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 [LPMH21, RRM14]. 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 $\inf_f \mathbb{E}_{x\in\mathbb{P}}[\ell_f(x)]$, 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.

*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), [DKK+19, ADKS22, LBSS21, OZS20, FB81] as well as for linear regression [BJKK17, MGJK19]. Robust Statistics has been studied extensively [DK23] for problems such as high-dimensional mean estimation [PBR19, CDGS20] and Robust Covariance Estimation [CDGW19, FWZ18]. Recently, there has been an interest in solving robust machine learning problems by gradient descent [PSBR18, DKK+19]. Subquantile minimization aims to address the shortcomings of standard ERM in applications of noisy/corrupted data [KLA18, JZL+18]. In many real-world applications, the covariates have a non-linear dependence on labels [AMMIL12, Section 3.4]. In which case it is suitable to transform the covariates to a different space utilizing kernels [HSS08]. Therefore, in this paper we consider the problem of Robust Learning for Kernel Learning.

Definition 1. (**Huber ϵ-Contamination Model** [HR09]). 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, [LBSS21]. 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 [HYwL20] where there has been an analysis in the Binary Classification Task. We show Subquantile Minimization reduces to the same objective in [HYwL20]. 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.

The study of Kernel Learning in the Gaussian Design is quite popular, [CLKZ21, Dic16]. In [CLKZ21], the feature space, $\phi(x_i) \sim \mathcal{N}(0, \Sigma)$ where Σ is a diagonal matrix of dimension p, where p can be infinite. In this work, we adopt a similar framework, and with the power of Mercer's Theorem [Mer09], we are able to say $\mathrm{Tr}(\Sigma) < \infty$. We use this fact extensively in our infinite-dimensional concentration inequalities.

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. Assume Subquantile Minimization returns $f_{\widehat{w}}$ for $n \geq \frac{(1-2\varepsilon)(C_k\|\Sigma\|_{\mathrm{op}}+\beta)}{(1-c_1)\lambda_{\min}(\Sigma)} + \sqrt{\beta}$ for a constant $c_1 \in (0,1)$ such that for*
Kernelized Regression:

$$\mathop{\mathbb{E}}_{\mathcal{D}\sim\hat{\mathbb{P}}} \|f_{\widehat{w}} - f_w^*\|_{\mathcal{H}} \leq O\left(\frac{\gamma\sigma}{\sqrt{\lambda_{\min}(\Sigma)}}\right) \tag{1}$$

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

$$\mathop{\mathbb{E}}_{\mathcal{D}\sim\hat{\mathbb{P}}} \|f_{\widehat{w}} - f_w^*\|_{\mathcal{H}} \leq O\left(\frac{\sqrt{\mathrm{Tr}(\Sigma)} + \sqrt{Q_k}}{\sqrt{n(1-2\varepsilon)\lambda_{\min}(\Sigma)}}\right) \tag{2}$$

Kernel Multi-Class Classification:

$$\mathop{\mathbb{E}}_{\mathcal{D}\sim\hat{\mathbb{P}}} \left\|f_{\hat{\mathrm{W}}} - f_{\mathrm{W}}^*\right\|_{\mathcal{H}} \leq O(\gamma) \tag{3}$$

1.1 Related Work

The idea of iterative thresholding algorithms for robust learning tasks dates back to 1806 by Legendre [Leg06]. From the popularity of Machine Learning, numerous algorithms have been developed in this idealogy. Therefore, we will dedicate this section to reviewing such works and to make clear our contributions to the iterative thresholding literature.

Robust Regression via Hard Thresholding [BJK15]. Bhatia et al. consider robust linear regression by considering an active set S, which contains the points with the lowest error. This set is updated each iteration in conjunction with either a full solve (Torrent-FC) or a gradient iteration (Torrent-GD). Torrent-GD is an unconstrained variant of our algorithm. The main limitation of this work is that only

the case of label corruption is considered. We pick up the result of Theorem 9 and Theorem 11 in [BJK15] (up to constants) for linear regression with and without feature corruption, which is one of our key contributions. Learning with bad training data via iterative trimmed loss minimization [SS19a]. This work considers optimizing over the bottom-k errors by choosing the αn points with smallest error and then updating the model from these αn. This general model is the same as ours. Theoretically, this work considers only general linear models. Experimentally, this work considers more general machine learning models such as GANS. Trimmed Maximum Likelihood Estimation for Robust Generalized Linear Model [ADKS22]. This work studies a different class of generalized linear models. Interestingly, they show for Gaussian Regression the iterative trimmed maximum likelihood estimator is able to achieve near minimax optimal error. This work does not consider feature corruption and primarily focuses on the covariates sampled with Gaussian Design from Identity covariance.
Sum of Ranked Range Loss for Supervised Learning [HYwL20]. Hu et al. proposed learning over the bottom k losses, this is an alternative formulation of our algorithm. This is an extension of previous work studying the learning of the top k losses, [FLYH17]. They solve their optimization problem with difference of sums convex solvers. This work considers only the classification task and does not give rigorous error bounds. Subsequent work on analyzing the middle k losses is analyzed in [HYW$^+$23].
The iterative trimmed loss framework with batch Stochastic Gradient Descent (SGD) is analyzed in [SS19b]. They experimentally test their design in deep learning applications such as image classification and Generative Adversarial Networks (GANs).

1.2 Contributions

We will now state our main contributions clearly.

1. We provide a novel theoretical framework using the Moreau Envelope for analyzing the iterative trimmed estimator for machine learning tasks.

2. We provide rigorous error bounds for subquantile minimization in the kernel regression, kernel binary classification, and kernel multi-class classification. Furthermore, we provide our bounds for both label and feature corruption with a general Gaussian Design.

3. We perform experiments on state-of-the-art matrices and show the effectiveness of our algorithm compared to other robust learning procedures. Furthermore, the experiments support the theory.

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.

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

The p-subquantile of the risk is then given by

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

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

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

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 [RHL$^+$20]. 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

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

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

We provide an algorithm for Subquantile Minimization of the ridge regression and classification kernel learning 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: λ

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$(w^{(k)}, \mathrm{X})$

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

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 $\ell(x_i; f_w, y_i) \leq t$

$\mathrm{S} \leftarrow \left(x_1^\top \quad \ldots \quad x_{np}^\top\right)^\top$

return S

3 Theory

To consider theoretical guarantees of Subquantile Minimization, we first analyze the inner and outer optimization problems. We first analyze kernel learning in the presence of corrupted data. Next, we provide error bounds for the two most important kernel learning problems, kernel ridge regression, and kernel classification. Now we will give our first result regarding kernel learning in the Huber ϵ-contamination model. Now we will analyze the two-step minimax optimization steps described in Equations (7) and (8).

