ROBUST LINEAR REGRESSION BY SUPER-QUANTILE OPTIMIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

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

1 INTRODUCTION

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

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

In this 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 $\mathbb{Q}$.
It is know the least squares optimal solution for $\boldsymbol{X}$ is equal to $(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$
Note $\boldsymbol{X} = \begin{pmatrix}\boldsymbol{P}\\ \boldsymbol{Q}\end{pmatrix}$ so $\boldsymbol{X}^T = \begin{pmatrix}\boldsymbol{P}^T & \boldsymbol{Q}^T\end{pmatrix}$

$$\boldsymbol{X}^T\boldsymbol{X} = \begin{pmatrix}\boldsymbol{P}^T & \boldsymbol{Q}^T\end{pmatrix}\begin{pmatrix}\boldsymbol{P}\\ \boldsymbol{Q}\end{pmatrix} \tag{1}$$

$$= \boldsymbol{P}^T\boldsymbol{P} + \boldsymbol{Q}^T\boldsymbol{Q} \tag{2}$$

$$(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T = (\boldsymbol{P}^T\boldsymbol{P} + \boldsymbol{Q}^T\boldsymbol{Q})^{-1}\begin{pmatrix}\boldsymbol{P}^T & \boldsymbol{Q}^T\end{pmatrix} \tag{3}$$

$$= \begin{pmatrix}(\boldsymbol{P}^T\boldsymbol{P} + \boldsymbol{Q}^T\boldsymbol{Q})^{-1}\boldsymbol{P}^T & (\boldsymbol{P}^T\boldsymbol{P} + \boldsymbol{Q}^T\boldsymbol{Q})^{-1}\boldsymbol{Q}^T\end{pmatrix} \tag{4}$$

$$\boldsymbol{X}^\dagger\boldsymbol{y} = \begin{pmatrix}(\boldsymbol{P}^T\boldsymbol{P} + \boldsymbol{Q}^T\boldsymbol{Q})^{-1}\boldsymbol{P}^T & (\boldsymbol{P}^T\boldsymbol{P} + \boldsymbol{Q}^T\boldsymbol{Q})^{-1}\boldsymbol{Q}^T\end{pmatrix}\begin{pmatrix}\boldsymbol{y}_P\\ \boldsymbol{y}_Q\end{pmatrix} \tag{5}$$

$$= (\boldsymbol{P}^T\boldsymbol{P} + \boldsymbol{Q}^T\boldsymbol{Q})^{-1}\boldsymbol{P}^T\boldsymbol{y}_P + (\boldsymbol{P}^T\boldsymbol{P} + \boldsymbol{Q}^T\boldsymbol{Q})^{-1}\boldsymbol{Q}^T\boldsymbol{y}_Q \tag{6}$$

Note the optimal solution for a linear regression model on $\mathbb{P}$ is $(\boldsymbol{P}^T\boldsymbol{P})^{-1}\boldsymbol{P}^T\boldsymbol{y}_P$
Often times in the case of corrupted data we have $\boldsymbol{P}$ and $\boldsymbol{Q}$ are sampled similarly however $\boldsymbol{y}_P$ and $\boldsymbol{y}_Q$ are very different. Thus $(\boldsymbol{P}^T\boldsymbol{P} + \boldsymbol{Q}^T\boldsymbol{Q})^{-1}\boldsymbol{Q}^T\boldsymbol{y}_Q$ could have a large effect on the optimal solution. This is why we propose Sub-Quantile Optimization, we seek to reduce the impact of $\boldsymbol{Q}^T\boldsymbol{Q}$ and $\boldsymbol{y}_Q$ by reducing the number of rows in $\boldsymbol{Q}$. Thus we reduce the condition number of $\boldsymbol{Q}^T\boldsymbol{Q}$ and the overall effect on the optimal solution for $\boldsymbol{P}$.

