Risk Bounds of SGD with Huber Loss for Robust Linear Regression

Arunachalam Chidambaram
University of California, Los Angeles
arun808@g.ucla.edu

Kunal Kishore
University of California, Los Angeles
kunalkishore@g.ucla.edu

Yingying Zhu
University of California, Los Angeles
yingyingz@g.ucla.edu

Abstract

In scenarios where the data samples are corrupted by an adversary, it's been established that Huber loss is preferred over l_2 loss due to it being less sensitive to outliers in data set. Current algorithms rely on having the whole data at hand in order to identify and remove the outliers. In this work, we consider a practical, online and highly scalable robust algorithm for the linear regression problem through constant step-size stochastic gradient descent with iterate averaging. We obtain an upper bound on the generalization error along with a finer analysis of the variance error stated in terms of the full eigenspectrum of the data covariance matrix.

1 Introduction

It is crucial to learn the underlying model despite corruptions by an adversary. Robust Learning to uncover the true parameters from a corrupted data set has become popular because of its applications in economics, biology, computer vision and safety critical systems. While efficient algorithms exist for robust learning, most commonly they require inspecting the entire data set, not feasible for large number of instances in the data set. There exists a void where in a robust online regression algorithm is required for practical use.

Stochastic gradient descent (SGD) is an iterative method for optimizing an objective function. It differs from gradient descent optimization, since it replaces the actual gradient calculated from the entire data set by a gradient from a randomly selected instance of the data. Through SGD, online updates to the parameters can be carried out avoiding the need to inspect the complete data. Linear Regression is one of the fundamental statistical model, so a robust linear regression to combat adversaries had gained traction over the years. In this project, we attempt to bound generalization error in over-parameterized linear regression model using Huber loss - a loss function which is preferred over l_2 loss due to it being less sensitive to outliers in data set through Stochastic Gradient Descent.

2 Related Work

2.1 Response corrupted robust regression

Given a simple contamination model that can only corrupt the responses and not the features, two approaches have been discussed: viewing the regression problem as a minimizing problem or rely on using a robust loss. Since the first approach is NP hard [6] and is non-convex, we consider using the

Huber loss as it is one of the classical examples of convex robust losses [7], which is a mix of l_1 and l_2 loss.

Existing research in online SGD has also shown that SGD on the l_1 loss converges to the true parameter vector at a rate independent of the values of the contaminated measurements [3]. This convergence rate proves that we don't need to rely on the whole data at hand in order to identify and remove the outliers existing in the robust regression problem. However, it does not use the Huber loss which we expect to work better in an adversarial setting. Also, it does not provide a generalization bound to the excess risk and just proves the convergence of the weights.

2.2 Excess risk bound in linear regression

In giving both an upper and an lower bound for excess risk, people have proven that unregularized SGD can generalize even in the infinite-dimensional setting [2]. As a result, if provided certain spectrum decay conditions on the data covariance, the benign-overfitting phenomenon can also be observed for SGD. We intend to extend this in an adversarial setting by using Huber loss in contrast to the l_2 loss used in the work.

3 Problem Formulation

In scenarios when we collect data samples from a noisy linear generative source where an unknown proportion of samples has been arbitrarily effected by an adversarial noise, the exact model recovery is not possible and several robust algorithms have been proposed. We focus on using the following model optimized using Stochastic Gradient Descent on Huber Loss - which is known to work well against adversarial noise in a fraction η of the samples. Consider we have a stream of independent observations $(\mathbf{x}_i, y_i)_{i \in N}$ from the following linear model:

$$y = \langle \mathbf{x}, \mathbf{w}^* \rangle + \epsilon + \mathbf{b}$$

where $\mathbf{w}^*$ is the true underlying parameter we wish to recover. We provide upper generalization bound to this model in an over-parameterized setting when using iterate averaging. Here are our assumptions that we shall be selectively using across two approaches:

