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, P when there exists contaminated samples, Q. We consider the Huber Contamination modeled as $\hat{P} = (1-\varepsilon)P + \varepsilon Q$ where $\varepsilon \in (0, 0.5)$. Traditional Least Squares Methods fit the empirical risk model to all training data in $\hat{P}$. 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. Sub-Quantile Optimization theoretically and empirically works in the case of both oblivious and adversarial outliers.

1 INTRODUCTION

Linear Regression is one of the most widely used statistical estimators throughout science. Robustness Learning in High Dimensions on Huber Contamination Models, Huber & Ronchetti (2009), has gained much attention in the last decade, Diakonikolas & Kane (2019). 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 in numerous datasets, including those featuring data in the medical, economic, and meteorological fields. Ultimately, in many real data sets, the samples may not be collected from even or fair distributions; thus, classical analyses such as standard regression or least-squares may not represent the actual distribution of the data well.

The quantile is a statistical measure that is distribution-agnostic, this makes it very suitable for robust estimation in the Huber Contamination Model.

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

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

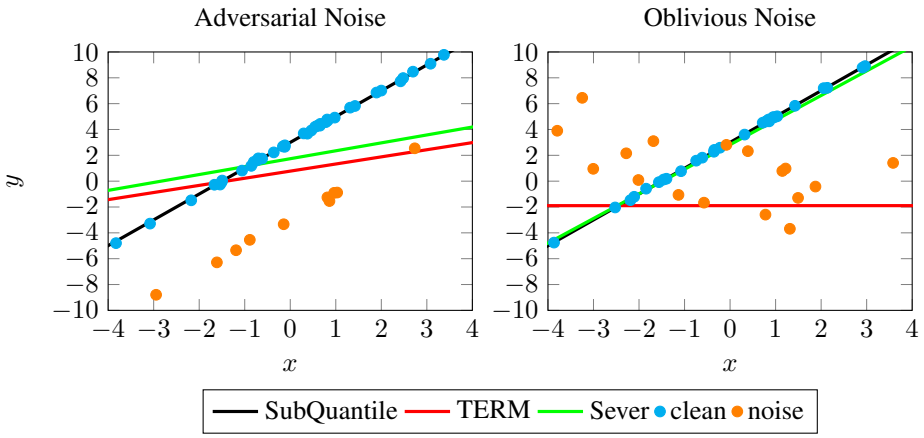

Figure 1: Sub-Quantile Performance on Adapative Outliers

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. In Table

Paper	Adversary	Threshold
Sever Diakonikolas et al. (2019)	Adaptive	Gradient of Loss
CRR Bhatia et al. (2017)	Oblivious	
This Paper	Adapative	Loss

Table 1: A comparison of different iterative thresholding algorithms for Robust Least Squares Regression

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 known as tilt

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

By using the tilt hyperparameter to change the individual impact of each specific loss, the model is more resistant to outliers found in the data.

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.

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

This estimator is proven to return near-optimal risk on a variety of linear models, including Gaussian regression, Poisson regression, and binomial regression; these achievements can be demonstrated on label and covariate corruptions.

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

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

This method is novel in that it is highly scalable, making it robust against high-dimension data with structured outliers. Similarly, SEVER is easily implemented with standard machine learning libraries and can be applied to many typical learning problems, including classification and regression. Despite this, the algorithm still falls short when features have high covariance or when features have low predictive power of the target. Moreover, SEVER requires approximate learners to be run after every iteration, making SEVER unfeasible for large-scale machine learning tasks.

Quantile Regression Yu et al. (2003) relies on splitting data into quantiles to better represent data that is not evenly distributed. The paper introduces various estimation methods for quantile regression and apply them to a multitude of datasets. In doing so, they prove quantile regression is suitable at estimating both linear and nonlinear response models.

Super-Quantile Optimization Rockafellar et al. (2014) aims to solve error minimization problems by building upon the aforementioned quantile regression by centering around a conditional value-at-risk, or a superquantile. For $\alpha \in [0, 1)$, the α-superquantile for a random variable Y is defined as

$$\bar{q}_\alpha(Y) := \frac{1}{1-\alpha}\int_\alpha^1 q_\beta(Y) d\beta \tag{4}$$

In doing so, more conservatively fitted curves are produced. As with quantile regression, such curves do require the solution of a linear program. This concept of superquantile error provides insight into tail behavior for quantities of error and an overall unique approach to linear regression.

Robust Risk Minimization Osama et al. (2020) is a method in which given an upper bound on the corrupted data fraction ϵ, the risk function can be minimized as follows:

$$\hat{\theta}_{RRM} = \operatorname*{argmin}_{\theta\in\Theta} \min_{\pi\in\Pi:\mathbf{H}(\pi)\geq ln|(1-\tilde{\epsilon})n|} R(\theta, \pi) \tag{5}$$

This method is popular as it does not require the removal of corrupted data points and does not rely on a specified corruption fraction.

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*

$$Q_p(p) = \inf\{x \in \mathbb{R} : p \leq F(x)\} \tag{6}$$

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

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

$$\hat{F}_n(t) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{X_i\leq t} \tag{7}$$

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

$$U = \mathbb{E}\left[\ell\left(f(\boldsymbol{x}; \boldsymbol{\theta}, \boldsymbol{y})\right)\right] \tag{8}$$

The **p-Quantile** of the Empirical Risk is given

$$\mathbb{L}_p(U) = \frac{1}{p}\int_0^p \mathcal{Q}_q(U)\, dq = \mathbb{E}\left[U|U \leq \mathcal{Q}_p(U)\right] = \max_{t\in\mathbb{R}}\left\{t - \frac{1}{p}\mathbb{E}\left[(t-U)^+\right]\right\} \tag{9}$$

In equation 9, 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

$$\boldsymbol{\theta}_{SM} = \arg\min_{\boldsymbol{\theta}\in\mathbb{R}^d} \max_{t\in\mathbb{R}} \left\{ t - \frac{1}{p}\mathbb{E}(t - \ell(f(\boldsymbol{x};\boldsymbol{\theta}), y))^+ \right\} \tag{10}$$

For the linear regression case, this equation becomes

$$\boldsymbol{\theta}_{SM} = \arg\min_{\boldsymbol{\theta}\in\mathbb{R}^d} \max_{t\in\mathbb{R}} \left\{ t - \frac{1}{np}\sum_{i=1}^{n}(t - (\boldsymbol{\theta}^\top \boldsymbol{x}_i - y_i)^2)^+ \right\} \tag{11}$$

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

$$t_{k+1} = \arg\max_t g(t, \boldsymbol{\theta}_k) \tag{12}$$

$$\boldsymbol{\theta}_{k+1} = \boldsymbol{\theta}_k + \alpha \nabla_{\boldsymbol{\theta}_k} g(t, \boldsymbol{\theta}_k) \tag{13}$$

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*

$$\boldsymbol{y}_P = \boldsymbol{\beta}_P^\top \boldsymbol{P} + \epsilon_P \tag{14}$$

and similarly

$$\boldsymbol{y}_Q = \boldsymbol{\beta}_Q^\top \boldsymbol{Q} + \epsilon_Q \tag{15}$$

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 $\hat{P}$. 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.*

$$\boldsymbol{P}_i, \boldsymbol{Q}_j \sim \mathcal{N}_p(\boldsymbol{0}, \boldsymbol{\Sigma}) \tag{16}$$

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

$$\mathbb{E}\left[\boldsymbol{X}^\dagger\boldsymbol{y}\right] = \boldsymbol{\beta}_P + \epsilon(\boldsymbol{\beta}_Q - \boldsymbol{\beta}_P) \tag{17}$$

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

$$\mathbb{E}\left[\boldsymbol{X}^{\dagger}\boldsymbol{y}\right] = \boldsymbol{\beta}_P - n(1-\epsilon)\boldsymbol{\Sigma}_P^{-1}\boldsymbol{\beta}_Q \tag{18}$$

The proof can be found in Appendix A.2.

