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 $\hat{\mathbb{P}} = (1-\varepsilon)\mathbb{P} + \varepsilon\mathbb{Q}$. 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.

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

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

$$\mathbb{L}_p(U) = \frac{1}{p}\int_0^p \mathcal{Q}_q(U)\, dq = \mathbb{E}\left[U | U \leq \mathcal{Q}_p(U)\right] = \max_{t \in \mathbb{R}} \left\{ t - \frac{1}{p}\mathbb{E}\left[(t-U)^+\right] \right\} \tag{7}$$

In equation 7, t represents the p-quantile of U. We also show that we can calculate t by a maximizing optimization function. The Sub-Quantile Optimization problem is posed as follows

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

For the linear regression case, this equation becomes

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

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). Theoretically, it has been proven to converge in research by Jin et al. (2019).

3.1 MOTIVATION

Assumption 1. *To provide theoretical bounds on the effectiveness of Sub-Quantile Minimization, we make the General Linear Model Assumption that*

$$\boldsymbol{y}_P = \boldsymbol{P}\boldsymbol{\beta}_P + \boldsymbol{\epsilon}_P \tag{12}$$

and similarly

$$\boldsymbol{y}_Q = \boldsymbol{Q}\boldsymbol{\beta}_Q + \boldsymbol{\epsilon}_Q \tag{13}$$

where $\boldsymbol{\beta}_P$ and $\boldsymbol{\beta}_Q$ the oracle regressors for $\mathbb{P}$ and $\mathbb{Q}$.

Since we are interested in learning the optimal model for distributions, our goal is to learn the parameters $\boldsymbol{\beta}_P$ from the distribution $\hat{\mathbb{P}}$. 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 2. *We assume the rows of $\boldsymbol{P}$ and $\boldsymbol{Q}$ are sampled from the same multivariate normal distribution.*

$$\boldsymbol{P}_i, \boldsymbol{Q}_j \sim \mathcal{N}_p(\boldsymbol{0}, \boldsymbol{\Sigma}) \tag{14}$$

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

Note $\boldsymbol{X} = \begin{pmatrix} \boldsymbol{P} \\ \boldsymbol{Q} \end{pmatrix}$ and $\boldsymbol{y} = \begin{pmatrix} \boldsymbol{y}_P \\ \boldsymbol{y}_Q \end{pmatrix}$ so $\boldsymbol{X}^\top = \begin{pmatrix} \boldsymbol{P}^\top & \boldsymbol{Q}^\top \end{pmatrix}$

Theorem 1. *The expected optimal parameters of the corrupted model* $\hat{\mathbb{P}}$

$$\mathbb{E}\left[\boldsymbol{X}^\dagger \boldsymbol{y}\right] = \boldsymbol{\beta}_P + \epsilon(\boldsymbol{\beta}_Q - \boldsymbol{\beta}_P) \tag{15}$$

The proof is reliant on assumption 2, this allows us to utilize the Wishart Distribution, $\mathcal{W}$, and the inverse Wishart Distribution, $\mathcal{W}^{-1}$. Please refer to Appendix B.1. By Theorem 1 we can see the level of corruption is dependent upon ϵ, which represents the percentage of corrupted samples, and the distance between the optimal parameters for $\mathbb{P}$, which is $\boldsymbol{\beta}_P$ and the optimal parameters for $\mathbb{Q}$, which is $\boldsymbol{\beta}_Q$.

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

Theorem 1 finds the optimal model when the corrupted distribution is sampled from the same distribution as the target distribution but has different optimal parameters. We will now look at the case of feature corruption. This is where the optimal parameters of the two distributions are the same but the data from $\mathbb{P}$ and $\mathbb{Q}$ are sampled differently.

Theorem 2. *In the case of* $\mathbb{P}$ *and* $\mathbb{Q}$ *being from different Normal Distributions. The expected optimal parameters of the corrupted model* $\hat{\mathbb{P}}$

$$\mathbb{E}\left[\boldsymbol{X}^\dagger \boldsymbol{y}\right] = \boldsymbol{\beta}_P - n(1-\epsilon)\boldsymbol{\Sigma}_P^{-1}\boldsymbol{\beta}_P \tag{16}$$

The proof can be found in Appendix B.2. We will show our results hold in Numerical Experiments. As seen in the results in table 1, the theory we provide is supported by Numerical Experimentation.

Dataset	$\epsilon = 0.2$		$\epsilon = 0.4$	
	$\mathbb{E}\left[\boldsymbol{X}^\dagger \boldsymbol{y}\right]$	Experimental	$\mathbb{E}\left[\boldsymbol{X}^\dagger \boldsymbol{y}\right]$	Experimental
`Quadratic Regression`	$(0.6\,, -0.6\,, 2.4)$	$\mathbf{0.777}_{(\mathbf{0.007})}$	$(0.2\,, -0.2\,, 2.8)$	$\mathbf{7.749}_{(\mathbf{0.009})}$
`Drug Discovery`	$0.895_{(0.009)}$	$0.775_{(0.006)}$	$8.944_{(0.007)}$	$7.742_{(0.006)}$

Table 1: Verification of Theorem 1 over `Quadratic Regression` Synthetic Dataset

In equation 15, note as $\epsilon \to 0$ we are returned $\boldsymbol{\beta}_P$. This is the intuition behind SubQuantile Minimization. By minimizing over the SubQuantile, we seek to reduce ϵ, and thus our model will return a model which is by expectation closer to $\boldsymbol{\beta}_P$.

4 THEORY

4.1 ANALYSIS OF $g(t, \theta)$

In this section, we will explore the fundamental aspects of $g(t, \boldsymbol{\theta})$. This will motivate the convergence analysis in the next section.

Lemma 4.1. *$g(t_{k+1}, \boldsymbol{\theta}_k)$ 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 4.2. *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 $\hat{\boldsymbol{\nu}} = sorted\left((\boldsymbol{\theta}^\top \boldsymbol{X} - \boldsymbol{y})^2\right)$ and note $0 < p < 1$

$$\partial(-g(t,\boldsymbol{\theta})) = -1 + \frac{1}{np}\sum_{i=1}^{n}\left\{\begin{array}{ll} 1, & \text{if } t > \hat{\boldsymbol{\nu}}_i \\ 0, & \text{if } t < \hat{\boldsymbol{\nu}}_i \\ [0,1]\,, & \text{if } t = \hat{\boldsymbol{\nu}}_i \end{array}\right\} \tag{17}$$

$$= 0 \text{ when } t = \hat{\boldsymbol{\nu}}_{np} \tag{18}$$

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

Lemma 4.3. *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.*

$$\nabla_{\boldsymbol{\theta}} g(t_{k+1}, \boldsymbol{\theta}_k) = \frac{1}{np}\sum_{i=1}^{np} 2\boldsymbol{x}_i(\boldsymbol{\theta}_k^\top \boldsymbol{x}_i - y_i) \tag{19}$$

We provide a proof in Appendix C.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.*

