Subquantile Minimization: Theory and Applications for Machine Learning

Arvind Rathnashyam* Alex Gittens†

September 8, 2023

Abstract

In this paper we propose Subquantile Minimization for learning with adversarial corruption in the training set. Superquantile objectives have been formed in the past in the context of fairness where one wants to learn an underrepresented distribution equally [12, 22]. Our intuition is to learn a more favorable representation of the *majority* class, thus we propose to optimize over the p-subquantile of the loss in the dataset. In particular, we study the Huber Contamination Problem for Kernel Learning 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), [4, 1, 13, 17, 6] as well as for linear regression [2, 16]. Sub-quantile minimization aims to address the shortcomings of standard ERM in applications of noisy/corrupted data [11, 9]. 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** [7]). 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 $n\varepsilon$ 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

In this section we will describe previous works in robust algorithms for the Huber ϵ-Contamination Model and works in minimax optimization that will be relevant to our theoretical analysis.

Robust Algorithms

[4] 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.

[13] 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.

[1] 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.

[3] 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

[10] 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.

[26] 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, $X = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n\}$. Uppercase bold ($\mathbf{X}, \mathbf{Y}, \mathbf{Z}, \dots$) are matrices. Uppercase Roman are sets (X, S, P, Q). Lowercase bold are vectors ($\mathbf{x}, \mathbf{y}, \mathbf{z}, \dots$). We represent the data matrix $\mathbf{X} = \begin{pmatrix}\mathbf{P}^\top & \mathbf{Q}^\top\end{pmatrix}^\top$, the labels vector as $\mathbf{y} = \begin{pmatrix}\mathbf{y}_P^\top & \mathbf{y}_Q^\top\end{pmatrix}^\top$, and the dataset $X = P \cup Q = \{(\mathbf{x}_i, y_i)\}_{i=1}^n = \{\mathbf{x}_i, y_i\}_{i \in P} \cup \{\mathbf{x}_i, y_i\}_{i \in Q}$.

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

We propose to optimize over the subquantile of the risk. The p-quantile of a random variable, U, is given as $\mathcal{Q}_p(U)$, this is the largest number, t, such that the probability of $U \leq t$ is at least p.

$$\mathcal{Q}_p(U) \leq t \iff \mathbb{P}\{U \leq t\} \geq p \tag{1}$$

The p-subquantile of the risk is then given by

$$\mathbb{L}_p(U) = \frac{1}{p}\int_0^p \mathcal{Q}_p(U)\, dq = \mathbb{E}\left[U | U \leq \mathcal{Q}_p(U)\right] = \max_{t \in \mathbb{R}} \left\{ t - \frac{1}{p}\mathbb{E}\left(t - U\right)^+ \right\} \tag{2}$$

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

$$\hat{\mathbf{w}} = \operatorname*{arg\,min}_{\mathbf{w} \in \mathbb{R}^d : \|\mathbf{w}\| \leq R} \max_{t \in \mathbb{R}} \left\{ g(t, \mathbf{w}) \triangleq \sum_{i=1}^n \left(t - (f(\mathbf{x}_i; \mathbf{w}) - y_i)^2\right)^+ \right\} \tag{3}$$

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 [19]. Let $\Pi_{\mathcal{K}}$ be the projection of a vector on to the convex set $\mathcal{K} \triangleq \{\mathbf{x} \in \mathbb{R}^n : \|\mathbf{x}\|_{\mathcal{H}} \leq R\}$, then our update steps are

$$t^{(k+1)} = \operatorname*{arg\,max}_{t \in \mathbb{R}} g(\mathbf{w}^{(k)}, t) \tag{4}$$

$$\mathbf{w}^{(k+1)} = \Pi_{\mathcal{K}}\left(\mathbf{w}^{(k)} - \alpha \nabla g(\mathbf{w}^{(k)}, t^{(k+1)})\right) \tag{5}$$

We note this is a non-convex concave minimax optimization problem. We provide an algorithm for Subquantile Minimization of the ridge regression and classification kernel learning algorithm. **??** is applicable to both kernel ridge regression and kernel classication.

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 $k \in 1, 2, \dots, T$ **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. $\hat{\boldsymbol{\nu}}_{i-1} \leq \hat{\boldsymbol{\nu}}_i \leq \hat{\boldsymbol{\nu}}_{i+1}$
$t \leftarrow \hat{\boldsymbol{\nu}}_{np}$
Let $\mathbf{x}_1, ..., \mathbf{x}_{np}$ be np points such that $f(\mathbf{x}_i; \mathbf{w}, y_i) \leq t$
$\mathbf{S} \leftarrow \begin{pmatrix}\mathbf{x}_1^\top & \dots & \mathbf{x}_{np}^\top\end{pmatrix}^\top$
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. Now we will give our first result regarding kernel learning in the Huber ϵ-contamination model.

