Subquantile Minimization: Theory and Applications for Machine Learning

Arvind Rathnashyam* Alex Gittens†

September 1, 2023

Abstract

In this paper we propose Subquantile Minimization for adversarial corruption in the training set. We study the Huber Contamination Problem for the Kernel Learning Problem where the distribution is formed as, $\hat{\mathbb{P}} = (1-\varepsilon)\mathbb{P} + \varepsilon\mathbb{Q}$, and we want to find the function $\inf_f \mathbb{E}_{\mathbf{x}\in\mathbb{P}}[\ell_f(\mathbf{x})]$, from the noisy distribution, $\hat{\mathbb{P}}$. We assume the adversary has knowledge of the true distribution of $\mathbb{P}$, and is able to corrupt the covariates and the labels of ε samples. To our knowledge, we are the first to study the problem of general kernel learning in the Huber Contamination Model. We theoretically analyze Kernel Ridge Regression and Kernel Classification and empirically show the strength of Subquantile Minimization. Furthermore, we run experiments on various datasets and compare with the state-of-the-art algorithms to show the superior performance of Subquantile Minimization.

*CS, Rensselear Polytechnic Institute, rathna@rpi.edu
†CS, Rensselaer Polytechnic Institute, gittea@rpi.edu

1 Introduction

There has been extensive study of algorithms to learn the target distribution from a Huber ε-Contaminated Model for a Generalized Linear Model (GLM), (Diakonikolas u. a., 2019), (Awasthi u. a., 2022), (Li u. a., 2021). In many real-world applications, linear models are insufficient to model the data. Therefore, we introduce the problem of Robust Learning for Kernel Learning.

Definition 1. (**Huber ϵ-Contamination Model**). Given a corruption parameter $0 < \epsilon < 0.5$, a data matrix, $\mathbf{X}$ and labels $\mathbf{y}$. An adversary is allowed to inspect all samples and modify $n(1-\epsilon)$ samples arbitrarily. The algorithm is then given the ϵ-corrupted data matrix $\mathbf{X}$ and $\mathbf{y}$ as training data.

Contributions

1. We propose a gradient-descent based algorithm for robust kernel learning in the Huber ϵ-Contamination Model which is fast.
2. We provide a theoretical analysis and give error bounds for kernel ridge regression and kernel classification.

1.1 Related Work

Robust Algorithms

(Diakonikolas u. a., 2019) proposed a robust meta-algorithm which filters points based their outlier likelihood score, which they define as the projection of the gradient of the point on to the top right singular vector of the Singular Value Decomposition of the Gradient of Losses. Empirically SEVER is strong in adversarially robust linear regression and Singular Vector Machines. SEVER however requires a base learner execution and SVD calculation for each iteration, thus it does not scale well for large scale applications.

(Li u. a., 2021) proposed optimization over the Tilted Empirical Loss. This is done by minimization of an exponentially weighted functional of the traditional Empirical Risk. Their involves a hyperparameter t, negative values of t trains more robustly, whereas positive values of t trains more fairly. This empirically works well in machine learning applications such as Noisy Annotation. The issue with introducing the exponential smoothing into the ERM function is the lack of interpretability.

(Awasthi u. a., 2022) theoretically analyzed the Trimmed Maximum Likelihood Estimator algorithm in General Linear Models, including Gaussian Regression. They were able to show the Trimmed Maximum Likelihood Estimator achieves near optimal error for Gaussian Regression.

(Cheng u. a., 2020) studied empirical covariance estimation by gradient descent. They use gradient descent on a minimax formulation of the estimation problem. Their theoretical analysis is based upon the Moreau envelope. They prove their algorithm results in the norm of the gradient of the Moreau Envelope, and the ensuing $\mathbf{w}$ is a good point in the search space.

Minimax Optimization

(Jin u. a., 2020) studied minimax optimization in the non-convex non-concave setting. Furthermore, they study convergence of alternating minimizing-maximizing algorithm with a maximizing oracle. Their research utilizes the Moreau Envelope.

(Yang u. a., 2022) studied minimax optimization in the case of non-strong concavity.

1.2 Notation

The data matrix $\mathbf{X}$ is a fixed $n \times d$ matrix, the matrix $\mathbf{K}$ is the Gram Matrix, where $\mathbf{K}_{ij} = k(\mathbf{x}_i, \mathbf{x}_j)$ and $k(\cdot, \cdot)$ represents a kernel function, e.g. Linear kernel: $k(\mathbf{x}, \mathbf{y}) = \mathbf{x}^\top \mathbf{y}$, RBF kernel: $k(\mathbf{x}, \mathbf{y}) = \exp\left(-\gamma \|\mathbf{x} - \mathbf{y}\|_2^2\right)$. We denote $\mathbf{X}^\top = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n]$ where $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ represent the data vectors of the data matrix. We often denote X as the set of all data vectors, $X = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$. Uppercase bold $(\mathbf{X}, \mathbf{Y}, \mathbf{Z}, \ldots)$ are matrices. Uppercase Roman are sets (X, S, P, Q). Lowercase bold are vectors $(\mathbf{x}, \mathbf{y}, \mathbf{z}, \ldots)$.

We denote $\mathbf{I}_{k \times k}$ as the $k \times k$ identity matrix. The spectral norm of $\mathbf{A}$ is $\|\mathbf{A}\|_2 = \max_{\|\mathbf{x}\|=1} \|\mathbf{A}\mathbf{x}\| = \sigma_{\max}(\mathbf{A})$.

We also denote $\triangleq$ as 'defined as', to be used when we are defining a variable. We will use $\overset{\text{def}}{=}$ to say a variable is defined as a quantity from previous literature.

2 Subquantile Minimization

Given a convex objective function, f, the Subquantile learning problem becomes:

$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w}\in\mathbb{R}^d} \max_{t\in\mathbb{R}} g(t,\mathbf{w}) \triangleq \sum_{i=1}^{n} \left(t - (f(\mathbf{x}_i;\mathbf{w}) - y_i)^2\right)^+ \tag{1}$$

where t is the p-quantile of the empirical risk. Note that for a fixed t therefore the objective is not concave with respect to $\mathbf{w}$. Thus, to solve this problem we use the iterations from equation 11 in (Razaviyayn u. a., 2020).

$$t^{(k+1)} = \arg\max_{t\in\mathbb{R}} g(\mathbf{w}^{(k)}, t) \tag{2}$$

$$\mathbf{w}^{(k+1)} = \mathbf{w}^{(k)} - \alpha \nabla g(\mathbf{w}^{(k)}, t^{(k+1)}) \tag{3}$$

We note this is a non-convex concave minimax optimization problem.

Remark 2. Define the function $\Phi(\mathbf{w}) \triangleq \max_{t\in\mathbb{R}} g(t,\mathbf{w})$. This function is a L-weakly convex function, i.e., $\Phi(\mathbf{w}) + \frac{L}{2}\|\mathbf{w}\|^2$ is a convex function over $\mathbf{w}$.

We provide an algorithm for Subquantile Minimization of the ridge regression kernel algorithm. We solve the kernel ridge regression with Functional Gradient Descent.

Algorithm 1: SUBQ-GRADIENT

Input: Iterations: T; Quantile: p; Data Matrix: $\mathbf{X}, (n \times d), n \gg d$; Learning schedule: $\alpha_1, \cdots, \alpha_T$; Ridge parameter: λ

Output: Trained Parameters, $\mathbf{w}_{(T)}$

$\mathbf{w}_{(0)} \leftarrow \mathcal{N}_d(0,\sigma)$
for $k \in 1, 2, \ldots, T$ **do**
  $\mathbf{S}_{(k)} \leftarrow$ SUBQUANTILE$(\mathbf{w}^{(k)}, \mathbf{X})$
  $\mathbf{w}^{(k+1)} \leftarrow \mathbf{w}^{(k)} - \alpha_{(k)} \nabla_{\mathbf{w}} g\left(t^{(k+1)}, \mathbf{w}^{(k)}\right)$
end
return $\mathbf{w}_{(T)}$

Algorithm 2: SUBQUANTILE

Input: Parameters $\mathbf{w}$, Data Matrix: $\mathbf{X}, (n \times d)$, Convex Loss Function f

Output: Subquantile Matrix S

$\hat{\boldsymbol{\nu}}_i \leftarrow f(\mathbf{x}_i;\mathbf{w},y_i)$ s.t. $\hat{\boldsymbol{\nu}}_{i-1} \le \hat{\boldsymbol{\nu}}_i \le \hat{\boldsymbol{\nu}}_{i+1}$
$t \leftarrow \hat{\boldsymbol{\nu}}_{np}$
Let $\mathbf{x}_1, ..., \mathbf{x}_{np}$ be np points such that $f(\mathbf{x}_i;\mathbf{w},y_i) \le t$
$\mathbf{S} \leftarrow \begin{pmatrix}\mathbf{x}_1^\top & \ldots & \mathbf{x}_{np}^\top\end{pmatrix}^\top$
return S

3 Theory

To consider theoretical guarantees of Subquantile Minimization, we first analyze the inner and outer optimization problems. We first analyze kernel learning in the presence of corrupted data. Next, we provide error bounds for the two most important kernel learning problems, kernel ridge regression, and kernel classification.