Lemma 3. *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 $\hat{\nu}_{np} \in \arg\max_{t \in \mathbb{R}} g(t, w)$.*

Proof is given in **??**.

Interpretation 4. From Lemma 3, we see the t will be greater than or equal to the errors of exactly np points. Thus, we are continuously updating over the np minimum errors.

Lemma 5. *Let $\hat{\nu}_i \triangleq f(x_i; w, y_i)$ 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 3, 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 is given in Appendix B.2.

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,

$$\zeta_\lambda(x) = \frac{1}{\lambda} \log\left(1 + e^{\lambda x}\right) \tag{9}$$

We then have the approximation of g as

$$\tilde{g}_\lambda(t, f_w) \triangleq t - \sum_{i=1}^{n} \zeta_\lambda\left(t - \ell\left(f_w; x_i, y_i\right)\right) \tag{10}$$

$$= t - \frac{1}{np} \sum_{i=1}^{n} \frac{1}{\lambda} \log\left(1 + \exp\left(\lambda\left(t - \ell\left(f_w; x_i, y_i\right)\right)\right)\right) \tag{11}$$

More details on the Softplus Approximation such as exact computations can be found in Appendix B.3. We can then calculate the Lipschitz constant of the approximation function with respect to f_w.

Lemma 6 (Lipschitz continuous gradient)**.** *Let $f_w, f_{\tilde{w}} \in \mathcal{K}$, then we have for any $\lambda > 0$,*

$$\left|\nabla_f \tilde{g}_\lambda(t, f_w) - \nabla_f \tilde{g}_\lambda(t, f_{\tilde{w}})\right| \leq \beta \left\|f_w - f_{\tilde{w}}\right\|_{\mathcal{H}} \tag{12}$$

where

$$\beta = \frac{1}{np} \sum_{i=1}^{n} \left\|\nabla_f^2 \ell\left(f_w; x_i, y_i\right)\right\|_{\mathrm{op}} \tag{13}$$

and β has no dependence on λ.

Proof is in Appendix B.4. This lemma is important as it states the β-smoothness constant is independent of the approximation term, λ. We will use this lemma in the next section by pushing $\lambda \to \infty$ and analyzing the resultant function.

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 7. (**Moreau Envelope on closed, convex set**, [Mor65]). 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:

$$\mathsf{M}_{\lambda\ell}(f_w) \triangleq \inf_{f_{\hat{w}} \in \mathcal{K}} \left\{\ell(f_{\hat{w}}) + \frac{1}{2\rho}\left\|f_w - f_{\hat{w}}\right\|_{\mathcal{H}}^2\right\} \tag{14}$$

Definition 8. 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 9 (First Order Stationary Point)**.** Let $f_{\hat{w}}$ be a first-order stationary point, then for any $f_w \in \mathcal{K}$, it follows

$$\left\langle \nabla_f g\left(f_{\hat{w}}\right), f_w - f_{\hat{w}}\right\rangle_{\mathcal{H}} \geq 0 \qquad \forall\, f_w \in \mathcal{K} \tag{15}$$

Definition 10 (Stationary Point of Moreau Envelope)**.** A point $f_{\hat{w}}$ is a stationary point of the Moreau Envelope defined in Definition 7 of Φ defined in Definition 8 if

$$f_{\hat{w}} = \operatorname*{arg\,inf}_{f_w \in \mathcal{K}} \left\{\Phi_\lambda\left(f_w\right) + \frac{1}{2\rho}\left\|f_w - f_{\hat{w}}\right\|_{\mathcal{H}}^2\right\} \tag{16}$$

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. $\|f_{\hat{w}} - f_w^*\|_{\mathcal{H}}$ is small.

Lemma 11 (Lower bound on distance from stationary point and optimal point)**.** *Let Φ_λ be defined as in Definition 8, then if $f_{\hat{w}}$ is a stationary point as defined in Definition 10 and $g(t, f_w)$ has β-Lipschitz Gradient, then*

$$\lim_{\lambda \to \infty} \left(\Phi_\lambda\left(f_{\hat{w}}\right) - \Phi_\lambda\left(f_w^*\right)\right) \leq \beta \left\|f_{\hat{w}} - f_w^*\right\|_{\mathcal{H}}^2 \tag{17}$$

We can now upper bound $\|f_{\hat{w}} - f_w^*\|_{\mathcal{H}}$. 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 11. This bound is different for each of the loss functions, so we must upper bound $\|f_{\hat{w}} - f_w^*\|_{\mathcal{H}}$ 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 $(x_i, y_i) \in \mathcal{D}$ is given by the following equation

$$\ell\left(f_w; x_i, y_i,\right) = \left(f_w(x_i) - y_i\right)^2 \tag{18}$$

It is important to note that β is upper bounded as

$$\beta = \frac{2}{np}\operatorname{Tr}(\mathrm{K}) \leq \frac{2}{np}(n(1-\varepsilon)\max_{i \in P} k(x_i, x_i) + n\varepsilon \max_{j \in Q} k(x_j, x_j) = 2p^{-1}((1-\varepsilon)P_k + \varepsilon Q_k) \tag{19}$$

which is independent of n. For our bounds, to be useful, we require the *Strong Projection Property.*

Definition 12 (Strong Projection Property)**.** Let f_w^* be the optimal function for the uncorrupted dataset, $\mathbb{P}$. Then, we have for a finite m there exists positive constants c_7 and c_8 such that the following holds,

$$\|\mathrm{Proj}_m f_w^*\|_{\mathcal{H}} \geq c_8 \left(\sum_{i \in P} k_m(x_i, x_i)\right) \left(\alpha LT + \left\|f_w^{(0)} - f_w^*\right\|_{\mathcal{H}}\right) + O\left(nT \operatorname{Tr}(\Sigma_m)\left(\sigma^2 \log(n) + \frac{\sigma^3}{\delta}\right)\right)$$
$$+ c_7 R \left(\sum_{i \in Q} k_m(x_i, x_i)\right) + c_7 \|y_Q\|_1 \left(\sum_{i \in Q} \sqrt{k_m(x_i, x_i)}\right)$$

The *Strong Projection Property* is important as $\lambda_{\min}(\Sigma)$ is not well defined for an infinite dimensional feature space, e.g. Gaussian Kernel. The implication of the *Strong Projection Property* is given in the following lemma.

