Robust Linear Regression by Sub-Quantile Optimization

Anonymous authors
Paper under double-blind review

Abstract

Robust Linear Regression is the problem of fitting data to a distribution, $\mathbb{P}$ when there exists contaminated samples, $\mathbb{Q}$. We model this as $\hat{\mathbb{P}} = (1-\epsilon)\mathbb{P} + \epsilon\mathbb{Q}$. Traditional Least Squares Methods fit the empirical risk model to all training data in $\hat{\mathbb{P}}$. In this paper we show theoretical and experimental results of sub-quantile optimization, where we optimize with respect to the p-quantile of the empirical loss.

1 Introduction

Linear Regression is one of the most widely used statistical estimators throughout Science. Although robustness is only a somewhat recent topic in machine learning, it has been a topic in statistics for many decades. Several popular methods have been very popular due to their simplicity and high effectiveness including quantile regression Koenker & Hallock (2001), Theil-Sen Estimator Sen (1968), and Huber Regression Huber & Ronchetti (2009).

Our goal is to provide a theoretic analysis and convergence conditions for sub-quantile optimization and offer practioners a method for robust linear regression.

In this paper we will show how Sub-Quantile Optimization can address the shortcomings of ERM in the case of corrupted data or imbalanced data, where there exists a majority class and a minority class.

2 Related Work

Least Trimmed Squares (LTS) Mount et al. (2014).

Tilted Empirical Risk Minimization (TERM) Li et al. (2020) is a framework built to similarly handle the shortcomings of ERM with respect to robustness. The TERM framework instead minimizes the following quantity, where t is a hyperparameter

$$\tilde{R}(t;\boldsymbol{\theta}) \coloneqq \frac{1}{t}\log\left(\frac{1}{N}\sum_{i\in[N]} e^{tf(\boldsymbol{x}_i;\boldsymbol{\theta})}\right) \tag{1}$$

SMART Awasthi et al. (2022) proposes the *iterative trimmed maximum likelihood estimator* against adversarially corrupted samples in General Linear Models (GLM). The estimator is defined as follows, where $S = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ represents the training data.

$$\hat{\boldsymbol{\theta}}(S) = \min_{\boldsymbol{\theta}} \min_{\hat{S}\subset S, |\hat{S}|=(1-\epsilon)n} \sum_{(\boldsymbol{x}_i,y_i)\in S} -\log f(y_i|\boldsymbol{\theta}^T\boldsymbol{x}_i) \tag{2}$$

SEVER Diakonikolas et al. (2019) is a gradient filtering algorithm which removes elements whose gradients have the furthest distance from the average gradient of all points

$$\tau_i = \left((\nabla f_i(\boldsymbol{w}) - \hat{\nabla})\cdot\boldsymbol{v}\right)^2 \tag{3}$$

Super-Quantile Optimization Rockafellar et al. (2014)

Robust Risk Minimization Osama et al. (2020)

3 Sub-quantile Optimization

Definition 1. *Let F_X represent the Cumulative Distribution Function (CDF) of the random variable X. The* **p-*Quantile*** *of a Random Variable X is defined as follows*

$$Q_p(p) = \inf\{x \in \mathbb{R} : p \leq F(x)\} \tag{4}$$

Note $Q_p(0.5)$ represents the median of the random variable.

Definition 2. *The* ***Empirical Distribution Function*** *is defined as follows*

$$\hat{F}_n(t) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{X_i \leq t} \tag{5}$$

Definition 3. *Let ℓ be the loss function.* ***Risk*** *is defined as follows*

$$U = \mathbb{E}\left[\ell\left(f(\boldsymbol{x};\boldsymbol{\theta},\boldsymbol{y})\right)\right] \tag{6}$$

The p-Quantile of the Empirical Risk is given

$$\mathbb{L}_p(U) = \frac{1}{p}\int_0^p Q_q(U)\,dq \tag{7}$$

The Sub-Quantile Optimization problem is posed as follows

$$\boldsymbol{\theta}_{SM} = \underset{\boldsymbol{\theta}\in\mathbb{R}^d}{\arg\min}\,\max_{t\in\mathbb{R}} \left\{ t - \frac{1}{p}\mathbb{E}(t - \ell(f(\boldsymbol{x};\boldsymbol{\theta}),y))^+ \right\} \tag{8}$$

For the linear regression case, this equation becomes

$$\boldsymbol{\theta}_{SM} = \underset{\boldsymbol{\theta}\in\mathbb{R}^d}{\arg\min}\,\max_{t\in\mathbb{R}} \left\{ t - \frac{1}{np}\sum_{i=1}^{n}(t - (\boldsymbol{\theta}^T\boldsymbol{x}_i - y_i)^2)^+ \right\} \tag{9}$$

The two-step optimization for Sub-Quantile optimization is given as follows

$$t_{k+1} = \arg\max_t g(t,\boldsymbol{\theta}_k) \tag{10}$$

$$\boldsymbol{\theta}_{k+1} = \boldsymbol{\theta}_k + \alpha\nabla_{\boldsymbol{\theta}_k} g(t,\boldsymbol{\theta}_k) \tag{11}$$

This algorithm is adopted from Razaviyayn et al. (2020).

4 Theoretical Analysis

Lemma 4.1. *The maximizing value of t in $g(t,\boldsymbol{\theta})$ in t-update step of optimization as described by Equation 10 is maximized when $t = Q_p(U)$*

Proof. Since $g(t,\boldsymbol{\theta})$ with respect to t is a concave function. Maximizing $g(t,\boldsymbol{\theta})$ is equivalent to minimizing $-g(t,\boldsymbol{\theta})$. We will find fermat's optimality condition for the function $-g(t,\boldsymbol{\theta})$, which is convex. Let $\hat{\boldsymbol{\nu}} = sorted\left((\boldsymbol{\theta}^T\boldsymbol{X} - \boldsymbol{y})^2\right)$ and note $0 < p < 1$

$$\partial(-g(t,\boldsymbol{\theta})) = -1 + \frac{1}{np}\sum_{i=1}^{n}\left\{\begin{array}{ll} 1, & \text{if } t > \hat{\boldsymbol{\nu}}_i \\ 0, & \text{if } t < \hat{\boldsymbol{\nu}}_i \\ [0,1]\,, & \text{if } t = \hat{\boldsymbol{\nu}}_i \end{array}\right\} \tag{12}$$

$$= 0 \text{ when } t = \hat{\boldsymbol{\nu}}_{np} \tag{13}$$

This is the p-quantile of U. A full proof is provided in Appendix B.1. □

Lemma 4.2. *Let $t = \hat{\boldsymbol{\nu}}_{np}$. The $\boldsymbol{\theta}$-update step described in Equation 9 is equivalent to minimizing the least squares loss of the np elements with the lowest squared loss.*

$$\nabla_{\boldsymbol{\theta}} g(t_{k+1},\boldsymbol{\theta}_k) = \frac{1}{np}\sum_{i=1}^{np} 2\boldsymbol{x}_i(\boldsymbol{\theta}_k^T\boldsymbol{x}_i - y_i) \tag{14}$$

We provide a proof in Appendix B.2. However, this result is quite intuitive as it shows we are optimizing over the p Sub-Quantile of the Risk.

Interpretation 1. *Sub-Quantile Minimization continously minimizes the risk over the p-quantile of the error. In each iteration, this means we reduce the error of the points within the lowest np errors.*

We are solving a min-max convex-concave problem, thus we are looking for a Nash Equilibrium Point.