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

Proof. We see by lemma 4.2 and interpretation 1, we are optimizing by the np points with the lowest squared error. Mathematically,

$$g(t_{k+1}, \boldsymbol{\theta}_k) = t_{k+1} - \frac{1}{np}\sum_{i=1}^{n}\left(t_{k+1} - (\boldsymbol{\theta}^\top \boldsymbol{x}_i - y_i)^2\right)^+ \tag{20}$$

$$= t_{k+1} - \frac{1}{np}\sum_{i=1}^{np}\left(t_{k+1} - (\boldsymbol{\theta}^\top \boldsymbol{x}_i - y_i)^2\right)^+ \tag{21}$$

$$= t - t + \frac{1}{np}\sum_{i=1}^{np}(\boldsymbol{\theta}^\top \boldsymbol{x}_i - y_i)^2 \tag{22}$$

$$= \frac{1}{np}\sum_{i=1}^{np}(\boldsymbol{\theta}^\top \boldsymbol{x}_i - y_i)^2 \tag{23}$$

Now we can make a simple argument for convexity. We have a non-negative multiple of the sum of the composition of an affine function with a convex function. Thus $g(t, \boldsymbol{\theta})$ is convex with respect to $\boldsymbol{\theta}$. □

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

4.2 Optimization

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 $(t, \boldsymbol{\theta}) \in \mathbb{R} \times \mathbb{R}^d$*

$$g(t^*, \boldsymbol{\theta}) \leq g(t^*, \boldsymbol{\theta}^*) \leq g(t, \boldsymbol{\theta}^*) \tag{24}$$

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:*

$$g(t^*, \boldsymbol{\theta}) \leq g(t^*, \boldsymbol{\theta}^*) \leq g(t, \boldsymbol{\theta}^*) \tag{25}$$

Proposition 1. *As g is first-order differentiable, 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. By Proposition 1, this means both first-order partial derivatives are equal to 0. By lemma 4.2, we have shown $\nabla_t g(t, \boldsymbol{\theta}) =$

0 when $\boldsymbol{\nu}_{np} \leq t < \boldsymbol{\nu}_{np+1}$. Furthermore, by lemma 4.3, we have shown $\nabla_{\boldsymbol{\theta}}(g, \boldsymbol{\theta}) = 0$ when the least squares error is minimized for the np points with lowest squared error. This means that for a subset of np points from $\boldsymbol{X}$, the least squares error is minimized. What we are interested in is how many points within those np points come from $\mathbb{P}$ and how many of those points from $\mathbb{Q}$. Our goal is to minimize the number of points within the np lowest squared losses from Q, as they will introduce error to our predictions on points from P.

Lemma 4.6. *If $t_{k+1} \leq t_k$ then $g(t_{k+1}, \boldsymbol{\theta}_k) = g(t_k) + \frac{1}{np}\sum_{i=np}^{n}(t_k - \boldsymbol{\nu}_i)^+$. If $t_{k+1} > t_k$, then $g(t_{k+1}, \boldsymbol{\theta}_k) = g(t_k) + \frac{1}{np}\sum_{i=n(p-\delta)}^{np}(t - \boldsymbol{\nu}_i)^+ - \delta t$. For a small δ.*

Proof Sketch. When $t_{k+1} \leq t_k$ this result is quite intuitive, as we are simply removing the error of the elements outside elements within the lowest np squared losses. We delegate the rest of the proof to Appendix D.1 □

4.3 Reducing ϵ

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

$$\phi(\boldsymbol{\theta}) = \frac{1}{np}\sum_{i=1}^{m}(\boldsymbol{\theta}^\top \boldsymbol{p}_i - y_i)^2 \tag{26}$$

$$\psi(\boldsymbol{\theta}) = \frac{1}{np}\sum_{i=1}^{l}(\boldsymbol{\theta}^\top \boldsymbol{q}_i - y_i)^2 \tag{27}$$

These two functions hold nice properties.

$$\nabla_{\boldsymbol{\theta}}\phi(\boldsymbol{\theta}) = \frac{1}{np}\sum_{i=1}^{m} 2\boldsymbol{p}_i(\boldsymbol{\theta}^\top \boldsymbol{p}_i - y_i) \tag{28}$$

$$\nabla_{\boldsymbol{\theta}}\psi(\boldsymbol{\theta}) = \frac{1}{np}\sum_{i=1}^{l} 2\boldsymbol{q}_i(\boldsymbol{\theta}^\top \boldsymbol{q}_i - y_i) \tag{29}$$

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

$$\nabla_{\boldsymbol{\theta}_k} g(t_{k+1}, \boldsymbol{\theta}_k) = \nabla_{\boldsymbol{\theta}_k}\phi(\boldsymbol{\theta}) + \nabla_{\boldsymbol{\theta}_k}\psi(\boldsymbol{\theta}) \tag{30}$$

Assumption 3. *Training on a subset of the from $\mathbb{P}$ or from $\mathbb{Q}$ generalizes well to the total data.*

From Assumption 3, we can make the assumption that if we take an optimization step with respect to the data in $\mathbb{P}$ or $\mathbb{Q}$ within the lowest np squared errors, it will generalize well to the data from the $n(1-p)$ highest squared errors.

Theorem 3. *In each iteration of the two-step optimization, by expectation the number of elements from $\mathbb{P}$ will increase.*

Corollary 3.1. *If the $\epsilon > 0.5$, i.e., the corruption is the majority class, Sub-Quantile Optimization is still able to able to converge to the optimal class when p is chosen to be less than 0.5.*

Corrolary 3.1 seems like a very unintuitive result, as the whole idea of subquantile minimization is to optimize over the majority class and leave out the tail of the distribution. We provide a proof in

5 Empirical Results

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

$$t_{k+1} = \hat{\boldsymbol{\nu}}_{np} \tag{31}$$

$$t_{k+1} = \frac{1}{np}\sum_{i=1}^{np}\hat{\boldsymbol{\nu}}_i \tag{32}$$

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 31. We also compare against the t-update as described in equation 32.

Algorithm 1: Sub-Quantile Minimization Optimization Algorithm

Input: Training iterations T, Quantile p, Corruption Percentage ϵ, 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**
  $\boldsymbol{\nu} = (\boldsymbol{X}\boldsymbol{\theta}_k - \boldsymbol{y})^2$
  $\hat{\boldsymbol{\nu}} = sorted(\boldsymbol{\nu})$
  $t_{k+1} = \hat{\boldsymbol{\nu}}_{np}$
  $t_{k+1} = \frac{1}{np}\sum_{i=1}^{np} \boldsymbol{\nu}_i$
  $L \coloneqq \sum_{i=1}^{np} \boldsymbol{x}_i^\top \boldsymbol{x}_i$
  $\alpha \coloneqq \frac{1}{2L}$
  $\boldsymbol{\theta}_{k+1} = \boldsymbol{\theta}_k - \alpha \nabla_{\boldsymbol{\theta}_k} g(t_{k+1}, \boldsymbol{\theta}_k)$
end
return $\boldsymbol{\theta}_T$

We also present a batch algorithm which improves training speed significantly. In accordance with Minibatch theory, if the subset I of all data is representative of all the data, then this will have similar results to Algorithm 1.

