Robust Linear Regression by Subquantile Minimization

Arvind Rathnashyam Fatih Orhan Josh Myers Jake Herman

Rensselaer Polytechnic Institute

(rathna, orhanf, myersj5, hermaj2)@rpi.edu

ML and Optimization Group U5

April 23, 2023

Huber Contamination Model

Problem

The Huber Contamination Model is the following:

$$\hat{P} = (1 - \varepsilon)P + \varepsilon Q \text{ where } \varepsilon \in (0, 0.5)$$

where P and Q represent the general linear models

$$\boldsymbol{y}_P = \mathbf{P}\boldsymbol{\beta}_P + \epsilon_P$$

$$\boldsymbol{y}_Q = \mathbf{Q}\boldsymbol{\beta}_Q + \epsilon_Q$$

$\boldsymbol{\beta}_P$ and $\boldsymbol{\beta}_\mathbf{Q}$ are oracle regressors and ϵ_P and ϵ_Q represent 0-centered gaussian noise.

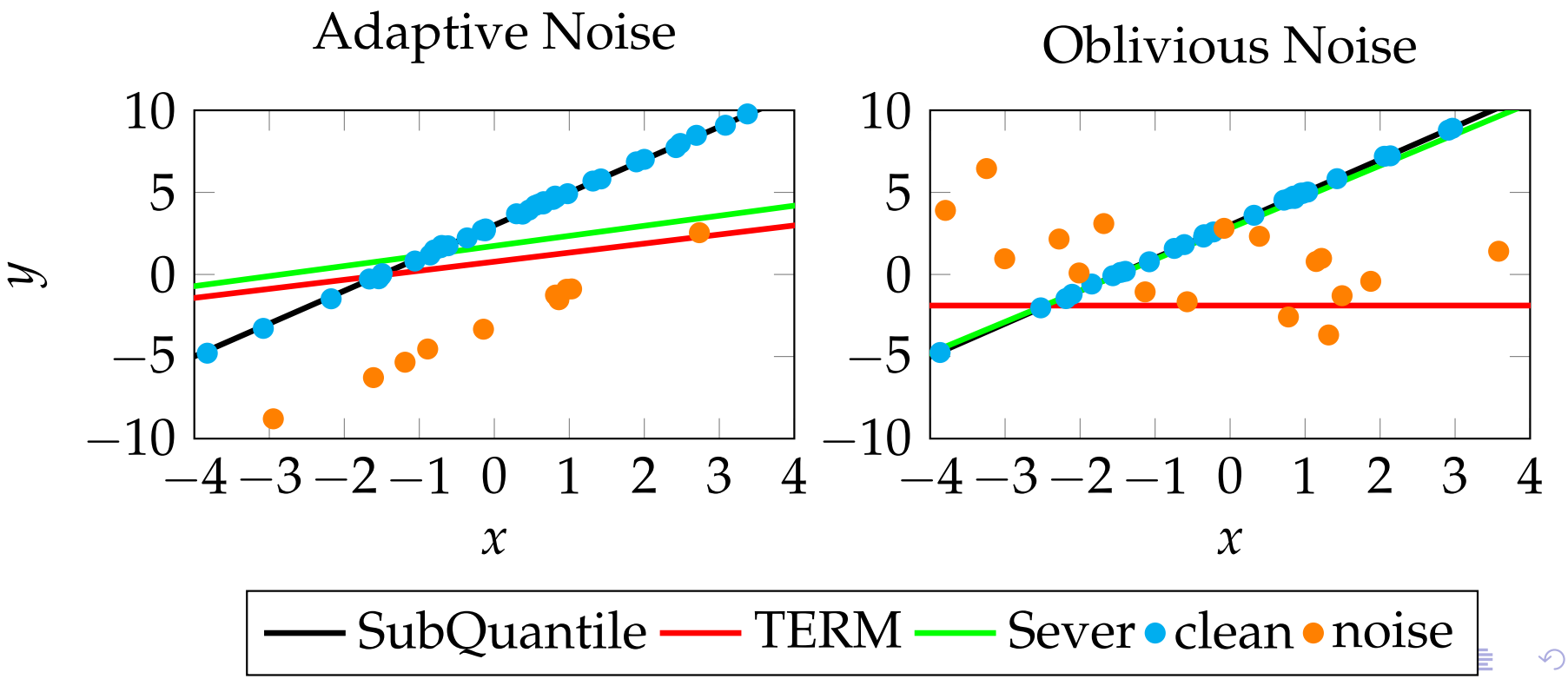
Our goal is to learn a model that learns a good distribution of P from $\hat{P}$

Motivation

Definition

Oblivious Noise is noise generated independent of the target distribution

Adaptive Noise is noise which is generated with knowledge of the target distribution.



