Subquantile Minimization for Kernel Learning in the Huber ϵ-Contamination Model

Arvind Rathnashyam* Alex Gittens†

October 14, 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 [19, 33]. 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. In our theoretical analysis, we analyze our non-convex concave objective function with the Moreau Envelope. We show (i) a stationary point with respect to the Moreau Envelope is a good point and (ii) we can reach a stationary point with gradient descent methods. Further, we analyze accelerated gradient methods for the non-convex concave minimax optimization problem. We empirically test Kernel Ridge Regression and Kernel Classification on various state of the art datasets and show Subquantile Minimization gives strong results. 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), [5, 1, 20, 24, 9] as well as for linear regression [2, 22]. Robust Statistics has been studied extensively [6] for problems such as high-dimensional mean estimation [26, 3] and Robust Covariance Estimation [4, 8]. Recently, there has been an interest in solving robust machine learning problems by gradient descent [27, 5]. Subquantile minimization aims to address the shortcomings of standard ERM in applications of noisy/corrupted data [17, 15]. 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** [13]). Given a corruption parameter $0 < \epsilon < 0.5$, a data matrix, 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 X and $\mathbf{y}$ as training data.

First, we will give our main error bounds.

Theorem 2. *(Informal). Let the dataset be given as $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$ such that εn are arbitrarily corrupted by an adversary. Let* K *be the kernel matrix and S be the points with the lowest error w.r.t $\widehat{\mathbf{w}}$, then Subquantile Minimization returns $\widehat{\mathbf{w}}$ such that for*
Kernel Ridge Regression:

$$\|\mathsf{f}_{\widehat{\mathbf{w}}} - \mathsf{f}_{\mathbf{w}^*}\|_{\mathcal{H}} \left(\frac{2L\left(\sum_{j\in P\setminus S} \eta_j^2\right)}{2L\sigma_{\min}\left(\sum_{i\in S\cap P} \phi(\mathbf{x}_i)\phi(\mathbf{x}_i)^\top\right) - 1} \right)^{1/2} + \frac{4L\left|\sum_{i\in S\cap P} \eta_i \sqrt{K(\mathbf{x}_i, \mathbf{x}_i)}\right|}{\sigma_{\min}\left(\sum_{i\in S\cap P} \phi(\mathbf{x}_i)\phi(\mathbf{x}_i)^\top\right)} + L\epsilon \tag{1}$$

Linear Regression:

$$\|\mathsf{f}_{\widehat{\mathbf{w}}} - \mathsf{f}_{\mathbf{w}^*}\|_{\mathcal{H}} \leq \tag{2}$$

Kernel Binary Classification:

$$\|\mathsf{f}_{\widehat{\mathbf{w}}} - \mathsf{f}_{\mathbf{w}^*}\|_{\mathcal{H}} \leq \tag{3}$$

Kernel Multi-Class Classification:

$$\|\mathsf{f}_{\widehat{\mathbf{w}}} - \mathsf{f}_{\mathbf{w}^*}\|_{\mathcal{H}} \leq \tag{4}$$

We will now state our main contributions clearly.
Contributions

1. We propose a gradient-descent based algorithm for robust kernel learning in the Huber ϵ-Contamination Model which is fast.
2. We rigorously analyze error bounds for subquantile minimization in the linear regression, kernel regression, kernel binary classification, and kernel multi-class classification tasks.
3. We give new bounds for accelerated gradient methods for accelerated gradient methods in nonconvex-concave minimax opimization.

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

[5] 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 data settings.

[20] 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 and lack of theoretical error bounds due to the nonlinearity induced by the exponential.

[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. We tend to follow their general framework but we adapt it the Reproducing Kernel Hilbert Space Norm and for our minimax objective.

Minimax Optimization

[16] 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.

[37] studied minimax optimization in the case of non-strong concavity.

1.2 Notation

The data matrix X is a fixed $n \times d$ matrix, the matrix K is the Gram Matrix, where $\mathrm{K}_{ij} = k(\mathbf{x}_i, \mathbf{x}_j)$ and $k(\cdot, \cdot)$ represents a kernel function, e.g. Linear kernel: $k(\mathbf{x}, \mathbf{y}) = \mathbf{x}^\top \mathbf{y}$, RBF kernel: $k(\mathbf{x}, \mathbf{y}) = \exp\left(-\gamma \left\|\mathbf{x} - \mathbf{y}\right\|_2^2\right)$. We denote $\mathrm{X}^\top = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n]$ where $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ represent the data vectors of the data matrix. We denote X as the set of all data vectors, $X = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$. We represent the data matrix $\mathrm{X} = \begin{pmatrix}\mathrm{P}^\top & \mathrm{Q}^\top\end{pmatrix}^\top$, 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 $\mathrm{I}_{k \times k}$ as the $k \times k$ identity matrix. The spectral norm of A is $\|\mathrm{A}\|_2 = \max_{\|\mathbf{x}\|=1} \|\mathrm{A}\mathbf{x}\| = \sigma_{\max}(\mathrm{A})$. The reproducing Hilbert Space Norm of f is given as $\|f\|_{\mathcal{H}} \triangleq \mathbf{w}^{\mathrm{T}} \mathrm{K} \mathbf{w}$ where $f(\cdot) = \sum_{i=1}^n w_i k(\mathbf{x}_i, \cdot)$.

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

Uppercase bold $(\mathrm{X}, \mathrm{Y}, \mathrm{Z}, \ldots)$ are matrices. Uppercase Roman are sets (X, S, P, Q). Lowercase bold are vectors $(\mathbf{x}, \mathbf{y}, \mathbf{z}, \ldots)$.

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.

$$\mathcal{Q}_p(U) \leq t \iff \mathbb{P}\{U \leq t\} \geq p \tag{5}$$

The p-subquantile of the risk is then given by

$$\mathbb{L}_p(U) = \frac{1}{p}\int_0^p \mathcal{Q}_p(U)\, dq = \mathbb{E}[U|U \leq \mathcal{Q}_p(U)] = \max_{t\in\mathbb{R}} \left\{ t - \frac{1}{p}\mathbb{E}(t-U)^+ \right\} \tag{6}$$

Given a convex objective function, ℓ, the kernelized learning poblem becomes:

$$\mathsf{f}_{\widehat{\mathbf{w}}} = \arg\min_{\mathsf{f}_{\mathbf{w}}\in\mathcal{K}} \max_{t\in\mathbb{R}} \left\{ g(t, \mathsf{f}_{\mathbf{w}}) \triangleq t - \sum_{i=1}^{n} \left(t - (\mathsf{f}_{\mathbf{w}}(\mathbf{x}_i) - y_i)^2\right)^+ \right\} \tag{7}$$

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 [30]. Let $\Pi_{\mathcal{K}}$ be the projection of a vector on to the convex set $\mathcal{K} \triangleq \{\mathsf{f} \in \mathcal{H} : \|\mathsf{f}\|_{\mathcal{H}} \leq R\}$, then our update steps are

$$t^{(k+1)} = \arg\max_{t\in\mathbb{R}} g(\mathsf{f}_{\mathbf{w}}^{(k)}, t) \tag{8}$$

$$\mathsf{f}_{\mathbf{w}}^{(k+1)} = \Pi_{\mathcal{K}}\left(\mathsf{f}_{\mathbf{w}}^{(k)} - \alpha\nabla_{\mathsf{f}} g(\mathsf{f}_{\mathbf{w}}^{(k)}, t^{(k+1)})\right) \tag{9}$$

Claim 3. *The function $g(t, \mathsf{f}_{\mathbf{w}})$ defined in Equation (7) is non-convex-concave, i.e. it is not convex with respect to $\mathsf{f}_{\mathbf{w}}$ and is concave with respect to t. Furthermore, $g(t, \mathsf{f}_{\mathbf{w}})$ is ρ-weakly convex where ρ is the β-smoothness factor of $g(t, \mathsf{f}_{\mathbf{w}})$ w.r.t $\mathsf{f}_{\mathbf{w}}$.*

Proof. We will show $-g(t, \mathsf{f}_{\mathbf{w}})$ is convex with respect to t. Let $\nu_i \triangleq (f_{\mathbf{w}}(\mathbf{x}_i) - y_i)^2$.

$$-g(\lambda t_1 + (1-\lambda)t_2, \mathsf{f}_{\mathbf{w}}) = -\lambda t_1 - (1-\lambda)t_2 + \sum_{i=1}^{n} (\lambda t_1 + (1-\lambda)t_2 - \nu_i)^+ \tag{10}$$

$$\leq -\lambda t_1 - (1-\lambda)t_2 + \sum_{i=1}^{n} \lambda(t_1 - \nu_i)^+ + (1-\lambda)(t_2 - \nu_i)^+ \tag{11}$$

$$= -\lambda g(t_1, \mathsf{f}_{\mathbf{w}}) - (1-\lambda) g(t_2, \mathsf{f}_{\mathbf{w}}) \tag{12}$$

Therefore we have $g(t, \mathsf{f}_{\mathbf{w}})$ is concave in t. Next we will prove $g(t, \mathsf{f}_{\mathbf{w}})$ is not convex in $\mathsf{f}_{\mathbf{w}}$. ■

Claim 4. *The function $g(t, \mathsf{f}_{\mathbf{w}})$ defined in Equation (7) is L-weakly convex in $\mathsf{f}_{\mathbf{w}}$, where L is lipschitz constant of the gradient of $g(t, \mathsf{f}_{\mathbf{w}})$ w.r.t $\mathsf{f}_{\mathbf{w}}$. This is true for Conditional Value at Risk, [**?**].*

Proof. Here we note that if we add $\max_{i\in[n]} |k(\mathbf{x}_i, \mathbf{x}_i) + f_{\mathbf{w}}(\mathbf{x}_i) - y_i|$ to each of the second derivatives, then we are pushing the trace to be negative . Note this is the L-lispchitz gradient. This is equivalent to $g(\mathsf{f}_{\mathbf{w}}, t) + \frac{L}{2}\|\mathbf{f}_{\mathbf{w}}\|_{\mathcal{H}}^2$. ■

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: $\mathrm{X}, (n \times d), n \gg d$; 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**

  $\mathrm{S}_{(k)} \leftarrow \textsc{Subquantile}(\mathbf{w}^{(k)}, \mathrm{X})$

  $\mathbf{w}^{(k+1)} \leftarrow \mathbf{w}^{(k)} - \alpha_{(k)} \nabla_{\mathbf{w}} g\left(t^{(k+1)}, \mathbf{w}^{(k)}\right)$

end

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

Algorithm 2: SUBQUANTILE

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

Output: Subquantile Matrix S

$\hat{\boldsymbol{\nu}}_i \leftarrow \ell(\mathbf{x}_i; \mathsf{f}_{\mathbf{w}}, y_i)$ 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 $\ell(\mathbf{x}_i; \mathsf{f}_{\mathbf{w}}, y_i) \leq t$

$\mathrm{S} \leftarrow \begin{pmatrix} \mathbf{x}_1^\top & \ldots & \mathbf{x}_{np}^\top \end{pmatrix}^\top$

return S

3 Structural Results

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. Now we will analyze the two-step minimax optimization steps described in Equations (8) and (9).

Lemma 5. *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.

$$\partial(-g(t, \mathsf{f}_{\mathbf{w}})) = \partial\left(-t + \frac{1}{np}\sum_{i=1}^{n}(t - \hat{\boldsymbol{\nu}}_i)^+\right) \tag{13}$$

$$= -1 + \frac{1}{np}\sum_{i=1}^{np}\begin{cases} 1 & \text{if } t > \hat{\boldsymbol{\nu}}_i \\ 0 & \text{if } t < \hat{\boldsymbol{\nu}}_i \\ [0,1] & \text{if } t = \hat{\boldsymbol{\nu}}_i \end{cases} \tag{14}$$

$$= 0 \text{ when } t = \hat{\boldsymbol{\nu}}_{np} \tag{15}$$

This is equivalent to the p-quantile of the Risk. ■
It therefore follows,

$$\sum_{i=1}^{n}\mathbb{I}\left\{\hat{\boldsymbol{\nu}}_{np} \geq \left(f_{\mathbf{w}}^{(k)}(\mathbf{x}_i) - y_i\right)^2\right\}\left(f_{\mathbf{w}}^{(k)}(\mathbf{x}_i) - y_i\right)^2 \in \max_{t \in \mathbb{R}} g(t, \mathsf{f}_{\mathbf{w}}^{(k)}) \tag{16}$$

Interpretation 6. From lemma 5, 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 7. *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 5, it then follows $\nabla_{\mathbf{w}} g(t^{(k)}, \mathsf{f}_{\mathbf{w}}^{(k)}) = \frac{1}{np}\sum_{i=1}^{np}\nabla f(\mathbf{x}_i; \mathsf{f}_{\mathbf{w}}^{(k)}, y_i)$*

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

$$\nabla_{\mathsf{f}} g(t^{(k+1)}, \mathsf{f}_{\mathbf{w}}^{(k)}) = \nabla_{\mathsf{f}}\left(\hat{\boldsymbol{\nu}}_{np} - \frac{1}{np}\sum_{i=1}^{n}\left(\hat{\boldsymbol{\nu}}_{np} - \ell(\mathbf{x}_i; \mathsf{f}_{\mathbf{w}}^{(k)}, y_i)\right)^+\right) \tag{17}$$

$$= -\frac{1}{np}\sum_{i=1}^{np}\nabla_{\mathsf{f}}\left(\hat{\boldsymbol{\nu}}_{np} - \ell(\mathbf{x}_i; \mathsf{f}_{\mathbf{w}}^{(k)}, y_i)\right)^+ \tag{18}$$

$$= \frac{1}{np}\sum_{i=1}^{n}\nabla_{\mathsf{f}}\ell(\mathbf{x}_i; \mathsf{f}_{\mathbf{w}}^{(k)}, y_i)\begin{cases} 1 & \text{if } t > \hat{\boldsymbol{\nu}}_i \\ 0 & \text{if } t < \hat{\boldsymbol{\nu}}_i \\ [0,1] & \text{if } t = \hat{\boldsymbol{\nu}}_i \end{cases} \tag{19}$$