Theorem 2. *Given a RKHS $\mathcal{H} \subset L^2(D)$ on a domain D with a proper kernel $K : D^2 \to \mathbb{R}$. Let $X = P \cup Q$ such that $|P| = (1-\varepsilon)n$ and $|Q| = \varepsilon n$ be our set of points. Further, assume $P \sim \mathbb{P} \subset D$ and $Q \sim \mathbb{Q} \subset D$. Let L be a loss function (e.g. regression or classification) and $\Omega : \mathbb{R}_+ \to \mathbb{R}$ be a strictly increasing regularization function. Let $f_{\mathbb{P}}$ be the minimizer for the points $\{\mathbf{x}_i\}_{i \in P}$ and $f_{\mathbb{Q}}$ be the minimizer for the points $\{\mathbf{x}_i\}_{i \in Q}$. Then for the minimizer of all the points $\{\mathbf{x}_i\}_{i=1}^n$*

$$\widehat{f} \in \arg\min_{f \in \mathcal{H}} L\left(\{f(\mathbf{x}_i)\}_{i=1}^n\right) + \Omega\left(\|f\|_{\mathcal{H}}^2\right) \tag{6}$$

it follows

$$\left\|\widehat{f} - f_{\mathbb{P}}\right\|_{\mathcal{H}} \leq \tag{7}$$

Now we will analyze the two-step minimax optimization steps described in Equations (4) and (5).

Lemma 3. *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 $f(\mathbf{x}_1; \mathbf{w}, y_1) \leq f(\mathbf{x}_2; \mathbf{w}, y_2) \leq \cdots \leq f(\mathbf{x}_n; \mathbf{w}, y_n)$. 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.

$$\partial(-g(t, \mathbf{w})) = \partial\left(-t + \frac{1}{np}\sum_{i=1}^{n}(t - \hat{\boldsymbol{\nu}}_i)\right) \tag{8}$$

$$= -1 + \frac{1}{np}\sum_{i=1}^{np}\begin{cases} 1 & \text{if } t > \hat{\boldsymbol{\nu}}_i \\ 0 & \text{if } t < \hat{\boldsymbol{\nu}}_i \\ [0,1] & \text{if } t = \hat{\boldsymbol{\nu}}_i \end{cases} \tag{9}$$

$$= 0 \text{ when } t = \hat{\boldsymbol{\nu}}_{np} \tag{10}$$

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

Interpretation 4. From lemma 3, 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 5. *Let $\hat{\boldsymbol{\nu}}_i \triangleq f(\mathbf{x}_i; \mathbf{w}, y_i)$ s.t. $\hat{\boldsymbol{\nu}}_{i-1} \leq \hat{\boldsymbol{\nu}}_i \leq \hat{\boldsymbol{\nu}}_{i+1}$, if we choose $t^{(k+1)} = \hat{\boldsymbol{\nu}}_{np}$ as by lemma 3, 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:

$$\nabla_{\mathbf{w}} g(t^{(k+1)}, \mathbf{w}^{(k)}) = \nabla_{\mathbf{w}}\left(\hat{\boldsymbol{\nu}}_{np} - \frac{1}{np}\sum_{i=1}^{n}\left(\hat{\boldsymbol{\nu}}_{np} - f(\mathbf{x}_i; \mathbf{w}, y_i)\right)^+\right) \tag{11}$$

$$= -\frac{1}{np}\sum_{i=1}^{np}\nabla_{\mathbf{w}}\left(\hat{\boldsymbol{\nu}}_{np} - f(\mathbf{x}_i; \mathbf{w}, y_i)\right)^+ \tag{12}$$

$$= \frac{1}{np}\sum_{i=1}^{n}\nabla_{\mathbf{w}} f(\mathbf{x}_i; \mathbf{w}^{(k)}, y_i)\begin{cases} 1 & \text{if } t > \hat{\boldsymbol{\nu}}_i \\ 0 & \text{if } t < \hat{\boldsymbol{\nu}}_i \\ [0,1] & \text{if } t = \hat{\boldsymbol{\nu}}_i \end{cases} \tag{13}$$

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

$$\nabla_{\mathbf{w}} g(t^{(k+1)}, \mathbf{w}^{(k)}) = \frac{1}{np}\sum_{i=1}^{np}\nabla_{\mathbf{w}} f(\mathbf{x}_i; \mathbf{w}, y_i) \tag{14}$$

This concludes the proof. ■

Lemma 6. *(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:*

$$|g(t, \mathbf{w}) - g(t, \widehat{\mathbf{w}})| \leq L \|\mathbf{w} - \widehat{\mathbf{w}}\| \tag{15}$$

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

Lemma 7. *(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:*

$$\|\nabla_{\mathbf{w}} g(t, \mathbf{w}) - \nabla_{\mathbf{w}} g(t, \widehat{\mathbf{w}})\| \leq \beta \|\mathbf{w} - \widehat{\mathbf{w}}\| \tag{16}$$

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 $t \geq \left(\mathbf{k}_{\mathbf{x}}^\top \mathbf{w} - y\right)^2$. Since g is twice differentiable, we will analyze the Hessian.

$$\left\|\nabla_{\mathbf{w}}^2 g(t, \mathbf{w})\right\|_2 = \left\|\frac{2}{np}\sum_{\mathbf{x}\in S}\mathbf{k}_{\mathbf{x}}\mathbf{k}_{\mathbf{x}}^\top\right\|_2 \overset{\text{eqn. (53)}}{\leq} \left\|\frac{2}{np}\sum_{\mathbf{x}\in X}\mathbf{k}_{\mathbf{x}}\mathbf{k}_{\mathbf{x}}^\top\right\|_2 = \frac{2}{np}\sigma_{\max}\left(\sum_{\mathbf{x}\in X}\mathbf{k}_{\mathbf{x}}\mathbf{k}_{\mathbf{x}}^\top\right) \overset{\text{lem. 13}}{\leq} \tag{17}$$

This concludes the proof. ■

Remark 8. 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 $\mathbf{w}$.