- **A.1** The features $\boldsymbol{x}$ are centered Gaussian $\sim \mathcal{N}(0, \mathbf{H})$ where $\mathbf{H}$ is a P.D. matrix. with finite trace.
- **A.2** The features $\boldsymbol{x}$ are scaled so that their norm is upper bounded as $||\boldsymbol{x}||_2 \leq Q$
- **A.3** The noise ϵ is centered Gaussian $\sim \mathcal{N}(0, \sigma^2)$ where ϵ is independent of $\boldsymbol{x}$.
- **A.4** The adversarial noise b is independent of $\boldsymbol{x}$, with $\mathbb{P}(b \neq 0) = \alpha \in [0, 1)$.
- **A.5** The magnitude of adversarial noise is lower bounded: $\mathbb{E}\left(|b| \mid b \neq 0\right) \geq M$
- **A.6** The weight iterates $\mathbf{w}_t$ remain in the ℓ_2 region of the Huber loss.
- **A.7** Assume $\mathbb{E}[\mathbf{x}\mathbf{x}^\top], \mathbb{E}[\mathbf{x} \otimes \mathbf{x} \otimes \mathbf{x} \otimes \mathbf{x}]$ and $\mathbb{E}[y^2]$ exist and are all finite.
- **A.8** There exists a constant $R > 0$ such that $\mathbb{E}[\mathbf{x}\mathbf{x}^\top \mathbf{x}\mathbf{x}^\top] \preccurlyeq R^2 \mathbf{H}$
- **A.9** Suppose that: $\boldsymbol{\Sigma} := \mathbb{E}\left[(y - \langle \mathbf{w}^*, \mathbf{x} \rangle)^2 \mathbf{x}\mathbf{x}^\top\right]$, $\sigma^2 := ||\mathbf{H}^{-\frac{1}{2}} \boldsymbol{\Sigma} \mathbf{H}^{-\frac{1}{2}}||_2$ exist and are finite. $\boldsymbol{\Sigma}$ is the covariance matrix of the gradient noise at $\mathbf{w}^*$.

4 Huber Loss For Robust Regression

The conventional approach of learning a linear model is to minimize the ℓ_2 loss function while using the contaminated data. But as shown in Scott et al.[3], doing so does not work well in the case of adversarially contaminated data, as shown in Figure 1. This is because the ℓ_2 loss is very sensitive to outliers in the data.

Therefore, in our approach, we use the Huber loss function and minimize it using the contaminated data. It's a convex function parameterized by δ and is defined as follows:

$$\ell(\mathbf{w}) := \begin{cases} \frac{1}{2}(y - \langle \mathbf{w}, \mathbf{x} \rangle)^2 & \text{for } |y - \langle \mathbf{w}, \mathbf{x} \rangle| \leq \delta, \\ \delta \cdot \left(|y - \langle \mathbf{w}, \mathbf{x} \rangle| - \frac{1}{2}\delta\right), & \text{otherwise} \end{cases}$$

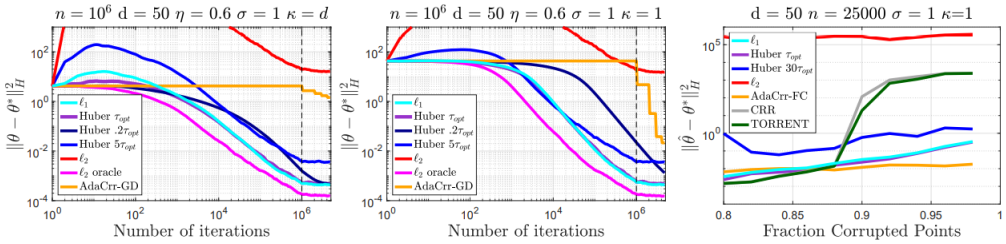

Figure 1: Online robust regression on synthetic data. Left and middle: Convergence rates for a fixed α and for two different conditioning of $\mathbf{H}$. The dashed line marks the first pass over the data. Right: Estimation performance when varying the portion of corruption α.