In equation 17, 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 2.1. *$g(t_{k+1}, \boldsymbol{\theta}_k)$ 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 2.2. *The maximizing value of t in $g(t, \boldsymbol{\theta})$ in t-update step of optimization as described by Equation 12 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 $0 < p < 1$

$$\partial(-g(t, \boldsymbol{\theta})) = -1 + \frac{1}{np}\sum_{i=1}^{n}\left\{\begin{array}{ll} 1, & \text{if } t > \hat{\boldsymbol{\nu}}_i \\ 0, & \text{if } t < \hat{\boldsymbol{\nu}}_i \\ [0,1], & \text{if } t = \hat{\boldsymbol{\nu}}_i \end{array}\right\} \tag{19}$$

$$= 0 \text{ when } t = \hat{\boldsymbol{\nu}}_{np} \tag{20}$$

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

Lemma 2.3. *Let $t = \hat{\boldsymbol{\nu}}_{np}$. The $\boldsymbol{\theta}$-update step described in Equation 11 is equivalent to minimizing the least squares loss of the np elements with the lowest squared loss.*

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

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

Proof. We see by lemma 2.2 and interpretation 1, we are optimizing by the np points with the lowest squared error. Mathematically,

$$g(t_{k+1}, \boldsymbol{\theta}_k) = t_{k+1} - \frac{1}{np} \sum_{i=1}^{n} \left(t_{k+1} - (\boldsymbol{\theta}^\top \boldsymbol{x}_i - y_i)^2\right)^+ \tag{22}$$

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

$$= t - t + \frac{1}{np} \sum_{i=1}^{np} (\boldsymbol{\theta}^\top \boldsymbol{x}_i - y_i)^2 \tag{24}$$

$$= \frac{1}{np} \sum_{i=1}^{np} (\boldsymbol{\theta}^\top \boldsymbol{x}_i - y_i)^2 \tag{25}$$

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 2.5. *$g(t, \boldsymbol{\theta})$ 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 12 and 13, respectively.

Lemma 2.6. *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 $(t, \boldsymbol{\theta}) \in \mathbb{R} \times \mathbb{R}^d$*

$$g(t^*, \boldsymbol{\theta}) \leq g(t^*, \boldsymbol{\theta}^*) \leq g(t, \boldsymbol{\theta}^*) \tag{26}$$

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:*

$$g(t^*, \boldsymbol{\theta}) \leq g(t^*, \boldsymbol{\theta}^*) \leq g(t, \boldsymbol{\theta}^*) \tag{27}$$

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 2.2, we have shown $\nabla_t g(t, \boldsymbol{\theta}) = 0$ when $\boldsymbol{\nu}_{np} \leq t < \boldsymbol{\nu}_{np+1}$. Furthermore, by lemma 2.3, 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.

Now we come to one of the most important theorems of this paper.

Theorem 3. *In the case of linear regression, SubQuantile Minimization will converge to a local or global minimum.*

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 $\boldsymbol{\beta}_Q = \mathbf{0}$.

Lemma 3.1. *The expected value of error on points in* $\mathbb{P}$ *will be lower than the expected value of error on points in* $\mathbb{Q}$ *if* $\text{proj}_{\boldsymbol{\beta}_P}(\boldsymbol{\theta}) - \boldsymbol{\beta}_P < \text{proj}_{\boldsymbol{\beta}_Q}(\boldsymbol{\theta}) - \boldsymbol{\beta}_Q$

Lemma 3.1 gives us an intuitive result, the proof is in appendix C.1. If in each optimization step, our projection on $\boldsymbol{\beta}_P$ is closer than our projection on to $\boldsymbol{\beta}_Q$, we know the number of steps from $\mathbb{Q}$ will increase from the previous iteration.

Theorem 4. *After an optimization step, there will be, by expectation, more elements from* $\mathbb{P}$ *in the SubQuantile Matrix than in the previous matrix if*

$$(1 - \varepsilon - \alpha_1)\boldsymbol{\beta}_P > (\varepsilon - \alpha_2)\boldsymbol{\beta}_Q \tag{28}$$

where α_1 *and* α_2 *represents the coefficients for the linear combination of* $\boldsymbol{\theta}$ *in the basis defined as* $\boldsymbol{B} = [\boldsymbol{\beta}_P \quad \boldsymbol{\beta}_Q \quad \boldsymbol{R}]$

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. Each time step we add a $\mathcal{O}(d)$ partitioning step, thus total added complexity would be $T\mathcal{O}(d)$.

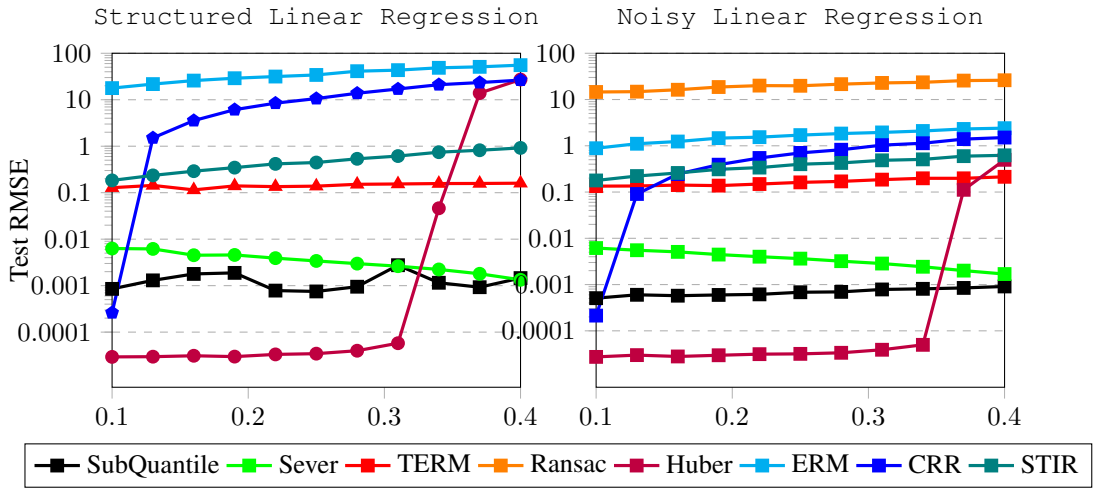

Figure 2: `Structured Linear Regression` & `Noisy Linear Regression` Datasets

5 Empirical Results

Algorithm 1: Sub-Quantile Minimization Optimization Algorithm

Input: Training iterations T, Quantile p, Corruption Percentage ϵ, Input Parameters m, SubQuantile Update: j

Output: Trained Parameters, $\boldsymbol{\theta}$

$\boldsymbol{\theta}_0 \leftarrow (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{y}$

for $k \in 1, 2, \ldots, m$ **do**

  compute the loss over the training dataset $\boldsymbol{\nu} = (\boldsymbol{X}\boldsymbol{\theta}_k - \boldsymbol{y})^2$

  if $k \% j = 0$ **then**

    $\hat{\boldsymbol{\nu}} \leftarrow sorted(\boldsymbol{\nu})$

    update the subquantile matrix $t_{k+1} \leftarrow \hat{\boldsymbol{\nu}}_{np}$

  end

  calculate the Lipschitz Constant of the Subquantile Matrix $L \leftarrow \frac{1}{np} \left\| \boldsymbol{S}^\top \boldsymbol{S} \right\|_2$

  $\alpha \leftarrow \frac{1}{2L}$

  gradient descent on theta $\boldsymbol{\theta}_{k+1} \leftarrow \boldsymbol{\theta}_k - \alpha \nabla_{\boldsymbol{\theta}_k} g(t_{k+1}, \boldsymbol{\theta}_k)$

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.

5.1 Synthetic Data

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

From the results we can see in Figure 2, 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

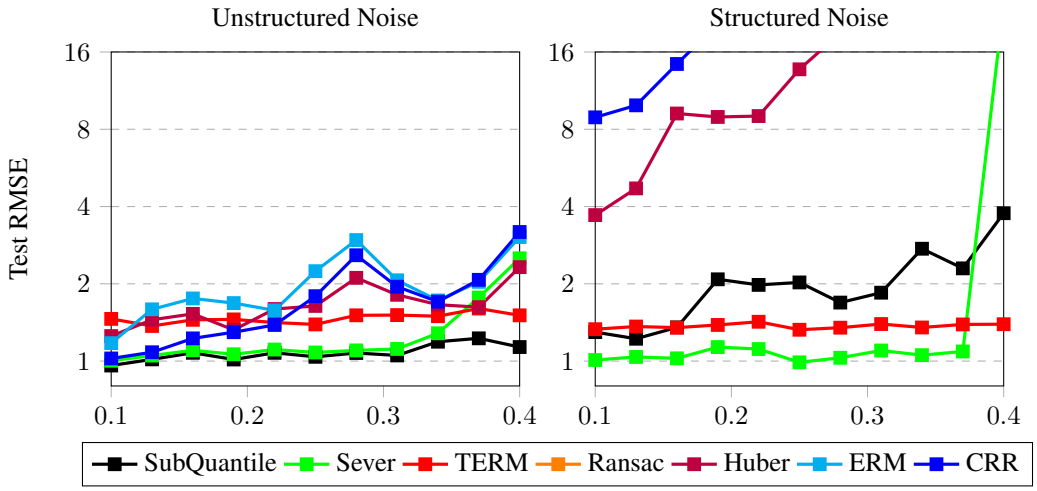

Figure 3: `Drug Discovery` Dataset with Normal Noise and Structured Noise

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 G 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).