3.1 Necessary Kernel Inequalities

We will first extend the idea of Resilience [25] to kernel learning.

Definition 9. (**Resilience**) from [25]. Let $\mathcal{H}$ represent a RKHS, then given the feature mapping $\phi : \mathcal{X} \to \mathcal{H}$, and the set $X = \{\mathbf{x}_i\}_{i=1}^n = P \cup Q$, such that $|P| = n(1-\epsilon)$ and $|Q| = n\epsilon$, 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*

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

Proof.

$$\left\|\frac{1}{np}\sum_{i\in S\cap P}\mathbf{k}_i\right\| = \frac{1}{np}\left\|\sum_{i\in S\cap P}\sum_{j\in X}\phi(\mathbf{x}_i)^\top\phi(\mathbf{x}_j)\right\| \tag{19}$$

■

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

$$\left\|\frac{1}{np}\sum_{i\in S\cap Q}\mathbf{k}_i\right\| \leq \mathcal{O}() \tag{20}$$

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

$$\left\| \frac{1}{np} \sum_{i \in S \cap P} \xi_i \mathbf{k}_i \right\| \leq \mathcal{O}\left(\right) \tag{21}$$

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

$$\left\| \frac{1}{np} \sum_{i \in S} \mathbf{k}_i \mathbf{k}_i^\top \right\| \leq \mathcal{O}\left(\right) \tag{22}$$

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: $f(\mathbf{x}^\star; \mathbf{w}) = \sum_{i=1}^{n} \mathbf{w}_i \kappa(\mathbf{x}_i, \mathbf{x}^\star)$

Theorem 14. *(Monotonically Decreasing). Let f be a convex loss function and $\mathbf{X}$ follows ?? with learning schedule $\alpha_{(1)}, \alpha_{(2)}, \cdots, \alpha_{(T)}$. Then it follows at any iteration $k \in \mathbb{N}$:*

$$g\left(t^{(k+1)}, \mathbf{w}^{(k+1)}\right) \leq g\left(t^{(k)}, \mathbf{w}^{(k)}\right) \tag{23}$$

From our definition of $S^{(k)}$ in theorem 14, we are interested in as $k \to \infty$ the quantities: $|\mathbf{x} \in S^{(k)} \cap P|$ and $|\mathbf{x} \in S^{(k)} \cap Q|$, where the latter cardinality represents the number of corrupted points in the subquantile set.

Definition 15. (**Moreau Envelope**). Let f be proper lower semi-continuous convex function $f : \mathcal{X} \to \mathbb{R}$, then the Moreau Envelope is defined as:

$$f_\lambda(\mathbf{x}) \triangleq \inf_{\widehat{\mathbf{x}} \in \mathcal{X}} \left(f(\widehat{\mathbf{x}}) + \frac{1}{2\lambda} \left\|\mathbf{x} - \widehat{\mathbf{x}}\right\|_2^2 \right) \tag{24}$$

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

Definition 16. (**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 15. Then we say that a point $\mathbf{w}$ is a first-order stationary point if $\left\|\Phi_\beta(\mathbf{w})\right\|_2 = 0$

Definition 17. (**First-Order Stationary Point**). Let $\Phi(\mathbf{w}) = \max_t g(t, \mathbf{w})$. Then $\mathbf{w}$ is a first-order stationary point if

$$\left(\nabla_{\mathbf{w}} \Phi(\mathbf{w})\right)^\top (\widetilde{\mathbf{w}} - \mathbf{w}) \geq 0 \, \forall \widetilde{\mathbf{w}} \in \mathcal{K} \tag{25}$$

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

Assumption 18. Define $\Phi(\cdot)$ as the function in remark 8. Then it follows $\arg\min_{\mathbf{w} \in \mathbb{R}^d} \Phi\left(\mathbf{w}\right) = \mathbf{w}^\star$

Theorem 19. *Let $\widehat{\mathbf{w}}$ be a First-Order Stationary Point defined in definition 16, it then follows:*

Proof. We will define the function Φ as in remark 8. The derivative of the Moreau Envelope is well known, $\nabla \Phi_\lambda(\mathbf{w}) = \mathbf{w} - \frac{1}{2\lambda} \mathrm{prox}_{\lambda \Phi}\left(\mathbf{w}\right)$. Thus, if $\left\|\nabla \Phi_{1/2\ell}\left(\mathbf{w}\right)\right\| = 0$, then it follows $\mathbf{w} = \ell \mathrm{prox}_{1/2\ell\Phi}\left(\mathbf{w}\right) \stackrel{\text{def}}{=} \arg\min_{\widehat{\mathbf{w}}} \left(\Phi\left(\mathbf{w}\right) + \ell \left\|\widehat{\mathbf{w}} - \widehat{\mathbf{w}}\right\|^2 \right)$. Thus it follows for any $\widehat{\mathbf{w}}$ s.t. $\Phi\left(\widehat{\mathbf{w}}\right) < \Phi\left(\mathbf{w}\right)$, then it holds $\ell \left\|\mathbf{w} - \widehat{\mathbf{w}}\right\|^2 > \Phi\left(\widehat{\mathbf{w}}\right) - \Phi\left(\mathbf{w}\right)$. Therefore, if we let $\mathbf{w} = \mathbf{w}^\star$, then we find $\Phi\left(\widehat{\mathbf{w}}\right) < \Phi\left(\mathbf{w}^\star\right) + \ell \left\|\mathbf{w}^\star - \widehat{\mathbf{w}}\right\|^2$.

