right] \\ &= \mathop{\mathbb{E}}_{x} \left[(x - \mathbb{E}[x])(x - \mathbb{E}[x])^T + (\mathbb{E}[x] - v)(\mathbb{E}[x] - v)^T \right] \\ &\quad + \underbrace{\mathop{\mathbb{E}}_{x} \left[(x - \mathbb{E}[x])(\mathbb{E}[x] - v)^T + (\mathbb{E}[x] - v)(x - \mathbb{E}[x])^T \right]}_{=0} \\ &= \Sigma + (\mathbb{E}[x] - v)(\mathbb{E}[x] - v)^T \\ &\succeq \Sigma \, , \end{aligned}$$

where the last equality follows from the definition of Σ and the fact that v is a fixed vector.

So, Assumption 2.5 and Property 2.6 assure that $F(\cdot)$ is a ρ-strongly convex function. Then, by the law of total expectation with respect to b:

$$G_R(w) = (1-\alpha)F(w) + \alpha H(w) \ .$$

Since $H(\cdot)$ is convex and $(1-\alpha)F(\cdot)$ is $(1-\alpha)\rho$-strongly convex, $G_R(\cdot)$ is $(1-\alpha)\rho$-strongly convex.

If $v = \mathbb{E}[x]$ we can also show that $w^* = \arg\min_{w\in\mathcal{B}} G_R(w)$. For any $w \in \mathcal{B}$

$$\begin{aligned}
G_R(w) &:= \mathop{\mathbb{E}}_{x,\epsilon,b}\left[h_R\left(\left\langle w - w^*, x - \mathbb{E}[x]\right\rangle - \langle w^*, q\rangle - \epsilon - b\right)\right] \\
&\overset{(a)}{\geq} \mathop{\mathbb{E}}_{\epsilon,b}\left[h_R\left(\mathop{\mathbb{E}}_{x}\left[\left\langle w - w^*, x - \mathbb{E}[x]\right\rangle - \langle w^*, q\rangle - \epsilon - b\right]\right)\right] \\
&\overset{(b)}{=} \mathop{\mathbb{E}}_{\epsilon,b}\left[h_R\left(\left\langle w - w^*, \mathop{\mathbb{E}}_{x}\left[x - \mathbb{E}[x]\right]\right\rangle - \langle w^*, q\rangle - \epsilon - b\right)\right] \\
&= \mathop{\mathbb{E}}_{\epsilon,b}\left[h_R\left(-\langle w^*, q\rangle - \epsilon - b\right)\right] \\
&= \mathop{\mathbb{E}}_{\epsilon,b}\left[h_R\left(\left\langle w^* - w^*, x - \mathbb{E}[x]\right\rangle - \langle w^*, q\rangle - \epsilon - b\right)\right] \\
&= G_R(w^*) \ ,
\end{aligned}$$

where (a) follows from convexity of the Huber loss and Jensen's inequality and (b) follows from the linearity of the inner product. Moreover, w^* is the unique minimizer of $G_R(w)$ in $\mathcal{B}$. This is because according to Property 2.7, if $G_R(w) = G_R(w^*)$ for some $w \in \mathcal{B}$ then

$$0 = G_R(w) - G_R(w^*) \geq \frac{(1-\alpha)\rho}{2}\|w - w^*\|^2 \ .$$

Since $\frac{(1-\alpha)\rho}{2} > 0$ is assumed, this implies $w = w^*$. □

E Proof of Theorem 5.4

Theorem E.1 (Theorem 5.4)**.** *Let $\bar{w}$ be the output of Algorithm 2 with input $\left(R, (1-\alpha)\rho, T, \mathbb{E}[x]\right)$, then,*

$$\mathbb{E}\left[\|\bar{w} - w^*\|^2\right] \leq \frac{72R^2}{((1-\alpha)\rho)^2 \cdot T} \ .$$

Proof. Note that the g_t that we employ in Algorithm 2 is an unbiased gradient estimate for the Expected Huber loss $\nabla L_R(w_t)$ (recall the definition in Equation (3)). Also, since $\phi_R(\cdot)$ is bounded by R and the features' norms are bounded by 1, then $\|g_t\| \leq 2R$, for all t. Moreover, Lemma 5.2 implies that $L_R(\cdot)$ is $(1-\alpha)\rho$-strongly-convex with optimum $w^* \in \mathcal{B}$.

Now, since Algorithm 2 applies an SGD algorithm with strongly-convex assumption, and $\frac{1}{2}$-suffix averaging, then Lemma 5.3 with $G := 2R$ yields,

$$\mathbb{E}[L_R(\bar{w})] - L_R(w^*) \leq \frac{36R^2}{(1-\alpha)\rho \cdot T} \ .$$

Combining the latter with the strong-convexity of $L_R(\cdot)$, while using Property 2.7 establishes the theorem. □

F Proof of Theorem 5.7

Prior to proving Theorem 5.7 we need to introduce some definitions and establish some auxiliary lemmas.

Lemma F.1. *$h_R(\cdot)$ is $2R$-Lipschitz with probability* 1.

Proof. For any $w \in \mathcal{B}$

$$\begin{aligned}&\left\|\nabla h_R\left(\left\langle w-w^*, x-\mathbb{E}\left[x\right]\right\rangle - \langle w^*, \mu\rangle - \epsilon - b\right)\right\| \\ &= \left\|\phi_R\left(\left\langle w-w^*, x-\mathbb{E}\left[x\right]\right\rangle - \langle w^*, \mu\rangle - \epsilon - b\right)\cdot\left(x-\mathbb{E}\left[x\right]\right)\right\| \\ &\leq 2R\,,\end{aligned}$$

where the equality follows from definition and the inequality follows from the definition of $\phi_R(\cdot)$ and Assumption 2.4. The above holds with probability 1. □

The following is a version of Hoeffding's inequality for random vectors,

Lemma F.2 (Kakade (2010, Theorem 2.1))**.** *Assume that $\left\{x_i \in \mathbb{R}^d\right\}_{i=1}^T$ are random variables sampled i.i.d and $\|x_i\| \leq K$ almost surly. Then with probability $\geq 1-\delta$,*

$$\left\|\frac{1}{T}\sum_{i=1}^T x_i - \mathbb{E}\left[x\right]\right\| \leq 6K\sqrt{\frac{\log 1/\delta}{T}}\,.$$

The main challenge is that now w^* is not necessarily the optimum of $L_R(\cdot)$ in $\mathcal{B}$. To circumvent this, we will analyze an auxiliary expected loss function defined as

$$\tilde{L}_R(w) := \mathop{\mathbb{E}}_{x,\epsilon,b}\left[h_R\left(\left\langle w-w^*, x-\mathbb{E}\left[x\right]\right\rangle - \langle w^*, \mu\rangle - \epsilon - b\right)\right]\,.$$

This function is not the one minimized by the algorithm, however, it has the following desirable property:

Lemma F.3. *$\forall \mu \in \mathbb{R}^d:\ \tilde{L}_R(w)$ is $(1-\alpha)\rho$-strongly convex in $\mathcal{B}$ and $w^* = \arg\min_{w\in\mathcal{B}} \tilde{L}_R(w)$.*

Proof. Because μ is given, the proof is immediate from the more general Lemma D.1 with $v = \mathbb{E}\left[x\right]$ and $q = \mu$. □

Now, we can use Property 2.7 to bound the estimation error: for a given μ,

$$\begin{aligned}\frac{(1-\alpha)\rho}{2}\|\bar{w}-w^*\|^2 &\leq \tilde{L}_R(\bar{w}) - \tilde{L}_R(w^*) \\ &= \tilde{L}_R(\bar{w}) - L_R(\bar{w}) \\ &\quad + L_R(\bar{w}) - L_R(w^*) \\ &\quad + \underbrace{L_R(w^*) - \tilde{L}_R(w^*)}_{=0}\,,\end{aligned}$$

where the last term equals zero by the definitions of $L_R(w^*)$ and $\tilde{L}_R(w^*)$. So, by taking expectation and defining $C = \frac{(1-\alpha)\rho}{2}$, we obtain,

$$C\cdot\mathbb{E}\left[\|\bar{w}-w^*\|^2\right] \leq \mathbb{E}\left[\tilde{L}_R(\bar{w}) - L_R(\bar{w})\right] + \mathbb{E}\left[L_R(\bar{w}) - L_R(w^*)\right]\,, \tag{5}$$

We now bound each of these terms.

For the second term, since Algorithm 3 applies SGD to the strongly-convex function $L_R(\cdot)$, then,

$$\mathbb{E}\left[L_R(\bar{w}) - L_R(w^*)\right] \leq \frac{36R^2}{(1-\alpha)\rho\cdot T}\,,$$

due to Lemma 5.3, with the parameter choice $F = L_R, G = 2R, \lambda = (1-\alpha)\rho$.

The first term above can be upper bounded as follows:

Lemma F.4. *Let $C = \frac{(1-\alpha)\rho}{2}$. Then for a given μ,*

$$\mathbb{E}\left[\tilde{L}_R(\bar{w}) - L_R(\bar{w})\right] \leq \frac{C}{2}\cdot\mathbb{E}\left[\|\bar{w}-w^*\|^2\right] + \frac{2R^2}{C}\cdot\frac{36\log T + 4}{T}\,.$$

Proof. Take $C = \frac{(1-\alpha)\rho}{2}$. Then,

$$\begin{aligned}
&\tilde{L}_R(\bar{w}) - L_R(\bar{w}) \\
&\quad \overset{(a)}{=} \mathop{\mathbb{E}}_{x,\epsilon,b}\left[h_R\left(\left\langle \bar{w} - w^*, x - \mathbb{E}\left[x\right]\right\rangle - \langle w^*, \mu\rangle - \epsilon - b\right) - h_R\left(\langle \bar{w} - w^*, x - \mu\rangle - \langle w^*, \mu\rangle - \epsilon - b\right)\right] \\
&\quad \overset{(b)}{\leq} 2R \cdot \left|\left\langle \bar{w} - w^*, \mu - \mathbb{E}\left[x\right]\right\rangle\right| \\
&\quad \overset{(c)}{\leq} C \cdot \frac{2R}{C} \cdot \left\|\bar{w} - w^*\right\| \cdot \left\|\mu - \mathbb{E}\left[x\right]\right\| \\
&\quad \overset{(d)}{\leq} \frac{C}{2}\left\|\bar{w} - w^*\right\|^2 + \frac{2R^2}{C}\left\|\mu - \mathbb{E}\left[x\right]\right\|^2 ,
\end{aligned} \tag{6}$$

where (a) follows from the definitions of $\tilde{L}_R(\cdot)$ and $L_R(\cdot)$, (b) follows from Lemma F.1, (c) follows from the Cauchy-Schwarz's inequality and (d) follows from Young's inequality: $a \cdot b \leq \frac{1}{2}(a^2 + b^2)$ where $a = \left\|\bar{w} - w^*\right\|$ and $b = \frac{2R}{C} \cdot \left\|\mu - \mathbb{E}\left[x\right]\right\|$.