Consistent Robust Regression: NeurIPS 2017

- CRR attempts to improve robust least squares regression by correcting the errors with hard thresholding
- This is applicable when there is a set of oblivious noise in the data of the form

$$\boldsymbol{y} = \boldsymbol{w}X + \boldsymbol{b}^*$$

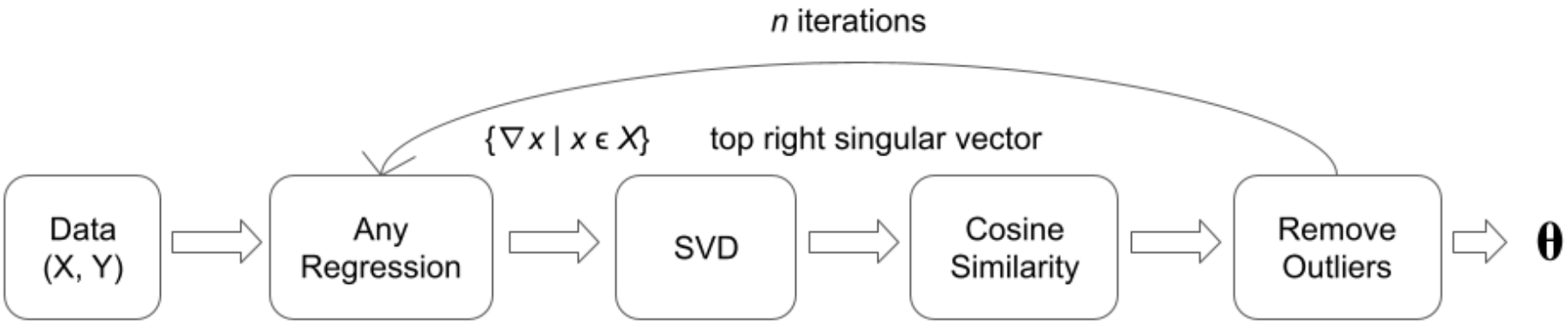
  They want to recover $\boldsymbol{w}$ where $\boldsymbol{b}^*$ is a k-sparse corruption vector.
- They solve the following optimization problem:

$$\min_{\|\boldsymbol{b}\|_0 \leq k^*} f(\boldsymbol{b}) = \frac{1}{2} \left\| (I - X(X^T X)^{-1} X^T)(\boldsymbol{y} - \boldsymbol{b}) \right\|_2^2$$

- For $\boldsymbol{v} \in \mathbb{R}^n$, define hard thresholding operator $\hat{\boldsymbol{v}} = HT_k(\boldsymbol{v})$ where $\hat{\boldsymbol{v}}_i = \boldsymbol{v}_i$ if $\sigma_v^{-1}(i) < k$ for where σ represents the descending elements of the $\boldsymbol{v}$ otherwise $\hat{v}_i = 0$.
- Order complexity is $\mathcal{O}(d^3 + nd)$
- Requires data to be heavy-tailed

Sever: ICML 2019

- For outliers to have an affect on data the gradient must be :
  - Large in magnitude
  - Pointing in a general direction
- Detect using singular-value decomposition on the gradients
- Small datasets, hyperparameter tuning, costly iterations: $\mathcal{O}(n^3 + nd^2)$



Tilted Empirical Risk Minimization: ICLR 2021

- Addresses the deficies of ERM such as learning in the presence of corrupted or imbalanced data
- Minimize the objective function:

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

- Empirically, this method requires approximately twice as many iterations as standard ERM.
- The method is not overly intuitive, as the convexity is lost when introducing the exponential.
- To choose t one must grid-search, however the authors have recommended $t = -2$ in the case of robust regression.