Definition 4. *$(t^*, \boldsymbol{\theta}^*)$ is a* ***Nash Equilibrium*** *of g if for any $(t, \boldsymbol{\theta}) \in \mathbb{R} \times \mathbb{R}^d$*

$$g(t^*, \boldsymbol{\theta}) \leq g(t^*, \boldsymbol{\theta}^*) \leq g(t, \boldsymbol{\theta}^*) \tag{15}$$

Definition 5. *$(t^*, \boldsymbol{\theta}^*)$ is a* ***Local Nash Equilibrium*** *of g if there exists $\delta > 0$ such that for any $t, \boldsymbol{\theta}$ $(t, \boldsymbol{\theta})$ satsifying $\|t - t^*\| \leq \delta$ and $\|\boldsymbol{\theta} - \boldsymbol{\theta}^*\| \leq \delta$ then:*

$$g(t^*, \boldsymbol{\theta}) \leq g(t^*, \boldsymbol{\theta}^*) \leq g(t, \boldsymbol{\theta}^*) \tag{16}$$

Proposition 1. *As g is first-order differentiable, thus any local Nash Equilibrium satisfies $\nabla_{\boldsymbol{\theta}} g(t, \boldsymbol{\theta}) = \mathbf{0}$ and $\nabla_t g(t, \boldsymbol{\theta}) = 0$*

We are now interested in what it means to be at a Local Nash Equilibrium.

Lemma 4.3. *$g(t_{k+1}, \boldsymbol{\theta}_k)$ is convex with respect to $\boldsymbol{\theta}_k$.*

Lemma 4.3 tells us we are solving a min-max concave-convex optimization problem. In Jin et al. (2019), the researchers examined the problem where they are given a max-oracle which is correct up to some value ϵ. In our case, we can set $\epsilon = 0$.

Lemma 4.4. *$g(t_{k+1}, \boldsymbol{\theta}_k)$ is concave with respect to t.*

Proof. We provide a simple argument for concavity. Note t is a concave and convex function. Also $(\cdot)^+$ is a convex strictly non-negative function. Therefore we have a concave function minus the non-negative multiple of a summation of an affine function composed with a convex function. Therefore this is a concave function with respect to t. □

Lemma 4.5. *$g(t, \boldsymbol{\theta})$ is L-smooth with respect to $\boldsymbol{\theta}$ with $L = \left\| \frac{2}{np} \sum_{i=1}^{np} \|\boldsymbol{x}_i\|^2 \right\|$*

Lemma 4.6. *Since $g(t, \boldsymbol{\theta})$ is L-smooth by Lemma 4.5 $g(t, \boldsymbol{\theta})$ is a monotonically decreasing function.*

Proof Sketch. We are looking to prove $g(t_{k+1}, \boldsymbol{\theta}_{k+1}) \leq g(t_k, \boldsymbol{\theta}_k)$. This is equivalent to proving

$$g(t_{k+1}, \boldsymbol{\theta}_k) - g(t_{k+1}, \boldsymbol{\theta}_{k+1}) \geq g(t_{k+1}, \boldsymbol{\theta}_k) - g(t_k, \boldsymbol{\theta}_k) \tag{17}$$

This is due to the ordering of our two-step optimization. □

Lemma 4.7. *$g(t, \boldsymbol{\theta})$ is bounded above by $\sum_{i=1}^{np} \boldsymbol{\nu}_i$ and below by 0.*

Theorem 1. *By Lemma 4.6 and 4.7, $g(t, \boldsymbol{\theta})$ converges to a local minimum.*

Proof Sketch. We need to prove $\lim_{k \to \infty} \|\nabla_{\boldsymbol{\theta}_k} g(t, \boldsymbol{\theta}_k)\| = 0$. By Lemma 4.5, $g(t, \boldsymbol{\theta})$ is L-smooth with respect to $\boldsymbol{\theta}$, thus if t is greater than the same np elements, then $\|\nabla_{\boldsymbol{\theta}_k} g(t, \boldsymbol{\theta}_k)\| \to 0$. □

4.1 Robustness

Assumption 1. *The parameters of $\boldsymbol{\theta}$ are sampled from a symmetrical continuous distribution around 0. Inspired by Lu (2020).*

Assumption 2. *To provide theoretical bounds on the effectiveness of Sub-Quantile Minimization, we assume the error of data from Q on a trained model over $\boldsymbol{P}$ is normally distributed. Let $\boldsymbol{\theta}_P^* = (\boldsymbol{P}^T\boldsymbol{P})^{-1}\boldsymbol{P}^T\boldsymbol{y}_P$ be the optimal linear regression model for $\boldsymbol{P}$. Then $\boldsymbol{\theta}_P^{*T}\boldsymbol{Q} - \boldsymbol{y}_q \sim \mathcal{N}(\mu, \sigma^2)$ where $\mu \neq 0$.*

We will provide experimental results in Section E that show in most real-world datasets, the corruption follows this assumption. We want to clarify the corruption is not adversarially chosen.
In this section we quantify the effect of corruption on the desired model. To introduce notation,

let $\boldsymbol{P}$ represent the data from distribution $\mathbb{P}$ and let $\boldsymbol{Q}$ represent the training data for $\mathbb{Q}$. Let $\boldsymbol{y}_P$ represent the target data for $\mathbb{P}$ and let $\boldsymbol{y}_Q$ represent the target data for $\mathbb{Q}$.

Assumption 3. *We assume $\boldsymbol{P}$ and $\boldsymbol{Q}$ are sampled from the same normal distribution.*

$$\boldsymbol{P}_i, \boldsymbol{Q}_j \sim \mathcal{N}_p(\boldsymbol{0}, \boldsymbol{\Sigma}) \tag{18}$$

We will use our assumptions to quantify the effect of the corrupted data on an optimal least squares regression model. We are interested in $(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y} - (\boldsymbol{P}^T\boldsymbol{P})^{-1}\boldsymbol{P}^T\boldsymbol{y}$ It is know the least squares optimal solution for $\boldsymbol{X}$ is equal to $(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$

Note $\boldsymbol{X} = \begin{pmatrix}\boldsymbol{P}\\ \boldsymbol{Q}\end{pmatrix}$ so $\boldsymbol{X}^T = \begin{pmatrix}\boldsymbol{P}^T & \boldsymbol{Q}^T\end{pmatrix}$

We will first calculate the pseudo-inverse

$$\boldsymbol{X}^T\boldsymbol{X} = \begin{pmatrix}\boldsymbol{P}^T & \boldsymbol{Q}^T\end{pmatrix}\begin{pmatrix}\boldsymbol{P}\\ \boldsymbol{Q}\end{pmatrix} \tag{19}$$

$$= \boldsymbol{P}^T\boldsymbol{P} + \boldsymbol{Q}^T\boldsymbol{Q} \tag{20}$$

Now we can calculate the Moore-Penrose Inverse

$$(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T = (\boldsymbol{P}^T\boldsymbol{P} + \boldsymbol{Q}^T\boldsymbol{Q})^{-1}\begin{pmatrix}\boldsymbol{P}^T & \boldsymbol{Q}^T\end{pmatrix} \tag{21}$$

$$= \begin{pmatrix}(\boldsymbol{P}^T\boldsymbol{P} + \boldsymbol{Q}^T\boldsymbol{Q})^{-1}\boldsymbol{P}^T & (\boldsymbol{P}^T\boldsymbol{P} + \boldsymbol{Q}^T\boldsymbol{Q})^{-1}\boldsymbol{Q}^T\end{pmatrix} \tag{22}$$

Now we solve for the optimal model

$$\boldsymbol{X}^\dagger\boldsymbol{y} = \begin{pmatrix}(\boldsymbol{P}^T\boldsymbol{P} + \boldsymbol{Q}^T\boldsymbol{Q})^{-1}\boldsymbol{P}^T & (\boldsymbol{P}^T\boldsymbol{P} + \boldsymbol{Q}^T\boldsymbol{Q})^{-1}\boldsymbol{Q}^T\end{pmatrix}\begin{pmatrix}\boldsymbol{y}_P\\ \boldsymbol{y}_Q\end{pmatrix} \tag{23}$$

$$= (\boldsymbol{P}^T\boldsymbol{P} + \boldsymbol{Q}^T\boldsymbol{Q})^{-1}\boldsymbol{P}^T\boldsymbol{y}_P + (\boldsymbol{P}^T\boldsymbol{P} + \boldsymbol{Q}^T\boldsymbol{Q})^{-1}\boldsymbol{Q}^T\boldsymbol{y}_Q \tag{24}$$