Upper bound of $\left\|\mathbf{w}^\star - \widehat{\mathbf{w}}\right\|^2$ The idea here is if $\left\|\widehat{\mathbf{w}} - \mathbf{w}^\star\right\|$ is large, then the norm of the derivative must be

correspondingly large. However, without any distribution assumptions on Q, how to show $\widehat{\mathbf{w}}$ is not a good weight vector for 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 20. *([21, 20]). Assume the function Φ is ℓ-weakly convex. Let $\lambda < \frac{1}{\ell}$, and denote $\widehat{\mathbf{w}} = \arg\min_{\mathbf{w}'} \left(\Phi(\mathbf{w}') + \frac{1}{2\lambda}\|\mathbf{w} - \mathbf{w}'\|^2\right)$, $\|\nabla\Phi_\lambda(\mathbf{w})\| \leq \epsilon$ implies:*

$$\|\widehat{\mathbf{w}} - \mathbf{w}\| = \lambda\epsilon \text{ and } \min_{\mathbf{g}\in\partial\Phi(\widehat{\mathbf{w}})+\partial\mathcal{I}_{\mathcal{K}}(\widehat{\mathbf{w}})} \|\mathbf{g}\| \leq \epsilon \tag{26}$$

Note the subdifferential of the support function is the normal cone, i.e.

$$\partial\mathcal{I}_{\mathcal{K}} = \mathcal{N}(\widehat{\mathbf{w}}) = \{\widetilde{\mathbf{w}} \in \mathbb{R}^n | \langle\widetilde{\mathbf{w}}, \bar{\mathbf{w}} - \widehat{\mathbf{w}}\rangle \leq \forall\bar{\mathbf{w}} \in \mathcal{K}\} \tag{27}$$

We will define a convex cone.

Definition 21. A set Ω is a cone if $\lambda x \in \Omega$ whenever $x \in \Omega$ and $\lambda \geq 0$, if Ω is convex then it is a convex cone.

Thus there exists $\mathbf{g} = \mathbf{u} + \mathbf{v}$ where $\mathbf{u} \in \partial\Phi(\widehat{\mathbf{w}})$ and $\mathbf{v} \in \mathcal{N}(\widehat{\mathbf{w}})$.

Theorem 22. *Let $\widehat{\mathbf{w}} = \arg\min_{\mathbf{w}'} \left(\Phi(\mathbf{w}') + \frac{1}{2\lambda}\|\mathbf{w} - \mathbf{w}'\|^2\right)$ s.t. $\|\nabla\Phi_\lambda(\mathbf{w})\| \leq \epsilon$, then it follows*

$$\|\widehat{\mathbf{w}} - \mathbf{w}^*\|_{\mathcal{H}} \leq \Xi \tag{28}$$

Definition 23. (**Approximate First-Order Stationary Point**) from [3]. For any function f and closed convex set $\mathcal{K}$ consider its associated Moreau envelope $f_\beta(\mathbf{w})$ in definition 15. 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 23 is used in the analysis of the minimax algorithm in [3]. First, if you can prove a stationary point is good, theorem 19, then using lemma 20, you can show an approximate stationary point is good.

We adopt the proof strategy of [1] and [3], 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.

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

Theorem 24. *(**Algorithm ?? 1 reaches a η-approximate stationary point**). Algorithm ?? 1 reaches a η-approximate stationary point in a polynomial number of iterations.*

Proof. From [15] Theorem 31 and [3] Lemma 4.2, it follows:

$$\mathbb{E}\left[\left\|\nabla\Phi_{1/2\ell}(\overline{\mathbf{w}})\right\|^2\right] \leq 2 \cdot \frac{\left(\Phi_{1/2\ell}(\mathbf{w}_0) - \min\Phi(\mathbf{w})\right) + \ell\beta^2\gamma^2}{\gamma\sqrt{T+1}} \tag{29}$$

where $\widehat{\mathbf{w}} = \arg\min_{\mathbf{w}'} \Phi(\mathbf{w}') + \ell\|\mathbf{w} - \mathbf{w}'\|^2$.
Let $\left\|\nabla\Phi_{1/2\ell}\left(\mathbf{w}^{(T)}\right)\right\| \leq \mu$, it then follows from lemma 20, $\left\|\widehat{\mathbf{w}} - \mathbf{w}^{(T)}\right\| = \mu/2\ell$. ■

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**

  $\mathbf{S}_{(k)} \leftarrow \text{SUBQUANTILE}(\mathbf{w}^{(k)}, \mathbf{X})$ ▷ ?? 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. (5)

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 [13]	∞	∞	∞	∞
SEVER [4]	$0.071_{(0.0106)}$	$0.015_{(0.0041)}$	$0.056_{(0.0513)}$	$0.101_{(0.0643)}$
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 [13]	$0.888_{(0.1360)}$	$0.891_{(0.1699)}$	$1.023_{(0.1329)}$	$0.931_{(0.0433)}$
SEVER [4]	$0.593_{(0.0478)}$	$0.573_{(0.0559)}$	$0.567_{(0.1191)}$	∞
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 [13]	$0.874_{(0.0205)}$	$0.916_{(0.0421)}$	$0.840_{(0.0249)}$	$0.878_{(0.0749)}$
SEVER [4]	$0.532_{(0.0134)}$	$0.516_{(0.0340)}$	$0.526_{(0.0217)}$	$0.552_{(0.0444)}$
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 $\lambda = 2$.