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

2 RELATED WORK

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

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

$$\tilde{R}(t;\theta) := \frac{1}{t}\log\left(\frac{1}{N}\sum_{i\in[N]} e^{tf(x_i;\theta)}\right) \tag{7}$$

SMART Awasthi et al. (2022)

SEVER Diakonikolas et al. (2019)

Gradient Filtering Approaches (Need Source)

Super-Quantile Optimization Rockafellar et al. (2014)

3 SUB-QUANTILE OPTIMIZATION

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

$$t_{k+1} = \arg\max_t g(t,\boldsymbol{\theta}_k) \tag{8}$$

$$\boldsymbol{\theta}_{k+1} = \boldsymbol{\theta}_k + \alpha\nabla_{\boldsymbol{\theta}_k} g(t,\boldsymbol{\theta}_k) \tag{9}$$

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

Theorem 3.1. *Sub-Quantile Optimization Converges Almost Surely*

Lemma 3.1.1. *$g(t,\boldsymbol{\theta})$ is maximized when $t = Q_p(U)$*

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

$$\partial(-g(t,\boldsymbol{\theta})) = \partial\left(-t + \frac{1}{np}\sum_{i=1}^{n}(t-\hat{\boldsymbol{\nu}}_i)^+\right) \tag{10}$$

$$= -1 + \frac{1}{np}\sum_{i=1}^{n}\left\{\begin{array}{ll} 1, & \text{if } t > \hat{\boldsymbol{\nu}}_i \\ 0, & \text{if } t < \hat{\boldsymbol{\nu}}_i \\ [0,1], & \text{if } t = \hat{\boldsymbol{\nu}}_i \end{array}\right\} \tag{11}$$

$$= 0 \text{ when } t = \hat{\boldsymbol{\nu}}_{np} \tag{12}$$

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

Lemma 3.1.2. *Let $t = \hat{\boldsymbol{\nu}}_{np}$. The second step in the optimization is the derivative with respect to the first np elements in the sorted squared losses, $\hat{\boldsymbol{\nu}}$. The derivative of $g(t,\boldsymbol{\theta})$ w.r.t $\nabla_{\boldsymbol{\theta}} g(t_{k+1},\boldsymbol{\theta}_k) = \frac{1}{np}\sum_{i=1}^{np} 2\boldsymbol{x}_i(\boldsymbol{\theta}_k^T\boldsymbol{x}_i - y_i)$*

We provide a proof in Appendix B.2. By our choice of t_{k+1} what we find is the terms cancel out and we are left minimizing the terms within the np lowest squared losses.

4 THEORETICAL ANALYSIS

4.1 ROBUSTNESS

Assumption 4.1. *The parameters of $\boldsymbol{\theta}$ are sampled from a symmetrical continuous distribution around 0. Inspired by Lu (2020).*

By Lemma 3.1.2, $\boldsymbol{\theta}$ is updated only on the np points with the smallest squared loss. To quantify how "Robust" our linear regressor is, we want to know how many data points from $\mathbb{Q}$ are within the lowest np squared losses as the number of iterations, $k \to \infty$. To do this, we will model the data sampled from $\mathbb{P}$ and $\mathbb{Q}$ as random variables. Let $\boldsymbol{p}_1, \boldsymbol{p}_2, \ldots, \boldsymbol{p}_{(1-\epsilon)n}$ be the $(1-\epsilon)n$ points sampled i.i.d from $\mathbb{P}$. Let $P_1, P_2, \ldots, P_m$ be random variables that represent the data sampled from $\mathbb{P}$, $\boldsymbol{p}_1, \boldsymbol{p}_2, \ldots, \boldsymbol{p}_{1-(\epsilon)n}$ such that

$$P_i = \begin{cases} 1 & \text{if } (\boldsymbol{\theta}_k^T \boldsymbol{p}_i - y_i)^2 \leq \hat{\boldsymbol{\nu}}_{np} \\ 0 & \text{if } (\boldsymbol{\theta}_k^T \boldsymbol{p}_i - y_i)^2 > \hat{\boldsymbol{\nu}}_{np} \end{cases} \tag{13}$$

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

$$Q_i = \begin{cases} 1 & \text{if } (\boldsymbol{\theta}_k^T \boldsymbol{q}_i - y_i)^2 \leq \hat{\boldsymbol{\nu}}_{np} \\ 0 & \text{if } (\boldsymbol{\theta}_k^T \boldsymbol{q}_i - y_i)^2 > \hat{\boldsymbol{\nu}}_{np} \end{cases} \tag{14}$$

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

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

$$Q_k^+ = \sum_{i=1}^{n\epsilon} Q_i \text{ and } P_k^+ = \sum_{i=1}^{n(1-\epsilon)} P_i \tag{15}$$