Note that $\bar{w}$ and μ are random as they depend on $\{(z_i, y_i)\}_{i=1}^T$ and $\{(x_t, y_t)\}_{t=1}^T$. By taking an expectation with respect to the $2T$ samples on both sides we have

$$\mathbb{E}\left[\tilde{L}_R(\bar{w}) - L_R(\bar{w})\right] \leq \mathbb{E}\left[\frac{C}{2}\left\|\bar{w} - w^*\right\|^2 + \frac{2R^2}{C}\left\|\mu - \mathbb{E}\left[x\right]\right\|^2\right] .$$

We conclude the proof by bounding $\mathbb{E}\left[\left\|\mu - \mathbb{E}\left[x\right]\right\|^2\right]$. We make use of Lemma F.2 by defining $\mathcal{U}$ as the event for which $\left\|\mu - \mathbb{E}\left[x\right]\right\| \leq 6\sqrt{\frac{\log 1/\delta}{T}}$. Then,

$$\begin{aligned}
\mathbb{E}\left[\left\|\mu - \mathbb{E}\left[x\right]\right\|^2\right] &\overset{(a)}{=} \mathop{\mathbb{E}}_{z_1,z_2,\ldots,z_T}\left[\left\|\mu - \mathbb{E}\left[x\right]\right\|^2\right] \\
&\overset{(b)}{\leq} \mathbb{P}\left(\mathcal{U}\right) \cdot \mathop{\mathbb{E}}_{z_1,z_2,\ldots,z_T}\left[\left\|\mu - \mathbb{E}\left[x\right]\right\|^2 \middle| \mathcal{U}\right] + \mathbb{P}\left(\mathcal{U}^c\right) \cdot \mathop{\mathbb{E}}_{z_1,z_2,\ldots,z_T}\left[\left\|\mu - \mathbb{E}\left[x\right]\right\|^2 \middle| \mathcal{U}^c\right] \\
&\overset{(c)}{\leq} (1-\delta) \cdot \frac{36\log 1/\delta}{T} + \delta \cdot 4 \\
&\overset{(d)}{\leq} \frac{36\log T + 4}{T} ,
\end{aligned}$$

where (a) follows from the i.i.d assumption on the features, (b) follows from the law of total expectation, (c) follows from the definition of $\mathcal{U}$ and Lemma F.2 (which is true for every $\delta \in (0,1)$) and a naive upper bound of $\left\|\mu - \mathbb{E}\left[x\right]\right\| \leq 2$ with probability 1 (which follows from Assumption 2.4) and (d) follows from taking $\delta = \frac{1}{T}$. Plugging the above into Equation (6) concludes the proof. □

We are now ready to prove Theorem 5.7,

Theorem F.5 (Theorem 5.7)**.** *Let $\bar{w}$ be the output of Algorithm 3 with input $(D, R, (1-\alpha)\rho, T)$, then,*

$$\mathbb{E}\left[\left\|\bar{w} - w^*\right\|^2\right] \leq \frac{16R^2 \cdot (36\log T + 13)}{((1-\alpha)\rho)^2 \cdot T} .$$

Proof. By plugging Lemma 5.3 and Lemma F.4 back into Equation (5), we obtain,

$$C \cdot \mathbb{E}\left[\left\|\bar{w} - w^*\right\|^2\right] \leq \frac{36R^2}{(1-\alpha)\rho \cdot T} + \frac{C}{2} \cdot \mathbb{E}\left[\left\|\bar{w} - w^*\right\|^2\right] + \frac{2R^2}{C} \cdot \frac{36\log T + 4}{T} .$$

Recalling $C = \frac{(1-\alpha)\rho}{2}$, the above implies,

$$\mathbb{E}\left[\left\|\bar{w} - w^*\right\|^2\right] \leq \frac{16R^2 \cdot (36\log T + 13)}{((1-\alpha)\rho)^2 \cdot T} ,$$

which establishes the theorem. □

G Proof of Theorem 6.3

Prior to proving Theorem 6.3 we need to introduce some definitions and establish some auxiliary lemmas.

Remark G.1. We have stated the sub-Gaussian norm assumptions (Assumption 6.1 and Assumption 6.2) in terms of the tails of the probability density functions. We refer the reader to Vershynin (2018, Chapter 2) for equivalent definitions of sub-Gaussian norm variables in terms of their moment generating function, or in terms of integer moments (all these definitions are essentially equivalent).

We will use the following bounds on the moments of sub-Gaussian norm random variables.

Lemma G.2. *Let z be a sub-Gaussian norm random variable with variance proxy λ^2, that is $\mathbb{P}\left(|z| > u\right) \leq 2e^{-\frac{u^2}{2\lambda^2}}$ for all $u \in \mathbb{R}$. Then, for any $p \geq 1$*

$$\mathbb{E}\left[|z|^p\right] \leq \sqrt{2\pi\lambda^2} \cdot p \cdot \lambda^p \frac{2^{p/2}\Gamma(\frac{p+1}{2})}{\sqrt{\pi}} \,,$$

where $\Gamma(\cdot)$ is the Gamma function. Specifically, $\mathbb{E}\left[|z|\right] \leq \sqrt{2\pi}\lambda$, $\mathbb{E}\left[|z|^2\right] \leq 4\lambda^2$, and $\mathbb{E}\left[|z|^4\right] \leq 24\lambda^4$.

Proof. Let $n \sim \mathcal{N}(0, \lambda^2)$. Then, it holds that

$$\begin{aligned}
\mathbb{E}\left[\|z\|^p\right] &\overset{(a)}{=} \int_0^\infty p u^{p-1} \mathbb{P}\left(\|z\| \geq u\right) du \\
&\overset{(b)}{\leq} \int_0^\infty p|u|^{p-1} 2e^{-\frac{u^2}{2\lambda^2}} du \\
&= \sqrt{2\pi\lambda^2} \cdot p \cdot \int_{-\infty}^\infty |u|^{p-1} \frac{1}{\sqrt{2\pi\lambda^2}} e^{-\frac{u^2}{2\lambda^2}} du \\
&= \sqrt{2\pi\lambda^2} \cdot p \cdot \mathbb{E}\left[|n|^{p-1}\right] \\
&\overset{(c)}{=} \sqrt{2\pi\lambda^2} \cdot p \cdot \lambda^p \frac{2^{p/2}\Gamma(\frac{p+1}{2})}{\sqrt{pi}} \,,
\end{aligned}$$

where (a) follows from the tail representation of the absolute moments $p \in (0, \infty)$ of a non-negative random variable z (Vershynin, 2018, Exercise 1.2.3), (b) follows from the sub-Gaussian assumption, (c) follows from the known formula of the central absolute moments of the Gaussian distribution. □

Lemma G.3. *Let $\mathcal{G}$ be an event such that $\mathbb{P}\left(\mathcal{G}\right) \leq \delta$ for some $\delta \in \left(0, \frac{1}{2}\right)$. Then,*

$$\tilde{\Sigma} := \mathbb{E}\left[\left(x - \mathbb{E}\left[x\middle|\mathcal{G}^c\right]\right)\left(x - \mathbb{E}\left[x\middle|\mathcal{G}^c\right]\right)^T\middle|\mathcal{G}^c\right] \succeq \left(\rho - 105\kappa^2\sqrt{\delta}\right) \cdot I \,.$$

Proof. Let $\tilde{\Sigma}$ be the conditional covariance matrix of the features. We first relate it to the unconditional covariance matrix by the decomposition

$$\begin{aligned}
\tilde{\Sigma} &= \mathbb{E}\left[\left(xx^T\right)\middle|\mathcal{G}^c\right] - \mathbb{E}\left[x\middle|\mathcal{G}^c\right]\mathbb{E}\left[x^T\middle|\mathcal{G}^c\right] \\
&\overset{(a)}{=} \frac{\mathbb{E}\left[xx^T\right] - \mathbb{E}\left[xx^T\mathbf{I}(\mathcal{G})\right]}{\mathbb{P}\left(\mathcal{G}^c\right)} - \mathbb{E}\left[x\middle|\mathcal{G}^c\right]\mathbb{E}\left[x^T\middle|\mathcal{G}^c\right] \\
&\overset{(b)}{=} \Sigma + \underbrace{\left(\frac{1}{\mathbb{P}\left(\mathcal{G}^c\right)} - 1\right)\mathbb{E}\left[xx^T\right]}_{:=G_1} + \underbrace{\mathbb{E}\left[x\right]\mathbb{E}\left[x^T\right] - \mathbb{E}\left[x\middle|\mathcal{G}^c\right]\mathbb{E}\left[x^T\middle|\mathcal{G}^c\right]}_{:=G_2} - \underbrace{\frac{\mathbb{E}\left[xx^T \cdot \mathbf{I}(\mathcal{G})\right]}{\mathbb{P}\left(\mathcal{G}^c\right)}}_{:=G_3} \,,
\end{aligned} \tag{7}$$