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

$$\nabla_{\mathsf{f}} g(t^{(k+1)}, \mathsf{f}_{\mathbf{w}}^{(k)}) = \frac{1}{np}\sum_{i=1}^{np}\nabla_{\mathsf{f}}\ell(\mathbf{x}_i; \mathsf{f}_{\mathbf{w}}^{(k)}, y_i) \tag{20}$$

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

3.1 On the Softplus Approximation

It is clear our objective function is non-smooth. Thus we propose to use the Softplus approximation to smooth the function.

$$\zeta_\gamma(x) = \frac{1}{\lambda}\log\left(1+e^{\lambda x}\right) \tag{21}$$

We then have the approximation of g as

$$\widetilde{g}_\lambda(t, f_{\mathbf{w}}) \triangleq t - \sum_{i=1}^{n} \zeta_\lambda\left(t - \ell\left(f_{\mathbf{w}}; \mathbf{x}_i, y_i\right)\right) \tag{22}$$

$$= t - \frac{1}{np}\sum_{i=1}^{n}\frac{1}{\lambda}\log\left(1+\exp\left(\lambda\left(t-\ell\left(f_{\mathbf{w}};\mathbf{x}_i,y_i\right)\right)\right)\right) \tag{23}$$

Now we compute the derivatives,

$$\nabla_t \widetilde{g}_\lambda(t, f_{\mathbf{w}}) = \nabla_t\left(t - \frac{1}{np}\sum_{i=1}^{n}\frac{1}{\lambda}\ln\left(1+\exp\left(\lambda\left(t-\ell(f_{\mathbf{w}};\mathbf{x}_i,y_i)\right)\right)\right)\right) \tag{24}$$

$$= 1 - \frac{1}{np}\sum_{i=1}^{n}\sigma\left(\lambda\left(t-\ell(f_{\mathbf{w}};\mathbf{x}_i,y_i)\right)\right) \tag{25}$$

where $\sigma(\cdot)$ is the sigmoid function. It therefore follows,

$$\lim_{\lambda\to\infty}\nabla_t\widetilde{g}_\lambda(t, f_{\mathbf{w}}) = 1 - \frac{1}{np}\sum_{i=1}^{n}\mathbb{I}\left\{t-\ell\left(f_{\mathbf{w}};\mathbf{x}_i,y_i\right)\right\} \tag{26}$$

$$\nabla_f \widetilde{g}_\lambda(t, f_{\mathbf{w}}) = \nabla_f\left(t - \frac{1}{np}\sum_{i=1}^{n}\frac{1}{\lambda}\ln\left(1+\exp\left(\lambda\left(t-\ell(f_{\mathbf{w}};\mathbf{x}_i,y_i)\right)\right)\right)\right) \tag{27}$$

$$= \frac{1}{np}\sum_{i=1}^{n}\nabla_f\ell\left(f_{\mathbf{w}};\mathbf{x}_i,y_i\right)\sigma\left(\lambda\left(t-\ell(f_{\mathbf{w}};\mathbf{x}_i,y_i)\right)\right) \tag{28}$$

We therefore similarly have,

$$\lim_{\lambda\to\infty}\nabla_f\widetilde{g}_\lambda\left(t, f_{\mathbf{w}}\right) = \frac{1}{np}\sum_{i=1}^{n}\mathbb{I}\left\{t-\ell\left(f_{\mathbf{w}};\mathbf{x}_i,y_i\right)\right\}\nabla_f\ell\left(f_{\mathbf{w}};\mathbf{x}_i,y_i\right) \tag{29}$$

Then the second derivative is given by

$$\left[\nabla_f^2\widetilde{g}_\lambda(t, f_{\mathbf{w}})\right]_{i,j} = \left[\nabla_f\right]_j\left(\frac{1}{np}\sum_{i=1}^{n}\left[\nabla_f\right]_i\ell\left(f_{\mathbf{w}};\mathbf{x}_i,y_i\right)\sigma\left(\lambda\left(t-\ell(f_{\mathbf{w}};\mathbf{x}_i,y_i)\right)\right)\right) \tag{30}$$

$$= \frac{1}{np}\sum_{i=1}^{n}\Bigg(\left[\nabla_f^2\right]_{i,j}\ell\left(f_{\mathbf{w}};\mathbf{x}_i,y_i\right)\sigma\left(\lambda\left(t-\ell(f_{\mathbf{w}};\mathbf{x}_i,y_i)\right)\right)$$

$$-\left[\nabla_f\ell\left(f_{\mathbf{w}};\mathbf{x}_i,y_i\right)\right]_j\left[\nabla_f\ell\left(f_{\mathbf{w}};\mathbf{x}_i,y_i\right)\right]_i\sigma\left(\lambda\left(t-\ell(f_{\mathbf{w}};\mathbf{x}_i,y_i)\right)\right)\left(1-\sigma\left(\lambda\left(t-\ell(f_{\mathbf{w}};\mathbf{x}_i,y_i)\right)\right)\right)\Bigg) \tag{31}$$

We then similarly have,

$$\lim_{\lambda\to\infty}\left[\nabla_f^2\widetilde{g}_\lambda(t, f_{\mathbf{w}})\right]_{j,k} = \frac{1}{np}\sum_{i=1}^{n}\mathbb{I}\left\{t-\ell\left(f_{\mathbf{w}};\mathbf{x}_i,y_i\right)\right\}\left[\nabla_f^2\ell\left(f_{\mathbf{w}};\mathbf{x}_i,y_i\right)\right]_{j,k} \tag{32}$$

3.2 Weakly Convex Concave Optimization Theory

With our smoothed function, we are now able to use the weakly-convex concave minimization literature to analyze g.

Definition 8. (**Moreau Envelope**, [21]). Let f be proper lower semi-continuous convex function $f : \mathcal{X} \to \mathbb{R}$, then the Moreau Envelope is defined as:

$$f_\lambda(\mathbf{x}) \triangleq \inf_{\widehat{\mathbf{x}} \in \mathcal{X}} \left(f(\widehat{\mathbf{x}}) + \frac{1}{2\lambda} \left\| \mathbf{x} - \widehat{\mathbf{x}} \right\|_2^2 \right) \tag{33}$$

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

Definition 9. (**Moreau Envelope on closed, convex set**, [3]). Let f be proper lower semi-continuous convex function $\ell : \mathcal{K} \to \mathbb{R}$, where $\mathcal{K} \subset \mathcal{X}$ is a closed and convex set, then the Moreau Envelope is defined as:

$$\mathsf{M}_{\lambda \ell}(\mathsf{f}_{\mathbf{w}}) \triangleq \inf_{\mathsf{f}_{\widehat{\mathbf{w}}} \in \mathcal{K}} \left\{ \ell(\mathsf{f}_{\widehat{\mathbf{w}}}) + \frac{1}{2\lambda} \left\| \mathsf{f}_{\mathbf{w}} - \mathsf{f}_{\widehat{\mathbf{w}}} \right\|_{\mathcal{H}}^2 \right\} \tag{34}$$

Lemma 10. *Let ℓ be a proper convex function, $\ell : \mathcal{K} \to \mathbb{R}$, where $\mathcal{K} \subset \mathcal{H}$ is a closed and convex set. Define a new function, $\bar{\ell} : \mathcal{H} \to \mathbb{R}$,*

$$\bar{\ell}(\mathsf{f}_{\mathbf{w}}) = \begin{cases} \ell(\mathsf{f}_{\mathbf{w}}) & \mathsf{f}_{\mathbf{w}} \in \mathcal{K} \\ \infty & \mathsf{f}_{\mathbf{w}} \notin \mathcal{K} \end{cases} = \ell(\mathsf{f}_{\mathbf{w}}) + \mathbb{I}\left\{ \mathsf{f}_{\mathbf{w}} \in \mathcal{K} \right\} \tag{35}$$

where $\mathbb{I}\{\cdot\}$ is the convex set indicator function. Then the derivative of the Moreau Envelope defined in Definition 9 is given by

$$\nabla \mathsf{M}_{\lambda \bar{\ell}}(\mathsf{f}_{\mathbf{w}}) = \frac{1}{\lambda} \left(\mathsf{f}_{\mathbf{w}} - \arg\min_{\mathsf{f}_{\widehat{\mathbf{w}}} \in \mathcal{K}} \left\{ \ell(\mathsf{f}_{\widehat{\mathbf{w}}}) + \frac{1}{2\lambda} \left\| \mathsf{f}_{\mathbf{w}} - \mathsf{f}_{\widehat{\mathbf{w}}} \right\|_{\mathcal{H}}^2 \right\} \right) \tag{36}$$

Proof. It follows

$$\mathsf{M}_{\lambda \bar{\ell}}(\mathsf{f}_{\mathbf{w}}) = \inf_{\mathsf{f}_{\widehat{\mathbf{w}}} \in \mathcal{H}} \left\{ \bar{\ell}(\mathsf{f}_{\widehat{\mathbf{w}}}) + \frac{1}{2\lambda} \left\| \mathsf{f}_{\mathbf{w}} - \mathsf{f}_{\widehat{\mathbf{w}}} \right\|_{\mathcal{H}}^2 \right\} = \inf_{\mathsf{f}_{\widehat{\mathbf{w}}} \in \mathcal{K}} \left\{ \ell(\mathsf{f}_{\widehat{\mathbf{w}}}) + \frac{1}{2\lambda} \left\| \mathsf{f}_{\mathbf{w}} - \mathsf{f}_{\widehat{\mathbf{w}}} \right\|_{\mathcal{H}}^2 \right\} \tag{37}$$

since we are assuming $\mathcal{K}$ is non-empty. Then, from standard results in convex analysis, we have

$$\nabla \mathsf{M}_{\lambda \bar{\ell}}(\mathsf{f}_{\mathbf{w}}) = \frac{1}{\lambda} \left(\mathsf{f}_{\mathbf{w}} - \mathrm{prox}_{\lambda \bar{\ell}}(\mathsf{f}_{\mathbf{w}}) \right) = \frac{1}{\lambda} \left(\mathsf{f}_{\mathbf{w}} - \arg\min_{\mathsf{f}_{\widehat{\mathbf{w}}} \in \mathcal{H}} \left\{ \bar{\ell}\left(\mathsf{f}_{\widehat{\mathbf{w}}}\right) + \frac{1}{2\lambda} \left\| \mathsf{f}_{\mathbf{w}} - \mathsf{f}_{\widehat{\mathbf{w}}} \right\|_{\mathcal{H}}^2 \right\} \right) \tag{38}$$

$$= \frac{1}{\lambda} \left(\mathsf{f}_{\mathbf{w}} - \arg\min_{\mathsf{f}_{\widehat{\mathbf{w}}} \in \mathcal{K}} \left\{ \ell(\mathsf{f}_{\widehat{\mathbf{w}}}) + \frac{1}{2\lambda} \left\| \mathsf{f}_{\mathbf{w}} - \mathsf{f}_{\widehat{\mathbf{w}}} \right\|_{\mathcal{H}}^2 \right\} \right) \tag{39}$$

This concludes the proof. ■

With Lemma 10, we now have a finite Lipschitz Constant for the Kernel Regression which will allow us to use theory in Moreau Envelopes for convergence analysis. Our analysis will continue by looking at the function $\bar{\ell}$ rather than ℓ, although since we are doing a projection on to the convex set, so $\bar{\ell}$ does not change our gradient descent.

Remark 11. Define the function $\Phi(\mathsf{f}_{\mathbf{w}}) \triangleq \max_{t \in \mathbb{R}} g\left(t, \mathsf{f}_{\mathbf{w}}\right)$. This function is a L-weakly convex function in $\mathcal{K}$, i.e., $\Phi(\mathsf{f}_{\mathbf{w}}) + \frac{L}{2} \left\| \mathbf{f}_{\mathbf{w}} \right\|_{\mathcal{H}}^2$ is a convex function over $\mathbf{w}$ in the convex and compact set $\mathcal{K}$.

Assumption 12. Define $\Phi(\cdot)$ as the function in remark 11. Then it follows $\arg\min_{\mathbf{w} \in \mathbb{R}^d} \Phi\left(\mathbf{w}\right) = \mathbf{w}^\star$

3.3 Kernel Regression

The loss for the Kernel Ridge Regression problem for a single training pair $(\mathbf{x}_i, y_i)$ is given by the following equation

$$\ell\left(\mathbf{x}, y_i, ; \mathsf{f}_{\mathbf{w}}\right) = \left(\mathsf{f}_{\mathbf{w}}(\mathbf{x}_i) - y_i\right)^2 \tag{40}$$

For our theory, we need the L-lipschitz constant and β-smoothness constant.

Lemma 13. *(L-Lipschitz of $g\left(t, \mathsf{f}_{\mathbf{w}}\right)$ w.r.t $\mathsf{f}_{\mathbf{w}}$). Let $\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n$, represent the data vectors. It then follows for any $\mathsf{f}_{\mathbf{w}}, \mathsf{f}_{\widehat{\mathbf{w}}} \in \mathcal{K}$:*

$$\left|g\left(t, \mathsf{f}_{\mathbf{w}}\right) - g\left(t, \mathsf{f}_{\widehat{\mathbf{w}}}\right)\right| \leq L \left\|\mathsf{f}_{\mathbf{w}} - \mathsf{f}_{\widehat{\mathbf{w}}}\right\|_{\mathcal{H}} \tag{41}$$

where $L = \frac{2R}{np}(\sum_{i=1}^{n} \sqrt{K(\mathbf{x}_i, \mathbf{x}_i)})^2 + \frac{2\|\mathbf{y}\|}{np}(\sum_{i=1}^{n} \sqrt{K(\mathbf{x}_i, \mathbf{x}_i)})$

Lemma 14. *(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* X. *It then follows:*

$$\left\|\nabla_{\mathbf{w}} g\left(t, \mathsf{f}_{\mathbf{w}}\right) - \nabla_{\mathsf{f}} g\left(t, \mathsf{f}_{\widehat{\mathbf{w}}}\right)\right\| \leq \beta \left\|\mathsf{f}_{\mathbf{w}} - \mathsf{f}_{\widehat{\mathbf{w}}}\right\|_{\mathcal{H}} \tag{42}$$

where $\beta = \frac{2}{np}\left\|\sum_{i \in X} \mathbf{k}_i\right\|_{\mathcal{H}}^2$

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

$$\left\|\nabla_{\mathsf{f}}^2 g\left(t, \mathsf{f}_{\mathbf{w}}\right)\right\|_{\mathcal{H}} = \tag{43}$$

This concludes the proof. ■

Similar results for the Lipschitz Constant for non-kernelized learning algorithms can be seen in [38].

