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

Arvind Rathnashyam
RPI Math and CS, rathna@rpi.edu

Alex Gittens
RPI CS, gittea@rpi.edu

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 [23, 37]. 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 $\inf_f \mathbb{E}_{x\in\mathbb{P}}\left[\ell_f(x)\right]$, 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.

[*] Preliminary Work

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), [7, 1, 24, 29, 11] as well as for linear regression [2, 27]. Robust Statistics has been studied extensively [8] for problems such as high-dimensional mean estimation [31, 3] and Robust Covariance Estimation [4, 10]. Recently, there has been an interest in solving robust machine learning problems by gradient descent [32, 7]. Subquantile minimization aims to address the shortcomings of standard ERM in applications of noisy/corrupted data [21, 19]. In many real-world applications, linear models are insufficient to model the data. Therefore, we consider the problem of Robust Learning for Kernel Learning.

Definition 1. (**Huber ϵ-Contamination Model** [17]). Given a corruption parameter $0 < \epsilon < 0.5$, a data matrix, X and labels 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 y as training data.

Current approaches for robust learning across various machine learning tasks often use gradient descent over a robust objective, [24]. These robust objectives tend to not be convex and therefore do not have a strong analysis on the error bounds for general classes of models.
We similarly propose a robust objective which has a nonconvex-concave objective. This objective has also been proposed recently in [16] where there has been an analysis in the Binary Classification Task. We show Subquantile Minimization reduces to the same objective in [16]. We use theory from the weakly-convex concave optimization literature for our error bounds. We are able to levarage this theory by analyzing the asymptotic distribution of a softplus approximation of the Subquantile objective.

Theorem 2. *(Informal). Let the dataset be given as $\{(x_i, y_i)\}_{i=1}^n$ such that the labels and features of εn samples are arbitrarily corrupted by an adversary. Let* K *be the kernel matrix and S be the points with the lowest error w.r.t $f_{\widehat{w}}$, then Subquantile Minimization returns $f_{\widehat{w}}$ for $n \geq \frac{(1-2\varepsilon)(C_k\|\Sigma\|_{\text{op}}+\beta)}{(1-c_1)\lambda_{\min}(\Sigma)} + \sqrt{\beta}$ for a constant $c_1 \in (0, 1)$ such that for*
Kernelized Regression:

$$\|f_{\widehat{w}} - f_{w^*}\|_{\mathcal{H}} \leq O\left(\sqrt{\frac{\varepsilon}{1-2\varepsilon}}\frac{\sigma}{\sqrt{\lambda_{\min}(\Sigma)}}\right) \tag{1}$$

where $\epsilon \to 0$ as number of gradient descenter iterations goes to ∞ and $\Sigma = \mathbb{E}[\phi(x) \otimes \phi(x)]$.
Kernel Binary Classification:

$$\|f_{\widehat{w}} - f_{w^*}\|_{\mathcal{H}} \leq O\left(\frac{\sqrt{2}L}{\beta}\right) \tag{2}$$

Kernel Multi-Class Classification:

$$\big\|f_{\widehat{\mathrm{W}}} - f_{\mathrm{W}^*}\big\|_{\mathcal{H}} \leq O\left(\frac{\sqrt{2}L}{\beta}\right) \tag{3}$$

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 provide rigorous error bounds for subquantile minimization in the kernel regression, kernel binary classification, and kernel multi-class classification tasks for the linear, polynomial, and gaussian kernel.

3. We give new bounds for accelerated gradient methods for accelerated gradient methods in nonconvex-concave minimax opimization.

4. We perform experiments on state-of-the-art matrices in kernel learning and show the effectiveness of our algorithm compared to other robust algorithms.

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

[7] 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.

[24] 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 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.

[16] proposed learning over the bottom k losses, this is an alternative formulation of our algorithm. They solve their optimization problem with difference of sums convex solvers. This work considers only the binary classification problem.

[15] proposed learning over the middle k losses, this is an extension of previous work [16]. Similarly, this work considers only the problem of binary classification and gives a generalization bound on the training error and out of sample error.

Minimax Optimization

[20] studied minimax optimization in the non-convex non-concave setting. Furthermore, they study convergence of alternating minimizing-maximizing algorithm with a maximizing oracle. Their key result is the convergence of the algorithm with infinite iterations.

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

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.

$$\mathcal{Q}_p(U) \leq t \iff \mathbb{P}\left\{U \leq t\right\} \geq p \tag{4}$$

The p-subquantile of the risk is then given by

$$\mathbb{L}_p\left(U\right) = \frac{1}{p}\int_0^p \mathcal{Q}_p\left(U\right)dq = \mathbb{E}\left[U | U \leq \mathcal{Q}_p\left(U\right)\right] = \max_{t\in\mathbb{R}}\left\{t - \frac{1}{p}\mathbb{E}\left(t-U\right)^+\right\} \tag{5}$$

Given an objective function, ℓ, the kernelized learning poblem becomes:

$$f_{\widehat{w}} = \arg\min_{f_w \in \mathcal{K}} \max_{t\in\mathbb{R}} \left\{ g(t, f_w) \triangleq t - \sum_{i=1}^{n} \left(t - (f_w(x_i) - y_i)^2\right)^+ \right\} \tag{6}$$

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 w. Thus, to solve this problem we use the iterations from equation 11 in [34]. Let $\Pi_{\mathcal{K}}$ be the projection of a vector on to the convex set $\mathcal{K} \triangleq \{f \in \mathcal{H} : \|f\|_{\mathcal{H}} \leq R\}$, then our update steps are

$$t^{(k+1)} = \arg\max_{t\in\mathbb{R}} g(f_w^{(k)}, t) \tag{7}$$

$$f_w^{(k+1)} = \Pi_{\mathcal{K}}\left(f_w^{(k)} - \alpha\nabla_f g(f_w^{(k)}, t^{(k+1)})\right) \tag{8}$$

Claim 3. *The function $g(t, f_w)$ defined in Equation (6) is non-convex-concave, i.e. it is not convex with respect to f_w and is concave with respect to t. Furthermore, $g(t, f_w)$ is ρ-weakly convex where ρ is the β-smoothness factor of $g(t, f_w)$ w.r.t f_w.*

Proof. We will show $-g(t, f_w)$ is convex with respect to t. Let $\nu_i \triangleq (f_w(x_i) - y_i)^2$.

$$-g\left(\lambda t_1 + (1-\lambda)t_2, f_w\right) = -\lambda t_1 - (1-\lambda)t_2 + \sum_{i=1}^{n}\left(\lambda t_1 + (1-\lambda)t_2 - \nu_i\right)^+ \tag{9}$$

$$\leq -\lambda t_1 - (1-\lambda)t_2 + \sum_{i=1}^{n}\lambda(t_1-\nu_i)^+ + (1-\lambda)(t_2-\nu_i)^+ \tag{10}$$

$$= -\lambda g(t_1, f_w) - (1-\lambda)g(t_2, f_w) \tag{11}$$

Therefore we have $g(t, f_w)$ is concave in t. Next we will prove $g(t, f_w)$ is not convex in f_w. ■

Claim 4. *The function $g(t, f_w)$ defined in Equation (6) is L-weakly convex in f_w, where L is lipschitz constant of the gradient of $g(t, f_w)$ w.r.t f_w. This is true for Conditional Value at Risk, [35].*

Proof. Here we note that if we add $\max_{i\in[n]}|k(x_i, x_i) + f_w(x_i) - y_i|$ 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 $g(f_w, t) + \frac{L}{2}\|f_w\|_{\mathcal{H}}^2$. ■