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

From here we can calculate the expected update rule

$$\boldsymbol{\theta}_1 = \boldsymbol{\theta}_0 + \alpha \sum_{i=1}^{m} 2\boldsymbol{p}_i(\boldsymbol{\theta}_0^T \boldsymbol{p}_i - y_i) + \alpha \sum_{i=1}^{l} 2\boldsymbol{q}_i(\boldsymbol{\theta}_0^T \boldsymbol{q}_i - y_i) \tag{16}$$

Lemma 4.0.1. *$f_i(\boldsymbol{\theta})$ is a Lipschitz Continuous with parameter $L = ||\boldsymbol{x}_i||_2^2$*

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

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

$$\phi(\boldsymbol{\theta}) = \frac{1}{np} \sum_{i=1}^{m} (\boldsymbol{\theta}^T \boldsymbol{p}_i - y_i)^2 \tag{17}$$

$$\psi(\boldsymbol{\theta}) = \frac{1}{np} \sum_{i=1}^{l} (\boldsymbol{\theta}^T \boldsymbol{q}_i - y_i)^2 \tag{18}$$

These two functions hold nice properties.

$$\nabla_{\boldsymbol{\theta}} \phi(\boldsymbol{\theta}) = \frac{1}{np} \sum_{i=1}^{m} 2\boldsymbol{p}_i(\boldsymbol{\theta}^T \boldsymbol{p}_i - y_i) \tag{19}$$

$$\nabla_{\boldsymbol{\theta}} \psi(\boldsymbol{\theta}) = \frac{1}{np} \sum_{i=1}^{l} 2\boldsymbol{q}_i(\boldsymbol{\theta}^T \boldsymbol{q}_i - y_i) \tag{20}$$

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

$$\nabla_{\boldsymbol{\theta}_k}(t_{k+1}, \boldsymbol{\theta}_k) = \nabla_{\boldsymbol{\theta}_k} \phi(\boldsymbol{\theta}) + \nabla_{\boldsymbol{\theta}_k} \psi(\boldsymbol{\theta}) \tag{21}$$

Assumption 4.2. *Since we randomly choose the parameters of $\boldsymbol{\theta}_0$ and we assume since $m > l$ by expectation,*

$$\nabla_{\boldsymbol{\theta}_0} \phi(\boldsymbol{\theta}_0) > \nabla_{\boldsymbol{\theta}_0} \psi(\boldsymbol{\theta}_0) \tag{22}$$

4.2 Convergence

Lemma 4.0.2. *$g(t_{k+1}, \boldsymbol{\theta}_k)$ is convex with respect to $\boldsymbol{\theta}_k$.*

Lemma 4.0.2 tells us we are solving a min-max concave-convex optimization problem. In Jin et al. (2019), the researchers examined the problem where they are given a max-oracle which is correct up to some value ϵ. In our case, we can set $\epsilon = 0$.

Lemma 4.0.3. *$g(t, \boldsymbol{\theta})$ is Lipschitz Continuous with respect to t*

Lemma 4.0.4. *$g(t, \boldsymbol{\theta})$ is L-smooth with respect to $\boldsymbol{\theta}$ with* $L = \left\| \frac{2}{np} \sum_{i=1}^{np} \|\boldsymbol{x}_i\|^2 \right\|$

Lemma 4.0.5. *Since $g(t, \boldsymbol{\theta})$ is L-smooth by Lemma 4.0.4 $g(t, \boldsymbol{\theta})$ is a monotonically decreasing function.*

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

$$g(t_{k+1}, \boldsymbol{\theta}_k) - g(t_{k+1}, \boldsymbol{\theta}_{k+1}) \geq g(t_{k+1}, \boldsymbol{\theta}_k) - g(t_k, \boldsymbol{\theta}_k) \tag{23}$$

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

Lemma 4.0.6. *$g(t, \boldsymbol{\theta})$ is bounded above by $\sum_{i=1}^{np} \boldsymbol{\nu}_i$ and below by 0.*

Theorem 4.1. *By Lemma 4.0.5 and 4.0.6, $g(t, \boldsymbol{\theta})$ converges to a local minimum.*

Proof Sketch. Note after the t-update and $\boldsymbol{\theta}$-update as described in equations 8 and 9, respectively, lemma 4.0.5 tells us $g(t_{k+1}, \boldsymbol{\theta}_{k+1}) \leq g(t_k, \boldsymbol{\theta}_k)$ □