By assumption 3, all rows of $\boldsymbol{P}$ and $\boldsymbol{Q}$ are sampled from a common Normal Distribution. Thus we are able to utilize properties of the Wishart Distribution, Nydick (2012).

$$\boldsymbol{P}^T\boldsymbol{P} = \sum_{i=1}^{n*(1-\epsilon)} \boldsymbol{P}_i\boldsymbol{P}_i^T \tag{25}$$

$$\boldsymbol{Q}^T\boldsymbol{Q} = \sum_{j=1}^{n\epsilon} \boldsymbol{Q}_i\boldsymbol{Q}_i^T \tag{26}$$

Thus we can say $\boldsymbol{P}^T\boldsymbol{P}$ and $\boldsymbol{Q}^T\boldsymbol{Q}$ are sampled from the Wishart distribution.

$$\boldsymbol{P}^T\boldsymbol{P} \sim \mathcal{W}(n(1-\epsilon), \boldsymbol{\Sigma}) \tag{27}$$

$$\boldsymbol{Q}^T\boldsymbol{Q} \sim \mathcal{W}(n\epsilon, \boldsymbol{\Sigma}) \tag{28}$$

We can now use the Expected Value of the Wishart Distribution.

$$\mathbb{E}(\boldsymbol{P}^T\boldsymbol{P}) = n(1-\epsilon)\boldsymbol{\Sigma} \tag{29}$$

$$\mathbb{E}(\boldsymbol{Q}^T\boldsymbol{Q}) = n\epsilon\boldsymbol{\Sigma} \tag{30}$$

It thus follows

$$\boldsymbol{P}^T\boldsymbol{P} + \boldsymbol{Q}^T\boldsymbol{Q} = n\boldsymbol{\Sigma} \tag{31}$$

Since we are interested in the pseudo-inverse, we will utilize the Inverse Wishart Distribution.

$$\left(\boldsymbol{P}^T\boldsymbol{P} + \boldsymbol{Q}^T\boldsymbol{Q}\right)^{-1} \sim \mathcal{W}^{-1}(n, \boldsymbol{\Sigma}) \tag{32}$$

It thus follows by the expectation of the Inverse Wishart Distribution

$$\mathbb{E}\left[\left(\boldsymbol{P}^T\boldsymbol{P} + \boldsymbol{Q}^T\boldsymbol{Q}\right)^{-1}\right] = n\boldsymbol{\Sigma}^{-1} \tag{33}$$

Now we will plug this into Equation 24:

$$\mathbb{E}\left[\boldsymbol{X}^\dagger\boldsymbol{y}\right] = \left(n\boldsymbol{\Sigma}^{-1}\right)\left(n(1-\epsilon)\boldsymbol{\Sigma}\right)^T\boldsymbol{y}_P + \left(n\boldsymbol{\Sigma}^{-1}\right)\left(n\epsilon\boldsymbol{\Sigma}\right)^T\boldsymbol{y}_Q \tag{34}$$

Note the optimal solution for a linear regression model on $\mathbb{P}$ is $(\boldsymbol{P}^T\boldsymbol{P})^{-1}\boldsymbol{P}^T\boldsymbol{y}_P$
Often times in the case of corrupted data we have $\boldsymbol{P}$ and $\boldsymbol{Q}$ are sampled similarly however $\boldsymbol{y}_P$ and

$\boldsymbol{y}_Q$ are very different. Thus $(\boldsymbol{P}^T\boldsymbol{P} + \boldsymbol{Q}^T\boldsymbol{Q})^{-1}\boldsymbol{Q}^T\boldsymbol{y}_Q$ could have a large effect on the optimal solution. This is why we propose Sub-Quantile Optimization, we seek to reduce the impact of $\boldsymbol{Q}^T\boldsymbol{Q}$ and $\boldsymbol{y}_Q$ by reducing the number of rows in $\boldsymbol{Q}$. Thus we reduce the condition number of $\boldsymbol{Q}^T\boldsymbol{Q}$ and the overall effect on the optimal solution for $\boldsymbol{P}$.

Here we utilize the idea of *influence* from McWilliams et al. (2014).

By Lemma 4.2, $\boldsymbol{\theta}$ is updated only on the np points with the smallest squared loss. To quantify how "Robust" our linear regressor is, we want to know how many data points from $\mathbb{Q}$ are within the lowest np squared losses as the number of iterations, $k \to \infty$. To do this, we will model the data sampled from $\mathbb{P}$ and $\mathbb{Q}$ as random variables. Let $\boldsymbol{p}_1, \boldsymbol{p}_2, \ldots, \boldsymbol{p}_{(1-\epsilon)n}$ be the $(1-\epsilon)n$ points sampled i.i.d from $\mathbb{P}$. Let $P_1, P_2, \ldots, P_m$ be random variables that represent the data sampled from $\mathbb{P}$, $\boldsymbol{p}_1, \boldsymbol{p}_2, \ldots, \boldsymbol{p}_{1-(\epsilon)n}$ such that

$$P_i = \begin{cases} 1 & \text{if } (\boldsymbol{\theta}_k^T\boldsymbol{p}_i - y_i)^2 \leq \hat{\boldsymbol{\nu}}_{np} \\ 0 & \text{if } (\boldsymbol{\theta}_k^T\boldsymbol{p}_i - y_i)^2 > \hat{\boldsymbol{\nu}}_{np} \end{cases} \tag{35}$$

Let $\boldsymbol{q}_1, \boldsymbol{q}_2, \ldots, \boldsymbol{q}_{\epsilon n}$ be the ϵn points sampled i.i.d from $\mathbb{Q}$. Let $Q_1, Q_2, \ldots, Q_m$ be random variables that represent the data sampled from $\mathbb{Q}$, $\boldsymbol{q}_1, \boldsymbol{q}_2, \ldots, \boldsymbol{q}_{\epsilon n}$ such that

$$Q_i = \begin{cases} 1 & \text{if } (\boldsymbol{\theta}_k^T\boldsymbol{q}_i - y_i)^2 \leq \hat{\boldsymbol{\nu}}_{np} \\ 0 & \text{if } (\boldsymbol{\theta}_k^T\boldsymbol{q}_i - y_i)^2 > \hat{\boldsymbol{\nu}}_{np} \end{cases} \tag{36}$$

It is clear that $\mathbb{P}\left[Q_i = 1\right] = 1 - \mathbb{P}\left[P_i = 1\right]$ So it is only necesary to calculate $\mathbb{P}\left[P_i = 1\right]$

Furthermore, we will define another random variable to determine the number of corrupted samples within the np lowest squared losses after optimization iteration k.

$$Q_k^+ = \sum_{i=1}^{n\epsilon} Q_i \text{ and } P_k^+ = \sum_{i=1}^{n(1-\epsilon)} P_i \tag{37}$$

Let $\boldsymbol{p}_1, \boldsymbol{p}_2, \ldots, \boldsymbol{p}_m$ represent the points sampled from $\mathbb{P}$ within the lowest np squared losses and let $\boldsymbol{q}_1, \boldsymbol{q}_2, \ldots, \boldsymbol{q}_l$ represent the points sampled from $\mathbb{Q}$ within the lowest np squared losses, where $m = \mathbb{E}\left[P_0^+\right]$ and $l = \mathbb{E}\left[Q_0^+\right]$. We will first note the following