Algorithm 2: Stochastic Sub-Quantile Minimization Optimization Algorithm

Input: Training iterations T, Quantile p, Corruption Percentage ϵ, 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
  $\boldsymbol{\nu} = (\boldsymbol{X}_I\boldsymbol{\theta}_k - \boldsymbol{y}_I)^2$
  $\hat{\boldsymbol{\nu}} = sorted(\boldsymbol{\nu})$
  $t_{k+1} = \hat{\boldsymbol{\nu}}_{mp}$
  $t_{k+1} = \frac{1}{mp}\sum_{i=1}^{mp} \boldsymbol{\nu}_i$
  $L \coloneqq \sum_{i=1}^{mp} \boldsymbol{x}_i^\top \boldsymbol{x}_i$
  $\alpha \coloneqq \frac{1}{2L}$
  $\boldsymbol{\theta}_{k+1} = \boldsymbol{\theta}_k - \alpha \nabla_{\boldsymbol{\theta}_k} g(t_{k+1}, \boldsymbol{\theta}_k)$
end
return $\boldsymbol{\theta}_T$

5.1 Synthetic Data

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 1. Our results show the near optimal performance of Sub-Quantile Minimization. The results and comparison with other methods can be seen in Table **??**. Note we are not interested in $\epsilon \geq 0.5$ 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 F 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) utilizing the noise procedure described in Li et al. (2020).

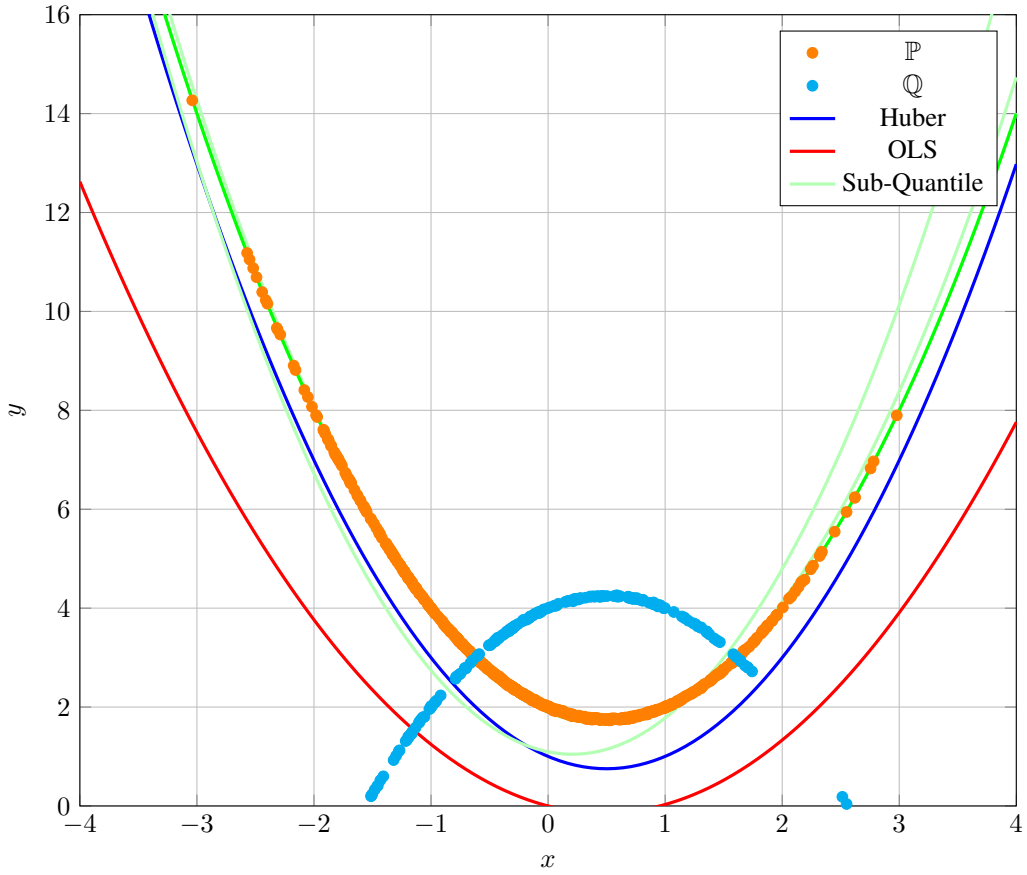

Figure 1: `Quadratic Regression` $n = 1000$ and $\epsilon = 0.2$

Objectives	Test RMSE (`Quadratic Regression`)		
	$\epsilon = 0$	$\epsilon = 0.2$	$\epsilon = 0.4$
OLS 119	$0.0099_{(0.0002)}$	$2.078_{(0.146)}$	$4.104_{(0.442)}$
Huber Huber & Ronchetti (2009)	$1.000_{(0.0002)}$	$1.000_{(0.0003)}$	$1.13_{(0.087)}$
RANSAC Fischler & Bolles (1981)	$0.010_{(0.0002)}$	$0.011_{(0.0002)}$	$0.061_{(0.053)}$
TERM Li et al. (2020)	$0.010_{(0.0001)}$	$0.012_{(0.0008)}$	$0.017_{(0.0016)}$
SEVER Diakonikolas et al. (2019)	$0.0166_{(0.007)}$	$0.011_{(0.0004)}$	$0.0267_{(0.036)}$
SubQuantile($p = 0.6$)	$\mathbf{0.0099_{(0.0002)}}$	$\mathbf{0.00998_{(0.0002)}}$	$\mathbf{0.010_{(0.0001)}}$
Genie ERM	$0.0099_{(0.0002)}$	$0.00997_{(0.0002)}$	$0.010_{(0.0001)}$

Table 2: `Qudatic Regression` Synthetic Dataset. Empirical Risk over $\mathbb{P}$

As we can see in Table 3, we obtain state of the art results in the lower range of range of noise, and futher more, we obtain results on par with the current state of the art. This makes our model the strongest among the tested, due to our strength throughout the whole range of noises.

6 Conclusion

In this work we provide a theoretical analysis for robust linear regression by minimizing the *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.

Objectives	Test RMSE (`Drug Discovery`)			
	$\epsilon = 0$	$\epsilon = 0.2$	$\epsilon = 0.4$	$\epsilon = 0.8$
OLS 119	$0.990_{(0.060)}$	$1.969_{(0.118)}$	$2.829_{(0.086)}$	$4.682_{(0.101)}$
Huber Huber & Ronchetti (2009)	$1.326_{(0.096)}$	$1.628_{(0.253)}$	$2.023_{(0.498)}$	$3.442_{(0.581)}$
RANSAC Fischler & Bolles (1981)	∞	∞	∞	∞
TERM Li et al. (2020)	$1.313_{(0.072)}$	$1.334_{(0.105)}$	$1.343_{(0.0740)}$	$\mathbf{1.428}_{(\mathbf{0.107})}$
SEVER Diakonikolas et al. (2019)	$1.079_{(0.059)}$	$1.076_{(0.048)}$	$\mathbf{1.067}_{(\mathbf{0.091})}$	$3.993_{(0.203)}$
SubQuantile($p = 1 - \epsilon$)	$\mathbf{1.052}_{(\mathbf{0.062})}$	$\mathbf{1.060}_{(\mathbf{0.065})}$	$1.073_{(0.101)}$	$1.479_{(0.0695)}$
Genie ERM	$0.990_{(0.060)}$	$1.038_{(0.041)}$	$1.037_{(0.086)}$	∞

Table 3: `Drug Discovery` Dataset. Empirical Risk over $\mathbb{P}$

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.

