ROBUST LINEAR REGRESSION BY SUB-QUANTILE OPTIMIZATION

Arvind Rathnashyam, Fatih Orhan, Joshua Myers, & Jake Herman *
Department of Computer Science
Rensselaer Polytechnic University
Troy, NY 12180, USA
{rathna, orhanf, myersj5, hermaj2}@rpi.edu

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. The key motivating factor in investigating robust linear regression is the sheer vastness of probability distributions that are not drawn from a normal distribution schema. Given that outliers in data sets occur so frequent, the ability for a linear regression model to be robust is necessary to compensate for the various distributions being analyzed.

1.1 MOTIVATIONS

The failure of classical regression techniques being unable to model data highly corrupted by outliers can be conveyed clearly by Yu et al. (2014) in predicting death rates (per million people) of a country given its cigarette consumption per capita twenty years prior. When numerous outliers have considerably high leverage on a data set, they can vastly skew a model's predictions if said outliers are not accounted for. When utilizing more robust techniques, including MM-estimation or robust and efficient weighted least squares estimator (REWLSE), the data can represented more accurately without being overwhelmed by outlying data.

One of the most prominent cases in which robust linear regression serves as a sufficiently resilient method of prediction is in Khan et al. (2021) analyzing economic growth, or the rise in gross domestic product (GDP) of a nation. The econometric application of robust regression techniques serves to identify how various robust regression techniques performs at assessing how the rapid population growth seen in India and Pakistan affect the economic growth of said countries. The outliers seen in the data sets analyzed by Khan et al. affect both independent and dependent variables, thus having a robust model that is resilient to the sheer number of outliers is crucial to produce accurate results.

The value and necessity of robust linear regression can also be demonstrated in meteorology and climatology by Muhlbauer et al. (2009). In analysing temperature and precipitation time-series data collected in Switzerland, classical regression techniques (namely least-squares regression) were compared to two robust regression techniques: least median of squares (LMS) and the aforementioned LTS estimators. In comparing the two techniques, it was apparent that classical regression techniques can be consequentially sensitive to outliers, while the rudimentary robust regression models proved to be resilient to outliers. Furthermore, with the ever-increasing likelihood of outlying

*Work done as a part of ML and Optimization Spring 2023 Group Project.

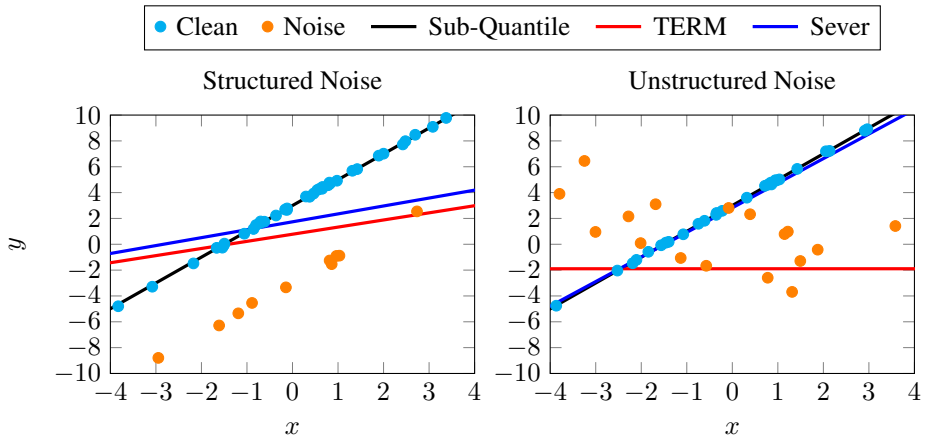


Figure 1: Sub-Quantile Performance on Linearly Dependent and Independent Noise

meteorological observations (heat waves, heavy precipitation, etc.), it is imperative that models be robust enough to quantify an accurate trend that accounts for the existence of outlier-corrupted data.

1.2 Contributions

Our goal is to provide a theoretic analysis and convergence conditions for sub-quantile optimization and offer practitioners a method for robust linear regression. Several popular methods have been utilized due to their simplicity and high effectiveness including quantile regression Koenker & Hallock (2001), Theil-Sen Estimator Sen (1968), and Huber Regression Huber & Ronchetti (2009). These methods, although rudimentary, serve to show the effectiveness of building resistance against outliers in data. By improving upon existing methods, namely least-squares estimation in these cases, models can be designed to better estimate data sets with considerably corruptive outliers.

Sub-Quantile Optimization aims to address the shortcomings of ERM in applications such as noisy/corrupted data (Khetan et al. (2018),Jiang et al. (2018)), classification with imbalanced classes, (Lin et al. (2017),He & Garcia (2009)), as well as fair learning (Corbett-Davies & Goel (2018)).

As seen in the above comparison, current models fail to estimate data sets corrupted by structured noise, with some models even failing to estimate trends plagued with unstructured noise. Through this, sub-quantile optimization is shown to prevail at overcoming these challenges current models currently face.

2 Related Work

Least Trimmed Squares (LTS) Mount et al. (2014) is an estimator that relies on minimizing the sum of the smallest h residuals given a $(d-1)$-dimension hyperplane calculated given n data points in $\mathbf{R}^d$ and an integer trimming parameter h. Given that the outliers comprise less than half the data, this algorithm is more efficient than the more common LMS estimator. However, this algorithm unfortunately suffers from the curse of dimensionality; the computational cost of the algorithm grows exponentially with increasing dimensions of the data. Thus, the necessity to design a more computationally efficient algorithm is expressed.

Tilted Empirical Risk Minimization (TERM) Li et al. (2020) is a framework built to similarly handle the shortcomings of emperical risk minimization (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) Laguel et al. (2021)

Robust Risk Minimization Osama et al. (2020)

Quantile Regression Yu et al. (2003)

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*

$$\boldsymbol{y}_P = \boldsymbol{\beta}_P^\top \boldsymbol{P} + \epsilon_P \tag{12}$$

and similarly

$$\boldsymbol{y}_Q = \boldsymbol{\beta}_Q^\top \boldsymbol{Q} + \epsilon_Q \tag{13}$$

where $\boldsymbol{\beta}_P$ and $\boldsymbol{\beta}_Q$ the oracle regressors for $\mathbb{P}$ and $\mathbb{Q}$ and ϵ_P and ϵ_Q are both Normally Distributed with mean 0*.*

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 A.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 $\hat{\mathbb{P}}$*

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

The proof can be found in Appendix A.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 B.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 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.*

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

Now we will state two properties regarding the effect of the t-update step and the $\boldsymbol{\theta}$-update step as described in Equations 10 and 11, respectively.

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 B.4 □

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. In other words:

$$\mathbb{E}\left[\nabla_{\boldsymbol{\theta}} g(t_{k+1}, \boldsymbol{\theta}_k)\right] = 0$$

$$2(\boldsymbol{\mu}\boldsymbol{\mu}^\top + \boldsymbol{\Sigma})(\boldsymbol{\theta}_k - (1-\varepsilon)\boldsymbol{\beta}_P - \varepsilon\boldsymbol{\beta}_Q) = 0$$

Since the first term is non-zero, the equality is satisfied when:

$$(\boldsymbol{\theta}_k - (1-\varepsilon)\boldsymbol{\beta}_P - \varepsilon\boldsymbol{\beta}_Q) = 0$$

$$\boldsymbol{\theta}_k = (1-\varepsilon)\boldsymbol{\beta}_P + \varepsilon\boldsymbol{\beta}_P$$

Note this aligns with the results of Theorem 1. 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 4.7. *We will utilize lemmas 4.3 and 4.5 to quantify the change of g after the theta-update.*

Proof Sketch. We are specifically interested in the case when g increases after a theta-update. □

4.3 Converging to β_P

We will start by defining the two types of noise we are interesting in.