Assumption 3. We represent the data matrix $\mathbf{X} = \begin{pmatrix}\widehat{\mathbf{P}}^\top & \widehat{\mathbf{Q}}^\top\end{pmatrix}^\top$ and the labels vector as $\mathbf{y} = \begin{pmatrix}\widehat{\mathbf{y}_P}^\top & \widehat{\mathbf{y}_Q}^\top\end{pmatrix}^\top$ where the rows of $\widehat{\mathbf{P}}$ and $\widehat{\mathbf{Q}}$ are sampled from the distribution, $\mathcal{N}_d(\mathbf{0},\mathbf{I})$

Lemma 4. *Let $f(\mathbf{x};\mathbf{w})$ be a convex loss function. Let $\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n$ denote the n data points ordered such that $f(\mathbf{x}_1;\mathbf{w},y_1) \le f(\mathbf{x}_2;\mathbf{w},y_2) \le \cdots \le f(\mathbf{x}_n;\mathbf{w},y_n)$. If we denote $\hat{\boldsymbol{\nu}}_i \triangleq f(\mathbf{x}_i;\mathbf{w},y_i)$, it then follows* $\arg\max_{t\in\mathbb{R}} g(t,\mathbf{w}) = \hat{\boldsymbol{\nu}}_{np}$.

Proof. First we can note, the max value of t for g is equivalent to the min value of t for g. We can now find

the Fermat Optimality Conditions for g.

$$\partial(-g(t,\mathbf{w})) = \partial\left(-t + \frac{1}{np}\sum_{i=1}^{n}(t-\hat{\boldsymbol{\nu}}_i)\right) \tag{4}$$

$$= -1 + \frac{1}{np}\sum_{i=1}^{np}\begin{cases}1 & \text{if } t > \hat{\boldsymbol{\nu}}_i \\ 0 & \text{if } t < \hat{\boldsymbol{\nu}}_i \\ [0,1] & \text{if } t = \hat{\boldsymbol{\nu}}_i\end{cases} \tag{5}$$

$$= 0 \text{ when } t = \hat{\boldsymbol{\nu}}_{np} \tag{6}$$

This is equivalent to the p-quantile of the Risk. ■

Interpretation 5. From lemma 4, we see the t will be greater than or equal to the errors of exactly np points. Thus, we are continuously updating over the np minimum errors.

Lemma 6. *Let $\hat{\boldsymbol{\nu}}_i \triangleq f(\mathbf{x}_i;\mathbf{w},y_i)$ s.t. $\hat{\boldsymbol{\nu}}_{i-1} \le \hat{\boldsymbol{\nu}}_i \le \hat{\boldsymbol{\nu}}_{i+1}$, if we choose $t^{(k+1)} = \hat{\boldsymbol{\nu}}_{np}$ as by lemma 4, it then follows $\nabla_{\mathbf{w}} g(t^{(k)},\mathbf{w}^{(k)}) = \frac{1}{np}\sum_{i=1}^{np}\nabla f(\mathbf{x}_i;\mathbf{w}^{(k)},y_i)$*

Proof. By our choice of $t^{(k+1)}$, it follows:

$$\nabla_{\mathbf{w}} g(t^{(k+1)},\mathbf{w}^{(k)}) = \nabla_{\mathbf{w}}\left(\hat{\boldsymbol{\nu}}_{np} - \frac{1}{np}\sum_{i=1}^{n}\left(\hat{\boldsymbol{\nu}}_{np} - f(\mathbf{x}_i;\mathbf{w},y_i)\right)^+\right) \tag{7}$$

$$= -\frac{1}{np}\sum_{i=1}^{np}\nabla_{\mathbf{w}}\left(\hat{\boldsymbol{\nu}}_{np} - f(\mathbf{x}_i;\mathbf{w},y_i)\right)^+ \tag{8}$$

$$= \frac{1}{np}\sum_{i=1}^{n}\nabla_{\mathbf{w}} f(\mathbf{x}_i;\mathbf{w}^{(k)},y_i)\begin{cases}1 & \text{if } t > \hat{\boldsymbol{\nu}}_i \\ 0 & \text{if } t < \hat{\boldsymbol{\nu}}_i \\ [0,1] & \text{if } t = \hat{\boldsymbol{\nu}}_i\end{cases} \tag{9}$$

Now we note $\boldsymbol{\nu}_{np} \le t^{(k+1)} \le \boldsymbol{\nu}_{np+1}$

$$\nabla_{\mathbf{w}} g(t^{(k+1)},\mathbf{w}^{(k)}) = \frac{1}{np}\sum_{i=1}^{np}\nabla_{\mathbf{w}} f(\mathbf{x}_i;\mathbf{w},y_i) \tag{10}$$

This concludes the proof. ■

Lemma 7. *(L-Lipschitz of $g(t,\mathbf{w})$ w.r.t $\mathbf{w}$). Let $\mathbf{x}_1,\mathbf{x}_2,\cdots,\mathbf{x}_n$, represent the data vectors. It then follows:*

$$|g(t,\mathbf{w}) - g(t,\widehat{\mathbf{w}})| \le L\|\mathbf{w}-\widehat{\mathbf{w}}\| \tag{11}$$

where $L = \frac{K}{np}\sigma_{\max}^2\left(\sum_{i=1}^{n}\left(\mathbf{k}_i\mathbf{k}_i^\top\right)\right) + \frac{2}{np}\sigma_{\max}\left(\sum_{i=1}^{n}\mathbf{k}_i\right)\|\mathbf{y}\|_2$

Lemma 8. *(β-Smoothness of $g(t,\mathbf{w})$ w.r.t $\mathbf{w}$). Let $\mathbf{x}_1,\mathbf{x}_2,\cdots,\mathbf{x}_n$ represent the rows of the data matrix $\mathbf{X}$. It then follows:*

$$\|\nabla_{\mathbf{w}} g(t,\mathbf{w}) - \nabla_{\mathbf{w}} g(t,\widehat{\mathbf{w}})\| \le \beta\|\mathbf{w}-\widehat{\mathbf{w}}\| \tag{12}$$

where $\beta = \frac{2}{np}\sigma_{\max}(\mathbf{X})$

Proof. W.L.O.G, let S be the set of points such that if $\mathbf{x}\in S$, then $t \ge \left(\mathbf{k}_{\mathbf{x}}^\top\mathbf{w} - y\right)^2$. Since g is twice differentiable, we will analyze the Hessian.

$$\left\|\nabla_{\mathbf{w}}^2 g(t,\mathbf{w})\right\|_2 = \left\|\frac{2}{np}\sum_{\mathbf{x}\in S}\mathbf{k}_{\mathbf{x}}\mathbf{k}_{\mathbf{x}}^\top\right\|_2 \overset{\text{eqn. (50)}}{\le} \left\|\frac{2}{np}\sum_{\mathbf{x}\in X}\mathbf{k}_{\mathbf{x}}\mathbf{k}_{\mathbf{x}}^\top\right\|_2 = \frac{2}{np}\sigma_{\max}\left(\sum_{\mathbf{x}\in X}\mathbf{k}_{\mathbf{x}}\mathbf{k}_{\mathbf{x}}^\top\right) \overset{\text{lem. 13}}{\le} \tag{13}$$

This concludes the proof. ■

3.1 Necessary Kernel Inequalities

We will first extend the idea of Resilience Steinhardt u. a. (2018) to kernel learning.

Definition 9. (**Resilience**) from (Steinhardt u. a., 2018). Let $\mathcal{H}$ represent a RKHS, then given the feature mapping $\phi : \mathcal{X} \to \mathcal{H}$, and the set $X = \{\mathbf{x}_i\}_{i=1}^n = \widehat{P} \cup \widehat{Q}$, such that $|\widehat{P}| = n(1-\epsilon)$ and $|\widehat{Q}| = n\epsilon$, it holds that for all $S \subseteq X$ s.t. $|S| \geq (1-\epsilon)n$, then $\left\| \frac{1}{|S|} \sum_{i \in S} \phi(\mathbf{x}_i) - \mu \right\| \leq \tau$ then we say the set X has (ϵ, τ)-resilience in the Hilbert Space.

Without the idea of resilience defined in definition 9, we will be unable to put error bounds on our algorithm.

Lemma 10. *Let S be the set of elements in the subquantile, then*

$$\left\| \frac{1}{np} \sum_{i \in S \cap \widehat{P}} \mathbf{k}_i \right\| \leq \mathcal{O}() \tag{14}$$

Proof.

$$\left\| \frac{1}{np} \sum_{i \in S \cap \widehat{P}} \mathbf{k}_i \right\| = \frac{1}{np} \left\| \sum_{i \in S \cap \widehat{P}} \sum_{j \in X} \phi(\mathbf{x}_i)^\top \phi(\mathbf{x}_j) \right\| \tag{15}$$

■

Lemma 11. *Under the same setting as lemma 10,*

$$\left\| \frac{1}{np} \sum_{i \in S \cap \widehat{Q}} \mathbf{k}_i \right\| \leq \mathcal{O}() \tag{16}$$

Lemma 12. *Under the same setting as lemma 10,*

$$\left\| \frac{1}{np} \sum_{i \in S \cap \widehat{P}} \xi_i \mathbf{k}_i \right\| \leq \mathcal{O}() \tag{17}$$

Lemma 13. *Under the same setting as lemma 10,*

$$\left\| \frac{1}{np} \sum_{i \in S} \mathbf{k}_i \mathbf{k}_i^\top \right\| \leq \mathcal{O}() \tag{18}$$

3.2 Kernel Ridge Regression

We denote the matrix $\mathbf{K}$ as the Gram Matrix where $K_{ij} = \kappa(\mathbf{x}_i, \mathbf{x}_j) \triangleq \exp\left(-\gamma \left\|\mathbf{x}_i - \mathbf{x}_j\right\|_2^2\right)$. Given a parameter set $\mathbf{w}$, the prediction for a new point will be: $f(\mathbf{x}^*; \mathbf{w}) = \sum_{i=1}^n \mathbf{w}_i \kappa(\mathbf{x}_i, \mathbf{x}^*)$

First let us define the kernels:

RBF Kernel: $\mathbf{k}_\mathbf{x} \triangleq \left(\exp\left(-\gamma \left\|\mathbf{x} - \mathbf{x}_1\right\|_2^2\right) \quad \cdots \quad \exp\left(-\gamma \left\|\mathbf{x} - \mathbf{x}_n\right\|_2^2\right)\right)^\top$ where $\gamma \triangleq \frac{1}{d}$

Linear Kernel: $\mathbf{k}_\mathbf{x} \triangleq \left(\mathbf{x}^\top \mathbf{x}_1 \quad \cdots \quad \mathbf{x}^\top \mathbf{x}_n\right)$

Assumption 14. (**Gauss-Markov Assumption**) Let $\widehat{P} = \{(\mathbf{p}_1, y_1), (\mathbf{p}_2, y_2), \cdots, (\mathbf{p}_m, y_m)\}$ and $K(\mathbf{x}, \mathbf{y})$ be a proper kernel. If $\mathbf{p} \in \widehat{P}$, then define $\mathbf{k}_{\mathbf{p}} \triangleq \begin{pmatrix} K(\mathbf{p}, \mathbf{p}_1) & K(\mathbf{p}, \mathbf{p}_2) & \cdots & K(\mathbf{p}, \mathbf{p}_m) \end{pmatrix}^\top$: we first assume y can be modeled as a non-linear function of the data to smooth the landscape of the function.

$$y = f(\mathbf{p}) + \widehat{\xi}_P \text{ where } \widehat{\xi}_P \sim \mathcal{N}(0, \widehat{\sigma}_P^2) \tag{19}$$

Furthermore, we assume this function can be represented by the Reproducing Kernel Hilbert Space:

$$y = \left(\mathbf{k}_{\mathbf{p}}^\top \mathbf{w}^* - f(\mathbf{p})\right) + \xi_P \text{ where } \xi_P \sim \mathcal{N}(0, \sigma_P^2) \tag{20}$$

Theorem 15. *(Monotonically Decreasing). Let f be a convex loss function and $\mathbf{X}$ follows assumption 3 with learning schedule $\alpha_{(1)}, \alpha_{(2)}, \cdots, \alpha_{(T)}$. Then it follows at any iteration $k \in \mathbb{N}$:*

$$g\left(t^{(k+1)}, \mathbf{w}^{(k+1)}\right) \leq g\left(t^{(k)}, \mathbf{w}^{(k)}\right) \tag{21}$$

From our definition of $S^{(k)}$ in theorem 15, we are interested in as $k \to \infty$ the quantities: $|\mathbf{x} \in S^{(k)} \cap \widehat{P}|$ and $|\mathbf{x} \in S^{(k)} \cap \widehat{Q}|$, where the latter cardinality represents the number of corrupted points in the subquantile set.

Definition 16. (**Moreau Envelope**). Let f be proper lower semi-continuous convex function $f : \mathcal{X} \to \mathbb{R}$, then the Moreau Envelope is defined as:

$$f_\lambda(\mathbf{x}) \triangleq \inf_{\widehat{\mathbf{x}} \in \mathcal{X}} \left(f(\widehat{\mathbf{x}}) + \frac{1}{2\lambda} \|\mathbf{x} - \widehat{\mathbf{x}}\|_2^2 \right) \tag{22}$$

The Moreau Envelope can be interpreted as an infimal convolution of the function f with a quadratic.

Definition 17. (**First-Order Stationary Point**). For any proper lower semi-continuous function f and closed convex set $\mathcal{K}$ consider its associated Moreau envelope $f_\beta(\mathbf{w})$ in definition 16. Then we say that a point $\mathbf{w}$ is a first-order stationary point if $\|\Phi_\beta(\mathbf{w})\|_2 = 0$

Definition 18. (**First-Order Stationary Point**). Let $\Phi(\mathbf{w}) = \max_t g(t, \mathbf{w})$. Then $\mathbf{w}$ is a first-order stationary point if

$$(\nabla_{\mathbf{w}} \Phi(\mathbf{w}))^\top (\widetilde{\mathbf{w}} - \mathbf{w}) \geq 0 \, \forall \widetilde{\mathbf{w}} \in \mathcal{K} \tag{23}$$

The idea of the First-Order Stationary Point will have significance in our analysis of the base learner algorithm in § appendix D.

Assumption 19. Define $\Phi(\cdot)$ as the function in remark 2. Then it follows $\arg\min_{\mathbf{w} \in \mathbb{R}^d} \Phi(\mathbf{w}) = \mathbf{w}^*$

Theorem 20. *Let $\widehat{\mathbf{w}}$ be a First-Order Stationary Point defined in definition 17, it then follows:*

Proof. We will define the function Φ as in remark 2. The derivative of the Moreau Envelope is well known, $\nabla \Phi_\lambda(\mathbf{w}) = \mathbf{w} - \frac{1}{2\lambda} \text{prox}_{\lambda \Phi}(\mathbf{w})$. Thus, if $\left\|\nabla \Phi_{1/2\ell}(\mathbf{w})\right\| = 0$, then it follows $\mathbf{w} = \ell \text{prox}_{1/2\ell\Phi}(\mathbf{w}) \stackrel{\text{def}}{=} \arg\min_{\widehat{\mathbf{w}}} \left(\Phi(\mathbf{w}) + \ell \|\widehat{\mathbf{w}} - \widehat{\mathbf{w}}\|^2 \right)$. Thus it follows for any $\widehat{\mathbf{w}}$ s.t. $\Phi(\widehat{\mathbf{w}}) < \Phi(\mathbf{w})$, then it holds $\ell \|\mathbf{w} - \widehat{\mathbf{w}}\|^2 > \Phi(\widehat{\mathbf{w}}) - \Phi(\mathbf{w})$. Therefore, if we let $\mathbf{w} = \mathbf{w}^*$, then we find $\Phi(\widehat{\mathbf{w}}) < \Phi(\mathbf{w}^*) + \ell \|\mathbf{w}^* - \widehat{\mathbf{w}}\|^2$.

Upper bound of $\|\mathbf{w}^* - \widehat{\mathbf{w}}\|^2$ The idea here is if $\|\widehat{\mathbf{w}} - \mathbf{w}^*\|$ is large, then the norm of the derivative must be correspondingly large. However, without any distribution assumptions on $\widehat{Q}$, how to show $\widehat{\mathbf{w}}$ is not a good weight vector for $\widehat{Q}$. ■

In practice, however, it is important to note that solving for $\|\nabla \Phi_\lambda\| = 0$ is NP-Hard. Thus, we will analyze the approximate stationary point.

Lemma 21. *((Rockafellar, 1970)). Assume the function Φ is ℓ-weakly convex. Let $\lambda < \frac{1}{\ell}$, and denote $\widehat{\mathbf{w}} = \arg\min_{\mathbf{w}'} \left(\Phi(\mathbf{w}') + \frac{1}{2\lambda} \|\mathbf{w} - \mathbf{w}'\|^2 \right)$, $\|\nabla \Phi_\lambda\| \leq \epsilon$ implies:*

$$\|\widehat{\mathbf{w}} - \mathbf{w}\| = \lambda \epsilon \text{ and } \min_{\mathbf{g} \in \partial \Phi(\widehat{\mathbf{w}})} \|\mathbf{g}\| \leq \epsilon \tag{24}$$

Definition 22. (**Approximate First-Order Stationary Point**) from (Cheng u. a., 2020). For any function f and closed convex set $\mathcal{K}$ consider its associated Moreau envelope $f_\beta(\mathbf{w})$ in definition 16. Then we say that a point $\mathbf{w}$ is a ρ-approximate stationary point if $\|f_\beta(\mathbf{w})\|_2 \leq \rho$.

The approximate stationary point in definition 22 is used in the analysis of the minimax algorithm in (Cheng u. a., 2020). First, if you can prove a stationary point is good, theorem 20, then using lemma 21, you can show an approximate stationary point is good.

Definition 23. (**Approximate First-Order Stationary Point**) from (Awasthi u. a., 2022). A regression coefficients vector, $\mathbf{w}$ is an approximate stationary point if:

$$\left(\sum_{\mathbf{x}\in S} \mathbf{k}_\mathbf{x}\left(\mathbf{k}_\mathbf{x}^\top \mathbf{w} - y\right)\right)^\top \left(\frac{(\mathbf{w}-\mathbf{w}^*)}{\|\mathbf{w}-\mathbf{w}^*\|}\right) \leq \eta \tag{25}$$

We adopt the proof strategy of (Awasthi u. a., 2022) and (Cheng u. a., 2020), and have a two-part proof strategy. First we show an approximate stationary point is close to the true distribution of $\mathbb{P}$. Then, we analyze the optimization to show ?? 1 converges to an approximate stationary point in a polynomial number of iterations.

Theorem 24. *Let $\widehat{\mathbf{w}}$ be the vector returned from ?? 1 after T iterations with $\alpha = \frac{1}{\beta}$, then we reach an approximation stationary point defined in ??,*

$$\left(\sum_{\mathbf{x}\in S} \mathbf{k}_\mathbf{x}\left(\mathbf{k}_\mathbf{x}^\top \mathbf{w} - y\right)\right)^\top \left(\frac{(\mathbf{w}-\mathbf{w}^*)}{\|\mathbf{w}-\mathbf{w}^*\|}\right) \leq \eta \tag{26}$$

Theorem 25. *(**Approximate Stationary Point is Good**). Let $\widehat{\mathbf{w}}$ be a η-stationary point as defined in definition 23. It then follows:*

$$\|\widehat{\mathbf{w}} - \mathbf{w}^*\| \leq \frac{\left\|\sum_{\mathbf{x}\in S\cap\widehat{P}} \xi_\mathbf{x}\mathbf{k}_\mathbf{x}\right\| + \sqrt{\frac{p}{1-p}}\left(\left\|\sum_{i\in\widehat{P}\cap S}\xi_i\right\|\left\|\sum_{\mathbf{x}\in S\cap\widehat{P}}\mathbf{k}_\mathbf{x}\right\|\right)}{\left\|\sum_{\mathbf{x}\in S\cap\widehat{P}}\mathbf{k}_\mathbf{x}\right\|^2 - \sqrt{\frac{p}{1-p}}\left\|\sum_{\mathbf{x}\in\widehat{P}\setminus S}\mathbf{k}_\mathbf{x}\right\|} \tag{27}$$

where $S^{(T)}$ represents the np set of points in the subquantile.