where (a) follows from the law of total expectation, (b) follows from the definition of the unconditional covariance matrix of the features $\Sigma := \mathbb{E}\left[xx^T\right] - \mathbb{E}\left[x\right]\mathbb{E}\left[x^T\right]$. For the matrix G_1 in Equation (7), the assumptions $\Sigma \succeq \rho \cdot I$ and $\delta \leq \frac{1}{2}$ imply that

$$G_1 \succeq 0 \,. \tag{8}$$

We next bound the maximal value of $\left|v^T G_2 v\right|$ and $\left|v^T G_3 v\right|$ over all unit vectors $v \in \mathbb{R}^d$ (with $\|v\| = 1$). For G_2, we further decompose to

$$\begin{aligned} G_2 &= \mathbb{E}[x]\,\mathbb{E}\left[x^T\right] - \mathbb{E}\left[x\middle|\mathcal{G}^c\right]\mathbb{E}\left[x^T\middle|\mathcal{G}^c\right] \\ &\overset{(a)}{=} \underbrace{\mathbb{E}[x]\left(\mathbb{E}\left[x^T\right] - \mathbb{E}\left[x^T\middle|\mathcal{G}^c\right]\right)}_{:=G_{2,1}} + \underbrace{\left(\mathbb{E}[x] - \mathbb{E}\left[x\middle|\mathcal{G}^c\right]\right)\mathbb{E}\left[x^T\middle|\mathcal{G}^c\right]}_{:=G_{2,2}}, \end{aligned}$$

where (a) follows by adding and subtracting the common term $\mathbb{E}[x]\,\mathbb{E}\left[x^T\middle|\mathcal{G}^c\right]$. Now, for any $v \in \mathbb{R}^d$ with $\|v\| = 1$, it holds that

$$\begin{aligned} \left|v^T G_{2,1} v\right| &= \left|v^T \mathbb{E}[x]\left(\mathbb{E}\left[x^T\right] - \mathbb{E}\left[x^T\middle|\mathcal{G}^c\right]\right)v\right| \\ &\overset{(a)}{\leq} \left\|\mathbb{E}[x]\right\| \cdot \left\|\mathbb{E}[x] - \mathbb{E}\left[x\middle|\mathcal{G}^c\right]\right\| \\ &\overset{(b)}{=} \mathbb{E}[x] \cdot \frac{\left\|(\mathbb{P}(\mathcal{G}^c) - 1)\mathbb{E}[x] - \mathbb{E}[x \cdot \mathbf{I}(\mathcal{G})]\right\|}{\mathbb{P}(\mathcal{G}^c)} \\ &\overset{(c)}{\leq} \mathbb{E}[\|x\|] \cdot \frac{\left\|(\mathbb{P}(\mathcal{G}^c) - 1)\mathbb{E}[x] - \mathbb{E}[x \cdot \mathbf{I}(\mathcal{G})]\right\|}{\mathbb{P}(\mathcal{G}^c)} \\ &\overset{(d)}{\leq} \mathbb{E}[\|x\|] \cdot \frac{(1 - \mathbb{P}(\mathcal{G}^c))\mathbb{E}[\|x\|] + \mathbb{E}[\|x\| \cdot \mathbf{I}(\mathcal{G})]}{\mathbb{P}(\mathcal{G}^c)} \\ &\overset{(e)}{\leq} \mathbb{E}[\|x\|] \cdot \frac{(1 - \mathbb{P}(\mathcal{G}^c))\mathbb{E}[\|x\|] + \sqrt{\mathbb{E}\left[\|x\|^2\right] \cdot \mathbb{P}(\mathcal{G})}}{\mathbb{P}(\mathcal{G}^c)} \\ &\overset{(f)}{\leq} 18\kappa^2 \cdot \sqrt{\delta}\,, \end{aligned} \tag{9}$$

where (a) follows from Cauchy-Schwarz's inequality, (b) follows from the law of total expectation, (c) follows from Jensen's inequality, (d) follows from the triangle inequality and Jensen's inequality, (e) follows from Cauchy-Schwarz's inequality, (f) follows from the assumptions that x is sub-Gaussian with variance parameter κ^2 and Lemma G.2, with the assumption $\mathbb{P}(G) = \delta \leq \frac{1}{2}$ (and as $\delta < \sqrt{\delta}$).

For $G_{2,2}$ we use a similar bounding method, except that now the bound on $\left\|\mathbb{E}[x]\right\|$ is replaced by a bound on $\left\|\mathbb{E}\left[x\middle|G^c\right]\right\|$ (in Equation (9)). This conditional expectation can be bounded as follows:

$$\begin{aligned} \left\|\mathbb{E}\left[x\middle|G^c\right]\right\| &\overset{(a)}{=} \frac{\left\|\mathbb{E}[x] + \mathbb{E}[x \cdot \mathbf{I}(\mathcal{G})]\right\|}{\mathbb{P}(\mathcal{G}^c)} \\ &\overset{(b)}{\leq} \frac{\mathbb{E}[\|x\|] + \sqrt{\mathbb{E}\left[\|x\|^2\right] \cdot \mathbb{P}(\mathcal{G})}}{\mathbb{P}(\mathcal{G}^c)} \\ &\overset{(c)}{\leq} \sqrt{8\pi}\kappa + 4\kappa\sqrt{\delta}\,, \end{aligned}$$

where (a) follows from the law of total expectation, (b) follows by similar steps in the analysis of $G_{2,1}$, using the triangle, Jensen's and Cauchy-Schwarz's inequalities, and (c) follows from Lemma G.2 and the assumptions. With this bound, as in the analysis of $G_{2,1}$, it holds for any $v \in \mathbb{R}^d$ with $\|v\| = 1$ that

$$\left|v^T G_{2,2} v\right| \leq 80\kappa^2\sqrt{\delta}\,,$$

using $\delta\sqrt{\delta} \leq \delta \leq \sqrt{\delta}$.

From the bounds on $\left|v^T G_{2,1} v\right|$ and $\left|v^T G_{2,2} v\right|$ we deduce that

$$\left|v^T G_2 v\right| = \left|v^T (G_{2,1} + G_{2,2}) v\right| \leq \left|v^T G_{2,1} v\right| + \left|v^T G_{2,2} v\right| \leq 98\kappa^2\sqrt{\delta}\,. \tag{10}$$

For G_3 it holds for any $v \in \mathbb{R}^d$ with $\|v\| = 1$ that

$$\begin{aligned}\left|v^T G_3 v\right| &= \frac{\left|\mathbb{E}\left[v^T x x^T v \cdot \mathbf{I}(\mathcal{G})\right]\right|}{\mathbb{P}\left(\mathcal{G}^c\right)} \\ &\overset{(a)}{\leq} \frac{\mathbb{E}\left[\|x\|^2 \cdot \mathbf{I}(\mathcal{G})\right]}{\mathbb{P}\left(\mathcal{G}^c\right)} \\ &\overset{(b)}{\leq} \frac{\sqrt{\mathbb{E}\left[\|x\|^4\right] \cdot \mathbb{P}\left(\mathcal{G}\right)}}{\mathbb{P}\left(\mathcal{G}^c\right)} \\ &\overset{(c)}{\leq} 7\kappa^2\sqrt{\delta}\,, \end{aligned} \tag{11}$$

where (a) follows from Cauchy-Schwarz's inequality in $\mathbb{R}^d$, (b) follows from Cauchy-Schwarz's inequality in L_2, and (c) follows from Lemma G.2 and the assumptions. Using the decomposition of $\tilde{\Sigma}$ in Equation (7) and the bounds on G_1, G_2, G_3 in Equation (8), Equation (10) and Equation (11), respectively, it holds for any $v \in \mathbb{R}^d$ with $\|v\| = 1$ that

$$v^T \Sigma v \geq \rho + 0 - 98\kappa^2\sqrt{\delta} - 7\kappa^2\sqrt{\delta}\,,$$

which directly implies to the stated claim. □

We are now ready to prove Theorem 6.3,

Theorem G.4 (Theorem 6.3)**.** *Given Assumption 6.1 and Assumption 6.2. Define*

$$R := 2\sqrt{8\kappa^2 \log\left(\left(\frac{21\kappa}{\sqrt{\rho}} + 8\right) \cdot T\right)} \cdot D + \sqrt{8\sigma^2 \log\left(\left(\frac{21\sigma}{\sqrt{\rho}} + 8\right) \cdot T\right)}\,.$$

Then,

$$\mathbb{E}\left[\|\bar{w} - w^*\|^2\right] \leq \left(D^2 + \frac{64R^2 \cdot (36\log T + 13)}{((1-\alpha)\rho)^2}\right) \cdot \frac{1}{T}\,.$$

Proof. Let a time T be given, and consider the events

$$\mathcal{F}_x(u) := \left\{ \bigcup_{i \in \{1,2,\ldots,T\}} \|x_i\| > u \right\}\,,$$

and

$$\mathcal{F}_\epsilon(u) := \left\{ \bigcup_{i \in \{1,2,\ldots,T\}} |\epsilon_i| > u \right\}\,.$$

Further let $u_1 := \sqrt{2\kappa^2 \log(\frac{4T}{\delta})}$ and $u_2 := \sqrt{2\sigma^2 \log(\frac{4T}{\delta})}$ and set

$$\mathcal{F} := \mathcal{F}_\epsilon(u_1) \cup \mathcal{F}_x(u_2)\,.$$

By a union bound over $i \in [1, 2, \ldots, T]$ and computing probabilities over ϵ and x, the sub-Gaussian assumptions implies that $\mathbb{P}\left(\mathcal{F}\right) \leq \delta$. We choose $\delta = \min\left\{\frac{1}{2}, \frac{\rho^2}{210^2\kappa^4} \cdot \frac{1}{T^2}\right\}$. Note, that with this choice of δ, and by identifying $\mathcal{G} = \mathcal{F}$, Lemma G.3 implies that $\tilde{\Sigma} \succeq \frac{\rho}{2} \cdot I$ for all $T \geq 1$.