Definition 6. ***Unstructured Noise** is noise that is not dependent on the input data, i.e., $\mathbb{P}[\boldsymbol{y}|\boldsymbol{X}] = \mathbb{P}[\boldsymbol{X}]$*

Definition 7. ***Linearly Structured Noise** is noise that is made from a linear combination of the input data, i.e. $\boldsymbol{y} = \boldsymbol{\beta}_Q \boldsymbol{X} + \boldsymbol{\epsilon}$*

Also note we often consider Gaussian Noise as Unstrucutred Noise, but it can be modeled as Structured Noise where $\boldsymbol{\beta}_Q = \mathbf{0}$.

Theorem 3. *The Expected Value of Epsilon and the p-Quantile can be calculated as:*

$$\mathbb{E}[\varepsilon] = \frac{\mathbb{E}[P^+]}{np} \tag{26}$$

$$\mathcal{Q}_p = \dots \tag{27}$$

We provide a proof in Appendix D.1. Furthermore, we validate our theoretical estimation of the number of elements of $\mathbb{P}$ less than $\mathcal{Q}_p$ as well as our estimation of $\mathcal{Q}_p$ in figures ... & .... Our strongest theoretical results for convergence come in the case of *Linearly Structured Noise*.

Theorem 4. *The convergence of SubQuantile Optimization is dependent on the initial weights $\boldsymbol{\theta}_0$ in the case of linearly structured noise.*

Proof Sketch. Let us consider a worst-case scenario where $\boldsymbol{\theta}_k = \boldsymbol{\beta}_Q$. Note this is not an impossible situation as if you naively choose $\boldsymbol{\theta}_0 = \mathbf{0}$ and $\mathbb{Q} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_Q)$. □

Theorem 4 tells us we can not randomly select weights and expect to converge to the majority class. We must choose the initial weights in a specific way as such so we can guarantee the majority elements by expectation have more representation within the lowest np squared losses after the loss iteration.

Theorem 5. *SubQuantile Optimization converges to a local minimum independent of initial $\boldsymbol{\theta}_0$ initialization.*

Proof Sketch. By Theorem 1, $\boldsymbol{\theta}_0 = (1-\epsilon)\boldsymbol{\beta}_P + \epsilon\boldsymbol{\beta}_Q$ □

We are now interested in theoretical guarantees with no distributional assumptions. First we will consider some intuition on why this problem is not as hard as compared to when the corruption is linearly structured. Let us say the noise is of non-linear regression, in other words $\boldsymbol{y}_Q \sim f(\boldsymbol{X}, \boldsymbol{\beta}_Q)$, where f is a non-linear combination of features of $\boldsymbol{X}$. In this case, it is not possible to model the non-linear regression by a linear combination of the features, thus, if we have elements from $\mathbb{Q}$ within the lowest np losses, then training on these points will not generalize well to points from $\mathbb{Q}$, so their error will not decrease.

4.4 Complexity of SubQuantile Optimization

In this section we will provide the expected complexity in the case of linearly structured noise.

5 Empirical Results

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

$$t_{k+1} = \hat{\boldsymbol{\nu}}_{np} \tag{28}$$

$$t_{k+1} = \frac{1}{np}\sum_{i=1}^{np} \hat{\boldsymbol{\nu}}_i \tag{29}$$

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 28. We also compare against the t-update as described in equation 29.

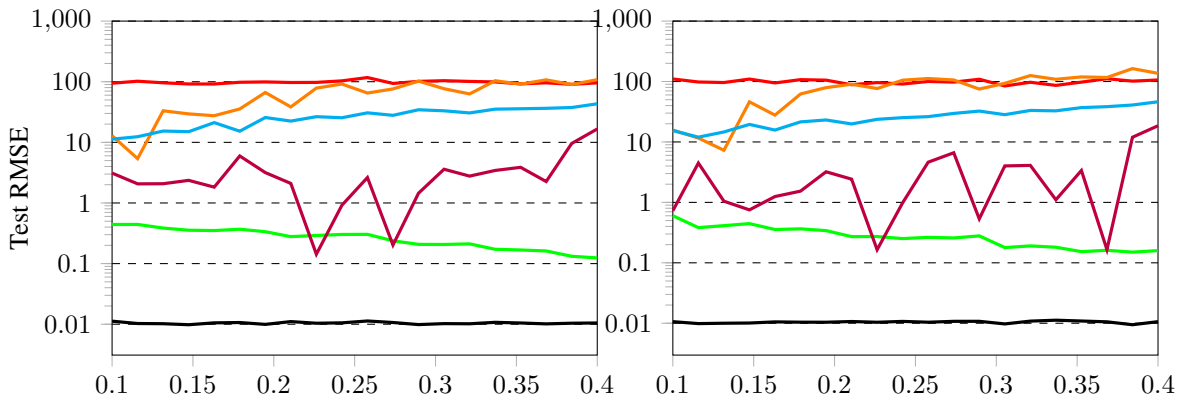
Algorithm 1: Sub-Quantile Minimization Optimization Algorithm

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: $y|x \sim \mathcal{N}(-x^2 + x + 4, 0.01)$

$\boldsymbol{\theta}_0 \leftarrow \frac{2}{L}\sum_{i=1}^{n}(\boldsymbol{x}_i y_i)$
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
if $\epsilon > 0.5$ **then**
  $\boldsymbol{P} = \hat{\boldsymbol{\nu}}\,[np:]$
  $\boldsymbol{y}_P = \boldsymbol{y}\,[np:]$
  $\boldsymbol{\theta}_T = (\boldsymbol{P}^\top \boldsymbol{P})^{-1}\boldsymbol{P}^\top \boldsymbol{y}_P$
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.

5.1 Synthetic Data



We now demonstrate SubQuantile Regression in the presence of Gaussian Random Noise.

From the results we can see in Figure **??**, Subquantile Minimization performs better throughout all noise ranges. The one struggle exists when ϵ is around 0.5, thus we face issues similar to the power method where there exists the top two eigenvalues such that $|\lambda_1| \approx |\lambda_2|$.

In our first synthetic experiment, we run Algorithm 1 on synthetically generated structured linear regression data, the noise is sampled from a linear distribution that is dependent on the vector of $\boldsymbol{X}$. The results of Sub-Quantile Minimization can be seen in Figure **??**. Our results show the near optimal performance of Sub-Quantile Minimization. The results and comparison with other methods can be seen in Table 3. Note we are not interested in $\epsilon \geq 0.5$ as the concept of corruptness becomes unclear. We see in Table 3, 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

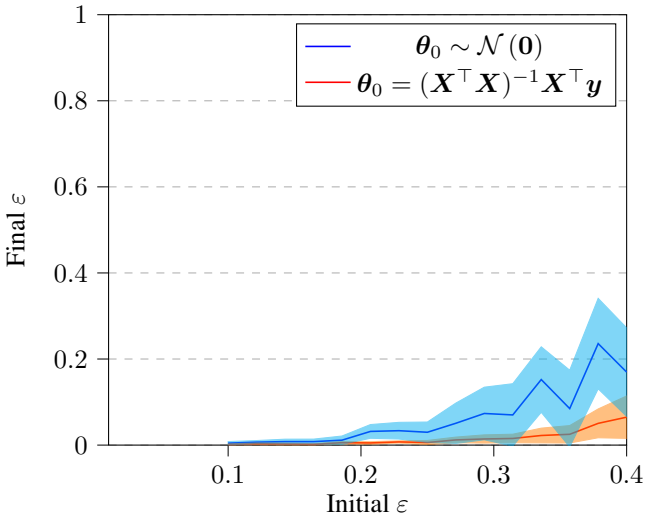


Figure 2: Probability of points from $\mathbb{Q}$ in the final subquantile

large data datasets. Please refer to Appendix F for more details on the `Structured Linear Regression` Dataset.

In our second synthetic experiment, we run Algorithm 1 similarly on synthetically generated linear regression data. However, in this experiment, the noise is sampled from a Gaussian that is independent of the $\boldsymbol{X}$ coordinates.