Table 2 is up to date. Figure 3 is up to date. We have implemented CRR Bhatia et al. (2017) and STIR Mukhoty et al. (2019).

Objectives	Test RMSE (`Drug Discovery`)			
	$\epsilon = 0.1$	$\epsilon = 0.2$	$\epsilon = 0.3$	$\epsilon = 0.4$
ERM **??**	$1.276_{(0.1240)}$	$1.924_{(0.8236)}$	$2.892_{(1.7534)}$	$2.511_{(1.2485)}$
CRR Bhatia et al. (2017)	$1.087_{(0.1164)}$	$1.552_{(0.5108)}$	$2.619_{(1.3739)}$	$2.590_{(1.1956)}$
STIR Mukhoty et al. (2019)	∞	∞	∞	∞
Huber Huber & Ronchetti (2009)	$1.412_{(0.0474)}$	$1.501_{(0.2918)}$	$2.231_{(0.9054)}$	$2.247_{(1.0399)}$
RANSAC Fischler & Bolles (1981)	∞	∞	∞	∞
TERM Li et al. (2020)	$1.368_{(0.0520)}$	$1.425_{(0.0810)}$	$1.369_{(0.1062)}$	$1.505_{(0.1964)}$
SEVER Diakonikolas et al. (2019)	$1.062_{(0.0891)}$	$1.098_{(0.0485)}$	$1.089_{(0.1119)}$	$2.318_{(0.8228)}$
SubQuantile($p = 1 - \epsilon$)	$\mathbf{1.023_{(0.1037)}}$	$\mathbf{1.059_{(0.0609)}}$	$\mathbf{1.049_{(0.0817)}}$	$\mathbf{1.257_{(0.2540)}}$
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 4 demonstrate the percentage of elements from $\mathbb{Q}$ in the final subquantile matrix $\boldsymbol{S}_T$ from different theta initializations after T iterations of subquantile minimization in the `Noisy`

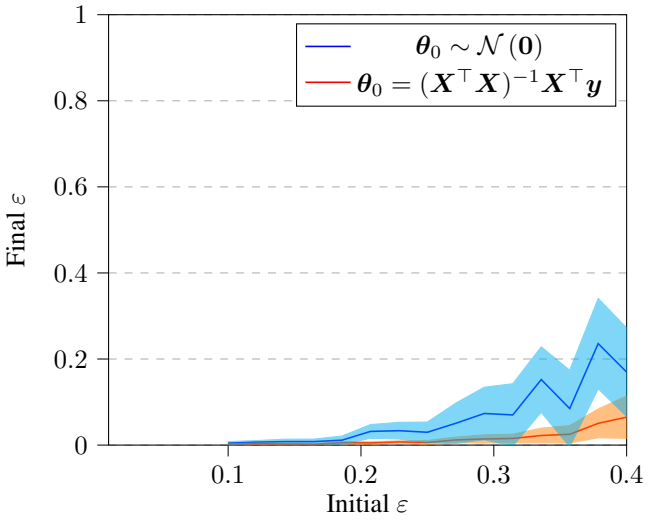

Figure 4: Probability of points from $\mathbb{Q}$ in the final subquantile

`Linear Regression` Dataset. 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. Since minimizing over the sub-quantile is a general machine learning framework, it is scalable to larger scale machine learning problems. In future work, more real world applications can be explored and the theory can be expanded beyond linear regression.

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.

Kush Bhatia, Prateek Jain, Parameswaran Kamalaruban, and Purushottam Kar. Consistent robust regression. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/e702e51da2c0f5be4dd354bb3e295d37-Paper.pdf.

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 and Daniel M. Kane. Recent advances in algorithmic high-dimensional robust statistics. *ArXiv*, abs/1911.05911, 2019.

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.

Dheeru Dua and Casey" Graff. UCI machine learning repository, 2017. URL http://archive.ics.uci.edu/ml.

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.

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.

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.

Bhaskar Mukhoty, Govind Gopakumar, Prateek Jain, and Purushottam Kar. Globally-convergent iteratively reweighted least squares for robust regression problems. In Kamalika Chaudhuri and Masashi Sugiyama (eds.), *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pp. 313–322. PMLR, 16–18 Apr 2019. URL https://proceedings.mlr.press/v89/mukhoty19a.html.

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.

Kelley Pace and Ronald Barry. Sparse spatial autoregressions. *Statistics 'I& Probability Letters*, 33(3):291–297, 1997. URL https://EconPapers.repec.org/RePEc:eee:stapro:v:33:y:1997:i:3:p:291-297.

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.

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

A.1 PROOF OF THEOREM 1

Proof.
We will first calculate the pseudo-inverse

$$\boldsymbol{X}^\top \boldsymbol{X} = \begin{pmatrix} \boldsymbol{P}^\top & \boldsymbol{Q}^\top \end{pmatrix} \begin{pmatrix} \boldsymbol{P} \\ \boldsymbol{Q} \end{pmatrix} \tag{29}$$

$$= \boldsymbol{P}^\top \boldsymbol{P} + \boldsymbol{Q}^\top \boldsymbol{Q} \tag{30}$$

Now we can calculate the Moore-Penrose Inverse

$$(\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top = (\boldsymbol{P}^\top \boldsymbol{P} + \boldsymbol{Q}^\top \boldsymbol{Q})^{-1} \begin{pmatrix} \boldsymbol{P}^\top & \boldsymbol{Q}^\top \end{pmatrix} \tag{31}$$

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

Now we solve for the optimal model

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

$$= (\boldsymbol{P}^\top \boldsymbol{P} + \boldsymbol{Q}^\top \boldsymbol{Q})^{-1} \boldsymbol{P}^\top \boldsymbol{y}_P + (\boldsymbol{P}^\top \boldsymbol{P} + \boldsymbol{Q}^\top \boldsymbol{Q})^{-1} \boldsymbol{Q}^\top \boldsymbol{y}_Q \tag{34}$$

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).

$$\boldsymbol{P}^\top \boldsymbol{P} = \sum_{i=1}^{n*(1-\epsilon)} \boldsymbol{P}_i \boldsymbol{P}_i^\top \tag{35}$$

$$\boldsymbol{Q}^\top \boldsymbol{Q} = \sum_{j=1}^{n\epsilon} \boldsymbol{Q}_j \boldsymbol{Q}_j^\top \tag{36}$$

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

$$\boldsymbol{P}^\top \boldsymbol{P} \sim \mathcal{W}(n(1-\epsilon), \boldsymbol{\Sigma}) \tag{37}$$

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

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

$$\mathbb{E}(\boldsymbol{P}^\top \boldsymbol{P}) = n(1-\epsilon)\boldsymbol{\Sigma} \tag{39}$$

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

It thus follows

$$\mathbb{E}\left[\boldsymbol{P}^\top \boldsymbol{P} + \boldsymbol{Q}^\top \boldsymbol{Q}\right] = n\boldsymbol{\Sigma} \tag{41}$$

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

$$\left(\boldsymbol{P}^\top \boldsymbol{P} + \boldsymbol{Q}^\top \boldsymbol{Q}\right)^{-1} \sim \mathcal{W}^{-1}(n, \boldsymbol{\Sigma}) \tag{42}$$

It thus follows by the expectation of the Inverse Wishart Distribution

$$\mathbb{E}\left[\left(\boldsymbol{P}^\top \boldsymbol{P} + \boldsymbol{Q}^\top \boldsymbol{Q}\right)^{-1}\right] = n\boldsymbol{\Sigma}^{-1} \tag{43}$$

Now we will plug this into Equation 34:

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

$$= \left(n\boldsymbol{\Sigma}^{-1}\right) \boldsymbol{P}^\top (\boldsymbol{P}\boldsymbol{\beta} + \boldsymbol{\epsilon}_P) + \left(n\boldsymbol{\Sigma}^{-1}\right) \boldsymbol{Q}^\top (\boldsymbol{Q}\boldsymbol{\beta}_Q^\top + \boldsymbol{\epsilon}_Q) \tag{45}$$

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

$$= \left(n\boldsymbol{\Sigma}^{-1}\right) \left((n(1-\epsilon)\boldsymbol{\Sigma})\boldsymbol{\beta}_P + n\epsilon\boldsymbol{\Sigma} (\boldsymbol{\beta}_P + \boldsymbol{\Psi})\right) \tag{47}$$

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

$$= \boldsymbol{\beta}_P + \epsilon(\boldsymbol{\Psi}) \tag{49}$$

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 $\mathbb{Q}$.