From here we can calculate the expected update rule

$$\boldsymbol{\theta}_1 = \boldsymbol{\theta}_0 + \alpha \sum_{i=1}^{m} 2\boldsymbol{p}_i(\boldsymbol{\theta}_0^T\boldsymbol{p}_i - y_i) + \alpha \sum_{i=1}^{l} 2\boldsymbol{q}_i(\boldsymbol{\theta}_0^T\boldsymbol{q}_i - y_i) \tag{38}$$

Lemma 4.8. *$f_i(\boldsymbol{\theta})$ is a Lipschitz Continuous with parameter $L = ||\boldsymbol{x}_i||_2^2$*

Proof. The proof is quite simple in optimization theory, we provide the full proof in Appendix B.4 □

Let us define two functions for the empirical loss on $\mathbb{P}$ and $\mathbb{Q}$

$$\phi(\boldsymbol{\theta}) = \frac{1}{np}\sum_{i=1}^{m}(\boldsymbol{\theta}^T\boldsymbol{p}_i - y_i)^2 \tag{39}$$

$$\psi(\boldsymbol{\theta}) = \frac{1}{np}\sum_{i=1}^{l}(\boldsymbol{\theta}^T\boldsymbol{q}_i - y_i)^2 \tag{40}$$

These two functions hold nice properties.

$$\nabla_{\boldsymbol{\theta}}\phi(\boldsymbol{\theta}) = \frac{1}{np}\sum_{i=1}^{m} 2\boldsymbol{p}_i(\boldsymbol{\theta}^T\boldsymbol{p}_i - y_i) \tag{41}$$

$$\nabla_{\boldsymbol{\theta}}\psi(\boldsymbol{\theta}) = \frac{1}{np}\sum_{i=1}^{l} 2\boldsymbol{q}_i(\boldsymbol{\theta}^T\boldsymbol{q}_i - y_i) \tag{42}$$

Here we note that the summation of these derivatives is equal to the theta update

$$\nabla_{\boldsymbol{\theta}_k}(t_{k+1}, \boldsymbol{\theta}_k) = \nabla_{\boldsymbol{\theta}_k}\phi(\boldsymbol{\theta}) + \nabla_{\boldsymbol{\theta}_k}\psi(\boldsymbol{\theta}) \tag{43}$$

Assumption 4. *Since we randomly choose the parameters of $\boldsymbol{\theta}_0$ and we assume since $m > l$ by expectation,*

$$\nabla_{\boldsymbol{\theta}_0}\phi(\boldsymbol{\theta}_0) > \nabla_{\boldsymbol{\theta}_0}\psi(\boldsymbol{\theta}_0) \tag{44}$$

4.1.1 Point Change Conditions

In this section we mathematically reason the conditions for a point *initially* outside the lowest np squared losses to come within the lowest np squared losses.

Let us take two data points $\boldsymbol{x}$ and $\boldsymbol{x}'$ such that $\boldsymbol{x} \leq t_0$ and $\boldsymbol{x}' > t_0$

Theorem 2. *The rate of decrease of $\boldsymbol{x}'$ is greater than $\boldsymbol{x}$ iff*

$$-2r' \left\|\boldsymbol{x}'\right\| \cos(\omega') + \alpha \left\|\nabla_{\boldsymbol{\theta}_k}\right\| \left\|\boldsymbol{x}'\right\|^2 \cos^2(\omega') \geq -2r \left\|\boldsymbol{x}\right\| \cos(\omega) + \alpha \left\|\nabla_{\boldsymbol{\theta}_k}\right\| \left\|\boldsymbol{x}\right\|^2 \cos^2(\omega) \tag{45}$$

Theorem 2 reveals to us the importance of $\cos(\omega)$ which represents the angle between $\nabla f_{\boldsymbol{x}}(\boldsymbol{\theta}_0)$ and $\nabla g(\boldsymbol{\theta}_0)$. By our initial assumption $|\boldsymbol{\theta}_0^T\boldsymbol{x}' - y'| > |\boldsymbol{\theta}_0^T\boldsymbol{x} - y|$. Let us look at the example where $||\boldsymbol{x}|| = ||\boldsymbol{x}'||$, in this case, while $\cos(\omega') > \cos(\omega)$, the rate of decrease of $\boldsymbol{x}'$ will be more than the rate of decrease of $\boldsymbol{x}$. What this means is there will be an iteration step where $(\boldsymbol{\theta}_k^T\boldsymbol{x}' - y')^2 < (\boldsymbol{\theta}_k^T\boldsymbol{x} - y)^2$. Thus $\boldsymbol{x}'$ will come within the lowest np squared losses and $\boldsymbol{x}$ will no longer be in the np lowest square losses. We can now formulate our convergence conditions.
For all points outside the np lowest squared losses. There exists no point within the np lowest squared losses such that for any $k \in \mathbb{N}$, the points within the np lowest squared losses do not change.

5 Numerical Experiments

The first experiment we will run will display the difference of the following two t updates

$$t_{k+1} = \hat{\boldsymbol{\nu}}_{np} \tag{46}$$

$$t_{k+1} = \frac{1}{np}\sum_{i=1}^{np}\hat{\boldsymbol{\nu}}_i \tag{47}$$

In general, if the $\hat{\boldsymbol{\nu}}_1, \hat{\boldsymbol{\nu}}_2, \ldots, \hat{\boldsymbol{\nu}}_{np}$ are closely distributed, then $\frac{1}{np}\sum_{i=1}^{np}\hat{\boldsymbol{\nu}}_i \approx \hat{\boldsymbol{\nu}}_{np}$. In Algorithm 1, we display our training method for Sub-quantile Optimization with the t update as described in equation 46. We also compare against the t-update as described in equation 47.

```
Algorithm 1: Sub-Quantile Minimization Optimization Algorithm
Input: Training iterations $T$, Quantile $p$, Corruption Percentage $\epsilon$, Input Parameters $m$
Output: Trained Parameters, $\boldsymbol{\theta}$
Data: Inliers: $y|x \sim \mathcal{N}(x^2 - x + 2, 0.01)$, Outliers: $y|x \sim \mathcal{N}(-x^2 + x + 4, 0.01)$
$\boldsymbol{\theta}_1 \leftarrow \mathcal{N}(0, \sigma)^d$
for $k \in 1, 2, \ldots, m$ do
    $\boldsymbol{\nu} = (\boldsymbol{X}\boldsymbol{\theta}_k - \boldsymbol{y})^2$
    $\hat{\boldsymbol{\nu}} = sorted(\boldsymbol{\nu})$
    $t_{k+1} = \hat{\boldsymbol{\nu}}_{np}$
    $t_{k+1} = \frac{1}{np}\sum_{i=1}^{np}\boldsymbol{\nu}_i$
    $L := \sum_{i=1}^{np}\boldsymbol{x}_i^T\boldsymbol{x}_i$
    $\alpha := \frac{1}{2L}$
    $\boldsymbol{\theta}_{k+1} = \boldsymbol{\theta}_k - \alpha\nabla_{\boldsymbol{\theta}_k}g(t_{k+1}, \boldsymbol{\theta}_k)$
end
return $\boldsymbol{\theta}_T$
```

We also present a batch algorithm which improves training speed significantly.