Lemma 13 (Strong Projection Property Implication)**.** *Let $f_w^{(t)}$ for $t \in [T]$ be iterates from* **??** *3. Then, it follows for a $m \in \mathbb{N}$ and a constant $C > 0$ then with probability $1 - \delta$ for $\delta \in (0, 1)$,*

$$\langle \Sigma, (f_w - f_w^*) \otimes (f_w - f_w^*) \rangle_{\mathrm{HS}} \geq c_4 \lambda_m \tag{20}$$

Proof is given in Appendix C.2. We will discuss the implication of Lemma 13 after stating the following theorem.

Theorem 14 (Stationary Point for Kernelized Regression is Good)**.** *Let $f_{\hat{w}}$ be a stationary point defined in Definition 10 for the function Φ defined in Definition 8. Then for a constant $c_1 \in (0, 1)$, if $n \geq \frac{8 \operatorname{Tr}(\Sigma)^2}{\lambda_{\min}(\Sigma)(1-c_1)^2(1-2\varepsilon)} + \frac{8\beta}{(1-c_1)^2(1-2\varepsilon)}$,*

$$\mathop{\mathbb{E}}_{\mathcal{D} \sim \hat{\mathbb{P}}} \|f_{\hat{w}} - f_{w^*}\|_{\mathcal{H}} \leq \sqrt{\frac{\gamma\sigma}{c_1 \lambda_{\min}(\Sigma)}} + \frac{O\left(\sigma\sqrt{\gamma \log(n(1-2\varepsilon)) \operatorname{Tr}(\Sigma)}\right)}{c_1\sqrt{n(1-2\varepsilon)}\, c \lambda_{\min}(\Sigma)} \tag{21}$$

where β is the Lipschitz Gradient Constant given in Lemma 27.

In Theorem 14, we have an upper bound on the expected distance from a stationary point to the optimal point over the distance of the dataset. The numerator of the second term grows in $O(\sqrt{\log(n)})$ and the denominator grows in $O(\sqrt{n})$ 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 [SST+18]. 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. The effect of Lemma 13 can be seen in the denominator of both terms. Instead of $\lambda_{\min}(\Sigma)$ we have $c_4\lambda_m$ for a finite m. This differene will be clear in the following corollary, where we utilize the theory develoepd for kernelized regression to imply a result for regularized linear regression.

Corollary 15 (Linear Regression Expected Erorr Bound)**.** *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* **??** *3, we have the following error bounds for robust kernelized linear regression. Given sufficient data*

$$\mathbb{E}\left\|w^{(T)} - w^*\right\|_2 \leq O\left(\frac{\gamma\sigma}{\sqrt{\lambda_{\min}(\Sigma)}}\right) \tag{22}$$

Proof given in Appendix C.4. It is important to note in all our bounds, $\gamma \leq \sqrt{\frac{\varepsilon}{1-2\varepsilon}}$ is a theoretical worst case bound when the Subquantile contains the minimum possible number of uncorrupted points. In other words, we have $\gamma \triangleq \frac{|P \setminus S|}{|S \cap P|} \leq \frac{n\varepsilon}{n(1-2\varepsilon)} = \frac{\varepsilon}{1-2\varepsilon}$. So, as $|S \cap P|$ increases, we have a better error bound as $|P \setminus S|$ decreases. As is typical in the robust statistics literature, we make no assumptions on the distribution of the corrupted data so we cannot say anything about $|S \cap P|$. We will have γ decreases if stationary points give high error for corrupt points as our optimization procedure moves toward a stationary point.

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)

$$\ell\left(x_i, y_i; f_w\right) = -\left(y_i \log\left(\sigma\left(f_w(x_i)\right)\right) - (1-y_i)\log\left(1-\sigma\left(f_w(x_i)\right)\right)\right) \tag{23}$$

Theorem 16. *[A stationary point is good for kernel binary classification] Let $f_{\hat{w}}$ be a statoinary point defined in Definition 9 for the function Φ defined in Definition 8. Then for a constant $c_4 \in (0,1)$, if $n \geq \frac{4\,\mathrm{Tr}(\Sigma)}{\lambda_{\min}(\Sigma)(1-2\varepsilon)(1-c_4)}$, then in expectation over the dataset distribution,*

$$\mathop{\mathbb{E}}_{\mathcal{D}\sim\hat{\mathbb{P}}} \|f_{\hat{w}} - f_w^*\|_{\mathcal{H}} \leq O\left(\frac{\sqrt{\mathrm{Tr}(\Sigma)} + \sqrt{Q_k}}{\sqrt{n(1-2\varepsilon)}\exp\left(-R\left(\mathrm{Tr}\left(\Sigma\right) + \log n\right)\right)\lambda_{\min}(\Sigma)}\right) \tag{24}$$

Proof is given in Appendix D.2. This result although shows consistency, i.e. when $n \to \infty$, then we have in expectation $\|f_w - f_w^*\| \to 0$, however it does crucially rely on the fact that Q_k is bounded, and in general when n is not large, a large Q_k does affect the error bounds.

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

$$\ell\left(x_i, y_i; \mathrm{W}\right) = -\sum_{j=1}^{|\mathcal{Y}|} \mathbb{I}\left\{j = y_i\right\} \log\left(\frac{\exp\left(f_{\mathrm{W}_j}(x_i)\right)}{\sum_{k=1}^{|\mathcal{Y}|}\exp\left(f_{\mathrm{W}_k}(x_i)\right)}\right) \tag{25}$$

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 17 ([Roc70, DD19])**.** *Assume the function Φ is β-weakly convex. Let $\lambda < \frac{1}{\beta}$, and let $f_{\hat{w}} = \arg\min_{f_w \in \mathcal{K}}(\Phi\left(f_w\right) + \frac{1}{2\lambda}\|f_w - f_{\hat{w}}\|_{\mathcal{H}}^2)$, then $\|\nabla\Phi_\lambda(f_w)\|_{\mathcal{H}} \leq \epsilon$ implies:*

$$\|f_{\hat{w}} - f_w\|_{\mathcal{H}} = \lambda\epsilon \quad \textit{and} \quad \min_{g \in \partial\Phi(f_{\hat{w}}) + \partial\mathcal{I}_{\mathcal{K}}(f_{\hat{w}})} \|g\|_{\mathcal{H}} \leq \epsilon \tag{26}$$

With Lemma 17 in hand, it suffices to show that $\|\nabla\Phi_\lambda(f_w)\|_{\mathcal{H}}$ is small, as it then follows that f_w is close to a stationary point of the Moreau Envelope. It has been shown in optimization theory that utilizing standard gradient descent, $\|\nabla\Phi_\lambda(f_w)\|_{\mathcal{H}}$ decreases at a rate of $O(T^{-1/2})$. The exact theorem and proof can be seen in [DD19] and a proof where the maximimum of the inner problem can be calculated to within $(1+\epsilon)$ can be seen in [JNJ20] and [CDGS20].