We next evaluate the error of the SGD algorithm by considering two events – the event $\mathcal{F}^c$ in which both $\|x_i\|$ and $|\epsilon_i|$ are bounded for all $i \in \{1, 2, \ldots, T\}$, and the event $\mathcal{F}$, which has a vanishing probability $\delta = O(T^{-2})$. Specifically, by the law of total expectation

$$\mathbb{E}\left[\|\bar{w} - w^*\|^2\right] = \mathbb{P}\left(\mathcal{F}\right) \cdot \mathbb{E}\left[\|\bar{w} - w^*\|^2 \,\middle|\, \mathcal{F}\right] + \mathbb{P}\left(\mathcal{F}^c\right) \cdot \mathbb{E}\left[\|\bar{w} - w^*\|^2 \,\middle|\, \mathcal{F}^c\right]\,, \tag{12}$$

where $\mathcal{F}^c$ is the complement of the event $\mathcal{F}$. The first term in Equation (12) is upper bounded as follows:

$$\begin{aligned}\mathbb{P}(\mathcal{F})\,\mathbb{E}\left[\|\bar{w}-w^*\|^2\,\middle|\,\mathcal{F}\right] &\overset{(a)}{\leq} \sqrt{\mathbb{P}(\mathcal{F})}\cdot 4D^2\\ &\overset{(b)}{\leq} \frac{4D^2\rho}{210\kappa^2}\cdot\frac{1}{T}\\ &\overset{(c)}{\leq} \frac{D^2}{T}\,,\end{aligned} \tag{13}$$

where (a) follows from the fact that $0\leq\mathbb{P}(\mathcal{F})\leq 1$ and $\bar{w},w^*\in\mathcal{B}$, (b) follows since $\sqrt{\mathbb{P}(\mathcal{F})}\leq\sqrt{\delta}$ and the choice of δ, and (c) follows since

$$\rho\leq\max_{v\in\mathbb{R}^d:\,\|v\|\leq 1}\mathbb{E}\left[\left\langle v,x-\mathbb{E}[x]\right\rangle^2\right]\leq\mathbb{E}\left[\left\|x-\mathbb{E}[x]\right\|^2\right]\leq\mathbb{E}\left[\|x\|^2\right]\leq 4\kappa^2\,.$$

For the second term in Equation (12), we note that conditioned on $\mathcal{F}^c$ the noise and the feature vectors are bounded, that is $\|x_i\|\leq u_1$ and $|\epsilon_i|\leq u_2$ for all $i\in\{1,2,\ldots,T\}$. This model is similar to the one discussed in previous sections, in particular to Model 5.5 in Section 5.2, where the expectation is unknown, with two differences. First, as said, by the choice of δ the conditional covariance matrix of the features has minimal eigenvalue of $\frac{\rho}{2}$, instead of ρ for the unconditional covariance matrix. The second difference is that $\mathbb{E}\left[\epsilon\middle|\mathcal{F}^c\right]$ may not equal zero. However, it can be easily verified that the result of Section 5.1 holds, since the noise related terms in the second derivative of the function $F(\cdot)$ therein vanish.

We will follow the same steps as in Section 5.2: computing μ with T samples conditioned on $\mathcal{F}^c$ and feed them to Algorithm 3 with different input, that is because the radius parameter R, is different and will depend on u_1 and u_2.

The derivation of the new radius parameter R is similar to derivation made in Section 3, that is bounding the norm of the features and noise by u_1 and u_2 respectively. By our choice of $\delta=\min\left\{\frac{1}{2},\frac{\rho^2}{210^2\kappa^4}\cdot\frac{1}{T^2}\right\}$ we may further bound

$$u_1=\sqrt{2\kappa^2\log\left(4\cdot\max\left\{2T,\frac{210^2\kappa^4T^3}{\rho^2}\right\}\right)}\leq\sqrt{8\kappa^2\log\left(\left(\frac{21\kappa}{\sqrt{\rho}}+8\right)\cdot T\right)}\,,$$

and similarly

$$u_2\leq\sqrt{8\sigma^2\log\left(\left(\frac{21\sigma}{\sqrt{\rho}}+8\right)\cdot T\right)}\,.$$

Then, by taking

$$R:=2\sqrt{8\kappa^2\log\left(\left(\frac{21\kappa}{\sqrt{\rho}}+8\right)\cdot T\right)}\cdot D+\sqrt{8\sigma^2\log\left(\left(\frac{21\sigma}{\sqrt{\rho}}+8\right)\cdot T\right)}\,,$$

it holds by Theorem 5.7 that

$$\begin{aligned}\mathbb{P}(\mathcal{F}^c)\,\mathbb{E}\left[\|\bar{w}-w^*\|^2\,\middle|\,\mathcal{F}^c\right] &\leq \frac{16R^2\cdot(36\log T+13)}{((1-\alpha)\frac{\rho}{2})^2\cdot T}\\ &= \frac{64R^2\cdot(36\log T+13)}{((1-\alpha)\rho)^2\cdot T}\,,\end{aligned} \tag{14}$$

where we use $0\leq\mathbb{P}(\mathcal{F}^c)\leq 1$.

Note that w^* is deterministic and does not change when conditioned of $\mathcal{F}^c$, then by combining the bounds in Equations (13) and (14) into Equation (12) completes the proof. □

H Discussing Our Results with respect to the Spreadness Assumption

The recent paper of D'Orsi et al. (2021) assumes that the empirical design matrix of the features satisfies a condition termed *spreadness*. This condition essentially states that the energy of each vector in the column space of the design matrix cannot be too concentrated on a small subset of its coordinates. It is stated in D'Orsi et al. (2021, Appendix A.2.1) that

the spreadness assumption is necessary in order to obtain meaningful guarantees for any algorithm, i.e., a hardness result. In contrast, our work shows that one can obtain meaningful guarantees without making this assumption. In this section, we settle the apparent discrepancy between the results, by showing that the worst-case distribution used by D'Orsi et al. (2021) to establish the hardness result, does not take into account the norm of the optimal solution $\|w^*\|$. We then show when this problem parameter is taken into account, an adversary cannot inflict an arbitrarily large error to a learner, but is rather limited to an error of $\mathcal{O}\left(\frac{\|w^*\|^2}{T}\right)$.

We will not elaborate further on the spreadness assumption since it is unnecessary for establishing our point. Instead, we will describe worst-case example used in D'Orsi et al. (2021) to prove the hardness result, and show that it is irrelevant when taking $\|w^*\|$ into account.

The Example of D'Orsi et al. (2021): Assume an oblivious contamination as in Model 2.1 such that $\epsilon = 0$ with probability 1, the features are 1 dimensional, i.e. $x \in \mathbb{R}$, and the contaminations are distributed as follows,[2]

$$\begin{cases} b_i \sim \mathcal{N}(0, \sigma^2) & \text{, with probability } \alpha \\ b_i = 0 & \text{, otherwise} \end{cases} \text{, } \forall i \in [1, \dots, T] \text{ ,}$$

here the fraction of contamination is α, and the adversary injects Gaussian noise with variance σ^2; note that the adversary may choose σ to be arbitrarily large. It is also assumed that the number of non-zero features in the training set is equal to $\frac{1}{C(1-\alpha)}$, [3] where $C > 0$ is a large enough constant that does not depend on α, T. With these assumptions, their hardness proof is based on the following statement (Lemma A.5 therein):

Lemma H.1. *[D'Orsi et al. (2021, Lemma A.5)] Consider the robust linear regression task with the above assumptions. Then for any estimate $\hat{w}$ which is based on the contaminated data, there exists an optimal solution w^* (of the problem with non-corrupted data) such that the following holds,*

$$\frac{1}{T}\sum_{i=1}^{T} \|\langle x_i, \hat{w} - w^* \rangle\|^2 \geq \Omega\left(\frac{\sigma^2}{T}\right) \ .$$

Seemingly, the above Lemma implies that since the adversary can choose σ to be arbitrarily large this means that one cannot obtain meaningful guarantees in this case.[4] Nevertheless, as we show below, by explicitly taking the norm of the optimal solution into account, we show that the adversary's power is limited to inflicting a bounded error of $\mathcal{O}\left(\frac{\|w^*\|^2}{T}\right)$.

Lemma H.2. *Consider the robust linear regression task with the above assumptions. Also assume that the features are bounded, i.e., $\|x\| \leq 1$ with probability 1. Then there exists a trivial estimator $\hat{w}$ such that the following holds,*

$$\frac{1}{T}\sum_{i=1}^{T} \|\langle x_i, \hat{w} - w^* \rangle\|^2 \leq \mathcal{O}\left(\frac{\|w^*\|^2}{(1-\alpha)T}\right) \ . \tag{15}$$

Conclusion: Importantly, the above lemma implies that for this example there exists an estimator that achieves a vanishing prediction error that *does not depend on σ^2, but rather depends on $\|w^*\|^2$*. Moreover, the above lemma implies that in the worst case example of D'Orsi et al. (2021) (see Lemma H.1 above), the choice of w^* given σ is such that $\sigma^2 \leq \|w^*\|^2$.

Thus, even without the spreadness assumption, an adversary cannot inflict unbounded errors, and meaningful guarantees are actually possible.

Next we prove the above lemma.

Proof of Lemma H.2. Consider the following trivial predictor $\hat{w} = 0$. In this case, since the features are bounded by 1, and

[2] We stick to the notations in our paper where α is the fraction of contaminated samples. Conversely, D'Orsi et al. (2021) denote the fraction of contaminated samples by $1 - \alpha$.

[3] In the example appearing in (D'Orsi et al., 2021, Appendix A.2.1) there is a typo where they mistakenly write that number of non-zero features is $\frac{T}{C(1-\alpha)}$, but it should actually be $\frac{1}{C(1-\alpha)}$. We validated this with the authors of D'Orsi et al. (2021).