Since we are solving a minimax objective, we want a relation between the norm of the gradient of the Moreau Envelope of Φ and $\left(\sum_{\mathbf{x}\in S^{(T)}}\mathbf{k}_\mathbf{x}\left(\mathbf{k}_\mathbf{x}^\top\mathbf{w}^{(T)} - y\right)\right)^\top\left(\frac{\mathbf{w}^{(T)}-\mathbf{w}^*}{\|\mathbf{w}^{(T)}-\mathbf{w}^*\|}\right)$. First, we will show using stepsize of $1/\beta$ returns a μ-approximate stationary point.

Lemma 26. *content...*

Proof. content... ■

Theorem 27. *(**Algorithm ?? 1 reaches a η-approximate stationary point**). Algorithm ?? 1 reaches a η-approximate stationary point in a polynomial number of iterations.*

Proof. From (Lin u. a., 2020) Theorem 31 and (Cheng u. a., 2020) Lemma 4.2, it follows:

$$\mathbb{E}\left[\left\|\nabla\Phi_{1/2\ell}\left(\overline{\mathbf{w}}\right)\right\|^2\right] \leq 2\cdot\frac{\left(\Phi_{1/2\ell}\left(\mathbf{w}_0\right) - \min\Phi\left(\mathbf{w}\right)\right) + \ell\beta^2\gamma^2}{\gamma\sqrt{T+1}} \tag{28}$$

where $\widehat{\mathbf{w}} = \arg\min_{\mathbf{w}'}\Phi\left(\mathbf{w}'\right) + \ell\left\|\mathbf{w}-\mathbf{w}'\right\|^2$.
Let $\left\|\nabla\Phi_{1/2\ell}\left(\mathbf{w}^{(T)}\right)\right\| \leq \mu$, it then follows from lemma 21, $\left\|\widehat{\mathbf{w}} - \mathbf{w}^{(T)}\right\| = \mu/2\ell$. ■

3.3 Kernel Classification

4 Experiments

4.1 Kernel Ridge Regression

Algorithm 3: SUBQ-KERNEL-RIDGE-REGRESSION

Input: Iterations: T; Quantile: p; Data Matrix: $\mathbf{X}, (n \times d), n \gg d$; Labels: $\mathbf{y}, (n \times 1)$; Learning schedule: $\alpha_1, \cdots, \alpha_T$; Ridge parameter: λ

Output: Trained Parameters: $\mathbf{w}_{(T)}$; Base Learner: $\mathcal{L}$

$\mathbf{w}_{(0)} \leftarrow \left(\mathbf{K}^\top \mathbf{K} + \lambda \mathbf{I}\right) \mathbf{K}^\top \mathbf{y}$ ▷ Base Learner

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

  $\mathbf{S}_{(k)} \leftarrow \text{SUBQUANTILE}(\mathbf{w}^{(k)}, \mathbf{X})$ ▷ Algorithm ?? 2

  $\nabla_{\mathbf{w}} g\left(t^{(k+1)}, \mathbf{w}^{(k)}\right) \leftarrow 2 \sum_{i \in S^{(k)}} \mathbf{k}_i \left(\mathbf{k}_i \mathbf{w}^{(k)} - y_i\right) + \lambda \mathbf{K} \mathbf{w}^{(k)}$ ▷ Gradient Calculation

  $\mathbf{w}^{(k+1)} \leftarrow \mathbf{w}^{(k)} - \alpha_{(k)} \nabla_{\mathbf{w}} g\left(t^{(k+1)}, \mathbf{w}^{(k)}\right)$ ▷ $\mathbf{w}$-update in eqn. (3)

end

return $\mathbf{w}_{(T)}$

Objectives	Test RMSE (`Polynomial Regression (Degree = 3)`)			
	$\epsilon = 0.1$	$\epsilon = 0.2$	$\epsilon = 0.3$	$\epsilon = 0.4$
KRR	$0.460_{(0.2143)}$	$1.171_{(0.7809)}$	$0.950_{(0.3053)}$	$1.230_{(0.4678)}$
TERM (Li u. a., 2021)	∞	∞	∞	∞
SEVER (Diakonikolas u. a., 2019)	$0.071_{(0.0106)}$	$0.015_{(0.0041)}$	$0.056_{(0.0513)}$	$0.101_{(0.0643)}$
SUBQUANTILE($p = (1 - \epsilon)$)	$\mathbf{0.010_{(0.0004)}}$	$\mathbf{0.010_{(0.0002)}}$	$\mathbf{0.010_{(0.0007)}}$	$\mathbf{0.012_{(0.0030)}}$
Genie ERM	∞	∞	∞	∞

Table 1: `Polynomial Regression` Synthetic Dataset. 1000 samples, $x \sim \mathcal{N}(0, 1)$, $y \sim \mathcal{N}(\sum_{i=0} a_i x^i, 0.01)$ where $a_i \sim \mathcal{N}(0, 1)$. Oblivious Noise is sampled from $\mathcal{N}(0, 5)$. Subquantile is capped at 10,000 iterations. Polynomial Kernel: $K(\mathbf{x}, \mathbf{y}) = \left(\mathbf{x}^\top \mathbf{y} + 1\right)^3$. Regularization parameters is chosen as $\lambda = 1$. SEVER is trained with 16 iterations and $p = 0.02$.

Objectives	Test RMSE (`Boston Housing Regression`)			
	$\epsilon = 0.1$	$\epsilon = 0.2$	$\epsilon = 0.3$	$\epsilon = 0.4$
KRR	$0.544_{(0.0712)}$	$0.862_{(0.1199)}$	$0.865_{(0.0811)}$	$1.049_{(0.2249)}$
TERM (Li u. a., 2021)	$0.888_{(0.1360)}$	$0.891_{(0.1699)}$	$1.023_{(0.1329)}$	$0.931_{(0.0433)}$
SEVER (Diakonikolas u. a., 2019)	$0.593_{(0.0478)}$	$0.573_{(0.0559)}$	$0.567_{(0.1191)}$	∞
SUBQUANTILE($p = (1 - \epsilon)$)	$\mathbf{0.427_{(0.0691)}}$	$\mathbf{0.534_{(0.1105)}}$	$\mathbf{0.510_{(0.0695)}}$	$\mathbf{0.549_{(0.1030)}}$
Genie ERM	∞	∞	∞	∞

Table 2: `Boston Housing Regression` Dataset. Oblivious Noise is sampled from $\mathcal{N}(0, 5)$. Subquantile is capped at 10,000 iterations. Regularization Parameter is chosen as $\lambda = 2$

In fig. 1, we see the final subquantile has significantly less outliers than the original corruption in the data set. Furthermore, we see there is a greater decrease in the higher outlier settings. Looking at table 1 and figures fig. 1, subquantile minimization has near optimal performance in the `Polynomial Regression` Synthetic Dataset.

Objectives	Test RMSE (`Concrete Data Regression`)			
	$\epsilon = 0.1$	$\epsilon = 0.2$	$\epsilon = 0.3$	$\epsilon = 0.4$
KRR	$0.802_{(0.0324)}$	$0.929_{(0.0209)}$	$0.993_{(0.0441)}$	$0.775_{(0.0514)}$
TERM (Li u. a., 2021)	$0.874_{(0.0205)}$	$0.916_{(0.0421)}$	$0.840_{(0.0249)}$	$0.878_{(0.0749)}$
SEVER (Diakonikolas u. a., 2019)	$0.532_{(0.0134)}$	$0.516_{(0.0340)}$	$0.526_{(0.0217)}$	$0.552_{(0.0444)}$
SUBQUANTILE($p = (1-\epsilon)$)	$\mathbf{0.468_{(0.0220)}}$	$\mathbf{0.491_{(0.0271)}}$	$\mathbf{0.555_{(0.0391)}}$	$\mathbf{0.566_{(0.0405)}}$
Genie ERM	∞	∞	∞	∞

Table 3: `Concrete Data Regression` Dataset. Oblivious Noise is sampled from $\mathcal{N}(0, 5)$. Subquantile is capped at 10,000 iterations. Regularization Parameter is chosen as $\lambda = 2$.