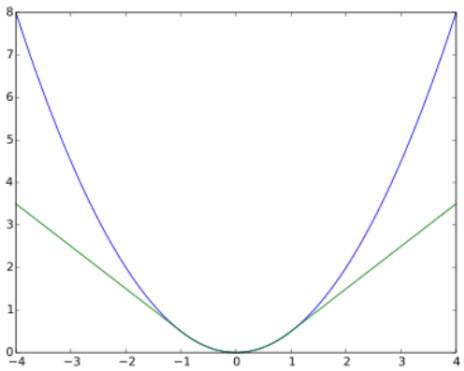

Figure 2: Huber loss (green, $\delta = 1$) and squared error loss (blue) as a function of $y - f(x)$

The plot of Huber loss and the ℓ_2 loss functions have been shown in Figure 2. It behaves just like the ℓ_2 loss function around the origin, and beyond a certain threshold, it acts like the ℓ_1 loss function so that the gradient of the Huber loss is overall bounded over $\mathbb{R}$.

We use a fixed learning rate SGD to optimize the above loss where the weight update at the t^{th} iteration can be written as:

$$\mathbf{w}_t = \begin{cases} \mathbf{w}_{t-1} + \gamma \cdot (y_t - \langle \mathbf{w}_{t-1}, \mathbf{x}_t \rangle)\mathbf{x}_t & \text{when } |y_t - \langle \mathbf{w}_{t-1}, \mathbf{x}_t \rangle| \leq \delta, \\ \mathbf{w}_{t-1} + \gamma \cdot \delta \cdot \text{sgn}(y_t - \langle \mathbf{w}_{t-1}, \mathbf{x}_t \rangle)\mathbf{x}_t & \text{otherwise} \end{cases}$$

Note that we use an averaged weight in the above expression which is defined as follows:

$$\overline{\mathbf{w}}_N := \frac{1}{N} \sum_{t=0}^{N-1} \mathbf{w}_t \text{ (or actively calculated as) } \overline{\mathbf{w}}_t = \frac{1}{t}\mathbf{w}_t + \frac{t-1}{t}\overline{\mathbf{w}}_{t-1}$$

5 Generalization Error - Loose Upper Bound

In this section, we intend to provide a loose upper bound under much less stricter assumptions (using assumptions **A.1**, **A.2**, **A.3**, **A.9**. Here we make use of the fact that the gradient of the Huber loss function is also bounded. Based on this observation, we directly use results from Theorem 14.8 in [4], to get the following upper bound on the generalization error:

$$\mathbb{E}[\ell(\mathbf{w}_t)] - \ell(\mathbf{w}^*) \leq \frac{1}{2\gamma}\|\mathbf{w}^*\|^2 + \frac{N}{2}\gamma\delta^2 Q^2$$

Proof. Here is the the update step in SGD, where $\mathbf{v}_t$ denotes the gradient:

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \gamma \mathbf{v}_t$$

We could rewrite the following term as

$$\langle \mathbf{w}_t - \mathbf{w}^*, \mathbf{v}_t \rangle = \frac{1}{\gamma} \langle \mathbf{w}_t - \mathbf{w}^*, \gamma \mathbf{v}_t \rangle$$

$$= \frac{1}{2\gamma} - \|\mathbf{w}_t - \mathbf{w}^* - \gamma \mathbf{v}_t\|^2 + \|\mathbf{w}_t - \mathbf{w}^*\|^2 + \gamma^2 \|\mathbf{v}_t\|^2)$$

$$= \frac{1}{2\gamma}(-\|\mathbf{w}_t - \mathbf{w}^*\|^2 + \|\mathbf{w}_t - \mathbf{w}^*\|^2) + \frac{\gamma}{2}\|\mathbf{v}_t\|^2$$