4 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 [GB10].

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.

4.1 Linear Regression

In this section, we give experimental results for datasets using the linear kernel. This section will serve as a comparison to the state of the art algorithms developed specifically for the Robust Linear Problem. In particular, we compare against Kernel Ridge Regression (KRR) implemented in the sklearn package [PVG+11], Consistent Robust Regression (CRR) [BJKK17], Globally-convergent iteratively reweighted least squares (STIR) [MGJK19]. We also compare with several robust meta-algorithms, i.e. algorithms which work for multiple robust learning tasks, e.g classification and regression. We compare with SEVER [DKK+19] and Tilted Empirical Risk Minimization [LBSS21].

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: λ

Output: Trained Parameters: $f_w^{(T)}$

$w_i^{(0)} \leftarrow \text{Unif}\left[-\sqrt{\frac{6}{n}}, \sqrt{\frac{6}{n}}\right], \forall i \in [n]$ ▷ Initialize Weights Randomly

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

  $S^{(k)} \leftarrow \textsc{Subquantile}(f_w^{(k)}, \mathrm{X})$ ▷ **??** 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

  $\nabla_f g\left(t^{(k+1)}, f_w^{(k)}\right) \leftarrow \sum_{i \in S^{(k)}} \left(\text{softmax}\left(f_{\mathrm{W}}^{(k)}(x_i)\right) - y_i\right) \cdot k(x_i, \cdot)$ ▷ Multi-Class 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 $f_w^{(t)}$

Algorithms	Test RMSE							
	Concrete [Yeh07]		Wine Quality [CR09]		Boston Housing [DG17]		Drug [OSB+18]	
	$\epsilon = 0.2(\downarrow)$	$\epsilon = 0.4(\downarrow)$	$\epsilon = 0.2(\downarrow)$	$\epsilon = 0.4(\downarrow)$	$\epsilon = 0.2(\downarrow)$	$\epsilon = 0.4(\downarrow)$	$\epsilon = 0.2(\downarrow)$	$\epsilon = 0.4(\downarrow)$
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	$\underline{0.533}_{(0.0347)}$	$\underline{0.592}_{(0.0548)}$	$\underline{0.915}_{(0.0343)}$	$\underline{0.841}_{(0.0413)}$	$\underline{0.526}_{(0.0287)}$	$\underline{0.720}_{(0.1147)}$	$\underline{1.172}_{(0.0542)}$	$\underline{1.215}_{(0.0536)}$
Subquantile	$\mathbf{0.396}_{(0.0216)}$	$\mathbf{0.442}_{(0.0468)}$	$\mathbf{0.808}_{(0.0389)}$	$\mathbf{0.827}_{(0.0216)}$	$\mathbf{0.446}_{(0.1230)}$	$\mathbf{0.456}_{(0.1055)}$	$\mathbf{1.074}_{(0.0378)}$	$\mathbf{1.132}_{(0.0892)}$
Oracle ERM	∞	∞	∞	∞	∞	∞	∞	∞

Table 1: Boston Housing, Concrete Data, Wine Quality, and Drug and Polynomial Synthetic Dataset. $R = 10000$ for all datasets. 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.

4.2 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.

4.3 Kernel Multi-Class Classification

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

Results. We will clearly state our main findings.

- **Label Noise vs. Label and Feature Noise**. As suggested by our developed theory, for linear regression or using unbounded kernels, a large multiplicative term increases β and therefore requires more gradient descent iterations to achieve the same distance from a Moreau stationary point. Therefore, from simply increasing the number of gradient descent iterations, we are able to achieve similar RMSE in practice. This happens because the distance from a stationary point and the optimal is not affected by feature noise. This is one of the strengths of our theoretical analysis.

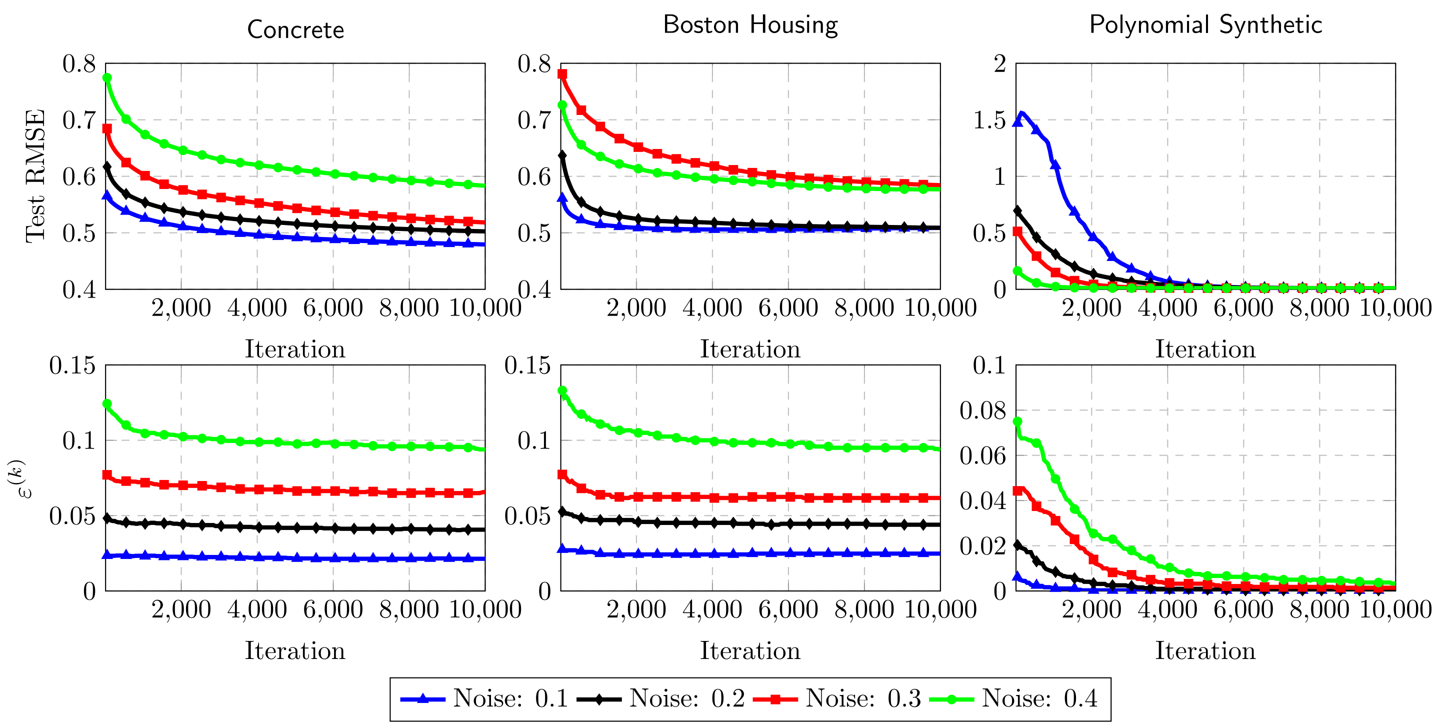

Figure 1: Test RMSE over the iterations in Concrete, Boston Housing, and Polynomial Datasets for SUBQUANTILE at different noise levels