3.4 Kernel Binary Classification

The Negative Log Likelihood for the the Kernel Classification problem is given by the following equation for a single training pair $(\mathbf{x}_i, y_i)$

$$\ell\left(\mathbf{x}_i, y_i; \mathsf{f}_{\mathbf{w}}\right) = -\left(y_i \log\left(\sigma\left(\mathsf{f}_{\mathbf{w}}(\mathbf{x}_i)\right)\right) + (1 - y_i) \log\left(1 - \sigma\left(\mathsf{f}_{\mathbf{w}}(\mathbf{x}_i)\right)\right)\right) \tag{44}$$

Similar to § section 3.3, we require the L-Lipschitz constant and β-smoothness constant.

Lemma 15. *(L-Lipschitz of $g\left(t, \mathbf{w}\right)$ w.r.t $\mathbf{w}$). Let $\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n$, represent the data vectors. It then follows:*

$$\left|g\left(t, \mathsf{f}_{\mathbf{w}}\right) - g\left(t, \mathsf{f}_{\widehat{\mathbf{w}}}\right)\right| \leq L \left\|\mathsf{f}_{\mathbf{w}} - \mathsf{f}_{\widehat{\mathbf{w}}}\right\|_{\mathcal{H}} \tag{45}$$

where $L =$

Lemma 16. *(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* X. *It then follows:*

$$\left\|\nabla_{\mathsf{f}} g\left(t, \mathsf{f}_{\mathbf{w}}\right) - \nabla_{\mathsf{f}} g\left(t, \mathsf{f}_{\widehat{\mathbf{w}}}\right)\right\| \leq \beta \left\|\mathsf{f}_{\mathbf{w}} - \mathsf{f}_{\widehat{\mathbf{w}}}\right\|_{\mathcal{H}} \tag{46}$$

where $\beta =$

3.5 Kernel Multi-Class Classification

The Negative Log-Likelihood Loss for the the Kernel Multi-Class Classification problem is given by the following equation for a single training pair $(\mathbf{x}_i, y_i)$, note W is now a matrix

$$\ell\left(\mathbf{x}_i, y_i; \mathrm{W}\right) = -\sum_{j=1}^{|\mathcal{C}|} \mathbb{I}\left\{y_i = j\right\} \log\left(\frac{\exp\left(\mathrm{W}_k^\top \mathbf{k}_i\right)}{\sum_{h=1}^{|\mathcal{C}|} \exp\left(\mathrm{W}_h^\top \mathbf{k}_i\right)}\right) \tag{47}$$

Lemma 17. *(L-Lipschitz of $g\left(t, \mathbf{w}\right)$ w.r.t $\mathbf{w}$). Let $\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n$, represent the data vectors. It then follows:*

$$\left|g\left(t, \mathbf{w}\right) - g\left(t, \widehat{\mathbf{w}}\right)\right| \leq L \left\|\mathbf{w} - \widehat{\mathbf{w}}\right\|_{\mathcal{H}} \tag{48}$$

where $L = \frac{2R}{np}\left\|\sum_{i=1}^{n} \mathbf{k}_i\right\|_{\mathcal{H}}^2 + \frac{2}{np}\left\|\sum_{i=1}^{n} \mathbf{k}_i\right\|_{\mathcal{H}} \|\mathbf{y}\|_2$

3.6 Necessary Kernel Inequalities

We will first extend the idea of Resilience [34] to kernel learning.

Definition 18. (**Resilience**) from [34]. Let $\mathcal{H}$ represent the RKHS associated with the proper kernel $K : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$, then given the feature mapping $\phi : \mathbb{R}^d \to \mathcal{H}$, and the set $X = \{\mathbf{x}_i\}_{i=1}^n = P \cup Q$, such that $|P| = n(1-\epsilon)$ and $|Q| = n\epsilon$, It then follows for any subset $T \subseteq P$ such that $|T| = (1-2\varepsilon)n$,

$$\left\| \frac{1}{|T|} \sum_{i \in T} \phi\left(\mathbf{x}_i\right) - \mu_{\mathbb{P}} \right\| \leq \tau$$

where $\mu_{\mathbb{P}} = \mathbb{E}_{\mathbf{x} \sim \mathbb{P}}[\phi(\mathbf{x})]$ is the kernel mean embedding for the distribution, $\mathbb{P}$. We say the set X has (ϵ, τ)-resilience in the Reproducing Kernel Hilbert Space.

Without the idea of resilience defined in definition 18, we will be unable to put error bounds on our algorithm.

Definition 19. (**First-Order Stationary Point**). Let $\Phi(\mathbf{w}) = \max_t g(t, \mathbf{w})$. Then $\mathbf{w}$ is a first-order stationary point if

$$\left\| \nabla \Phi_\lambda\left(\mathbf{w}\right) \right\|_{\mathcal{H}} = 0 \tag{49}$$

i.e.

$$\mathsf{f}_{\mathbf{w}} = \arg\min_{\mathsf{f}_{\widehat{\mathbf{w}}} \in \mathcal{K}} \left(\Phi\left(\mathsf{f}_{\widehat{\mathbf{w}}}\right) + \frac{1}{2\lambda} \left\| \mathsf{f}_{\mathbf{w}} - \mathsf{f}_{\widehat{\mathbf{w}}} \right\|_{\mathcal{H}}^2 \right) \tag{50}$$

Lemma 20. *If* $\|\mathsf{f}_{\mathbf{w}} - \mathsf{f}_{\mathbf{w}^*}\| \geq \eta$, *then it follows*

$$\Phi\left(\mathsf{f}_{\mathbf{w}}\right) - \Phi\left(\mathsf{f}_{\mathbf{w}^*}\right) \geq \eta^2 \sigma_{\min} \left(\sum_{i \in S \cap P} \phi(\mathbf{x}_i) \phi(\mathbf{x}_i)^\top \right) - 2\eta \left(\sum_{i \in S \cap P} K(\mathbf{x}_i, \mathbf{x}_i) \right)^{1/2} \left(\sum_{i \in S \cap P} \eta_i^2 \right)^{1/2} - \sum_{j \in P \setminus S} \eta_j^2 \tag{51}$$

The proof is deferred to § **??**.

Theorem 21. *Let* $\mathsf{f}_{\widehat{\mathbf{w}}}$ *be a stationary point defined in* **??** *for the function* Φ *defined in* **??**. *Then,*

$$\eta \leq \left(\frac{2L \left(\sum_{j \in P \setminus S} \eta_j^2 \right)}{2L\sigma_{\min}\left(\sum_{i \in S \cap P} \phi(\mathbf{x}_i) \phi(\mathbf{x}_i)^\top \right) - 1} \right)^{1/2} + \frac{4L \left| \sum_{i \in S \cap P} \eta_i \sqrt{K(\mathbf{x}_i, \mathbf{x}_i)} \right|}{\sigma_{\min}\left(\sum_{i \in S \cap P} \phi(\mathbf{x}_i) \phi(\mathbf{x}_i)^\top \right)} \tag{52}$$

where L *is the Lipschitz Constant given in Lemma 13.*

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

Lemma 22. *([32, 31?]). Assume the function* Φ *is* ℓ*-weakly convex. Let* $\lambda < \frac{1}{\ell}$, *and denote* $\widehat{\mathbf{w}} = \arg\min_{\mathbf{w}'} \left(\Phi\left(\mathbf{w}'\right) + \frac{1}{2\lambda} \|\mathbf{w} - \mathbf{w}'\|_{\mathcal{H}}^2 \right)$, $\|\nabla \Phi_\lambda(\mathbf{w})\|_{\mathcal{H}} \leq \epsilon$ *implies:*

$$\|\widehat{\mathbf{w}} - \mathbf{w}\| = \lambda\epsilon \text{ and } \min_{\mathbf{g} \in \partial\Phi(\widehat{\mathbf{w}}) + \partial\mathcal{I}_{\mathcal{K}}(\widehat{\mathbf{w}})} \|\mathbf{g}\| \leq \epsilon \tag{53}$$

How to extend this to Hilbert Space Norm?

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

$$\partial\mathcal{I}_{\mathcal{K}} = \mathcal{N}(\widehat{\mathbf{w}}) = \{\mathsf{f}_{\widetilde{\mathbf{w}}} \in \mathbb{R}^n | \langle \widetilde{\mathbf{w}}, \bar{\mathbf{w}} - \widehat{\mathbf{w}} \rangle \leq \forall \bar{\mathbf{w}} \in \mathcal{K}\} \tag{54}$$

We will define a convex cone.

Definition 23. 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 24. *Let* $\widehat{\mathbf{w}} = \arg\min_{\mathbf{w}'} \left(\Phi\left(\mathbf{w}'\right) + \frac{1}{2\lambda} \left\| \mathbf{w} - \mathbf{w}' \right\|^2 \right)$ *s.t.* $\left\| \nabla \Phi_\lambda(\mathbf{w}) \right\|_{\mathcal{H}} \leq \epsilon$, *then it follows*

$$\left\| \widehat{\mathbf{w}} - \mathsf{f}_{\mathbf{w}^*} \right\|_{\mathcal{H}} \leq \Xi \tag{55}$$

Definition 25. (**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 8. Then we say that a point $\mathbf{w}$ is a ρ-approximate stationary point if $\left\| f_\beta(\mathbf{w}) \right\|_2 \leq \rho$.

The approximate stationary point in definition 25 is used in the analysis of the minimax algorithm in [3]. First, if you can prove a stationary point is good, theorem 21, then using lemma 22, 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 algorithm 1 converges to an approximate stationary point in a polynomial number of iterations.

4 Optimization Results

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. However, since our methods are in the kernelized setting. The 2-norm, $\|\mathbf{w} - \mathbf{w}^*\|$ is not sufficient, we want $\|\mathbf{w} - \mathbf{w}^*\|_{\mathcal{H}}$ to be close, as the RKHS being small indicates the function $f(\mathbf{x})$ and $f^*(\mathbf{x})$ will be close.

4.1 Optimization in the Reproducing Kernel Hilbert Space

In this section, we will discuss and give necessary optimization results in the RKHS norm.

4.2 Accelerated Gradient Methods

When working with big data it is often the case we need faster gradient methods as the gradient can be expensive to obtain. In this section, we give results on the convergence rate of accelerated gradient methods on the update of $\mathbf{w}$. We will analyze the convergence of three popular accelerated gradient methods.

4.2.1 Momentum Accelerated Gradient Descent

In this section we study Momentum Accelerated Gradient Descent [28, 25] with our non-convex-concave optimization algorithm.

$$\mathbf{b}^{(t)} = \mu \mathbf{b}^{(t-1)} + \nabla_{\mathsf{f}} \Phi\left(\mathsf{f}_{\mathbf{w}}^{(t-1)}\right) \tag{56}$$

$$\mathsf{f}_{\mathbf{w}}^{(t)} = \mathsf{f}_{\mathbf{w}}^{(t-1)} - \alpha \mathbf{b}^{(t)} \tag{57}$$

Theorem 26. *Momentum Accelerated Gradient Descent given in Equations* (56) *and* (57) *reaches a η-approximate stationary point. Algorithm 1 reaches a η-approximate stationary point in a polynomial number of iterations.*

$$\mathbb{E}\left[\left\|\nabla \Phi_{1/2\ell}\left(\mathsf{f}_{\mathbf{w}}\right)\right\|_{\mathcal{H}}^2\right] \leq \tag{58}$$

4.2.2 Nesterov Accelerated Gradient Descent

In this section we study Nesterov Accelerated Gradient Descent [23] with our non-convex-concave optimization algorithm.

$$\mathbf{b}^{(t+1)} = (1+\mu)\, \mathsf{f}_{\mathbf{w}}^{(t)} - \mu\, \mathsf{f}_{\mathbf{w}}^{(t-1)} \tag{59}$$

$$\mathsf{f}_{\mathbf{w}}^{(t+1)} = \mathbf{b}^{(t+1)} - \alpha \nabla_{\mathsf{f}} \Phi\left(\mathsf{f}_{\mathbf{w}}^{(t)}\right) \tag{60}$$

Theorem 27. *Nesterov Accelerated Gradient Descent given in Equations* (59) *and* (60) *reaches a η-approximate stationary point. Algorithm 1 reaches a η-approximate stationary point in a polynomial number of iterations.*

$$\mathbb{E}\left[\left\|\nabla \Phi_{1/2\ell}\left(\mathsf{f}_{\mathbf{w}}\right)\right\|_{\mathcal{H}}^2\right] \leq \tag{61}$$

5 Experiments

We perform numerical experiments on state of the art datasets comparing with other state of the art methods. We initialize the weights parameterizing $f_{\mathbf{w}}$ with the Glorot Initialization Scheme [10].