[4] Note that the choice of the optimal solution w^* is allowed to depends on the magnitude of the adversary's noise σ (see D'Orsi et al. (2021, Fact A.1)).

only $\frac{1}{C(1-\alpha)}$ of them are non-zero we obtain,

$$\begin{aligned}\frac{1}{T}\sum_{i=1}^{T}\|\langle x_i,\hat{w}-w^*\rangle\|^2 &= \frac{1}{T}\sum_{i=1}^{T}\|\langle x_i,w^*\rangle\|^2\\ &\le \frac{1}{T}\sum_{i=1}^{T}\|x_i\|^2\cdot\|w^*\|^2\\ &\le \frac{\|w^*\|^2}{T}\sum_{i=1}^{T}\|x_i\|^2\\ &\le \frac{C\|w^*\|^2}{(1-\alpha)T}\,,\end{aligned}$$

here the first line uses $\hat{w}=0$, the second line uses Cauchy-Schwarz's inequality, and the last line uses $\|x_i\|\le 1$, and the fact that only $1/C(1-\alpha)$ of them are non zero. □

Combining Lemma H.1 and Lemma H.2, we obtain

$$\Omega\left(\frac{\sigma^2}{T}\right)\le\frac{1}{T}\sum_{i=1}^{T}\|\langle x_i,\hat{w}-w^*\rangle\|^2\le\mathcal{O}\left(\frac{\|w^*\|^2}{(1-\alpha)T}\right)\,.$$

Thus, taking the norm of w^* into account reveals that the adversary's power is restricted in this example, and fast rates of convergence are possible, even without the spreadness assumption.

I Adaptivity to the Norm of the Optimal Solution

In this section, we extend our results to the strongly-convex case (i.e., $\Sigma\succeq\rho\cdot I$) where *the norm of the optimal solution* $\|w^*\|$ *is unknown* and $\mathbb{E}[x]=0$. The main result of this section appears in Theorem M.1, and our overall adaptive algorithm is depicted in Algorithm 6.

The High Level Approach When a bound D on the norm of w^* was assumed to be known, we had set the radius of the Huber loss to be $R:=6D+\sigma$. This ensured that the expected Huber loss $L_R(\cdot)$ is strongly-convex in the ball of radius D around the origin. This knowledge of D simplified the algorithm's operation and its analysis.

In this section, we take a slightly different approach. Since we do not have a bound on $\|w^*\|$, we take $R:=6+\sigma$. This ensures that $L_R(\cdot)$ is strongly-convex in a ball of radius 1 around w^*. This choice complicates our algorithm and analysis, yet it enables to obtain adaptivity to $\|w^*\|$, as well as to gain better dependence on $\|w^*\|$ compared to the analysis in the main text (i.e. the case where a bound on $\|w^*\|$ is known).

At a high level, our algorithmic approach is to apply SGD in two phases (see Algorithm 6):

1. In the *first phase* we start at $w_0=0$ and apply a recent variant of SGD (Algorithm 4) that implicitly adapts to the norm of the optimal solution (Carmon and Hinder, 2022). At the end of this phase we are ensured to lie in a small enough ball around w^* where $L_R(\cdot)$ is strongly-convex.

2. In the *second phase* we apply a variant of SGD for the strongly-convex case (Algorithm 7) that ensures fast convergence to w^*. This phase is initiated from the output of the first phase.

Next we provide the full details of algorithms and analysis.

Choice of R and strong-convexity of $L_R(\cdot)$ As we mentioned, we choose $R:=6+\sigma$. The next lemma shows that in this case, $L_R(\cdot)$ is $(1-\alpha)\rho$ strongly convex in a ball of radius 1 around w^*, that is, in $\mathcal{B}^*:=\left\{w\in\mathbb{R}^d:\|w-w^*\|\le 1\right\}$, and that w^* is its global minimum.

Lemma I.1. *Let $R:=6+\sigma$, then, $L_R(\cdot)$ is $(1-\alpha)\rho$-strongly convex in $\mathcal{B}^*$ and $w^*=\arg\min L_R(w)$.*

The proof of the above Lemma appears in Appendix I.1.

Algorithm 4 Parameter-Free SGD

Input: $\eta_\epsilon > 0, T \in \mathbb{N}_+, \left\{\gamma^{(k)}, \beta^{(k)}\right\}_{k=2,4,8,\ldots}$
for $k = 2, 4, 8, \ldots$ **do**
  if $k > T/4$ **then return** w_0
  $T_k = \left\lfloor \frac{T}{2k} \right\rfloor$
  $\eta_o :=$ Algorithm 5 with input $\left(\eta_\epsilon, 2^{2^k}\eta_\epsilon; T_k, \gamma^{(k)}, \beta^{(k)}\right)$
  if $\eta_o < \infty$ **then return** $\frac{1}{T_k}\sum_{i<T_k} w_i(\eta_o)$
end for

Algorithm 5 Root Finding Bisection

Input: $\eta_{lo}, \eta_{hi}; \tau, \gamma, \beta$
if $\eta_{hi} \le \phi(\eta_{hi})$ **then return** ∞ ▷ $\phi(\cdot)$ defined in Equation (18)
if $\eta_{lo} > \phi(\eta_{lo})$ **then return** η_{lo}
while $\eta_{hi} > 2\eta_{lo}$ **do**
  $\eta_{mid} := \sqrt{\eta_{lo}\eta_{hi}}$
  if $\eta_{mid} \le \phi(\eta_{mid})$ **then** $\eta_{lo} := \eta_{mid}$ **else** $\eta_{hi} := \eta_{mid}$
end while
if $\bar{r}_\tau(\eta_{hi}) \le \bar{r}_\tau(\eta_{lo})\frac{\phi(\eta_{hi})}{\eta_{hi}}$ **then return** η_{hi} **else return** η_{lo} ▷ $\bar{r}_\tau(\cdot)$ defined in Equation (16)

Phase 1 SGD. In the first phase of the optimization we apply a recent variant of SGD for stochastic convex optimization problems that implicitly adapts to $\|w_0 - w^*\|$, where w_0 is the initialization point and w^* is the optimal solution. This variant is due to Carmon and Hinder (2022), and is depicted in Algorithm 4.

Below we provide a version of Theorem 2 of Carmon and Hinder (2022) that applies to $L_R(\cdot)$. We plug in the fact that $w_0 = 0$, and also use the convexity $L_R(\cdot)$, as well as the fact that the stochastic gradients that we employ (based on the contaminated samples) have a norm bounded by R, which is due to the use of the Huber loss.

Root Finding Bisection: Here we explain the *Root Finding Bisection* procedure used in Algorithm 4, as well as the notations therein.

Whenever the function $\phi(\eta)$ is mentioned in the algorithm, the following is performed:

- Unconstrained SGD with a learning rate of η is being applied for τ steps starting at $w_0 = 0$. We denote its iterates by $w_i(\eta)$, and the stochastic gradient that it queries by $g_i(\eta)$. Thus, $w_0(\eta) = 0$, and,

$$w_{i+1}(\eta) := w_i(\eta) - \eta g_i(\eta) \ ,$$

  where $g_i(\eta)$ is a stochastic gradient estimate taken at $w_i(\eta)$.

- Based on this SGD run we compute the following expressions,

$$\bar{r}_\tau(\eta) := \max_{i \le \tau} \|w_i(\eta)\| \ . \tag{16}$$

$$G_\tau(\eta) := \sum_{i<\tau} \|g_i(\eta)\|^2 \ . \tag{17}$$

$$\phi(\eta) := \frac{\bar{r}_\tau(\eta)}{\sqrt{\gamma G_\tau(\eta) + \beta}} \quad \text{for } \gamma, \beta > 0 \ . \tag{18}$$

We are now ready to state the guarantees of Algorithm 4 applied to the Huber loss $L_R(\cdot)$ (while using samples from the contaminated data). The theorem below is a simplification of the more general (Carmon and Hinder, 2022, Theorem 2).

Theorem I.2. *Let $L_R : \mathbb{R}^d \mapsto \mathbb{R}$ be the expected Huber loss with parameter $R > 0$. Then applying Algorithm 4 to $L_R(\cdot)$ using $T/2$ samples, initial point $w_0 = 0$, and with the following choices:*

- $\eta_\epsilon := \frac{\epsilon}{R^2 T}$
- $\gamma^{(k)} = 32^2 C_k$, *and* $\beta^{(k)} = (32RC_k)^2$, *where* $C_k = 2k + \log\left(60\log^2(3T)/\delta\right)$

ensures that the following holds with probability $\geq 1-\delta$,

$$L_R(\bar{w}_1) - L_R(w^*) \leq M_0 \left(\frac{R\|w^*\|}{\sqrt{T}} + \frac{\epsilon}{T}\right) \cdot \chi^2 \,,$$

where M_0 *is a universal constant, independent of the problem's parameters,* $\chi := \log\left(\frac{1}{\delta}\log_+\left(R\|w^*\|T/\epsilon\right)\right)$ *and* $\log_+(z) := \max\{2, \log(z)\}$.

The parameter $\epsilon > 0$ can be tuned by the algorithm, and we choose $\epsilon = 1/T$.[5] We also remark that Carmon and Hinder (2022, Theorem 2) is more general, yet above we stated a simplified version that is relevant to our application.

Combing the above Theorem with the strong-convexity of $L_R(\cdot)$ in $\mathcal{B}^*$ immediately implies that after phase 1 the iteration is ensured to lie in a ball radius of $1/3$ around w^*. This is formalized below.

Corollary I.3. *Assume*

$$T \geq \Omega\left(\sqrt{\frac{\log^2\left(\log(T\|w^*\|)/\delta\right)}{(1-\alpha)\rho}} + \frac{\|w^*\|^2 \log^2\left(\log(T\|w^*\|)/\delta\right)}{(1-\alpha)^2\rho^2}\right) \,.$$

Then, applying Algorithm 4 using $T/2$ *samples with the choices mentioned in Theorem I.2 and* $\epsilon = 1/T$, *ensures that* $\bar{w}_1$, *the output of Algorithm 4, satisfies*

$$\|\bar{w}_1 - w^*\| \leq \frac{1}{3}$$

with probability $\geq 1-\delta$.