Probability Theory Preliminaries of the Subquantile

1. The quantile is given as the following:

$$\mathcal{Q}_p = \inf \{x \in \mathbb{R} : p \leq F(x)\}$$

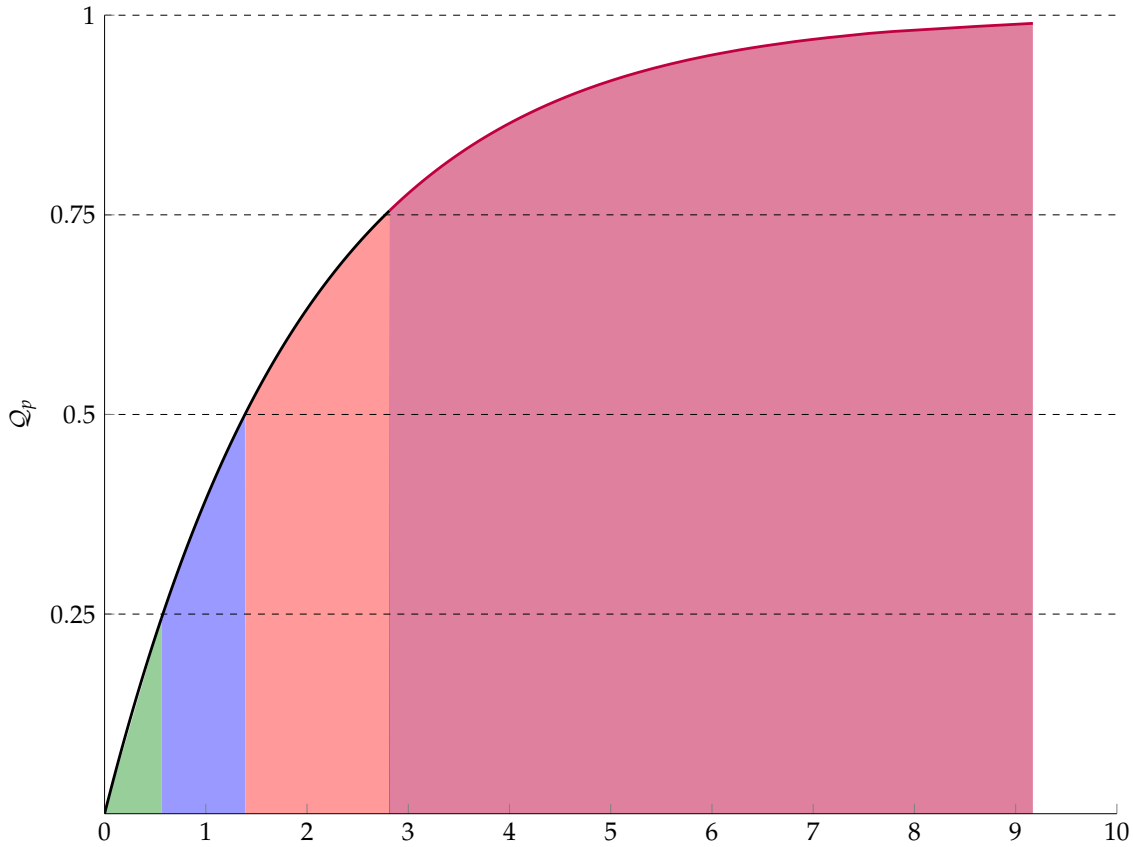


Figure: Quantile plot of the χ^2 distribution with 2 degrees of freedom

Subquantile Optimization Problem

We are now able to define the optimization problem we will solve:

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

The objective function is:

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

Algorithm:

$$t_{(k+1)} = \underset{t\in\mathbb{R}}{\operatorname{argmax}}\, g(t, \boldsymbol{\theta}_{(k)})$$

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

Lemma

$$\nabla_{\boldsymbol{\theta}} g(t, \boldsymbol{\theta}_{(k)}) = \frac{1}{np} \sum_{i=1}^{np} 2\boldsymbol{x}_i \left(\boldsymbol{\theta}_{(k)}^{\top} \boldsymbol{x}_i - y_i \right)$$

where $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^{np}$ *represent the np points in the dataset with the lowest loss.*

Lemma

$$\underset{t \in \mathbb{R}}{\operatorname{argmax}}\, g(t, \boldsymbol{\theta}_{(k)}) = \boldsymbol{\nu}_{np}$$

where $\boldsymbol{\nu}_{np}$ *represents the npth highest loss in the dataset.*

Here we are able to see the true nature of Subquantile Optimization. Each iteration we are optimizing over the points within the lowest np errors.