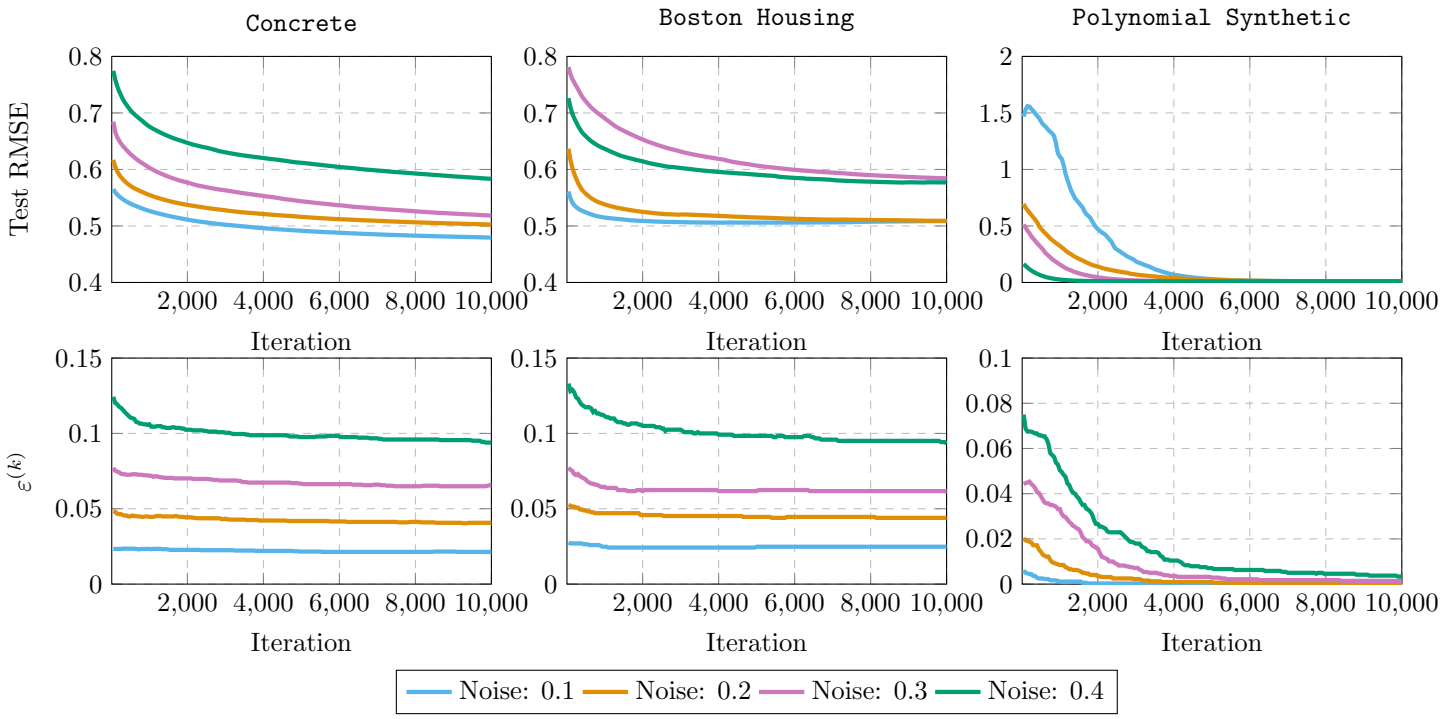

Figure 1: Test RMSE over the iterations in `Concrete`, `Boston Housing`, and `Polynomial` Datasets for SUBQUANTILE at different noise levels

4.2 Kernel Classification

Algorithm 4: SUBQ-KERNEL-CLASSIFICATION

Input: Iterations: T; Quantile: p; Data Matrix: $\mathbf{X}, (n \times d), n \gg d$; Labels: $\mathbf{y}, (n \times 1)$; Learning schedule: $\alpha_1, \cdots, \alpha_T$; Ridge parameter: λ

Output: Trained Parameters: $\mathbf{w}_{(T)}$; Base Learner: $\mathcal{L}$

$\mathbf{w}_{(0)} \leftarrow \left(\mathbf{K}^\top \mathbf{K} + \lambda \mathbf{I}\right) \mathbf{K}^\top \mathbf{y}$ ▷ Base Learner

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

  $\mathbf{S}_{(k)} \leftarrow$ SUBQUANTILE$(\mathbf{w}^{(k)}, \mathbf{X})$ ▷ Algorithm ?? 2

  $\nabla_{\mathbf{w}} g\left(t^{(k+1)}, \mathbf{w}^{(k)}\right) \leftarrow -\sum_{i \in S^{(k)}} y_i \mathbf{k}_i + \lambda \mathbf{K} \mathbf{w}^{(k)}$ ▷ Gradient Calculation

  $\mathbf{w}^{(k+1)} \leftarrow \mathbf{w}^{(k)} - \alpha_{(k)} \nabla_{\mathbf{w}} g\left(t^{(k+1)}, \mathbf{w}^{(k)}\right)$ ▷ $\mathbf{w}$-update in eqn. (3)

end

return $\mathbf{w}_{(T)}$

5 Discussion

The main contribution of this paper is the study of a nonconvex-concave optimization algorithm for the robust learning problem.

Interpretability. One of the strengths in Subquantile Optimization is the high interpretability. Once training is finished, we can see the $n(1-p)$ points with highest error to find the outliers. Furthermore, there is only hyperparameter p, which should be chosen to be approximately the percentage of inliers in the data and thus is not very difficult to tune for practical purposes.

General Assumptions. The general assumption is the majority of the data should inliers. This is not a very strong assumption, as by the definition of outlier it should be in the minority.

In future work, the analysis of Subquantile Minimization can be extended to neural networks and other learning algorithms.

References

[Awasthi u. a. 2022] Awasthi, Pranjal ; Das, Abhimanyu ; Kong, Weihao ; Sen, Rajat: Trimmed Maximum Likelihood Estimation for Robust Generalized Linear Model. In: Oh, Alice H. (Hrsg.) ; Agarwal, Alekh (Hrsg.) ; Belgrave, Danielle (Hrsg.) ; Cho, Kyunghyun (Hrsg.): *Advances in Neural Information Processing Systems*, URL https://openreview.net/forum?id=VHmdFPy4U_u, 2022

[Cheng u. a. 2020] Cheng, Yu ; Diakonikolas, Ilias ; Ge, Rong ; Soltanolkotabi, Mahdi: High-dimensional Robust Mean Estimation via Gradient Descent. In: III, Hal D. (Hrsg.) ; Singh, Aarti (Hrsg.): *Proceedings of the 37th International Conference on Machine Learning* Bd. 119, PMLR, 13–18 Jul 2020, S. 1768–1778. – URL https://proceedings.mlr.press/v119/cheng20a.html

[Diakonikolas u. a. 2019] Diakonikolas, Ilias ; Kamath, Gautam ; Kane, Daniel M. ; Li, Jerry ; Steinhardt, Jacob ; Stewart, Alistair: Sever: A Robust Meta-Algorithm for Stochastic Optimization. In: *Proceedings of the 36th International Conference on Machine Learning*, JMLR, Inc., 2019 (ICML '19), S. 1596–1606

[Dua und Graff 2017] Dua, Dheeru ; Graff, Casey": *UCI Machine Learning Repository*. 2017. – URL http://archive.ics.uci.edu/ml

[Jambulapati u. a. 2020] Jambulapati, Arun ; Li, Jerry ; Tian, Kevin: Robust Sub-Gaussian Principal Component Analysis and Width-Independent Schatten Packing. In: Larochelle, H. (Hrsg.) ; Ranzato, M. (Hrsg.) ; Hadsell, R. (Hrsg.) ; Balcan, M.F. (Hrsg.) ; Lin, H. (Hrsg.): *Advances in Neural Information Processing Systems* Bd. 33, Curran Associates, Inc., 2020, S. 15689–15701. – URL https://proceedings.neurips.cc/paper_files/paper/2020/file/b58144d7e90b5a43edcce1ca9e642882-Paper.pdf

[Jin u. a. 2020] Jin, Chi ; Netrapalli, Praneeth ; Jordan, Michael: What is Local Optimality in Nonconvex-Nonconcave Minimax Optimization? In: III, Hal D. (Hrsg.) ; Singh, Aarti (Hrsg.): *Proceedings of the 37th International Conference on Machine Learning* Bd. 119, PMLR, 13–18 Jul 2020, S. 4880–4889. – URL https://proceedings.mlr.press/v119/jin20e.html

[Li u. a. 2021] Li, Tian ; Beirami, Ahmad ; Sanjabi, Maziar ; Smith, Virginia: Tilted Empirical Risk Minimization. In: *International Conference on Learning Representations*, URL https://openreview.net/forum?id=K5YasWXZT3O, 2021

[Liao u. a. 2020] Liao, Zhenyu ; Couillet, Romain ; Mahoney, Michael W.: A random matrix analysis of random Fourier features: beyond the Gaussian kernel, a precise phase transition, and the corresponding double descent. In: Larochelle, H. (Hrsg.) ; Ranzato, M. (Hrsg.) ; Hadsell, R. (Hrsg.) ; Balcan, M.F. (Hrsg.) ; Lin, H. (Hrsg.): *Advances in Neural Information Processing Systems* Bd. 33, Curran Associates, Inc., 2020, S. 13939–13950. – URL https://proceedings.neurips.cc/paper_files/paper/2020/file/a03fa30821986dff10fc66647c84c9c3-Paper.pdf

[Lin u. a. 2020] Lin, Tianyi ; Jin, Chi ; Jordan, Michael I.: Near-Optimal Algorithms for Minimax Optimization. In: Abernethy, Jacob (Hrsg.) ; Agarwal, Shivani (Hrsg.): *Proceedings of Thirty Third Conference on Learning Theory* Bd. 125, PMLR, 09–12 Jul 2020, S. 2738–2779. – URL https://proceedings.mlr.press/v125/lin20a.html

[Rahimi und Recht 2007] Rahimi, Ali ; Recht, Benjamin: Random Features for Large-Scale Kernel Machines. In: Platt, J. (Hrsg.) ; Koller, D. (Hrsg.) ; Singer, Y. (Hrsg.) ; Roweis, S. (Hrsg.): *Advances in Neural Information Processing Systems* Bd. 20, Curran Associates, Inc., 2007. – URL https://proceedings.neurips.cc/paper_files/paper/2007/file/013a006f03dbc5392effeb8f18fda755-Paper.pdf