Algorithm 3: SUBQUANTILE-KERNEL

Input: Iterations: T; Quantile: p; Data Matrix: $\mathrm{X} \in \mathbb{R}^{n \times d}, n \gg d$; Labels: $\mathbf{y} \in \mathbb{R}^{n \times 1}$; Learning Rate schedule: $\alpha_1, \cdots, \alpha_T$; Ridge parameter: λ

Output: Trained Parameters: $\mathsf{f}_{\mathbf{w}}^{(T)}$

$\mathbf{w}_i^{(0)} \leftarrow \text{Unif}\left[-\sqrt{\frac{6}{n}}, \sqrt{\frac{6}{n}}\right], \forall i \in [n]$ ▷ Base Learner

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

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

  $\nabla_{\mathsf{f}} g\left(t^{(k+1)}, \mathsf{f}_{\mathbf{w}}^{(k)}\right) \leftarrow 2 \sum_{i \in S^{(k)}} \left(\mathsf{f}_{\mathbf{w}}^{(k)}(\mathbf{x}_i) - y_i\right) \cdot K(\mathbf{x}_i, \cdot)$ ▷ Regression

  $\nabla_{\mathsf{f}} g\left(t^{(k+1)}, \mathsf{f}_{\mathbf{w}}^{(k)}\right) \leftarrow \sum_{i \in S^{(k)}} \left(\sigma\left(\mathsf{f}_{\mathbf{w}}^{(k)}(\mathbf{x}_i)\right) - y_i\right) \cdot K(\mathbf{x}_i, \cdot)$ ▷ Binary Classification

  $\mathsf{f}_{\mathbf{w}}^{(k+1)} \leftarrow \mathsf{f}_{\mathbf{w}}^{(k)} - \alpha_{(k)} \nabla_{\mathsf{f}} g\left(t^{(k+1)}, \mathsf{f}_{\mathbf{w}}^{(k)}\right)$ ▷ $\mathbf{w}$-update in eqn. (9)

end

return $\mathsf{f}_{\mathbf{w}}^{(T)}$

Algorithms	Test RMSE							
	`Concrete`		`Wine Quality`		`Boston Housing`		`Drug`	
	$\epsilon = 0.2(\downarrow)$	$\epsilon = 0.4(\downarrow)$	$\epsilon = 0.2(\downarrow)$	$\epsilon = 0.4(\downarrow)$	$\epsilon = 0.2(\downarrow)$	$\epsilon = 0.4(\downarrow)$	$\epsilon = 0.2(\downarrow)$	$\epsilon = 0.4(\downarrow)$
KRR	$1.355_{(0.0934)}$	$2.282_{(0.2063)}$	$1.437_{(0.0979)}$	$2.272_{(0.1088)}$	$1.285_{(0.0896)}$	$2.266_{(0.0686)}$	$1.478_{(0.0533)}$	$2.381_{(0.0203)}$
TERM [20]	$0.829_{(0.0422)}$	$0.928_{(0.0197)}$	$1.854_{(0.7437)}$	$1.069_{(0.1001)}$	$0.879_{(0.0178)}$	$0.875_{(0.0711)}$	∞	∞
SEVER [5]	$\underline{0.533}_{(0.0347)}$	$\underline{0.592}_{(0.0548)}$	$\underline{0.915}_{(0.0343)}$	$\underline{0.841}_{(0.0413)}$	$\underline{0.526}_{(0.0287)}$	$\underline{0.720}_{(0.1147)}$	$\mathbf{1.172}_{(0.0542)}$	$\mathbf{1.215}_{(0.0536)}$
SUBQUANTILE	$\mathbf{0.519}_{(0.0134)}$	$\mathbf{0.547}_{(0.0174)}$	$\mathbf{0.808}_{(0.0389)}$	$\mathbf{0.827}_{(0.0216)}$	$\mathbf{0.468}_{(0.0896)}$	$\mathbf{0.458}_{(0.0662)}$	$\underline{1.280}_{(0.0568)}$	$\underline{1.372}_{(0.0294)}$
Genie ERM	∞	∞	∞	∞	∞	∞	∞	∞

Table 1: `Boston Housing`, `Concrete Data`, `Wine Quality`, and `Drug` and `Polynomial Synthetic` Dataset. Label Noise: $y_{\text{noise}} \sim \mathcal{N}(5, 5)$. Feature Noise: $y_{\text{noise}} = 10000 y_{\text{original}}$ and $\mathbf{x}_{\text{noise}} = 100 \mathbf{x}_{\text{original}}$. `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 $a_i \sim \mathcal{N}(0, 1)$. The Radial Basis Function is used in first three experiments and polynomail kernel with degree 3 and $C = 1$ is used in the last experiment.

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 **??** and figures fig. 1, subquantile minimization has near optimal performance in the `Polynomial Regression` Synthetic Dataset.

5.1 Linear Regression

In this section, we give experimental results for datasets using the linear kernel. This section will serve as a comparison to the various Robust Linear Regression Algorithms developed which are not meta-algorithms.

5.2 Kernel Binary Classification

In this section we will give the algorithm for subquantile minimization for the kernel classification problem and then give some experimental results on state of the art datasets comparing against other state of the art robust algorithms.

Now we will give some experimental results.

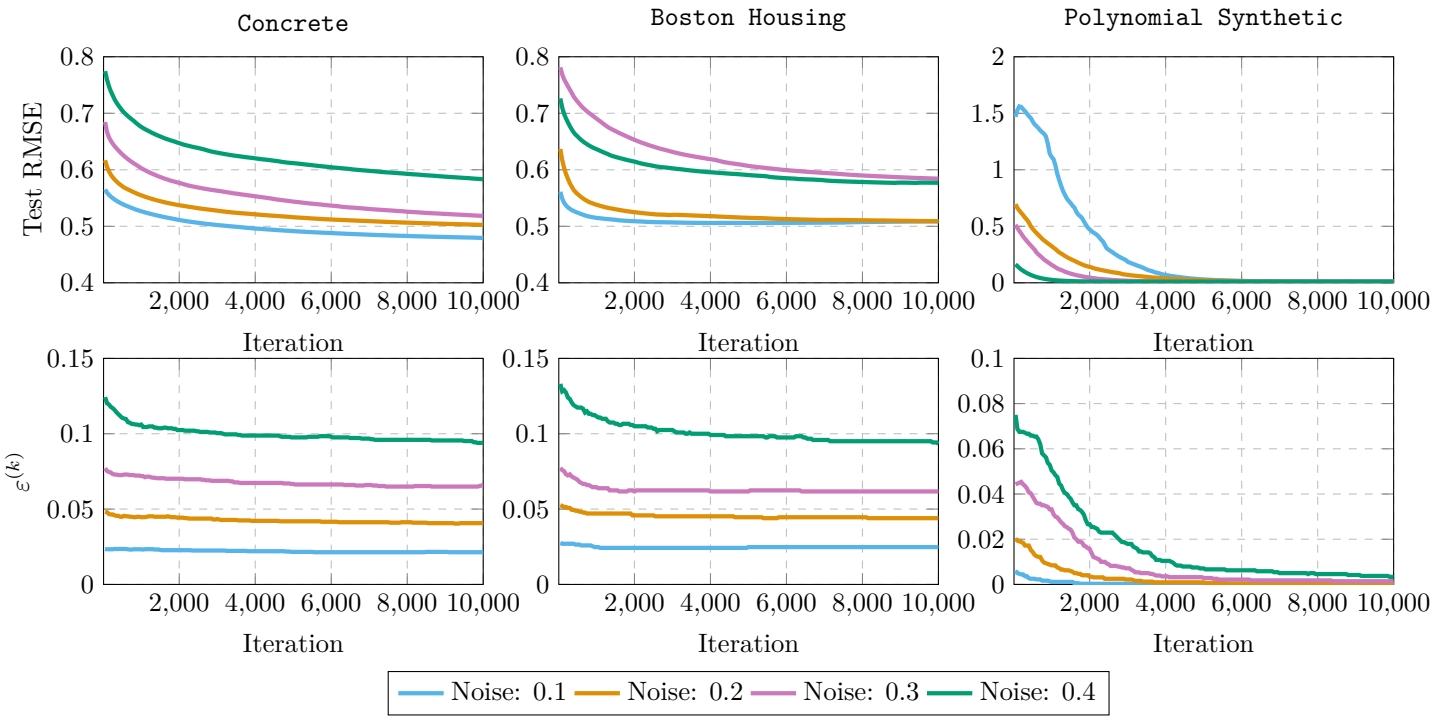

Figure 1: Test RMSE over the iterations in `Concrete`, `Boston Housing`, and `Polynomial` Datasets for SUBQUANTILE at different noise levels

Algorithms	Test RMSE							
	`Boston Housing`		`Wine Quality`		`Concrete`		`Drug`	
	Label(↓)	Label+Feature	Label	Label+Feature	Label	Label+Feature	Label	Label+Feature
KRR	$0.907_{(0.2724)}$	$90.799_{(5.7170)}$	$0.894_{(0.0404)}$	$62.913_{(7.4959)}$	$0.747_{(0.0465)}$	$77.383_{(5.5692)}$	$2.679_{(0.1286)}$	$141.690_{(3.5297)}$
RANSAC [9]	$1.167_{(0.6710)}$	$22.460_{(19.1987)}$	$1.489_{(0.2730)}$	$39.630_{(13.0294)}$	$0.870_{(0.2308)}$	$23.629_{(16.1023)}$	$2.801_{(0.2004)}$	$117.389_{(8.3915)}$
CRR [2]	$0.636_{(0.0905)}$	$88.626_{(5.7380)}$	$\mathbf{0.814}_{(0.0481)}$	$58.488_{(3.5612)}$	$0.710_{(0.0919)}$	$73.932_{(4.7867)}$	$1.887_{(0.1463)}$	$152.827_{(6.6038)}$
STIR [22]	$\underline{0.562}_{(0.0626)}$	$78.878_{(8.0164)}$	$0.828_{(0.0293)}$	$58.352_{(4.6700)}$	$\mathbf{0.684}_{(0.0245)}$	$76.555_{(4.5927)}$	$1.721_{(0.1520)}$	$144.975_{(5.4953)}$
SEVER [5]	$0.601_{(0.0979)}$	$5.980_{(8.2603)}$	$\underline{0.818}_{(0.0373)}$	$9.065_{(13.7632)}$	$\mathbf{0.684}_{(0.0438)}$	$4.119_{(8.2436)}$	$1.469_{(0.1162)}$	$156.043_{(4.5543)}$
TERM [20]	$0.608_{(0.1357)}$	$\underline{0.569}_{(0.0620)}$	$0.840_{(0.0563)}$	$\underline{0.827}_{(0.0255)}$	$0.830_{(0.0934)}$	$\underline{0.808}_{(0.0726)}$	$\mathbf{1.185}_{(0.1077)}$	$\mathbf{1.147}_{(0.1258)}$
SUBQUANTILE	$\mathbf{0.503}_{(0.0470)}$	$\mathbf{0.560}_{(0.0373)}$	$0.838_{(0.0260)}$	$\mathbf{0.821}_{(0.0305)}$	$0.757_{(0.1174)}$	$\mathbf{0.630}_{(0.0269)}$	$\underline{1.244}_{(0.1091)}$	$\underline{2.413}_{(0.6737)}$
Genie ERM	$0.630_{(0.1015)}$	$0.665_{(0.1134)}$	$0.838_{(0.0130)}$	$0.865_{(0.0222)}$	$0.763_{(0.0390)}$	$0.768_{(0.0181)}$	$0.988_{(0.0823)}$	$0.985_{(0.0838)}$

Table 2: For only Label Noise, $y_{\text{noisy}} \sim \mathcal{N}(5, 5)$. For Label and Feature Noise $\mathbf{x}_{\text{noisy}} = 100\mathbf{x}_{\text{original}}$ and $y_{\text{noisy}} = 10000 y_{\text{original}}$.

5.3 Kernel Multi-Class Classification

In this section we will provide some experimental results on the multi-class classification task.

5.4 Accelerated Gradient Methods

In this section we give some empirical results on using different accelerated gradient methods described in § Section 4. In Figure 2, we see using momentum can give significantly faster convergence.

Algorithms	Test Accuracy							
	`Heart Disease`				`Breast Cancer`			
	Label		Label+Feature		Label		Label+Feature	
	$\epsilon = 0.2(\uparrow)$	$\epsilon = 0.4(\uparrow)$	$\epsilon = 0.2(\uparrow)$	$\epsilon = 0.4(\uparrow)$	$\epsilon = 0.2(\uparrow)$	$\epsilon = 0.4(\uparrow)$	$\epsilon = 0.2(\uparrow)$	$\epsilon = 0.4(\uparrow)$
SVM	$\underline{0.826}_{(0.0482)}$	$\mathbf{0.625}_{(0.0475)}$	$0.549_{(0.0763)}$	$0.513_{(0.0344)}$	$\underline{0.940}_{(0.0140)}$	$0.781_{(0.0496)}$	$0.651_{(0.0370)}$	$0.633_{(0.0428)}$
SEVER [5]	$0.728_{(0.1134)}$	$0.531_{(0.0321)}$	$0.807_{(0.0557)}$	$\mathbf{0.677}_{(\mathbf{0.1552})}$	$0.918_{(0.0439)}$	$0.575_{(0.1456)}$	$0.954_{(0.0146)}$	$0.809_{(0.0947)}$
TERM [20]	$0.790_{(0.0420)}$	$\underline{0.610}_{(0.0667)}$	$\underline{0.816}_{(0.0485)}$	$\underline{0.607}_{(0.1318)}$	$0.937_{(0.0151)}$	$\mathbf{0.793}_{(\mathbf{0.0691})}$	$\mathbf{0.972}_{(\mathbf{0.0151})}$	$\mathbf{0.973}_{(\mathbf{0.0190})}$
SUBQUANTILE	$\mathbf{0.846}_{(\mathbf{0.0382})}$	$0.597_{(0.0794)}$	$\mathbf{0.830}_{(0.0471)}$	$0.516_{(0.0623)}$	$\mathbf{0.956}_{(0.0055)}$	$\underline{0.785}_{(0.1059)}$	$\underline{0.955}_{(0.0159)}$	$\underline{0.823}_{(0.1445)}$
Genie ERM	∞	∞	∞	∞	∞	∞	∞	∞

Table 3: `Heart Disease` and `Breast Cancer` Dataset. Label Noise: $y_{\text{noise}} = \mathbb{I}\{y_{\text{original}} = 0\}$. Feature Noise: $\mathbf{x}_{\text{noise}} = 100\mathbf{x}_{\text{original}}$. The Radial Basis Function is used in all experiments.