General Theory

Lemma

Let $\hat{\boldsymbol{\nu}}$ be the losses of all the data ordered in ascending order. Then it follows:

$$\arg\max_{t\in\mathbb{R}} g(t,\boldsymbol{\theta}) = \hat{\boldsymbol{\nu}}_{np} \tag{2}$$

Therefore, in each maximizing step we take the element with the npth largest loss as $t_{(k+1)}$. With this choice of t_{k+1} it then follows:

Lemma

The derivative with respect to θ at the kth iteration step is:

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

where $\boldsymbol{x_1}, \ldots, \boldsymbol{x}_{np}$ represent the np points with the lowest squared error.

Optimization

Definition

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

Lemma

Any Local Nash Equlibrium satisfies $\nabla_{\boldsymbol{\theta}} g(t_{(k)}, \boldsymbol{\theta}_{(k)}) = \mathbf{0}$ *and* $\nabla_t g(t_{(k)}, \boldsymbol{\theta}_{(k)}) = 0$

We first give intuition on what it means to be at a Local Nash Equilibrium. It means we have a $\boldsymbol{\theta}$ that gives minimizes ERM over the points within the lowest np errors.

Optimization

Subquantile Optimization continuously optimizes over the np data points with the lowest squared error. In other words, we are trying to minimize the min-loss over the p-quantile. This gives a nice characterization of the optimization problem.

Theorem

The Subquantile min-max optimization problem is equivalent to the following min-min optimization problem:

$$\boldsymbol{\theta}_{SM} = \arg\min_{\boldsymbol{\theta} \in \mathbb{R}^d} \min_{S \in \prod(X)} \|S\boldsymbol{\theta} - y_S\|_2^2$$

where $\prod$ *represents the* $\binom{n}{np}$ *matrices of* np *rows of* X

Proof.

If we let the np elements with error less than $t_{(k+1)}$ be rows of the matrix S, we see we have the same optimization problem. □

The reasoning for this characterization of this optimization algorithm is now it is easier to show convergence. There are multiple ways to solve for this algorithm. In the ridge regression case, we solve with the following algorithm:

Algorithm 1: Sub-Quantile Minimization for Ridge Regression

Input: Training Iterations T,Quantile p

Output: Trained Parameters, $\boldsymbol{\theta}_{(T)}$

$\boldsymbol{\theta}_{(0)} \leftarrow (X^T X + \lambda I)^{-1} X^T y$

for $k \in \{1, 2, \ldots, T\}$ **do**

  $\hat{\boldsymbol{\nu}} \leftarrow (X\boldsymbol{\theta}_{(k)} - y)^2$

  $t_{(k+1)} \leftarrow \hat{\boldsymbol{\nu}}_{np}$

  Subquantile Matrix minimization step

  $S_{(k)} \leftarrow \Big\| \boldsymbol{x}_i$ if $(\boldsymbol{x}_i\boldsymbol{\theta}_{(k)} - y_i)^2 \leq$

  $t_{(k+1)}$ where $\Big\|$ represents the concatenation operator

  $\boldsymbol{\theta}$ minimization step $\boldsymbol{\theta}_{(k+1)} \leftarrow (S^T S + \lambda I)^{-1} S^T y_S$

end

return $\boldsymbol{\theta}_{(T)}$

Proof of Convergence for Linear Regression

Proof.