We provide an algorithm for Subquantile Minimization of the ridge regression and classification kernel learning algorithm. Algorithm 1 is applicable to both kernel ridge regression and kernel classication.

```
Algorithm 1: SUBQ-GRADIENT
  Input: Iterations: $T$; Quantile: $p$; Data
         Matrix: X, $(n \times d), n \gg d$; Learning
         schedule: $\alpha_1, \cdots, \alpha_T$; Ridge
         parameter: $\lambda$
  Output: Trained Parameters, $w_{(T)}$
  $w_{(0)} \leftarrow \mathcal{N}_d(0, \sigma)$
  for $k \in 1, 2, \ldots, T$ do
      $\mathrm{S}_{(k)} \leftarrow$ SUBQUANTILE($w^{(k)}$, X)
      $w^{(k+1)} \leftarrow w^{(k)} - \alpha_{(k)}\nabla_w g\left(t^{(k+1)}, w^{(k)}\right)$
  end
  return $w_{(T)}$
```

```
Algorithm 2: SUBQUANTILE
  Input: Parameters $w$, Data Matrix:
         X, $(n \times d)$, Convex Loss Function $f$
  Output: Subquantile Matrix $S$
  $\hat{\nu}_i \leftarrow \ell(x_i; f_w, y_i)$ s.t. $\hat{\nu}_{i-1} \leq \hat{\nu}_i \leq \hat{\nu}_{i+1}$
  $t \leftarrow \hat{\nu}_{np}$
  Let $x_1, ..., x_{np}$ be $np$ points such that
      $\ell(x_i; f_w, y_i) \leq t$
  $\mathrm{S} \leftarrow \begin{pmatrix} x_1^\top & \ldots & x_{np}^\top \end{pmatrix}^\top$
  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 (7) and (8).

Lemma 5. *Let $f(x; w)$ be a convex loss function. Let $x_1, x_2, \cdots, x_n$ denote the n data points ordered such that $f(x_1; w, y_1) \leq f(x_2; w, y_2) \leq \cdots \leq f(x_n; w, y_n)$. If we denote $\hat{\nu}_i \triangleq f(x_i; w, y_i)$, it then follows* $\arg\max_{t\in\mathbb{R}} g(t, w) = \hat{\nu}_{np}$.

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.

$$\begin{aligned}
\partial(-g(t, f_w)) &= \partial\left(-t + \frac{1}{np}\sum_{i=1}^{n}(t - \hat{\nu}_i)^+\right) && (12)\\
&= -1 + \frac{1}{np}\sum_{i=1}^{np}\begin{cases}1 & \text{if } t > \hat{\nu}_i\\ 0 & \text{if } t < \hat{\nu}_i\\ [0,1] & \text{if } t = \hat{\nu}_i\end{cases} && (13)\\
&= 0 \text{ when } t = \hat{\nu}_{np} && (14)
\end{aligned}$$

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

It therefore follows,

$$\sum_{i=1}^{n}\mathbb{I}\left\{\hat{\nu}_{np} \geq \left(f_w^{(k)}(x_i) - y_i\right)^2\right\}\left(f_w^{(k)}(x_i) - y_i\right)^2 \in \max_{t\in\mathbb{R}} g(t, f_w^{(k)}) \tag{15}$$

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 $\hat{\nu}_i \triangleq f(x_i; w, y_i)$ s.t. $\hat{\nu}_{i-1} \leq \hat{\nu}_i \leq \hat{\nu}_{i+1}$, if we choose $t^{(k+1)} = \hat{\nu}_{np}$ as by Lemma 5, it then follows $\nabla_w g(t^{(k)}, f_w^{(k)}) = \frac{1}{np}\sum_{i=1}^{np}\nabla f(x_i; f_w^{(k)}, y_i)$*

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

$$\begin{aligned}
\nabla_f g(t^{(k+1)}, f_w^{(k)}) &= \nabla_f\left(\hat{\nu}_{np} - \frac{1}{np}\sum_{i=1}^{n}\left(\hat{\nu}_{np} - \ell(x_i; f_w^{(k)}, y_i)\right)^+\right) && (16)\\
&= -\frac{1}{np}\sum_{i=1}^{np}\nabla_f\left(\hat{\nu}_{np} - \ell(x_i; f_w^{(k)}, y_i)\right)^+ && (17)\\
&= \frac{1}{np}\sum_{i=1}^{n}\nabla_f \ell(x_i; f_w^{(k)}, y_i)\begin{cases}1 & \text{if } t > \hat{\nu}_i\\ 0 & \text{if } t < \hat{\nu}_i\\ [0,1] & \text{if } t = \hat{\nu}_i\end{cases} && (18)
\end{aligned}$$

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

$$\nabla_f g(t^{(k+1)}, f_w^{(k)}) = \frac{1}{np}\sum_{i=1}^{np}\nabla_f \ell(x_i; f_w^{(k)}, y_i) \tag{19}$$

This concludes the proof. ■

We denote the matrix K as the Gram Matrix where $[\mathrm{K}]_{ij} = k(x_i, x_j) \triangleq \exp(-\rho\|x_i - x_j\|_2^2)$. Given a parameter set w, the prediction for a new point will be: $f(x^\star; w) = \sum_{i=1}^{n} w_i \kappa(x_i, x^\star)$

From our definition of $S^{(k)}$ in **??**, we are interested in as $k \to \infty$ the quantities: $|x \in S^{(k)} \cap P|$ and $|x \in S^{(k)} \cap Q|$, 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. The main ideas is to *first* approximate ReLU, consider the theory with respect to the approximation, and then take the limit as the approximation goes to the ReLU. The softplus approximation is given as follows,

$$\zeta_\lambda(x) = \frac{1}{\lambda}\log\left(1 + e^{\lambda x}\right) \tag{20}$$

We then have the approximation of g as

$$\widetilde{g}_\lambda(t, f_w) \triangleq t - \sum_{i=1}^{n} \zeta_\lambda \left(t - \ell\left(f_w; x_i, y_i\right)\right) \tag{21}$$

$$= t - \frac{1}{np} \sum_{i=1}^{n} \frac{1}{\lambda} \log\left(1 + \exp\left(\lambda\left(t - \ell\left(f_w; x_i, y_i\right)\right)\right)\right) \tag{22}$$

Now we compute the derivatives w.r.t to the softplus approximation, and then we consider the limit of the derivative as $\lambda \to \infty$.

$$\nabla_t \widetilde{g}_\lambda(t, f_w) = \nabla_t \left(t - \frac{1}{np} \sum_{i=1}^{n} \frac{1}{\lambda} \ln\left(1 + \exp\left(\lambda\left(t - \ell(f_w; x_i, y_i)\right)\right)\right)\right) \tag{23}$$

$$= 1 - \frac{1}{np} \sum_{i=1}^{n} \sigma\left(\lambda\left(t - \ell(f_w; x_i, y_i)\right)\right) \tag{24}$$

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

$$\lim_{\lambda\to\infty} \nabla_t \widetilde{g}_\lambda(t, f_w) = 1 - \frac{1}{np} \sum_{i=1}^{n} \mathbb{I}\left\{t - \ell\left(f_w; x_i, y_i\right)\right\} \tag{25}$$

$$\nabla_f \widetilde{g}_\lambda(t, f_w) = \nabla_f \left(t - \frac{1}{np} \sum_{i=1}^{n} \frac{1}{\lambda} \ln\left(1 + \exp\left(\lambda\left(t - \ell(f_w; x_i, y_i)\right)\right)\right)\right) \tag{26}$$

$$= \frac{1}{np} \sum_{i=1}^{n} \nabla_f \ell\left(f_w; x_i, y_i\right) \sigma\left(\lambda\left(t - \ell(f_w; x_i, y_i)\right)\right) \tag{27}$$

We therefore similarly have,

$$\lim_{\lambda\to\infty} \nabla_f \widetilde{g}_\lambda\left(t, f_w\right) = \frac{1}{np} \sum_{i=1}^{n} \mathbb{I}\left\{t - \ell\left(f_w; x_i, y_i\right)\right\} \nabla_f \ell\left(f_w; x_i, y_i\right) \tag{28}$$

Then the second derivative is given by

$$\nabla_f^2 \widetilde{g}_\lambda(t, f_w) \ = \nabla_f \left(\frac{1}{np} \sum_{i=1}^{n} \nabla_f \ell\left(f_w; x_i, y_i\right) \sigma\left(\lambda\left(t - \ell(f_w; x_i, y_i)\right)\right)\right) \tag{29}$$

$$= \frac{1}{np} \sum_{i=1}^{n} \Bigg(\nabla_f^2 \ell\left(f_w; x_i, y_i\right) \sigma\left(\lambda\left(t - \ell(f_w; x_i, y_i)\right)\right)$$

$$- \left(\nabla_f \ell\left(f_w; x_i, y_i\right)\right)^2 \sigma\left(\lambda\left(t - \ell(f_w; x_i, y_i)\right)\right)\left(1 - \sigma\left(\lambda\left(t - \ell(f_w; x_i, y_i)\right)\right)\right)\Bigg) \tag{30}$$

We then similarly have,

$$\lim_{\lambda\to\infty} \nabla_f^2 \widetilde{g}_\lambda(t, f_w) = \frac{1}{np} \sum_{i=1}^{n} \mathbb{I}\left\{t - \ell\left(f_w; x_i, y_i\right)\right\} \nabla_f^2 \ell\left(f_w; x_i, y_i\right) \tag{31}$$

We can then calculate the Lipschitz constant of the approximation function with respect to f_w.

Lemma 8 (Lipschitz continuous gradient)**.** *Let $f_w, f_{\widetilde{w}} \in \mathcal{K}$, then we have*

$$\lim_{\lambda\to\infty} \left|\nabla_f \widetilde{g}_\lambda(t, f_w) - \nabla_f \widetilde{g}_\lambda(t, f_{\widetilde{w}})\right| \leq \beta \left\|f_w - f_{\widetilde{w}}\right\|_{\mathcal{H}} \tag{32}$$

where

$$\beta = \frac{1}{np} \sum_{i=1}^{n} \left| \nabla_f^2 \ell \left(f_w; x_i, y_i \right) \right| \tag{33}$$

Proof. We will upper bound the second derivative.

$$\lim_{\lambda \to \infty} \left| \nabla_f \widetilde{g}_\lambda(t, f_w) - \nabla_f \widetilde{g}_\lambda(t, f_{\widetilde{w}}) \right| \leq \lim_{\lambda \to \infty} \sup \left\{ \nabla_f^2 \widetilde{g}_\lambda(t, f_w) \right\} \left\| f_w - f_{\widetilde{w}} \right\|_{\mathcal{H}} \tag{34}$$

$$\leq \lim_{\lambda \to \infty} \sup \left\{ \frac{1}{np} \sum_{i=1}^{n} \nabla_f^2 \ell \left(f_w; x_i, y_i \right) \sigma \left(\lambda \left(t - \ell(f_w; x_i, y_i) \right) \right) \right\} \left\| f_w - f_{\widetilde{w}} \right\|_{\mathcal{H}} \tag{35}$$

$$\leq \lim_{\lambda \to \infty} \left(\frac{1}{np} \sum_{i=1}^{n} \left| \nabla_f^2 \ell \left(f_w; x_i, y_i \right) \right| \right) \left\| f_w - f_{\widetilde{w}} \right\|_{\mathcal{H}} \tag{36}$$

$$= \frac{1}{np} \sum_{i=1}^{n} \left| \nabla_f^2 \ell \left(f_w; x_i, y_i \right) \right| \left\| f_w - f_{\widetilde{w}} \right\|_{\mathcal{H}} \tag{37}$$

There is no dependence on λ. ■

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. The Moreau Envelope can be interpreted as an infimal convolution of the function f. When f is ρ-weakly convex, if $\lambda \leq \rho^{-1}$, then the Moreau Envelope is smooth.

Definition 9. (**Moreau Envelope on closed, convex set**, [26]). 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:

$$\mathsf{M}_{\lambda \ell}(f_w) \triangleq \inf_{f_{\widehat{w}} \in \mathcal{K}} \left\{ \ell(f_{\widehat{w}}) + \frac{1}{2\rho} \left\| f_w - f_{\widehat{w}} \right\|_{\mathcal{H}}^2 \right\} \tag{38}$$

Definition 10. Define the function $\Phi(f_w) \triangleq \max_{t \in \mathbb{R}} g\left(t, f_w\right)$. This function is a L-weakly convex function in $\mathcal{K}$, i.e., $\Phi(f_w) + \frac{L}{2} \left\| f_w \right\|_{\mathcal{H}}^2$ is a convex function over w in the convex and compact set $\mathcal{K}$.

Definition 11 (Stationary Point of Moreau Envelope)**.** A point $f_{\widehat{w}}$ is a stationary point of the Moreau Envelope defined in Definition 9 of Φ defined in Definition 10 if

$$f_{\widehat{w}} = \arg\inf_{f_w \in \mathcal{K}} \left\{ \Phi_\lambda \left(f_w \right) + \frac{1}{2\rho} \left\| f_w - f_{\widehat{w}} \right\|_{\mathcal{H}}^2 \right\} \tag{39}$$

We will show that if a point f_w is a stationary point then this point is close to the optimal point for the uncorrupted distribution, i.e. $\left\| f_{\widehat{w}} - f_w^* \right\|_{\mathcal{H}}$ is small.

Lemma 12 (Lower bound on distance from stationary point and optimal point)**.** *Let Φ_λ be defined as in Definition 10, then if $f_{\widehat{w}}$ is a stationary point as defined in Definition 11, then*

$$\lim_{\lambda \to \infty} \left(\Phi_\lambda \left(f_{\widehat{w}} \right) - \Phi_\lambda \left(f_w^* \right) \right) \leq \beta \left\| f_{\widehat{w}} - f_w^* \right\|_{\mathcal{H}}^2 \tag{40}$$

Proof. By the definition of stationary point, we have

$$f_{\widehat{w}} = \lim_{\lambda \to \infty} \arg\inf_{f_w \in \mathcal{K}} \left\{ \Phi_\lambda \left(f_w \right) + \frac{1}{2\rho} \left\| f_w - f_{\widehat{w}} \right\|_{\mathcal{H}}^2 \right\} \tag{41}$$

$$\overset{\zeta_1}{=} \arg\inf_{f_w \in \mathcal{K}} \left\{ \lim_{\lambda \to \infty} \Phi_\lambda \left(f_w \right) + \frac{1}{2\rho} \left\| f_w - f_{\widehat{w}} \right\|_{\mathcal{H}}^2 \right\} \tag{42}$$

(ζ_1) holds as we ρ is independent of λ as shown in the proof of Lemma 8. This implies then for any $f_w \in \mathcal{K}$ and noting $\rho \leq \beta^{-1}$, it follows