Martin A. Fischler and Robert C. Bolles. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Commun. ACM*, 24(6):381–395, jun 1981. ISSN 0001-0782. doi: 10.1145/358669.358692. URL https://doi.org/10.1145/358669.358692.

Gene H. Golub and Charles F. Van Loan. *Matrix Computations (3rd Ed.)*. Johns Hopkins University Press, USA, 1996. ISBN 0801854148.

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.

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 PROOFS IN SECTION 3

B.1 PROOF OF THEOREM 1

Proof.
We will first calculate the pseudo-inverse

$$\boldsymbol{X}^\top \boldsymbol{X} = \begin{pmatrix} \boldsymbol{P}^\top & \boldsymbol{Q}^\top \end{pmatrix} \begin{pmatrix} \boldsymbol{P} \\ \boldsymbol{Q} \end{pmatrix} \tag{33}$$

$$= \boldsymbol{P}^\top \boldsymbol{P} + \boldsymbol{Q}^\top \boldsymbol{Q} \tag{34}$$

Now we can calculate the Moore-Penrose Inverse

$$(\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top = (\boldsymbol{P}^\top \boldsymbol{P} + \boldsymbol{Q}^\top \boldsymbol{Q})^{-1} \begin{pmatrix} \boldsymbol{P}^\top & \boldsymbol{Q}^\top \end{pmatrix} \tag{35}$$

$$= \begin{pmatrix} (\boldsymbol{P}^\top \boldsymbol{P} + \boldsymbol{Q}^\top \boldsymbol{Q})^{-1} \boldsymbol{P}^\top & (\boldsymbol{P}^\top \boldsymbol{P} + \boldsymbol{Q}^\top \boldsymbol{Q})^{-1} \boldsymbol{Q}^\top \end{pmatrix} \tag{36}$$

Now we solve for the optimal model

$$\boldsymbol{X}^\dagger \boldsymbol{y} = \begin{pmatrix} (\boldsymbol{P}^\top \boldsymbol{P} + \boldsymbol{Q}^\top \boldsymbol{Q})^{-1} \boldsymbol{P}^\top & (\boldsymbol{P}^\top \boldsymbol{P} + \boldsymbol{Q}^\top \boldsymbol{Q})^{-1} \boldsymbol{Q}^\top \end{pmatrix} \begin{pmatrix} \boldsymbol{y}_P \\ \boldsymbol{y}_Q \end{pmatrix} \tag{37}$$

$$= (\boldsymbol{P}^\top \boldsymbol{P} + \boldsymbol{Q}^\top \boldsymbol{Q})^{-1} \boldsymbol{P}^\top \boldsymbol{y}_P + (\boldsymbol{P}^\top \boldsymbol{P} + \boldsymbol{Q}^\top \boldsymbol{Q})^{-1} \boldsymbol{Q}^\top \boldsymbol{y}_Q \tag{38}$$

By assumption 2, 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).

$$\boldsymbol{P}^\top \boldsymbol{P} = \sum_{i=1}^{n*(1-\epsilon)} \boldsymbol{P}_i \boldsymbol{P}_i^\top \tag{39}$$

$$\boldsymbol{Q}^\top \boldsymbol{Q} = \sum_{j=1}^{n\epsilon} \boldsymbol{Q}_j \boldsymbol{Q}_j^\top \tag{40}$$

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

$$\boldsymbol{P}^\top \boldsymbol{P} \sim \mathcal{W}(n(1-\epsilon), \boldsymbol{\Sigma}) \tag{41}$$

$$\boldsymbol{Q}^\top \boldsymbol{Q} \sim \mathcal{W}(n\epsilon, \boldsymbol{\Sigma}) \tag{42}$$

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

$$\mathbb{E}(\boldsymbol{P}^\top \boldsymbol{P}) = n(1-\epsilon)\boldsymbol{\Sigma} \tag{43}$$

$$\mathbb{E}(\boldsymbol{Q}^\top \boldsymbol{Q}) = n\epsilon\boldsymbol{\Sigma} \tag{44}$$

It thus follows

$$\mathbb{E}\left[\boldsymbol{P}^\top \boldsymbol{P} + \boldsymbol{Q}^\top \boldsymbol{Q}\right] = n\boldsymbol{\Sigma} \tag{45}$$

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

$$\left(\boldsymbol{P}^\top \boldsymbol{P} + \boldsymbol{Q}^\top \boldsymbol{Q}\right)^{-1} \sim \mathcal{W}^{-1}(n, \boldsymbol{\Sigma}) \tag{46}$$

It thus follows by the expectation of the Inverse Wishart Distribution

$$\mathbb{E}\left[\left(\boldsymbol{P}^\top \boldsymbol{P} + \boldsymbol{Q}^\top \boldsymbol{Q}\right)^{-1}\right] = n\boldsymbol{\Sigma}^{-1} \tag{47}$$

Now we will plug this into Equation 38:

$$\mathbb{E}\left[\boldsymbol{X}^\dagger \boldsymbol{y}\right] = \left(n\boldsymbol{\Sigma}^{-1}\right) \boldsymbol{P}^\top \boldsymbol{y}_P + \left(n\boldsymbol{\Sigma}^{-1}\right) \boldsymbol{Q}^\top \boldsymbol{y}_Q \tag{48}$$

$$= \left(n\boldsymbol{\Sigma}^{-1}\right) \boldsymbol{P}^\top (\boldsymbol{P}\boldsymbol{\beta} + \boldsymbol{\epsilon}_P) + \left(n\boldsymbol{\Sigma}^{-1}\right) \boldsymbol{Q}^\top (\boldsymbol{Q}\boldsymbol{\beta}_Q^\top + \boldsymbol{\epsilon}_Q) \tag{49}$$

$$= \left(n\boldsymbol{\Sigma}^{-1}\right) \left((\boldsymbol{P}^\top \boldsymbol{P})\boldsymbol{\beta}_P + (\boldsymbol{Q}^\top \boldsymbol{Q}) (\boldsymbol{\beta}_P + (\boldsymbol{\beta}_Q - \boldsymbol{\beta}_P))\right) \tag{50}$$

$$= \left(n\boldsymbol{\Sigma}^{-1}\right) \left((n(1-\epsilon)\boldsymbol{\Sigma})\boldsymbol{\beta}_P + n\epsilon\boldsymbol{\Sigma} (\boldsymbol{\beta}_P + \boldsymbol{\Psi})\right) \tag{51}$$

$$= \left(n\boldsymbol{\Sigma}^{-1}\right) (n\boldsymbol{\Sigma}\boldsymbol{\beta}_P + n\epsilon\boldsymbol{\Sigma}\boldsymbol{\Psi}) \tag{52}$$

$$= \boldsymbol{\beta}_P + \epsilon(\boldsymbol{\Psi}) \tag{53}$$

This concludes the proof. □

B.2 PROOF OF THEOREM 2

Proof. The first half of the proof follows from Appendix B.1. We start by noting new notation. $\mathbf{\Sigma}_P$ represents the covariance matrix for $\mathbb{P}$ and $\mathbf{\Sigma}_Q$ represents the covariance matrix for $\mathbb{Q}$.