4.2.1 Point Change Conditions

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

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

Theorem 4.2. *The rate of decrease of $\boldsymbol{x}'$ is greater than $\boldsymbol{x}$ iff*

$$||\boldsymbol{x}'|||\boldsymbol{\theta}_0^T \boldsymbol{x}' - y| \cos(\omega') > ||\boldsymbol{x}|||\boldsymbol{\theta}_0^T \boldsymbol{x}' - y| \cos(\omega) \tag{24}$$

Theorem 4.2 reveals to us the importance of $\cos(\omega)$ which represents the angle between $\nabla f_{\boldsymbol{x}}(\boldsymbol{\theta}_0)$ and $\nabla g(\boldsymbol{\theta}_0)$. By our initial assumption $|\boldsymbol{\theta}_0^T \boldsymbol{x}' - y'| > |\boldsymbol{\theta}_0^T \boldsymbol{x} - y|$. Let us look at the example where $||\boldsymbol{x}|| = ||\boldsymbol{x}'||$, in this case, while $\cos(\omega') > \cos(\omega)$, the rate of decrease of $\boldsymbol{x}'$ will be more than the rate of decrease of $\boldsymbol{x}$. What this means is there will be an iteration step where $(\boldsymbol{\theta}_k^T \boldsymbol{x}' - y')^2 < (\boldsymbol{\theta}_k^T \boldsymbol{x} - y)^2$. Thus $\boldsymbol{x}'$ will come within the lowest np squared losses and $\boldsymbol{x}$ will no longer be in the np lowest square losses. We can now formulate our convergence conditions.
For all points outside the np lowest squared losses. There exists no point within the np lowest squared losses such that for any $k \in \mathbb{N}$, the points within the np lowest squared losses do not change.

5 Optimizing for the Sub-Quantile

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

$$t_{k+1} = \hat{\boldsymbol{\nu}}_{np} \tag{25}$$

$$t_{k+1} = \frac{1}{np} \sum_{i=1}^{np} \hat{\boldsymbol{\nu}}_i \tag{26}$$

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 25. In Algorithm 2, we use the same training procedure but modify the t-update as described in equation 26.

Algorithm 1: Sub-Quantile Optimization where $t_{k+1} = \boldsymbol{\nu}_{np}$

Input: Training iterations m, Quantile p, Corruption Percentage ϵ, Input Parameters d
Output: Trained Parameters, $\boldsymbol{\theta}$
Data: Inliers: $y|x \sim \mathcal{N}(x^2 - x + 2, 0.01)$, Outliers: $y|x \sim \mathcal{N}(-x^2 + x + 4, 0.01)$

$\boldsymbol{\theta}_1 \leftarrow \mathcal{N}(0, \sigma)^d$
for $k \in 1, 2, \ldots, m$ **do**
  $\boldsymbol{\nu} = (\boldsymbol{X}\boldsymbol{\theta}_k - \boldsymbol{y})^2$
  $\hat{\boldsymbol{\nu}} = sorted(\boldsymbol{\nu})$
  $t_{k+1} = \hat{\boldsymbol{\nu}}_{np}$
  $L \coloneqq \sum_{i=1}^{np} \boldsymbol{x}_i^T \boldsymbol{x}_i$
  $\alpha \coloneqq \frac{1}{2L}$
  $\boldsymbol{\theta}_{k+1} = \boldsymbol{\theta}_k - \alpha \nabla_{\boldsymbol{\theta}_k} g(t_{k+1}, \boldsymbol{\theta}_k)$
end
return $\frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{\theta}_m^T \boldsymbol{x}_i - y_i)^2$

Algorithm 2: Sub-Quantile Optimization where $t_{k+1} = \frac{1}{np}\sum_{i=1}^{np} \boldsymbol{\nu}_i$

Input: Training iteration, m, Quantile p, Corruption Percentage ϵ, Input Parameters d
Output: Trained Parameters, $\boldsymbol{\theta}$
Data: Inliers: $y|x \sim \mathcal{N}(x^2 - x + 2, 0.01)$, Outliers: $y|x \sim \mathcal{N}(-x^2 + x + 4, 0.01)$

$\boldsymbol{\theta}_1 \leftarrow \mathcal{N}(0, \sigma)^d$
for $k \in 1, 2, \ldots, m$ **do**
  $\boldsymbol{\nu} = (\boldsymbol{X}\boldsymbol{\theta}_k - \boldsymbol{y})^2$
  $\hat{\boldsymbol{\nu}} = sorted(\boldsymbol{\nu})$
  $t_{k+1} = \frac{1}{np}\sum_{i=1}^{np} \hat{\boldsymbol{\nu}}_i$
  $\boldsymbol{\theta}_{k+1} = \boldsymbol{\theta}_k - \alpha \nabla_{\boldsymbol{\theta}_k} g(t_{k+1}, \boldsymbol{\theta}_k)$
end
return $\frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{\theta}_m^T \boldsymbol{x}_i - y_i)^2$