Summing the equality over t, we get the following

$$\sum_{t=0}^{N-1} \langle \mathbf{w}_t - \mathbf{w}^*, \mathbf{v}_t \rangle = \frac{1}{2\gamma} \sum_{t=0}^{N-1} (-\|\mathbf{w}_{t+1} - \mathbf{w}^*\|^2 + \|\mathbf{w}_t - \mathbf{w}^*\|^2) + \frac{\gamma}{2} \sum_{t=0}^{N-1} \|\mathbf{v}_t\|^2$$

$$= \frac{1}{2\gamma}(\|\mathbf{w}_0 - \mathbf{w}^*\|^2 - \|\mathbf{w}_N - \mathbf{w}^*\|^2) + \frac{\gamma}{2} \sum_{t=0}^{N-1} \|\mathbf{v}_t\|^2$$

$$\leq \frac{1}{2\gamma}\|\mathbf{w}^*\|^2 + \frac{\gamma}{2} \sum_{t=0}^{N-1} \|\mathbf{v}_t\|^2$$

Define the notation $\mathbf{v}_{1:t}$ to denote the sequence $\mathbf{v}_1$, ..., $\mathbf{v}_t$. Put the expectation function inside

$$\mathop{\mathbb{E}}_{\mathbf{v}_{1:N}} \frac{1}{n} \sum_{t=1}^{N} \langle \mathbf{w}_t - \mathbf{w}^*, \mathbf{v}_t \rangle = \frac{1}{N} \sum_{t=1}^{N} \mathop{\mathbb{E}}_{\mathbf{v}_{1:N}} \langle \mathbf{w}_t - \mathbf{w}^*, \mathbf{v}_t \rangle$$

$$\mathop{\mathbb{E}}_{\mathbf{v}_{1:N}} [\langle \mathbf{w}_t - \mathbf{w}^*, \mathbf{v}_t \rangle] = \mathop{\mathbb{E}}_{\mathbf{v}_{1:t}} [\langle \mathbf{w}_t - \mathbf{w}^*, \mathbf{v}_t \rangle] = \mathop{\mathbb{E}}_{\mathbf{v}_{1:t-1}} [\mathop{\mathbb{E}}_{\mathbf{v}_{1:N}} [\langle \mathbf{w}_t - \mathbf{w}^*, \mathbf{v}_t \rangle | \mathbf{v}_{1:t-1}]]$$

$$= \mathop{\mathbb{E}}_{\mathbf{v}_{1:t-1}} [\langle \mathbf{w}_t - \mathbf{w}^*, \mathop{\mathbb{E}}_{\mathbf{v}_{1:t}} [\mathbf{v}_t | \mathbf{v}_{1:t-1}] \rangle] \geq \mathop{\mathbb{E}}_{\mathbf{v}_{1:t-1}} [\ell(\mathbf{w}_t) - \ell(\mathbf{w}^*)]$$

Combining the above two results, we get

$$\mathbb{E}[\ell(\mathbf{w}_t)] - \ell(\mathbf{w}^*) \leq \frac{1}{2\gamma}\|\mathbf{w}^*\|^2 + \frac{\gamma}{2} \sum_{t=0}^{N-1} \|\mathbf{v}_t\|^2$$

In case of Huber loss, we have the following upper bound on gradient $||\mathbf{v}_t||_2 \leq \delta Q$, using which we get the final result:

$$\boxed{\mathbb{E}[\ell(\mathbf{w}_t)] - \ell(\mathbf{w}^*) \leq \frac{1}{2\gamma}\|\mathbf{w}^*\|^2 + \frac{N}{2}\gamma\delta^2 Q^2}$$