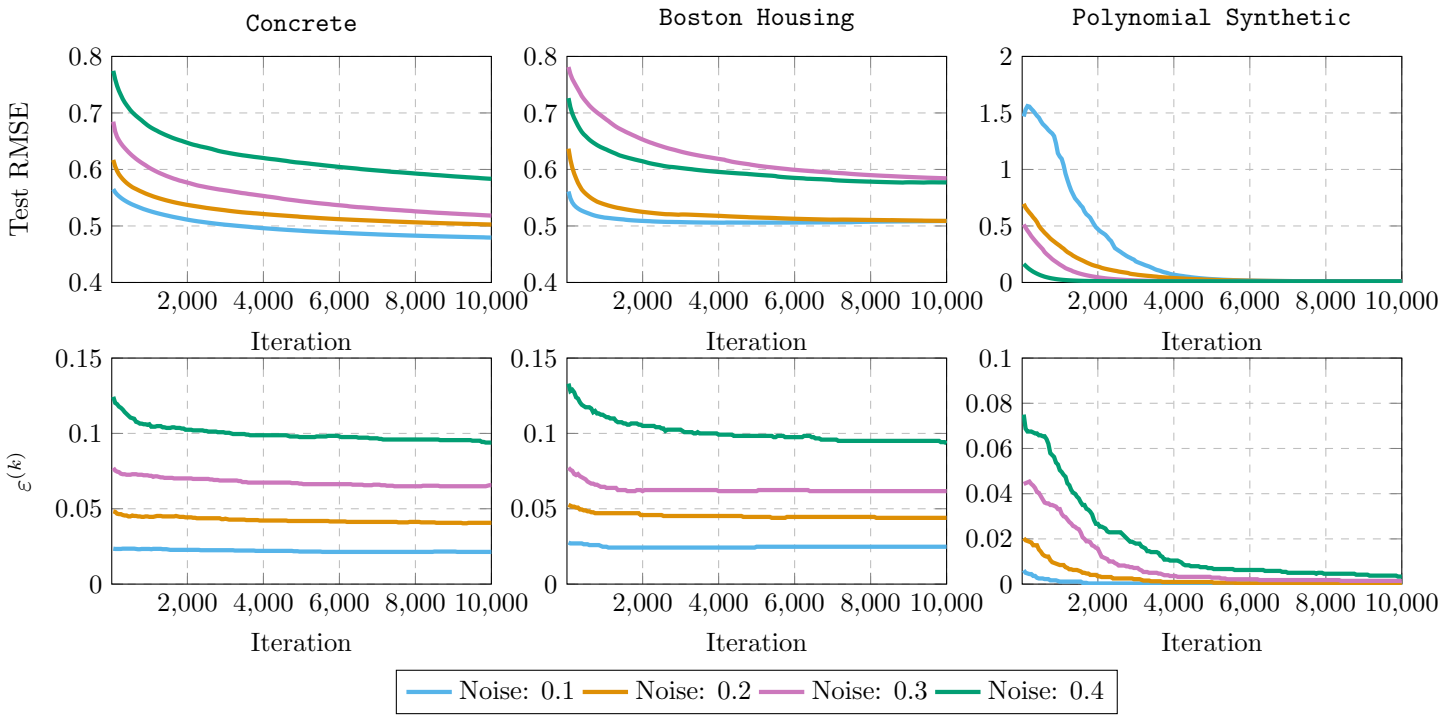

Figure 1: Test RMSE over the iterations in `Concrete`, `Boston Housing`, and `Polynomial` Datasets for SUBQUANTILE at different noise levels

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. (5)

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 for kernel ridge regression and kernel classification.

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

[1] Pranjal Awasthi, Abhimanyu Das, Weihao Kong, and Rajat Sen. Trimmed maximum likelihood estimation for robust generalized linear model. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022.

[2] Kush Bhatia, Prateek Jain, Parameswaran Kamalaruban, and Purushottam Kar. Consistent robust regression. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[3] Yu Cheng, Ilias Diakonikolas, Rong Ge, and Mahdi Soltanolkotabi. High-dimensional robust mean estimation via gradient descent. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 1768–1778. PMLR, 13–18 Jul 2020.

[4] 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, pages 1596–1606. JMLR, Inc., 2019.

[5] Dheeru Dua and Casey" Graff. UCI machine learning repository, 2017.

[6] Martin A. Fischler and Robert C. Bolles. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Commun. ACM*, 24(6):381–395, jun 1981.

[7] Peter J. Huber and Elvezio. Ronchetti. *Robust statistics*. Wiley series in probability and statistics. Wiley, Hoboken, N.J., 2nd ed. edition, 2009.

[8] Arun Jambulapati, Jerry Li, and Kevin Tian. Robust sub-gaussian principal component analysis and width-independent schatten packing. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 15689–15701. Curran Associates, Inc., 2020.

[9] Lu Jiang, Zhengyuan Zhou, Thomas Leung, Li-Jia Li, and Li Fei-Fei. Mentornet: Learning data-driven curriculum for very deep neural networks on corrupted labels. In *ICML*, 2018.