Methods such as TERM, Li et al. (2020), are unable to capture the target distribution through structurally generated noise, which can also be called *adversarial*. SubQuantile Optimization, on the other hand, is robust to such adversarial attacks.

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).

Objectives	Test RMSE (`Drug Discovery`)			
	$\epsilon = 0$	$\epsilon = 0.2$	$\epsilon = 0.4$	$\epsilon = 0.8$
OLS 125	$0.990_{(0.060)}$	$1.969_{(0.118)}$	$2.829_{(0.086)}$	$4.682_{(0.101)}$
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)	$1.313_{(0.072)}$	$1.334_{(0.105)}$	$1.343_{(0.0740)}$	$\mathbf{1.428_{(0.107)}}$
SEVER Diakonikolas et al. (2019)	$1.079_{(0.059)}$	$1.076_{(0.048)}$	$\mathbf{1.067_{(0.091)}}$	$3.993_{(0.203)}$
SubQuantile($p = 1 - \epsilon$)	$\mathbf{1.052_{(0.062)}}$	$\mathbf{1.060_{(0.065)}}$	$1.073_{(0.101)}$	$1.479_{(0.0695)}$
Genie ERM	$0.990_{(0.060)}$	$1.038_{(0.041)}$	$1.037_{(0.086)}$	∞

Table 2: `Drug Discovery` Dataset. Empirical Risk over $\mathbb{P}$

The results in figure 2 demonstrate the number of elements from $\mathbb{Q}$ in the final subquantile matrix $\boldsymbol{S}_T$ from different theta initializations after T iterations of subquantile minimization. As we can see in Table 2, 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. This dataset is also

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.

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.

Sam Corbett-Davies and Sharad Goel. The measure and mismeasure of fairness: A critical review of fair machine learning. *CoRR*, abs/1808.00023, 2018. URL http://arxiv.org/abs/1808.00023.

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.

Haibo He and Edwardo A. Garcia. Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, 21(9):1263–1284, 2009. doi: 10.1109/TKDE.2008.239.

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.

Lu Jiang, Zhengyuan Zhou, Thomas Leung, Li-Jia Li, and Li Fei-Fei. Mentornet: Learning data-driven curriculum for very deep neural networks on corrupted labels. In *ICML*, 2018.

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.

Dost Muhammad Khan, Anum Yaqoob, Seema Zubair, Muhammad Azam Khan, Zubair Ahmad, and Osama Abdulaziz Alamri. Applications of robust regression techniques: An econometric approach. *Mathematical Problems in Engineering*, 2021:6525079, May 2021.

Ashish Khetan, Zachary C. Lipton, and Anima Anandkumar. Learning from noisy singly-labeled data. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=H1sUHgb0Z.

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.

Yassine Laguel, Krishna Pillutla, Jérôme Malick, and Zaid Harchaoui. Superquantiles at work: Machine learning applications and efficient subgradient computation. *Set-Valued and Variational Analysis*, 29(4):967–996, Dec 2021. ISSN 1877-0541. doi: 10.1007/s11228-021-00609-w. URL https://doi.org/10.1007/s11228-021-00609-w.

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

T. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar. Focal loss for dense object detection. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 2999–3007, Los Alamitos, CA, USA, oct 2017. IEEE Computer Society. doi: 10.1109/ICCV.2017.324. URL https://doi.ieeecomputersociety.org/10.1109/ICCV.2017.324.

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.

Andreas Muhlbauer, Peter Spichtinger, and Ulrike Lohmann. Application and comparison of robust linear regression methods for trend estimation. *Journal of Applied Meteorology and Climatology*, 48(9):1961 – 1970, 2009. doi: https://doi.org/10.1175/2009JAMC1851.1. URL https://journals.ametsoc.org/view/journals/apme/48/9/2009jamc1851.1.xml.

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.

Chun Yu, Weixin Yao, and Xue Bai. Robust linear regression: A review and comparison, 2014.

Keming Yu, Zudi Lu, and Julian Stander. Quantile regression: applications and current research areas. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 52(3):331–350, 2003. doi: https://doi.org/10.1111/1467-9884.00363. URL https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/1467-9884.00363.

A Proofs on the effect of Linear Corruption **14**
A.1 Proof of Theorem 1 14
A.2 Proof of Theorem 2 15

B General Properties of Sub-Quantile Minimization **16**
B.1 Derivation of Lemma 4.2 16
B.2 Derivation of Lemma 4.3 16
B.3 Derivation of Lemma 4.5 17
B.4 Proof of Lemma 4.6 17

C Stochastic Sub-Quantile Optimization **19**

D Theory for Linear Corruption **20**
D.1 Proof of Theorem 3 20
D.2 Proof of Theorem 4 21
D.3 Proof of Theorem 5 23

E Additional Experiments **24**

F Experimental Details **25**
F.1 `Structured Linear Regression` Dataset 25
F.2 `Noisy Linear Regression` Dataset 25
F.3 `Quadratic Regression` Dataset 25
F.4 `Drug Discovery` Dataset 25
F.5 Baseline Methods in Section 5 25

A PROOFS ON THE EFFECT OF LINEAR CORRUPTION

A.1 PROOF OF THEOREM 1

Proof.
We will first calculate the pseudo-inverse

$$\boldsymbol{X}^\top \boldsymbol{X} = \begin{pmatrix} \boldsymbol{P}^\top & \boldsymbol{Q}^\top \end{pmatrix} \begin{pmatrix} \boldsymbol{P} \\ \boldsymbol{Q} \end{pmatrix} \tag{30}$$

$$= \boldsymbol{P}^\top \boldsymbol{P} + \boldsymbol{Q}^\top \boldsymbol{Q} \tag{31}$$

Now we can calculate the Moore-Penrose Inverse

$$(\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top = (\boldsymbol{P}^\top \boldsymbol{P} + \boldsymbol{Q}^\top \boldsymbol{Q})^{-1} \begin{pmatrix} \boldsymbol{P}^\top & \boldsymbol{Q}^\top \end{pmatrix} \tag{32}$$

$$= \begin{pmatrix} (\boldsymbol{P}^\top \boldsymbol{P} + \boldsymbol{Q}^\top \boldsymbol{Q})^{-1} \boldsymbol{P}^\top & (\boldsymbol{P}^\top \boldsymbol{P} + \boldsymbol{Q}^\top \boldsymbol{Q})^{-1} \boldsymbol{Q}^\top \end{pmatrix} \tag{33}$$

Now we solve for the optimal model

$$\boldsymbol{X}^\dagger \boldsymbol{y} = \begin{pmatrix} (\boldsymbol{P}^\top \boldsymbol{P} + \boldsymbol{Q}^\top \boldsymbol{Q})^{-1} \boldsymbol{P}^\top & (\boldsymbol{P}^\top \boldsymbol{P} + \boldsymbol{Q}^\top \boldsymbol{Q})^{-1} \boldsymbol{Q}^\top \end{pmatrix} \begin{pmatrix} \boldsymbol{y}_P \\ \boldsymbol{y}_Q \end{pmatrix} \tag{34}$$

$$= (\boldsymbol{P}^\top \boldsymbol{P} + \boldsymbol{Q}^\top \boldsymbol{Q})^{-1} \boldsymbol{P}^\top \boldsymbol{y}_P + (\boldsymbol{P}^\top \boldsymbol{P} + \boldsymbol{Q}^\top \boldsymbol{Q})^{-1} \boldsymbol{Q}^\top \boldsymbol{y}_Q \tag{35}$$

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).

$$\boldsymbol{P}^\top \boldsymbol{P} = \sum_{i=1}^{n*(1-\epsilon)} \boldsymbol{P}_i \boldsymbol{P}_i^\top \tag{36}$$

$$\boldsymbol{Q}^\top \boldsymbol{Q} = \sum_{j=1}^{n\epsilon} \boldsymbol{Q}_j \boldsymbol{Q}_j^\top \tag{37}$$