Proof. **Part (a).** Here we show that for any u such $\|u - w^*\| = 1/3$ then its sub-optimality is bounded from below. Indeed, since $L_R(\cdot)$ is $(1-\alpha)\rho$-strongly convex in $\mathcal{B}^*$, and its global minimum is w^* (Lemma I.1), then for any such u we have,

$$L_R(u) - L_R(w^*) \geq \frac{(1-\alpha)\rho}{2}\|u - w^*\|^2 = \frac{(1-\alpha)\rho}{18} \,.$$

Part (b). Here we show that the sub-optimality of any w such $w \notin \mathcal{B}_{1/3}(w^*) := \left\{w \in \mathbb{R}^d : \|w - w^*\| \leq 1/3\right\}$ is bounded from below. Indeed, for any such w there exists u such that $\|u - w^*\| = 1/3$, and also such that,

$$w - u = (\theta - 1)(u - w^*) \,, \tag{19}$$

for some $\theta > 1$. This vector u is actually the point on the line segment $[w^*, w]$ whose distance from w^* is exactly $1/3$. We can now show that $L_R(w) \geq L_R(u)$,

$$\begin{aligned} L_R(w) - L_R(u) &\overset{(a)}{\geq} \langle \nabla L_R(u), w - u\rangle \\ &\overset{(b)}{=} (\theta - 1)\langle \nabla L_R(u), u - w^*\rangle \\ &\overset{(c)}{\geq} 0 \,, \end{aligned}$$

where (a) follows from the convexity of $L_R(\cdot)$, (b) follows from Equation (19) and (c) follows from the optimality of w^* and the fact that $\theta > 1$.

Part (c). From parts **(a)** and **(b)** it follows that

$$L_R(w) - L_R(w^*) \geq \frac{(1-\alpha)\rho}{18}$$

[5] There is nothing special about this choice and we could alternatively take $\epsilon = 1$ or say $\epsilon = 1/T^5$ without qualitatively affecting the result by much.

for any $w \notin \mathcal{B}_{1/3}(w^*)$. On the other hand, Theorem I.2 ensures that $\bar{w}_1$ satisfies with probability $\geq 1-\delta$ (taking $\epsilon = 1/T$),

$$L_R(\bar{w}_1) - L_R(w^*) \leq M_0 \left(\frac{R\,\|w^*\|}{\sqrt{T}} + \frac{1}{T^2} \right) \cdot \chi^2 \,.$$

Combining the above immediately implies that

$$\|\bar{w}_1 - w^*\| \leq \frac{1}{3} \,,$$

with probability $\geq 1-\delta$, for all large enough T, and concretely, for

$$T \geq \Omega \left(\sqrt{\frac{\log^2\left(\log(T\|w^*\|)/\delta\right)}{(1-\alpha)\rho}} + \frac{\|w^*\|^2 \log^2\left(\log(T\|w^*\|)/\delta\right)}{(1-\alpha)^2\rho^2} \right) \,.$$

□

Algorithm 6 Diameter Adaptive Huber SGD

Input: $\eta_\epsilon, \left\{\gamma^{(k)}, \beta^{(k)}\right\}_{k=2,4,8,\ldots}, \sigma > 0, \lambda > 0, T \in \mathbb{N}_+$

Phase 1: $\bar{w}_1 :=$ output of Algorithm 4 with input $\left(\eta_\epsilon, \frac{T}{2}, \left\{\gamma^{(k)}, \beta^{(k)}\right\}_{k=2,4,8,\ldots}\right)$ ▷ applied to Huber Loss $L_R(\cdot)$, and using samples from the contaminated model

Phase 2: $\bar{w}_2 :=$ output of Algorithm 7 with input $\left(\bar{w}_1, 6+\sigma, \lambda, \frac{T}{2}\right)$ ▷ applied to Huber Loss $L_R(\cdot)$, and using samples from the contaminated model

Return: $\bar{w}_2$

Algorithm 7 SGD for Strongly Convex Function

Input: $w_1 \in \mathbb{R}^d, R > 0, \lambda > 0, T \in \mathbb{N}_+$
$\mathcal{W} := \left\{w \in \mathbb{R}^d : \|w - w_1\| \leq \frac{2}{3}\right\}$
for $t = 1$ **to** T **do**
  Draw (x_t, y_t) from Model 2.1
  $\eta_t := 1/\lambda t$
  $g_t := \phi_R\left(\langle w_t, x_t\rangle - y_t\right) \cdot x_t$
  $w_{t+1} := \Pi_{\mathcal{W}}\left(w_t - \eta_t \cdot g_t\right)$
end for
Return: $\bar{w} := \frac{1}{T}\sum_{t=1}^{T} w_t$

Phase 2 SGD. Corollary I.3 ensures that the output of Phase 1, i.e. $\bar{w}_1$, is inside the ball of radius $1/3$ around w^*, with probability $\geq 1-\delta$. As can be seen from Algorithm 7, in Phase 2 we take a ball of radius $2/3$ around $\bar{w}_1$ and apply SGD for strongly-convex functions constrained to that ball. The constraint to this ball has two desired properties: **(i)** it contains the global minimum w^* (due to the guarantee of phase 1), and **(ii)** $L_R(\cdot)$ is $(1-\alpha)\rho$-strongly-convex in this ball (since $L_R(\cdot)$ is strongly-convex over $\mathcal{B}^*$). Thus, we have fast convergence to w^*. This is formalized in the next theorem, which is the main theorem of this section.

Theorem I.4. *Denote $\bar{w}_2$ as the output of Algorithm 6, then, with probability $\geq (1-\delta)\cdot(1-4\delta\log T)$, with T as defined in Corollary I.3, then*

$$\|\bar{w}_2 - w^*\|^2 \leq \mathcal{O}\left(\left(\frac{R\left(\sqrt{\log T} + \sqrt{\log 1/\delta}\right)}{(1-\alpha)\rho} \right)^2 \right) \cdot \frac{1}{T} \,.$$

Proof. For $R := 6+\sigma$, it holds that $L_R(\cdot)$ is R-Lipschitz and $(1-\alpha)\rho$-strongly convex in $\mathcal{B}^*$ with w^* as its minimum

(Lemma I.1). Then, with probability $(1-\delta)\cdot(1-4\delta\log T)$,

$$\begin{aligned}\|\bar{w}_2 - w^*\|^2 &\leq \frac{2}{(1-\alpha)\rho}\left(L_R(\bar{w}_2) - L_R(w^*)\right)\\ &\leq \mathcal{O}\left(\left(\frac{R\left(\sqrt{\log T}+\sqrt{\log 1/\delta}\right)}{(1-\alpha)\rho}\right)^2\right)\cdot\frac{1}{T}\,,\end{aligned}$$

where the first inequality follows from the fact that $\bar{w}_2 \in \mathcal{B}^*$, and the strong convexity of $L_R(\cdot)$, combined with Property 2.7, and the second inequality follows from Corollary J.2 with $L := R, \nu := (1-\alpha)\rho, w_1 := \bar{w}$ and $D := 2/3$. □

Theorem I.5. *Let $\epsilon > 0, \delta \in (0,1)$ and denote $\bar{w}_2$ as the output of Algorithm 6. If*

$$T \geq \tilde{\Omega}\left(\frac{(6+\sigma)^2}{((1-\alpha)\rho)^2}\left(\|w^*\|^2\log^2(1/\delta) + \frac{\log(1/\delta)}{\epsilon}\right)\right)\,,$$

Then, $\|\bar{w}_2 - w^*\|^2 \leq \epsilon$ with probability larger than $(1-\delta)\cdot(1-4\delta\log T)$, where $\tilde{\Omega}$ hides logarithmic dependence in $1/\epsilon$.

Proof. This follows immediately from Corollary I.3 and Theorem I.4. □

I.1 Proof of Lemma I.1

Proof. We show that $L_R(\cdot)$ is a sum of a $(1-\alpha)\rho$-strongly function and a convex function and as such it is a $(1-\alpha)\rho$-strongly convex. This is done in a similar way to the proof of Lemma D.1.

Define

$$H(w) := \mathop{\mathbb{E}}_{x,\epsilon,b}\left[h_R\left(\langle w - w^*, x\rangle - \epsilon - b\right)\Big| b \neq 0\right]\,,$$

which is convex as an average of convex functions.

Given that $b = 0$, it holds for any $w \in \mathcal{B}^*$ that

$$|\langle w - w^*, x\rangle - \epsilon| \leq \|w - w^*\|\cdot\|x\| + |\epsilon| \leq 6 + \sigma\,, \tag{20}$$

which follows from the definition of $\mathcal{B}^*$ and the boundness of x, ϵ. Now, define

$$\begin{aligned}F(w) &:= \mathop{\mathbb{E}}_{x,\epsilon,b}\left[h_R\left(\langle w - w^*, x\rangle - \epsilon - b\right)\Big| b = 0\right]\\ &\overset{(a)}{=} \mathop{\mathbb{E}}_{x,\epsilon,b}\left[\frac{1}{2}\left(\langle w - w^*, x\rangle - \epsilon\right)^2\Big| b = 0\right]\\ &\overset{(b)}{=} \mathop{\mathbb{E}}_{x,\epsilon}\left[\frac{1}{2}\left(\langle w - w^*, x\rangle - \epsilon\right)^2\right]\\ &= \mathop{\mathbb{E}}_{x,\epsilon}\left[\frac{1}{2}\langle w - w^*, x\rangle^2 - \epsilon\cdot\langle w - w^*, x\rangle + \frac{1}{2}\epsilon^2\right]\,,\end{aligned}$$

where (a) follows from Equation (20) and $R := 6 + \sigma$ and (b) follows from the assumption that x, ϵ, b are statistically independent. The Hessian matrix of $F(\cdot)$ is given by $\nabla^2 F(w) = \mathop{\mathbb{E}}_{x}\left[xx^T\right] = \Sigma$ which is, by assumption, positive definite.