6 Generalization Error - Variance Upper Bound

6.1 Preliminaries

We use $\otimes$ to denote the kronecker/tensor product. We define the following linear operators:

$$\mathcal{I} = \mathbf{I} \otimes \mathbf{I}, \quad \mathcal{M} = \mathbb{E}[\mathbf{x} \otimes \mathbf{x} \otimes \mathbf{x} \otimes \mathbf{x}], \quad \widetilde{\mathcal{M}} = \mathbf{H} \otimes \mathbf{H},$$

$$\mathcal{T} = \mathbf{H} \otimes \mathbf{I} + \mathbf{I} \otimes \mathbf{H} - \gamma\mathcal{M}, \quad \widetilde{\mathcal{T}} = \mathbf{H} \otimes \mathbf{I} + \mathbf{I} \otimes \mathbf{H} - \gamma\mathbf{H} \otimes \mathbf{H}.$$

We use the notation $\mathcal{O} \circ \mathbf{A}$ to denote the operator acting on a symmetric matrix $\mathbf{A}$.

$$\mathcal{I} \circ \mathbf{A} = \mathbf{A}, \quad \mathcal{M} \circ \mathbf{A} = \mathbb{E}[(\mathbf{x}^\top \mathbf{A}\mathbf{x})\mathbf{x}\mathbf{x}^\top], \quad \widetilde{\mathcal{M}} \circ \mathbf{A} = \mathbf{H}\mathbf{A}\mathbf{H},$$

$$(\mathcal{I} - \gamma\mathcal{T}) \circ \mathbf{A} = \mathbb{E}[(\mathbf{I} - \gamma\mathbf{x}\mathbf{x}^\top)\mathbf{A}(\mathbf{I} - \gamma\mathbf{x}\mathbf{x}^\top)], \quad (\mathcal{I} - \gamma\widetilde{\mathcal{T}}) \circ \mathbf{A} = (\mathbf{I} - \gamma\mathbf{H})\mathbf{A}(\mathbf{I} - \gamma\mathbf{H}).$$

6.2 The Bias-Variance Decomposition

Let's define $\boldsymbol{\eta}_t = \mathbf{w}_t - \mathbf{w}^*$. We decompose this into bias and variance term as done in [2]:

$$\boldsymbol{\eta}_t = \boldsymbol{\eta}_t^{\text{bias}} + \boldsymbol{\eta}_t^{\text{variance}}$$

If the iterates are initialized from the optimal $\mathbf{w}^*$, than the obtained SGD iterates $\eta_t^{\textbf{variance}}$ reveal the *variance error*,

$$\boldsymbol{\eta}_t^{\textbf{variance}} = (\mathbf{I} - \gamma \mathbf{x}_t \mathbf{x}_t^\top)\boldsymbol{\eta}_{t-1}^{\textbf{variance}} + \gamma \epsilon_t \mathbf{x}_t + \gamma b_t \mathbf{x}_t$$

We then define $\mathbf{C}_t \coloneqq \mathbb{E}[\boldsymbol{\eta}_t^{\textbf{variance}} \otimes \boldsymbol{\eta}_t^{\textbf{variance}}]$ and use the update equation from above as follows

$$\begin{aligned}
\mathbf{C}_t &= \mathbb{E}[(\mathbf{I} - \gamma \mathbf{x}_t \mathbf{x}_t^\top)\boldsymbol{\eta}_{t-1}^{\textbf{variance}} \otimes (\mathbf{I} - \gamma \mathbf{x}_t \mathbf{x}_t^\top)\boldsymbol{\eta}_{t-1}^{\textbf{variance}}] \\
&\quad + \mathbb{E}[\gamma^2(\epsilon_t + b_t)^2 \mathbf{x}_t \otimes \mathbf{x}_t] \\
&\quad + \mathbb{E}[\gamma(\epsilon_t + b_t)(\mathbf{I} - \gamma \mathbf{x}_t \mathbf{x}_t^\top)\boldsymbol{\eta}_{t-1}^{\textbf{variance}} \otimes \mathbf{x}_t] \\
&\quad + \mathbb{E}[\gamma(\epsilon_t + b_t)\mathbf{x}_t \otimes (\mathbf{I} - \gamma \mathbf{x}_t \mathbf{x}_t^\top)] \\
&= (\mathbf{I} - \gamma \mathcal{T}) \circ \mathbf{C}_{t-1} + \gamma^2 \Sigma + \gamma^2 \mathbf{V} - \gamma^2 \mathbf{U} - \gamma^2 \mathbf{U}
\end{aligned}$$