$$\lim_{\lambda \to \infty} \Phi_\lambda \left(f_{\widehat{w}} \right) \leq \lim_{\lambda \to \infty} \Phi_\lambda \left(f_w \right) + \beta \left\| f_w - f_{\widehat{w}} \right\|_{\mathcal{H}}^2 \tag{43}$$

We can then plug in the optimal, f_w^* for f_w and rearrange and we have the desired result. ■

We can now upper bound $\|f_{\widehat{w}} - f_w^*\|_{\mathcal{H}}$. We proceed by contradiction, i.e. if a stationary point is sufficiently far from the optimal point, then this will break the stationary property proved in Lemma 12. This bound is different for each of the loss functions, so we must upper bound $\|f_{\widehat{w}} - f_w^*\|_{\mathcal{H}}$ seperately for each loss function with the same high level overview.

3.3 Kernelized Regression

The loss for the Kernel Ridge Regression problem for a single training pair (x_i, y_i) is given by the following equation

$$\ell\left(f_w; x_i, y_i,\right) = \left(f_w(x_i) - y_i\right)^2 \tag{44}$$

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

Lemma 13. *(L-Lipschitz of $g\left(t, f_w\right)$ w.r.t f_w). Let $x_1, x_2, \cdots, x_n$, represent the data vectors. It then follows for any $f_w, f_{\widehat{w}} \in \mathcal{K}$:*

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

where

$$L = \frac{2R}{np}\left(\sum_{i=1}^{n} \sqrt{k(x_i, x_i)}\right)^2 + \frac{2\left\|y\right\|}{np}\left(\sum_{i=1}^{n} \sqrt{k(x_i, x_i)}\right) \tag{46}$$

Proof is given in Appendix B.1.

Lemma 14. *(β-Smoothness of $g(t, w)$ w.r.t w). Let $x_1, x_2, \cdots, x_n$ represent the rows of the data matrix* X. *It then follows:*

$$\left\|\nabla_w g\left(t, f_w\right) - \nabla_f g\left(t, f_{\widehat{w}}\right)\right\| \leq \beta \left\|f_w - f_{\widehat{w}}\right\|_{\mathcal{H}} \tag{47}$$

where

$$\beta = \frac{2}{np} \sum_{i \in X} k(x_i, x_i) = \frac{2}{np} \operatorname{Tr}\left(\mathrm{K}\right) \tag{48}$$

Proof. W.L.O.G, let S be the set of points such that if $x \in S$, then $t \geq \left(f_w(x) - y\right)^2$. Since g is twice differentiable, we will analyze the second derivative.

$$\left\|\nabla_f^2 g\left(t, f_w\right)\right\|_{\mathcal{H}} = \left\|\frac{2}{np}\sum_{i=1}^{n} \mathbb{I}\left\{t \geq \left(f_w(x_i) - y_i\right)^2\right\} k(x_i, x_i)\right\| \leq \frac{2}{np}\sum_{i=1}^{n} k(x_i, x_i) = \frac{2}{np}\operatorname{Tr}\left(\mathrm{K}\right) \tag{49}$$

This conclude the proof. ■

Similar results for the Lipschitz Constant for non-kernelized learning algorithms can be seen in [43]. It is important to note that β is upper bounded as

$$\beta = \frac{2}{np}\operatorname{Tr}(\mathrm{K}) \leq \frac{2}{np}\left(n(1-\varepsilon)\max_{i \in P} k(x_i, x_i) + n\varepsilon \max_{j \in Q} k(x_j, x_j)\right) = 2p^{-1}\left((1-\varepsilon)P_k + \varepsilon Q_k\right) \tag{50}$$

which is independent of n.

Lemma 15. *If $\|f_w - f_{w^*}\| \geq \eta$, then it follows*

$$\begin{aligned}\Phi\left(f_w\right) - \Phi\left(f_{w^*}\right) \geq \eta^2 \big(n(1-2\varepsilon)\lambda_{\min}\left(\Sigma\right) - O\left(P_k \left\|\Sigma\right\|_{\mathrm{HS}} \sqrt{n(1-2\varepsilon)}\right)\big) \\ - 2\eta \left\|\sum_{i \in S \cap P} \eta_i \phi(x_i)\right\|_{\mathcal{H}} - \sum_{j \in P \setminus S} \eta_j^2\end{aligned} \tag{51}$$

The proof is deferred to Appendix B.2.

Theorem 16. *Let $f_{\widehat{w}}$ be a stationary point defined in Definition 11 for the function Φ defined in Definition 10.*

Then for a constant $c_1 \in (0,1)$, if $n \geq \frac{(1-2\varepsilon)(C_k\|\Sigma\|_{\mathrm{op}}+\beta)}{(1-c_1)\lambda_{\min}(\Sigma)} + \sqrt{\beta}$,

$$\eta \leq \sqrt{\frac{\sum_{j \in P \cap S} \eta_j^2}{c_1 n \lambda_{\min}(\Sigma)}} + \frac{2\left\|\sum_{i \in S \cap P} \eta_i \phi(x_i)\right\|_{\mathcal{H}}}{c_1 n \lambda_{\min}(\Sigma)} \tag{52}$$

where β is the Lipschitz Gradient Constant given in Lemma 14.

The proof is given in Appendix B.3. In Theorem 16, we have an upper bound on the distance from a stationary point to the optimal point. The numerator of the second term grows in $O(\sqrt{n})$ and the denominator grows in $O(n)$ as can be shown by choosing sufficiently large n. Asymptotically the second term will then go to 0. In the first term, we have both the numerator and denominator scale in $O(n)$. Furthermore, when we consider the case of feature noise, e.g. a large multiplicative term on the features, we simply require more data to obtain the same bounds. Such a result is corroborated in [39]. For the linear and polynomial kernel, we then have β increases, therefore to obtain the same bound on η as with no feature noise, we simply need more data.

Corollary 17. *Consider Subquantile Minimization for Linear Regression on the data X with optimal parameters w^*. Assume $x_i \sim \mathcal{N}(0, \Sigma)$ for $i \in [n]$. Then after T iterations of Algorithm 1, we have the following error bounds for robust kernelized linear regression. Given sufficient data*

$$\mathbb{E}\left\|w^{(T)} - w^*\right\|_2 \leq \tilde{O}\left(\frac{\sigma}{\sqrt{\lambda_{\min}(\Sigma)}}\right) \tag{53}$$

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 (x_i, y_i)

$$\ell(x_i, y_i; f_w) = -\left(y_i \log\left(\sigma\left(f_w(x_i)\right)\right) - (1-y_i)\log\left(1 - \sigma\left(f_w(x_i)\right)\right)\right) \tag{54}$$

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

Lemma 18. *(L-Lipschitz of $g(t,w)$ w.r.t w). Let $x_1, x_2, \cdots, x_n$, represent the data vectors. It then follows:*

$$|g(t, f_w) - g(t, f_{\widehat{w}})| \leq L\,\|f_w - f_{\widehat{w}}\|_{\mathcal{H}} \tag{55}$$

where

$$L = \frac{1}{np}\sum_{i \in X}\sqrt{k(x_i, x_i)} = \frac{1}{np}\operatorname{Tr}(\mathrm{K}) \tag{56}$$

Proof is given in Appendix B.4.

Lemma 19. *(β-Smoothness of $g(t,w)$ w.r.t w). Let $x_1, x_2, \cdots, x_n$ represent the rows of the data matrix X. It then follows:*

$$\|\nabla_f g(t, f_w) - \nabla_f g(t, f_{\widehat{w}})\| \leq \beta\,\|f_w - f_{\widehat{w}}\|_{\mathcal{H}} \tag{57}$$

where

$$\beta = \frac{1}{4p}\sum_{i=1}^{n} k(x_i, x_i) = \frac{1}{4p}\operatorname{Tr}(\mathrm{K}) \tag{58}$$

Proof is given in Appendix B.5.

Lemma 20. [1] *If $\|f_w - f_{w^*}\| \geq \eta$, then it follows*

$$\Phi(f_w) - \Phi(f_{w^*}) \geq 2\,\|f_w - f_w^*\|_{\mathcal{H}} - 1.386 + \sum_{P \setminus S} y_i \log\left(\sigma\left(f_w^*(x_i)\right)\right) + (1-y_i)\log\left(1 - \sigma\left(f_w^*(x_i)\right)\right) \tag{59}$$

Proof is given in Appendix B.6.

[1] In Progress

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 (x_i, y_i), note W is now a matrix

$$\ell\left(x_i, y_i; \mathrm{W}\right) = -\sum_{j=1}^{|\mathcal{C}|} \mathbb{I}\left\{y_i = j\right\} \log\left(\frac{\exp\left(f_{\mathrm{W}_k}(x_i)\right)}{\sum_{h=1}^{|\mathcal{C}|} \exp\left(f_{\mathrm{W}_h}(x_i)\right)}\right) \tag{60}$$

Lemma 21. *(L-Lipschitz of $g\left(t, w\right)$ w.r.t w). Let $x_1, x_2, \cdots, x_n$, represent the data vectors. It then follows:*

$$\left|g\left(t, f_w\right) - g\left(t, f_{\widehat{w}}\right)\right| \le L \left\|f_w - f_{\widehat{w}}\right\|_{\mathcal{H}} \tag{61}$$

where

$$L = \tag{62}$$

3.6 Necessary Kernel Inequalities

We will first extend the idea of Resilience [40] to kernel learning.

Definition 22. (**Resilience**) from [40]. 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 $X = \{x_i\}_{i=1}^n = P \cup Q$, such that $|P| = n(1-\epsilon)$ and $|Q| = n\epsilon$, It then follows for any subset $T \subseteq P$ such that $|T| = (1-2\varepsilon)n$,

$$\left\|\frac{1}{|T|}\sum_{i \in T} \phi\left(x_i\right) - \mu_{\mathbb{P}}\right\| \le \tau$$

where $\mu_{\mathbb{P}} = \mathbb{E}_{x\sim\mathbb{P}}[\phi(x)]$ 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 22, we will be unable to put error bounds on our algorithm. 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 23 ([36, 6]). *Assume the function Φ is β-weakly convex. Let $\lambda < \frac{1}{\beta}$, and let $f_{\widehat{w}} = \arg\min_{f_w \in \mathcal{K}}(\Phi\left(f_w\right) + \frac{1}{2\lambda}\left\|f_w - f_{\widehat{w}}\right\|_{\mathcal{H}}^2)$, then $\|\nabla\Phi_\lambda(f_w)\|_{\mathcal{H}} \le \epsilon$ implies:*

$$\|f_{\widehat{w}} - f_w\|_{\mathcal{H}} = \lambda\epsilon \quad \textit{and} \quad \min_{g \in \partial\Phi(f_{\widehat{w}}) + \partial\mathcal{I}_{\mathcal{K}}(f_{\widehat{w}})} \|g\|_{\mathcal{H}} \le \epsilon \tag{63}$$

4 Optimization Results

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, $\|w - w^*\|$ is not sufficient, we want $\|w - w^*\|_{\mathcal{H}}$ to be close, as the RKHS being small indicates the function $f(x)$ and $f^*(x)$ 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. In the analysis given in [20], given sufficient iterations, if we choose a sufficiently small learning rate, we can always reach a stationary point given infinite iterations. In practice, and in theory, this convergence rate is a square root factor slower. We prove, if we instead converge to a weaker stationary point, we can converge a factor of square root faster.

Theorem 24. *Let $\Psi(f_w) \triangleq \max_{t\in\mathbb{R}} g(f_w, t)$ where $g(f_w, t)$ is β-smooth in f_w and L-Lipschitz in f_w and is concave in t but not necesarily convex in f_w. Then Algorithm 1 with stepsize $\eta = \frac{1}{\beta}$ reaches a $(\sqrt{2}L + O(1))$-stationary point in $O(16\beta R + 4\beta np\sigma^2))$ iterations.*

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 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 [33, 30] with our non-convex-concave optimization algorithm.

$$b^{(t)} = \mu b^{(t-1)} + \nabla_f \Phi\left(f_w^{(t-1)}\right) \tag{64}$$

$$f_w^{(t)} = f_w^{(t-1)} - \alpha b^{(t)} \tag{65}$$

Theorem 25. *Momentum Accelerated Gradient Descent given in Equations (64) and (65) reaches a η-approximate stationary point. Algorithm 1 reaches a η-approximate stationary point in a polynomial number of iterations.*

$$\mathbb{E}\left[\left\|\nabla \Phi_{1/2\ell}\left(f_w\right)\right\|_{\mathcal{H}}^2\right] \leq \tag{66}$$

4.2.2 Nesterov Accelerated Gradient Descent

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

$$b^{(t+1)} = (1+\mu)\, f_w^{(t)} - \mu f_w^{(t-1)} \tag{67}$$

$$f_w^{(t+1)} = b^{(t+1)} - \alpha \nabla_f \Phi\left(f_w^{(t)}\right) \tag{68}$$

Theorem 26. *Nesterov Accelerated Gradient Descent given in Equations (67) and (68) reaches a η-approximate stationary point. Algorithm 1 reaches a η-approximate stationary point in a polynomial number of iterations.*

$$\mathbb{E}\left[\left\|\nabla \Phi_{1/2\ell}\left(f_w\right)\right\|_{\mathcal{H}}^2\right] \leq \tag{69}$$

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_w with the Glorot Initialization Scheme [12].

Algorithm 3: Subquantile-Kernel