So, Assumption 2.5 and Property 2.6 assure that $F(\cdot)$ is a ρ-strongly convex function. Then, by the law of total expectation with respect to b:

$$L_R(w) = (1-\alpha)F(w) + \alpha H(w)\,.$$

Since $H(\cdot)$ is convex and $(1-\alpha)F(\cdot)$ is $(1-\alpha)\rho$-strongly convex, $L_R(\cdot)$ is $(1-\alpha)\rho$-strongly convex.

Next we show that $w^* = \arg\min L_R(w)$. Indeed For any $w \in \mathbb{R}^d$

$$\begin{aligned}
L_R(w) &:= \mathop{\mathbb{E}}_{x,\epsilon,b}\left[h_R\left(\langle w - w^*, x\rangle - \epsilon - b\right)\right] \\
&\overset{(a)}{\geq} \mathop{\mathbb{E}}_{\epsilon,b}\left[h_R\left(\mathop{\mathbb{E}}_{x}\left[\langle w - w^*, x\rangle - \epsilon - b\right]\right)\right] \\
&\overset{(b)}{=} \mathop{\mathbb{E}}_{\epsilon,b}\left[h_R\left(\left\langle w - w^*, \mathbb{E}\left[x\right]\right\rangle - \epsilon - b\right)\right] \\
&\overset{(c)}{=} \mathop{\mathbb{E}}_{\epsilon,b}\left[h_R\left(-\epsilon - b\right)\right] \\
&= \mathop{\mathbb{E}}_{\epsilon,b}\left[h_R\left(\langle w^* - w^*, x\rangle - \epsilon - b\right)\right] \\
&= L_R(w^*)\,,
\end{aligned}$$

where (a) follows from convexity of the Huber loss and Jensen's inequality, (b) follows from the linearity of the inner product and (c) from the fact that $\mathbb{E}\left[x\right] = 0$. Moreover, due to the strong-convexity of $L_R(w)$ in $\mathcal{B}^*$, then w^* is the unique minimizer of $L_R(w)$.

□

J High Probability SGD for Strongly Convex Function

Lemma J.1 (Adjusted Theorem 2 from Kakade and Tewari (2008)). *Define $\mathcal{W} := \left\{w \in \mathbb{R}^d : \|w - w_1\| \leq D\right\}$ for some $w_1 \in \mathbb{R}^d$. Let Z be a random variable taking values in some space $\mathcal{Z}$. Let $f : \mathcal{W} \times \mathcal{Z} \mapsto \mathbb{R}$ be L-Lipschitz and let $F(w) := \mathbb{E}\left[f(w; z)\right]$ be ν-strongly convex. Denote $w^* := \arg\min_{w\in\mathcal{W}} F(w)$ as the minimizer of F. Also, assume that we apply a regret minimization algorithm $\mathcal{A}$ over the sequence of losses $\{f(\cdot, Z_t)\}_{t=1}^T$, and $\mathcal{A}$ yields predictions $w_1, \ldots, w_T$. Then, for $\delta \in [0, 1)$, with probability $\geq 1 - 4\delta \log T$,*

$$\frac{1}{T}\sum_{t=1}^{T} F(w_t) - F(w^*) \leq \frac{C(L, \nu)}{T}$$

where

$$C(L, \nu) := Reg_T + 4L\sqrt{\frac{\log 1/\delta \cdot Reg_T}{\nu}} + \max\left\{\frac{15L^2}{\nu}, 24LD\right\} \cdot \log 1/\delta$$

and $Reg_T := \sum_{t=1}^T f(w_t; Z_t) - \min_{w\in\mathcal{W}} \sum_{t=1}^T f(w; Z_t)$ is the regret of the online $\mathcal{A}$ algorithm being used.

Further, using Jensen's inequality, $\frac{1}{T}\sum_{t=1}^T F(w_t)$ can be replaced by $F\left(\frac{1}{T}\sum_{t=1}^T w_t\right)$.

The following is an immediate corollary of the above lemma.

Corollary J.2. *Under the same conditions of Lemma J.1, and by taking the online algorithm $\mathcal{A}$ to be OGD with a learning rate of $\eta_t = \frac{1}{\nu t}$, it holds with probability $1 - 4\delta \log T$ that*

$$F(\bar{w}_T) - F(w^*) \leq \mathcal{O}\left(\frac{L^2}{\nu}\log T + \frac{L^2}{\nu}\sqrt{\log T \cdot \log 1/\delta} + \max\left\{\frac{L^2}{\nu}, LD\right\} \cdot \log 1/\delta\right) \cdot \frac{1}{T}\,,$$

where $\bar{w}_T := \frac{1}{T}\sum_{t=1}^T w_t$.

The corollary follows immediately from combining Lemma J.1 with the regret guarantees of OGD for the strongly-convex case (which uses $\eta_t = \frac{1}{\nu t}$), that ensures $\text{Reg}_T \leq O\left(\frac{L^2}{\nu}\log T\right)$ (see e.g. Hazan (2021, Theorem 3.3)).

Proof of Lemma J.1. This lemma is a re-statement of the original theorem in Kakade and Tewari (2008), with the minor difference that in Kakade and Tewari (2008) it is assumed that each $f(\cdot; Z)$ is strongly-convex, whereas here we only assume the strong-convexity of $F(\cdot)$. All other assumptions are identical.

We thus next describe the (minor) changes needed for this adjusted theorem:
Proof of (Kakade and Tewari, 2008, Lemma 1)
The argument stated therein

$$\frac{f(w;Z)+f(w';Z)}{2} \geq f\left(\frac{w+w'}{2};Z\right)+\frac{\nu}{8}\left\|w-w'\right\|^2,$$

is not needed, since we assume here that F is strongly convex, and so the line that follows in their proof, to wit,

$$\frac{F(w)+F(w')}{2} \geq f\left(\frac{w+w'}{2};Z\right)+\frac{\nu}{8}\left\|w-w'\right\|^2,$$

holds by the assumed strong convexity of F.

Proof of (Kakade and Tewari, 2008, Theorem 2): The proof hinges on proving various properties of the random variable

$$\xi_t := F(w_t)-F(w^*)-\left(f(w_t;Z_t)-f(w^*;Z_t)\right) .$$

Specifically, the assumption therein is that $f:\mathcal{W}\times\mathcal{Z}\mapsto[0,B]$, and this immediately implies that $|\xi_t|\leq 2B$.

Here, we do not make a direct assumption that f is bounded. However, the requires modification is minor. We know that $f(\cdot)$ is L-Lipschitz, and thus so is $F(\cdot)$. Then, since $\|w^*\|\leq D$, the triangle inequality implies

$$|\xi_t| \leq |F(w_t)-F(w^*)|+|f(w_t;Z_t)-f(w^*;Z_t)| \leq 2L\|w_t-w^*\| \leq 4LD ,$$

which is the required boundedness property of the ξ_t's. □

K Extension of Impossibility result to General α

Here we show that in the general convex case, when $\mathbb{E}[x]\neq 0$, then for any value of $\alpha\in[0,1]$, there does not exist a learner that can ensure a prediction error better than $\Omega(\alpha^2)$.

Recall that in our setting we assume that the learner does not know the distribution of the features nor of the measurement noise ϵ, or of the contamination b. Thus, in order to prove the impossibility result we will construct two different models with two different measurement noise distributions and two different contamination distributions and show that after being contaminated, no learner can distinguish between these models. We then show the latter to imply that no learner can obtain a better prediction error than $\Omega(\alpha^2)$. Next we provide the details.

The impossibility result. Let $\alpha\geq 0$ be the fraction of samples that the adversary may contaminate. Now consider two one-dimensional models with parameter vectors w_1^*, w_2^*, which will be determined later. For both models we assume that $x=1$, thus $\mathbb{E}[x]$ is known and equals 1. Next we show how to construct these two models such that after being contaminated by an adversary, one cannot distinguish between them.

Model 1: for model $m=1$ we take $w_1^*=1$, and $\epsilon_i^{(1)}=0\ ;\forall i$. We also define the adversarial perturbation $b_i^{(1)}$ to be a Bernoulli random variable with parameter α, thus $b_i^{(1)}=1$ with probability α, and is equal to zero otherwise. Therefore for any sample i in model 1 we have,

$$y_i^{(1)} = w_1^*\cdot x_i+\epsilon_i^{(1)}+b_i^{(1)} = 1+b_i^{(1)} , \tag{21}$$

where $b_i^{(1)}\sim \text{Ber}(\alpha)$ is the adversarial perturbation injected to the sample i.

Model 2: Now for model $m=2$ we take $w_2^*=1+\alpha$, and define noise for this model as follows,

$$\epsilon_i^{(2)} = Z_i-\alpha$$

where $Z_i\sim\text{Ber}(\alpha)$, and therefore $\mathbb{E}[\epsilon_i^{(2)}]=0$. For this model we also choose the adversarial perturbations to be always zero, i.e., $b_i^{(2)}=0\ ;\forall i$. Consequently, for any sample i in model 2 we can write,

$$y_i^{(2)} = w_2^*\cdot x_i+\epsilon_i^{(2)}+b_i^{(2)} = (1+\alpha)+(Z_i-\alpha)+0 = 1+Z_i , \tag{22}$$

where $Z_i \sim \text{Ber}(\alpha)$.

Thus, from Equations (21) and (22), it is clear that one cannot distinguish between models 1 and 2 from their α-contaminated samples.