[10] Chi Jin, Praneeth Netrapalli, and Michael Jordan. What is local optimality in nonconvex-nonconcave minimax optimization? In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 4880–4889. PMLR, 13–18 Jul 2020.

[11] Ashish Khetan, Zachary C. Lipton, and Anima Anandkumar. Learning from noisy singly-labeled data. In *International Conference on Learning Representations*, 2018.

[12] Yassine Laguel, Krishna Pillutla, Jérôme Malick, and Zaid Harchaoui. Superquantiles at work: Machine learning applications and efficient subgradient computation. *Set-Valued and Variational Analysis*, 29(4):967–996, Dec 2021.

[13] Tian Li, Ahmad Beirami, Maziar Sanjabi, and Virginia Smith. Tilted empirical risk minimization. In *International Conference on Learning Representations*, 2021.

[14] Zhenyu Liao, Romain Couillet, and Michael W Mahoney. A random matrix analysis of random fourier features: beyond the gaussian kernel, a precise phase transition, and the corresponding double descent. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 13939–13950. Curran Associates, Inc., 2020.

[15] Tianyi Lin, Chi Jin, and Michael I. Jordan. Near-optimal algorithms for minimax optimization. In Jacob Abernethy and Shivani Agarwal, editors, *Proceedings of Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pages 2738–2779. PMLR, 09–12 Jul 2020.

[16] Bhaskar Mukhoty, Govind Gopakumar, Prateek Jain, and Purushottam Kar. Globally-convergent iteratively reweighted least squares for robust regression problems. In Kamalika Chaudhuri and Masashi Sugiyama, editors, *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pages 313–322. PMLR, 16–18 Apr 2019.

[17] Muhammad Osama, Dave Zachariah, and Petre Stoica. Robust risk minimization for statistical learning from corrupted data. *IEEE Open Journal of Signal Processing*, 1:287–294, 2020.

[18] Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. In J. Platt, D. Koller, Y. Singer, and S. Roweis, editors, *Advances in Neural Information Processing Systems*, volume 20. Curran Associates, Inc., 2007.

[19] Meisam Razaviyayn, Tianjian Huang, Songtao Lu, Maher Nouiehed, Maziar Sanjabi, and Mingyi Hong. Nonconvex min-max optimization: Applications, challenges, and recent theoretical advances. *IEEE Signal Processing Magazine*, 37(5):55–66, 2020.

[20] R Tyrrell Rockafellar. Convex analaysis. 2015.

[21] Ralph Tyrell Rockafellar. *Convex Analysis*. Princeton University Press, Princeton, 1970.

[22] 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.

[23] Markus Schneider. Probability inequalities for kernel embeddings in sampling without replacement. In *International Conference on Artificial Intelligence and Statistics*, 2016.

[24] Matthew Staib and Stefanie Jegelka. Distributionally robust optimization and generalization in kernel methods. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[25] Jacob Steinhardt, Moses Charikar, and Gregory Valiant. Resilience: A criterion for learning in the presence of arbitrary outliers. In Anna R. Karlin, editor, *9th Innovations in Theoretical Computer Science Conference, ITCS 2018, January 11-14, 2018, Cambridge, MA, USA*, volume 94 of *LIPIcs*, pages 45:1–45:21. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2018.

[26] Junchi Yang, Antonio Orvieto, Aurelien Lucchi, and Niao He. Faster single-loop algorithms for minimax optimization without strong concavity. In Gustau Camps-Valls, Francisco J. R. Ruiz, and Isabel Valera, editors, *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pages 5485–5517. PMLR, 28–30 Mar 2022.

Contents

A Kernel Embedding Inequalities

Lemma 25. *(Resilience on Inlier Samples). Let* $X = \{\mathbf{x}_i\}_{i=1}^{n}$ *and* $P = \{\mathbf{x}_i\}_{i=1}^{np}$*, and* $[\mathbf{k}_i]_j = \phi(\mathbf{x}_i)^\top \phi(\mathbf{x}_j)$*. If the conditions in assumption definition 9 it then follows:*

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

where $\|\phi(\mathbf{x})\| \leq d$ *for any* $\mathbf{x} \in \mathcal{X}$ *a.s. A similar inequality can be found in [23], Theorem 2*

Lemma 26. *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:*

$$\left\|\sum_{i=1}^{n}\xi_i\right\| \leq C \text{ with high probability} \tag{30}$$

First, we can note,

$$\sum_{i=1}^{n}\xi_i \sim \mathcal{N}(n\mu, n\sigma^2) \tag{31}$$

With Hoeffding's Inequality, we have:

$$\mathbb{P}\left\{\left\|\frac{1}{n}\sum_{i=1}^{n}\xi_i - \mu\right\| \geq \epsilon\right\} \leq 2\exp\left(-\frac{2n^2\epsilon^2}{\sum_{i=1}^{n}\|\xi_i\|_{\psi_2}^2}\right) \tag{32}$$

Lemma 27. *(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 Deferred Proofs

B.1 Proof of Theorem 2

Proof. We have

$$\left\|\widehat{f} - f_{\mathbb{P}}\right\|_{\mathcal{H}} = \left\|\widehat{f} - f_{\mathbb{Q}} + f_{\mathbb{Q}} - f_{\mathbb{P}}\right\|_{\mathcal{H}} \tag{33}$$

$$\leq \left\|\widehat{f}\right\|_{\mathcal{H}} \tag{34}$$

■

B.2 Proof of theorem 14

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 3.