Let us consider the objective function:

$$f(\boldsymbol{\theta}, S) = \|S\boldsymbol{\theta} - y_S\|_2^2$$

To show convergence it is sufficient to show

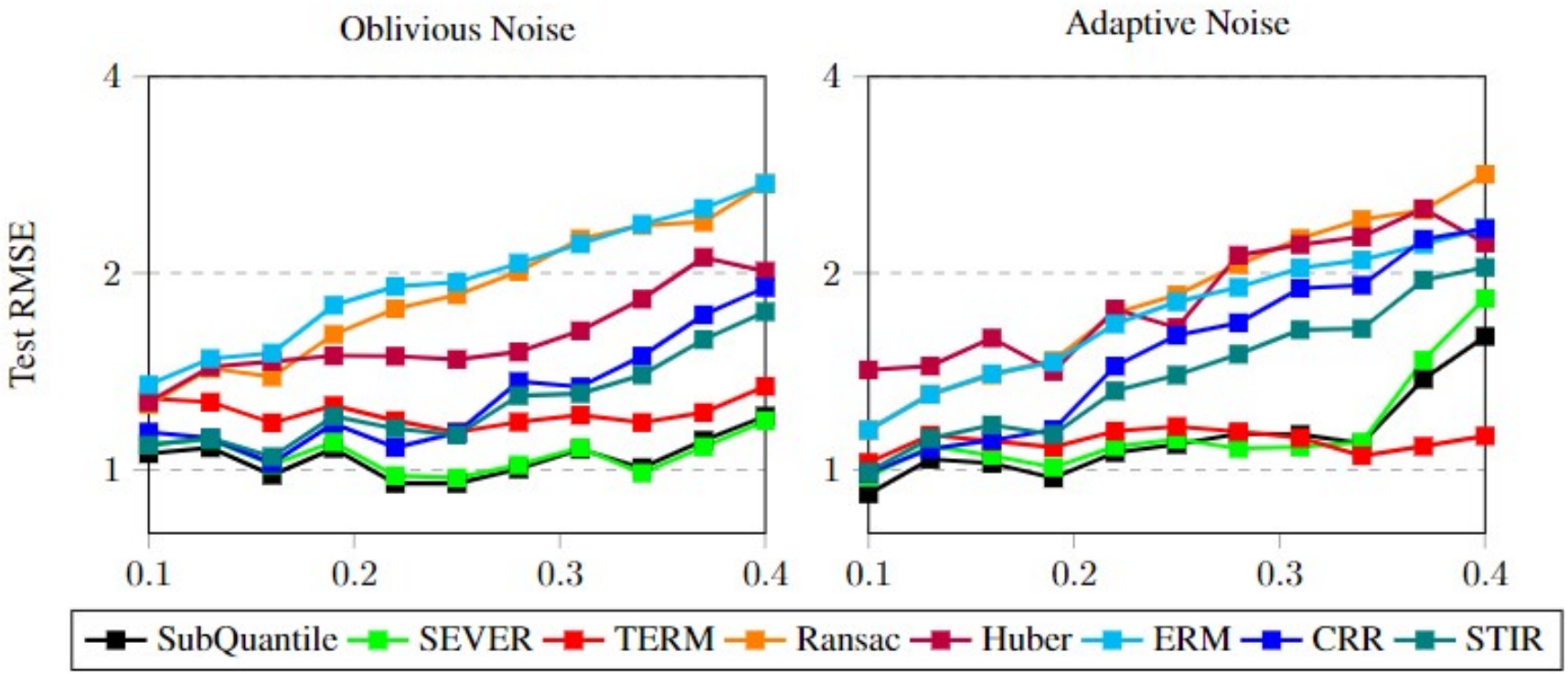
$$f(\boldsymbol{\theta}_{(t+1)}, S_{(t+1)}) \leq f(\boldsymbol{\theta}_{(t)}, S_{(t)})$$

We know $S_{(t+1)} = \arg\min_{S \in \prod(X)} f(\boldsymbol{\theta}_{(t)}, S)$, thus
$f(\boldsymbol{\theta}_{(t)}, S_{(t+1)}) \leq f(\boldsymbol{\theta}_{(t)}, S_{(t)})$
Since $\boldsymbol{\theta}_{(t+1)} = (S_{(t+1)}^T S_{(t+1)})^{-1} S_{(t+1)}^T \boldsymbol{y}_S$, and linear regression is a convex problem, thus $f(\boldsymbol{\theta}_{(t+1)}, S_{t+1}) \leq f(\boldsymbol{\theta}_{(t)}, S_{t+1})$.
Therefore we have proved $f(\boldsymbol{\theta}_{(t)}, S_{(t+1)}) \leq f(\boldsymbol{\theta}_{(t)}, S_{(t)})$ and
$f(\boldsymbol{\theta}_{(t+1)}, S_{(t+1)}) \leq f(\boldsymbol{\theta}_{(t)}, S_{(t+1)})$
It thus follows $f(\boldsymbol{\theta}_{(t+1)}, S_{(t+1)}) \leq f(\boldsymbol{\theta}_{(t)}, S_{(t)})$ □