Algorithms	Test RMSE							
	Boston Housing [DG17]		Wine Quality [CR09]		Concrete [Yeh07]		Drug [OSB+18]	
	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	$\underline{0.562}_{(0.0626)}$	$78.878_{(8.0164)}$	$0.828_{(0.0293)}$	$58.352_{(4.6700)}$	$\underline{0.684}_{(0.0245)}$	$76.555_{(4.5927)}$	$1.721_{(0.1520)}$	$144.975_{(5.4953)}$
SEVER	$0.601_{(0.0979)}$	$5.980_{(8.2603)}$	$\mathbf{0.814}_{(0.0207)}$	$9.065_{(13.7632)}$	$\underline{0.684}_{(0.0438)}$	$4.119_{(8.2436)}$	$1.469_{(0.1162)}$	$156.043_{(4.5543)}$
TERM	$0.608_{(0.1357)}$	$\underline{0.569}_{(0.0620)}$	$0.840_{(0.0563)}$	$\underline{0.827}_{(0.0255)}$	$0.780_{(0.0734)}$	$\underline{0.808}_{(0.0726)}$	$\underline{1.185}_{(0.1077)}$	$\mathbf{1.147}_{(0.1258)}$
SUBQUANTILE	$\mathbf{0.503}_{(0.0470)}$	$\mathbf{0.548}^{*}_{(0.0286)}$	$\mathbf{0.813}_{(\mathbf{0.0357})}$	$\mathbf{0.821}_{(0.0305)}$	$\mathbf{0.632}_{(0.0275)}$	$\mathbf{0.703}_{(0.0427)}$	$\mathbf{1.074}_{(0.1848)}$	$\underline{2.413}_{(0.6737)}$
Oracle 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 $y_{\text{noisy}} = 10000y_{\text{original}}$. * As indicated by the theory, when encountering feature noise, we require more gradient descent iterations to achieve the same bound between the returned point and the stationary point. Therefore, we train the label noise perturbed dataset for 10000 iterations, and the feature noise perturbed dataset for 100000 iterations.

- **Error vs. ε.** We find approximately linear increase in the error with increasing ε. This can be seen in the γ term, which is upper bounded $\sqrt{\varepsilon/(1-2\varepsilon)}$. When $\varepsilon \to 0.5$, the denominator approaches 0 and therefore our worst case bound increases.
- **Kernel.** Our error bounds are stronger when the dimension of the kernel is lower, i.e. we need more data to obtain the same error bounds. However, in practice, we find many datasets are better approximated by polynomial or RBF kernels, and therefore the γ term is significantly lower.

Algorithms	Test Accuracy							
	Heart Disease [JD88]				Breast Cancer [WS95]			
	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	$\underline{0.793}_{(0.0422)}$	$\underline{0.695}_{(0.0636)}$	$0.784_{(0.0432)}$	$\mathbf{0.816}_{(0.0562)}$	$0.904_{(0.0356)}$	$0.575_{(0.1456)}$	$0.956_{(0.0164)}$	$\underline{0.974}_{(0.0062)}$
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	$\mathbf{0.803}_{(0.0293)}$	$\mathbf{0.790}_{(0.0350)}$	$\mathbf{0.833}_{(0.0318)}$	$\underline{0.807}_{(0.0468)}$	$\mathbf{0.928}_{(0.0129)}$	$\mathbf{0.916}_{(0.0185)}$	$\underline{0.972}_{(0.0187)}$	$0.963_{(0.0170)}$
Oracle 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.

Algorithms	Test Accuracy							
	Iris [Fis88]		Glass [Ger87]		Wine [AF91]		Satimage [Sri93]	
	$\epsilon = 0.2(\uparrow)$	$\epsilon = 0.4(\uparrow)$	$\epsilon = 0.2(\uparrow)$	$\epsilon = 0.4(\uparrow)$	$\epsilon = 0.2(\uparrow)$	$\epsilon = 0.4(\uparrow)$	$\epsilon = 0.2(\uparrow)$	$\epsilon = 0.4(\uparrow)$
SVC	$0.977_{(0.0300)}$	$0.757_{(0.1155)}$	$0.553_{(0.0969)}$	$0.435_{(0.0721)}$	$0.928_{(0.0484)}$	$0.678_{(0.1368)}$	$0.882_{(0.0056)}$	$0.732_{(0.0168)}$
TERM	∞	∞	∞	∞	∞	∞	∞	∞
SEVER	∞	∞	∞	∞	∞	∞	∞	∞
SUBQUANTILE	$\mathbf{0.987}_{(0.0163)}$	$\mathbf{0.820}_{(0.1720)}$	$\mathbf{0.656}_{(0.0804)}$	$\mathbf{0.598}_{(0.0889)}$	$\mathbf{0.975}_{(0.0262)}$	$\mathbf{0.867}_{(0.1971)}$	$\mathbf{0.899}_{(0.0076)}$	$\mathbf{0.861}_{(0.0297)}$
Oracle ERM	∞	∞	∞	∞	∞	∞	∞	∞

Table 4: Iris ($R = 1$), Glass ($R = 10$), Wine ($R = 100$), and Satimage ($R = 10000$) Datasets. Label Noise is a randomly chosen incorrect label. Feature Noise: $y_{\text{noise}} = 10000y_{\text{original}}$ and $x_{\text{noise}} = 100x_{\text{original}}$. The Radial Basis Function is used in all experiments.

5 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. Our theoretical analysis is novel in that it utilizes the Moreau Envelope from a min-max formulation of the iterative thresholding algorithm.

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.

Future Work. The analysis of Subquantile Minimization can be extended to neural networks as kernel learning can be seen as a one-layer network. This generalization will be appear in subsequent work. Another interesting direction work in optimization is for accelerated methods for optimizing non-convex concave min-max problems with a maximization oracle. The current theory analyzes standard gradient descent for the minimization. Ideas such as Momentum and Nesterov Acceleration in conjunction with the maximum oracle are interesting and can be analyzed in future work.