$$\mathbb{E}\left[\boldsymbol{P}^\top \boldsymbol{P}\right] = n(1-\epsilon)\mathbf{\Sigma}_P \tag{50}$$

$$\mathbb{E}\left[\boldsymbol{Q}^\top \boldsymbol{Q}\right] = n\epsilon\mathbf{\Sigma}_Q \tag{51}$$

It thus follows

$$\mathbb{E}\left[\boldsymbol{P}^\top \boldsymbol{P} + \boldsymbol{Q}^\top \boldsymbol{Q}\right] = (n(1-\epsilon)\mathbf{\Sigma}_P + n\epsilon\mathbf{\Sigma}_Q) \tag{52}$$

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

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

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

$$= n(1-\epsilon)\mathbf{\Sigma}_P^{-1} - n(1-\epsilon)\mathbf{\Sigma}_P^{-1}(n\epsilon\mathbf{\Sigma}_Q^{-1}) \tag{54}$$

We will now calculate the expected optimal parameters by plugging this into Equation 34:

$$\mathbb{E}\left[\boldsymbol{X}^\dagger \boldsymbol{y}\right] = n(1-\epsilon)\mathbf{\Sigma}_P^{-1}(\boldsymbol{P}^T\boldsymbol{P})\boldsymbol{\beta}_P - n(1-\epsilon)\mathbf{\Sigma}_P^{-1}(n\epsilon\mathbf{\Sigma}_Q^{-1})(\boldsymbol{Q}^T\boldsymbol{Q})\boldsymbol{\beta}_Q \tag{55}$$

$$= n(1-\epsilon)\mathbf{\Sigma}_P^{-1}(n(1-\epsilon)\mathbf{\Sigma}_P)\boldsymbol{\beta}_P - n(1-\epsilon)\mathbf{\Sigma}_P^{-1}(n\epsilon\mathbf{\Sigma}_Q^{-1})(n\epsilon\mathbf{\Sigma}_Q)\boldsymbol{\beta}_Q \tag{56}$$

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

This concludes the proof. □

B General Properties of Sub-Quantile Minimization

B.1 Derivation of Lemma 2.2

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 $0 < p < 1$

$$\partial(-g(t, \boldsymbol{\theta})) = \partial\left(-t + \frac{1}{np}\sum_{i=1}^{n}(t - \hat{\boldsymbol{\nu}}_i)^+\right) \tag{58}$$

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

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

$$= -1 + \frac{1}{np}\sum_{i=1}^{n}\left\{\begin{array}{ll} 1, & \text{if } t > \hat{\boldsymbol{\nu}}_i \\ 0, & \text{if } t < \hat{\boldsymbol{\nu}}_i \\ [0, 1], & \text{if } t = \hat{\boldsymbol{\nu}}_i \end{array}\right\} \tag{61}$$

$$= 0 \text{ when } t = \hat{\boldsymbol{\nu}}_{np} \tag{62}$$

This is the p-quantile of $\boldsymbol{\nu}$. Assuming no two points are equal in the dataset, this means the minimizing value for t has a range of values, $\hat{\boldsymbol{\nu}}_{np} \leq t < \hat{\boldsymbol{\nu}}_{np+1}$. This means $g(t, \boldsymbol{\theta})$ is not strongly convex with respect to t.

B.2 Derivation of Lemma 2.3

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

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

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

B.3 Derivation of Lemma 2.5

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

$$||\nabla_{\boldsymbol{\theta}} g(\boldsymbol{\theta}', t) - \nabla_{\boldsymbol{\theta}} g(\boldsymbol{\theta}, t)|| \leq L||\boldsymbol{\theta}' - \boldsymbol{\theta}|| \tag{63}$$

$$\left\|\nabla_{\boldsymbol{\theta}} g(\boldsymbol{\theta}^{'}, t) - \nabla_{\boldsymbol{\theta}} g(\boldsymbol{\theta}, t)\right\| = \left\| \frac{1}{np} \sum_{i=1}^{np} 2\boldsymbol{x}_i(\boldsymbol{\theta}_k^{'\top} \boldsymbol{x}_i - y_i) - \frac{1}{np} \sum_{i=1}^{np} 2\boldsymbol{x}_i(\boldsymbol{\theta}_k^{\top} \boldsymbol{x}_i - y_i) \right\| \tag{64}$$

$$= \left\| \frac{1}{np} \sum_{i=1}^{np} 2\boldsymbol{x}_i(\boldsymbol{\theta}_k^{'\top} \boldsymbol{x}_i - \boldsymbol{\theta}_k^{\top} \boldsymbol{x}_i) \right\| \tag{65}$$

$$= \left\| \frac{1}{np} \sum_{i=1}^{np} 2\boldsymbol{x}_i \boldsymbol{x}_i^{\top}(\boldsymbol{\theta}_k^{'\top} - \boldsymbol{\theta}_k^{\top}) \right\| \tag{66}$$

$$\overset{\text{Cauchy}-\text{Schwarz}}{\leq} \left\| \frac{2}{np} \sum_{i=1}^{np} \boldsymbol{x}_i \boldsymbol{x}_i^{\top} \right\| \left\| \boldsymbol{\theta}_k^{'\top} - \boldsymbol{\theta}_k^{\top} \right\| \tag{67}$$

$$= L \left\| \boldsymbol{\theta}_k^{'\top} - \boldsymbol{\theta}_k^{\top} \right\| \tag{68}$$

where $L = \left\| \frac{2}{np} \boldsymbol{X}^{\top} \boldsymbol{X} \right\|$
This concludes the derivation.

B.4 Proof of Lemma 2.6

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 $t_k \geq \hat{\boldsymbol{\nu}_k}$

$$g(t_k, \boldsymbol{\theta}_k) = t_k - \frac{1}{np} \sum_{i=1}^{n} (t_k - \boldsymbol{\nu}_i)^+ \tag{69}$$

$$= t_k - \frac{1}{np}(np)t_k + \frac{1}{np} \sum_{i=1}^{np} \boldsymbol{\nu}_i + \frac{1}{np} \sum_{i=np}^{n} (t_k - \boldsymbol{\nu}_i)^+ \tag{70}$$

$$= \frac{1}{np} \sum_{i=1}^{np} \boldsymbol{\nu}_i + \frac{1}{np} \sum_{i=np}^{n} (t_k - \boldsymbol{\nu}_i)^+ \tag{71}$$

$$g(t_{k+1}, \boldsymbol{\theta}_k) - g(t_k, \boldsymbol{\theta}_k) = \frac{1}{np} \sum_{i=1}^{np} \boldsymbol{\nu}_i - \left(\frac{1}{np} \sum_{i=1}^{np} \boldsymbol{\nu}_i + \frac{1}{np} \sum_{i=np}^{n} (t_k - \boldsymbol{\nu}_i)^+ \right) \tag{72}$$

$$= -\frac{1}{np} \sum_{i=np}^{n} (t_k - \boldsymbol{\nu}_i)^+ \tag{73}$$

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

$$g(t_k, \boldsymbol{\theta}_k) = t_k - \frac{1}{np} \sum_{i=1}^{n} (t_k - \boldsymbol{\nu}_i)^+ \tag{74}$$

$$= t_k - \frac{1}{np} \sum_{i=1}^{n(p-\delta)} (t_k - \boldsymbol{\nu}_i)^+ \tag{75}$$

$$= t_k - \frac{1}{np}(n(p-\delta))t_k + \frac{1}{np} \sum_{i=1}^{n(p-\delta)} \boldsymbol{\nu}_i \tag{76}$$

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

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

This concludes the proof. □

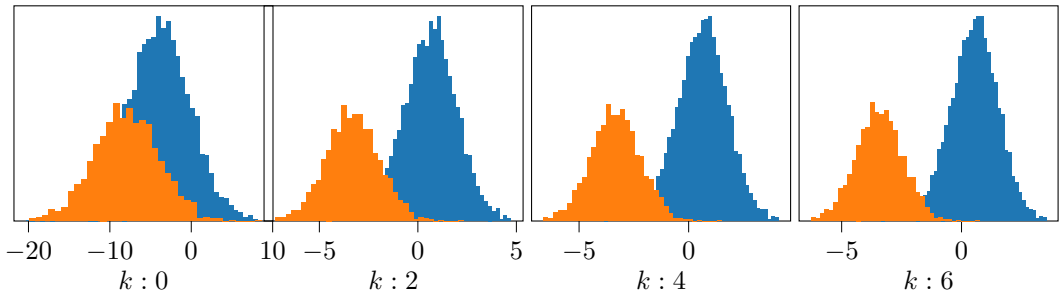

Figure 5: Residuals with respect to $\mathbb{P}$ and $\mathbb{Q}$, k represents optimization step.