```
Input: Iterations: $T$; Quantile: $p$; Data Matrix: $\mathrm{X} \in \mathbb{R}^{n \times d}, n \gg d$; Labels: $y \in \mathbb{R}^{n \times 1}$; Learning Rate
       schedule: $\alpha_1, \cdots, \alpha_T$; Ridge parameter: $\lambda$
Output: Trained Parameters: $f_w^{(T)}$
$w_i^{(0)} \leftarrow \mathrm{Unif}\left[-\sqrt{\frac{6}{n}}, \sqrt{\frac{6}{n}}\right], \forall i \in [n]$            ▷ Base Learner
for $k = 1, 2, \ldots, T$ do
    $S^{(k)} \leftarrow$ SUBQUANTILE$(f_w^{(k)}, \mathrm{X})$            ▷ Algorithm 2
    $\nabla_{\mathrm{f}} g\left(t^{(k+1)}, f_w^{(k)}\right) \leftarrow 2 \sum_{i \in S^{(k)}} \left(f_w^{(k)}(x_i) - y_i\right) \cdot k(x_i, \cdot)$            ▷ Regression
    $\nabla_f g\left(t^{(k+1)}, f_w^{(k)}\right) \leftarrow \sum_{i \in S^{(k)}} \left(\sigma\left(f_w^{(k)}(x_i)\right) - y_i\right) \cdot k(x_i, \cdot)$            ▷ Binary Classification
    $f_w^{(k+1)} \leftarrow f_w^{(k)} - \alpha_{(k)} \nabla_f g\left(t^{(k+1)}, f_w^{(k)}\right)$            ▷ $f_w$-update in Equation (8)
end
Pick $t$ uniformly at random from $[T]$
return $f_w^{(t)}$
```

In Figure 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.

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

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	$0.829_{(0.0422)}$	$0.928_{(0.0197)}$	$1.854_{(0.7437)}$	$1.069_{(0.1001)}$	$0.879_{(0.0178)}$	$0.875_{(0.0711)}$	∞	∞
SEVER	$\underline{0.533}_{(0.0347)}$	$\underline{0.592}_{(0.0548)}$	$\underline{0.915}_{(0.0343)}$	$\underline{0.841}_{(0.0413)}$	$\underline{0.526}_{(0.0287)}$	$\underline{0.720}_{(0.1147)}$	$\mathbf{1.172}_{(0.0542)}$	$\underline{1.215}_{(0.0536)}$
Subquantile	$\mathbf{0.519}_{(0.0134)}$	$\mathbf{0.547}_{(0.0174)}$	$\mathbf{0.808}_{(0.0389)}$	$\mathbf{0.827}_{(0.0216)}$	$\mathbf{0.468}_{(0.0896)}$	$\mathbf{0.458}_{(0.0662)}$	$\underline{1.280}_{(0.0568)}$	$\mathbf{1.132}_{(0.0892)}$
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 $x_{\text{noise}} = 100 x_{\text{original}}$. `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 polynomial kernel with degree 3 and $C = 1$ is used in the last experiment.

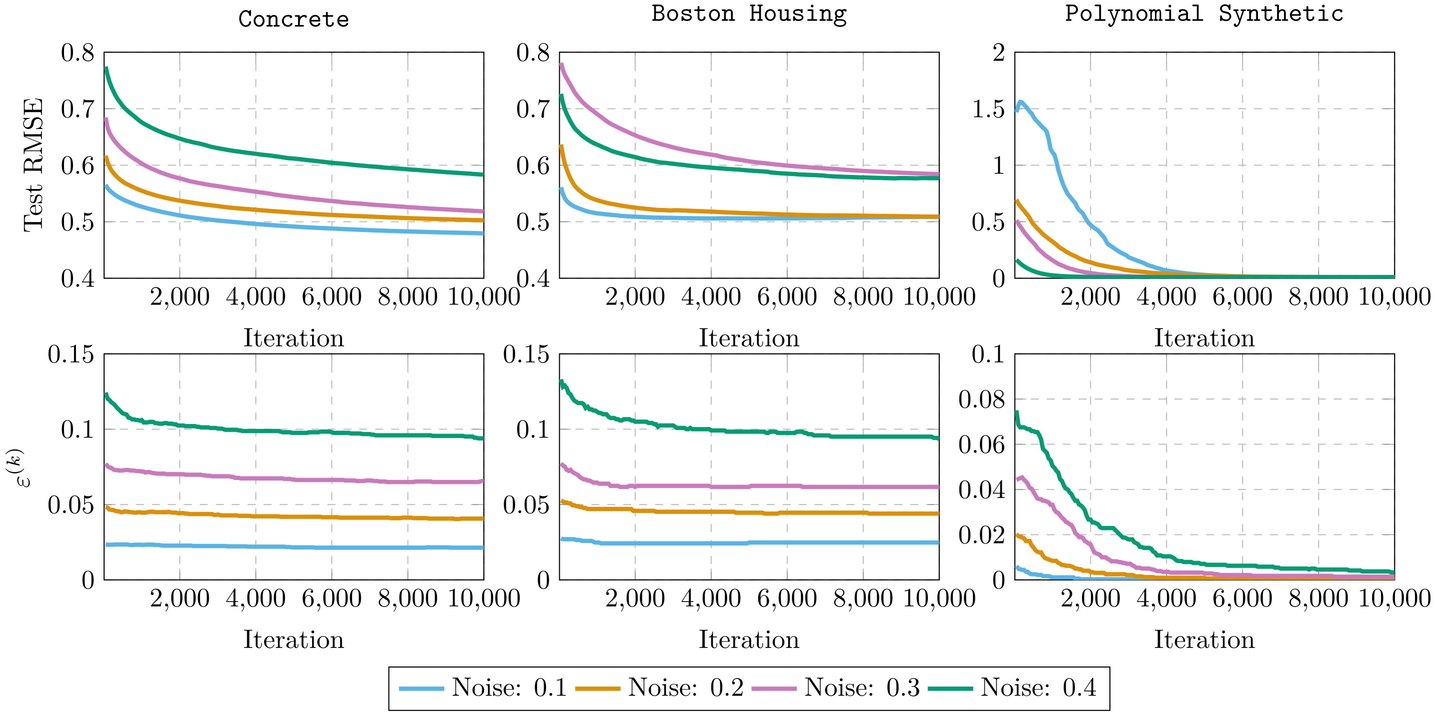

Figure 1: Test RMSE over the iterations in `Concrete`, `Boston Housing`, and `Polynomial` Datasets for Subquantile at different noise levels

art robust algorithms.

Now we will give some experimental results.

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 **??**, we see using momentum can give significantly faster convergence.

Algorithms	Test RMSE							
	`Boston Housing`		`Wine Quality`		`Concrete`		`Drug`	
	Label(↓)	Label+Feature	Label	Label+Feature	Label	Label+Feature	Label	Label+Feature
KRR	$0.907_{(0.2724)}$	$90.799_{(5.7170)}$	$0.894_{(0.0404)}$	$62.913_{(7.4959)}$	$0.825_{(0.0943)}$	$77.383_{(5.5692)}$	$2.679_{(0.1286)}$	$141.690_{(3.5297)}$
RANSAC	$1.167_{(0.6710)}$	$22.460_{(19.1987)}$	$1.489_{(0.2730)}$	$39.630_{(13.0294)}$	$0.870_{(0.2308)}$	$23.629_{(16.1023)}$	$2.801_{(0.2004)}$	$117.389_{(8.3915)}$
CRR	$0.636_{(0.0905)}$	$88.626_{(5.7380)}$	$0.818_{(0.0224)}$	$58.488_{(3.5612)}$	$0.710_{(0.0919)}$	$73.932_{(4.7867)}$	$1.887_{(0.1463)}$	$152.827_{(6.6038)}$
STIR	$\underline{0.562}_{(0.0626)}$	$78.878_{(8.0164)}$	$0.828_{(0.0293)}$	$58.352_{(4.6700)}$	$\mathbf{0.684}_{(0.0245)}$	$76.555_{(4.5927)}$	$1.721_{(0.1520)}$	$144.975_{(5.4953)}$
SEVER	$0.601_{(0.0979)}$	$5.980_{(8.2603)}$	$\mathbf{0.814}_{(0.0207)}$	$9.065_{(13.7632)}$	$\mathbf{0.684}_{(0.0438)}$	$4.119_{(8.2436)}$	$1.469_{(0.1162)}$	$156.043_{(4.5543)}$
TERM	$0.608_{(0.1357)}$	$\underline{0.569}_{(0.0620)}$	$0.840_{(0.0563)}$	$\underline{0.827}_{(0.0255)}$	$0.830_{(0.0934)}$	$\underline{0.808}_{(0.0726)}$	$\mathbf{1.185}_{(0.1077)}$	$\mathbf{1.147}_{(0.1258)}$
Subquantile	$\mathbf{0.503}_{(0.0470)}$	$\underline{0.560}_{(0.0373)}$	$\mathbf{0.813}_{\mathbf{(0.0357)}}$	$\mathbf{0.821}_{(0.0305)}$	$0.720_{(0.1092)}$	$\mathbf{0.630}_{(0.0269)}$	$\underline{1.244}_{(0.1091)}$	$\underline{2.413}_{(0.6737)}$
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 $x_{\text{noisy}} = 100x_{\text{original}}$ and $y_{\text{noisy}} = 10000y_{\text{original}}$.

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	$0.777_{(0.0396)}$	$0.639_{(0.0762)}$	$0.534_{(0.0766)}$	$0.538_{(0.0626)}$	$0.926_{(0.0331)}$	$0.548_{(0.1194)}$	$0.649_{(0.0254)}$	$0.618_{(0.0507)}$
SEVER	$\underline{0.793}_{(0.0422)}$	$\underline{0.695}_{(0.0636)}$	$0.784_{(0.0432)}$	$\mathbf{0.816}_{(0.0562)}$	$0.904_{(0.0356)}$	$0.575_{(0.1456)}$	$\underline{0.956}_{(0.0164)}$	$\underline{0.974}_{(0.0062)}$
TERM	$0.741_{(0.0393)}$	$0.620_{(0.0699)}$	$\underline{0.803}_{(0.0613)}$	$\underline{0.810}_{(0.0286)}$	$\underline{0.940}_{(0.0378)}$	$\underline{0.763}_{(0.0364)}$	$\mathbf{0.986}_{(0.0143)}$	$\mathbf{0.986}_{(0.0119)}$
Subquantile	$\mathbf{0.803}_{(0.0293)}$	$\mathbf{0.790}_{(0.0350)}$	$\mathbf{0.833}_{(0.0318)}$	$\underline{0.807}_{(0.0468)}$	$\mathbf{0.928}_{(0.0129)}$	$\mathbf{0.916}_{(0.0185)}$	$0.951_{(0.0212)}$	$0.953_{(0.0197)}$
Genie ERM	∞	∞	∞	∞	∞	∞	∞	∞

Table 3: `Heart Disease` and `Breast Cancer` Dataset. Label Noise: $y_{\text{noise}} = \mathbb{I}\{y_{\text{original}} = 0\}$. Feature Noise: $x_{\text{noise}} = 100x_{\text{original}}$. The Linear Kernel is used in all experiments.

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. We develop strong theoretical bounds on the normed difference between the function returned by Subquantile Minimization and the optimal function for data in the target distribution, $\mathbb{P}$, in the Gaussian Design. In expectation and with high probability, given sufficient data dependent on the kernel, we obtain a near minimax optimal error bound for a general positive definite continuous kernel.

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 and the features follow Gaussian Design. 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. Our theory

suggests for a problem where the amount of corruptions is unknown,
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. Furthermore, we assume the feature maps have a Gaussian Design. Such a design in many prior works in kernel learning and we therefore find it suitable.
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. 1, 1.1

[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. 1

[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. 1, 1.1

[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. 1

[5] Hugo Cui, Bruno Loureiro, Florent Krzakala, and Lenka Zdeborová. Generalization error rates in kernel regression: The crossover from the noiseless to noisy regime. *Advances in Neural Information Processing Systems*, 34:10131–10143, 2021. B.2

[6] Damek Davis and Dmitriy Drusvyatskiy. Stochastic model-based minimization of weakly convex functions. *SIAM Journal on Optimization*, 29(1):207–239, 2019. 23

