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

Presentation Overview

1. Robust Regression
   - Huber Contamination Model
   - Blocks
   - Preliminaries

2. Empirical Results
   - Table

3. Mathematics

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 $\boldsymbol{\beta}_Q$ 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 $\hat{P}$

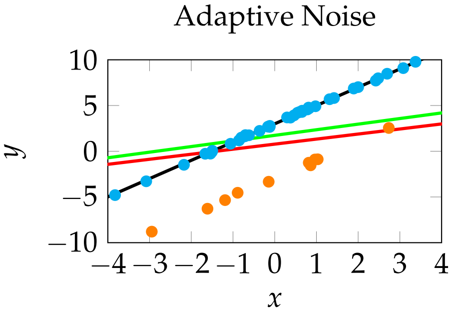
Motivation

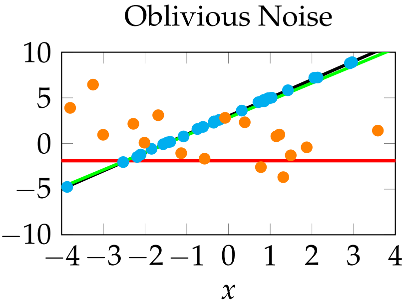
Definition

In the Robust Statistics literature:

Oblivious Noise is noise generated independent of the target distribution

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





Theorem

The expected optimal parameters of the corrupted model $\hat{P}$

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

where $\boldsymbol{X}^{\dagger} \triangleq \left(\boldsymbol{X}^{\top}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\top}$, i.e. the Moore-Penrose Inverse.

This theorem motivates our reasoning for optimizing over the Subquantile. We want a method to reduce ε.

Statistical Preliminaries of the Subquantile

1. The quantile is given as the following:

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

2. Let ℓ be the loss functions. We can now define risk as:

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

3. The p-Quantile of the Empirical Risk is given by:

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

4. For the least squares regression case:

$$\mathbb{L}_p = \max_{t\in\mathbb{R}}\left\{t - \frac{1}{np}\sum_{i=1}^{n}\left(t - \left(\boldsymbol{\theta}^\top \boldsymbol{x}_i - y_i\right)\right)^+\right\}$$

Subquantile Optimization Problem

We are know 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:

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

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

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

where y_{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.

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 $\nabla_{\boldsymbol{\theta}} g(t_{(k)}, \boldsymbol{\theta}_{(\mathbf{k})}) = \mathbf{0}$ *and* $\nabla_{t} g(t_{(k)}, \boldsymbol{\theta}_{(\mathbf{k})}) = 0$

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.

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 *np*th 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.

Prior Works

- [1] Sever computes the gradient of losses in each iteration. This incurs a $\mathcal{O}(dn^2)$ time complexity per iteration. Furthermore, points thrown out in early iterations are not resampled.
- [2] CRR runs in order complexity $\mathcal{O}(d^3 + nd)$. The theoretical guarantees of CRR are given only in the case of Oblivious Noise.
- Our work computes a np partition of the loss in each iteration. This incurs only a $\mathcal{O}(n)$ time complexity per iteration. Subquantile Minimization is novel in that it resamples points that may not have been in previous Subquantiles.

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]	$\mathbf{1.029}_{(\mathbf{0.0707})}$	$1.126_{(0.0776)}$	$1.191_{(0.1091)}$	$1.201_{(0.1409)}$
SEVER [1]	$1.043_{(0.0970)}$	$\mathbf{1.067}_{(\mathbf{0.0457})}$	$\mathbf{1.071}_{(\mathbf{0.0807})}$	$\mathbf{1.138}_{(\mathbf{0.1162})}$
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($p = 1 - \epsilon$)	$\mathbf{0.966}_{(\mathbf{0.1119})}$	$\mathbf{1.002}_{(\mathbf{0.1025})}$	$\mathbf{1.010}_{(\mathbf{0.0630})}$	$\mathbf{1.082}_{(\mathbf{0.1066})}$
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

Definitions & Examples

Definition

A prime number is a number that has exactly two divisors.

Example

- 2 is prime (two divisors: 1 and 2).
- 3 is prime (two divisors: 1 and 3).
- 4 is not prime (three divisors: 1, 2, and 4).

You can also use the `theorem`, `lemma`, `proof` and `corollary` environments.

Verbatim

Example (Theorem Slide Code)

```
\begin{frame}
\frametitle{Theorem}
\begin{theorem}[Mass--energy equivalence]
$E = mc^2$
\end{theorem}
\end{frame}
```

Slide without title.

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.

The End

Questions? Comments?