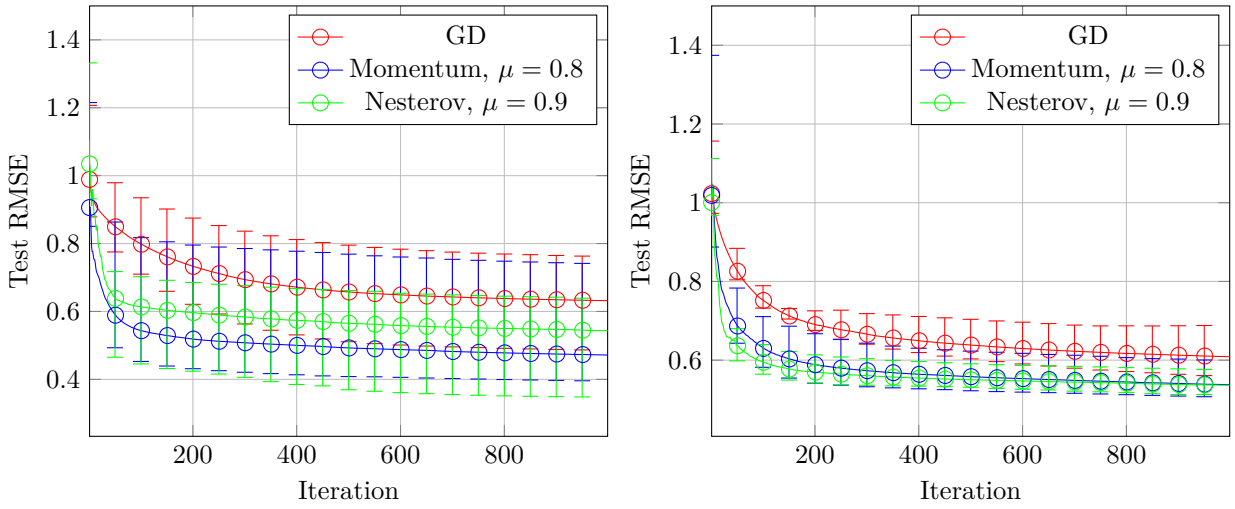

Figure 2: The effect of Momentum for Kernel Regression in the `Boston Housing` dataset *(Left)*and `Concrete` dataset *(Right)*. We observe faster convergence. In both experiments, we use the Radial Basis Function.

6 Discussion

The main contribution of this paper is the study of a nonconvex-concave formulation of Subquantile minimization for the robust learning problem for kernel ridge regression and kernel classification. We present an algorithm to solve the nonconvex-concave formulation and prove rigorous error bounds which show that the more good data that is given decreases the error bounds. We also present accelerated gradient methods for the two-step algorithm to solve the nonconvex-concave optimization problem and give novel theoretical bounds.

Theory. From our theoretical bounds, we see the more good data we obtain, the closer the resultant function is to the optimal function in the RKHS, an important idea in learning theory. We see we obtain an approximately close function to the optimal with high probability. Furthermore, we extend minimax optimization with maximizing oracle gradient descent literature with novel bounds for accelerated gradient descent.

Experiments. From our experiments, we see Subquantile Minimization is competitive with algorithms developed solely for robust linear regression as well as other meta-algorithms. Our theoretical analysis is through the lens of kernel-learning, but the generalization to linear regression from a non-kernel perspective can be done. In kernelized regression, we see SUBQUANTILE is the strongest of the meta-algorithms. Furthemore, in binary and multi-class classification, SUBQUANTILE is very strong. Thus, we can see empirically SUBQUANTILE is the strongest meta-algorithm across all kernelized regression and classification tasks and also the strongest algorithm in linear regression.

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.

The analysis of Subquantile Minimization can be extended to neural networks. This generalization will be appear in subsequent work.

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] Yu Cheng, Ilias Diakonikolas, Rong Ge, and David P Woodruff. Faster algorithms for high-dimensional robust covariance estimation. In *Conference on Learning Theory*, pages 727–757. PMLR, 2019.

[5] 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.

[6] Ilias Diakonikolas and Daniel M Kane. *Algorithmic high-dimensional robust statistics*. Cambridge University Press, 2023.

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

[8] Jianqing Fan, Weichen Wang, and Yiqiao Zhong. An l∞ eigenvector perturbation bound and its application to robust covariance estimation. *Journal of Machine Learning Research*, 18(207):1–42, 2018.

[9] 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.

[10] Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 249–256. JMLR Workshop and Conference Proceedings, 2010.

[11] Arthur Gretton. Introduction to rkhs, and some simple kernel algorithms. *Adv. Top. Mach. Learn. Lecture Conducted from University College London*, 16(5-3):2, 2013.

[12] David Gross. Recovering low-rank matrices from few coefficients in any basis. *IEEE Transactions on Information Theory*, 57(3):1548–1566, 2011.

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

[14] Steinbrunn William Pfisterer Matthias Janosi, Andras and Robert Detrano. Heart Disease. UCI Machine Learning Repository, 1988. DOI: https://doi.org/10.24432/C52P4X.

[15] 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.

[16] 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.

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

[18] Jonas Moritz Kohler and Aurelien Lucchi. Sub-sampled cubic regularization for non-convex optimization. In *International Conference on Machine Learning*, pages 1895–1904. PMLR, 2017.

[19] 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.

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

[21] Jean-Jacques Moreau. Proximité et dualité dans un espace hilbertien. *Bulletin de la Société mathématique de France*, 93:273–299, 1965.

[22] 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.

[23] Yurii Evgen'evich Nesterov. A method of solving a convex programming problem with convergence rate o\bigl(k^2\bigr). In *Doklady Akademii Nauk*, volume 269, pages 543–547. Russian Academy of Sciences, 1983.

[24] 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.

[25] Boris T Polyak. Some methods of speeding up the convergence of iteration methods. *Ussr computational mathematics and mathematical physics*, 4(5):1–17, 1964.

[26] Adarsh Prasad, Sivaraman Balakrishnan, and Pradeep Ravikumar. A unified approach to robust mean estimation. *arXiv preprint arXiv:1907.00927*, 2019.

[27] Adarsh Prasad, Arun Sai Suggala, Sivaraman Balakrishnan, and Pradeep Ravikumar. Robust estimation via robust gradient estimation. *arXiv preprint arXiv:1802.06485*, 2018.

[28] Ning Qian. On the momentum term in gradient descent learning algorithms. *Neural networks*, 12(1):145–151, 1999.

[29] 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.

[30] 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.

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

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

[33] 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.

[34] 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.

[35] Joel A Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of computational mathematics*, 12:389–434, 2012.

[36] Joel A Tropp et al. An introduction to matrix concentration inequalities. *Foundations and Trends® in Machine Learning*, 8(1-2):1–230, 2015.

[37] 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.

[38] Rahul Yedida, Snehanshu Saha, and Tejas Prashanth. Lipschitzlr: Using theoretically computed adaptive learning rates for fast convergence. *Applied Intelligence*, 51:1460–1478, 2021.

A Concentration Inequalities

In this section we will give various concentration inequalities on the inlier data. We first restate our assumptions.

Assumption 28. (Sub-Gaussian Design of Covariates). We assume each covariate is drawn i.i.d from a zero-mean covariance $\mathbf{\Sigma}$ sub-Gaussian distribution with sub-Gaussian norm $K_1 \in \mathbb{R}_{++}$.

$$\mathbb{E}\left[\mathbf{x}_i\right] = \mathbf{0} \tag{62}$$

$$\mathbb{E}\left[\mathbf{x}_i \mathbf{x}_i^\top\right] = \mathbf{\Sigma} \tag{63}$$

For all $i \in [n]$,

$$\mathbb{P}\left\{\mathbf{v}^\top \mathbf{x}_i \geq t\right\} \leq \exp\left(-\frac{t^2}{K_1^2}\right) \tag{64}$$

Is this equivalent to

$$\mathbb{P}\left\{\mathbf{v}^\top \mathbf{x}_i \geq t\right\} \leq \exp\left(-\frac{t^2}{K_1^2}\right) \tag{65}$$

Assumption 29. (Sub-Gaussian Design of Optimal Residuals). Recall the residual is defined as $\eta_i \triangleq \mathsf{f}^*_{\mathbf{w}}(\mathbf{x}_i) - y_i$. Then we assume for some $K_2 \in \mathbb{R}_{++}$

$$\mathbb{E}\left[\eta_i\right] = 0 \tag{66}$$

$$\mathbb{P}\left\{|\eta_i| \geq t\right\} \leq 2\exp\left(-\frac{t^2}{K_2^2}\right) \tag{67}$$

A.1 Linear Kernel

We will first begin with the Linear Kernel Case.

Assumption 30. Let the data have dimension d such that Assumption 28 holds. Then assume

$$n > \frac{1}{18}\left(\sqrt{\left(18 - 24dK\right)^2 - 36d^2K^2\log\left(2d\right)} - \left(18 - 24dK\right)\log\left(2d\right)\right) = \Omega\left(dK\log\left(2d\right)\right) \tag{68}$$

Lemma 31. *Let $\mathbf{x}_1, \ldots, \mathbf{x}_n$ have sub-Gaussian design described in Assumption 28 and $\eta_i, \ldots, \eta_n$ have sub-Gaussian design described in Assumption 29. It then follows*

$$\mathbb{E}\left\|\sum_{i=1}^n \eta_i \mathbf{x}_i\right\| \leq \frac{1}{2}K_1K_2e^{1/4}\sqrt{8\pi nd} = \mathcal{O}\left(K_1K_2\sqrt{nd}\right) \tag{69}$$

and with probability at least 0.99 *we have*

$$\left\|\sum_{i=1}^n \eta_i \mathbf{x}_i\right\|_2 \leq 7K_1K_2\sqrt{nd} \tag{70}$$

Proof. We will first upper bound the event $\|\sum_{i=1}^n \eta_i \mathbf{x}_i\| \geq t$. To this we utilize the vector Bernstein Inequality.

$$\|\mathbf{x}\|_2 = \sqrt{\sum_{i=1}^d x_i^2} \leq \sqrt{dK_1^2} \leq \sqrt{d}K_1 \tag{71}$$

$$\mathbb{E}\left[\left\|\eta_i\mathbf{x}_i\right\|^2\right] = \mathbb{E}\left[\eta_i^2\left\|\mathbf{x}_i\right\|^2\right] = \mathbb{E}\left[\eta_i^2\right]\mathbb{E}\left[\sum_{i=1}^d x_i^2\right] \leq dK_1^2K_2^2 \tag{72}$$

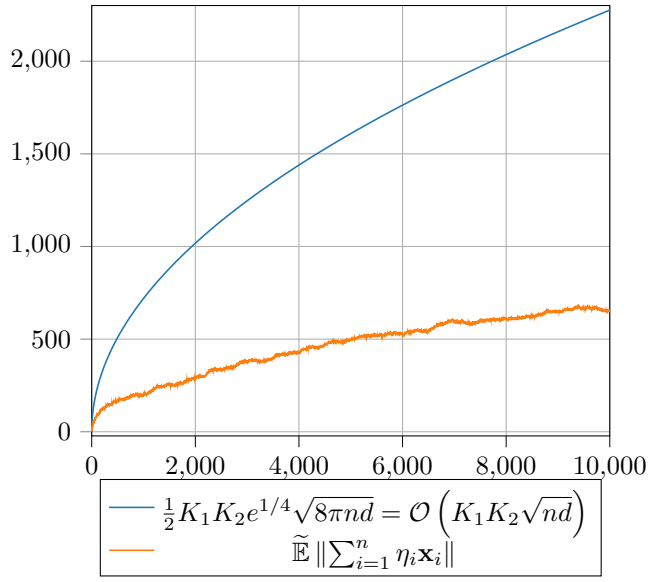

Figure 3: The bound for **??** compared to an average over 10 trials. We set $\mathbf{\Sigma} = \mathrm{I}$, $K = 1$, $d = 20$.

$$\mathbb{E}\left[\left\|\sum_{i=1}^n \eta_i\mathbf{x}_i\right\|\right] = \int_0^\infty \mathbb{P}\left\{\left\|\sum_{i=1}^n \eta_i\mathbf{x}_i\right\| > t\right\} \tag{73}$$

$$\leq \int_0^\infty \exp\left(-\frac{t^2}{8ndK_1^2K_2^2} + \frac{1}{4}\right) \tag{74}$$

$$= \frac{1}{2}K_1K_2e^{1/4}\sqrt{8\pi nd} \tag{75}$$

In probability, we apply the Vector Bernstein Inequality and obtain with probability at least $1-\delta$

$$\left\|\sum_{i=1}^n \eta_i\mathbf{x}_i\right\|_2 \leq K_1K_2\sqrt{nd} + 2K_1K_2\sqrt{nd\ln\left(\frac{1}{\delta}\right)} \tag{76}$$

This completes the proof ■

Lemma 32. *Let $\mathbf{x}_1, \ldots, \mathbf{x}_n$ have sub-Gaussian design described in Assumption 28 such that $n > d\log(d)$. It then follows*

$$\mathbb{E}\left[\sigma_{\min}\left(\sum_{k\in[n]}\mathbf{x}_k\mathbf{x}_k^\top\right)\right] \geq \sigma_{\min}\left(\mathbf{\Sigma}\right)\left(n - \left(\sqrt{2(dK-1)n\log(2d)} + \frac{1}{3}dK\log(2d)\right)\right) = \Omega\left(\sigma_{\min}\left(\mathbf{\Sigma}\right)n\right) \tag{77}$$

and with probability $(1-\delta)$

$$\sigma_{\min}\left(\sum_{k\in[n]}\mathbf{x}_k\mathbf{x}_k^\top\right) \geq \frac{1}{1-\delta}\sigma_{\min}\left(\mathbf{\Sigma}\right)\left(n - \left(\sqrt{2(dK-1)n\log(2d)} + \frac{1}{3}dK\log(2d)\right)\right) \tag{78}$$

Proof. In expectation we have

$$\mu_{\min} \triangleq \sigma_{\min}\left(\sum_{i=1}^n \mathbb{E}\left[\mathbf{x}_i\mathbf{x}_i^\top\right]\right) = \sigma_{\min}\left(\sum_{i=1}^n \mathbf{\Sigma}\right) = n\sigma_{\min}\left(\mathbf{\Sigma}\right) \tag{79}$$

Then we have by the sub-Gaussian design in Assumption 28.