References

[ADKS22] 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, 2

[AF91] Stefan Aeberhard and M. Forina. Wine. UCI Machine Learning Repository, 1991. DOI: https://doi.org/10.24432/C5PC7J. 9

[AMMIL12] Yaser S Abu-Mostafa, Malik Magdon-Ismail, and Hsuan-Tien Lin. *Learning from data*, volume 4. AMLBook New York, 2012. 1

[BJK15] Kush Bhatia, Prateek Jain, and Purushottam Kar. Robust regression via hard thresholding. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015. 1, 2

[BJKK17] 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, 6

[CDGS20] 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, 6

[CDGW19] Yu Cheng, Ilias Diakonikolas, Rong Ge, and David P. Woodruff. Faster algorithms for high-dimensional robust covariance estimation. In Alina Beygelzimer and Daniel Hsu, editors, *Proceedings of the Thirty-Second Conference on Learning Theory*, volume 99 of *Proceedings of Machine Learning Research*, pages 727–757. PMLR, 25–28 Jun 2019. 1

[Che52] Herman Chernoff. A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations. *The Annals of Mathematical Statistics*, pages 493–507, 1952. 15

[CLKZ21] 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. 1, 23

[CR09] Cerdeira A. Almeida F. Matos T. Cortez, Paulo and J. Reis. Wine Quality. UCI Machine Learning Repository, 2009. DOI: https://doi.org/10.24432/C56S3T. 7, 8

[DD19] Damek Davis and Dmitriy Drusvyatskiy. Stochastic model-based minimization of weakly convex functions. *SIAM Journal on Optimization*, 29(1):207–239, 2019. 6

[DG17] Dheeru Dua and Casey" Graff. UCI machine learning repository, 2017. 7, 8

[Dic16] Lee H Dicker. Ridge regression and asymptotic minimax estimation over spheres of growing dimension. 2016. 1

[DK23] Ilias Diakonikolas and Daniel M Kane. *Algorithmic high-dimensional robust statistics*. Cambridge University Press, 2023. 1

[DKK+19] 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, 6

[FB81] 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

[Fis88] R. A. Fisher. Iris. UCI Machine Learning Repository, 1988. DOI: https://doi.org/10.24432/C56C76. 9

[FLYH17] Yanbo Fan, Siwei Lyu, Yiming Ying, and Baogang Hu. Learning with average top-k loss. 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. 2

[FWZ18] 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

[GB10] 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. 6

[Ger87] B. German. Glass Identification. UCI Machine Learning Repository, 1987. DOI: https://doi.org/10.24432/C5WW2P. 9

[GP17] Bolin Gao and Lacra Pavel. On the properties of the softmax function with application in game theory and reinforcement learning. *arXiv preprint arXiv:1704.00805*, 2017. 28

[Gre13a] 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. 16

[Gre13b] 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. 20

[H8̈9] O. Hölder. Ueber einen mittelwerthabsatz. *Nachrichten von der Königl. Gesellschaft der Wissenschaften und der Georg-Augusts-Universität zu Göttingen*, 1889:38–47, 1889. 28

[HR09] Peter J. Huber and Elvezio. Ronchetti. *Robust statistics*. Wiley series in probability and statistics. Wiley, Hoboken, N.J., 2nd ed. edition, 2009. 1

[HSS08] Thomas Hofmann, Bernhard Schölkopf, and Alexander J. Smola. Kernel methods in machine learning. *The Annals of Statistics*, 36(3):1171 – 1220, 2008. 1

[HYW+23] Shu Hu, Zhenhuan Yang, Xin Wang, Yiming Ying, and Siwei Lyu. Outlier robust adversarial training. *arXiv preprint arXiv:2309.05145*, 2023. 2

[HYwL20] Shu Hu, Yiming Ying, xin wang, and Siwei Lyu. Learning by minimizing the sum of ranked range. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 21013–21023. Curran Associates, Inc., 2020. 1, 2

[JD88] Steinbrunn William Pfisterer Matthias Janosi, Andras and Robert Detrano. Heart Disease. UCI Machine Learning Repository, 1988. DOI: https://doi.org/10.24432/C52P4X. 9

[JNJ20] 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. 6

[JZL+18] 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

[KLA18] Ashish Khetan, Zachary C. Lipton, and Anima Anandkumar. Learning from noisy singly-labeled data. In *International Conference on Learning Representations*, 2018. 1

[LBSS21] Tian Li, Ahmad Beirami, Maziar Sanjabi, and Virginia Smith. Tilted empirical risk minimization. In *International Conference on Learning Representations*, 2021. 1, 6

[Leg06] Adrien M Legendre. *Nouvelles methodes pour la determination des orbites des comtes: avec un supplement contenant divers perfectionnemens de ces methodes et leur application aux deux cometes de 1805*. Courcier, 1806. 1

[LPMH21] 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.

[Mer09] 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. 1

[MGJK19] 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, 6

[Mor65] Jean-Jacques Moreau. Proximité et dualité dans un espace hilbertien. *Bulletin de la Société mathématique de France*, 93:273–299, 1965. 4

[OSB+18] Ivan Olier, Noureddin Sadawi, G Richard Bickerton, Joaquin Vanschoren, Crina Grosan, Larisa Soldatova, and Ross D King. Meta-qsar: a large-scale application of meta-learning to drug design and discovery. *Machine Learning*, 107:285–311, 2018. 7, 8

[OZS20] 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

[PBR19] Adarsh Prasad, Sivaraman Balakrishnan, and Pradeep Ravikumar. A unified approach to robust mean estimation. *arXiv preprint arXiv:1907.00927*, 2019. 1

[PSBR18] Adarsh Prasad, Arun Sai Suggala, Sivaraman Balakrishnan, and Pradeep Ravikumar. Robust estimation via robust gradient estimation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 82, 2018. 1

[PVG+11] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830, 2011. 6

[RH23] Philippe Rigollet and Jan-Christian Hütter. High-dimensional statistics. *arXiv preprint arXiv:2310.19244*, 2023. 17

[RHL+20] 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

[Roc70] Ralph Tyrell Rockafellar. *Convex Analysis*. Princeton University Press, Princeton, 1970. 6

[RRM14] 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.