[7] 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. 1, 1.1

[8] Ilias Diakonikolas and Daniel M Kane. *Algorithmic high-dimensional robust statistics.* Cambridge University Press, 2023. 1

[9] Dheeru Dua and Casey" Graff. UCI machine learning repository, 2017. E.1

[10] 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. 1

[11] 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. 1

[12] 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. 5

[13] 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. B.1

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

[15] Shu Hu, Zhenhuan Yang, Xin Wang, Yiming Ying, and Siwei Lyu. Outlier robust adversarial training. *arXiv preprint arXiv:2309.05145*, 2023. 1.1

[16] Shu Hu, Yiming Ying, Siwei Lyu, et al. Learning by minimizing the sum of ranked range. *Advances in Neural Information Processing Systems*, 33:21013–21023, 2020. 1, 1.1

[17] Peter J. Huber and Elvezio. Ronchetti. *Robust statistics.* Wiley series in probability and statistics. Wiley, Hoboken, N.J., 2nd ed. edition, 2009. 1

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

[19] 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. 1

[20] 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. 1.1, 4.1, C.1

[21] Ashish Khetan, Zachary C. Lipton, and Anima Anandkumar. Learning from noisy singly-labeled data. In *International Conference on Learning Representations*, 2018. 1

[22] 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.

[23] 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. (document)

[24] Tian Li, Ahmad Beirami, Maziar Sanjabi, and Virginia Smith. Tilted empirical risk minimization. In *International Conference on Learning Representations*, 2021. 1, 1, 1.1

[25] James Mercer. Xvi. functions of positive and negative type, and their connection the theory of integral equations. *Philosophical transactions of the royal society of London. Series A, containing papers of a mathematical or physical character*, 209(441-458):415–446, 1909. 28

[26] Jean-Jacques Moreau. Proximité et dualité dans un espace hilbertien. *Bulletin de la Société mathématique de France*, 93:273–299, 1965. 9

[27] 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. 1

[28] 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. 4.2.2

[29] 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. 1

[30] Boris T Polyak. Some methods of speeding up the convergence of iteration methods. *Ussr computational mathematics and mathematical physics*, 4(5):1–17, 1964. 4.2.1

[31] Adarsh Prasad, Sivaraman Balakrishnan, and Pradeep Ravikumar. A unified approach to robust mean estimation. *arXiv preprint arXiv:1907.00927*, 2019. 1

[32] Adarsh Prasad, Arun Sai Suggala, Sivaraman Balakrishnan, and Pradeep Ravikumar. Robust estimation via robust gradient estimation. *arXiv preprint arXiv:1802.06485*, 2018. 1

[33] Ning Qian. On the momentum term in gradient descent learning algorithms. *Neural networks*, 12(1):145–151, 1999. 4.2.1

[34] 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. 2

[35] R Tyrrell Rockafellar and Stanislav Uryasev. Conditional value-at-risk for general loss distributions. *Journal of banking & finance*, 26(7):1443–1471, 2002. 4

[36] Ralph Tyrell Rockafellar. *Convex Analysis*. Princeton University Press, Princeton, 1970. 23

[37] 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. (document)

[38] Gabriele Santin and Robert Schaback. Approximation of eigenfunctions in kernel-based spaces. *Advances in Computational Mathematics*, 42:973–993, 2016. B.2

[39] Ludwig Schmidt, Shibani Santurkar, Dimitris Tsipras, Kunal Talwar, and Aleksander Madry. Adversarially robust generalization requires more data. *Advances in neural information processing systems*, 31, 2018. 3.3

[40] 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. 3.6, 22

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

[42] 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. 1.1

[43] Rahul Yedida, Snehanshu Saha, and Tejas Prashanth. Lipschitzlr: Using theoretically computed adaptive learning rates for fast convergence. *Applied Intelligence*, 51:1460–1478, 2021. 3.3

A Concentration Inequalities

In this section we will give various concentration inequalities on the inlier data.

Lemma 27 (Infinite Dimensional Covariance Estimation in the Hilbert-Schmidt Norm)**.** *Let* $\Sigma \triangleq \mathbb{E}_{\phi(x_i)\sim\mathbb{P}}[\phi(x_i)\otimes\phi(x_i)]$*. Then let* $x_1,\ldots,x_n$ *be i.i.d sampled from* $\mathbb{P}$ *such that* $\phi(x_i)\sim\mathcal{N}(0,\Sigma)$*, we then have*

$$\mathbb{E}_{\phi(x_i)\sim\mathcal{N}(0,\Sigma)}\left\|\frac{1}{n}\sum_{i=1}^{n}\Phi(x_i)\otimes\Phi(x_i)-\Sigma\right\|_{\mathrm{HS}}\le O\left(n^{-1/2}\left\|\Sigma\right\|_{\mathrm{HS}}\right) \tag{70}$$

Proof. Our proof follows standard ideas from High-Dimensional Probability. Let ξ_i for $i\in[n]$ denote i.i.d Rademacher variables such that for $\xi_i\sim\mathcal{R}$, it follows $\mathbb{P}\{\xi_i=1\}=\mathbb{P}\{\xi_i=-1\}=\frac{1}{2}$. We then have,

$$\mathbb{E}_{\phi(x_i)\sim\mathcal{N}(0,\Sigma)}\left\|\frac{1}{n}\sum_{i=1}^{n}\phi(x_i)\otimes\phi(x_i)-\Sigma\right\|_{\mathrm{HS}}$$

$$\overset{\zeta_1}{\le}\mathbb{E}_{\phi(x_i)\sim\mathcal{N}(0,\Sigma)}\mathbb{E}_{\tilde{\phi}\sim\mathcal{N}(0,\Sigma)}\left\|\frac{1}{n}\sum_{i=1}^{n}\left(\phi(x_i)\otimes\phi(x_i)-\tilde{\phi}(x_i)\otimes\tilde{\phi}(x_i)\right)\right\|_{\mathrm{HS}} \tag{71}$$

$$=\mathbb{E}_{\phi(x_i)\sim\mathcal{N}(0,\Sigma)}\mathbb{E}_{\tilde{\phi}(x_i)\sim\mathcal{N}(0,\Sigma)}\mathbb{E}_{\xi_i\sim\mathcal{R}}\left\|\frac{1}{n}\sum_{i=1}^{n}\xi_i\left(\phi(x_i)\otimes\phi(x_i)-\tilde{\phi}(x_i)\otimes\tilde{\phi}(x_i)\right)\right\|_{\mathrm{HS}} \tag{72}$$

$$\overset{\zeta_2}{\le}\frac{2}{n}\mathbb{E}_{\phi(x_i)\sim\mathcal{N}(0,\Sigma)}\mathbb{E}_{\xi_i\sim\mathcal{R}}\left\|\sum_{i=1}^{n}\xi_i\phi(x_i)\otimes\phi(x_i)\right\|_{\mathrm{HS}} \tag{73}$$

$$\overset{\zeta_3}{\le}\frac{2}{n}\left(\mathbb{E}_{\phi(x_i)\sim\mathcal{N}(0,\Sigma)}\mathbb{E}_{\xi_i\sim\mathcal{R}}\left\|\sum_{i=1}^{n}\xi_i\phi(x_i)\otimes\phi(x_i)\right\|_{\mathrm{HS}}^2\right)^{1/2} \tag{74}$$

(ζ_1) follows from noticing $\phi(x_i)\otimes\phi(x_i)-\Sigma$ is a mean 0 covariance operator and then applying Jensen's Inequality. (ζ_2) follows from the triangle inequality. (ζ_3) follows from Jensen's Inequality. Let e_k for $k\in[p]$ represent an orthonormal basis for the tensor product space $\mathcal{H}\otimes\mathcal{H}$. By expanding out the Hilbert-Schmidt Norm, we then have

$$\frac{2}{n}\left(\mathbb{E}_{\phi(x_i)\sim\mathcal{N}(0,\Sigma)}\mathbb{E}_{\xi_i\sim\mathcal{R}}\left\|\sum_{i=1}^{n}\xi_i\phi(x_i)\otimes\phi(x_i)\right\|_{\mathrm{HS}}^2\right)^{1/2}$$