$$F(\mathbf{w}^{(k+1)}, S^{(k+1)}) \leq F(\mathbf{w}^{(k)}, S^{(k)}) \tag{35}$$

$$F(\mathbf{w}^{(k+1)}, S^{(k+1)}) - F(\mathbf{w}^{(k)}, S^{(k+1)}) \leq F(\mathbf{w}^{(k)}, S^{(k)}) - F(\mathbf{w}^{(k)}, S^{(k+1)}) \tag{36}$$

Upper Bound of LHS

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

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

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

$$= -\alpha^{(k)} \left\|\nabla_{\mathbf{w}} F(\mathbf{w}^{(k)}, S^{(k+1)})\right\|_2^2 + \frac{L}{2} \left\|\mathbf{w}^{(k+1)} - \mathbf{w}^{(k)}\right\|_2^2 \tag{39}$$

$$= -\alpha^{(k)} \left\|\nabla_{\mathbf{w}} F(\mathbf{w}^{(k)}, S^{(k+1)})\right\|_2^2 + \frac{L\alpha_{(k)}^2}{2} \left\|\nabla_{\mathbf{w}} F(\mathbf{w}^{(k)}, S^{(k+1)})\right\|_2^2 \tag{40}$$

$$= \left(\frac{L\alpha_{(k)}^2}{2} - \alpha_{(k)}\right) \left\|\nabla_{\mathbf{w}} F(\mathbf{w}^{(k)}, S^{(k+1)})\right\|_2^2 \tag{41}$$

Lower Bound of RHS

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

$$F(\mathbf{w}^{(k)}, S^{(k)}) - F(\mathbf{w}^{(k)}, S^{(k+1)}) = \sum_{\mathbf{x} \in S^{(k)} \setminus S^{(k+1)}} f(\mathbf{x}; \mathbf{w}^{(k)}, y) - \sum_{\mathbf{x} \in S^{(k+1)} \setminus S^{(k)}} f(\mathbf{x}; \mathbf{w}^{(k)}, y) \tag{42}$$

$$\geq |S^{(k)} \setminus S^{(k+1)}| \inf\left\{f(\mathbf{x}; \mathbf{w}^{(k)}, y) : \mathbf{x} \in S^{(k)} \setminus S^{(k+1)}\right\} - |S^{(k+1)} \setminus S^{(k)}| \sup\left\{f(\mathbf{x}; \mathbf{w}^{(k)}, y) : \mathbf{x} \in S^{(k+1)} \setminus S^{(k)}\right\} \tag{43}$$

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

$$= \eta \left(\inf\left\{f(\mathbf{x}; \mathbf{w}^{(k)}, y) : \mathbf{x} \in S^{(k)} \setminus S^{(k+1)}\right\} - \sup\left\{f(\mathbf{x}; \mathbf{w}^{(k)}, y) : \mathbf{x} \in S^{(k+1)} \setminus S^{(k)}\right\}\right) \tag{44}$$

$$= \eta \left(\hat{\boldsymbol{\nu}}_{np+1}^{(k)} - \hat{\boldsymbol{\nu}}_{np}^{(k)}\right) \tag{45}$$

$$\geq 0 \tag{46}$$

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.3 Proof of lemma 6

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

$$\left\|\nabla_{\mathbf{w}} g\left(t, \mathbf{w}\right)\right\|_2 = \left\| \frac{2}{np} \sum_{i=1}^{n} \mathbb{1}_{t \geq (\mathbf{x}_i^\top \mathbf{w} - y_i)^2} \left(\mathbf{x}_i \left(\mathbf{x}_i^\top \mathbf{w} - y_i\right)\right) \right\|_2 \tag{47}$$

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 $t \geq \left(\mathbf{k}_i^\top \mathbf{w} - y_i\right)^2$.

$$= \left\| \frac{2}{np} \sum_{i=1}^{m} \mathbf{k}_i \left(\mathbf{k}_i^\top \mathbf{w} - y_i\right) \right\|_2 \tag{48}$$

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

$$\overset{(b)}{\leq} \frac{2}{np} \left(\left\| \sum_{i=1}^{m} \mathbf{k}_i \mathbf{k}_i^\top \right\|_2 \|\mathbf{w}\|_2 + \left\| \sum_{i=1}^{m} \mathbf{k}_i \right\|_2 \|\mathbf{y}\|_2 \right) \tag{50}$$

$$\overset{(c)}{\leq} \frac{2}{np} \left(\sigma_{\max} \left(\sum_{i=1}^{n} \mathbf{k}_i \mathbf{k}_i^\top \right) \|\mathbf{w}\|_2 + \left\| \sum_{i=1}^{m} \mathbf{k}_i \right\|_2 \|\mathbf{y}\|_2 \right) \tag{51}$$

$$\overset{(d)}{\leq} \frac{R}{np} \sigma_{\max} \left(\sum_{i=1}^{n} \mathbf{k}_i \mathbf{k}_i^\top \right) + \frac{2}{np} \left\| \sum_{i=1}^{m} \mathbf{k}_i \right\|_2 \|\mathbf{y}\|_2 \tag{52}$$

where (a) follows from Triangle Inequality, (b) follows from Cauchy-Schwarz Inequality, (d) follows from assuming $\|\mathbf{w}\|_{\mathcal{H}} \leq R$, where $R \in \mathbb{R}_+ < \infty$ is some positive constant. In (c), we note that $\mathbf{X}_m$ is positive semi-definite.