5.1 Synthetic Data

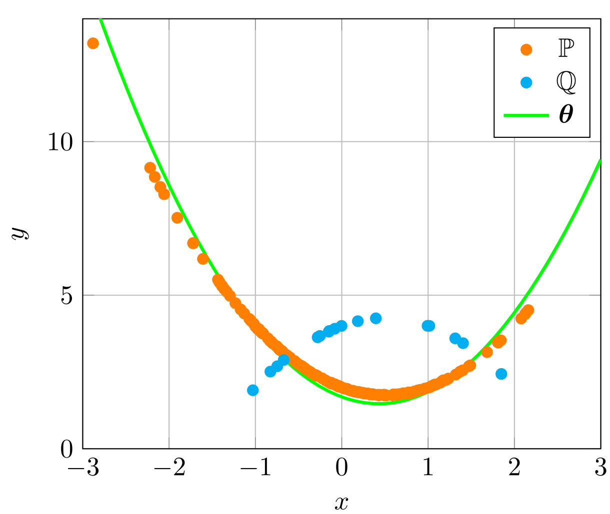

Figure 1: Quadratic Regression, $p = 0.9$

In our first synthetic experiment, we run Algorithm 1 on synthetically generated quadratic data.

5.2 Real Data

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

Author Contributions

Acknowledgments

References

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

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

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

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

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

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

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

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`.

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

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

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

A Assumptions

B General Properties of Sub-Quantile Linear Regression

B.1 Proof of Lemma 3.1.1

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

$$\partial(-g(t,\boldsymbol{\theta})) = \partial\left(-t+\frac{1}{np}\sum_{i=1}^{n}(t-\hat{\boldsymbol{\nu}}_i)^+\right) \tag{27}$$

$$= \partial(-t)+\partial\left(\frac{1}{np}\sum_{i=1}^{n}(t-\hat{\boldsymbol{\nu}}_i)^+\right) \tag{28}$$

$$= -1+\frac{1}{np}\sum_{i=1}^{n}\partial(t-\hat{\boldsymbol{\nu}}_i)^+ \tag{29}$$

$$= -1+\frac{1}{np}\sum_{i=1}^{n}\left\{\begin{array}{ll}1, & \text{if } t>\hat{\boldsymbol{\nu}}_i\\ 0, & \text{if } t<\hat{\boldsymbol{\nu}}_i\\ \left[0,1\right], & \text{if } t=\hat{\boldsymbol{\nu}}_i\end{array}\right\} \tag{30}$$

$$= 0 \text{ when } t=\hat{\boldsymbol{\nu}}_{np} \tag{31}$$

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

B.2 Proof of Lemma 3.1.2

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

$$\nabla_{\boldsymbol{\theta}_k} g(t_{k+1},\boldsymbol{\theta}_k) = \nabla_{\boldsymbol{\theta}_k}\left(\boldsymbol{\nu}_{np}-\frac{1}{np}\sum_{i=1}^{n}(\boldsymbol{\nu}_{np}-(\boldsymbol{\theta}_k^T\boldsymbol{x}_i-y_i)^2)^+\right) \tag{32}$$

$$= \nabla_{\boldsymbol{\theta}_k}\left((\boldsymbol{\theta}_k^T\boldsymbol{x}_{np}-y_{np})^2-\frac{1}{np}\sum_{i=1}^{n}\left((\boldsymbol{\theta}_k^T\boldsymbol{x}_{np}-y_{np})^2-(\boldsymbol{\theta}_k^T\boldsymbol{x}_i-y_i)^2\right)^+\right) \tag{33}$$

$$= \nabla_{\boldsymbol{\theta}_k}(\boldsymbol{\theta}_k^T\boldsymbol{x}_{np}-y_{np})^2-\frac{1}{np}\sum_{i=1}^{n}\nabla_{\boldsymbol{\theta}_k}\left((\boldsymbol{\theta}_k^T\boldsymbol{x}_{np}-y_{np})^2-(\boldsymbol{\theta}_k^T\boldsymbol{x}_i-y_i)^2\right)^+ \tag{34}$$