where we define $\mathbf{U} \coloneqq \mathbb{E}[b_t]\mathbb{E}[\boldsymbol{\eta}_{t-1}^\top \mathbf{x}_t \mathbf{x}_t \mathbf{x}_t^\top]$, $\mathbf{V} \coloneqq \mathbb{E}[b_t^2]\mathbf{H}$

And finally we obtain the following iterative form of $\mathbf{C}_t$:

$$\mathbf{C}_t = (\mathcal{I} - \gamma \mathcal{T}) \circ \mathbf{C}_{t-1} + \gamma^2(\Sigma + \mathbf{V} - 2\mathbf{U})$$

We then solve the recursion equation of $\mathbf{C}_t$ to get:

$$\mathbf{C}_t = \gamma^2 \sum_{k=0}^{t-1} (\mathcal{I} - \gamma \mathcal{T})^k \circ (\Sigma + \mathbf{V} - 2\mathbf{U})$$

$$\Rightarrow \mathbf{C}_t = \gamma^2 \sum_{k=0}^{t-1} (\mathcal{I} - \gamma \mathcal{T})^k \circ \Sigma + \gamma^2 \sum_{k=0}^{t-1} (\mathcal{I} - \gamma \mathcal{T})^k \circ (\mathbf{V} - 2\mathbf{U})$$

We now claim that under assumptions **A.3**, **A.4**, **A.5**, **A.6**, the second term in the following expression for $\mathbf{C}_t$ is *not* a PSD.

$$\mathbf{C}_t = \underbrace{\gamma^2 \sum_{k=0}^{t-1} (\mathcal{I} - \gamma \mathcal{T})^k \circ \Sigma}_{\text{First term:} = \widetilde{\mathbf{C}}_t} + \underbrace{\gamma^2 \sum_{k=0}^{t-1} (\mathcal{I} - \gamma \mathcal{T})^k \circ (\mathbf{V} - 2\mathbf{U})}_{\text{Second term} \preccurlyeq \mathbf{0}}$$

This means that $\mathbf{C}_t$ will have the same upper bound (if any) as the first term - which we define as $\widetilde{\mathbf{C}}_t$. Under certain condition on δ, Lemma 2 shows that $\mathbf{V} - 2\mathbf{U} \preccurlyeq 0$

From *Cauchy-Schwarz* inequality, we have the following bias-variance decomposition of the excess risk term (see Jain et al. (2017b) [5]):

$$\mathbb{E}[\ell(\bar{\mathbf{w}}_N)] - \ell(\mathbf{w}^*) = \frac{1}{2}\langle \mathbf{H}, \mathbb{E}[\bar{\boldsymbol{\eta}}_N \otimes \bar{\boldsymbol{\eta}}_N] \rangle \leq (\sqrt{\text{bias}} + \sqrt{\text{variance}})^2$$

where bias := $\frac{1}{2}\langle \mathbf{H}, \mathbb{E}[\bar{\boldsymbol{\eta}}_N^{\text{bias}} \otimes \bar{\boldsymbol{\eta}}_N^{\text{bias}}] \rangle$, variance := $\frac{1}{2}\langle \mathbf{H}, \mathbb{E}[\bar{\boldsymbol{\eta}}_N^{\text{variance}} \otimes \bar{\boldsymbol{\eta}}_N^{\text{variance}}] \rangle$