Algorithm 2: Stochastic Sub-Quantile Minimization Optimization Algorithm

```
Input: Training iterations $T$, Quantile $p$, Corruption Percentage $\epsilon$, Input Parameters $d$, Batch
       Size $m$
Output: Trained Parameters, $\boldsymbol{\theta}$
Data: Inliers: $y|x \sim \mathcal{N}(x^2 - x + 2, 0.01)$, Outliers: $y|x \sim \mathcal{N}(-x^2 + x + 4, 0.01)$
$\boldsymbol{\theta}_1 \leftarrow \mathcal{N}(0, \sigma)^d$
for $k \in 1, 2, \ldots, T$ do
    $I \subseteq [n]$ of size $m$
    $\boldsymbol{\nu} = (\boldsymbol{X}_I \boldsymbol{\theta}_k - \boldsymbol{y}_I)^2$
    $\hat{\boldsymbol{\nu}} = sorted(\boldsymbol{\nu})$
    $t_{k+1} = \hat{\boldsymbol{\nu}}_{mp}$
    $t_{k+1} = \frac{1}{mp} \sum_{i=1}^{mp} \boldsymbol{\nu}_i$
    $L := \sum_{i=1}^{mp} \boldsymbol{x}_i^T \boldsymbol{x}_i$
    $\alpha := \frac{1}{2L}$
    $\boldsymbol{\theta}_{k+1} = \boldsymbol{\theta}_k - \alpha \nabla_{\boldsymbol{\theta}_k} g(t_{k+1}, \boldsymbol{\theta}_k)$
end
return $\boldsymbol{\theta}_T$
```

5.1 Synthetic Data

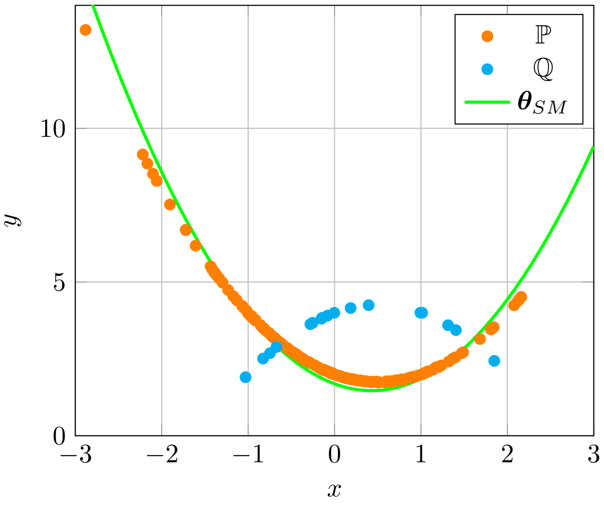

Figure 1: Quadratic Regression, $p = 0.9$, $t_{k+1} = \hat{\boldsymbol{\nu}}_{np}$

In our first synthetic experiment, we run Algorithm 1 on synthetically generated quadratic data. The results of Sub-Quantile Minimization can be seen in Figure **??**. Our results compared with the current State of the Art and Baseline Methods can be seen in Table **??**. Note we are not interested in $\epsilon > 0.5$ as the concept of corruptness becomes unclear. We see in Table **??**, Sub-Quantile Minimization produces State of the Art Results in the Quadratic Regression Case. Furthermore, it performs significantly better than baseline methods in the high-noise regimes ($\epsilon = 0.4$), this is confirmed in both the small data and large data datasets. Please refer to Appendix E for more details on the Quadratic Regression Dataset.

5.2 Real Data

We provide results on the *Drug Discovery* Dataset in Diakonikolas et al. (2019)

Table 1: Quadratic Regression Synthetic Dataset. Empirical Risk over $\mathbb{P}$

Objectives	$n = 10^4$		$n = 10^6$	
	$\epsilon = 0.2$	$\epsilon = 0.4$	$\epsilon = 0.2$	$\epsilon = 0.4$
OLS 104	$60.06_{(3.472)}$	$107.78_{(4.87)}$	$621.46_{(4.32)}$	$1076.3_{(2.9)}$
Huber Huber & Ronchetti (2009)	$0.942_{(0.008)}$	$26.27_{(7.50)}$	$9.419_{(0.005)}$	$261.48_{(3.96)}$
TERM Li et al. (2020)	∞	∞	∞	∞
SubQuantile	$\mathbf{0.896}_{(\mathbf{0.007})}$	$\mathbf{0.777}_{(\mathbf{0.007})}$	$\mathbf{8.946}_{(\mathbf{0.005})}$	$\mathbf{7.75}_{(\mathbf{0.009})}$

6 Conclusion

In this work we provide a theoretical analysis for robust linear regression by minimizing *Sub-Quantile* of the Empirical Risk. Furthermore, we run various numerical experiments and compare against the current State of the Art in Robust Linear Regression.

Author Contributions

Acknowledgments

REFERENCES

Pranjal Awasthi, Abhimanyu Das, Weihao Kong, and Rajat Sen. Trimmed maximum likelihood estimation for robust learning in generalized linear models, 2022. URL https://arxiv.org/abs/2206.04777.

Ilias Diakonikolas, Gautam Kamath, Daniel M. Kane, Jerry Li, Jacob Steinhardt, and Alistair Stewart. Sever: A robust meta-algorithm for stochastic optimization. In *Proceedings of the 36th International Conference on Machine Learning*, ICML '19, pp. 1596–1606. JMLR, Inc., 2019.

Peter J. Huber and Elvezio. Ronchetti. *Robust statistics*. Wiley series in probability and statistics. Wiley, Hoboken, N.J., 2nd ed. edition, 2009. URL http://catdir.loc.gov/catdir/toc/ecip0824/2008033283.html.

Chi Jin, Praneeth Netrapalli, and Michael I. Jordan. What is local optimality in nonconvex-nonconcave minimax optimization?, 2019. URL https://arxiv.org/abs/1902.00618.

Roger Koenker and Kevin F. Hallock. Quantile regression. *Journal of Economic Perspectives*, 15(4): 143–156, December 2001. doi: 10.1257/jep.15.4.143. URL https://www.aeaweb.org/articles?id=10.1257/jep.15.4.143.

Tian Li, Ahmad Beirami, Maziar Sanjabi, and Virginia Smith. Tilted empirical risk minimization. *arXiv preprint arXiv:2007.01162*, 2020.

Lu Lu. Dying ReLU and initialization: Theory and numerical examples. *Communications in Computational Physics*, 28(5):1671–1706, 2020. doi: 10.4208/cicp.oa-2020-0165. URL https://doi.org/10.4208%2Fcicp.oa-2020-0165.

Brian McWilliams, Gabriel Krummenacher, Mario Lucic, and Joachim M. Buhmann. Fast and robust least squares estimation in corrupted linear models. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'14, pp. 415–423, Cambridge, MA, USA, 2014. MIT Press.

David M. Mount, Nathan S. Netanyahu, Christine D. Piatko, Ruth Silverman, and Angela Y. Wu. On the least trimmed squares estimator. *Algorithmica*, 69(1):148–183, 2014. doi: 10.1007/s00453-012-9721-8. URL https://doi.org/10.1007/s00453-012-9721-8.

Steven W Nydick. The wishart and inverse wishart distributions. *Electronic Journal of Statistics*, 6 (1-19), 2012.

Muhammad Osama, Dave Zachariah, and Petre Stoica. Robust risk minimization for statistical learning from corrupted data. *IEEE Open Journal of Signal Processing*, 1:287–294, 2020.

Meisam Razaviyayn, Tianjian Huang, Songtao Lu, Maher Nouiehed, Maziar Sanjabi, and Mingyi Hong. Non-convex min-max optimization: Applications, challenges, and recent theoretical advances, 06 2020.

R.T. Rockafellar, J.O. Royset, and S.I. Miranda. Superquantile regression with applications to buffered reliability, uncertainty quantification, and conditional value-at-risk. *European Journal of Operational Research*, 234(1):140–154, 2014. ISSN 0377-2217. doi: https://doi.org/10.1016/j.ejor.2013.10.046. URL https://www.sciencedirect.com/science/article/pii/S0377221713008692.

Pranab Kumar Sen. Estimates of the regression coefficient based on kendall's tau. *Journal of the American Statistical Association*, 63(324):1379–1389, 1968. doi: 10.1080/01621459.1968.10480934. URL https://www.tandfonline.com/doi/abs/10.1080/01621459.1968.10480934.