$$\mathbb{E}\left[\boldsymbol{P}^\top \boldsymbol{P}\right] = n(1-\epsilon)\mathbf{\Sigma}_P \tag{54}$$

$$\mathbb{E}\left[\boldsymbol{Q}^\top \boldsymbol{Q}\right] = n\epsilon\mathbf{\Sigma}_Q \tag{55}$$

It thus follows

$$\mathbb{E}\left[\boldsymbol{P}^\top \boldsymbol{P} + \boldsymbol{Q}^\top \boldsymbol{Q}\right] = (n(1-\epsilon)\mathbf{\Sigma}_P + n\epsilon\mathbf{\Sigma}_Q) \tag{56}$$

This is where the structure of the proof because we can no longer follow the Inverse Wishart Distribution.

$$\mathbb{E}\left[\left(\boldsymbol{P}^\top \boldsymbol{P} + \boldsymbol{Q}^\top \boldsymbol{Q}\right)^{-1}\right] = (n(1-\epsilon)\mathbf{\Sigma}_P + n\epsilon\mathbf{\Sigma}_Q)^{-1} \tag{57}$$

Now we can use the Woodbury Formula Golub & Van Loan (1996)

$$= n(1-\epsilon)\mathbf{\Sigma}_P^{-1} - n(1-\epsilon)\mathbf{\Sigma}_P^{-1}(n\epsilon\mathbf{\Sigma}_Q^{-1}) \tag{58}$$

We will now calculate the expected optimal parameters by plugging this into Equation 38:

$$\mathbb{E}\left[\boldsymbol{X}^\dagger \boldsymbol{y}\right] = n(1-\epsilon)\mathbf{\Sigma}_P^{-1}(\boldsymbol{P}^T\boldsymbol{P})\boldsymbol{\beta}_P - n(1-\epsilon)\mathbf{\Sigma}_P^{-1}(n\epsilon\mathbf{\Sigma}_Q^{-1})(\boldsymbol{Q}^T\boldsymbol{Q})\boldsymbol{\beta}_P \tag{59}$$

$$= n(1-\epsilon)\mathbf{\Sigma}_P^{-1}(n(1-\epsilon)\mathbf{\Sigma}_P)\boldsymbol{\beta}_P - n(1-\epsilon)\mathbf{\Sigma}_P^{-1}(n\epsilon\mathbf{\Sigma}_Q^{-1})(n\epsilon\mathbf{\Sigma}_Q)\boldsymbol{\beta}_P \tag{60}$$

$$= \boldsymbol{\beta}_P - n(1-\epsilon)\mathbf{\Sigma}_P^{-1}\boldsymbol{\beta}_P \tag{61}$$

This concludes the proof. □

C General Properties of Sub-Quantile Minimization

C.1 Proof of Lemma 4.2

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 $\hat{\boldsymbol{\nu}} = sorted\left((\boldsymbol{\theta}^\top \boldsymbol{X} - \boldsymbol{y})^2\right)$ and note $0 < p < 1$

$$\partial(-g(t, \boldsymbol{\theta})) = \partial\left(-t + \frac{1}{np}\sum_{i=1}^{n}(t - \hat{\boldsymbol{\nu}}_i)^+\right) \tag{62}$$

$$= \partial(-t) + \partial\left(\frac{1}{np}\sum_{i=1}^{n}(t - \hat{\boldsymbol{\nu}}_i)^+\right) \tag{63}$$

$$= -1 + \frac{1}{np}\sum_{i=1}^{n}\partial(t - \hat{\boldsymbol{\nu}}_i)^+ \tag{64}$$

$$= -1 + \frac{1}{np}\sum_{i=1}^{n}\left\{\begin{array}{ll} 1, & \text{if } t > \hat{\boldsymbol{\nu}}_i \\ 0, & \text{if } t < \hat{\boldsymbol{\nu}}_i \\ [0, 1], & \text{if } t = \hat{\boldsymbol{\nu}}_i \end{array}\right\} \tag{65}$$

$$= 0 \text{ when } t = \hat{\boldsymbol{\nu}}_{np} \tag{66}$$

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

C.2 Proof of Lemma 4.3

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

$$\nabla_{\boldsymbol{\theta}_k} g(t_{k+1}, \boldsymbol{\theta}_k) = \nabla_{\boldsymbol{\theta}_k}\left(\boldsymbol{\nu}_{np} - \frac{1}{np}\sum_{i=1}^{n}(\boldsymbol{\nu}_{np} - (\boldsymbol{\theta}_k^\top \boldsymbol{x}_i - y_i)^2)^+\right) \tag{67}$$

$$= \nabla_{\boldsymbol{\theta}_k}\left((\boldsymbol{\theta}_k^\top \boldsymbol{x}_{np} - y_{np})^2 - \frac{1}{np}\sum_{i=1}^{n}\left((\boldsymbol{\theta}_k^\top \boldsymbol{x}_{np} - y_{np})^2 - (\boldsymbol{\theta}_k^\top \boldsymbol{x}_i - y_i)^2\right)^+\right) \tag{68}$$

$$= \nabla_{\boldsymbol{\theta}_k}(\boldsymbol{\theta}_k^\top \boldsymbol{x}_{np} - y_{np})^2 - \frac{1}{np}\sum_{i=1}^{n}\nabla_{\boldsymbol{\theta}_k}\left((\boldsymbol{\theta}_k^\top \boldsymbol{x}_{np} - y_{np})^2 - (\boldsymbol{\theta}_k^\top \boldsymbol{x}_i - y_i)^2\right)^+ \tag{69}$$

$$= 2\boldsymbol{x}_{np}(\boldsymbol{\theta}_k^\top \boldsymbol{x}_{np} - y_{np}) - \frac{1}{np}\sum_{i=1}^{n}2\boldsymbol{x}_{np}(\boldsymbol{\theta}_k^\top \boldsymbol{x}_{np} - y_{np}) - 2\boldsymbol{x}_i(\boldsymbol{\theta}_k^\top \boldsymbol{x}_i - y_i)\left\{\begin{array}{ll} 1, & \text{if } t > v_i \\ 0, & \text{if } t < v_i \\ [0, 1], & \text{if } t = v_i \end{array}\right\} \tag{70}$$

$$= 2\boldsymbol{x}_{np}(\boldsymbol{\theta}_k^\top \boldsymbol{x}_{np} - y_{np}) - \frac{1}{np}\sum_{i=1}^{np}2\boldsymbol{x}_{np}(\boldsymbol{\theta}_k^\top \boldsymbol{x}_{np} - y_{np}) - 2\boldsymbol{x}_i(\boldsymbol{\theta}_k^\top \boldsymbol{x}_i - y_i) \tag{71}$$

$$= 2\boldsymbol{x}_{np}(\boldsymbol{\theta}_k^\top \boldsymbol{x}_{np} - y_{np}) - 2\boldsymbol{x}_{np}(\boldsymbol{\theta}_k^\top \boldsymbol{x}_{np} - y_{np}) + \frac{1}{np}\sum_{i=1}^{np}2\boldsymbol{x}_i(\boldsymbol{\theta}_k^\top \boldsymbol{x}_i - y_i) \tag{72}$$

$$= \frac{1}{np}\sum_{i=1}^{np}2\boldsymbol{x}_i(\boldsymbol{\theta}_k^\top \boldsymbol{x}_i - y_i) \tag{73}$$