C Theory for Adaptive 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 5. Even after multiple iteration steps the residuals with respect to $\mathbb{P}$ and $\mathbb{Q}$ are still normal. Thus it follows by decreasing $\|\boldsymbol{\theta} - \boldsymbol{\beta}_P\|_1$ more relative to $\|\boldsymbol{\theta} - \boldsymbol{\beta}_Q\|_1$ then the SubQuantile will contain more points from $\boldsymbol{P}$ by expectation.

C.1 Proof of Lemma 3.1

Proof.

Assumption 4. *We assume $\boldsymbol{\beta}_P \neq \alpha \boldsymbol{\beta}_Q \ \forall \alpha \in \mathbb{R}$.*

First let us define the projections of $\boldsymbol{\theta}$ onto both $\boldsymbol{\beta}_P$ and $\boldsymbol{\beta}_Q$. We will define a basis of the $\boldsymbol{\theta}$-space which is $\mathbb{R}^d$. Let the basis,

$$\boldsymbol{B} = [\boldsymbol{\beta}_P \quad \boldsymbol{\beta}_Q \quad \boldsymbol{R}] \text{ where } \boldsymbol{R} = [\boldsymbol{r}_3 \quad \boldsymbol{r}_4 \quad \dots \quad \boldsymbol{r}_d] \tag{79}$$

Thus we can represent $\boldsymbol{\theta}$ as a linear combination of the basis

$$\boldsymbol{\theta} = \alpha_1 \boldsymbol{\beta}_P + \alpha_2 \boldsymbol{\beta}_Q + \sum_{i=3}^{d} \alpha_i \boldsymbol{r}_i \tag{80}$$

Now we will calculate the projections of $\boldsymbol{\theta}$ on to $\boldsymbol{\beta}_P$ and $\boldsymbol{\beta}_Q$.

$$\begin{aligned}
\text{proj}_{\boldsymbol{\beta}_P}(\boldsymbol{\theta}) &= \left(\frac{\boldsymbol{\theta}^\top \boldsymbol{\beta}_P}{\|\boldsymbol{\beta}_P\|_2^2} \right) \boldsymbol{\beta}_P \\
&= \left(\frac{\|\boldsymbol{\theta}\| \, \|\boldsymbol{\beta}_P\| \cos(\omega_P)}{\|\boldsymbol{\beta}_P\|_2^2} \right) \boldsymbol{\beta}_P \\
&= \left(\frac{\|\boldsymbol{\theta}\| \cos(\omega_P)}{\|\boldsymbol{\beta}_P\|} \right) \boldsymbol{\beta}_P
\end{aligned}$$

We can use the same derivation for $\boldsymbol{\beta}_Q$:

$$\text{proj}_{\boldsymbol{\beta}_Q}(\boldsymbol{\theta}) = \left(\frac{\|\boldsymbol{\theta}\| \cos(\omega_Q)}{\|\boldsymbol{\beta}_Q\|} \right) \boldsymbol{\beta}_Q \tag{81}$$

Now we can prove the lemma.

Let us set $\alpha_1 = \left(\frac{\|\boldsymbol{\theta}\| \cos(\omega_P)}{\|\boldsymbol{\beta}_P\|} \right)$ and $\alpha_2 = \left(\frac{\|\boldsymbol{\theta}\| \cos(\omega_Q)}{\|\boldsymbol{\beta}_Q\|} \right)$. Therefore we can rewrite $\boldsymbol{\theta}$ as the following:

$$\boldsymbol{\theta} = \left(\frac{\|\boldsymbol{\theta}\| \cos(\omega_P)}{\|\boldsymbol{\beta}_P\|} \right) \boldsymbol{\beta}_P + \left(\frac{\|\boldsymbol{\theta}\| \cos(\omega_Q)}{\|\boldsymbol{\beta}_Q\|} \right) \boldsymbol{\beta}_Q + \sum_{i=3}^{d} \alpha_i \boldsymbol{r}_i \tag{82}$$

From a simple algebraic manipulation we have:

$$\boldsymbol{\theta} - \boldsymbol{\beta}_P = \left(\frac{\|\boldsymbol{\theta}\|\cos(\omega_P)}{\|\boldsymbol{\beta}_P\|} - 1\right)\boldsymbol{\beta}_P + \left(\frac{\|\boldsymbol{\theta}\|\cos(\omega_Q)}{\|\boldsymbol{\beta}_Q\|}\right)\boldsymbol{\beta}_Q + \sum_{i=3}^{d}\alpha_i \boldsymbol{r}_i \tag{83}$$

It thus follows, if $\left|\left(\frac{\|\boldsymbol{\theta}'\|\cos(\omega_P)}{\|\boldsymbol{\beta}_P\|}\right) - 1\right| < \left|\left(\frac{\|\boldsymbol{\theta}\|\cos(\omega_P)}{\|\boldsymbol{\beta}_P\|}\right) - 1\right|$ then $|\boldsymbol{\theta}' - \boldsymbol{\beta}_P| < |\boldsymbol{\theta} - \boldsymbol{\beta}_P|$ due to our formation of the basis. This concludes the proof. □

C.2 Probability ε decreasing

Proof. To calculate the probability $\varepsilon^{(t+1)} < \varepsilon^{(t)}$, we will first calculate $\mathbb{E}\left[\varepsilon^{(t+1)}\right]$. We will start by calculating the expectation of the loss.

$$\begin{aligned}
\mathbb{E}\left[\left(\boldsymbol{\theta}_{(t+1)}^\top \boldsymbol{p}_i - y_i\right)^2\right] &= \mathbb{E}\left[\boldsymbol{\theta}_{(t+1)}^\top \boldsymbol{p}_i - y_i\right]^2 + \mathrm{Var}\left(\boldsymbol{\theta}_{(t+1)}^\top \boldsymbol{p}_i - y_i\right) \\
&= \mathbb{E}\left[\boldsymbol{\theta}_{(t+1)}^\top \boldsymbol{p}_i - \boldsymbol{\beta}_P^\top \boldsymbol{p}_i - \epsilon_P\right]^2 + \mathrm{Var}\left(\boldsymbol{\theta}_{(t+1)}^\top \boldsymbol{p}_i - \boldsymbol{\beta}_P^\top \boldsymbol{p}_i - \epsilon_P\right) \\
&= \left(\left(\boldsymbol{\theta}_{(t+1)}^\top - \boldsymbol{\beta}_P^\top\right)\mathbb{E}\left[\boldsymbol{p}_i\right]\right)^2 + \mathrm{Var}\left(\left(\boldsymbol{\theta}_{(t+1)}^\top - \boldsymbol{\beta}_P^\top\right)\boldsymbol{p}_i\right) + \mathrm{Var}\left(\epsilon_P\right) \\
&= \left(\left(\boldsymbol{\theta}_{(t+1)}^\top - \boldsymbol{\beta}_P^\top\right)\boldsymbol{\mu}\right)^2 + \left(\boldsymbol{\theta}_{(t+1)}^\top - \boldsymbol{\beta}_P^\top\right)\boldsymbol{\Sigma}\left(\boldsymbol{\theta}_{(t+1)} - \boldsymbol{\beta}_P\right) + \mathrm{Var}\left(\epsilon_P\right)
\end{aligned}$$

It thus follows similarly:

$$\mathbb{E}\left[\left(\boldsymbol{\theta}_{(t+1)}^\top \boldsymbol{q}_i - y_i\right)^2\right] = \left(\left(\boldsymbol{\theta}_{(t+1)}^\top - \boldsymbol{\beta}_Q^\top\right)\boldsymbol{\mu}\right)^2 + \left(\boldsymbol{\theta}_{(t+1)}^\top - \boldsymbol{\beta}_Q^\top\right)\boldsymbol{\Sigma}\left(\boldsymbol{\theta}_{(t+1)} - \boldsymbol{\beta}_Q\right) + \mathrm{Var}\left(\epsilon_Q\right)$$

To simplify our notation, let $\zeta_P \coloneqq \mathbb{E}\left[\left(\boldsymbol{\theta}_{(t+1)}^\top \boldsymbol{p}_i - y_i\right)^2\right]$ and $\zeta_Q \coloneqq \mathbb{E}\left[\left(\boldsymbol{\theta}_{(t+1)}^\top \boldsymbol{q}_i - y_i\right)^2\right]$ and let $\eta_P \coloneqq \mathbb{E}\left[\boldsymbol{\theta}_{(t+1)}^\top \boldsymbol{p}_i - y_i\right]$ and $\eta_Q \coloneqq \mathbb{E}\left[\boldsymbol{\theta}_{(t+1)}^\top \boldsymbol{q}_i - y_i\right]$ We are now interested in calculating the variance of the loss.