$$\mathbf{v}^\top \mathbf{X}_m \mathbf{v} = \mathbf{v}^\top \left(\sum_{i=1}^{m} \mathbf{k}_i \mathbf{k}_i^\top \right) \mathbf{v} = \sum_{i=1}^{m} \mathbf{v}^\top \mathbf{k}_i \mathbf{k}_i^\top \mathbf{v} = \sum_{i=1}^{m} \left(\mathbf{v}^\top \mathbf{k}_i\right)^2 \leq \sum_{i=1}^{n} \left(\mathbf{v}^\top \mathbf{k}_i\right)^2 \tag{53}$$

This concludes the proof. ■

C 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 [1]. This is because the Trimmed Maximum Likelihood Estimator is equivalent to minimizing over the subquantile of the likelihood.

Remark 28. Define the function $\Psi\left(t\right) \triangleq \min_{\mathbf{w}} g\left(t, \mathbf{w}\right)$

D Experimental Details

Our datasets are synthetic and are sourced from [5]

Dataset	**Dimension** d	**Sample Size** n	**Source**
`Polynomial`	3	1000	Ours
`Boston Housing`	13	506	[5]
`Concrete Data`	8	1030	[5]
`Wine Quality`	11	1599	[5]

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.

E Detailed Related Works

In this section we will give a detailed analysis of the relevant works.

E.1 High-dimensional Robust Mean Estimation via Gradient Descent [3]

In this work, Cheng et al. study high dimensional mean estimation when there exists an ϵ-fraction of adversarially corrupted data. They form a non-convex optimization problem based on a lemma from a previous paper of theirs minimize the objective with gradient descent. Let F be the objective function. First they define stationary points. Let $u \in \arg\max f(w)$, then a stationary point is defined as

$$\left(\nabla_w F(w,u)\right)^\top (\widetilde{w} - w) \geq 0 \quad \forall \widetilde{w} \in K \tag{54}$$

where K is a closed convex set. They show that any stationary point is a good point, i.e. $\|\mu_w - \mu^\star\| = \mathcal{O}(\epsilon\sqrt{\log(1/\epsilon)})$. Next, they show any approximate stationary point is a good point, i.e. if $\|\nabla f_\beta(w)\| = \mathcal{O}(\log(1/\epsilon))$, then $\|\mu_w - \mu^\star\| = \mathcal{O}(\epsilon\sqrt{\log(1/\epsilon)})$. Next, they show gradient descent converges to an approximate stationary point in a polynomial number of iterations.

Technical Results:

1. F is L-lipschitz, and β-smooth
2. To prove all stationary points are good, they prove by contradiction by showing if $\|\mu_w - \mu^\star\| > \mathcal{O}(\epsilon\sqrt{\log(1/\epsilon)})$, then there exists a corrupted point with a high gradient and a good point with a low gradient.
3. Let $f(w) \triangleq \max_u F(u,w)$ and $f_\beta(w) \triangleq \min_{\widetilde{w}} f(\widetilde{w}) + \beta \|w - \widetilde{w}\|_2^2$ be the Moreau envelope. They then prove $\|\nabla f_\beta(w)\| = \mathcal{O}(\log(1/\epsilon))$.
4. Then prove $\|\nabla f_\beta(w)\| = \mathcal{O}(\log(1/\epsilon))$ in a polynomial number of iterations w.r.t to n the sample size, and d the sample dimension.

E.2 Trimmed Maximum Likelihood Estimation for Robust Generalized Linear Model [1]

First we will give the algorithm

$$S^{(t)} = \underset{T \subset S^{(0)}: |T| = (1-2\epsilon)n}{\arg\min} \sum_{i \in T} -\log f(y_i | \langle \boldsymbol{\beta}^{(t)}, \mathbf{x}_i \rangle) \tag{55}$$

$$\boldsymbol{\beta}^{(t+1)} = \underset{\beta, \|\boldsymbol{\beta}\| \leq R}{\arg\min} \sum_{i \in S^{(t)}} -\log f(y_i | \langle \boldsymbol{\beta}^{(t)}, \mathbf{x}_i \rangle) \tag{56}$$

In Equation (55), the algorithm chooses the $(1-2\epsilon)n$ points giving the least error and put this in the set $S^{(t)}$. Next, in Equation (56), the algorithm then finds β that minimizes the negative log likelihood error for all the points in $S^{(t)}$ s.t. $\|\boldsymbol{\beta}\| \leq R$. For the theoretical analysis, Awasthi et al. consider a different approximation stationary point from [3].

$$\frac{1}{n} \sum_{i \in S} \nabla_{\boldsymbol{\beta}} \log f(y_i | \langle \boldsymbol{\beta}, \mathbf{x}_i \rangle)^\top \frac{(\boldsymbol{\beta}^\star - \boldsymbol{\beta})}{\|\boldsymbol{\beta}^\star - \boldsymbol{\beta}\|} \leq \gamma \tag{57}$$

We see Equation (57) is an upper bound, instead of a lower bound, of Equation (54). Next, they prove their algorithm reaches a η stationary point. Their proof does not use Moreau Envelopes or ideas in concave-non-convex optimization, rather they use the fact their algorithm terminates after it reaches a point when it can no longer make η improvement.