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

C.3 PROOF OF LEMMA 4.5

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

$$||\nabla_{\boldsymbol{\theta}} g(\boldsymbol{\theta}', t) - \nabla_{\boldsymbol{\theta}} g(\boldsymbol{\theta}, t)|| \leq L||\boldsymbol{\theta}' - \boldsymbol{\theta}|| \tag{74}$$

$$\left\|\nabla_{\boldsymbol{\theta}} g(\boldsymbol{\theta}^{'}, t) - \nabla_{\boldsymbol{\theta}} g(\boldsymbol{\theta}, t)\right\| = \left\| \frac{1}{np} \sum_{i=1}^{np} 2\boldsymbol{x}_i(\boldsymbol{\theta}_k^{'T}\boldsymbol{x}_i - y_i) - \frac{1}{np} \sum_{i=1}^{np} 2\boldsymbol{x}_i(\boldsymbol{\theta}_k^{\top}\boldsymbol{x}_i - y_i) \right\| \tag{75}$$

$$= \left\| \frac{1}{np} \sum_{i=1}^{np} 2\boldsymbol{x}_i(\boldsymbol{\theta}_k^{'T}\boldsymbol{x}_i - \boldsymbol{\theta}_k^{\top}\boldsymbol{x}_i) \right\| \tag{76}$$

$$= \left\| \frac{1}{np} \sum_{i=1}^{np} 2\boldsymbol{x}_i^{\top}\boldsymbol{x}_i(\boldsymbol{\theta}_k^{'T} - \boldsymbol{\theta}_k^{\top}) \right\| \tag{77}$$

$$\overset{\text{Cauchy}-\text{Schwarz}}{\leq} \left\| \frac{2}{np} \sum_{i=1}^{np} \|\boldsymbol{x}_i\|^2 \right\| \left\| \boldsymbol{\theta}_k^{'T} - \boldsymbol{\theta}_k^{\top} \right\| \tag{78}$$

$$= L \left\| \boldsymbol{\theta}_k^{'T} - \boldsymbol{\theta}_k^{\top} \right\| \tag{79}$$

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

This concludes the proof.

D PROOFS FOR CONVERGENCE

D.1 PROOF OF LEMMA 4.6

Proof. We will investigate the two cases $t_{k+1} \leq t$ and $t_{k+1} > t_k$.
Case (i) $t_{k+1} \leq t_k$
Let us first expand out $g(t_k, \boldsymbol{\theta}_k)$ with the knowledge that $t_k \geq \hat{\boldsymbol{\nu}_k}$

$$g(t_k, \boldsymbol{\theta}_k) = t_k - \frac{1}{np}\sum_{i=1}^{n}(t_k - \boldsymbol{\nu}_i)^+ \tag{80}$$

$$= t_k - \frac{1}{np}(np)t_k + \frac{1}{np}\sum_{i=1}^{np}\boldsymbol{\nu}_i + \frac{1}{np}\sum_{i=np}^{n}(t_k - \boldsymbol{\nu}_i)^+ \tag{81}$$

$$= \frac{1}{np}\sum_{i=1}^{np}\boldsymbol{\nu}_i + \frac{1}{np}\sum_{i=np}^{n}(t_k - \boldsymbol{\nu}_i)^+ \tag{82}$$

$$g(t_{k+1}, \boldsymbol{\theta}_k) - g(t_k, \boldsymbol{\theta}_k) = \frac{1}{np}\sum_{i=1}^{np}\boldsymbol{\nu}_i - \left(\frac{1}{np}\sum_{i=1}^{np}\boldsymbol{\nu}_i + \frac{1}{np}\sum_{i=np}^{n}(t_k - \boldsymbol{\nu}_i)^+\right) \tag{83}$$

$$= -\frac{1}{np}\sum_{i=np}^{n}(t_k - \boldsymbol{\nu}_i)^+ \tag{84}$$

Case (ii) $t_{k+1} > t_k$
Since we know t_k is less than $\boldsymbol{\nu}_{np}$, WLOG we will say t_k is greater than the lowest $n(p-\delta)$ elements, where $\delta \in (0, p)$.

$$g(t_k, \boldsymbol{\theta}_k) = t_k - \frac{1}{np}\sum_{i=1}^{n}(t_k - \boldsymbol{\nu}_i)^+ \tag{85}$$

$$= t_k - \frac{1}{np}\sum_{i=1}^{n(p-\delta)}(t_k - \boldsymbol{\nu}_i)^+ \tag{86}$$

$$= t_k - \frac{1}{np}(n(p-\delta))t_k + \frac{1}{np}\sum_{i=1}^{n(p-\delta)}\boldsymbol{\nu}_i \tag{87}$$

$$g(t_k, \boldsymbol{\theta}_{k+1}) - g(t_k, \boldsymbol{\theta}_k) = \frac{1}{np}\sum_{i=1}^{np}\boldsymbol{\nu}_i - \left(\delta t_k + \frac{1}{np}\sum_{i=1}^{n(p-\delta)}\boldsymbol{\nu}_i\right) \tag{88}$$

$$= \left(\frac{1}{np}\sum_{i=n(p-\delta)}^{n}\boldsymbol{\nu}_i\right) - \delta t_k \tag{89}$$

□

E PROOFS FOR REDUCING ε

E.1 PROOF OF THEOREM 3

Proof. Note Assumption 2 tells us the whole distribution is sampled from the same distribution.

$$\mathbb{E}\left[\nabla_{\boldsymbol{\theta}}\phi(\boldsymbol{\theta})\right] = \frac{1}{np}\sum_{i=1}^{np(1-\epsilon)} 2\mathbb{E}\left[\boldsymbol{p}_i\right]\mathbb{E}\left[\boldsymbol{\theta}^\top \boldsymbol{p}_i - y_i\right] \tag{90}$$

$$= \frac{1}{np}\sum_{i=1}^{np(1-\epsilon)} 2\mathbb{E}\left[\boldsymbol{p}_i\right]\mathbb{E}\left[\left(\boldsymbol{\theta}^\top - \boldsymbol{\beta}_P^\top\right)\boldsymbol{p}_i\right] \tag{91}$$

$$= \frac{2}{np}\sum_{i=1}^{np(1-\epsilon)} \boldsymbol{\mu}(\boldsymbol{\theta}^\top - \boldsymbol{\beta}_P^\top)\boldsymbol{\mu} \tag{92}$$

$$= 2(1-\epsilon)\boldsymbol{\mu}\boldsymbol{\mu}^\top(\boldsymbol{\theta} - \boldsymbol{\beta}_P) \tag{93}$$