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

$$\boldsymbol{P}^\top \boldsymbol{P} \sim \mathcal{W}(n(1-\epsilon), \boldsymbol{\Sigma}) \tag{38}$$

$$\boldsymbol{Q}^\top \boldsymbol{Q} \sim \mathcal{W}(n\epsilon, \boldsymbol{\Sigma}) \tag{39}$$

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

$$\mathbb{E}(\boldsymbol{P}^\top \boldsymbol{P}) = n(1-\epsilon)\boldsymbol{\Sigma} \tag{40}$$

$$\mathbb{E}(\boldsymbol{Q}^\top \boldsymbol{Q}) = n\epsilon\boldsymbol{\Sigma} \tag{41}$$

It thus follows

$$\mathbb{E}\left[\boldsymbol{P}^\top \boldsymbol{P} + \boldsymbol{Q}^\top \boldsymbol{Q}\right] = n\boldsymbol{\Sigma} \tag{42}$$

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

$$\left(\boldsymbol{P}^\top \boldsymbol{P} + \boldsymbol{Q}^\top \boldsymbol{Q}\right)^{-1} \sim \mathcal{W}^{-1}(n, \boldsymbol{\Sigma}) \tag{43}$$

It thus follows by the expectation of the Inverse Wishart Distribution

$$\mathbb{E}\left[\left(\boldsymbol{P}^\top \boldsymbol{P} + \boldsymbol{Q}^\top \boldsymbol{Q}\right)^{-1}\right] = n\boldsymbol{\Sigma}^{-1} \tag{44}$$

Now we will plug this into Equation 35:

$$\mathbb{E}\left[\boldsymbol{X}^\dagger \boldsymbol{y}\right] = \left(n\boldsymbol{\Sigma}^{-1}\right) \boldsymbol{P}^\top \boldsymbol{y}_P + \left(n\boldsymbol{\Sigma}^{-1}\right) \boldsymbol{Q}^\top \boldsymbol{y}_Q \tag{45}$$

$$= \left(n\boldsymbol{\Sigma}^{-1}\right) \boldsymbol{P}^\top (\boldsymbol{P}\boldsymbol{\beta} + \boldsymbol{\epsilon}_P) + \left(n\boldsymbol{\Sigma}^{-1}\right) \boldsymbol{Q}^\top (\boldsymbol{Q}\boldsymbol{\beta}_Q^\top + \boldsymbol{\epsilon}_Q) \tag{46}$$

$$= \left(n\boldsymbol{\Sigma}^{-1}\right) \left((\boldsymbol{P}^\top \boldsymbol{P})\boldsymbol{\beta}_P + (\boldsymbol{Q}^\top \boldsymbol{Q}) (\boldsymbol{\beta}_P + (\boldsymbol{\beta}_Q - \boldsymbol{\beta}_P))\right) \tag{47}$$

$$= \left(n\boldsymbol{\Sigma}^{-1}\right) \left((n(1-\epsilon)\boldsymbol{\Sigma})\boldsymbol{\beta}_P + n\epsilon\boldsymbol{\Sigma} (\boldsymbol{\beta}_P + \boldsymbol{\Psi})\right) \tag{48}$$

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

$$= \boldsymbol{\beta}_P + \epsilon(\boldsymbol{\Psi}) \tag{50}$$

This concludes the proof. □

A.2 PROOF OF THEOREM 2

Proof. The first half of the proof follows from Appendix A.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 $\mathbb{Q}$.

$$\mathbb{E}\left[\boldsymbol{P}^\top \boldsymbol{P}\right] = n(1-\epsilon)\mathbf{\Sigma}_P \tag{51}$$

$$\mathbb{E}\left[\boldsymbol{Q}^\top \boldsymbol{Q}\right] = n\epsilon\mathbf{\Sigma}_Q \tag{52}$$

It thus follows

$$\mathbb{E}\left[\boldsymbol{P}^\top \boldsymbol{P} + \boldsymbol{Q}^\top \boldsymbol{Q}\right] = (n(1-\epsilon)\mathbf{\Sigma}_P + n\epsilon\mathbf{\Sigma}_Q) \tag{53}$$

This is where the structure of the proof differs from Theorem 1 because we can no longer follow the Inverse Wishart Distribution.

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

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

$$= n(1-\epsilon)\mathbf{\Sigma}_P^{-1} - n(1-\epsilon)\mathbf{\Sigma}_P^{-1}(n\epsilon\mathbf{\Sigma}_Q^{-1}) \tag{55}$$

We will now calculate the expected optimal parameters by plugging this into Equation 35:

$$\mathbb{E}\left[\boldsymbol{X}^\dagger \boldsymbol{y}\right] = n(1-\epsilon)\mathbf{\Sigma}_P^{-1}(\boldsymbol{P}^T\boldsymbol{P})\boldsymbol{\beta}_P - n(1-\epsilon)\mathbf{\Sigma}_P^{-1}(n\epsilon\mathbf{\Sigma}_Q^{-1})(\boldsymbol{Q}^T\boldsymbol{Q})\boldsymbol{\beta}_Q \tag{56}$$

$$= n(1-\epsilon)\mathbf{\Sigma}_P^{-1}(n(1-\epsilon)\mathbf{\Sigma}_P)\boldsymbol{\beta}_P - n(1-\epsilon)\mathbf{\Sigma}_P^{-1}(n\epsilon\mathbf{\Sigma}_Q^{-1})(n\epsilon\mathbf{\Sigma}_Q)\boldsymbol{\beta}_Q \tag{57}$$

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

This concludes the proof. □

B General Properties of Sub-Quantile Minimization

B.1 Derivation of Lemma 4.2

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 $0 < p < 1$

$$\partial(-g(t, \boldsymbol{\theta})) = \partial\left(-t + \frac{1}{np}\sum_{i=1}^{n}(t - \hat{\boldsymbol{\nu}}_i)^+\right) \tag{59}$$

$$= \partial(-t) + \partial\left(\frac{1}{np}\sum_{i=1}^{n}(t - \hat{\boldsymbol{\nu}}_i)^+\right) \tag{60}$$

$$= -1 + \frac{1}{np}\sum_{i=1}^{n}\partial(t - \hat{\boldsymbol{\nu}}_i)^+ \tag{61}$$

$$= -1 + \frac{1}{np}\sum_{i=1}^{n}\left\{\begin{array}{ll} 1, & \text{if } t > \hat{\boldsymbol{\nu}}_i \\ 0, & \text{if } t < \hat{\boldsymbol{\nu}}_i \\ [0,1], & \text{if } t = \hat{\boldsymbol{\nu}}_i \end{array}\right\} \tag{62}$$

$$= 0 \text{ when } t = \hat{\boldsymbol{\nu}}_{np} \tag{63}$$

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

B.2 Derivation of Lemma 4.3

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

$$\nabla_{\boldsymbol{\theta}_k} g(t_{k+1}, \boldsymbol{\theta}_k) = \nabla_{\boldsymbol{\theta}_k}\left(\boldsymbol{\nu}_{np} - \frac{1}{np}\sum_{i=1}^{n}(\boldsymbol{\nu}_{np} - (\boldsymbol{\theta}_k^\top \boldsymbol{x}_i - y_i)^2)^+\right) \tag{64}$$

$$= \nabla_{\boldsymbol{\theta}_k}\left((\boldsymbol{\theta}_k^\top \boldsymbol{x}_{np} - y_{np})^2 - \frac{1}{np}\sum_{i=1}^{n}\left((\boldsymbol{\theta}_k^\top \boldsymbol{x}_{np} - y_{np})^2 - (\boldsymbol{\theta}_k^\top \boldsymbol{x}_i - y_i)^2\right)^+\right) \tag{65}$$

$$= \nabla_{\boldsymbol{\theta}_k}(\boldsymbol{\theta}_k^\top \boldsymbol{x}_{np} - y_{np})^2 - \frac{1}{np}\sum_{i=1}^{n}\nabla_{\boldsymbol{\theta}_k}\left((\boldsymbol{\theta}_k^\top \boldsymbol{x}_{np} - y_{np})^2 - (\boldsymbol{\theta}_k^\top \boldsymbol{x}_i - y_i)^2\right)^+ \tag{66}$$

$$= 2\boldsymbol{x}_{np}(\boldsymbol{\theta}_k^\top \boldsymbol{x}_{np} - y_{np}) - \frac{1}{np}\sum_{i=1}^{n}2\boldsymbol{x}_{np}(\boldsymbol{\theta}_k^\top \boldsymbol{x}_{np} - y_{np}) - 2\boldsymbol{x}_i(\boldsymbol{\theta}_k^\top \boldsymbol{x}_i - y_i)\left\{\begin{array}{ll} 1, & \text{if } t > v_i \\ 0, & \text{if } t < v_i \\ [0,1], & \text{if } t = v_i \end{array}\right\} \tag{67}$$