$$\sigma_{\min}\left(\sum_{k\in[n]}\mathbf{x}_k\mathbf{x}_k^\top\right) = \sigma_{\min}\left(n\mathbf{\Sigma} + \sum_{k\in[n]}\mathbf{x}_k\mathbf{x}_k^\top - n\mathbf{\Sigma}\right) \tag{80}$$

$$\overset{\text{Weyl's}}{\geq} \sigma_{\min}\left(n\boldsymbol{\Sigma}\right) - \sigma_{\max}\left(\sum_{k\in[n]} \mathbf{x}_k\mathbf{x}_k^\top - n\boldsymbol{\Sigma}\right) \tag{81}$$

$$= \mu_{\min} - \underbrace{\left\|\sum_{k\in[n]} \mathbf{x}_k\mathbf{x}_k^\top - \boldsymbol{\Sigma}\right\|_2}_{\text{A}} \tag{82}$$

We assume $\boldsymbol{\Sigma} \succcurlyeq \mathbf{0}$ and symmetric, therefore we can represent $\boldsymbol{\Sigma} = (\boldsymbol{\Sigma}^{1/2})^\top\boldsymbol{\Sigma}^{1/2} = \boldsymbol{\Sigma}^{1/2}(\boldsymbol{\Sigma}^{1/2})^\top$. Next we will upper bound A.

$$\text{A} \triangleq \left\|\sum_{k\in[n]} \mathbf{x}_k\mathbf{x}_k^\top - \boldsymbol{\Sigma}\right\|_2 \tag{83}$$

$$= \left\|\sum_{k\in[n]} \left(\boldsymbol{\Sigma}^{1/2}\widehat{\mathbf{x}}_k\right)\left(\boldsymbol{\Sigma}^{1/2}\widehat{\mathbf{x}}_k\right)^\top - \boldsymbol{\Sigma}\right\|_2 \tag{84}$$

$$= \left\|\sum_{k\in[n]} \boldsymbol{\Sigma}^{1/2}\widehat{\mathbf{x}}_k\widehat{\mathbf{x}}_k^\top\left(\boldsymbol{\Sigma}^{1/2}\right)^\top - \boldsymbol{\Sigma}\right\|_2 \tag{85}$$

$$= \left\|\boldsymbol{\Sigma}^{1/2}\left(\sum_{k\in[n]} \mathbf{x}_k\mathbf{x}_k^\top - \mathrm{I}\right)\left(\boldsymbol{\Sigma}^{1/2}\right)^\top\right\|_2 \tag{86}$$

$$\leq \underbrace{\left\|\sum_{k\in[n]} \widehat{\mathbf{x}}_k\widehat{\mathbf{x}}_k^\top - \mathrm{I}\right\|}_{\text{B}} \|\boldsymbol{\Sigma}\| \tag{87}$$

It then suffices to upper bound B. First, let $\widehat{\mathbf{x}} \sim \mathcal{N}(\mathbf{0}, \mathrm{I})$ and define $K \geq 1$ s.t. $\widehat{\mathbf{x}} \leq dK$, then

$$\mathbb{V}\left(\widehat{\mathbf{x}}\widehat{\mathbf{x}}^\top - \mathrm{I}\right) = \mathbb{E}\left[\left(\widehat{\mathbf{x}}\widehat{\mathbf{x}}^\top - \mathrm{I}\right)^\top\left(\widehat{\mathbf{x}}\widehat{\mathbf{x}}^\top - \mathrm{I}\right)\right] \tag{88}$$

$$= \mathbb{E}\left[\widehat{\mathbf{x}}\widehat{\mathbf{x}}^\top\widehat{\mathbf{x}}\widehat{\mathbf{x}}^\top\right] - 2\mathbb{E}\left[\widehat{\mathbf{x}}\widehat{\mathbf{x}}^\top\right] + \mathrm{I} \tag{89}$$

$$= \mathbb{E}\left[\widehat{\mathbf{x}}\widehat{\mathbf{x}}^\top\widehat{\mathbf{x}}\widehat{\mathbf{x}}^\top\right] - \mathrm{I} \tag{90}$$

$$= \mathbb{E}\left[\|\widehat{\mathbf{x}}\|^2\widehat{\mathbf{x}}\widehat{\mathbf{x}}^\top\right] - \mathrm{I} \tag{91}$$

$$\leq dK\mathbb{E}\left[\widehat{\mathbf{x}}\widehat{\mathbf{x}}^\top\right] - \mathrm{I} \tag{92}$$

$$= (dK - 1)\,\mathrm{I} \tag{93}$$

Then from Matrix Bernstein, we have

$$\mathbb{E}\left[\text{B}\right] \leq \sqrt{2(dK-1)n\log(2d)} + \frac{1}{3}dK\log(2d) \tag{94}$$

Then we have in expectation

$$\mathbb{E}\left[\sigma_{\min}\left(\sum_{k\in[n]} \mathbf{x}_k\mathbf{x}_k^\top\right)\right] \geq n\sigma_{min}\left(\boldsymbol{\Sigma}\right) - \|\boldsymbol{\Sigma}\|\left(\sqrt{2(dK-1)n\log(2d)} + \frac{1}{3}dK\log(2d)\right) \tag{95}$$

Assume the spectrum of $\boldsymbol{\Sigma}$ is flat, then we have

$$\mathbb{E}\left[\sigma_{\min}\left(\sum_{k\in[n]} \mathbf{x}_k\mathbf{x}_k^\top\right)\right] \geq \sigma_{\min}\left(\boldsymbol{\Sigma}\right)\left(n - \left(\sqrt{2(dK-1)n\log(2d)} + \frac{1}{3}dK\log(2d)\right)\right) \tag{96}$$

Similarly from the Matrix Bernstein Inequality in [36], we have

$$\mathbb{P}\left\{\sigma_{\min}\left(\sum_{k\in[n]}\mathbf{x}_k\mathbf{x}_k^\top\right) > t\right\} \leq 2d\exp\left(\frac{-t^2}{2n(dK-1)+Kt/3}\right) \tag{97}$$

We also like to note if the singular values have linear decay rate th ■

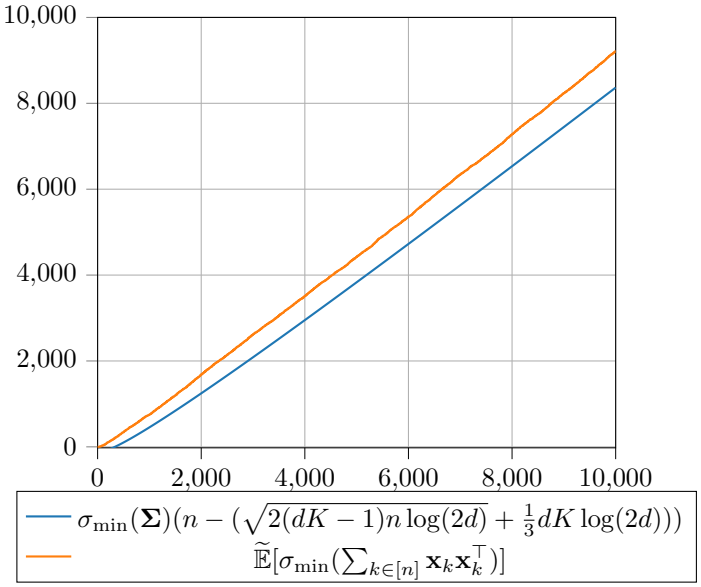

Figure 4: The bound for Lemma 32 compared to an average over 10 trials. We set $\mathbf{\Sigma} = \mathrm{I}$, $K = 4$, $d = 20$.

Lemma 33. *Let $\eta_i \in P \setminus S$ be defined as in Assumption 29, then it follows*

$$\mathbb{E}\left[\sum_{i\in P\setminus S}\eta_i^2\right] \leq eK_3^2 n(1-2\varepsilon) \tag{98}$$

Proof. Note $|P \setminus S| = (1-2\varepsilon)n$, then

$$\mathbb{P}\left\{\sum_{i\in P\setminus S}\eta_i^2 \geq t\right\} = \mathbb{P}\left\{\sum_{i\in P\setminus S}\left(\eta_i^2 - \mathbb{V}\left(\eta_i\right)\right) + \mathbb{V}\left(\eta_i\right) \geq t\right\} \tag{99}$$

$$= \mathbb{P}\left\{\sum_{i\in P\setminus S}\left(\eta_i^2 - \mathbb{E}\left[\eta_i^2\right]\right) \geq t - n(1-2\varepsilon)\mathbb{E}\left[\eta_i^2\right]\right\} \tag{100}$$

$$\overset{\text{Bernstein}}{\leq} \exp\left(-\frac{\left(t-n(1-2\varepsilon)\mathbb{E}\left[\eta_i^2\right]\right)^2}{2\left(\sum_{i\in P\cap S}\mathbb{E}\left[\left(t-n(1-2\varepsilon)\mathbb{E}\left[\eta_i^2\right]\right)^2\right] + \frac{1}{3}Mt\right)}\right) \tag{101}$$

Alternative Proof. We will utilize Chernoff's Inequality.

$$\mathbb{P}\left\{\sum_{i\in P\setminus S}\eta_i^2 \geq t\right\} \overset{\zeta_1}{\leq} \mathbb{P}\left\{\exp\left(\lambda\sum_{i\in P\setminus S}\eta_i^2\right) \geq \exp\left(\lambda t\right)\right\} \tag{102}$$

$$\overset{\text{Markov}}{\leq} \exp\left(-\lambda t\right)\mathbb{E}\left[\exp\left(\lambda\sum_{i\in P\setminus S}\eta_i^2\right)\right] \tag{103}$$

$$= \exp\left(-\lambda t\right) \mathbb{E}\left[\prod_{i \in P \setminus S} \exp\left(\lambda \eta_i^2\right)\right] \tag{104}$$

$$= \exp\left(-\lambda t\right) \prod_{i \in P \setminus S} \mathbb{E}\left[\exp\left(\lambda \eta_i^2\right)\right] \tag{105}$$

$$\overset{\zeta_2}{\leq} \exp\left(-\lambda t\right) \prod_{i \in P \setminus S} \exp\left(\lambda K_3^2\right) \tag{106}$$

$$= \exp\left(-\lambda t + n\left(1 - 2\varepsilon\right) \lambda K_3^2\right) \tag{107}$$

$$\overset{\zeta_3}{=} \exp\left(-\frac{t}{K_3^2 n(1-2\varepsilon)} + 1\right) \tag{108}$$

(ζ_1) holds for any $\lambda > 0$. (ζ_2) holds for any $\lambda \leq K_3$. Then we can calculate the expectation by the following,

$$\mathbb{E}\left[\sum_{i \in P \setminus S} \eta_i^2 \geq t\right] = \int_0^\infty \mathbb{P}\left\{\sum_{i \in P \setminus S} \eta_i^2 \geq t\right\} dt \tag{109}$$

$$\leq \int_0^\infty \exp\left(-\frac{t}{K_3^2 n(1-2\varepsilon)} + 1\right) dt \tag{110}$$

$$= e K_3^2 n(1-2\varepsilon) \tag{111}$$

This concludes the proof. ■

A.2 Polynomial Kernel

The polynomial kernel is given by

$$K(\mathbf{x}_1, \mathbf{x}_2) = \left(\mathbf{x}_1^\top \mathbf{x}_2 + C\right)^p \tag{112}$$

The feature map for the polynomial kernel is given as

$$\phi_{\mathsf{poly}}(\mathbf{x}) = \left[x_1, \ldots, x_d, x_1^2, \ldots, x_d^2, \ldots, x_1^p, \ldots, x_d^p, x_1 x_2, \ldots, x_{d-1} x_d\right] \in \mathbb{R}^{d^p} \tag{113}$$

Lemma 34. *Let $\mathbf{x}_1, \ldots, \mathbf{x}_n$ have sub-Gaussian design described in Assumption 28 and $\eta_i, \ldots, \eta_n$ have sub-Gaussian design described in Assumption 29. It then follows*

$$\mathbb{E}\left\|\sum_{i=1}^n \eta_i \phi_{\mathsf{poly}}\left(\mathbf{x}_i\right)\right\| \leq \Xi \tag{114}$$

Lemma 35. *Let $\mathbf{x}_1, \ldots, \mathbf{x}_n$ have sub-Gaussian design described in Assumption 28 such that $n > d\log(d)$. It then follows*

$$\sigma_{\min}\left(\sum_{i=1}^n \phi_{\mathsf{poly}}(\mathbf{x}_i) \phi_{\mathsf{poly}}(\mathbf{x}_i)^\top\right) \geq \Xi \tag{115}$$

A.3 Gaussian Kernel

The gaussian kernel is given for $\gamma > 0$ by

$$K(\mathbf{x}_1, \mathbf{x}_2) = \exp\left(-\gamma \left\|\mathbf{x}_1 - \mathbf{x}_2\right\|_2^2\right) \tag{116}$$

Lemma 36. *Let $\eta_1, \ldots, \eta_n \sim \mathcal{N}(0, \sigma^2)$ be sub-Gaussian. Then it follows*

$$\mathbb{E}\left|\sum_{i=1}^{n} \eta_i\right| \leq \Xi \tag{117}$$

We will utilize the Random Fourier Features (RFF) representation [29]. Let $\mathbf{w}_1, \ldots, \mathbf{w}_d \sim \mathcal{N}_d(\mathbf{0}, \frac{2}{\gamma}\mathrm{I})$, then the RFF reprsentation is given by

$$\phi_{\mathsf{RFF}}(\mathbf{x}) = \frac{1}{\sqrt{d}}\left[\cos\left(\mathbf{w}_1^\top \mathbf{x}\right), \sin\left(\mathbf{w}_1^\top \mathbf{x}\right), \cdots, \cos\left(\mathbf{w}_d^\top \mathbf{x}\right), \sin\left(\mathbf{w}_d^\top \mathbf{x}\right)\right] \tag{118}$$

The key idea behind the Randomized Fourier Features is

$$\mathbb{E}\left[\phi_{\mathsf{RFF}}(\mathbf{x}_1)^\top \phi_{\mathsf{RFF}}(\mathbf{x}_2)\right] = \exp\left(-\gamma \left\|\mathbf{x}_1 - \mathbf{x}_2\right\|_2^2\right) \tag{119}$$

We will first derive a relation between $\mathbb{E}\left[\phi_{\mathsf{RFF}}(\mathbf{x}_1)^\top \phi_{\mathsf{RFF}}(\mathbf{x}_2)\right]$ and $\mathbb{E}\left[\phi_{\mathsf{RFF}}(\mathbf{x}_1)\right]^\top \mathbb{E}\left[\phi_{\mathsf{RFF}}(\mathbf{x}_2)\right]$.