We will refer to these two terms as the *bias error* and *variance error* respectively. As a part of this project, we shall be focusing on bounding the variance error only. From [2], we have the following upper bound on variance error:

$$\text{variance error} := \frac{1}{2}\langle \mathbf{H}, \mathbb{E}[\bar{\boldsymbol{\eta}}_N^{\text{variance}} \otimes \bar{\boldsymbol{\eta}}_N^{\text{variance}}] \rangle \leq \frac{1}{N^2} \sum_{t=0}^{N-1} \sum_{k=t}^{N-1} \langle (\mathbf{I} - \gamma \mathbf{H})^{k-t}, \mathbf{C}_t \rangle$$

As mentioned earlier, since we have the same upper bound for $\mathbf{C}_t$ as $\widetilde{\mathbf{C}}_t$ (the condition for which is proved in Lemma 2), using the above result, we can arrive at the same final upper bound of variance as [2]:

$$\text{variance error} \leq \frac{\sigma^2}{N(1-\gamma R^2)} \cdot \sum_i (1-(1-\gamma\lambda_i)^N)^2$$

where λ_i are the eigenvalues of $\mathbf{H}$ sorted in decreasing order.

Setting $k^* = \max\{k : \lambda_k \geq \frac{1}{\gamma N}\}$, we can approximate and re-write the previous result as:

$$\boxed{\text{Variance of excess risk} \leq \frac{\sigma^2}{1-\gamma R^2}\left(\frac{k^*}{N} + \gamma^2 N \cdot \textstyle\sum_{i>k^*} \lambda_i^2\right)}$$

Lemma 1. Under assumption **A.6**, we have the following inequalities:

$$-\delta\mathbb{E}[|\mathbf{b}_t|]\mathbf{H} \preccurlyeq \mathbf{U} - \mathbf{V} \preccurlyeq \delta\mathbb{E}[|\mathbf{b}_t|]\mathbf{H}$$

Proof. We can rewrite the condition of when the weight update lies in the l_2 region as

$$|\langle \boldsymbol{\eta}_{t-1}, \mathbf{x}_t\rangle - b_t - \epsilon_t| \leq \delta$$

$$\Rightarrow |b_t\boldsymbol{\eta}_{t-1}^\top \mathbf{x}_t\mathbf{x}_t^\top - b_t^2\mathbf{x}_t\mathbf{x}_t^\top - \epsilon_t b_t\mathbf{x}_t\mathbf{x}_t^\top| \preccurlyeq \delta|b_t\mathbf{x}_t\mathbf{x}_t^\top|$$

$$\Rightarrow -\delta|b_t|\mathbf{x}_t\mathbf{x}_t^\top \preccurlyeq b_t\boldsymbol{\eta}_{t-1}^\top \mathbf{x}_t\mathbf{x}_t\mathbf{x}_t^\top - b_t^2\mathbf{x}_t\mathbf{x}_t^\top - \epsilon_t b_t\mathbf{x}_t\mathbf{x}_t^\top \preccurlyeq \delta|b_t|\mathbf{x}_t\mathbf{x}_t^\top$$

Then take expectation $\mathbb{E}(\cdot)$ on all sides, and since $\mathbb{E}(\epsilon_t) = 0$, we obtain the following

$$-\delta\mathbb{E}[|b_t|]\mathbf{x}_t\mathbf{x}_t^\top \preccurlyeq \mathbb{E}[b_t]\mathbb{E}[\boldsymbol{\eta}_{t-1}^\top \mathbf{x}_t\mathbf{x}_t\mathbf{x}_t^\top] - \mathbb{E}[b_t^2]\mathbf{x}_t\mathbf{x}_t^\top \preccurlyeq \delta\mathbb{E}[|b_t|]\mathbf{x}_t\mathbf{x}_t^\top$$

From the previous definitions of $\mathbf{U}$ and $\mathbf{V}$, we obtain the desired result:

$$-\delta\mathbb{E}[|b_t|]\mathbf{H} \preccurlyeq \mathbf{U} - \mathbf{V} \preccurlyeq \delta\mathbb{E}[|b_t|]\mathbf{H}$$