Drug Discovery

Objectives	Test RMSE (`Drug Discovery`)			
	$\epsilon = 0.1$	$\epsilon = 0.2$	$\epsilon = 0.3$	$\epsilon = 0.4$
ERM	$1.303_{(0.0665)}$	$1.790_{(0.0849)}$	$2.198_{(0.0645)}$	$2.623_{(0.1010)}$
CRR [2]	$1.079_{(0.0899)}$	$1.125_{(0.0832)}$	$1.385_{(0.1372)}$	$1.725_{(0.1136)}$
STIR [4]	$1.087_{(0.1256)}$	$1.167_{(0.0750)}$	$1.403_{(0.0987)}$	$1.668_{(0.1142)}$
Robust Risk [3]	$1.176_{(0.1110)}$	$1.336_{(0.1882)}$	$1.437_{(0.1723)}$	$1.800_{(0.0820)}$
SMART [5]	$1.094_{(0.1065)}$	$1.323_{(0.0758)}$	$1.578_{(0.0799)}$	$1.984_{(0.2020)}$
TERM [6]	$\mathbf{1.029_{(0.0707)}}$	$1.126_{(0.0776)}$	$1.191_{(0.1091)}$	$1.201_{(0.1409)}$
SEVER [1]	$1.043_{(0.0970)}$	$\mathbf{1.067_{(0.0457)}}$	$\mathbf{1.071_{(0.0807)}}$	$\mathbf{1.138_{(0.1162)}}$
Huber [7]	$1.412_{(0.0474)}$	$1.501_{(0.2918)}$	$2.231_{(0.9054)}$	$2.247_{(1.0399)}$
RANSAC [8]	$1.238_{(0.0529)}$	$1.643_{(0.1331)}$	$2.092_{(0.1935)}$	$2.679_{(0.1365)}$
SubQuantile($p = 1 - \epsilon$)	$\mathbf{0.966_{(0.1119)}}$	$\mathbf{1.002_{(0.1025)}}$	$\mathbf{1.010_{(0.0630)}}$	$\mathbf{1.082_{(0.1066)}}$
Genie ERM	$0.960_{(0.0845)}$	$0.982_{(0.0842)}$	$1.006_{(0.0879)}$	$1.030_{(0.0578)}$

Table: `Drug Discovery` Dataset. Empirical Risk over P with oblivious noise


Oblivious Noise
Adaptive Noise
Test RMSE
4
2
1
0.1
0.2
0.3
0.4
SubQuantile
SEVER
TERM
Ransac
Huber
ERM
CRR
STIR