[Sri93] Ashwin Srinivasan. Statlog (Landsat Satellite). UCI Machine Learning Repository, 1993. DOI: https://doi.org/10.24432/C55887. 9

[SS16] Gabriele Santin and Robert Schaback. Approximation of eigenfunctions in kernel-based spaces. *Advances in Computational Mathematics*, 42:973–993, 2016. 23

[SS19a] Yanyao Shen and Sujay Sanghavi. Learning with bad training data via iterative trimmed loss minimization. In *International Conference on Machine Learning*, pages 5739–5748. PMLR, 2019. 2

[SS19b] Yanyao Shen and Sujay Sanghavi. Learning with bad training data via iterative trimmed loss minimization. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 5739–5748. PMLR, 09–15 Jun 2019. 2

[SST+18] 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. 5

[Wey12] Hermann Weyl. Das asymptotische verteilungsgesetz der eigenwerte linearer partieller differentialgleichungen (mit einer anwendung auf die theorie der hohlraumstrahlung). *Mathematische Annalen*, 71(4):441–479, 1912. 19

[WS95] Mangasarian Olvi Street Nick Wolberg, William and W. Street. Breast Cancer Wisconsin (Diagnostic). UCI Machine Learning Repository, 1995. DOI: https://doi.org/10.24432/C5DW2B. 9

[Yeh07] I-Cheng Yeh. Concrete Compressive Strength. UCI Machine Learning Repository, 2007. DOI: https://doi.org/10.24432/C5PK67. 7, 8

[YSP21] Rahul Yedida, Snehanshu Saha, and Tejas Prashanth. Lipschitzlr: Using theoretically computed adaptive learning rates for fast convergence. *Applied Intelligence*, 51:1460–1478, 2021. 20

A Concentration Inequalities

In this section we will give various concentration inequalities on the inlier data for functions in the Reproducing Kernel Hilbert Space. We will first give our assumptions for robust kernelized regression.

Assumption 18 (Gaussian Design)**.** We assume for $x_i \sim \mathbb{P} \in \mathcal{X}$, then it follows for the feature map, $\phi(\cdot) : \mathcal{X} \to \mathcal{H}$,

$$\phi(x_i) \sim \mathcal{N}(0, \Sigma) \tag{27}$$

where Σ is a possibly infinite dimensional covariance operator.

Assumption 19 (Normal Residuals)**.** The residual is defined as $\mu_i \triangleq f_w^*(x_i) - y_i$. Then we assume for some $\sigma > 0$, it follows

$$\mu_i \sim \mathcal{N}(0, \sigma^2) \tag{28}$$

Lemma 20 (Maximum of Gaussians)**.** *Let $\mu_1, \ldots, \mu_n \sim \mathcal{N}(0, \sigma^2)$ for some $\sigma > 0$. Then it follows*

$$\mathop{\mathbb{E}}_{\mu_i \sim \mathcal{N}(0,\sigma^2)} \max_{i \in [n]} |\mu_i| \leq O\left(\sigma \sqrt{\log n}\right) \tag{29}$$

Proof. We will integrate over the CDF to make our claim.

$$\mathop{\mathbb{E}}_{\mu_i \sim \mathcal{N}(0,\sigma^2)} \max_{i \in [n]} |\mu_i| = \int_0^\infty \mathop{\mathbb{P}}_{\mu_i \sim \mathcal{N}(0,\sigma^2)} \left\{ \max_{i \in [n]} |\mu_i| > t \right\} dt \overset{(i)}{\leq} c_1 + n \int_{c_1}^\infty \mathop{\mathbb{P}}_{\mu_i \sim \mathcal{N}(0,\sigma^2)} \{|\mu_i| \geq t\}\, dt \tag{30}$$

$$\overset{(ii)}{=} c_1 + 2n \int_{c_1}^\infty \mathop{\mathbb{P}}_{\mu_i \sim \mathcal{N}(0,\sigma^2)} \{\mu_i \geq t\}\, dt = c_1 + 2n \int_{c_1}^\infty \int_t^\infty \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x}{\sigma}\right)^2} dx dt \tag{31}$$

$$\leq c_1 + \frac{n}{\sigma}\sqrt{\frac{2}{\pi}} \int_{c_1}^\infty \int_t^\infty \left(\frac{x}{t}\right) e^{-\frac{1}{2}\left(\frac{x}{\sigma}\right)^2} dx dt = c_1 + n\sigma \sqrt{\frac{2}{\pi}} \int_{c_1}^\infty \frac{e^{-\frac{1}{2}\left(\frac{t}{\sigma}\right)^2}}{t} dt \tag{32}$$

$$\leq c_1 + n\sigma \sqrt{\frac{2}{\pi}} \int_{c_1}^\infty \left(\frac{t}{c_1}\right) e^{-\frac{1}{2}\left(\frac{t}{\sigma}\right)^2} dt = c_1 + n\sigma^3 \sqrt{\frac{2}{\pi}} \frac{e^{-\frac{1}{2}\left(\frac{c_1}{\sigma}\right)^2}}{c_1} \tag{33}$$

(i) follows from a union bound and noting for a i.i.d sequence of random variables $\{X_i\}_{i \in [n]}$ and a constant C, it follows $\mathbb{P}\{\max_{i \in [n]} X_i \geq C\} = n\mathbb{P}\{X \geq C\}$ where X is sampled from the same distribution as each X_i. (ii) follows from the symmetricity of the Gaussian distribution about zero. From here, we choose $c_1 \triangleq \sigma\sqrt{2 \log n}$. Then we have,

$$\mathop{\mathbb{E}}_{\mu_i \sim \mathcal{N}(0,\sigma^2)} \max_{i \in [n]} |\mu_i| \leq \sigma\sqrt{2 \log n} + \frac{\sigma^2}{\sqrt{\pi \log n}} \tag{34}$$

This completes the proof. ■

Lemma 21 (Maximum of Squared Gaussians)**.** *Let $\mu_1, \ldots, \mu_n \sim \mathcal{N}(0, \sigma^2)$ for $\sigma > 0$, $n > 1$. Then it follows*

$$\mathop{\mathbb{E}}_{\mu_i \sim \mathcal{N}(0,\sigma^2)} \max_{i \in [n]} \mu_i^2 \leq O\left(\sigma^2 \log(n) + \frac{\sigma^3}{\log(n)}\right) \tag{35}$$

Proof. Our proof follows similarly to the proof for Lemma 20.