B Proofs for Section 3

In this section we give the deferred proofs of main results.

B.1 Proof of Lemma 13

Proof. For any $\mathsf{f}_{\mathbf{w}_1}, \mathsf{f}_{\mathbf{w}_2} \in \mathcal{K}$, we want to make sure the gradient is bounded.

$$|g(t, \mathsf{f}_{\mathbf{w}_1}) - g(t, \mathsf{f}_{\mathbf{w}_2})| = \left| \int_0^1 \nabla_{\mathsf{f}} g(t, (1-\lambda)\, \mathsf{f}_{\mathbf{w}_1} + \lambda\, \mathsf{f}_{\mathbf{w}_2})(\mathsf{f}_{\mathbf{w}_1} - \mathsf{f}_{\mathbf{w}_2}) d\lambda \right| \tag{120}$$

$$\leq \|\mathsf{f}_{\mathbf{w}_1} - \mathsf{f}_{\mathbf{w}_2}\|_{\mathcal{H}} \left| \int_0^1 \nabla_{\mathsf{f}} g(t, (1-\lambda)\, \mathsf{f}_{\mathbf{w}_1} + \lambda\, \mathsf{f}_{\mathbf{w}_2}) d\lambda \right| \tag{121}$$

$$\overset{(a)}{\leq} \|\mathsf{f}_{\mathbf{w}_1} - \mathsf{f}_{\mathbf{w}_2}\|_{\mathcal{H}} \max_{\mathbf{w} \in \mathcal{K}} \|\nabla_{\mathsf{f}} g(t, \mathsf{f}_{\mathbf{w}})\|_{\mathcal{H}} \tag{122}$$

In (a), we note that since $\mathcal{K}$ is convex, then by definition as $\mathsf{f}_{\mathbf{w}_1}, \mathsf{f}_{\mathbf{w}_2} \in \mathcal{K}$, we have for $\lambda \in [0,1]$, the convex combination $(1-\lambda)\, \mathsf{f}_{\mathbf{w}_1} + \lambda \mathsf{f}_{\mathbf{w}_2} \in \mathcal{K}$. We use the $\mathcal{H}$ norm of the gradient to bound L from above for an element in the convex closed set $\mathcal{K}$.

$$\|\nabla g\,(t, \mathsf{f}_{\mathbf{w}})\|_{\mathcal{H}} = \left\| \frac{2}{np} \sum_{i=1}^{n} \mathbb{I}\left\{ t \geq (\mathsf{f}_{\mathbf{w}}(\mathbf{x}_i) - y_i)^2 \right\} (\mathsf{f}_{\mathbf{w}}(\mathbf{x}_i) - y_i) \cdot K(\mathbf{x}_i, \cdot) \right\|_{\mathcal{H}} \tag{123}$$

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 $t \geq (\mathsf{f}_{\mathbf{w}}(\mathbf{x}_i) - y_i)^2$.

$$= \left\| \frac{2}{np} \sum_{i=1}^{m} (\mathsf{f}_{\mathbf{w}}(\mathbf{x}_i) - y_i) \cdot K\left(\mathbf{x}_i, \cdot\right) \right\|_{\mathcal{H}} \tag{124}$$

$$\leq \frac{2}{np} \left(\left\| \sum_{i=1}^{m} \mathsf{f}_{\mathbf{w}}(\mathbf{x}_i) \cdot K\left(\mathbf{x}_i, \cdot\right) \right\|_{\mathcal{H}} + \left\| \sum_{i=1}^{m} y_i K\left(\mathbf{x}_i, \cdot\right) \right\|_{\mathcal{H}} \right) \tag{125}$$

$$\leq \frac{2}{np} \left(\left\| \sum_{i=1}^{m} \left(\sum_{j=1}^{n} w_i K(\mathbf{x}_i, \mathbf{x}_j) \right) K(\mathbf{x}_i, \cdot) \right\|_{\mathcal{H}} + \left\| \sum_{i=1}^{m} y_i \right\| \left\| \sum_{i=1}^{m} K(\mathbf{x}_i, \cdot) \right\|_{\mathcal{H}} \right) \tag{126}$$

$$\overset{(a)}{=} \frac{2}{np} \left(\left\| \sum_{i=1}^{m} \left\langle \sum_{j=1}^{n} w_j K(\mathbf{x}_j, \cdot), K(\mathbf{x}_i, \cdot) \right\rangle_{\mathcal{H}} \cdot K(\mathbf{x}_i, \cdot) \right\|_{\mathcal{H}} + \left\| \sum_{i=1}^{m} y_i \right\| \sum_{i=1}^{m} \sqrt{K(\mathbf{x}_i, \mathbf{x}_i)} \right) \tag{127}$$

$$\leq \frac{2}{np} \left(\left| \left\langle \sum_{j=1}^{n} w_j K(\mathbf{x}_j, \cdot), \sum_{i=1}^{m} K(\mathbf{x}_i, \cdot) \right\rangle_{\mathcal{H}} \right| \left\| \sum_{i=1}^{m} K(\mathbf{x}_i, \cdot) \right\|_{\mathcal{H}} + \left\| \sum_{i=1}^{m} y_i \right\| \sum_{i=1}^{m} \sqrt{K(\mathbf{x}_i, \mathbf{x}_i)} \right) \tag{128}$$

$$\leq \frac{2}{np} \left(\left\| \sum_{j=1}^{n} w_j K(\mathbf{x}_j, \cdot) \right\|_{\mathcal{H}} \left(\sum_{i=1}^{m} \sqrt{K(\mathbf{x}_i, \mathbf{x}_i)} \right)^2 + \left\| \sum_{i=1}^{n} y_i \right\| \left(\sum_{i=1}^{m} \sqrt{K(\mathbf{x}_i, \mathbf{x}_i)} \right) \right) \tag{129}$$

$$\leq \frac{2}{np} \left(\|\mathsf{f}_{\mathbf{w}}\|_{\mathcal{H}} \left(\sum_{i=1}^{m} \sqrt{K(\mathbf{x}_i, \mathbf{x}_i)} \right)^2 + \|\mathbf{y}\|_1 \left(\sum_{i=1}^{n} \sqrt{K(\mathbf{x}_i, \mathbf{x}_i)} \right) \right) \tag{130}$$

$$\leq \frac{2R}{np} \left(\sum_{i=1}^{n} \sqrt{K(\mathbf{x}_i, \mathbf{x}_i)} \right)^2 + \frac{2\,\|\mathbf{y}\|_1}{np} \left(\sum_{i=1}^{n} \sqrt{K(\mathbf{x}_i, \mathbf{x}_i)} \right) \tag{131}$$

(a) follows form the reproducing property for RKHS [11]. If we have a normalized kernel such as the Gaussian Kernel, then we have the Lipschitz Constant is finite. Furthermore, if the adversary introduces label corruption that tends to ∞, then these points will not be in the Subquantile as $\mathsf{f}_{\mathbf{w}}$ has bounded norm, so it will have infinite error. This concludes the proof. ■

B.2 Proof of Lemma 15

Proof. We use the $\mathcal{H}$ norm of the gradient to bound L from above. Let S be denoted as the subquantile set. Define the sigmoid function as $\sigma(x) = \frac{1}{1+e^{-x}}$.

$$\|\nabla_{\mathsf{f}} g(t, \mathsf{f}_{\mathbf{w}})\|_{\mathcal{H}} = \left\| \frac{1}{np} \sum_{i=1}^{n} \mathbb{I}\{t \geq (1-y_i)\log(\mathsf{f}_{\mathbf{w}}(\mathbf{x}_i))\} \left(y_i - \sigma\left(\mathsf{f}_{\mathbf{w}}(\mathbf{x}_i)\right)\right) \cdot K(\mathbf{x}_i, \cdot) \right\|_{\mathcal{H}} \tag{132}$$

$$\leq \frac{1}{np} \left\| \sum_{i \in S} \left(y_i - \sigma\left(\mathsf{f}_{\mathbf{w}}(\mathbf{x}_i)\right)\right) \right\| \left\| \sum_{i \in S} K(\mathbf{x}_i, \cdot) \right\|_{\mathcal{H}} \tag{133}$$

$$\overset{(a)}{\leq} \sum_{i \in S} \sqrt{K(\mathbf{x}_i, \mathbf{x}_i)} \tag{134}$$

(a) follows from the fact that $y_i \in \{0, 1\}$ and range$(\sigma) \in [0, 1]$. This completes the proof.

■

B.3 Proof of Lemma 17

Proof. We use the spectral norm of the gradient to bound L from above. Let S be denoted as the subquantile set.

$$\|\nabla_{\mathrm{W}} g(t, \mathrm{W})\|_2 = \tag{135}$$

■

B.4 Proof of Lemma 20

Proof.

Let S be the set containing the points with the minimum error from X w.r.t to the weights vector $\mathbf{w}$. Define $\eta_i \triangleq \left(\mathsf{f}_{\mathbf{w}^*}(\mathbf{x}_i) - y_i\right)$ where $i \in P$.

$$\Phi\left(\mathsf{f}_{\mathbf{w}}\right) - \Phi\left(\mathsf{f}_{\mathbf{w}^*}\right) = \sum_{i \in S} \left(\mathsf{f}_{\mathbf{w}}(\mathbf{x}_i) - y_i\right)^2 - \sum_{j \in P} \left(\mathsf{f}_{\mathbf{w}^*}(\mathbf{x}_j) - y_j\right)^2 \tag{136}$$

$$= \sum_{i \in S \cap P} \left(\mathsf{f}_{\mathbf{w}}(\mathbf{x}_i) - y_i\right)^2 + \sum_{i \in S \cap Q} \left(\mathsf{f}_{\mathbf{w}}(\mathbf{x}_i) - y_i\right)^2 - \sum_{j \in P} \left(\mathsf{f}_{\mathbf{w}^*}(\mathbf{x}_j) - y_j\right)^2 \tag{137}$$

$$\geq \sum_{i \in S \cap P} \left(\mathsf{f}_{\mathbf{w}}(\mathbf{x}_i) - y_i\right)^2 - \sum_{j \in P} \left(\mathsf{f}_{\mathbf{w}^*}(\mathbf{x}_j) - y_j\right)^2 \tag{138}$$

$$= \sum_{i \in S \cap P} \left(\mathsf{f}_{\mathbf{w}}(\mathbf{x}_i) - \mathsf{f}_{\mathbf{w}^*}(\mathbf{x}_i) - \eta_i\right)^2 - \sum_{j \in P} \eta_j^2 \tag{139}$$

$$= \sum_{i \in S \cap P} \left(\left(\mathsf{f}_{\mathbf{w}} - \mathsf{f}_{\mathbf{w}^*}\right)(\mathbf{x}_i) - \eta_i\right)^2 - \sum_{j \in P} \eta_j^2 \tag{140}$$

$$\geq \sum_{i \in S \cap P} \underbrace{\left(\left(\mathsf{f}_{\mathbf{w}} - \mathsf{f}_{\mathbf{w}^*}\right)(\mathbf{x}_i)\right)^2}_{\mathrm{A}_1} - 2 \underbrace{\sum_{i \in S \cap P} \eta_i \left(\mathsf{f}_{\mathbf{w}} - \mathsf{f}_{\mathbf{w}^*}\right)(\mathbf{x}_i)}_{\mathrm{A}_2} - \underbrace{\sum_{j \in P \setminus S} \eta_j^2}_{\mathrm{A}_3} \tag{141}$$

Now we will upper bound A_1.

$$\sum_{i \in S \cap P} \left(\left(\mathsf{f}_{\mathbf{w}} - \mathsf{f}_{\mathbf{w}^*}\right)(\mathbf{x}_i)\right)^2 \overset{(a)}{=} \sum_{i \in S \cap P} \left\langle \sum_{j \in X} \left(w_j - w_j^*\right) K(\mathbf{x}_j, \cdot), K(\mathbf{x}_i, \cdot) \right\rangle_{\mathcal{H}}^2 \tag{142}$$

$$= \sum_{i \in S \cap P} \left\langle \sum_{j \in X} \left(w_j - w_j^*\right) \phi(\mathbf{x}_j), \phi(\mathbf{x}_i) \right\rangle \left\langle \phi(\mathbf{x}_i), \sum_{j \in X} \left(w_j - w_j^*\right) \phi(\mathbf{x}_j) \right\rangle \tag{143}$$

$$\geq \|\mathsf{f}_{\mathbf{w}} - \mathsf{f}_{\mathbf{w}^*}\|_{\mathcal{H}}^2 \underbrace{\sigma_{\min}\left(\sum_{i\in S\cap P}\phi(\mathbf{x}_i)\phi(\mathbf{x}_i)^\top\right)}_{B_1} \tag{144}$$

Next we will upper ound A_2,

$$\begin{aligned}
\mathrm{A}_2 &= \sum_{i\in S\cap P}\eta_i\left(\mathsf{f}_{\mathbf{w}} - \mathsf{f}_{\mathbf{w}^*}\right)(\mathbf{x}_i) && (145)\\
&= \sum_{i\in S\cap P}\left\langle\sum_{j\in X}(w_j - w_j^*)K(\mathbf{x}_j,\cdot), \eta_i K(\mathbf{x}_i,\cdot)\right\rangle_{\mathcal{H}} && (146)\\
&= \left\langle\sum_{j\in X}(w_j - w_j^*)K(\mathbf{x}_j,\cdot), \sum_{i\in S\cap P}\eta_i K(\mathbf{x}_i,\cdot)\right\rangle_{\mathcal{H}} && (147)\\
&\leq \|\mathsf{f}_{\mathbf{w}} - \mathsf{f}_{\mathbf{w}^*}\|_{\mathcal{H}}\left\|\sum_{i\in S\cap P}\eta_i K(\mathbf{x}_i,\cdot)\right\|_{\mathcal{H}} = \|\mathsf{f}_{\mathbf{w}} - \mathsf{f}_{\mathbf{w}^*}\|_{\mathcal{H}}\left\|\sum_{i\in S\cap P}\eta_i\phi(\mathbf{x}_i)\right\|_2 && (148)
\end{aligned}$$