[Razaviyayn u. a. 2020] Razaviyayn, Meisam ; Huang, Tianjian ; Lu, Songtao ; Nouiehed, Maher ; Sanjabi, Maziar ; Hong, Mingyi: Nonconvex Min-Max Optimization: Applications, Challenges, and Recent Theoretical Advances. In: *IEEE Signal Processing Magazine* 37 (2020), Nr. 5, S. 55–66

[Rockafellar 1970] Rockafellar, Ralph T.: *Convex Analysis*. Princeton : Princeton University Press, 1970. – URL https://doi.org/10.1515/9781400873173. – ISBN 9781400873173

[Schneider 2016] Schneider, Markus: Probability Inequalities for Kernel Embeddings in Sampling without Replacement. In: *International Conference on Artificial Intelligence and Statistics*, 2016

[Staib und Jegelka 2019] Staib, Matthew ; Jegelka, Stefanie: Distributionally Robust Optimization and Generalization in Kernel Methods. In: Wallach, H. (Hrsg.) ; Larochelle, H. (Hrsg.) ; Beygelzimer, A. (Hrsg.) ; Alché-Buc, F. d'(Hrsg.) ; Fox, E. (Hrsg.) ; Garnett, R. (Hrsg.): *Advances in Neural Information Processing Systems* Bd. 32, Curran Associates, Inc., 2019. – URL https://proceedings.neurips.cc/paper_files/paper/2019/file/1770ae9e1b6bc9f5fd2841f141557ffb-Paper.pdf

[Steinhardt u. a. 2018] Steinhardt, Jacob ; Charikar, Moses ; Valiant, Gregory: Resilience: A Criterion for Learning in the Presence of Arbitrary Outliers. In: Karlin, Anna R. (Hrsg.): *9th Innovations in Theoretical Computer Science Conference, ITCS 2018, January 11-14, 2018, Cambridge, MA, USA* Bd. 94, Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2018, S. 45:1–45:21. – URL https://doi.org/10.4230/LIPIcs.ITCS.2018.45

[Yang u. a. 2022] Yang, Junchi ; Orvieto, Antonio ; Lucchi, Aurelien ; He, Niao: Faster Single-loop Algorithms for Minimax Optimization without Strong Concavity. In: Camps-Valls, Gustau (Hrsg.) ; Ruiz, Francisco J. R. (Hrsg.) ; Valera, Isabel (Hrsg.): *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics* Bd. 151, PMLR, 28–30 Mar 2022, S. 5485–5517. – URL https://proceedings.mlr.press/v151/yang22b.html

A Kernel Embedding Inequalities

Lemma 28. *(Resilience on Inlier Samples). Let* $X = \{\mathbf{x}_i\}_{i=1}^{n}$ *and* $\widehat{P} = \{\mathbf{x}_i\}_{i=1}^{np}$*, and* $[\mathbf{k}_i]_j = \phi(\mathbf{x}_i)^\top \phi(\mathbf{x}_j)$*. If the conditions in assumption definition 9 it then follows:*

$$\mathbb{P}\left\{\left\|\frac{1}{np}\sum_{i\in\widehat{P}}\phi(\mathbf{x}) - \mu_{\mathbb{P}}\right\| > \epsilon\right\} \leq 2\exp\left(-\frac{2np\epsilon^2}{d^2}\right)$$

where $\|\phi(\mathbf{x})\| \leq d$ *for any* $\mathbf{x} \in \mathcal{X}$ *a.s. A similar inequality can be found in (Schneider, 2016), Theorem 2*

Lemma 29. *Let* $\{\xi_i\}_{i=1}^{n}$ *represent realizations of a random variable,* $\xi_i \sim \mathcal{N}(\mu, \sigma^2)$*. It then follows with high probability:*

$$\left\|\sum_{i=1}^{n}\xi_i\right\| \leq C \text{ with high probability} \tag{29}$$

First, we can note,

$$\sum_{i=1}^{n}\xi_i \sim \mathcal{N}(n\mu, n\sigma^2) \tag{30}$$

With Hoeffding's Inequality, we have:

$$\mathbb{P}\left\{\left\|\frac{1}{n}\sum_{i=1}^{n}\xi_i - \mu\right\| \geq \epsilon\right\} \leq 2\exp\left(-\frac{2n^2\epsilon^2}{\sum_{i=1}^{n}\|\xi_i\|_{\psi_2}^2}\right) \tag{31}$$

Lemma 30. *(Resilience of corrupted sample). Let* $X = \left\{\mathbf{x}_i \in \mathbb{R}^d\right\}_{i=1}^{n}$*. It then follows for any subset* $S \subset X$*, s.t.* $|S| \geq np\varepsilon^{(k)}$*.*

B Proofs

B.1 Proof of theorem 15

Proof. We will introduce new notation, let $S^{(k)}$ denote the set of np data points from $\mathbf{X}$ with the lowest objective value $f(\mathbf{x};\mathbf{w}^{(k)},y) \triangleq \left(f\left(\mathbf{x};\mathbf{w}^{(k)}\right)-y\right)^2$. We will also define $F(\mathbf{w},S) \triangleq \sum_{\mathbf{x}\in S}\left(f\left(\mathbf{x};\mathbf{w}^{(k)}\right)-y\right)^2$. Note this is an equivalent characterization of g from lemma 4.

$$F(\mathbf{w}^{(k+1)},S^{(k+1)}) \leq F(\mathbf{w}^{(k)},S^{(k)}) \tag{32}$$
$$F(\mathbf{w}^{(k+1)},S^{(k+1)}) - F(\mathbf{w}^{(k)},S^{(k+1)}) \leq F(\mathbf{w}^{(k)},S^{(k)}) - F(\mathbf{w}^{(k)},S^{(k+1)}) \tag{33}$$

Upper Bound of LHS
Note that Kernel Ridge Regression is Lipschitz Continuous, let L denote the Lipschitz-Constant.

$$F(\mathbf{w}^{(k+1)},S^{(k+1)}) - F(\mathbf{w}^{(k)},S^{(k+1)}) \leq \langle \nabla_{\mathbf{w}} F(\mathbf{w}^{(k)},S^{(k+1)}), \mathbf{w}^{(k+1)}-\mathbf{w}^{(k)}\rangle + \frac{L}{2}\left\|\mathbf{w}^{(k+1)}-\mathbf{w}^{(k)}\right\|_2^2 \tag{34}$$
$$= \langle \nabla_{\mathbf{w}} F(\mathbf{w}^{(k)},S^{(k+1)}), -\alpha^{(k)}\nabla_{\mathbf{w}} F(\mathbf{w}^{(k)},S^{(k+1)})\rangle + \frac{L}{2}\left\|\mathbf{w}^{(k+1)}-\mathbf{w}^{(k)}\right\|_2^2 \tag{35}$$
$$= -\alpha^{(k)}\left\|\nabla_{\mathbf{w}} F(\mathbf{w}^{(k)},S^{(k+1)})\right\|_2^2 + \frac{L}{2}\left\|\mathbf{w}^{(k+1)}-\mathbf{w}^{(k)}\right\|_2^2 \tag{36}$$
$$= -\alpha^{(k)}\left\|\nabla_{\mathbf{w}} F(\mathbf{w}^{(k)},S^{(k+1)})\right\|_2^2 + \frac{L\alpha_{(k)}^2}{2}\left\|\nabla_{\mathbf{w}} F(\mathbf{w}^{(k)},S^{(k+1)})\right\|_2^2 \tag{37}$$
$$= \left(\frac{L\alpha_{(k)}^2}{2}-\alpha_{(k)}\right)\left\|\nabla_{\mathbf{w}} F(\mathbf{w}^{(k)},S^{(k+1)})\right\|_2^2 \tag{38}$$

Lower Bound of RHS
We first note that the points in $S^{(k+1)}$ and not in $S^{(k)}$ have lower residuals.

$$F(\mathbf{w}^{(k)},S^{(k)}) - F(\mathbf{w}^{(k)},S^{(k+1)}) = \sum_{\mathbf{x}\in S^{(k)}\setminus S^{(k+1)}} f(\mathbf{x};\mathbf{w}^{(k)},y) - \sum_{\mathbf{x}\in S^{(k+1)}\setminus S^{(k)}} f(\mathbf{x};\mathbf{w}^{(k)},y) \tag{39}$$
$$\geq |S^{(k)}\setminus S^{(k+1)}|\inf\left\{f(\mathbf{x};\mathbf{w}^{(k)},y):\mathbf{x}\in S^{(k)}\setminus S^{(k+1)}\right\} - |S^{(k+1)}\setminus S^{(k)}|\sup\left\{f(\mathbf{x};\mathbf{w}^{(k)},y):\mathbf{x}\in S^{(k+1)}\setminus S^{(k)}\right\} \tag{40}$$

Let $\eta \triangleq |S^{(k)}\setminus S^{(k+1)} = |S^{(k+1)}\setminus S^{(k)}|$

$$= \eta\left(\inf\left\{f(\mathbf{x};\mathbf{w}^{(k)},y):\mathbf{x}\in S^{(k)}\setminus S^{(k+1)}\right\} - \sup\left\{f(\mathbf{x};\mathbf{w}^{(k)},y):\mathbf{x}\in S^{(k+1)}\setminus S^{(k)}\right\}\right) \tag{41}$$
$$= \eta\left(\hat{\boldsymbol{\nu}}_{np+1}^{(k)} - \hat{\boldsymbol{\nu}}_{np}^{(k)}\right) \tag{42}$$
$$\geq 0 \tag{43}$$

Therefore, if $\alpha_{(k)} \leq \frac{2}{L}$, then it follows $F(\mathbf{w}^{(k+1)},S^{(k+1)}) \leq F(\mathbf{w}^{(k)},S^{(k)})$. This concludes the proof as we shown descent lemma. ■