$$\begin{aligned}&= 2\boldsymbol{x}_{np}(\boldsymbol{\theta}_k^T\boldsymbol{x}_{np}-y_{np})-\frac{1}{np}\sum_{i=1}^{n}2\boldsymbol{x}_{np}(\boldsymbol{\theta}_k^T\boldsymbol{x}_{np}-y_{np})\\ &\qquad -2\boldsymbol{x}_i(\boldsymbol{\theta}_k^T\boldsymbol{x}_i-y_i)\left\{\begin{array}{ll}1, & \text{if } t>v_i\\ 0, & \text{if } t<v_i\\ \left[0,1\right], & \text{if } t=v_i\end{array}\right\}\end{aligned} \tag{35}$$

$$= 2\boldsymbol{x}_{np}(\boldsymbol{\theta}_k^T\boldsymbol{x}_{np}-y_{np})-\frac{1}{np}\sum_{i=1}^{np}2\boldsymbol{x}_{np}(\boldsymbol{\theta}_k^T\boldsymbol{x}_{np}-y_{np})-2\boldsymbol{x}_i(\boldsymbol{\theta}_k^T\boldsymbol{x}_i-y_i) \tag{36}$$

$$= 2\boldsymbol{x}_{np}(\boldsymbol{\theta}_k^T\boldsymbol{x}_{np}-y_{np})-2\boldsymbol{x}_{np}(\boldsymbol{\theta}_k^T\boldsymbol{x}_{np}-y_{np})+\frac{1}{np}\sum_{i=1}^{np}2\boldsymbol{x}_i(\boldsymbol{\theta}_k^T\boldsymbol{x}_i-y_i) \tag{37}$$

$$= \frac{1}{np}\sum_{i=1}^{np}2\boldsymbol{x}_i(\boldsymbol{\theta}_k^T\boldsymbol{x}_i-y_i) \tag{38}$$

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

B.3 Proof of Lemma ??

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

$$\mu_P = \frac{1}{n(1-\epsilon)} \sum_{i=1}^{n(1-\epsilon)} (\boldsymbol{\theta}^T \boldsymbol{p}_i - y_i)^2 \tag{39}$$

$$\mu_Q = \frac{1}{n\epsilon} \sum_{i=1}^{n\epsilon} (\boldsymbol{\theta}^T \boldsymbol{q}_i - y_i)^2 \tag{40}$$

$$\sigma_P^2 = \frac{1}{n(1-\epsilon)-1} \sum_{i=1}^{n(1-\epsilon)} (\mu_P - (\boldsymbol{\theta}^T \boldsymbol{p}_i - y_i)^2)^2 \tag{41}$$

$$\sigma_Q^2 = \frac{1}{n\epsilon - 1} \sum_{i=1}^{n\epsilon} (\mu_Q - (\boldsymbol{\theta}^T \boldsymbol{q}_i - y_i)^2)^2 \tag{42}$$

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

$$\mu_{\hat{P}} = (1-\epsilon)\mu_P + \epsilon\mu_Q \tag{43}$$

$$\sigma_{\hat{P}} = (1-\epsilon)^2\sigma_P^2 + \epsilon^2\sigma_Q^2 + \epsilon(1-\epsilon)Cov(P, Q) \tag{44}$$

$$= (1-\epsilon)^2\sigma_P^2 + \epsilon^2\sigma_Q^2 \tag{45}$$

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

$$Q_p(\hat{\mathbb{P}}) = \mu_{\hat{P}} + \Phi^{-1}(p)\sigma_{\hat{P}} \tag{46}$$

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

$$\mathbb{P}\left[P_i < Q_p(\hat{\mathbb{P}})\right] = \Phi\left(\frac{Q_p(\hat{\mathbb{P}}) - \mu_P}{\sigma_P}\right) \tag{47}$$

$$= \Phi\left(\frac{(1-\epsilon)\mu_P + \epsilon\mu_Q + \Phi^{-1}(p)((1-\epsilon)^2\sigma_P^2 + \epsilon^2\sigma_Q^2) - \mu_P}{\sigma_P}\right) \tag{48}$$

$$= \Phi\left(\frac{-\epsilon\mu_P + \epsilon\mu_Q + \Phi^{-1}(p)((1-\epsilon)^2\sigma_P^2 + \epsilon^2\sigma_Q^2)}{\sigma_P}\right) \tag{49}$$