$$=\frac{2}{n}\left(\mathbb{E}_{\phi(x_i)\sim\mathcal{N}(0,\Sigma)}\mathbb{E}_{\xi_i\sim\mathcal{R}}\sum_{k=1}^{p}\left\langle\sum_{i=1}^{n}\xi_i\phi(x_i)\otimes\phi(x_i)e_k,\sum_{j=1}^{n}\xi_j\phi(x_j)\otimes\phi(x_j)e_k\right\rangle\right)^{1/2} \tag{75}$$

$$=\frac{2}{n}\left(\mathbb{E}_{\phi(x_i)\sim\mathcal{N}(0,\Sigma)}\mathbb{E}_{\xi_i\sim\mathcal{R}}\sum_{k=1}^{p}\sum_{i=1}^{n}\sum_{j=1}^{n}\xi_i\xi_j\left\langle\phi(x_i)\otimes\phi(x_i)e_k,\phi(x_j)\otimes\phi(x_j)e_k\right\rangle\right)^{1/2} \tag{76}$$

$$\overset{\zeta_4}{\le}\frac{2}{n}\left(\mathbb{E}_{\phi(x_i)\sim\mathcal{N}(0,\Sigma)}\sum_{k=1}^{p}\sum_{i=1}^{n}\left\langle\phi(x_i)\otimes\phi(x_i)e_k,\phi(x_i)\otimes\phi(x_i)e_k\right\rangle\right)^{1/2} \tag{77}$$

$$=\frac{2}{n}\left(\sum_{i=1}^{n}\mathbb{E}_{\phi(x_i)\sim\mathcal{N}(0,\Sigma)}\left\|\phi(x_i)\otimes\phi(x_i)\right\|_{\mathrm{HS}}^2\right)^{1/2}\overset{\zeta_5}{=}\frac{2}{n}\left(\sum_{i=1}^{n}\mathbb{E}_{\phi(x_i)\sim\mathcal{N}(0,\Sigma)}\left\|\phi(x_i)\right\|_{\mathcal{H}}^4\right)^{1/2} \tag{78}$$

$$=\frac{2}{n}\left(\sum_{i=1}^{n}\mathbb{E}_{\phi(x_i)\sim\mathcal{N}(0,\Sigma)}k^2(x_i,x_i)\right)^{1/2}=\frac{2\left\|\Sigma\right\|_{\mathrm{HS}}}{\sqrt{n}} \tag{79}$$

(ζ_4) follows from noticing $\mathbb{E}_{\xi_i,\xi_j\sim\mathcal{R}}[\xi_i\xi_j] = \delta_{i,j}$. (ζ_5) follows from expanding the Hilbert-Schmidt Norm and applying Parseval's Identity. This completes the proof. ■

B Proofs for Section 3

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

B.1 Proof of Lemma 13

Proof. For any $f_{w_1}, f_{w_2} \in \mathcal{K}$, we will first show the gradient is bounded.

$$\begin{aligned}
|g(t,f_{w_1}) - g(t,f_{w_2})| &= \left|\int_0^1 \nabla_f g(t,(1-\lambda)f_{w_1} + \lambda f_{w_2})(f_{w_1} - f_{w_2})d\lambda\right| && (80)\\
&\leq \|f_{w_1} - f_{w_2}\|_{\mathcal{H}} \left|\int_0^1 \nabla_f g(t,(1-\lambda)f_{w_1} + \lambda f_{w_2})d\lambda\right| && (81)\\
&\overset{(a)}{\leq} \|f_{w_1} - f_{w_2}\|_{\mathcal{H}} \max_{f_w\in\mathcal{K}} \|\nabla_f g(t,f_w)\|_{\mathcal{H}} && (82)
\end{aligned}$$

In (a), we note that since $\mathcal{K}$ is convex, then by definition as $f_{w_1}, f_{w_2} \in \mathcal{K}$, we have for $\lambda \in [0,1]$, the convex combination $(1-\lambda)f_{w_1} + \lambda f_{w_2} \in \mathcal{K}$. We use the $\mathcal{H}$ norm of the gradient to bound L from above for an element in the convex closed set $\mathcal{K}$.

$$\|\nabla g(t,f_w)\|_{\mathcal{H}} = \left\|\frac{2}{np}\sum_{i=1}^n \mathbb{I}\left\{t \geq (f_w(x_i) - y_i)^2\right\}(f_w(x_i) - y_i)\cdot k(x_i,\cdot)\right\|_{\mathcal{H}} \tag{83}$$

W.L.O.G, let $x_1, x_2, \cdots, x_m$ where $0 \leq m \leq n$, represent the data vectors such that $t \geq (f_w(x_i) - y_i)^2$.

$$\begin{aligned}
&= \left\|\frac{2}{np}\sum_{i=1}^m (f_w(x_i) - y_i)\cdot k(x_i,\cdot)\right\|_{\mathcal{H}} && (84)\\
&\leq \frac{2}{np}\left(\left\|\sum_{i=1}^m f_w(x_i)\cdot k(x_i,\cdot)\right\|_{\mathcal{H}} + \left\|\sum_{i=1}^m y_i k(x_i,\cdot)\right\|_{\mathcal{H}}\right) && (85)\\
&\overset{(a)}{\leq} \frac{2}{np}\left(\left\|\sum_{i=1}^m \left\langle\sum_{j=1}^n w_j k(x_j,\cdot), k(x_i,\cdot)\right\rangle_{\mathcal{H}}\cdot k(x_i,\cdot)\right\|_{\mathcal{H}} + \left\|\sum_{i=1}^m y_i\right\| \left\|\sum_{i=1}^m k(x_i,\cdot)\right\|_{\mathcal{H}}\right) && (86)\\
&\leq \frac{2}{np}\left(\left|\left\langle\sum_{j=1}^n w_j k(x_j,\cdot), \sum_{i=1}^m k(x_i,\cdot)\right\rangle_{\mathcal{H}}\right| \left\|\sum_{i=1}^m k(x_i,\cdot)\right\|_{\mathcal{H}} + \left\|\sum_{i=1}^m y_i\right\| \sum_{i=1}^m \sqrt{k(x_i,x_i)}\right) && (87)\\
&\leq \frac{2}{np}\left(\|f_w\|_{\mathcal{H}}\left(\sum_{i=1}^m \sqrt{k(x_i,x_i)}\right)^2 + \sqrt{n}\,\|y\|_2\left(\sum_{i=1}^n \sqrt{k(x_i,x_i)}\right)\right) && (88)\\
&\leq \frac{2R}{np}\left(\sum_{i=1}^n \sqrt{k(x_i,x_i)}\right)^2 + \frac{2\,\|y\|_2}{p\sqrt{n}}\left(\sum_{i=1}^n \sqrt{k(x_i,x_i)}\right) && (89)
\end{aligned}$$

(a) follows form the reproducing property for RKHS [13]. 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 f_w has bounded norm, so it will have infinite error. This concludes the proof. ■

B.2 Proof of Lemma 15

Proof.
Let S be the set containing the points with the minimum error from X w.r.t to the weights vector w. Define

$\eta_i \triangleq (f_{w^*}(x_i) - y_i)$ where $i \in P$.

$$\lim_{\lambda\to\infty} \left(\Phi_\lambda(f_w) - \Phi_\lambda(f_{w^*})\right) = \sum_{i\in S} \left(f_w(x_i) - y_i\right)^2 - \sum_{j\in P} \left(f_{w^*}(x_j) - y_j\right)^2 \tag{90}$$

$$= \sum_{i\in S\cap P} \left(f_w(x_i) - y_i\right)^2 + \sum_{i\in S\cap Q} \left(f_w(x_i) - y_i\right)^2 - \sum_{j\in P} \left(f_{w^*}(x_j) - y_j\right)^2 \tag{91}$$

$$\geq \sum_{i\in S\cap P} \left(f_w(x_i) - y_i\right)^2 - \sum_{j\in P} \left(f_{w^*}(x_j) - y_j\right)^2 \tag{92}$$

$$= \sum_{i\in S\cap P} \left(f_w(x_i) - f_{w^*}(x_i) - \eta_i\right)^2 - \sum_{j\in P} \eta_j^2 \tag{93}$$

$$= \sum_{i\in S\cap P} \left((f_w - f_{w^*})(x_i) - \eta_i\right)^2 - \sum_{j\in P} \eta_j^2 \tag{94}$$

$$\geq \sum_{i\in S\cap P} \underbrace{\left((f_w - f_{w^*})(x_i)\right)^2}_{\textcircled{A_1}} \underbrace{-2 \sum_{i\in S\cap P} \eta_i (f_w - f_{w^*})(x_i)}_{\textcircled{A_2}} - \underbrace{\sum_{j\in P\setminus S} \eta_j^2}_{\textcircled{A_3}} \tag{95}$$

Now we will upper bound A_1. Similar to [5] Let $\mathbb{E}_{x\sim\mathbb{P}}[\varphi(x)\otimes\varphi(x)] = \mathbb{I}_m$ where $\varphi(x) = \{\varphi(x)\}_{k=1}^m$ and m is possibly infinite. We can then rescale the basis features. Then let $\phi(x) = \Sigma^{1/2}\varphi(x)$. We therefore have $\Sigma = \mathbb{E}_{x\sim\mathbb{P}}[\phi(x)\otimes\phi(x)] = \mathrm{diag}(\xi_1, \dots, \xi_n)$. This is the eigenfunction basis described in [38].

$$\textcircled{A_1} \triangleq \sum_{i\in S\cap P} \left((f_w - f_{w^*})(x_i)\right)^2 \overset{(a)}{=} \sum_{i\in S\cap P} \left\langle \sum_{j\in X} \left(w_j - w_j^*\right) k(x_j, \cdot), k(x_i, \cdot) \right\rangle_{\mathcal{H}}^2 \tag{96}$$

$$= \sum_{i\in S\cap P} \left\langle \sum_{j\in X} \left(w_j - w_j^*\right) \phi(x_j), \phi(x_i) \right\rangle_{\mathcal{H}} \left\langle \phi(x_i), \sum_{j\in X} \left(w_j - w_j^*\right) \phi(x_j) \right\rangle_{\mathcal{H}} \tag{97}$$

$$= \sum_{i\in S\cap P} \left\langle \sum_{j\in X} \left(w_j - w_j^*\right) \phi(x_j), \phi(x_i)\otimes\phi(x_i) \sum_{j\in X} \left(w_j - w_j^*\right) \phi(x_j) \right\rangle_{\mathcal{H}} \tag{98}$$

$$= \sum_{i\in S\cap P} \left\langle \phi(x)\otimes\phi(x), (f_w - f_w^*)\otimes(f_w - f_w^*) \right\rangle_{\mathrm{HS}} \tag{99}$$

$$= \sum_{i\in S\cap P} \left\langle \Sigma + \phi(x)\otimes\phi(x) - \Sigma, (f_w - f_w^*)\otimes(f_w - f_w^*) \right\rangle_{\mathrm{HS}} \tag{100}$$