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

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

$$= \frac{1}{np}\sum_{i=1}^{np}2\boldsymbol{x}_i(\boldsymbol{\theta}_k^\top \boldsymbol{x}_i - y_i) \tag{70}$$

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

B.3 Derivation of Lemma 4.5

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

$$||\nabla_{\boldsymbol{\theta}} g(\boldsymbol{\theta}', t) - \nabla_{\boldsymbol{\theta}} g(\boldsymbol{\theta}, t)|| \leq L||\boldsymbol{\theta}' - \boldsymbol{\theta}|| \tag{71}$$

$$\left\|\nabla_{\boldsymbol{\theta}} g(\boldsymbol{\theta}^{'}, t) - \nabla_{\boldsymbol{\theta}} g(\boldsymbol{\theta}, t)\right\| = \left\| \frac{1}{np} \sum_{i=1}^{np} 2\boldsymbol{x}_i(\boldsymbol{\theta}_k^{'T}\boldsymbol{x}_i - y_i) - \frac{1}{np} \sum_{i=1}^{np} 2\boldsymbol{x}_i(\boldsymbol{\theta}_k^{\top}\boldsymbol{x}_i - y_i) \right\| \tag{72}$$

$$= \left\| \frac{1}{np} \sum_{i=1}^{np} 2\boldsymbol{x}_i(\boldsymbol{\theta}_k^{'T}\boldsymbol{x}_i - \boldsymbol{\theta}_k^{\top}\boldsymbol{x}_i) \right\| \tag{73}$$

$$= \left\| \frac{1}{np} \sum_{i=1}^{np} 2\boldsymbol{x}_i^{\top}\boldsymbol{x}_i(\boldsymbol{\theta}_k^{'T} - \boldsymbol{\theta}_k^{\top}) \right\| \tag{74}$$

$$\overset{\text{Cauchy}-\text{Schwarz}}{\leq} \left\| \frac{2}{np} \sum_{i=1}^{np} \|\boldsymbol{x}_i\|^2 \right\| \left\| \boldsymbol{\theta}_k^{'T} - \boldsymbol{\theta}_k^{\top} \right\| \tag{75}$$

$$= L \left\| \boldsymbol{\theta}_k^{'T} - \boldsymbol{\theta}_k^{\top} \right\| \tag{76}$$

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

This concludes the proof.

B.4 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 $t_k \geq \hat{\boldsymbol{\nu}_k}$

$$g(t_k, \boldsymbol{\theta}_k) = t_k - \frac{1}{np} \sum_{i=1}^{n} (t_k - \boldsymbol{\nu}_i)^+ \tag{77}$$

$$= t_k - \frac{1}{np}(np)t_k + \frac{1}{np} \sum_{i=1}^{np} \boldsymbol{\nu}_i + \frac{1}{np} \sum_{i=np}^{n} (t_k - \boldsymbol{\nu}_i)^+ \tag{78}$$

$$= \frac{1}{np} \sum_{i=1}^{np} \boldsymbol{\nu}_i + \frac{1}{np} \sum_{i=np}^{n} (t_k - \boldsymbol{\nu}_i)^+ \tag{79}$$

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

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

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 $\delta \in (0, p)$.

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

$$= t_k - \frac{1}{np} \sum_{i=1}^{n(p-\delta)} (t_k - \boldsymbol{\nu}_i)^+ \tag{83}$$

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

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

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

This concludes the proof. □

C STOCHASTIC SUB-QUANTILE OPTIMIZATION

In the age of big data, stochastic methods are necessary for fast training of models to handle large amounts of data. In this section we will provide an algorithm for Stochastic Sub-Quantile Optimization and prove convergence.

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 $\boldsymbol{\theta}_T$

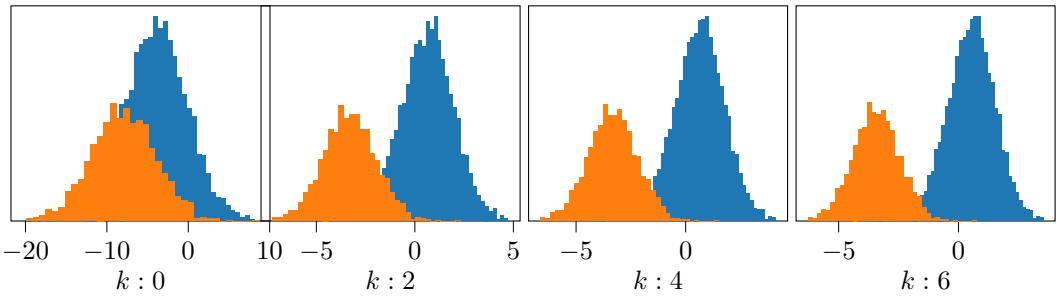


Figure 3: Residuals with respect to $\mathbb{P}$ and $\mathbb{Q}$

D Theory for Linear Corruption

In this section, we provide rigorous theory for why Sub-Quantile Minimization works so well in the case of corruption of the form $\boldsymbol{\beta}_Q^\top \boldsymbol{\mu} = y_p + \epsilon_Q$.

Assumption 3. *The residuals of $\boldsymbol{\theta}_k$ are normally distributed with respect to $\mathbb{P}$ and $\mathbb{Q}$. In other words, $\boldsymbol{\theta}_k \boldsymbol{p} - y_P$ and $\boldsymbol{\theta}_k \boldsymbol{q} - y_Q$ are normally distributed.*

Assumption 3 can be visually verified in figure 3. Even after multiple iteration steps the residuals with respect to $\mathbb{P}$ and $\mathbb{Q}$ are still normal.

D.1 Proof of Theorem 3

Proof. By Assumptions 1 and 2 we can do the following manipulations. Furthermore, please note $\boldsymbol{P}$ is a random matrix, $\boldsymbol{\theta}$, $\boldsymbol{\beta}_P$ are non-random vectors, where $\boldsymbol{\beta}^\top \boldsymbol{P}$ is a random vector-matrix product and $\mathbb{E}\left[\boldsymbol{\beta}_P^\top \boldsymbol{p}\right] = y_P$ and $\mathrm{Var}\left(\boldsymbol{\beta}_T^\top \boldsymbol{p}\right) = \epsilon_P$. In this theoretical analysis, we assume no feature corruption, i.e., $\mathbb{E}\left[\boldsymbol{p}_i\right] = \mathbb{E}\left[\boldsymbol{q}_i\right] = \boldsymbol{\mu}$. We will now analytically calculate the expected least squares error:

Recalculate expected value by using $\mathbb{E}\left[X^2\right] = \mathbb{E}\left[X\right]^2 + \mathrm{Var}(X)$

$$\begin{aligned}
\mathbb{E}\left[\boldsymbol{\theta}^\top \boldsymbol{\mu} - y_P\right] &= \mathbb{E}\left[\boldsymbol{\theta}^\top \boldsymbol{\mu} - (\boldsymbol{\beta}_P^\top \boldsymbol{\mu} + \epsilon_P)\right] \\
&= \mathbb{E}\left[(\boldsymbol{\theta}^\top - \boldsymbol{\beta}_P^\top)\boldsymbol{\mu} - \epsilon_P\right] \\
&= \mathbb{E}\left[(\boldsymbol{\theta}^\top - \boldsymbol{\beta}_P^\top)\boldsymbol{\mu}\right] + \mathbb{E}\left[-\epsilon_P\right] \\
&= \mathbb{E}\left[\boldsymbol{\theta}^\top - \boldsymbol{\beta}_P^\top\right]\mathbb{E}\left[\boldsymbol{\mu}\right] \\
&= (\boldsymbol{\theta}^\top - \boldsymbol{\beta}_P^\top)\boldsymbol{\mu} \\
\mathrm{Var}\left(\boldsymbol{\theta}^\top \boldsymbol{\mu} - y_P\right) &= \mathrm{Var}\left(\boldsymbol{\theta}^\top \boldsymbol{\mu} - (\boldsymbol{\beta}_P^\top \boldsymbol{\mu} + \epsilon_P)\right) \\
&= \mathrm{Var}\left((\boldsymbol{\theta}^\top - \boldsymbol{\beta}_P^\top)\boldsymbol{\mu} + \epsilon_P\right) \\
&= \mathrm{Var}\left((\boldsymbol{\theta}^\top - \boldsymbol{\beta}_P^\top)\boldsymbol{\mu}\right) + \mathrm{Var}\left(\epsilon_P\right) + \mathrm{Cov}\left((\boldsymbol{\theta}^\top - \boldsymbol{\beta}_P^\top)\boldsymbol{\mu}, \epsilon_P\right) \\
&= (\boldsymbol{\theta}^\top - \boldsymbol{\beta}_P^\top)\boldsymbol{\mu}(\boldsymbol{\theta} - \boldsymbol{\beta}_P) + \mathrm{Var}(\epsilon_P)
\end{aligned}$$

It thus follows:

$$\begin{aligned}
\mathbb{E}\left[\left(\boldsymbol{\theta}^\top \boldsymbol{\mu} - y_P\right)^2\right] &= \mathbb{E}\left[\left(\boldsymbol{\theta}^\top \boldsymbol{\mu} - y_P\right)^2\right]^2 + \mathrm{Var}\left(\boldsymbol{\theta}^\top \boldsymbol{\mu} - y_P\right) \\
&= \left((\boldsymbol{\theta}^\top - \boldsymbol{\beta}_P^\top)\boldsymbol{\mu}\right)^2 + (\boldsymbol{\theta}^\top - \boldsymbol{\beta}_P^\top)\boldsymbol{\mu}(\boldsymbol{\theta} - \boldsymbol{\beta}_P) + \mathrm{Var}(\epsilon_P)
\end{aligned}$$

This result is accurate as it shows when $\boldsymbol{\theta} = \boldsymbol{\beta}_P$ the expectation is equal to the variance in $\boldsymbol{\beta}_P$.

$$\Xi = \Xi$$

Therefore we can find the expectation over all of the data with:

$$\mathbb{E}\left[\boldsymbol{\theta}^\top \boldsymbol{\mu} - y\right] = (1-\varepsilon)(\boldsymbol{\theta}^\top - \boldsymbol{\beta}_P^\top)\boldsymbol{\mu} + \varepsilon(\boldsymbol{\theta}^\top - \boldsymbol{\beta}_Q^\top)\boldsymbol{\mu}$$

Furthermore, we can find the variance over all of the data with:

$$\begin{aligned}
\mathrm{Var}(\boldsymbol{\theta}^\top \boldsymbol{\mu} - y) = (1-\varepsilon)&\left(\left((\boldsymbol{\theta}^\top - \boldsymbol{\beta}_P^\top)\boldsymbol{\mu}\right)^2 + (\boldsymbol{\theta}^\top - \boldsymbol{\beta}_P^\top)\boldsymbol{\mu}(\boldsymbol{\theta} - \boldsymbol{\beta}_P) + \mathrm{Var}(\epsilon_P)\right) \\
+\varepsilon&\left(\left((\boldsymbol{\theta}^\top - \boldsymbol{\beta}_Q^\top)\boldsymbol{\mu}\right)^2 + (\boldsymbol{\theta}^\top - \boldsymbol{\beta}_Q^\top)\boldsymbol{\mu}(\boldsymbol{\theta} - \boldsymbol{\beta}_Q) + \mathrm{Var}(\epsilon_Q)\right)
\end{aligned}$$

Furthermore, this is empirically shown to be the correct derivation by optimal error in Figure **??** converging to 0.1 which we see in Appendix F is equal to ϵ_P. Therefore we can similarly say,

$$\mathbb{E}\left[\left\|\boldsymbol{\theta}^\top\boldsymbol{\mu} - y_Q\right\|_2^2\right] = \left((\boldsymbol{\theta}^\top\boldsymbol{\mu}) - (\boldsymbol{\beta}_Q^\top\boldsymbol{\mu})\right)^2 + \boldsymbol{\theta}^\top\boldsymbol{\Sigma}\boldsymbol{\theta} - 2\boldsymbol{\theta}^\top\boldsymbol{\Sigma}\boldsymbol{\beta}_Q + \boldsymbol{\beta}^\top\boldsymbol{\Sigma}\boldsymbol{\beta} + \mathrm{Var}(\epsilon_Q) \tag{87}$$

By assumption 3 we assume normal error, thus we are now interested in calculating the variance. Therefore we assume the distribution of squared errors is a non-central chi-squared distribution

Now that we have the Expected Value and Variance of Squared Error we now want to calculate the expected number of points in the p sub-quantile. This is equivalent to:

$$n(1-\varepsilon)\mathbb{P}\left[\left\|\boldsymbol{\theta}^\top\boldsymbol{\mu} - y_P\right\|_2^2 \leq \mathcal{Q}_p\right] \tag{88}$$

We thus want to calculate the p-quantile of the Joint Distribution. We will define $\mathbb{Z} = (1+\varepsilon)\mathbb{P} + \varepsilon\mathbb{Q}$. It thus follows:

$$\begin{aligned}\mathbb{E}\left[\mathbb{Z}\right] &= (1-\varepsilon)\mathbb{E}\left[\mathbb{P}\right] + \varepsilon\mathbb{E}\left[\mathbb{Q}\right] \\ \mathrm{Var}(\mathbb{Z}) &= (1-\varepsilon)^2\mathbb{E}\left[\mathbb{P}\right] + \varepsilon^2\mathbb{E}\left[\mathbb{Q}\right]\end{aligned}$$

By assumption 3 we will assume $\mathbb{Z}$ follows a chi-squared distribution. We can estimate the parameter k by the following:

$$k = \frac{\mathbb{E}\left[\mathbb{Z}\right]}{2\mathrm{Var}(\mathbb{Z})} \tag{89}$$

It thus follows by the definition of the Chi-Squared Distribution:

$$\mathcal{Q}_p = \inf\left\{x \in \mathbb{R} : p \leq \frac{1}{\Gamma(k/2)}\gamma\left(\frac{k}{2}, \frac{x}{2}\right)\right\} \tag{90}$$

This is efficiently calculated with computational solvers.
Since we assume $\mathbb{P}$ and $\mathbb{Q}$ are also chi-squared distribution.

$$\mathbb{P}\left[\left\|\boldsymbol{\theta}^\top\boldsymbol{\mu} - y_P\right\|_2^2\right] \leq \mathcal{Q}_p \tag{91}$$