A Assumptions

B General Properties of Sub-Quantile Linear Regression

B.1 Proof of Lemma 4.1

Proof. Since $g(t, \boldsymbol{\theta})$ is a concave function. Maximizing $g(t, \boldsymbol{\theta})$ is equivalent to minimizing $-g(t, \boldsymbol{\theta})$. We will find fermat's optimality condition for the function $-g(t, \boldsymbol{\theta})$, which is convex. Let $\hat{\boldsymbol{\nu}} = sorted\left((\boldsymbol{\theta}^T \boldsymbol{X} - \boldsymbol{y})^2\right)$ and note $0 < p < 1$

$$\begin{aligned}
\partial(-g(t, \boldsymbol{\theta})) &= \partial\left(-t + \frac{1}{np}\sum_{i=1}^{n}(t - \hat{\boldsymbol{\nu}}_i)^+\right) \tag{48}\\
&= \partial(-t) + \partial\left(\frac{1}{np}\sum_{i=1}^{n}(t - \hat{\boldsymbol{\nu}}_i)^+\right) \tag{49}\\
&= -1 + \frac{1}{np}\sum_{i=1}^{n}\partial(t - \hat{\boldsymbol{\nu}}_i)^+ \tag{50}\\
&= -1 + \frac{1}{np}\sum_{i=1}^{n}\left\{\begin{array}{ll} 1, & \text{if } t > \hat{\boldsymbol{\nu}}_i \\ 0, & \text{if } t < \hat{\boldsymbol{\nu}}_i \\ \left[0, 1\right], & \text{if } t = \hat{\boldsymbol{\nu}}_i \end{array}\right\} \tag{51}\\
&= 0 \text{ when } t = \hat{\boldsymbol{\nu}}_{np} \tag{52}
\end{aligned}$$

This is the p-quantile of $\boldsymbol{\nu}$. Not necesarily the p-quantile of $Q_p(U)$ □

B.2 Proof of Lemma 4.2

Proof. Note that $t_k = \boldsymbol{\nu}_{np}$ which is equivalent to $(\boldsymbol{\theta}_k^T \boldsymbol{x}_{np} - y_{np})^2$

$$\begin{aligned}
\nabla_{\boldsymbol{\theta}_k} g(t_{k+1}, \boldsymbol{\theta}_k) &= \nabla_{\boldsymbol{\theta}_k}\left(\boldsymbol{\nu}_{np} - \frac{1}{np}\sum_{i=1}^{n}(\boldsymbol{\nu}_{np} - (\boldsymbol{\theta}_k^T \boldsymbol{x}_i - y_i)^2)^+\right) \tag{53}\\
&= \nabla_{\boldsymbol{\theta}_k}\left((\boldsymbol{\theta}_k^T \boldsymbol{x}_{np} - y_{np})^2 - \frac{1}{np}\sum_{i=1}^{n}\left((\boldsymbol{\theta}_k^T \boldsymbol{x}_{np} - y_{np})^2 - (\boldsymbol{\theta}_k^T \boldsymbol{x}_i - y_i)^2\right)^+\right) \tag{54}\\
&= \nabla_{\boldsymbol{\theta}_k}(\boldsymbol{\theta}_k^T \boldsymbol{x}_{np} - y_{np})^2 - \frac{1}{np}\sum_{i=1}^{n}\nabla_{\boldsymbol{\theta}_k}\left((\boldsymbol{\theta}_k^T \boldsymbol{x}_{np} - y_{np})^2 - (\boldsymbol{\theta}_k^T \boldsymbol{x}_i - y_i)^2\right)^+ \tag{55}\\
&= 2\boldsymbol{x}_{np}(\boldsymbol{\theta}_k^T \boldsymbol{x}_{np} - y_{np}) - \frac{1}{np}\sum_{i=1}^{n} 2\boldsymbol{x}_{np}(\boldsymbol{\theta}_k^T \boldsymbol{x}_{np} - y_{np}) \nonumber\\
&\qquad - 2\boldsymbol{x}_i(\boldsymbol{\theta}_k^T \boldsymbol{x}_i - y_i)\left\{\begin{array}{ll} 1, & \text{if } t > v_i \\ 0, & \text{if } t < v_i \\ \left[0, 1\right], & \text{if } t = v_i \end{array}\right\} \tag{56}\\
&= 2\boldsymbol{x}_{np}(\boldsymbol{\theta}_k^T \boldsymbol{x}_{np} - y_{np}) - \frac{1}{np}\sum_{i=1}^{np} 2\boldsymbol{x}_{np}(\boldsymbol{\theta}_k^T \boldsymbol{x}_{np} - y_{np}) - 2\boldsymbol{x}_i(\boldsymbol{\theta}_k^T \boldsymbol{x}_i - y_i) \tag{57}\\
&= 2\boldsymbol{x}_{np}(\boldsymbol{\theta}_k^T \boldsymbol{x}_{np} - y_{np}) - 2\boldsymbol{x}_{np}(\boldsymbol{\theta}_k^T \boldsymbol{x}_{np} - y_{np}) + \frac{1}{np}\sum_{i=1}^{np} 2\boldsymbol{x}_i(\boldsymbol{\theta}_k^T \boldsymbol{x}_i - y_i) \tag{58}\\
&= \frac{1}{np}\sum_{i=1}^{np} 2\boldsymbol{x}_i(\boldsymbol{\theta}_k^T \boldsymbol{x}_i - y_i) \tag{59}
\end{aligned}$$

This is the derivative of the np samples with lowest error with respect to $\boldsymbol{\theta}$. □

B.3 Proof of Lemma ??

Proof. The probability a point from $\mathbb{P}$ is within the np lowest squared points is equivalent to the probability of a point from $\mathbb{P}$ being within the p quantile of the combined distribution of $\mathbb{P}$ and $\mathbb{Q}$. Let μ_P be the average loss over all points in $\mathbb{P}$ and μ_Q be the average loss over all points in $\mathbb{Q}$. Similarly let σ_P^2 and σ_Q^2 be the respective variances.

$$\mu_P = \frac{1}{n(1-\epsilon)} \sum_{i=1}^{n(1-\epsilon)} (\boldsymbol{\theta}^T \boldsymbol{p}_i - y_i)^2 \tag{60}$$

$$\mu_Q = \frac{1}{n\epsilon} \sum_{i=1}^{n\epsilon} (\boldsymbol{\theta}^T \boldsymbol{q}_i - y_i)^2 \tag{61}$$

$$\sigma_P^2 = \frac{1}{n(1-\epsilon)-1} \sum_{i=1}^{n(1-\epsilon)} (\mu_P - (\boldsymbol{\theta}^T \boldsymbol{p}_i - y_i)^2)^2 \tag{62}$$

$$\sigma_Q^2 = \frac{1}{n\epsilon - 1} \sum_{i=1}^{n\epsilon} (\mu_Q - (\boldsymbol{\theta}^T \boldsymbol{q}_i - y_i)^2)^2 \tag{63}$$

Let us now calculate the combined distribution, which by our problem statement is $\hat{\mathbb{P}}$.

$$\mu_{\hat{P}} = (1-\epsilon)\mu_P + \epsilon\mu_Q \tag{64}$$

$$\begin{aligned} \sigma_{\hat{P}} &= (1-\epsilon)^2\sigma_P^2 + \epsilon^2\sigma_Q^2 + \epsilon(1-\epsilon)Cov(P,Q) && (65) \\ &= (1-\epsilon)^2\sigma_P^2 + \epsilon^2\sigma_Q^2 && (66) \end{aligned}$$

Notice the Covariance is 0 because the samples are i.i.d. Now we will calculate the p-quantile of $\mathbb{Z}$. Let $\Phi \sim \mathcal{N}(0,1)$

$$Q_p(\hat{\mathbb{P}}) = \mu_{\hat{P}} + \Phi^{-1}(p)\sigma_{\hat{P}} \tag{67}$$