Let $\sigma_P^2 \coloneqq \mathrm{Var}\left(\boldsymbol{\theta}_{(t+1)}^\top \boldsymbol{p}_i - y_i\right) = \left(\boldsymbol{\theta}_{(t+1)}^\top - \boldsymbol{\beta}_P^\top\right)\boldsymbol{\Sigma}\left(\boldsymbol{\theta}_{(t+1)} - \boldsymbol{\beta}_P\right) + \mathrm{Var}\left(\epsilon_P\right)$.

$$\begin{aligned}
\mathrm{Var}\left(\left(\boldsymbol{\theta}_{(t+1)}^\top \boldsymbol{p}_i - y_i\right)^2\right) &= \mathbb{E}\left[\left(\boldsymbol{\theta}_{(t+1)}^\top \boldsymbol{p}_i - y_i\right)^4\right] - \mathbb{E}\left[\left(\boldsymbol{\theta}_{(t+1)}^\top \boldsymbol{p}_i - y_i\right)^2\right]^2 \\
&= \mathbb{E}\left[\boldsymbol{\theta}_{(t+1)}^\top \boldsymbol{p}_i - y_i\right]^4 + 6\mathbb{E}\left[\boldsymbol{\theta}_{(t+1)}^\top \boldsymbol{p}_i - y_i\right]^2 \sigma_P^2 + 3\sigma_P^4 - \zeta_P^2 \\
&= \eta_P^4 + 6\eta_P^2\sigma_P^2 + 3\sigma_P^4 - \zeta_P^2
\end{aligned}$$

It similarly follows:

$$\mathrm{Var}\left(\left(\boldsymbol{\theta}_{(t+1)}^\top \boldsymbol{q}_i - y_i\right)^2\right) = \eta_Q^4 + 6\eta_Q^2\sigma_Q^2 + 3\sigma_Q^4 - \zeta_Q^2$$

Recall that the expected value of $\varepsilon^{(t+1)}$ is equal to the expected number of elements from Q within the p-Quantile, $\mathcal{Q}_p$. So we will now calculate $\mathcal{Q}_p$. To this we will utilize the loss of the distribution.

$$\begin{aligned}
\mathbb{E}\left[\left(\boldsymbol{\theta}_{(t+1)}^\top \boldsymbol{x}_i - y_i\right)^2\right] &= (1-\varepsilon^{(t)})\mathbb{E}\left[\left(\boldsymbol{\theta}_{(t+1)}^\top \boldsymbol{p}_i - y_i\right)^2\right] + \varepsilon^{(t)}\mathbb{E}\left[\left(\boldsymbol{\theta}_{(t+1)}^\top \boldsymbol{q}_i - y_i\right)^2\right] \\
&= (1-\varepsilon^{(t)})\zeta_P + \varepsilon^{(t)}\zeta_Q \\
\mathrm{Var}\left(\left(\boldsymbol{\theta}_{(t+1)}^\top \boldsymbol{x}_i - y_i\right)^2\right) &= (1-\varepsilon^{(t)})^2\mathrm{Var}\left(\left(\boldsymbol{\theta}_{(t+1)}^\top \boldsymbol{p}_i - y_i\right)^2\right) + \varepsilon^{(t)2}\mathrm{Var}\left(\left(\boldsymbol{\theta}_{(t+1)}^\top \boldsymbol{q}_i - y_i\right)^2\right) \\
&= (1-\varepsilon^{(t)})^2\left(\eta_P^4 + 6\eta_P^2\sigma_P^2 + 3\sigma_P^4 - \zeta_P^2\right) + (\varepsilon^{(t)})^2\left(\eta_Q^4 + 6\eta_Q^2\sigma_Q^2 + 3\sigma_Q^4 - \zeta_Q^2\right)
\end{aligned}$$

It thus follows $\left(\boldsymbol{\theta}_{(t+1)}^\top \boldsymbol{x}_i - y_i\right)$ follows a χ^2 distribution with 2 degrees of freedom and

$$\lambda = \frac{(1-\varepsilon^{(t)})\zeta_P + \varepsilon^{(t)}\zeta_Q}{(1-\varepsilon^{(t)})^2\left(\eta_P^4 + 6\eta_P^2\sigma_P^2 + 3\sigma_P^4 - \zeta_P^2\right) + (\varepsilon^{(t)})^2\left(\eta_Q^4 + 6\eta_Q^2\sigma_Q^2 + 3\sigma_Q^4 - \zeta_Q^2\right)}$$

Let $\mathcal{Q}_p^{(t+1)}$ be given as the np greatest loss of the data, i.e. $\mathcal{Q}_p^{(t+1)} = \boldsymbol{\nu}_{np}^{(t+1)}$. It then follows:

$$\mathbb{P}\left[\left(\boldsymbol{\theta}_{(t+1)}^\top \boldsymbol{q}_i - y_i\right)^2 \leq \mathcal{Q}_p\right] = \frac{\gamma(1, \mathcal{Q}_p/2)}{\Gamma(1)}$$

It thus follows:

$$\mathbb{E}\left[\varepsilon^{(t+1)}\right] = \frac{\gamma(1, \mathcal{Q}_p/2)}{\Gamma(1)}$$

Now we can calculate the probability of improvement per iteration by invoking Markov's Inequality:

$$\begin{aligned}\mathbb{P}\left[\varepsilon^{(t+1)} \geq \varepsilon^{(t)}\right] &\leq \frac{\mathbb{E}\left[\varepsilon^{(t+1)}\right]}{\varepsilon^{(t)}} \\ &= \frac{\gamma(1, \mathcal{Q}_p/2)}{\varepsilon^{(t)}\Gamma(1)}\end{aligned}$$

where γ represents the lower incomplete gamma function. In many works in robust least squares regression, it is common for data to be sampled from $\mathbf{0}$ centered sub-Gaussian data with $\boldsymbol{\Sigma} = \boldsymbol{I}$ covariance matrix. This simplifies our theory greatly. Now, $\eta_P = \eta_Q = 0$ and $\sigma_P^2 = \left\|\boldsymbol{\theta}_{(t+1)} - \boldsymbol{\beta}_P\right\|_2^2 + \mathrm{Var}(\epsilon_P)$. It futher follows $\zeta_P = \sigma_P^2$. Thus □

We will now verify the theory by looking at the noise procedure described in Bhatia et al. (2017), where a bias is added to a random subset of the data.

C.3 Proof of Theorem 4

Proof. To show the change in ε we will first calculate the expected change in $\boldsymbol{\theta}$ by the $\boldsymbol{\theta}$-update described in Equation 13. We will also introduce some notation, $\mathcal{S}$ represents all $\boldsymbol{x} \in \boldsymbol{X}$ that are within the lowest np losses, i.e. within the subquantile, $\boldsymbol{p} \in \mathcal{S}$ represent all data vectors from $\mathbb{P}$ that are within the SubQuantile, similarly $\boldsymbol{q} \in \mathbb{Q}$ represent all data vectors from $\mathbb{Q}$ that are within the SubQuantile. Furthermore, $|\mathcal{S}| = np$, there are εn points from $\mathbb{Q}$ in $\mathcal{S}$ and $(1-\varepsilon)n$ points from $\mathbb{P}$ within $\mathcal{S}$.

$$\mathbb{E}\left[\boldsymbol{\theta}_{k+1}\right] = \boldsymbol{\theta}_k - \mathbb{E}\left[\alpha \nabla g(\boldsymbol{\theta}_k, t_{k+1})\right] \tag{84}$$

$$= \boldsymbol{\theta}_k - \alpha \mathbb{E}\left[\sum_{\boldsymbol{x} \in \mathcal{S}} \boldsymbol{x}(\boldsymbol{\theta}^\top \boldsymbol{x} - y)\right] \tag{85}$$

$$= \boldsymbol{\theta}_k - \alpha \mathbb{E}\left[\sum_{\boldsymbol{x} \in \mathcal{S}} \boldsymbol{x}\boldsymbol{x}^\top \boldsymbol{\theta}_k - \boldsymbol{x}y\right] \tag{86}$$

$$= \boldsymbol{\theta}_k - \alpha \mathbb{E}\left[\sum_{\boldsymbol{p} \in \mathcal{S}} \boldsymbol{p}\boldsymbol{p}^\top \boldsymbol{\theta}_k - \boldsymbol{p}y_p + \sum_{\boldsymbol{q} \in \mathcal{S}} \boldsymbol{q}\boldsymbol{q}^\top \boldsymbol{\theta}_k - \boldsymbol{q}y_q\right] \tag{87}$$