□

B.4 Proof of Lemma 4.0.1

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

$$||\nabla g_i(\boldsymbol{\theta}') - \nabla g_i(\boldsymbol{\theta})|| \leq \beta||\boldsymbol{\theta}' - \boldsymbol{\theta}|| \tag{50}$$

The proof is as follows

$$||\nabla g_i(\boldsymbol{\theta}') - \nabla g_i(\boldsymbol{\theta})|| = ||2\boldsymbol{x}_i(\boldsymbol{\theta}'^T \boldsymbol{x}_i - y_i) - 2\boldsymbol{x}_i(\boldsymbol{\theta}^T \boldsymbol{x}_i - y_i)|| \tag{51}$$

$$= ||2\boldsymbol{x}_i(\boldsymbol{\theta}'^T \boldsymbol{x}_i) - 2\boldsymbol{x}_i y_i - 2\boldsymbol{x}_i(\boldsymbol{\theta}^T \boldsymbol{x}_i) + 2\boldsymbol{x}_i y_i|| \tag{52}$$

$$= ||2\boldsymbol{x}_i(\boldsymbol{\theta}'^T \boldsymbol{x}_i - \boldsymbol{\theta}^T \boldsymbol{x}_i)|| \tag{53}$$

$$= ||2\boldsymbol{x}_i||\,||\boldsymbol{\theta}'^T \boldsymbol{x}_i - \boldsymbol{\theta}^T \boldsymbol{x}_i|| \tag{54}$$

$$= 2||\boldsymbol{x}_i||^2||\boldsymbol{\theta}' - \boldsymbol{\theta}|| \tag{55}$$

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

B.5 Proof of Theorem 4.2

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

$$f_{\boldsymbol{x}}(\boldsymbol{\theta}_1) - f_{\boldsymbol{x}}(\boldsymbol{\theta}_0) = (\boldsymbol{\theta}_1^T \boldsymbol{x} - y)^2 - (\boldsymbol{\theta}_0^T - y)^2 \tag{56}$$

$$= (\boldsymbol{\theta}_1^T \boldsymbol{x})^2 - 2(\boldsymbol{\theta}_1^T \boldsymbol{x})y - (\boldsymbol{\theta}_0^T \boldsymbol{x})^2 + 2(\boldsymbol{\theta}_0^T \boldsymbol{x})y \tag{57}$$

$$= ((\boldsymbol{\theta}_0 - \alpha \nabla g(\boldsymbol{\theta}_0))^T \boldsymbol{x})^2 - 2((\boldsymbol{\theta}_0 - \alpha \nabla g(\boldsymbol{\theta}_0))^T \boldsymbol{x})y - (\boldsymbol{\theta}_0^T \boldsymbol{x})^2 + 2(\boldsymbol{\theta}_0^T \boldsymbol{x})y \tag{58}$$

Note $\boldsymbol{\theta}_1 = \boldsymbol{\theta}_0 - \alpha \nabla g(t_1, \boldsymbol{\theta})$ by Equation 9

$$= (\boldsymbol{\theta}_0^T \boldsymbol{x} - \alpha \nabla g(\boldsymbol{\theta}_0)^T \boldsymbol{x})^2 - 2(\boldsymbol{\theta}_0^T \boldsymbol{x})y + 2\alpha(\nabla g(\boldsymbol{\theta})^T \boldsymbol{x})y - (\boldsymbol{\theta}_0^T \boldsymbol{x})^2 + 2(\boldsymbol{\theta}_0^T \boldsymbol{x})y \tag{59}$$

$$= (\boldsymbol{\theta}_0^T \boldsymbol{x})^2 - 2\alpha(\boldsymbol{\theta}_0^T)(\nabla g(\boldsymbol{\theta})^T \boldsymbol{x}) + \alpha^2(\nabla g(\boldsymbol{\theta})^T \boldsymbol{x})^2 + 2\alpha(\nabla g(\boldsymbol{\theta})^T \boldsymbol{x})y - (\boldsymbol{\theta}_0^T \boldsymbol{x})^2 \tag{60}$$

$$= -2\alpha(\boldsymbol{\theta}_0^T)(\nabla g(\boldsymbol{\theta})^T \boldsymbol{x}) + \alpha^2(\nabla g(\boldsymbol{\theta})^T \boldsymbol{x})^2 + 2\alpha(\nabla g(\boldsymbol{\theta})^T \boldsymbol{x})y \tag{61}$$