B.2 Proof of lemma 7

Proof. We use the ℓ_2 norm of the gradient to bound L from above.

$$\left\|\nabla_{\mathbf{w}} g\left(t,\mathbf{w}\right)\right\|_2 = \left\|\frac{2}{np}\sum_{i=1}^{n}\mathbb{1}_{t\geq(\mathbf{x}_i^\top\mathbf{w}-y_i)^2}\left(\mathbf{x}_i\left(\mathbf{x}_i^\top\mathbf{w}-y_i\right)\right)\right\|_2 \tag{44}$$

W.L.O.G, let $\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_m$ where $0 \leq m \leq n$, represent the data vectors such that $t \geq \left(\mathbf{k}_i^\top \mathbf{w} - y_i\right)^2$.

$$= \left\| \frac{2}{np} \sum_{i=1}^{m} \mathbf{k}_i \left(\mathbf{k}_i^\top \mathbf{w} - y_i\right) \right\|_2 \tag{45}$$

$$\overset{(a)}{\leq} \frac{2}{np} \left(\left\| \sum_{i=1}^{m} \left(\mathbf{k}_i \mathbf{k}_i^\top\right) \mathbf{w} \right\|_2 + \left\| \sum_{i=1}^{m} \mathbf{k}_i y \right\|_2 \right) \tag{46}$$

$$\overset{(b)}{\leq} \frac{2}{np} \left(\left\| \sum_{i=1}^{m} \left(\mathbf{k}_i \mathbf{k}_i^\top\right) \right\|_2 \|\mathbf{w}\|_2 + \left\| \sum_{i=1}^{m} \mathbf{k}_i \right\|_2 \|\mathbf{y}\|_2 \right) \tag{47}$$

$$\overset{(c)}{\leq} \frac{2}{np} \left(\sigma_{\max} \left(\sum_{i=1}^{n} \left(\mathbf{k}_i \mathbf{k}_i^\top\right) \right) \|\mathbf{w}\|_2 + \left\| \sum_{i=1}^{m} \mathbf{k}_i \right\|_2 \|\mathbf{y}\|_2 \right) \tag{48}$$

$$\overset{(d)}{\leq} \frac{K}{np} \sigma_{\max} \left(\sum_{i=1}^{n} \left(\mathbf{k}_i \mathbf{k}_i^\top\right) \right) + \frac{2}{np} \left\| \sum_{i=1}^{m} \mathbf{k}_i \right\|_2 \|\mathbf{y}\|_2 \tag{49}$$

where (a) follows from Triangle Inequality, (b) follows from Cauchy-Schwarz Inequality, (d) follows from assuming $\|\mathbf{w}\|_2 \leq K$, where K is some positive constant. In (c), we note that $\mathbf{X}_m$ is positive semi-definite.

$$\mathbf{v}^\top \mathbf{X}_m \mathbf{v} = \mathbf{v}^\top \left(\sum_{i=1}^{m} \mathbf{k}_i \mathbf{k}_i^\top \right) \mathbf{v} = \sum_{i=1}^{m} \mathbf{v}^\top \mathbf{k}_i \mathbf{k}_i^\top \mathbf{v} = \sum_{i=1}^{m} \left(\mathbf{v}^\top \mathbf{k}_i\right)^2 \leq \sum_{i=1}^{n} \left(\mathbf{v}^\top \mathbf{k}_i\right)^2 \tag{50}$$

This concludes the proof. ■

B.3 Proof of theorem 25

Proof. This proof follows a similar structure to (Awasthi u. a., 2022), Lemma A.1

$$\left(\frac{1}{np} \sum_{\mathbf{x} \in S} \mathbf{k}_\mathbf{x} \left(\mathbf{k}_\mathbf{x}^\top \mathbf{w} - y\right) \right)^\top (\mathbf{w} - \mathbf{w}^*) \leq \eta \|\widehat{\mathbf{w}} - \mathbf{w}^*\| \tag{51}$$

$$\iff \left(\frac{1}{np} \sum_{\mathbf{x} \in S \cap \widehat{P}} \mathbf{k}_\mathbf{x} \left(\mathbf{k}_\mathbf{x}^\top \widehat{\mathbf{w}} - y\right) \right)^\top (\widehat{\mathbf{w}} - \mathbf{w}^*) \leq \left(-\frac{1}{np} \sum_{\mathbf{x} \in S \cap \widehat{Q}} \mathbf{k}_\mathbf{x} \left(\mathbf{k}_\mathbf{x}^\top \widehat{\mathbf{w}} - y\right) \right)^\top (\widehat{\mathbf{w}} - \mathbf{w}^*) + \eta \|\widehat{\mathbf{w}} - \mathbf{w}^*\| \tag{52}$$

Lower Bound on LHS

$$\left(\frac{1}{np} \sum_{\mathbf{x} \in S \cap \widehat{P}} \mathbf{k}_\mathbf{x} \left(\mathbf{k}_\mathbf{x}^\top \widehat{\mathbf{w}} - y\right) \right)^\top (\widehat{\mathbf{w}} - \mathbf{w}^*) \tag{53}$$

$$\overset{\text{asm. 14}}{=} \left(\frac{1}{np} \sum_{\mathbf{x} \in S \cap \widehat{P}} \mathbf{k}_\mathbf{x} \left(\mathbf{k}_\mathbf{x}^\top \mathbf{w} - \mathbf{k}_\mathbf{x}^\top \mathbf{w}^* - \xi_i\right) \right)^\top (\mathbf{w} - \mathbf{w}^*) \tag{54}$$

$$= \frac{1}{np} \sum_{\mathbf{x} \in S \cap \widehat{P}} (\mathbf{w} - \mathbf{w}^*)^\top \left(\mathbf{k}_\mathbf{x} \mathbf{k}_\mathbf{x}^\top\right) (\mathbf{w} - \mathbf{w}^*) - \xi_i \mathbf{k}_\mathbf{x}^\top (\mathbf{w} - \mathbf{w}^*) \tag{55}$$

$$\geq \frac{1}{np} \sum_{\mathbf{x} \in S \cap \widehat{P}} \left(\mathbf{k}_\mathbf{x}^\top (\mathbf{w} - \mathbf{w}^*)\right)^2 - \|\xi_i \mathbf{k}_\mathbf{x}\| \|\mathbf{w} - \mathbf{w}^*\| \tag{56}$$

$$\geq \frac{1}{np} \left(K \|\mathbf{w} - \mathbf{w}^*\|^2 - \left\| \sum_{\mathbf{x} \in S \cap \widehat{P}} \xi_i \mathbf{k}_\mathbf{x} \right\| \|\mathbf{w} - \mathbf{w}^*\|_\mathcal{H} \right) \tag{57}$$

where K is a lower bound on $\left\|\sum_{\mathbf{x}\in S\cap\widehat{P}}\mathbf{k}_{\mathbf{x}}\mathbf{k}_{\mathbf{x}}^{\top}-\mathbb{E}\left[\mathbf{k}_{\mathbf{x}}\mathbf{k}_{\mathbf{x}}^{\top}\right]\right\|$. Note this term will be significantly greater than the other term in the denominator.

Upper Bound on RHS

$$\left(-\frac{1}{np}\sum_{\mathbf{x}\in S\cap\widehat{Q}}\mathbf{k}_{\mathbf{x}}\left(\mathbf{k}_{\mathbf{x}}^{\top}\widehat{\mathbf{w}}-y\right)\right)^{\top}\left(\widehat{\mathbf{w}}-\mathbf{w}^{*}\right) \tag{58}$$

$$\overset{\text{Cauchy-Schwarz}}{\leq}\left(\frac{1}{np}\sum_{\mathbf{x}\in S\cap\widehat{Q}}\left(\mathbf{k}_{\mathbf{x}}^{\top}\mathbf{w}-y\right)^{2}\right)^{1/2}\left(\frac{1}{np}\sum_{\mathbf{x}\in S\cap\widehat{Q}}\left(\mathbf{k}_{\mathbf{x}}^{\top}\left(\mathbf{w}-\mathbf{w}^{*}\right)\right)^{2}\right)^{1/2} \tag{59}$$

$$\leq\frac{1}{np}\left(\sum_{\mathbf{x}\in S\cap\widehat{Q}}\left(\mathbf{k}_{\mathbf{x}}^{\top}\mathbf{w}-y\right)^{2}\right)^{1/2}\left(\left(\sum_{\mathbf{x}\in S\cap\widehat{Q}}\left\|\mathbf{k}_{\mathbf{x}}\right\|^{2}\right)\left\|\mathbf{w}-\mathbf{w}^{*}\right\|^{2}\right)^{1/2} \tag{60}$$

$$\overset{(a)}{\leq}\frac{1}{np}\left(\sum_{\mathbf{x}\in S\cap\widehat{Q}}\left(\mathbf{k}_{\mathbf{x}}^{\top}\mathbf{w}-y\right)^{2}\right)^{1/2}\left(\left\|\sum_{\mathbf{x}\in S\cap\widehat{Q}}\mathbf{k}_{\mathbf{x}}\right\|^{2}\left\|\mathbf{w}-\mathbf{w}^{*}\right\|^{2}\right)^{1/2} \tag{61}$$

$$\overset{(b)}{\leq}\frac{1}{np}\left(\frac{p}{1-p}\sum_{\mathbf{x}\in\widehat{P}\setminus S}\left(\mathbf{k}_{\mathbf{x}}^{\top}\left(\mathbf{w}-\mathbf{w}^{*}\right)-\xi_{i}\right)^{2}\right)^{1/2}\left(\left\|\sum_{\mathbf{x}\in S\cap\widehat{Q}}\mathbf{k}_{\mathbf{x}}\right\|\left\|\mathbf{w}-\mathbf{w}^{*}\right\|\right) \tag{62}$$