We will use Assumption 1 to rewrite y_p and y_q

$$= \boldsymbol{\theta}_k - \alpha \mathbb{E}\left[\sum_{\boldsymbol{p} \in \mathcal{S}} \boldsymbol{p}\boldsymbol{p}^\top \boldsymbol{\theta}_k - \boldsymbol{p}(\boldsymbol{\beta}_P \boldsymbol{p} + \epsilon_P) + \sum_{\boldsymbol{q} \in \mathcal{S}} \boldsymbol{q}\boldsymbol{q}^\top \boldsymbol{\theta}_k - \boldsymbol{q}(\boldsymbol{\beta}_Q \boldsymbol{p} + \epsilon_Q)\right] \tag{88}$$

$$= \boldsymbol{\theta}_k - \alpha \left(\sum_{\boldsymbol{p} \in \mathcal{S}} \left(\boldsymbol{\mu}\boldsymbol{\mu}^\top + \boldsymbol{\Sigma}\right)\boldsymbol{\theta}_k - \left(\boldsymbol{\mu}\boldsymbol{\mu}^\top + \boldsymbol{\Sigma}\right)\boldsymbol{\beta}_P - \sum_{\boldsymbol{q} \in \mathcal{S}} \left(\boldsymbol{\mu}\boldsymbol{\mu}^\top + \boldsymbol{\Sigma}\right)\boldsymbol{\theta}_k + \left(\boldsymbol{\mu}\boldsymbol{\mu}^\top + \boldsymbol{\Sigma}\right)\boldsymbol{\beta}_Q\right) \tag{89}$$

Let $\boldsymbol{C} = \boldsymbol{\mu}\boldsymbol{\mu}^\top + \boldsymbol{\Sigma}$ for notational simplicity

$$= \boldsymbol{\theta}_k - \alpha np \boldsymbol{C}\left(\boldsymbol{\theta}_k - (1-\varepsilon)\boldsymbol{\beta}_P - \varepsilon \boldsymbol{\beta}_Q\right) \tag{90}$$

Now that we have the expected update for $\boldsymbol{\theta}$ in terms of the linear regression coefficients, we now want to utilize Lemma 3.1.

Let $\boldsymbol{\theta}_k = \alpha_1 \boldsymbol{\beta}_P + \alpha_2 \boldsymbol{\beta}_Q + \sum_{i=3}^{d} \alpha_i \boldsymbol{r}_i$ in the same basis $\boldsymbol{B}$ defined in Lemma 3.1. Note in the case of data that is normally distributed about 0 with no covariance amongst the predictor variables, then $\boldsymbol{C}$ is a multiple of the identity. If we proprocess the data using a standard normal scalar such that all features follow a $\mathcal{N}(0,1)$ distribution, then we can assume it follows $\boldsymbol{C} = (\mathbf{0}\mathbf{0}^\top + \boldsymbol{I}) = \boldsymbol{I}$. Then

the following manipulations hold:

$$\boldsymbol{\theta}_{k+1} = \boldsymbol{\theta}_k - \alpha np\left(\boldsymbol{\theta}_k - (1-\varepsilon)\boldsymbol{\beta}_P - \varepsilon\boldsymbol{\beta}_Q\right) \tag{91}$$

$$= \boldsymbol{\theta}_k(1-\alpha np) + \alpha np(1-\varepsilon)\boldsymbol{\beta}_P + (1-\alpha np)\varepsilon\boldsymbol{\beta}_Q \tag{92}$$

$$= (1-\alpha np)\left(\alpha_1\boldsymbol{\beta}_P + \alpha_2\boldsymbol{\beta}_Q + \sum_{i=3}^{d}\alpha_i\boldsymbol{r}_i\right) + \alpha np(1-\varepsilon)\boldsymbol{\beta}_P + \alpha np\varepsilon\boldsymbol{\beta}_Q \tag{93}$$

$$= \alpha np(1-\varepsilon-\alpha_1)\boldsymbol{\beta}_P + \alpha np(\varepsilon-\alpha_2)\boldsymbol{\beta}_Q + \alpha_1\boldsymbol{\beta}_P + \alpha_2\boldsymbol{\beta}_Q + (1-\alpha np)\sum_{i=3}^{d}\alpha_i\boldsymbol{r}_i \tag{94}$$

Thus the conditions for ε to decrease by expectation are $\|(1-\varepsilon-\alpha_1)\boldsymbol{\beta}_P\| > \|(\varepsilon-\alpha_2)\boldsymbol{\beta}_Q\|$. This concludes the proof. Still need to verify last two steps. □

D PROOFS FOR CONVERGENCE

D.1 PROOF OF THEOREM 3

Proof. We will first start by introducing new notation. Let $\boldsymbol{S}$ represent a matrix with np data points from $\boldsymbol{X}$, in other words it is a possible SubQuantile Matrix. Let $\boldsymbol{\Pi}$ represent the set of all such possible matrices $\boldsymbol{S}$ of $\boldsymbol{X}$. Note $|\boldsymbol{\Pi}| = \binom{n}{np}$. We can now redefine the min-max optimization problem of g to a min-min optimization problem. Let us define the function $f(\boldsymbol{\theta}, \boldsymbol{S}) = \left\|\boldsymbol{\theta}^\top \boldsymbol{S} - \boldsymbol{y}_S\right\|_2^2$

$$\boldsymbol{\theta}^*, \boldsymbol{S}^* = \arg\min_{\boldsymbol{\theta}\in\mathbb{R}^d} \arg\min_{\boldsymbol{S}\in\boldsymbol{\Pi}} \left\|\boldsymbol{\theta}^\top \boldsymbol{S} - \boldsymbol{y}_S\right\|_2^2 \tag{95}$$

I think this optimization problem is easier to show to converge. Note we have a $\mathcal{O}(n)$ oracle for the $\arg\min_{S\in\boldsymbol{\Pi}} f(\boldsymbol{\theta}_T, \boldsymbol{S}_T)$.

Lemma 4.1. *The resultant* $\widetilde{\boldsymbol{S}} = \arg\min_{\boldsymbol{S}\in\boldsymbol{\Pi}} f(\boldsymbol{\theta}, \boldsymbol{S})$ *is a unique minimizer iff all points in* $\boldsymbol{X}$ *are different.*

We will now show f is a monotonically decreasing function.
First let us define $\phi(\cdot) = \min_{\boldsymbol{S}\in\boldsymbol{\Pi}}(\cdot, \boldsymbol{S})$.

Lemma 4.2. *If*

Let us also note f is ℓ smooth with respect to $\boldsymbol{\theta}$. This is following notation from Jin et al. (2019).It thus follows:

$$\begin{aligned}
f(\boldsymbol{\theta}_{k+1}, \boldsymbol{S}_k) &\leq f(\boldsymbol{\theta}_k, \boldsymbol{S}_k) + \langle \nabla_{\boldsymbol{\theta}} f(\boldsymbol{\theta}_k, \boldsymbol{S}_k), \boldsymbol{\theta}_{k+1} - \boldsymbol{\theta}_k \rangle + \frac{\ell}{2} \left\|\boldsymbol{\theta}_{k+1} - \boldsymbol{\theta}_k\right\|_2^2 \\
&= \phi(\boldsymbol{\theta}_k) + \langle \nabla_{\boldsymbol{\theta}} f(\boldsymbol{\theta}_k, \boldsymbol{S}_k), -\frac{1}{\ell} \nabla_{\boldsymbol{\theta}} f(\boldsymbol{\theta}_k, \boldsymbol{S}_k) \rangle + \frac{\ell}{2} \left\| \frac{1}{\ell} \nabla_{\boldsymbol{\theta}} f(\boldsymbol{\theta}_k, \boldsymbol{S}_k) \right\|_2^2 \\
&= \phi(\boldsymbol{\theta}_k) - \frac{1}{\ell} \left(\nabla_{\boldsymbol{\theta}} f(\boldsymbol{\theta}_k, \boldsymbol{S}_k)\right)^2 + \frac{1}{2\ell} \left(\nabla_{\boldsymbol{\theta}} f(\boldsymbol{\theta}_k, \boldsymbol{S}_k)\right)^2 \\
&= \phi(\boldsymbol{\theta}_k) - \frac{1}{2\ell} \left(\nabla_{\boldsymbol{\theta}} f(\boldsymbol{\theta}_k, \boldsymbol{S}_k)\right)^2
\end{aligned}$$

Thus we have proved the inner optimization problem is monotonically decreasing. Since the outer minimization is strictly less than or equal to the result from the inner optimization, it follows after each two step optimization:

$$\phi(\boldsymbol{\theta}_{k+1}) \leq \phi(\boldsymbol{\theta}_k)$$

Since f is lower bounded by 0. We can invoke Montonicity Convergence Theorem, since f is a monotically decreasing function and is lower bounded, it therefore converges to either a local or global minimum. □