We will now lower bound B_1. For the linear kernel, for B_1 to be greater than 0 , than if $\mathbf{x}\in\mathbb{R}^d$, then we must have $\frac{d}{1-2\varepsilon} < n$, otherwise we will have a rank-deficient matrix which will thus have singular values of value 0. For the polynomial kernel, for B_1 to be greater than 0, than $n > d^r$ where r is the polynomial degree. We thus have

$$\Phi\left(\mathsf{f}_{\mathbf{w}}\right) - \Phi\left(\mathsf{f}_{\mathbf{w}^*}\right) \geq \eta^2\sigma_{\min}\left(\sum_{i\in S\cap P}\phi(\mathbf{x}_i)\phi(\mathbf{x}_i)^\top\right) - 2\eta\left\|\sum_{i\in S\cap P}\eta_i\phi\left(\mathbf{x}_i\right)\right\| - \sum_{j\in P\setminus S}\eta_j^2 \tag{149}$$

This completes the proof. ■

B.5 Proof of Theorem 21

Proof. First,

$$\left\|\nabla\Phi_\lambda\left(\mathsf{f}_{\mathbf{w}}\right)\right\|_{\mathcal{H}} = \left\|\frac{1}{\lambda}\left(\mathsf{f}_{\mathbf{w}} - \underset{\mathsf{f}_{\widehat{\mathbf{w}}}\in\mathcal{K}}{\arg\min}\left(\Phi\left(\mathsf{f}_{\widehat{\mathbf{w}}}\right) + \frac{1}{2\lambda}\left\|\mathsf{f}_{\mathbf{w}} - \mathsf{f}_{\widehat{\mathbf{w}}}\right\|_{\mathcal{H}}^2\right)\right)\right\|_{\mathcal{H}} = 0 \tag{150}$$

This implies for any $\widetilde{\mathbf{w}}\in\mathcal{K}$, it follows

$$\Phi\left(\mathsf{f}_{\widehat{\mathbf{w}}}\right) < \Phi\left(\mathsf{f}_{\widetilde{\mathbf{w}}}\right) + \frac{1}{2\lambda}\left\|\mathsf{f}_{\widetilde{\mathbf{w}}} - \mathsf{f}_{\widehat{\mathbf{w}}}\right\|_{\mathcal{H}}^2 \tag{151}$$

For any $\mathsf{f}_{\widehat{\mathbf{w}}}$ satisfying above, then the distance from the optimal must be low. Let $\widetilde{\mathbf{w}} = \mathbf{w}^*$, then we have

$$\Phi\left(\mathsf{f}_{\widehat{\mathbf{w}}}\right) - \Phi\left(\mathsf{f}_{\mathbf{w}^*}\right) \leq \frac{1}{2\lambda}\left\|\mathsf{f}_{\widehat{\mathbf{w}}} - \mathsf{f}_{\mathbf{w}^*}\right\|_{\mathcal{H}}^2 \tag{152}$$

We proceed by proof by contradiction. Assume $\|\mathsf{f}_{\widehat{\mathbf{w}}} - \mathbf{w}^*\| > \eta$, then if $\Phi(\mathsf{f}_{\widehat{\mathbf{w}}}) - \Phi(\mathbf{w}^*) > \frac{1}{2}\eta^2$, then we will have $\mathsf{f}_{\widehat{\mathbf{w}}}$ is not a stationary point, which will imply $\|\mathsf{f}_{\widehat{\mathbf{w}}} - \mathbf{w}^*\|_{\mathcal{H}} \leq \eta$. Therefore, we attempt to find the minimum value for η. From Lemma 20, we have we have

$$\Phi\left(\mathbf{w}\right) - \Phi\left(\mathbf{w}^*\right) \geq \eta^2\sigma_{\min}\left(\sum_{i\in S\cap P}\phi(\mathbf{x}_i)\phi(\mathbf{x}_i)^\top\right) - 2\eta\left\|\sum_{i\in S\cap P}\eta_i\phi\left(\mathbf{x}_i\right)\right\| - \sum_{j\in P\setminus S}\eta_j^2 \tag{153}$$

From the definition of stationary point, we have

$$\eta^2\left(\sigma_{\min}\left(\sum_{i\in S\cap P}\phi(\mathbf{x}_i)\phi(\mathbf{x}_i)^\top\right) - \frac{1}{2}L\right) - 2\eta\left\|\sum_{i\in S\cap P}\eta_i\phi\left(\mathbf{x}_i\right)\right\| - \sum_{j\in P\setminus S}\eta_j^2 \geq 0 \tag{154}$$

From lower bounding the positive solution of the quadratric equation, we have

$$\eta \geq \Omega\left(\frac{2L\left(\sum_{j\in P\setminus S}\eta_j^2\right)}{2L\sigma_{\min}\left(\sum_{i\in S\cap P}\phi(\mathbf{x}_i)\phi(\mathbf{x}_i)^\top\right)-1}\right)^{1/2}+\frac{4L\left\|\sum_{i\in S\cap P}\eta_i\phi(\mathbf{x}_i)\right\|}{\sigma_{\min}\left(\sum_{i\in S\cap P}\phi(\mathbf{x}_i)\phi(\mathbf{x}_i)^\top\right)} \tag{155}$$

Therefore, when **??** holds, we have a contradiction. It thus follows

$$\eta \leq \left(\frac{2L\left(\sum_{j\in P\setminus S}\eta_j^2\right)}{2L\sigma_{\min}\left(\sum_{i\in S\cap P}\phi(\mathbf{x}_i)\phi(\mathbf{x}_i)^\top\right)-1}\right)^{1/2}+\frac{4L\left\|\sum_{i\in S\cap P}\eta_i\phi(\mathbf{x}_i)\right\|}{\sigma_{\min}\left(\sum_{i\in S\cap P}\phi(\mathbf{x}_i)\phi(\mathbf{x}_i)^\top\right)} \tag{156}$$

This completes the proof. ■

C Proofs for Section 4

In this section we give the optimization results from § section 4.

C.1 Proof of Theorem 26

We will start with the definition of the Moreau Envelope.

$$\bar{\Phi}_\lambda(\mathsf{f}_{\mathbf{w}}) = \frac{1}{\lambda}\left(\mathsf{f}_{\mathbf{w}} - \underset{\mathsf{f}_{\widehat{\mathbf{w}}}\in\mathcal{K}}{\arg\min}\left\{\Phi(\mathsf{f}_{\widehat{\mathbf{w}}}) + \frac{1}{2\lambda}\left\|\mathsf{f}_{\mathbf{w}} - \mathsf{f}_{\widehat{\mathbf{w}}}\right\|_{\mathcal{H}}^2\right\}\right) \tag{157}$$

Note $g(t, \mathsf{f}_{\mathbf{w}})$ is L-lipschitz in $\mathsf{f}_{\mathbf{w}}$. Let $\mathsf{f}_{\widehat{\mathbf{w}}}^{(t)} = \arg\min_{\mathsf{f}_{\widetilde{\mathbf{w}}}\in\mathcal{K}}\{\bar{\Phi}(\mathsf{f}_{\mathbf{w}}) + \|\mathsf{f}_{\widetilde{\mathbf{w}}} - \mathsf{f}_{\mathbf{w}}\|_{\mathcal{H}}^2\}$. Then we have,

$$\bar{\Phi}_\lambda\left(\mathsf{f}_{\mathbf{w}}^{(t+1)}\right) \leq \Phi\left(\mathsf{f}_{\widehat{\mathbf{w}}}^{(t)}\right) + \left\|\mathsf{f}_{\mathbf{w}}^{(t+1)} - \mathsf{f}_{\widehat{\mathbf{w}}}^{(t)}\right\|_{\mathcal{H}}^2 \tag{158}$$

$$= \Phi\left(\mathsf{f}_{\widehat{\mathbf{w}}}^{(t)}\right) + \left\|\mathsf{f}_{\widehat{\mathbf{w}}}^{(t)} - \Pi_{\mathcal{K}}\left(\mathsf{f}_{\mathbf{w}}^{(t)} - \alpha\left(\mu\mathbf{b}^{(t)}\right)\right)\right\|_{\mathcal{H}}^2 \tag{159}$$

$$= \Phi\left(\mathsf{f}_{\widehat{\mathbf{w}}}^{(t)}\right) + \left\|\mathsf{f}_{\widehat{\mathbf{w}}}^{(t)} - \Pi_{\mathcal{K}}\left(\mathsf{f}_{\mathbf{w}}^{(t)} - \alpha\left(\mu\left(\mu\mathbf{b}^{(t-1)} + \nabla_{\mathsf{f}}\Phi\left(\mathsf{f}_{\mathbf{w}}^{(t-1)}\right)\right)\right)\right)\right\|_{\mathcal{H}}^2 \tag{160}$$

$$= \Phi\left(\mathsf{f}_{\widehat{\mathbf{w}}}^{(t)}\right) + \left\|\mathsf{f}_{\widehat{\mathbf{w}}}^{(t)} - \Pi_{\mathcal{K}}\left(\mathsf{f}_{\mathbf{w}}^{(t)} - \alpha\left(\sum_{i=k}^{t}\mu^k(1-\mu)\nabla_{\mathsf{f}}\,\mathsf{f}_{\mathbf{w}}^{(t-k)}\right)\right)\right\|_{\mathcal{H}}^2 \tag{161}$$

$$\leq \Phi\left(\mathsf{f}_{\widehat{\mathbf{w}}}^{(t)}\right) + \left\|\mathsf{f}_{\widehat{\mathbf{w}}}^{(t)} - \mathsf{f}_{\mathbf{w}}^{(t)} - \alpha\left(\sum_{k=1}^{t}\mu^k(1-\mu)\nabla_{\mathsf{f}}\,\mathsf{f}_{\mathbf{w}}^{(t-k)}\right)\right\|_{\mathcal{H}}^2 \tag{162}$$

C.2 Proof of Theorem 27

D Necessary Lemmas

Lemma 37 (MGF of Sub-Exponential Random Variable)**.** *Let x be a Sub-exponetial variable, then we have*

$$\mathbb{E}\left[\exp\left(\lambda x\right)\right] \leq \exp\left(C\lambda^2\right) \tag{163}$$

Theorem 38 (Matrix Chernoff, [35])**.** *Let X_k be a sequence of independent, random, self-adjoint matrices with dimension d s.t.*

$$\mathrm{X}_k \succcurlyeq \mathbf{0} \quad \textit{and} \quad \lambda_{\max}\left(\mathrm{X}_k\right) \leq R \quad \textit{almost surely} \tag{164}$$

Define

$$\mu_{\min} \triangleq \lambda_{\min}\left(\sum_k \mathbb{E}\mathrm{X}_k\right) \tag{165}$$

Then for $\delta \in [0,1]$

$$\mathbb{P}\left\{\lambda_{\min}\left(\sum_k \mathrm{X}_k\right) \geq (1-\delta)\mu_{\min}\right\} \leq d \cdot \left[\frac{e^{-\delta}}{(1-\delta)^{1-\delta}}\right]^{\mu_{\min}/R} \tag{166}$$

Theorem 39 (Matrix Bernstein, [36])**.** *Let X_k be a sequence of independent, random, self-adjoing matrices with dimension d s.t.*

$$\mathbb{E}\mathrm{X}_k = \mathbf{0} \tag{167}$$

Theorem 40 (Vector Bernstein, [12, 18])**.** *Let $\mathbf{x}_k$ be a sequence of independent, random vectors such that*

$$\mathbb{E}\mathbf{x}_k = \mathbf{0} \quad \forall k \tag{168}$$

E Experimental Details

In this section we give details on datasets and hyperparameters.

E.1 Kernel Regression

Our datasets are synthetic and are sourced from [7]

Dataset	**Dimension** d	**Sample Size** n	**Source**
`Polynomial`	3	1000	Ours
`Boston Housing`	13	506	[7]
`Concrete Data`	8	1030	[7]
`Wine Quality`	11	1599	[7]

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.2 Kernel Binary Classification

Dataset	**Dimension** d	**Sample Size** n	**Source**
`Heart Disease`	13	303	[14]
`Breast Cancer`	32	569	Kaggle

Table 5: Datasets for Kernel Binary Classification.

E.3 Kernel Multi-Class Classification

Dataset	**Dimension** d	**Sample Size** n	**Source**

Table 6: Datasets for Kernel Multi-Class Classification.

E.4 Linear Regression

Dataset	**Dimension** d	**Sample Size** n	**Source**
`Boston Housing`	14	506	Kaggle
`Wine Quality`	11	1599	[7]
`Concrete`	8	1030	[7]
`Drug`			

Table 7: Datasets for Linear Regression.

F Detailed Related Works

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

F.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

$$\left(\nabla_w F(w,u)\right)^\top (\widetilde{w} - w) \geq 0 \quad \forall \widetilde{w} \in K \tag{169}$$

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.

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

First we will give the algorithm

$$S^{(t)} = \underset{T \subset S^{(0)}: |T| = (1-2\epsilon)n}{\arg\min} \sum_{i \in T} -\log f(y_i | \langle \boldsymbol{\beta}^{(t)}, \mathbf{x}_i \rangle) \tag{170}$$

$$\boldsymbol{\beta}^{(t+1)} = \underset{\beta, \|\boldsymbol{\beta}\| \leq R}{\arg\min} \sum_{i \in S^{(t)}} -\log f(y_i | \langle \boldsymbol{\beta}^{(t)}, \mathbf{x}_i \rangle) \tag{171}$$

In Equation (170), the algorithm chooses the $(1-2\epsilon)n$ points giving the least error and put this in the set $S^{(t)}$. Next, in Equation (171), 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].

$$\frac{1}{n} \sum_{i \in S} \nabla_{\boldsymbol{\beta}} \log f(y_i | \langle \boldsymbol{\beta}, \mathbf{x}_i \rangle)^\top \frac{(\boldsymbol{\beta}^\star - \boldsymbol{\beta})}{\|\boldsymbol{\beta}^\star - \boldsymbol{\beta}\|} \leq \gamma \tag{172}$$

We see Equation (172) is an upper bound, instead of a lower bound, of Equation (169). 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.