$$\overset{(c)}{\leq}\frac{1}{np}\left(\frac{p}{1-p}\left(\left\|\sum_{\mathbf{x}\in\widehat{P}\setminus S}\mathbf{k}_{\mathbf{x}}\right\|\left\|\mathbf{w}-\mathbf{w}^{*}\right\|+\left\|\sum_{\mathbf{x}\in\widehat{P}\setminus S}\xi_{i}\right\|\right)^{2}\right)^{1/2}\left(\left\|\sum_{\mathbf{x}\in S\cap\widehat{Q}}\mathbf{k}_{\mathbf{x}}\right\|\left\|\mathbf{w}-\mathbf{w}^{*}\right\|\right) \tag{63}$$

$$=\frac{1}{np}\sqrt{\frac{p}{1-p}}\left(\left\|\sum_{\mathbf{x}\in\widehat{P}\setminus S}\mathbf{k}_{\mathbf{x}}\right\|\left\|\mathbf{w}-\mathbf{w}^{*}\right\|+\left\|\sum_{\mathbf{x}\in\widehat{P}\setminus S}\xi_{i}\right\|\right)\left(\left\|\sum_{\mathbf{x}\in S\cap\widehat{Q}}\mathbf{k}_{\mathbf{x}}\right\|\left\|\mathbf{w}-\mathbf{w}^{*}\right\|\right) \tag{64}$$

(a) holds because all entries in $\mathbf{k}_{\mathbf{x}}$ are strictly non-negative.
(b) holds because the points outside of the subquantile will have greater average error than the points within the subquantile, we use this step as we do not make distributional assumptions on $\widehat{Q}$. We also introduce the $\frac{p}{1-p}$ since this is the fraction of points within the subquantile to outside.
(c) holds because of the inequality, $\|\mathbf{x}-\mathbf{y}\|^{2}=\|\mathbf{x}\|^{2}-2\mathbf{x}^{\top}\mathbf{y}+\|\mathbf{y}\|^{2}\leq\|\mathbf{x}\|^{2}+2\|\mathbf{x}\|\|\mathbf{y}\|+\|\mathbf{y}\|^{2}=(\|\mathbf{x}\|+\|\mathbf{y}\|)^{2}$.
We know analyze the LHS and RHS together to upper bound $\|\mathbf{w}-\mathbf{w}^{*}\|_{\mathcal{H}}$.

$$K\left\|\mathbf{w}-\mathbf{w}^{*}\right\|-\left\|\sum_{\mathbf{x}\in S\cap\widehat{P}}\xi_{\mathbf{x}}\mathbf{k}_{\mathbf{x}}\right\|\leq\sqrt{\frac{p}{1-p}}\left(\left\|\sum_{\mathbf{x}\in\widehat{P}\cap S}\mathbf{k}_{\mathbf{x}}\right\|\left\|\mathbf{w}-\mathbf{w}^{*}\right\|+\left\|\sum_{i\in\widehat{P}\setminus S}\xi_{i}\right\|\right)\left\|\sum_{\mathbf{x}\in S\cap\widehat{Q}}\mathbf{k}_{\mathbf{x}}\right\|+\eta \tag{65}$$

$$\left\|\mathbf{w}-\mathbf{w}^{*}\right\|\leq\frac{\left\|\sum_{\mathbf{x}\in S\cap\widehat{P}}\xi_{\mathbf{x}}\mathbf{k}_{\mathbf{x}}\right\|+\sqrt{\frac{p}{1-p}}\left(\left\|\sum_{i\in\widehat{P}\setminus S}\xi_{i}\right\|\left\|\sum_{\mathbf{x}\in S\cap\widehat{P}}\mathbf{k}_{\mathbf{x}}\right\|\right)+\eta}{K-\sqrt{\frac{p}{1-p}}\left\|\sum_{\mathbf{x}\in\widehat{P}\setminus S}\mathbf{k}_{\mathbf{x}}\right\|} \tag{66}$$

Can we extend the notion of resilience in (Steinhardt u. a., 2018)(Definition 1, Proposition 2), (Awasthi u. a., 2022)(Proposition D.1, D.3, D.4), and (Jambulapati u. a., 2020)(Lemma 8, Corollary 3, Corollary 4) to the kernel feature space $\phi:\mathcal{X}\to\mathcal{H}$. (Staib und Jegelka, 2019) has some analysis on gaussian kernels which might be useful.
In (Liao u. a., 2020), equation 1 and equation 2, there is an approximation for the Kernel Gram Matrix for Gaussian Kernel, $\mathbf{K}$. The original idea is in (Rahimi und Recht, 2007). This could be easier to model.

$$\left[\mathbf{K}_{\mathbf{X}}\right]_{ij}=e^{-\frac{1}{2}\left(\|\mathbf{x}_{i}\|^{2}+\|\mathbf{x}_{j}\|^{2}\right)}\left(\cosh\left(\mathbf{x}_{i}^{\top}\mathbf{x}_{j}\right)+\sinh\left(\mathbf{x}_{i}^{\top}\mathbf{x}_{j}\right)\right) \tag{67}$$

This concludes the proof. ■

C Robust Linear Regression

Theorem 31. *(**Approximate Stationary Point is Good**). Let $\widehat{\mathbf{w}}$ be a η-stationary point as defined in definition 23. It then follows:*

$$\|\widehat{\mathbf{w}} - \mathbf{w}^*\| \leq \frac{\left\|\sum_{\mathbf{x}\in S\cap\widehat{P}} \xi_{\mathbf{x}}\mathbf{k}_{\mathbf{x}}\right\| + \sqrt{\frac{p}{1-p}}\left(\left\|\sum_{i\in\widehat{P}\cap S}\xi_i\right\|\left\|\sum_{\mathbf{x}\in S\cap\widehat{P}}\mathbf{k}_{\mathbf{x}}\right\|\right)}{\left\|\sum_{\mathbf{x}\in S\cap\widehat{P}}\mathbf{k}_{\mathbf{x}}\right\|^2 - \sqrt{\frac{p}{1-p}}\left\|\sum_{\mathbf{x}\in\widehat{P}\setminus S}\mathbf{k}_{\mathbf{x}}\right\|} \tag{68}$$

where $S^{(T)}$ represents the np set of points in the subquantile.

Proof. We start from the proof of theorem 25.

$$\|\widehat{\mathbf{w}} - \mathbf{w}^*\| \leq \frac{\left\|\sum_{i\in S\cap\widehat{P}} \xi_i\mathbf{x}_i\right\| + \sqrt{\frac{p}{1-p}}\left(\left\|\sum_{i\in\widehat{P}\cap S}\xi_i\right\|\left\|\sum_{i\in S\cap\widehat{P}}\mathbf{x}_i\right\|\right)}{\left\|\sum_{i\in S\cap\widehat{P}}\mathbf{x}_i\right\|^2 - \sqrt{\frac{p}{1-p}}\left\|\sum_{i\in\widehat{P}\setminus S}\mathbf{x}_i\right\|} \tag{69}$$

$$\leq \frac{\sigma\varepsilon\log(1/\varepsilon) + \sqrt{\frac{p}{1-p}}\left(\varepsilon + \varepsilon\sqrt{\log(1/\varepsilon)}\right)}{(1-\varepsilon\log(1/\varepsilon))^2 - \sqrt{\frac{p}{1-p}}\varepsilon\sqrt{\log(1/\varepsilon)}} \tag{70}$$

■

D Base Learner Algorithm

Algorithm 5: SUBQ-BASE-LEARNER

Input: Iterations: T; Quantile: p; Data Matrix: $\mathbf{X}, (n \times d), n \gg d$; Labels: $\mathbf{y}, (n \times 1)$; Learning schedule: $\alpha_1, \cdots, \alpha_T$; Ridge parameter: λ

Output: Trained Parameters: $\mathbf{w}_{(T)}$; Base Learner: $\mathcal{L}$

$\mathbf{w}_{(0)} \leftarrow \mathcal{L}(\mathbf{X}, \mathbf{y})$ ▷ Base Learner

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

  $\mathbf{S}_{(k)} \leftarrow$ SUBQUANTILE$(\mathbf{w}^{(k)}, \mathbf{X})$ ▷ Algorithm ?? 2

  $\mathbf{w}^{(k+1)} \leftarrow \mathcal{L}\left(\mathbf{S}^{(k)}, \mathbf{y}_S\right)$ ▷ $\mathbf{w}$-update by base learner

end

return $\mathbf{w}_{(T)}$

Here we can note the similarity of Algorithm ?? 5 to the algorithm described in (Awasthi u. a., 2022). This is because the Trimmed Maximum Likelihood Estimator is equivalent to minimizing over the subquantile of the likelihood.

Remark 32. Define the function $\Psi(t) \triangleq \min_{\mathbf{w}} g(t, \mathbf{w})$

E Experimental Details

Our datasets are synthetic and are sourced from (Dua und Graff, 2017)

Dataset	**Dimension** d	**Sample Size** n	**Source**
`Polynomial`	3	1000	Ours
`Boston Housing`	13	506	(Dua und Graff, 2017)
`Concrete Data`	8	1030	(Dua und Graff, 2017)
`Wine Quality`	11	1599	(Dua und Graff, 2017)

Table 4: `Polynomial Regression` Synthetic Dataset. 1000 samples, $x \sim \mathcal{N}(0,1)$, $y \sim \mathcal{N}(\sum_{i=0} a_i x^i, 0.01)$ where $a_i \sim \mathcal{N}(0,1)$. Oblivious Noise is sampled from $\mathcal{N}(0,5)$. Subquantile is capped at 10,000 iterations.