We will use similar logic to find the Expectation of the derivative w.r.t $\mathbb{Q}$

$$\mathbb{E}\left[\nabla_{\boldsymbol{\theta}}\psi(\boldsymbol{\theta})\right] = 2\epsilon\boldsymbol{\mu}\boldsymbol{\mu}^\top(\boldsymbol{\theta} - \boldsymbol{\beta}_Q) \tag{94}$$

Thus we can calculate the expected derivative

$$\mathbb{E}\left[\nabla_{\boldsymbol{\theta}} g(t_{k+1}, \boldsymbol{\theta}_k)\right] = \mathbb{E}\left[\nabla_{\boldsymbol{\theta}}\phi(\boldsymbol{\theta}_k)\right] + \mathbb{E}\left[\nabla_{\boldsymbol{\theta}}\psi(\boldsymbol{\theta}_k)\right] \tag{95}$$

$$= 2(1-\epsilon)\boldsymbol{\mu}\boldsymbol{\mu}^\top(\boldsymbol{\theta} - \boldsymbol{\beta}_P) + 2\epsilon\boldsymbol{\mu}\boldsymbol{\mu}^\top(\boldsymbol{\theta}_k - \boldsymbol{\beta}_Q) \tag{96}$$

$$= 2\boldsymbol{\mu}\boldsymbol{\mu}^\top\left(\boldsymbol{\theta}_k - (1-\epsilon)\boldsymbol{\beta}_P - \epsilon\boldsymbol{\beta}_Q\right) \tag{97}$$

Definition 6. *A function f is L smooth if*

$$\|f(\boldsymbol{x}) - f(\boldsymbol{y})\| \leq L\|x - y\| \;\forall \boldsymbol{x}, \boldsymbol{y}$$

We will now find the expected change in loss with respect to the distributions of $\mathbb{P}$ and $\mathbb{Q}$ in each iteration. Furthermore, let us note $g(t, \boldsymbol{\theta})$ is L-smooth with respect to $\boldsymbol{\theta}$. Thus we can use the properties of Definition 6.

$$\mathbb{E}\left[g(t_{k+1}, \boldsymbol{\theta}_{k+1})\right] \leq \mathbb{E}\left[g(t, \boldsymbol{\theta}_k)\right] + \left\langle \nabla_{\boldsymbol{\theta}}\mathbb{E}\left[g(t_k, \boldsymbol{\theta}_k)\right], \boldsymbol{\theta}_{k+1} - \boldsymbol{\theta}_k\right\rangle + \frac{L}{2}\|\boldsymbol{\theta}_{k+1} - \boldsymbol{\theta}_k\|_2^2 \tag{98}$$

$$= \mathbb{E}\left[g(t, \boldsymbol{\theta}_k)\right] + \left\langle \nabla_{\boldsymbol{\theta}}\phi(\boldsymbol{\theta}) + \nabla_{\boldsymbol{\theta}}\psi(\boldsymbol{\theta}), \boldsymbol{\theta}_{k+1} - \boldsymbol{\theta}_k\right\rangle + \frac{L}{2}\|\boldsymbol{\theta}_{k+1} - \boldsymbol{\theta}_k\|_2^2 \tag{99}$$

We can use the $\boldsymbol{\theta}$ update rule described in Equation 11

$$= \mathbb{E}\left[g(t, \boldsymbol{\theta}_k)\right] + \left\langle \nabla_{\boldsymbol{\theta}}\phi(\boldsymbol{\theta}) + \nabla_{\boldsymbol{\theta}}\psi(\boldsymbol{\theta}), -\frac{1}{L}\left(\nabla_{\boldsymbol{\theta}}\phi(\boldsymbol{\theta}_k) + \nabla_{\boldsymbol{\theta}}\psi(\boldsymbol{\theta}_k)\right)\right\rangle + \frac{L}{2}\|\phi(\boldsymbol{\theta}) + \nabla_{\boldsymbol{\theta}}\psi(\boldsymbol{\theta})\|_2^2 \tag{100}$$

We can now use the expected derivative described in Equation 97

$$= \mathbb{E}\left[g(t, \boldsymbol{\theta}_k)\right] + \left\langle 2\boldsymbol{\mu}\boldsymbol{\mu}^\top\left(\boldsymbol{\theta}^\top - (1-\epsilon)\boldsymbol{\beta}_P^\top - \epsilon\boldsymbol{\beta}_Q^\top\right), -\frac{1}{L}2\boldsymbol{\mu}\boldsymbol{\mu}^\top\left(\boldsymbol{\theta}^\top - (1-\epsilon)\boldsymbol{\beta}_P^\top - \epsilon\boldsymbol{\beta}_Q^\top\right)\right\rangle + \frac{1}{2L}\left(2\boldsymbol{\mu}\boldsymbol{\mu}^\top\left(\boldsymbol{\theta}^\top - (1-\epsilon)\boldsymbol{\beta}_P^\top - \epsilon\boldsymbol{\beta}_Q^\top\right)\right)^\top\left(2\boldsymbol{\mu}\boldsymbol{\mu}^\top\left(\boldsymbol{\theta}^\top - (1-\epsilon)\boldsymbol{\beta}_P^\top - \epsilon\boldsymbol{\beta}_Q^\top\right)\right) \tag{101}$$

$$= \mathbb{E}\left[g(t, \boldsymbol{\theta}_k)\right] - \frac{1}{2L}\left(2\boldsymbol{\mu}\boldsymbol{\mu}^\top\left(\boldsymbol{\theta}^\top - (1-\epsilon)\boldsymbol{\beta}_P^\top - \epsilon\boldsymbol{\beta}_Q^\top\right)\right)^\top\left(2\boldsymbol{\mu}\boldsymbol{\mu}^\top\left(\boldsymbol{\theta}^\top - (1-\epsilon)\boldsymbol{\beta}_P^\top - \epsilon\boldsymbol{\beta}_Q^\top\right)\right) \tag{102}$$

$$\tag{103}$$

Let us now calculate the expected change in $\boldsymbol{\theta}$ in each iteration. We will start with the $\boldsymbol{\theta}$-update as described in Equation 11.

$$\mathbb{E}\left[\boldsymbol{\theta}_{k+1}\right] = \boldsymbol{\theta}_k - \frac{1}{2L}\mathbb{E}\left[\nabla_{\boldsymbol{\theta}} g(t_{k+1}, \boldsymbol{\theta}_k)\right] \tag{104}$$

We can now use the expected derivative in equation 97

$$= \boldsymbol{\theta}_k - \frac{1}{2L}\left(2\boldsymbol{\mu}\boldsymbol{\mu}^\top\left(\boldsymbol{\theta}_k - (1-\varepsilon)\boldsymbol{\beta}_P - \varepsilon\boldsymbol{\beta}_Q\right)\right) \tag{105}$$

$$= \boldsymbol{\theta}_k - \frac{1}{L}\left(\boldsymbol{\mu}\boldsymbol{\mu}^\top\left(\boldsymbol{\theta}_k - (1-\varepsilon)\boldsymbol{\beta}_P - \varepsilon\boldsymbol{\beta}_Q\right)\right) \tag{106}$$

$$= \left(\boldsymbol{I} - \frac{1}{L}\boldsymbol{\mu}\boldsymbol{\mu}^\top\right)\boldsymbol{\theta}_k + \frac{1-\varepsilon}{L}\boldsymbol{\mu}\boldsymbol{\mu}^\top\boldsymbol{\beta}_P + \frac{\varepsilon}{L}\boldsymbol{\mu}\boldsymbol{\mu}^\top\boldsymbol{\beta}_Q \tag{107}$$