□

Lemma 2. Under assumptions **A.1**, **A.4** and **A.5**, following the result of Lemma 1, the following inequality holds when $\delta \leq \frac{\alpha M}{2}$:

$$\mathbf{V} - 2\mathbf{U} \preccurlyeq 0$$

Proof. From Lemma 1, we have

$$2\mathbf{V} - 2\mathbf{U} \preccurlyeq 2\delta\mathbb{E}[|b_t|]\mathbf{H}$$

Subtracting $\mathbf{V}$ on both sides, we get

$$\mathbf{V} - 2\mathbf{U} \preccurlyeq (2\delta\mathbb{E}[|b_t|] - \mathbb{E}[b_t^2])\mathbf{H}$$

Using Jensen's Inequality, and using assumptions **A.4** and **A.5** we have:

$$\mathbb{E}[b_t^2] \geq (\mathbb{E}[|b_t|])^2 \geq \mathbb{E}[|b_t|]\alpha M$$

$$\Rightarrow \frac{\mathbb{E}[b_t^2]}{2\mathbb{E}[|b_t|]} \geq \frac{\alpha M}{2}$$

If $\delta \leq \frac{\alpha M}{2}$, then from above inequality, we get $(2\delta\mathbb{E}[|b_t|] - \mathbb{E}[b_t^2]) \leq 0$, and since $\mathbf{H}$ is a P.D. from assumption **A.1**, we get the result

$$\mathbf{V} - 2\mathbf{U} \preccurlyeq 0$$

□

7 Conclusion and Future Work

We have successfully given generalization bound for Huber loss in an adversarial setting. We have also provided a finer analysis by giving the upper bound for variance error operating in the ℓ_2 region. We believe that extending our analysis to include the bias error will lead to a sharp upper bound on the generalization error of linear regression in the presence of adversary. The non-linearity in the gradient of the Huber loss makes bounding the bias error more demanding which we leave as future work. We have considered constant step-size in our work, but we believe that it's worthwhile to consider a decaying step-size which is more commonly used in stochastic gradient descent implementations. While this work on robust linear regression model is important, a step for future work is to extend these results to more complex machine learning models such as neural networks.

References

[1] Aymeric Dieuleveut, Nicolas Flammarion, and Francis Bach, *Harder, better, faster, stronger convergence rates for least-squares regression*. The Journal of Machine Learning Research, 18(1):3520–3570, 2017.

[2] Difan Zou, Jingfeng Wu, Vladimir Braverman, Quanquan Gu, and Sham Kakade, *Benign Overfitting of Constant-Stepsize SGD for Linear Regression*. Proceedings of Thirty Fourth Conference on Learning Theory, in Proceedings of Machine Learning Research, 2021.

[3] Pesme, S., and Flammarion, N. (2020, July 1). Online robust regression via SGD on the L1 loss. arXiv.org. Retrieved April 28, 2022, from https://arxiv.org/abs/2007.00399

[4] Shai Shalev-Shwartz and Shai Ben-David. 2014. Understanding Machine Learning: From Theory to Algorithms. Cambridge University Press, USA.

[5] Prateek Jain, Praneeth Netrapalli, Sham M Kakade, Rahul Kidambi, and Aaron Sidford. Parallelizing stochastic gradient descent for least squares regression: mini-batching, averaging, and model misspecification. The Journal of Machine Learning Research, 18(1): 8258–8299, 2017b.

[6] C. Studer, P. Kuppinger, G. Pope, and H. Bolcskei. Recovery of sparsely corrupted signals. IEEE Transactions on Information Theory, 58(5):3115–3130, 2012.

[7] P. J. Huber. Robust estimation of a location parameter. Ann. Math. Statist., 35:73–101, 1964.