By definition of the chi-squared distribution, this is equivalent to the CDF of chi-squared with

$$k = \frac{\mathbb{E}\left[\mathbb{P}\right]}{2\mathrm{Var}(\mathbb{P})} \tag{92}$$

$$\left(\mathbb{P}\left[\left\|\boldsymbol{\theta}^\top\boldsymbol{\mu} - y_P\right\|_2^2\right] \leq \mathcal{Q}_p\right) = \frac{1}{\Gamma(k/2)}\gamma\left(\frac{k}{2}, \frac{\mathcal{Q}_p}{2}\right) \tag{93}$$

We test our theoretical results in figure ... □

D.2 Proof of Theorem 4

Proof. Note Assumption 2 tells us the whole distribution is sampled from the same distribution. Let us define two functions for the empirical loss on $\mathbb{P}$ and $\mathbb{Q}$

$$\phi(\boldsymbol{\theta}) = \frac{1}{np}\sum_{i=1}^{m}(\boldsymbol{\theta}^\top\boldsymbol{p}_i - y_i)^2 \tag{94}$$

$$\psi(\boldsymbol{\theta}) = \frac{1}{np}\sum_{i=1}^{l}(\boldsymbol{\theta}^\top\boldsymbol{q}_i - y_i)^2 \tag{95}$$

These two functions hold nice properties.

$$\nabla_{\boldsymbol{\theta}}\phi(\boldsymbol{\theta}) = \frac{1}{np}\sum_{i=1}^{m}2\boldsymbol{p}_i(\boldsymbol{\theta}^\top\boldsymbol{p}_i - y_i) \tag{96}$$

$$\nabla_{\boldsymbol{\theta}}\psi(\boldsymbol{\theta}) = \frac{1}{np}\sum_{i=1}^{l}2\boldsymbol{q}_i(\boldsymbol{\theta}^\top\boldsymbol{q}_i - y_i) \tag{97}$$

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

$$\nabla_{\boldsymbol{\theta}_k}g(t_{k+1}, \boldsymbol{\theta}_k) = \nabla_{\boldsymbol{\theta}_k}\phi(\boldsymbol{\theta}) + \nabla_{\boldsymbol{\theta}_k}\psi(\boldsymbol{\theta}) \tag{98}$$

For a given optimization step, we will calculate the expected derivatives.
Recalculate expectation of derivative, should have the covariance matrix of $\boldsymbol{p}$

$$\mathbb{E}\left[\nabla_{\boldsymbol{\theta}}\phi(\boldsymbol{\theta})\right] = \frac{1}{np}\sum_{i=1}^{np(1-\varepsilon)} 2\mathbb{E}\left[\boldsymbol{p}_i(\boldsymbol{\theta}^\top \boldsymbol{p}_i - y_i)\right] \tag{99}$$

$$= \frac{2}{np}\sum_{i=1}^{np(1-\varepsilon} \mathbb{E}\left[\boldsymbol{p}_i\boldsymbol{p}_i^\top\boldsymbol{\theta} - \boldsymbol{p}_i(\boldsymbol{\beta}_P^\top\boldsymbol{p}_i + \epsilon_P)\right] \tag{100}$$

$$= \frac{2}{np}\sum_{i=1}^{np(1-\varepsilon)} \left(\boldsymbol{\mu}\boldsymbol{\mu}^\top + \boldsymbol{\Sigma}_P\right)\boldsymbol{\theta} - \left(\boldsymbol{\mu}\boldsymbol{\mu}^\top + \boldsymbol{\Sigma}_P\right)\boldsymbol{\beta}_P \tag{101}$$

$$= 2(1-\varepsilon)\left(\boldsymbol{\mu}\boldsymbol{\mu}^\top + \boldsymbol{\Sigma}_P\right)(\boldsymbol{\theta} - \boldsymbol{\beta}_P) \tag{102}$$

It thus similarly follows

$$\mathbb{E}\left[\nabla_{\boldsymbol{\theta}}\psi(\boldsymbol{\theta})\right] = 2(\varepsilon)\left(\boldsymbol{\mu}\boldsymbol{\mu}^\top + \boldsymbol{\Sigma}_Q\right)(\boldsymbol{\theta} - \boldsymbol{\beta}_Q) \tag{103}$$

Since we assume both $\mathbb{P}$ and $\mathbb{Q}$ are sampled from the same distribution. It follows $\boldsymbol{\Sigma}_P = \boldsymbol{\Sigma}_Q$, therefore we can calculate the expected derivative with respect to theta as:

$$\mathbb{E}\left[\nabla_{\boldsymbol{\theta}}g(t_{k+1}, \boldsymbol{\theta}_k)\right] = 2\left(\boldsymbol{\mu}\boldsymbol{\mu}^\top + \boldsymbol{\Sigma}\right)(2\boldsymbol{\theta} - (1-\varepsilon)\boldsymbol{\beta}_P - \varepsilon\boldsymbol{\beta}_Q)$$

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.

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

We can now use the expected derivative in equation 102

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

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

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

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

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

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

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

Note $\mathbb{E}\left[\boldsymbol{\mu}^\top\boldsymbol{\mu}\right] = \mathrm{Var}\left[\boldsymbol{\mu}\right] + \boldsymbol{\mu}^2 = C$ 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.

We are thus interested in $\mathbb{P}[\epsilon \to 0]$. The fact that $\boldsymbol{\theta}_0 = \mathbf{0}$ is vital to the convergence analysis. Furthermore, the coefficients of $\boldsymbol{\beta}_P$ and $\boldsymbol{\beta}_Q$ are also important for this analysis. This is why for the first iteration, it is vital not to randomize $\boldsymbol{\theta}_0$, but to randomize the first np elements trained on. As initial $\boldsymbol{\theta}_0$ is deterministic of the final distribution.

Definition 8. *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 8.

$$\mathbb{E}[g(t_{k+1}, \boldsymbol{\theta}_{k+1})] \leq \mathbb{E}[g(t, \boldsymbol{\theta}_k)] + \langle \nabla_{\boldsymbol{\theta}} \mathbb{E}[g(t_k, \boldsymbol{\theta}_k)], \boldsymbol{\theta}_{k+1} - \boldsymbol{\theta}_k \rangle + \frac{L}{2}\|\boldsymbol{\theta}_{k+1} - \boldsymbol{\theta}_k\|_2^2 \tag{119}$$

$$= \mathbb{E}[g(t, \boldsymbol{\theta}_k)] + \langle \nabla_{\boldsymbol{\theta}}\phi(\boldsymbol{\theta}) + \nabla_{\boldsymbol{\theta}}\psi(\boldsymbol{\theta}), \boldsymbol{\theta}_{k+1} - \boldsymbol{\theta}_k \rangle + \frac{L}{2}\|\boldsymbol{\theta}_{k+1} - \boldsymbol{\theta}_k\|_2^2 \tag{120}$$

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

$$= \mathbb{E}[g(t, \boldsymbol{\theta}_k)] + \langle \nabla_{\boldsymbol{\theta}}\phi(\boldsymbol{\theta}) + \nabla_{\boldsymbol{\theta}}\psi(\boldsymbol{\theta}), -\frac{1}{L}(\nabla_{\boldsymbol{\theta}}\phi(\boldsymbol{\theta}_k) + \nabla_{\boldsymbol{\theta}}\psi(\boldsymbol{\theta}_k)) \rangle + \frac{L}{2}\|\phi(\boldsymbol{\theta}) + \nabla_{\boldsymbol{\theta}}\psi(\boldsymbol{\theta})\|_2^2 \tag{121}$$