Let us now calculate $\mathbb{E}\left[\boldsymbol{\theta}_{k+1}\boldsymbol{\mu}\right]$

$$\mathbb{E}\left[\boldsymbol{\theta}_{k+1}^\top\boldsymbol{\mu}\right] = \left(\left(\boldsymbol{I} - \frac{1}{L}\boldsymbol{\mu}\boldsymbol{\mu}^\top\right)\boldsymbol{\theta}_k + \frac{1-\varepsilon}{L}\boldsymbol{\mu}\boldsymbol{\mu}^\top\boldsymbol{\beta}_P + \frac{\varepsilon}{L}\boldsymbol{\mu}\boldsymbol{\mu}^\top\boldsymbol{\beta}_Q\right)^\top\boldsymbol{\mu} \tag{108}$$

$$= \left(\boldsymbol{\theta}_k^\top\left(\boldsymbol{I} - \frac{1}{L}\boldsymbol{\mu}\boldsymbol{\mu}^\top\right) + \frac{1-\varepsilon}{L}\boldsymbol{\beta}_P^T\boldsymbol{\mu}\boldsymbol{\mu}^\top + \frac{\varepsilon}{L}\boldsymbol{\beta}_Q^\top\boldsymbol{\mu}\boldsymbol{\mu}^\top\right)\boldsymbol{\mu} \tag{109}$$

$$= \left(\boldsymbol{\theta}_k^\top\boldsymbol{\mu} - \frac{1}{L}\boldsymbol{\theta}_k^\top\boldsymbol{\mu}\boldsymbol{\mu}^\top\boldsymbol{\mu} + \frac{1-\varepsilon}{L}\boldsymbol{\beta}_P^\top\boldsymbol{\mu}\boldsymbol{\mu}^\top\boldsymbol{\mu} + \frac{\varepsilon}{L}\boldsymbol{\beta}_Q^\top\boldsymbol{\mu}\boldsymbol{\mu}^T\boldsymbol{\mu}\right) \tag{110}$$

let $\mathbb{E}\left[\boldsymbol{\mu}^\top\boldsymbol{\mu}\right] = C$ for simplicity

$$= \boldsymbol{\theta}_k^\top\boldsymbol{\mu} - \frac{C}{L}\boldsymbol{\theta}_k^\top\boldsymbol{\mu} + \frac{C(1-\varepsilon)}{L}\boldsymbol{\beta}_P^\top\boldsymbol{\mu} + \frac{C\varepsilon}{L}\boldsymbol{\beta}_Q^\top\boldsymbol{\mu} \tag{111}$$

$$\boldsymbol{\theta}_{k+1}^\top\boldsymbol{\mu} - \boldsymbol{\theta}_k^\top\boldsymbol{\mu} = \left(-\frac{C}{L}\boldsymbol{\theta}_k^\top + \frac{C(1-\varepsilon)}{L}\boldsymbol{\beta}_P^\top + \frac{C\varepsilon}{L}\boldsymbol{\beta}_Q^\top\right)\boldsymbol{\mu} \tag{112}$$

It thus follows nicely

$$\boldsymbol{\theta}_{k+1} - \boldsymbol{\theta}_k = -\frac{C}{L}\boldsymbol{\theta}_k + \frac{C(1-\varepsilon)}{L}\boldsymbol{\beta}_P + \frac{C\varepsilon}{L}\boldsymbol{\beta}_Q \tag{113}$$

This gives us insight into how $\boldsymbol{\theta}_{k+1}$ changes. We can now calculate a telescopic sum to see the change in $\boldsymbol{\theta}$ after T iterations. We will first assume ε does not change. Furthermore, let us note we start at $\boldsymbol{\theta}_0 = \mathbf{0}$. Let us display a couple of iterations to show how $\boldsymbol{\theta}_k$ changes

$$\boldsymbol{\theta}_0 = \mathbf{0} \tag{114}$$

$$\boldsymbol{\theta}_1 = \frac{C(1-\varepsilon)}{L}\boldsymbol{\beta}_P + \frac{C\varepsilon}{L}\boldsymbol{\beta}_Q \tag{115}$$

$$\boldsymbol{\theta}_2 = -\frac{C}{L}\left(\frac{C(1-\varepsilon)}{L}\boldsymbol{\beta}_P + \frac{C\varepsilon}{L}\boldsymbol{\beta}_Q\right) + \frac{C(1-\varepsilon)}{L}\boldsymbol{\beta}_P + \frac{C\varepsilon}{L}\boldsymbol{\beta}_Q \tag{116}$$

$$= \frac{CL(1-\varepsilon)}{L^2}\boldsymbol{\beta}_P + \frac{CL\varepsilon}{L^2}\boldsymbol{\beta}_Q - \frac{C^2(1-\varepsilon)}{L^2}\boldsymbol{\beta}_P - \frac{C^2\epsilon}{L^2}\boldsymbol{\beta}_Q \tag{117}$$

$$= \frac{(CL-C^2)(1-\varepsilon)}{L^2}\boldsymbol{\beta}_P + \frac{(CL-C^2)\epsilon}{L^2}\boldsymbol{\beta}_Q \tag{118}$$

This only converges to $\boldsymbol{\beta}_P$ when ε is decreasing. Otherwise it will simply converge to the OLS optimal solution we proved in theorem 1. □

F EXPERIMENTAL DETAILS

F.1 QUADRATIC REGRESSION DATASET

We will describe $\mathbb{P}$ and $\mathbb{Q}$ in the `Quadratic Regression` dataset.
$x \sim \mathcal{N}(0, 1)$
$n_{\text{train}} \in \{10\text{e}4, 10\text{e}6\}$
$\mathbb{P} : y|x \sim \mathcal{N}(x^2 - x + 2, 0.01)$
$\mathbb{Q} : y|x \sim \mathcal{N}(-x^2 + x + 4, 0.01)$

F.2 DRUG DISCOVERY DATASET

This dataset is downloaded from Diakonikolas et al. (2019). We utilize the same noise procedure as in Li et al. (2020).
$\mathbb{P}$ is given from an 80/20 train test split from the dataset.
$\mathbb{Q}$ is random noise sampled from $\mathcal{N}(5, 5)$

F.3 BASELINE METHODS 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.

$$\boldsymbol{X}^* = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{y} \tag{119}$$

Huber Regression is solved with the following objective function.

$$L_\delta(y, f(\boldsymbol{x})) = \begin{cases} \frac{1}{2}(y - f(\boldsymbol{x}))^2 \\ \delta \cdot \left(|y - f(\boldsymbol{x})| - \frac{1}{2}\delta\right) \quad \text{otherwise} \end{cases} \tag{120}$$

RANSAC