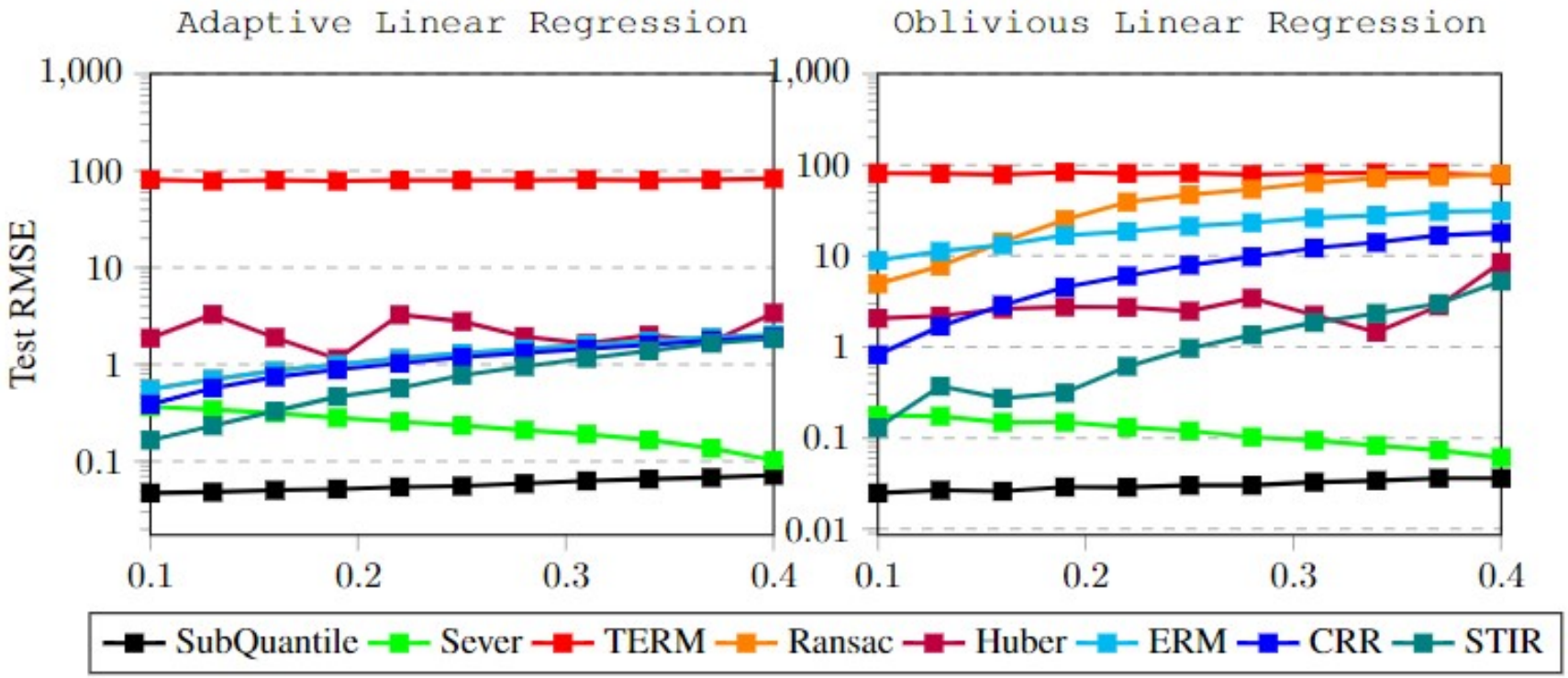


Figure 2: Structured Linear Regression & Noisy Linear Regression Datasets

Further Work

1. Based on our emprical results, SubQuantile optimization converges very fast (within less than 20 iterations). However, we have not yet shown the convergence guarantee theoretically. We will be working on finding an upper bound in terms of β_P, β_Q, and ϵ.
2. Applications of Subquantile Minimization in other Machine Learning domains such as robust classification in datasets such as CIFAR-10.
3. Proof of convergence for a stochatic Subquantile Minimization Algorithm.

References

- Diakonikolas, I., Kamath, G., Kane, D., Li, J., Steinhardt, J. & Stewart, A. Sever: A Robust Meta-Algorithm for Stochastic Optimization. *Proceedings Of The 36th International Conference On Machine Learning*. pp. 1596-1606 (2019)
- Bhatia, K., Jain, P., Kamalaruban, P. & Kar, P. Consistent Robust Regression. *Advances In Neural Information Processing Systems*. **30** (2017), https://proceedings.neurips.cc/paper-files/paper/2017/file/e702e51da2c0f5be4dd354bb3e295d37-Paper.pdf
- Osama, M., Zachariah, D. & Stoica, P. Robust Risk Minimization for Statistical Learning from Corrupted Data. *IEEE Open Journal Of Signal Processing*. **1** pp. 287-294 (2020)
- Mukhoty, B., Gopakumar, G., Jain, P. & Kar, P. Globally-convergent Iteratively Reweighted Least Squares for Robust Regression Problems. *Proceedings Of The Twenty-Second International Conference On Artificial Intelligence And Statistics*. **89** pp. 313-322 (2019,4,16), https://proceedings.mlr.press/v89/mukhoty19a.html
- Awasthi, P., Das, A., Kong, W. & Sen, R. Trimmed Maximum Likelihood Estimation for Robust Learning in Generalized Linear Models. (arXiv,2022), https://arxiv.org/abs/2206.04777
- Li, T., Beirami, A., Sanjabi, M. & Smith, V. Tilted empirical risk minimization.

Please refer to the paper for more details