Now since this is the case, any learner will output the same solution irrespective of the model from which he observes the contaminated samples. Let us denote this solution by w. Now if the model that we present to the learner is either model 1 or 2 with equal probability, then his expected prediction error (expected excess loss) in this case would be,

$$\begin{aligned}\text{ExpectedExccessLoss}(w) &= \frac{1}{2}(F_1(w) - \min_v F_1(v)) + \frac{1}{2}(F_2(w) - \min_v F_2(v)) \\ &= \frac{1}{2}\cdot\frac{1}{2}(w-1)^2 + \frac{1}{2}\cdot\frac{1}{2}(w-(1+\alpha))^2\end{aligned}$$

where $F_1(\cdot)$ and $F_2(\cdot)$ are the expected losses for models 1 and 2 over the clean (non-contaminated) data, and the second line follows by a straightforward computation. Now if we minimize the above objective as a function of w, we conclude that its minimum is obtained in $w_{\text{opt}} = 1 + \alpha/2$. Thus for any solution w we have,

$$\text{ExpectedExccessLoss}(w) \geq \text{ExpectedExccessLoss}(w_{\text{opt}}) = \frac{1}{8}\alpha^2 \,.$$

Therefore, any learner must incur a prediction error of at least $\Omega(\alpha^2)$ in this case.

L Experiment With More Algorithms

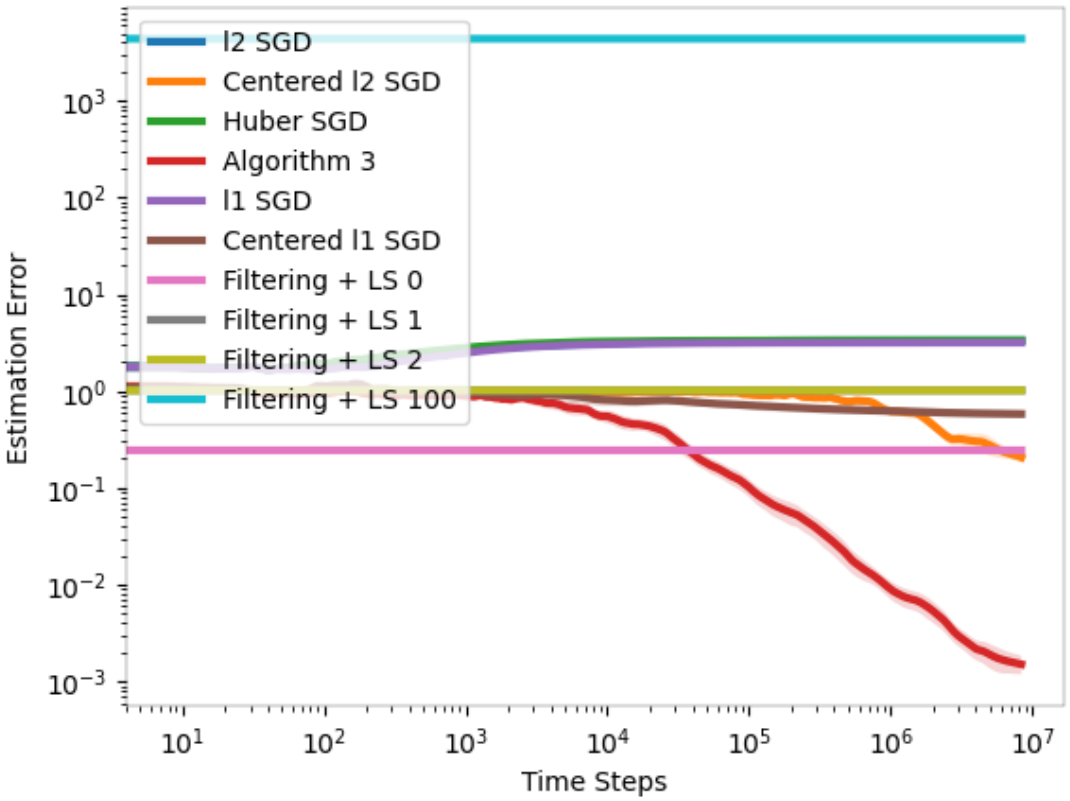

Figure 3: Results for $\alpha = 0.7$.

We perform one more experiment, similar to the experiments in Section 7, where $\alpha = 0.7$, but now we change the distribution of b and consider more algorithms: Centered + Non-Centered projected SGD over the ℓ_1 loss and Filtering + Least Squares (with different filtering thresholds).

Adversary The corruption b is chosen as such that if $b \neq 0$, then $b = 1$ w.p. $\frac{1}{2}$ and 100 otherwise.

Filtering + Least Squares Any sample whose (absolute value of its) label is larger than the filtering threshold is removed (we apply the filter approach with 5 different threshold levels, appearing as "Filtering + LS" algorithms. The threshold value is chosen to be $a \cdot D + \sigma$ where a is the number appearing in the legend).[6] Then, we use the remaining samples to compute the least squares solution: $(XX^T)^{-1}Xy$. This gives us one solution but we chose to show this as a line in the figure in order to see the result compared to the other algorithms across all steps.

Note that in Figure 3, the error for $a \in \{1, 2, 5, 10, 50\}$ is the same as the same samples are filtered.

[6] The result for $a = 100$ was omitted as its error is larger than 1000, and distorts the graph.

M Adaptivity to α

Algorithm 8 Adapting to Curvature

Require: Baseline Online learning algorithm $\mathcal{A}$
Initialize: W a convex close set in a reflexive Banach space, $\bar{x}_0$ an arbitrary point in W.
for $t = 1$ to T **do**
  Get point w_t from $\mathcal{A}$
  Set $z_t := w_t + \bar{x}_{t-1}$
  Play $x_t \in \Pi_W(z_t)$, receive subgradient $g_t \in \partial \ell_t(x_t)$
  Set $\tilde{g}_t \in g_t + \|g_t\|_* \partial S_W(z_t)$; where $S_W(z) := \inf_{u \in W} \|u - z\|$
  Set $\bar{x}_t := \frac{\bar{x}_0 + \sum_{i=1}^t \|\tilde{g}_i\|_*^2 x_i}{1 + \sum_{i=1}^t \|\tilde{g}\|_*^2}$
  Send $\tilde{g}_t$ to $\mathcal{A}$ as the t-th subgradient
end for

As mentioned in Section 6.2, the expected Huber loss $L_R(\cdot)$ is $(1-\alpha)\rho$-strongly-convex, and we encode this information into the learning rate of the SGD variants that we employ (Algorithms 2 & 3). In the recent paper of Cutkosky and Orabona (2018), an online convex optimization algorithm was presented that adapts to the strong-convexity parameter, and so does not need to know the value of α. In our analysis of Algorithm 3 the choice of the $\frac{1}{2}$-suffix averaging SGD can be easily switched with any SGD for strongly-convex function that has an error of $\tilde{\mathcal{O}}\left(\frac{1}{T}\right)$, such is Algorithm 8, which has the following guarantees:

Theorem M.1 (Cutkosky and Orabona 2018, Theorem 7)**.** *Let $\{\ell_t(\cdot)\}_{t=1,\ldots}$ be a sequence of convex loss functions. Let $\mathcal{A}$ be an online linear optimization algorithm that outputs w_t in response to g_t. Also assume that $\mathcal{W}$ is a convex closed set of diameter D. Suppose $\mathcal{A}$ guarantees for all T and $\mathring{v}$:*

$$\sum_{t=1}^{T} \left\langle w_t - \mathring{v}, \tilde{g}_t \right\rangle \leq \eta + \left\|\mathring{v}\right\| A \sqrt{\sum_{t=1}^{T} \|\tilde{g}_t\|_*^2 \left(1 + \ln\left(1 + \frac{\left\|\mathring{v}\right\|^2 T^C}{\eta^2}\right)\right)} + B \left\|\mathring{v}\right\| \ln\left(\frac{\left\|\mathring{v}\right\| T^C}{\eta} + 1\right) ,$$

for constants A, B and C and η independent of T. Then for all $\mathring{w} \in \mathcal{W}$, Algorithm 8 guarantees

$$\sum_{t=1}^{T} \ell_t(x_t) - \ell_t\left(\mathring{w}\right) \leq \sum_{t=1}^{T} \left\langle x_t - \mathring{v}, g_t \right\rangle \leq \mathcal{O}\left(\sqrt{V_T\left(\mathring{w}\right) \ln \frac{TD}{\eta} \ln(T)} + \ln \frac{TD}{\eta} \ln(T) + \eta\right),$$

where

$$V_T\left(\mathring{w}\right) := \left\|\bar{x}_0 - \mathring{w}\right\|^2 + \sum_{t=1}^{T} \|\tilde{g}_t\|_*^2 \left\|x_t - \mathring{w}\right\|^2 \leq D^2 \sum_{t=1}^{T} \|g_t\|_*^2 \left\|x_t - \mathring{w}\right\|^2 .$$

Note that for $\mathbb{R}^d$: $\|\cdot\|_* = \|\cdot\|$.

Adaptivity to Strong-Convexity: As Cutkosky and Orabona 2018 further state in their paper, if the loss function (denoted by $\ell_t(\cdot)$) is λ-strongly convex, that is $\ell_t(v_t) - \ell_t\left(\mathring{v}\right) \leq \left\langle v_t - \mathring{v}, g_t \right\rangle - \frac{\lambda}{2} \left\|v_t - \mathring{v}\right\|$, then the above theorem implies that,

$$\sum_{t=1}^{T} \ell_t(w_t) - \ell_t\left(\mathring{v}\right) \leq \mathcal{O}\left(\log^2(DT)\left(1 + \frac{1}{\lambda}\right)\right) ,$$

with the simplifying assumption $\|g_t\|_* \leq 1$. and the above demonstrates the adaptivity to the strong-convexity parameter as promised.

Note: To obtain this result one could employ Algorithm **3** from Cutkosky and Orabona 2018 as the baseline online algorithm $\mathcal{A}$ used as an input to Algorithm 8. This choice satisfies the conditions in Theorem M.1.

Online to Batch: The output of Algorithm 8 can be easily converted to the stochastic settings by standard online-to-batch conversion Cesa-Bianchi and Lugosi (2006) to achieve a stochastic algorithm that is adaptive to the strong-convexity parameter.