$$\geq n(1-2\varepsilon)\mathbb{E}_{x\sim\mathbb{P}}\left[(f_w - f_w^*)(x)^2\right] - \left\| \sum_{i\in S\cap P} \phi(x)\otimes\phi(x) - \Sigma \right\|_{\mathrm{HS}} \|f_w - f_w^*\|_{\mathcal{H}}^2 \tag{101}$$

$$\overset{\text{lem. 27}}{\geq} n(1-2\varepsilon)\left(\lambda_{\min}(\Sigma) - O\left(\frac{C_k \|\Sigma\|_{\mathrm{HS}}}{\sqrt{n(1-2\varepsilon)}}\right)\right) \|f_w - f_w^*\|_{\mathcal{H}}^2 \tag{102}$$

Next we will upper bound $\textcircled{$\mathrm{A}_2$}$,

$$\textcircled{A_2} \triangleq \sum_{i\in S\cap P} \eta_i (f_w - f_{w^*})(x_i) \tag{103}$$

$$= \sum_{i\in S\cap P} \left\langle \sum_{j\in X} (w_j - w_j^*) k(x_j, \cdot), \eta_i k(x_i, \cdot) \right\rangle_{\mathcal{H}} \tag{104}$$

$$= \left\langle \sum_{j\in X} (w_j - w_j^*) k(x_j, \cdot), \sum_{i\in S\cap P} \eta_i k(x_i, \cdot) \right\rangle_{\mathcal{H}} \tag{105}$$

$$\leq \|f_w - f_{w^*}\|_{\mathcal{H}} \left\| \sum_{i\in S\cap P} \eta_i k(x_i,\cdot)\right\|_{\mathcal{H}} = \|f_w - f_{w^*}\|_{\mathcal{H}} \left\|\sum_{i\in S\cap P}\eta_i\phi(x_i)\right\|_{\mathcal{H}} \tag{106}$$

We will now lower bound B_1. For the linear kernel, for B_1 to be greater than 0 , than if $x \in \mathbb{R}^d$, 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

$$\begin{aligned}\lim_{\lambda\to\infty}\left(\Phi_\lambda\left(f_w\right) - \Phi_\lambda\left(f_{w^*}\right)\right) \geq \eta^2 n(1-2\varepsilon)\left(\lambda_{\min}\left(\mathbb{E}_{x\sim\mathbb{P}}\left[\phi(x)\otimes\phi(x)\right]\right) - O\left(\frac{C_k\left\|\Sigma\right\|_{\text{HS}}}{\sqrt{n\left(1-2\varepsilon\right)}}\right)\right) \\ - 2\eta\left\|\sum_{i\in S\cap P}\eta_i\phi\left(x_i\right)\right\| - \sum_{j\in P\setminus S}\eta_j^2\end{aligned} \tag{107}$$

This completes the proof. ■

B.3 Proof of Theorem 16

Proof. First, we give the definiton of the Moreau stationary point.

$$\left\|\nabla \mathsf{M}_{\Phi_\lambda,\rho}\left(f_w\right)\right\|_{\mathcal{H}} = \left\|\frac{1}{\gamma}\left(f_w - \operatorname*{arg\,min}_{f_{\widehat{w}}\in\mathcal{K}}\left(\Phi\left(f_{\widehat{w}}\right) + \frac{1}{2\gamma}\left\|f_w - f_{\widehat{w}}\right\|_{\mathcal{H}}^2\right)\right)\right\|_{\mathcal{H}} = 0 \tag{108}$$

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

$$\lim_{\lambda\to\infty}\left(\Phi_\lambda\left(f_{\widehat{w}}\right)\right) < \lim_{\lambda\to\infty}\left(\Phi_\lambda\left(f_{\widetilde{w}}\right)\right) + \frac{1}{2\gamma}\left\|f_{\widetilde{w}} - f_{\widehat{w}}\right\|_{\mathcal{H}}^2 \tag{109}$$

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

$$\lim_{\lambda\to\infty}\left(\Phi_\lambda\left(f_{\widehat{w}}\right) - \Phi_\lambda\left(f_{w^*}\right)\right) \leq \frac{1}{2\gamma}\left\|f_{\widehat{w}} - f_{w^*}\right\|_{\mathcal{H}}^2 \tag{110}$$

We proceed by proof by contradiction. Assume $\|f_{\widehat{w}} - f_w^*\| > \eta$, then if $\Phi(f_{\widehat{w}}) - \Phi(f_w^*) > \frac{\eta^2}{2\gamma}$, then we will have $f_{\widehat{w}}$ is not a stationary point, which will imply $\|f_{\widehat{w}} - f_w^*\|_{\mathcal{H}} \leq \eta$. Therefore, we attempt to find the minimum value for η. From Lemma 15, we have we have

$$\lim_{\lambda\to\infty}\left(\Phi\left(f_w\right) - \Phi\left(f_w^*\right)\right) \geq \eta^2\left(n(1-2\varepsilon)\lambda_{\min}\left(\Sigma\right) - \left\|\Sigma\right\|_{\text{HS}}\sqrt{n(1-2\varepsilon)}\right) - 2\eta\left\|\sum_{i\in S\cap P}\eta_i\phi\left(x_i\right)\right\| - \sum_{j\in P\setminus S}\eta_j^2 \tag{111}$$

From the definition of stationary point, we have

$$\eta^2\left(n(1-2\varepsilon)\lambda_{\min}\left(\Sigma\right) - \left\|\Sigma\right\|_{\text{HS}}\sqrt{n(1-2\varepsilon)} - \beta\right) - 2\eta\left\|\sum_{i\in S\cap P}\eta_i\phi\left(x_i\right)\right\| - \sum_{j\in P\setminus S}\eta_j^2 \leq 0 \tag{112}$$

Therefore, when Equation (112) does not hold, we have a contradiction. It thus follows from upper bounding the positive solution of the quadratic equation,

$$\begin{aligned}\eta \leq & \left(\sum_{j\in P\setminus S}\eta_j^2\right)^{1/2}\left(n(1-2\varepsilon)\left(\lambda_{\min}\left(\Sigma\right)-O\left(\frac{C_k\left\|\Sigma\right\|_{\text{HS}}}{\sqrt{n\left(1-2\varepsilon\right)}}\right)\right)-\beta\right)^{-1/2} \\ & +2\left\|\sum_{i\in S\cap P}\eta_i\phi(x_i)\right\|_{\mathcal{H}}\left(n(1-2\varepsilon)\left(\lambda_{\min}\left(\Sigma\right)-O\left(\frac{C_k\left\|\Sigma\right\|_{\text{HS}}}{\sqrt{n\left(1-2\varepsilon\right)}}\right)\right)-\beta\right)^{-1}\end{aligned} \tag{113}$$

Then for some constant $c_1 \in (0,1)$, if $n \geq \frac{(1-2\varepsilon)(C_k\|\Sigma\|_{\text{HS}}+\beta)}{(1-c_1)\lambda_{\min}(\Sigma)}+\sqrt{\beta}$, we have

$$\eta \leq \sqrt{\frac{\sum_{j\in P\cap S}\eta_j^2}{c_1 n\lambda_{\min}\left(\Sigma\right)}}+\frac{2\left\|\sum_{i\in S\cap P}\eta_i\phi(x_i)\right\|_{\mathcal{H}}}{c_1 n\lambda_{\min}\left(\Sigma\right)} \tag{114}$$

This completes the proof. ■

B.4 Proof of Lemma 18

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 $\sigma(x) = \frac{1}{1+e^{-x}}$.

$$\begin{aligned}\left\|\nabla_{\mathsf{f}}g(t,f_w)\right\|_{\mathcal{H}} &= \left\|\frac{1}{np}\sum_{i=1}^{n}\mathbb{I}\{t\geq(1-y_i)\log(f_w(x_i))\}\left(y_i-\sigma\left(f_w(x_i)\right)\right)\cdot k(x_i,\cdot)\right\|_{\mathcal{H}} && (115)\\ &\leq \frac{1}{np}\sum_{i\in S}\left\|\left(y_i-\sigma\left(f_w(x_i)\right)\right)\cdot k(x_i,\cdot)\right\|_{\mathcal{H}} && (116)\\ &\overset{(a)}{\leq} \frac{1}{np}\sum_{i=1}^{n}\sqrt{k(x_i,x_i)} && (117)\end{aligned}$$

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

B.5 Proof of Lemma 19

We use the Hilbert Space norm of second derivative to bound β from above. Let S be the subquantile set.

$$\begin{aligned}\left\|\nabla_f^2 g(t,f_w)\right\|_{\mathcal{H}} &= \frac{1}{np}\sum_{i=1}^{n}\mathbb{I}\left\{t\geq(1-y_i)\log(f_w(x_i))\right\}\sigma\left(f_w(x_i)\right)\left(1-\sigma\left(f_w(x_i)\right)\right)k(x_i,x_i) && (118)\\ &\overset{\zeta_1}{\leq} \frac{1}{4p}\sum_{i=1}^{n}k(x_i,x_i) = \frac{1}{4p}\operatorname{Tr}\left(\mathrm{K}\right) && (119)\end{aligned}$$

(ζ_1) follows as for a scaler $\alpha \in [0,1]$, the maximum value of $\alpha(1-\alpha)$ is obtained at $\frac{1}{4}$. This completes the proof. ■

B.6 Proof of Lemma 20

Proof. Let S be the Subquantile set for f_w, then we have

$$\begin{aligned}\lim_{\lambda\to\infty}\Phi(f_w)-\Phi(f_w^*) &\geq -\sum_{i\in S}y_i\log\left(\sigma\left(f_w(x_i)\right)\right)+(1-y_i)\log\left(1-\sigma\left(f_w(x_i)\right)\right)\\ &\quad+\sum_{i\in P}y_i\log\left(\sigma\left(f_w^*(x_i)\right)\right)+(1-y_i)\log\left(1-\sigma\left(f_w^*(x_i)\right)\right) && (120)\\ &\overset{\zeta_1}{\geq} -\sum_{i\in S\cap P}y_i\log\left(\sigma\left(f_w(x_i)\right)\right)+(1-y_i)\log\left(1-\sigma\left(f_w(x_i)\right)\right)\\ &\quad+\sum_{i\in P}y_i\log\left(\sigma\left(f_w^*(x_i)\right)\right)+(1-y_i)\log\left(1-\sigma\left(f_w^*(x_i)\right)\right) && (121)\end{aligned}$$