$$= \alpha(\nabla g(\boldsymbol{\theta}_0)^T \boldsymbol{x})(-2(\boldsymbol{\theta}_0)^T \boldsymbol{x} + \alpha \nabla g(\boldsymbol{\theta}_0)^T \boldsymbol{x} + 2y) \tag{62}$$

$$= \alpha(\nabla g(\boldsymbol{\theta}_0)^T \boldsymbol{x})(-2(\boldsymbol{\theta}_0^T \boldsymbol{x} - y)) + \alpha \nabla g(\boldsymbol{\theta}_0)^T \boldsymbol{x}) \tag{63}$$

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

$$= -\alpha \nabla g(\boldsymbol{\theta}_0)^T \nabla f_{\boldsymbol{x}}(\boldsymbol{\theta}_0) + \alpha^2(\nabla g(\boldsymbol{\theta}_0)^T \boldsymbol{x})^2 \tag{64}$$

$$= -\alpha \left(||\nabla g(\boldsymbol{\theta}_0)|| ||\nabla f_{\boldsymbol{x}}(\boldsymbol{\theta}_0)|| \cos(\omega) - \alpha ||\nabla g(\boldsymbol{\theta}_0)||^2 ||\boldsymbol{x}||^2 \cos^2(\eta) \right) \tag{65}$$

$$= -\alpha ||\nabla g(\boldsymbol{\theta}_0)|| \left(||f_{\boldsymbol{x}}(\boldsymbol{\theta}_0)|| \cos(\omega) - \alpha ||\nabla g(\boldsymbol{\theta}_0)|| ||\boldsymbol{x}||^2 \cos^2(\eta) \right) \tag{66}$$

$$= -\alpha ||\nabla g(\boldsymbol{\theta}_0)|| \left(2||\boldsymbol{x}|| |\boldsymbol{\theta}_0^T \boldsymbol{x} - y| \cos(\omega) - \alpha ||\nabla g(\boldsymbol{\theta}_0)|| ||\boldsymbol{x}||^2 \cos^2(\eta) \right) \tag{67}$$

$$= -\alpha ||\nabla g(\boldsymbol{\theta}_0)|| ||\boldsymbol{x}|| \left(2 ||\boldsymbol{\theta}_0^T \boldsymbol{x} - y| \cos(\omega) - \alpha ||\nabla g(\boldsymbol{\theta}_0)|| ||\boldsymbol{x}|| \cos^2(\eta) \right) \tag{68}$$

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

$$||\boldsymbol{x}'||(2|\boldsymbol{\theta}_0^T \boldsymbol{x}' - y| \cos(\omega') - \alpha ||\nabla g(\boldsymbol{\theta})|| ||\boldsymbol{x}'|| \cos^2(\eta')) > ||\boldsymbol{x}||(2|\boldsymbol{\theta}_0^T \boldsymbol{x} - y| \cos(\omega) - \alpha ||\nabla g(\boldsymbol{\theta})|| ||\boldsymbol{x}|| \cos^2(\eta)) \tag{69}$$

Here we note that α is a very small term, $\alpha = \frac{1}{2L}$ where $L = ||\boldsymbol{X}^T \boldsymbol{X}||$ So equation 69 can be approximately simplied.

$$||\boldsymbol{x}'|| |\boldsymbol{\theta}_0^T \boldsymbol{x}' - y'| \cos(\omega') > ||\boldsymbol{x}|| |\boldsymbol{\theta}_0^T \boldsymbol{x} - y| \cos(\omega) \tag{70}$$

This completes the proof. □

C PROOFS FOR CONVERGENCE

C.1 PROOF OF LEMMA 4.0.4

This is a standard proof in Optimization Theory. The objective function $g(\boldsymbol{\theta}, t)$ is L-smooth w.r.t $\boldsymbol{\theta}$ iff

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

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

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

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

$$\leq \left\|\frac{2}{np}\sum_{i=1}^{np} \|\boldsymbol{x}_i\|^2\right\| \left\|\boldsymbol{\theta}_k^{'T} - \boldsymbol{\theta}_k^{T}\right\| \quad (75)$$

$$= L\left\|\boldsymbol{\theta}_k^{'T} - \boldsymbol{\theta}_k^{T}\right\| \quad (76)$$

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

This concludes the proof.