D.2 EXPECTATION OF IMPROVEMENT

Let us define the function $h(\boldsymbol{\theta}_k) = f(\boldsymbol{\theta}_k, \boldsymbol{S}_{k-1}) - f(\boldsymbol{\theta}_k, \boldsymbol{S}_k)$, a strictly non-negative function. From our results above it follows:

$$\begin{aligned}
\phi(\boldsymbol{\theta}_{k+1}) &\leq \phi(\boldsymbol{\theta}_k) - \frac{1}{2\ell} \left(\nabla_{\boldsymbol{\theta}} f(\boldsymbol{\theta}_k, \boldsymbol{S}_k)\right)^2 \\
&= f(\boldsymbol{\theta}_k, \boldsymbol{S}_{k-1}) - h(\boldsymbol{\theta}_k) + -\frac{\ell}{2} \left\|\boldsymbol{\theta}_{k+1} - \boldsymbol{\theta}_k\right\|_2^2 \\
&= f(\boldsymbol{\theta}_k, \boldsymbol{S}_{k-1}) - h(\boldsymbol{\theta}_k) - \frac{\ell}{2} \left\| \boldsymbol{\theta}_{k+1} - \frac{1}{\ell} \nabla_{\boldsymbol{\theta}} f(\boldsymbol{\theta}_k, \boldsymbol{S}_k) - \boldsymbol{\theta}_k \right\|_2^2 \\
&= f(\boldsymbol{\theta}_k, \boldsymbol{S}_{k-1}) - h(\boldsymbol{\theta}_k) - \frac{\ell}{2} \left(\left\|\boldsymbol{\theta}_{k+1} - \boldsymbol{\theta}_k\right\|_2^2 + \frac{2}{\ell} \left\|\boldsymbol{\theta}_{k+1} - \boldsymbol{\theta}_k\right\| \nabla_{\boldsymbol{\theta}} f(\boldsymbol{\theta}_k, \boldsymbol{S}_k) + \frac{1}{\ell^2} \nabla_{\boldsymbol{\theta}} f(\boldsymbol{\theta}_k, \boldsymbol{S}_k) \right)
\end{aligned}$$

The plan is to take a telescopic sum over $f(\boldsymbol{\theta}_{k+1}, \boldsymbol{S}_{k+1}) - f(\boldsymbol{\theta}_k, \boldsymbol{S}_k)$. This is in the works for the final version of the paper

E 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

  $\boldsymbol{\nu} = (\boldsymbol{X}_I \boldsymbol{\theta}_k - \boldsymbol{y}_I)^2$

  $\hat{\boldsymbol{\nu}} = sorted(\boldsymbol{\nu})$

  $t_{k+1} = \hat{\boldsymbol{\nu}}_{mp}$

  $L \coloneqq \sum_{i=1}^{mp} \boldsymbol{x}_i^\top \boldsymbol{x}_i$

  $\alpha \coloneqq \frac{1}{2L}$

  $\boldsymbol{\theta}_{k+1} = \boldsymbol{\theta}_k - \alpha \nabla_{\boldsymbol{\theta}_k} g(t_{k+1}, \boldsymbol{\theta}_k)$

end

return $\boldsymbol{\theta}_T$

I think a proof of convergence for a stochastic batch algorithm would be nice to add for the final paper if possible

F Additional Experiments

F.1 Quadratic Regression

Objectives	Test RMSE (Quadratic Regression)		
	$\epsilon = 0$	$\epsilon = 0.2$	$\epsilon = 0.4$
ERM	$0.0099_{(0.0002)}$	$2.078_{(0.146)}$	$4.104_{(0.442)}$
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($p = 0.6$,$j = 1$)	$\mathbf{0.0099_{(0.0002)}}$	$\mathbf{0.00998_{(0.0002)}}$	$\mathbf{0.010_{(0.0001)}}$
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.2 Abalone

We now provide results on Abalone Dataset introduced in Dua & Graff (2017). This experiment

Objectives	Test RMSE (Abalone Linear Regression)	
	Clean	Noisy
ERM	$2.213_{(0.0528)}$	$4845.335_{(117.5557)}$
CRR Bhatia et al. (2017)	$2.345_{(0.0430)}$	$396.872_{(96.5632)}$
STIR Mukhoty et al. (2019)	$2.240_{(0.0473)}$	$931.845_{(32.0864)}$
Huber Huber & Ronchetti (2009)	$5.535_{(0.0665)}$	$971.362_{(28.8863)}$
RANSAC Fischler & Bolles (1981)	$2.522_{(0.1407)}$	$2.621_{(0.1719)}$
TERM Li et al. (2020)	$10.686_{(0.2616)}$	$10.853_{(0.4245)}$
SEVER Diakonikolas et al. (2019)	$\mathbf{2.238_{(0.0901)}}$	$\mathbf{2.287_{(0.0757)}}$
SubQuantile($p = 0.8$,$j = 100$)	$\mathbf{2.292_{(0.0413)}}$	$\mathbf{2.261_{(0.0790)}}$
Genie ERM	$2.213_{(0.0528)}$	$2.238_{(0.0901)}$

Table 4: Abalone Regression Real Dataset. Empirical Risk over $\mathbb{P}$

has both feature and label noise in the Noisy Data. SubQuantile minimization no longer always converges to the $\mathbb{P}$ SubQuantile. I think the theory on this can be expanded on in Sections C.1 and C.3

F.3 Cal-Housing

We now provide results on Cal-Housing Dataset introduced in Pace & Barry (1997). This experiment has both feature and label noise in the Noisy Data.

In both the Cal-Housing and Abalone datasets there exists feature and label noise that exist with 5% probability. In this the case, the probability is low, however since the noise is very large, even having a few points from $\mathbb{Q}$ in the final subquantile matrix can largely the bias the predictions away from the optimal parameters for $\mathbb{P}$. Therefore, we reduce p, the size of the subquantile to reduce the probability of obtaining corrupted samples within the subquantile. However, what we get in a decrease in variance, we do increase the bias error, albeit very slightly. We theoretically justify the reduction of p in variance reduction and the effect of the parameter j in sections . . .

Objectives	Test RMSE (`Cal-Housing Linear Regression`)	
	Clean	Noisy
ERM	$0.598_{(0.0077)}$	$81.758_{(2.6230)}$
CRR Bhatia et al. (2017)	$\mathbf{0.602}_{(\mathbf{0.0081})}$	$75.777_{(2.9403)}$
STIR Mukhoty et al. (2019)	$\mathbf{0.604}_{(\mathbf{0.0070})}$	$65.555_{(2.1899)}$
Huber Huber & Ronchetti (2009)	$\mathbf{0.601}_{(\mathbf{0.0077})}$	$71.813_{(2.0755)}$
RANSAC Fischler & Bolles (1981)	$0.681_{(0.0389)}$	$0.679_{(0.0253)}$
TERM Li et al. (2020)	$0.737_{(0.0070)}$	$0.741_{(0.0155)}$
SEVER Diakonikolas et al. (2019)	$0.640_{(0.0067)}$	$0.642_{(0.0088)}$
SubQuantile($p = 0.9$, $j = 100$)	$\mathbf{0.615}_{(\mathbf{0.0076})}$	$\mathbf{0.612}_{(\mathbf{0.0096})}$
Genie ERM	$0.598_{(0.0077)}$	$0.603_{(0.0068)}$

Table 5: `Cal-Housing Regression` Real Dataset. Empirical Risk over $\mathbb{P}$

G Experimental Details

G.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. The noise is added after normalization of the dataset to the standard normal $\mathcal{N}(0,1)$.

G.2 `Noisy Linear Regression` Dataset

We will describe $\mathbb{P}$ and $\mathbb{Q}$ in the `Noisy Linear Regression` Dataset.
$\boldsymbol{x} \sim \mathcal{N}(0,3)^{500}$
$\boldsymbol{m} \sim \mathcal{N}(4,4)^{500}$
$b \sim \mathcal{N}(4,4)$
$\boldsymbol{m}' = \mathbf{0}$
$b' \sim \mathcal{N}(5,5)$
$n_{\text{train}} = 5\text{e}3$
$n_{\text{test}} = 1e2$
$\mathbb{P} : y|\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{m}^\top \boldsymbol{x} + b, 0.1)$
$\mathbb{Q} : y|\boldsymbol{x} \sim \mathcal{N}(b', 4)$
Please note $\boldsymbol{m}$, b, $\boldsymbol{m}'$, b', are all sampled independently. The noise is added after normalization of the dataset to the standard normal.

G.3 `Quadratic Regression` Dataset

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

G.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

G.5 FEATURE NOISE

Take 5% of the training data and multiply features by 100 and responses by 10000.