$$\begin{aligned} &= \sum_{i\in S\cap P} y_i \underbrace{\log\left(\frac{\sigma\left(f_w^*(x_i)\right)}{\sigma\left(f_w(x_i)\right)}\right)}_{\textcircled{A_1}} + (1-y_i) \underbrace{\log\left(\frac{1-\sigma\left(f_w^*(x_i)\right)}{1-\sigma\left(f_w(x_i)\right)}\right)}_{\textcircled{A_2}} \\ &\quad + \sum_{P\setminus S} y_i \log\left(\sigma\left(f_w^*(x_i)\right)\right) + (1-y_i)\log\left(1-\sigma\left(f_w^*(x_i)\right)\right) \end{aligned} \tag{122}$$

(ζ_1) follows from the optimality of the subquantile set, S. First we bound $\textcircled{A_1}$.

$$\textcircled{A_1} \triangleq \log\left(\frac{\sigma\left(f_w^*(x_i)\right)}{\sigma\left(f_w(x_i)\right)}\right) = \log\left(\frac{1+e^{-f_w(x_i)}}{1+e^{-f_w^*(x_i)}}\right) = \log\left(\left(\frac{1+e^{-f_w(x_i)}}{1+e^{-f_w^*(x_i)}}\right)\left(\frac{e^{f_w^*(x_i)}}{e^{f_w^*(x_i)}}\right)\right) \tag{123}$$

$$= \log\left(\frac{e^{f_w^*(x_i)}+e^{(f_w^*-f_w)(x_i)}}{1+e^{f_w^*(x_i)}}\right) = \log\left(e^{f_w^*(x_i)}+e^{(f_w^*-f_w)(x_i)}\right) - \log\left(1+e^{f_w^*(x_i)}\right) \tag{124}$$

$$= \left(f_w^*-f_w\right)(x_i) + \log\left(1+e^{-f_w(x_i)}\right) - \log\left(1+e^{f_w^*(x_i)}\right) \tag{125}$$

$$\geq \left(f_w^*-f_w\right)(x_i) - \left(f_w+f_w^*\right)(x_i) - 0.693 \tag{126}$$

Next we will upper bound $\textcircled{A_2}$.

$$\textcircled{A_2} \triangleq \log\left(\frac{1-\sigma\left(f_w^*(x_i)\right)}{1-\sigma\left(f_w(x_i)\right)}\right) = \log\left(\frac{\sigma\left(-f_w^*(x_i)\right)}{\sigma\left(-f_w(x_i)\right)}\right) = \log\left(\frac{1+e^{f_w(x_i)}}{1+e^{f_w^*(x_i)}}\right) \tag{127}$$

$$= \log\left(\left(\frac{1+e^{f_w(x_i)}}{1+e^{f_w^*(x_i)}}\right)\left(\frac{e^{-f_w^*(x_i)}}{e^{-f_w^*(x_i)}}\right)\right) = \log\left(\frac{e^{-f_w^*(x_i)}+e^{(f_w-f_w^*)(x_i)}}{1+e^{-f_w^*(x_i)}}\right) \tag{128}$$

$$= \log\left(e^{-f_w^*(x_i)}+e^{(f_w-f_w^*)(x_i)}\right) - \log\left(1+e^{-f_w^*(x_i)}\right) \tag{129}$$

$$= \left(f_w-f_w^*\right)(x_i) + \log\left(1+e^{-f_w(x_i)}\right) - \log\left(1+e^{-f_w^*(x_i)}\right) \tag{130}$$

$$\geq \left(f_w-f_w^*\right)(x_i) + \left(f_w+f_w^*\right)(x_i) - 0.693 \tag{131}$$

Combining Equations (126) and (131), we have

$$\lim_{\lambda\to\infty} \Phi(f_w) - \Phi(f_w^*) \geq 2y_i\left(\left(f_w-f_w^*\right)(x_i) + \left(f_w+f_w^*\right)(x_i) - 1.386\right) \tag{132}$$

B.7 Proof of Lemma 21

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

$$\|\nabla_{\mathrm{W}} g(t,\mathrm{W})\|_{\mathcal{H}} = \tag{133}$$

■

C Proofs for Section 4

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

C.1 Proof of Theorem 24

From Theorem 31 in [20], we have

$$\mathbb{E}\left[\|\nabla\Phi_\beta\left(f_{\bar{w}}\right)\|_{\mathcal{H}}\right] \leq \left(2\cdot\frac{\left(\Phi_\beta\left(f_w^{(0)}\right)-\min\Phi(f_w)\right)+\beta L^2\gamma^2}{\gamma\sqrt{T+1}}\right)^{1/2}$$

$$\leq \left(2 \cdot \frac{\left(\min_{f_{\bar{w}} \in \mathcal{K}} \left\{ \Phi\left(f_{\bar{w}}\right) + \frac{1}{2} \left\| f_w^{(0)} - f_{\bar{w}} \right\|_{\mathcal{H}}^2 \right\} - \min_{f_w \in \mathcal{K}} \Phi(f_w) \right) + \beta L^2 \gamma^2}{\gamma \sqrt{T+1}} \right)^{1/2}$$

$$\leq \left(2 \cdot \frac{4R^2 + \sum_{i \in S} \eta_i^2 + \beta L^2 \gamma^2}{\gamma \sqrt{T+1}} \right)^{1/2}$$

If we are to choose $\gamma = \frac{\sqrt{T+1}}{\beta}$, then with probability $1 - \delta$ for $\delta \in (0,1)$, we have

$$\mathbb{E}\left[\left\| \nabla \Phi_\beta \left(f_{\bar{w}} \right) \right\|_{\mathcal{H}} \right] \leq \left(\frac{8R^2 + 2np(1+\delta)\sigma^2}{\beta} + 2L \right)^{1/2} \tag{134}$$

$$\leq \left(\frac{\beta \left(8R^2 + 2 \sum_{i \in S} \eta_i^2 \right)}{T+1} \right)^{1/2} + \sqrt{2}L \tag{135}$$

$$\leq \frac{2\sqrt{2\beta}R + \sqrt{2\beta}\sqrt{\sum_{i \in S} \eta_i^2}}{\sqrt{T+1}} + \sqrt{2}L \tag{136}$$

An application of a tail bound for the norm of a Gaussian vector gives us with probability $1-\delta$ for $\delta \in (0,1)$ the following

$$\mathbb{E}\left[\left\| \nabla \Phi_\beta \left(f_{\bar{w}} \right) \right\|_{\mathcal{H}} \right] \leq \frac{\sqrt{2\beta} \left(2R + \sigma\sqrt{np} + \sqrt{2\sqrt{np}\sigma \log\left(\frac{1}{2\delta}\right)} \right)}{\sqrt{T+1}} + \sqrt{2}L \tag{137}$$

Setting T to the the square of the numerator gives the desired result.

C.2 Proof of Theorem 25

We will start with the definition of the Moreau Envelope.

$$\bar{\Phi}_\lambda(f_w) = \frac{1}{\lambda} \left(f_w - \operatorname*{arg\,min}_{f_{\widehat{w}} \in \mathcal{K}} \left\{ \Phi(f_{\widehat{w}}) + \frac{1}{2\lambda} \left\| f_w - f_{\widehat{w}} \right\|_{\mathcal{H}}^2 \right\} \right) \tag{138}$$

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

$$\bar{\Phi}_\lambda \left(f_w^{(t+1)} \right) \leq \Phi\left(f_{\widehat{w}}^{(t)} \right) + \left\| f_w^{(t+1)} - f_{\widehat{w}}^{(t)} \right\|_{\mathcal{H}}^2 \tag{139}$$

$$= \Phi\left(f_{\widehat{w}}^{(t)} \right) + \left\| f_{\widehat{w}}^{(t)} - \Pi_{\mathcal{K}} \left(f_w^{(t)} - \alpha \left(\mu b^{(t)} \right) \right) \right\|_{\mathcal{H}}^2 \tag{140}$$

$$= \Phi\left(f_{\widehat{w}}^{(t)} \right) + \left\| f_{\widehat{w}}^{(t)} - \Pi_{\mathcal{K}} \left(f_w^{(t)} - \alpha \left(\mu \left(\mu b^{(t-1)} + \nabla_f \Phi \left(f_w^{(t-1)} \right) \right) \right) \right) \right\|_{\mathcal{H}}^2 \tag{141}$$

$$= \Phi\left(f_{\widehat{w}}^{(t)} \right) + \left\| f_{\widehat{w}}^{(t)} - \Pi_{\mathcal{K}} \left(f_w^{(t)} - \alpha \left(\sum_{i=k}^{t} \mu^k (1-\mu) \nabla_f f_w^{(t-k)} \right) \right) \right\|_{\mathcal{H}}^2 \tag{142}$$

$$\leq \Phi\left(f_{\widehat{w}}^{(t)} \right) + \left\| f_{\widehat{w}}^{(t)} - f_w^{(t)} - \alpha \left(\sum_{k=1}^{t} \mu^k (1-\mu) \nabla_f f_w^{(t-k)} \right) \right\|_{\mathcal{H}}^2 \tag{143}$$

C.3 Proof of Theorem 26

D Necessary Lemmas

Theorem 28 (Mercer, [25]). *If k is a positive definite and continuous kernel on a compact set Ω, th eoperator T has a countable set of eigenvalues $\lambda_1 \geq \lambda_2 \geq \ldots \geq 0$ and eigenfunctions $\{\varphi\}_{j \in \mathbb{N}}$ with $T\varphi_j = \lambda\varphi_j$ s.t.*

$\|\varphi_j\| = \lambda_j^{-1}$. *The kernel then admits the following representation,*

$$k(x,y) = \sum_{j=1}^{\infty} \lambda_j \varphi_j(x)\varphi(y) \tag{144}$$

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 [9]

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

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	
`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	
`Concrete`	8	1030	
`Drug`			

Table 7: Datasets for Linear Regression.