$Q_p(\hat{\mathbb{P}})$ represents the npth squared loss. We know want to know what is the probability a point from $\mathbb{P}$ is below this.

$$\begin{aligned} \mathbb{P}\left[P_i < Q_p(\hat{\mathbb{P}})\right] &= \Phi\left(\frac{Q_p(\hat{\mathbb{P}}) - \mu_P}{\sigma_P}\right) && (68) \\ &= \Phi\left(\frac{(1-\epsilon)\mu_P + \epsilon\mu_Q + \Phi^{-1}(p)((1-\epsilon)^2\sigma_P^2 + \epsilon^2\sigma_Q^2) - \mu_P}{\sigma_P}\right) && (69) \\ &= \Phi\left(\frac{-\epsilon\mu_P + \epsilon\mu_Q + \Phi^{-1}(p)((1-\epsilon)^2\sigma_P^2 + \epsilon^2\sigma_Q^2)}{\sigma_P}\right) && (70) \end{aligned}$$

□

B.4 Proof of Lemma 4.8

Proof. Note we defined $g_i(\boldsymbol{\theta}) = (\boldsymbol{\theta}^T \boldsymbol{x}_i - y_i)^2$, thus $\nabla g_i(\boldsymbol{\theta}) = 2\boldsymbol{x}_i(\boldsymbol{\theta}^T \boldsymbol{x}_i - y_i)^2$. We will prove there exists β such that

$$||\nabla g_i(\boldsymbol{\theta}') - \nabla g_i(\boldsymbol{\theta})|| \leq \beta||\boldsymbol{\theta}' - \boldsymbol{\theta}|| \tag{71}$$

The proof is as follows

$$\begin{aligned} ||\nabla g_i(\boldsymbol{\theta}') - \nabla g_i(\boldsymbol{\theta})|| &= ||2\boldsymbol{x}_i(\boldsymbol{\theta}'^T \boldsymbol{x}_i - y_i) - 2\boldsymbol{x}_i(\boldsymbol{\theta}^T \boldsymbol{x}_i - y_i)|| && (72) \\ &= ||2\boldsymbol{x}_i(\boldsymbol{\theta}'^T \boldsymbol{x}_i) - 2\boldsymbol{x}_i y_i - 2\boldsymbol{x}_i(\boldsymbol{\theta}^T \boldsymbol{x}_i) + 2\boldsymbol{x}_i y_i|| && (73) \\ &= ||2\boldsymbol{x}_i(\boldsymbol{\theta}'^T \boldsymbol{x}_i - \boldsymbol{\theta}^T \boldsymbol{x}_i)|| && (74) \\ &= ||2\boldsymbol{x}_i||\,||\boldsymbol{\theta}'^T \boldsymbol{x}_i - \boldsymbol{\theta}^T \boldsymbol{x}_i|| && (75) \\ &= 2||\boldsymbol{x}_i||^2||\boldsymbol{\theta}' - \boldsymbol{\theta}|| && (76) \end{aligned}$$

Thus $g_i(\boldsymbol{\theta})$ is $||\boldsymbol{x}_i||^2$-smooth □

B.5 Proof of Theorem 2

Proof. As given in the assumption, $f_{\boldsymbol{x}}(\boldsymbol{\theta}) < f_{\boldsymbol{x}'}(\boldsymbol{\theta})$. So we are interested in the condition for $f_{\boldsymbol{x}'}(\boldsymbol{\theta}_1) - f_{\boldsymbol{x}'}(\boldsymbol{\theta}_0) < f_{\boldsymbol{x}}(\boldsymbol{\theta}_1) - f_{\boldsymbol{x}}(\boldsymbol{\theta}_0)$. We will calculate $f_{\boldsymbol{x}}(\boldsymbol{\theta}_1) - f_{\boldsymbol{x}}(\boldsymbol{\theta}_0)$ and generalize the results for $\boldsymbol{x}'$.

$$f_{\boldsymbol{x}}(\boldsymbol{\theta}_1) - f_{\boldsymbol{x}}(\boldsymbol{\theta}_0) = (\boldsymbol{\theta}_1^T \boldsymbol{x} - y)^2 - (\boldsymbol{\theta}_0^T - y)^2 \tag{77}$$

$$= (\boldsymbol{\theta}_1^T \boldsymbol{x})^2 - 2(\boldsymbol{\theta}_1^T \boldsymbol{x})y - (\boldsymbol{\theta}_0^T \boldsymbol{x})^2 + 2(\boldsymbol{\theta}_0^T \boldsymbol{x})y \tag{78}$$

$$= ((\boldsymbol{\theta}_0 - \alpha\nabla g(\boldsymbol{\theta}_0))^T \boldsymbol{x})^2 - 2((\boldsymbol{\theta}_0 - \alpha\nabla g(\boldsymbol{\theta}_0))^T \boldsymbol{x})y \\ - (\boldsymbol{\theta}_0^T \boldsymbol{x})^2 + 2(\boldsymbol{\theta}_0^T \boldsymbol{x})y \tag{79}$$

Note $\boldsymbol{\theta}_1 = \boldsymbol{\theta}_0 - \alpha\nabla g(t_1, \boldsymbol{\theta})$ by Equation 11

$$= (\boldsymbol{\theta}_0^T \boldsymbol{x} - \alpha\nabla g(\boldsymbol{\theta}_0)^T \boldsymbol{x})^2 - 2(\boldsymbol{\theta}_0^T \boldsymbol{x})y + 2\alpha(\nabla g(\boldsymbol{\theta})^T \boldsymbol{x})y \\ - (\boldsymbol{\theta}_0^T \boldsymbol{x})^2 + 2(\boldsymbol{\theta}_0^T \boldsymbol{x})y \tag{80}$$

$$= (\boldsymbol{\theta}_0^T \boldsymbol{x})^2 - 2\alpha(\boldsymbol{\theta}_0^T)(\nabla g(\boldsymbol{\theta})^T \boldsymbol{x}) + \alpha^2(\nabla g(\boldsymbol{\theta})^T \boldsymbol{x})^2 + 2\alpha(\nabla g(\boldsymbol{\theta})^T \boldsymbol{x})y \\ - (\boldsymbol{\theta}_0^T \boldsymbol{x})^2 \tag{81}$$

$$= -2\alpha(\boldsymbol{\theta}_0^T)(\nabla g(\boldsymbol{\theta})^T \boldsymbol{x}) + \alpha^2(\nabla g(\boldsymbol{\theta})^T \boldsymbol{x})^2 + 2\alpha(\nabla g(\boldsymbol{\theta})^T \boldsymbol{x})y \tag{82}$$

$$= \alpha(\nabla g(\boldsymbol{\theta}_0)^T \boldsymbol{x})(-2(\boldsymbol{\theta}_0)^T \boldsymbol{x} + \alpha\nabla g(\boldsymbol{\theta}_0)^T \boldsymbol{x} + 2y) \tag{83}$$

$$= \alpha(\nabla g(\boldsymbol{\theta}_0)^T \boldsymbol{x})(-2(\boldsymbol{\theta}_0^T \boldsymbol{x} - y)) + \alpha\nabla g(\boldsymbol{\theta}_0)^T \boldsymbol{x}) \tag{84}$$