We can now use the expected derivative described in Equation 102

$$= \mathbb{E}[g(t, \boldsymbol{\theta}_k)] + \langle 2\boldsymbol{\mu}\boldsymbol{\mu}^\top\left(\boldsymbol{\theta}^\top - (1-\epsilon)\boldsymbol{\beta}_P^\top - \epsilon\boldsymbol{\beta}_Q^\top\right), -\frac{1}{L}2\boldsymbol{\mu}\boldsymbol{\mu}^\top\left(\boldsymbol{\theta}^\top - (1-\epsilon)\boldsymbol{\beta}_P^\top - \epsilon\boldsymbol{\beta}_Q^\top\right) \rangle + \frac{1}{2L}\left(2\boldsymbol{\mu}\boldsymbol{\mu}^\top\left(\boldsymbol{\theta}^\top - (1-\epsilon)\boldsymbol{\beta}_P^\top - \epsilon\boldsymbol{\beta}_Q^\top\right)\right)^\top\left(2\boldsymbol{\mu}\boldsymbol{\mu}^\top\left(\boldsymbol{\theta}^\top - (1-\epsilon)\boldsymbol{\beta}_P^\top - \epsilon\boldsymbol{\beta}_Q^\top\right)\right) \tag{122}$$

$$= \mathbb{E}[g(t, \boldsymbol{\theta}_k)] - \frac{1}{2L}\left(2\boldsymbol{\mu}\boldsymbol{\mu}^\top\left(\boldsymbol{\theta}^\top - (1-\epsilon)\boldsymbol{\beta}_P^\top - \epsilon\boldsymbol{\beta}_Q^\top\right)\right)^\top\left(2\boldsymbol{\mu}\boldsymbol{\mu}^\top\left(\boldsymbol{\theta}^\top - (1-\epsilon)\boldsymbol{\beta}_P^\top - \epsilon\boldsymbol{\beta}_Q^\top\right)\right) \tag{123}$$

□

D.3 Proof of Theorem 5

Proof. First we will introduce some notation. Let $S \subseteq X$ such that $|S| = p|X|$ represent the np points within X that have the lowest error with respect to $\boldsymbol{\theta}$. To formalize this in our optimization, let $\boldsymbol{A}_k$ be the matrix whose rows represent all $s \in S_k$. Now we can focus on a one-step optimization problem

$$\boldsymbol{\theta}_{k+1} = \boldsymbol{\theta}_k - \alpha\frac{1}{np}\sum_{i \in S}\nabla f_i(\boldsymbol{\theta}_k) \tag{124}$$

□

E Additional Experiments

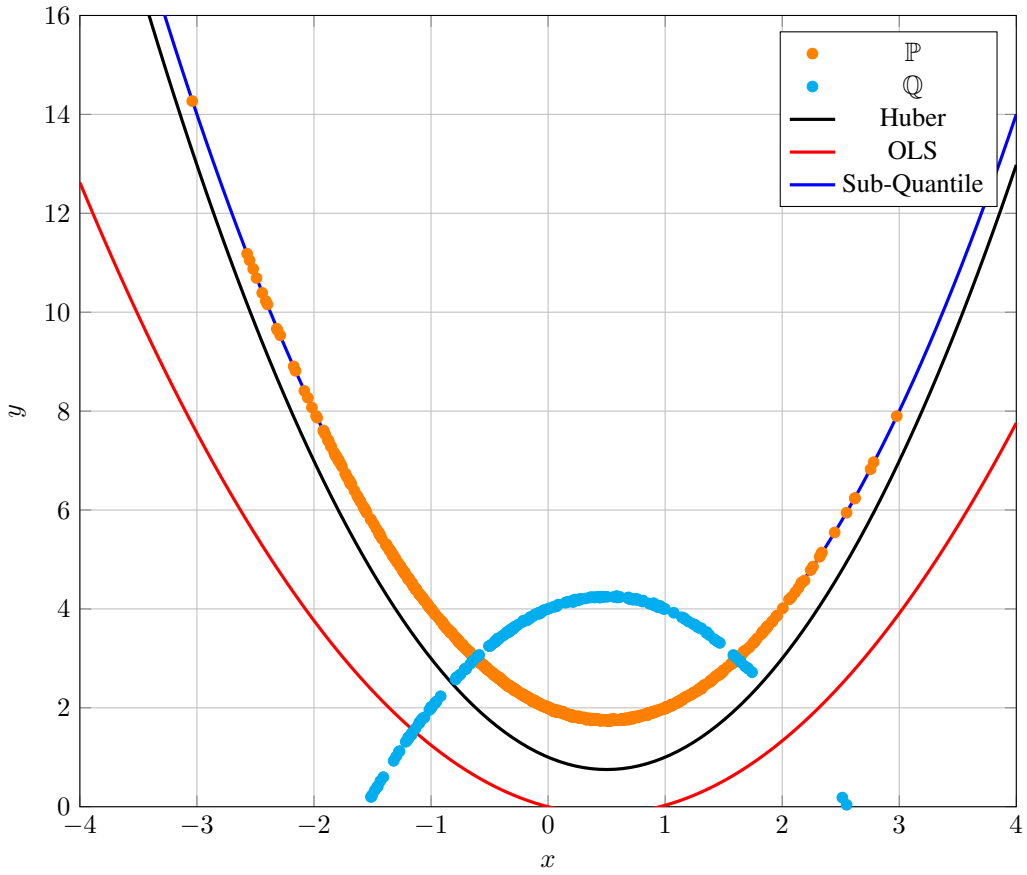


Figure 4: `Quadratic Regression` $n = 1000$ and $\epsilon = 0.2$

Objectives	Test RMSE (`Quadratic Regression`)		
	$\epsilon = 0$	$\epsilon = 0.2$	$\epsilon = 0.4$
OLS 125	$0.0099_{(0.0002)}$	$2.078_{(0.146)}$	$4.104_{(0.442)}$
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 3: `Quadratic Regression` Synthetic Dataset. Empirical Risk over $\mathbb{P}$

F Experimental Details

F.1 Structured Linear Regression Dataset

We will describe $\mathbb{P}$ and $\mathbb{Q}$ in the `Structured Linear Regression` Dataset.
$\boldsymbol{x} \sim \mathcal{N}(4,4)^{100}$
$\boldsymbol{m} \sim \mathcal{N}(4,4)^{100}$
$b \sim \mathcal{N}(4,4)$
$\boldsymbol{m}' \sim \mathcal{N}(4,4)^{100}$
$b' \sim \mathcal{N}(4,4)$
$n_{\text{train}} = 2\text{e}3$
$\mathbb{P} : y|\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{m}^\top \boldsymbol{x} + b, 0.1)$
$\mathbb{Q} : y|\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{m}'^\top \boldsymbol{x} + b', 0.1)$
Please note $\boldsymbol{m}$, b, $\boldsymbol{m}'$, b', are all sampled independently.

F.2 Noisy Linear Regression Dataset

We will describe $\mathbb{P}$ and $\mathbb{Q}$ in the `Noisy Linear Regression` Dataset.
$\boldsymbol{x} \sim \mathcal{N}(0,4)^{100}$
$\boldsymbol{m} \sim \mathcal{N}(0,4)^{100}$
$b \sim \mathcal{N}(0,4)$
$\boldsymbol{m}' = \mathbf{0}$
$b' \sim \mathcal{N}(4,4)$
$n_{\text{train}} = 2\text{e}3$
$\mathbb{P} : y|\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{m}^\top \boldsymbol{x} + b, 0.1)$
$\mathbb{Q} : y|\boldsymbol{x} \sim \mathcal{N}(b', 4)$
Please note $\boldsymbol{m}$, b, $\boldsymbol{m}'$, b', are all sampled independently.

F.3 Quadratic Regression Dataset

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

F.4 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 $\mathcal{N}(5,5)$.
The noise represents a noisy worker

F.5 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.

$$\boldsymbol{X}^* = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{y} \tag{125}$$

Huber Regression is solved with the following objective function.

$$L_\delta(y, f(\boldsymbol{x})) = \begin{cases} \frac{1}{2}(y - f(\boldsymbol{x}))^2 \\ \delta \cdot \left(|y - f(\boldsymbol{x})| - \frac{1}{2}\delta\right) \quad \text{otherwise} \end{cases} \tag{126}$$

RANSAC