Note $\nabla_{\boldsymbol{\theta}} f_{\boldsymbol{x}}(\boldsymbol{\theta}) = 2\boldsymbol{x}(\boldsymbol{\theta}^T \boldsymbol{x} - y)$

$$= -\alpha\nabla g(\boldsymbol{\theta}_0)^T \nabla f_{\boldsymbol{x}}(\boldsymbol{\theta}_0) + \alpha^2(\nabla g(\boldsymbol{\theta}_0)^T \boldsymbol{x})^2 \tag{85}$$

$$= -\alpha\left(||\nabla g(\boldsymbol{\theta}_0)||||\nabla f_{\boldsymbol{x}}(\boldsymbol{\theta}_0)||\cos(\omega) - \alpha||\nabla g(\boldsymbol{\theta}_0)||^2||\boldsymbol{x}||^2\cos^2(\eta)\right) \tag{86}$$

$$= -\alpha||\nabla g(\boldsymbol{\theta}_0)||\left(||f_{\boldsymbol{x}}(\boldsymbol{\theta}_0)||\cos(\omega) - \alpha||\nabla g(\boldsymbol{\theta}_0)||||\boldsymbol{x}||^2\cos^2(\eta)\right) \tag{87}$$

$$= -\alpha||\nabla g(\boldsymbol{\theta}_0)||\left(2||\boldsymbol{x}|||\boldsymbol{\theta}_0^T \boldsymbol{x} - y|\cos(\omega) - \alpha||\nabla g(\boldsymbol{\theta}_0)||||\boldsymbol{x}||^2\cos^2(\eta)\right) \tag{88}$$

$$= -\alpha||\nabla g(\boldsymbol{\theta}_0)||||\boldsymbol{x}||\left(2|||\boldsymbol{\theta}_0^T \boldsymbol{x} - y|\cos(\omega) - \alpha||\nabla g(\boldsymbol{\theta}_0)||||\boldsymbol{x}||\cos^2(\eta)\right) \tag{89}$$

Now we will generalize our results to the inequality $f_{\boldsymbol{x}'}(\boldsymbol{\theta}_1) - f_{\boldsymbol{x}'}(\boldsymbol{\theta}_0) < f_{\boldsymbol{x}}(\boldsymbol{\theta}_1) - f_{\boldsymbol{x}'}(\boldsymbol{\theta}_0)$

$$||\boldsymbol{x}'||(2|\boldsymbol{\theta}_0^T \boldsymbol{x}' - y|\cos(\omega') - \alpha||\nabla g(\boldsymbol{\theta})||||\boldsymbol{x}'||\cos^2(\eta')) \\ > ||\boldsymbol{x}||(2|\boldsymbol{\theta}_0^T \boldsymbol{x} - y|\cos(\omega) - \alpha||\nabla g(\boldsymbol{\theta})||||\boldsymbol{x}||\cos^2(\eta)) \tag{90}$$

Here we note that α is a very small term, $\alpha = \frac{1}{2L}$ where $L = ||\boldsymbol{X}^T\boldsymbol{X}||$ So equation 90 can be approximately simplied.

$$||\boldsymbol{x}'|||\boldsymbol{\theta}_0^T \boldsymbol{x}' - y'|\cos(\omega') > ||\boldsymbol{x}|||\boldsymbol{\theta}_0^T \boldsymbol{x} - y|\cos(\omega) \tag{91}$$

This completes the proof. □

C Proofs for Convergence

C.1 Proof of Lemma 4.5

The objective function $g(\boldsymbol{\theta}, t)$ is L-smooth w.r.t $\boldsymbol{\theta}$ iff

$$||\nabla_{\boldsymbol{\theta}} g(\boldsymbol{\theta}', t) - \nabla_{\boldsymbol{\theta}} g(\boldsymbol{\theta}, t)|| \leq L||\boldsymbol{\theta}' - \boldsymbol{\theta}|| \tag{92}$$

$$\left\|\nabla_{\boldsymbol{\theta}} g(\boldsymbol{\theta}^{'}, t) - \nabla_{\boldsymbol{\theta}} g(\boldsymbol{\theta}, t)\right\| = \left\|\frac{1}{np}\sum_{i=1}^{np} 2\boldsymbol{x}_i(\boldsymbol{\theta}_k^{'T}\boldsymbol{x}_i - y_i) - \frac{1}{np}\sum_{i=1}^{np} 2\boldsymbol{x}_i(\boldsymbol{\theta}_k^{T}\boldsymbol{x}_i - y_i)\right\| \tag{93}$$

$$= \left\|\frac{1}{np}\sum_{i=1}^{np} 2\boldsymbol{x}_i(\boldsymbol{\theta}_k^{'T}\boldsymbol{x}_i - \boldsymbol{\theta}_k^{T}\boldsymbol{x}_i)\right\| \tag{94}$$

$$= \left\|\frac{1}{np}\sum_{i=1}^{np} 2\boldsymbol{x}_i^T\boldsymbol{x}_i(\boldsymbol{\theta}_k^{'T} - \boldsymbol{\theta}_k^{T})\right\| \tag{95}$$

$$\overset{\text{Cauchy}-\text{Schwarz}}{\leq} \left\|\frac{2}{np}\sum_{i=1}^{np} \|\boldsymbol{x}_i\|^2\right\| \left\|\boldsymbol{\theta}_k^{'T} - \boldsymbol{\theta}_k^{T}\right\| \tag{96}$$

$$= L\left\|\boldsymbol{\theta}_k^{'T} - \boldsymbol{\theta}_k^{T}\right\| \tag{97}$$

where $L = \left\|\frac{2}{np}\sum_{i=1}^{np} \|\boldsymbol{x}_i\|^2\right\|$
This concludes the proof.

C.2 Proof of Lemma 4.6

As we noted in the proof sketch, proving $g(t_{k+1}, \boldsymbol{\theta}_{k+1}) < g(t_k, \boldsymbol{\theta}_k)$ is equivalent to proving

$$g(t_{k+1}, \boldsymbol{\theta}_k) - g(t_{k+1}, \boldsymbol{\theta}_{k+1}) \geq g(t_{k+1}, \boldsymbol{\theta}_{k+1}) - g(t_k, \boldsymbol{\theta}_k) \tag{98}$$

By equation 11 and lemma 4.5

$$g(t_{k+1}, \boldsymbol{\theta}_{k+1}) \leq g(t_{k+1}, \boldsymbol{\theta}) + \frac{1}{2L}\left\|\nabla_{\boldsymbol{\theta}_k} g(t_{k+1}, \boldsymbol{\theta}_k)\right\|^2 \tag{99}$$

Upper bound on the RHS

$$g(t_{k+1}, \boldsymbol{\theta}_k) - g(t_k, \boldsymbol{\theta}_k) = t_{k+1} - t_k - \frac{1}{np}\sum_{i=1}^{n}(t_{k+1} - \boldsymbol{\nu}_i)^+ + (t_k - \boldsymbol{\nu}_i)^+ \tag{100}$$

$$= -t_k + \frac{1}{np}\sum_{i=1}^{np}\boldsymbol{\nu}_i + \frac{1}{np}\sum_{i=1}^{n}(t_k - \boldsymbol{\nu}_i)^+ \tag{101}$$

$$\leq \frac{1}{np}\sum_{i=1}^{np}\boldsymbol{\nu}_i + \frac{1}{np}\sum_{i=1}^{n}(t_k - \boldsymbol{\nu}_i)^+ \tag{102}$$

Lower bound on the LHS

$$g \tag{103}$$

D Proofs for Convergence Rate

E Experimental Details

E.1 Experiments in Section 5

Here we will describe the objective functions used in the synthetic data experiments.

Ordinary Least Squares (OLS) can be solved utilizing the Moore Penrose Inverse.

$$\boldsymbol{X}^* = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y} \tag{104}$$

Huber Regression is solved with the following objective function.

$$L_\delta(y, f(\boldsymbol{x})) = \begin{cases} \frac{1}{2}(y - f(\boldsymbol{x}))^2 & \\ \delta \cdot \left(|y - f(\boldsymbol{x})| - \frac{1}{2}\delta\right) & \text{otherwise} \end{cases} \tag{105}$$