$$\mathop{\mathbb{E}}_{\mu_i \sim \mathcal{N}(0,\sigma^2)} \max_{i \in \mathbb{N}} \mu_i^2 = \int_0^\infty \mathop{\mathbb{P}}_{\mu_i \sim \mathcal{N}(0,\sigma^2)} \left\{ \max_{i \in [n]} \mu_i^2 \geq t \right\} dt \leq c_2 + n \int_{c_2}^\infty \mathop{\mathbb{P}}_{\mu_i \sim \mathcal{N}(0,\sigma^2)} \left\{ |\mu_i| \geq \sqrt{t} \right\} dt \tag{36}$$

$$= c_2 + 2n \int_{c_2}^\infty \mathop{\mathbb{P}}_{\mu_i \sim \mathcal{N}(0,\sigma^2)} \left\{ \mu_i \geq \sqrt{t} \right\} dt = c_2 + 2n \int_{c_2}^\infty \int_{\sqrt{t}}^\infty \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x}{\sigma}\right)^2} dx dt \tag{37}$$

$$= c_2 + n\sigma \sqrt{\frac{2}{\pi}} \int_{c_2}^\infty \frac{e^{-\frac{1}{2}\left(\frac{t}{\sigma^2}\right)}}{\sqrt{t}} dt \overset{(i)}{\leq} c_2 + n\sigma \sqrt{\frac{2}{\pi}} \int_{c_2}^\infty \left(\frac{t}{c_2}\right) e^{-\frac{1}{2}\left(\frac{t}{\sigma^2}\right)} dt \tag{38}$$

$$\leq c_2 + \left(\sqrt{\frac{2}{\pi}}\right) \frac{n\sigma \left(4\sigma^4 + 2c_2\sigma^2\right) e^{-\frac{c_2}{2\sigma^2}}}{c_2} \tag{39}$$

(i) holds for $c_2 > 1$. Then, setting $c_2 \triangleq 2\sigma^2 \log(n)$, we have

$$\mathop{\mathbb{E}}_{\mu_i \sim \mathcal{N}(0,\sigma^2)} \max_{i \in [n]} \mu_i^2 \leq 2\sigma^2 \log(n) + \left(2\sigma^3 \sqrt{\frac{2}{\pi}}\right) \left(1 + \frac{1}{\log(n)}\right) \tag{40}$$

This completes the proof. ∎

Lemma 22 ([1]Expected Maximum P_k)**.** *Let $x_i \sim \mathbb{P}$ such that $\phi(x_i) \sim \mathcal{N}(0, \Sigma)$ from Assumption 18. Then it follows*

$$\mathop{\mathbb{E}}_{\phi(x_i) \sim \mathcal{N}(0,\Sigma)} \left[\max_{i \in [n]} k(x_i, x_i)\right] \leq O\left(\mathrm{Tr}\left(\Sigma\right) + \log n\right) \tag{41}$$

Proof. We once integrate over the CDF to make our claim.

$$\begin{aligned}
\mathop{\mathbb{E}}_{\phi(x_i) \sim \mathcal{N}(0,\Sigma)} \left[\max_{i \in [n]} k(x_i, x_i)\right] &= c_2 + \int_{c_2}^{\infty} \mathop{\mathbb{P}}_{\phi(x_i) \sim \mathcal{N}(0,\Sigma)} \left\{\max_{i \in [n]} k(x_i, x_i) \geq t\right\} dt && (42)\\
&\overset{(i)}{\leq} c_2 + n \int_{c_2}^{\infty} \mathop{\mathbb{P}}_{\phi(x) \sim \mathcal{N}(0,\Sigma)} \left\{k(x, x) \geq t\right\} dt && (43)\\
&\overset{(ii)}{\leq} c_2 + n \int_{c_2}^{\infty} \inf_{\theta > 0} e^{-t\theta} \mathbb{E}\left[e^{\theta k(x,x)}\right] dt && (44)\\
&= c_2 + n \int_{c_2}^{\infty} \inf_{\theta > 0} e^{-t\theta} \mathbb{E}\left[e^{\theta \sum_{i \in [d]} x_i^2}\right] dt && (45)\\
&= c_2 + n \int_{c_2}^{\infty} \inf_{0 < \theta < 1/2} (1 - 2\theta)^{-d/2} \exp\left(\frac{\theta\, \mathrm{Tr}\left(\Sigma\right)}{1 - 2\theta} - t\theta\right) dt && (46)\\
&\leq c_2 + n \inf_{0 < \theta < 1/2} \int_{c_2}^{\infty} (1 - 2\theta)^{-d/2} \exp\left(\frac{\theta\, \mathrm{Tr}\left(\Sigma\right)}{1 - 2\theta} - t\theta\right) dt && (47)\\
&\overset{(iii)}{\leq} c_2 + n \int_{c_2}^{\infty} \left(1 - 2^{1-p}\right)^{-d/2} \exp\left(\left(\frac{2^{-p}}{1 - 2^{1-p}}\right) \mathrm{Tr}\left(\Sigma\right) - 2^{-p}\theta\right) dt && (48)\\
&= c_2 + n \left(1 - 2^{-p}\right)^{-d/2} \left(\frac{\exp\left(\left(\frac{2^p}{1-2^{1-p}}\right) \mathrm{Tr}\left(\Sigma\right)\right)}{2^{-p} \exp\left(2^{-p} c_2\right)}\right) && (49)\\
&\overset{(iv)}{=} \left(1 - 2^{1-p}\right)^{-1} \mathrm{Tr}\left(\Sigma\right) + 2^p \log n + \left(1 - 2^{1-p}\right)^{-d/2} && (50)
\end{aligned}$$

See (i) from the proof of Lemma 20. (ii) follows from a Chernoff bound [Che52]. (iii) follows from setting $\theta \triangleq 2^{-p}$ for $p \in \mathbb{R}_{++}$ such that $p > 1$ and $p < \infty$. (iv) follows from setting $c_2 \triangleq (1 - 2^{1-p})^{-1} \mathrm{Tr}(\Sigma) + 2^p \log n + (1 - 2^{1-p})^{-d/2}$. Further optimization can be done over p dependent on $\mathrm{Tr}(\Sigma)$ and $\log(n)$. ∎

Lemma 23 (Norm of Functions with Gaussian Design in the Reproducing Kernel Hilbert Space)**.** *Let $x_i \sim \mathbb{P}$ such that $\phi(x_i) \sim \mathcal{N}(0, \Sigma)$ from Assumption 18 and Assumption 19. Then, it follows*

$$\mathop{\mathbb{E}}_{\phi(x_i) \sim \mathcal{N}(0,\Sigma)} \mathop{\mathbb{E}}_{\mu_i \sim \mathcal{N}(0,\sigma^2)} \left\|\sum_{i=1}^n \mu_i \phi(x_i)\right\|_{\mathcal{H}} \leq O\left(\sigma \sqrt{n \log n\, \mathrm{Tr}\left(\Sigma\right)}\right) \tag{51}$$

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,

$$\begin{aligned}
\mathop{\mathbb{E}}_{\phi(x_i) \sim \mathcal{N}(0,\Sigma)} \mathop{\mathbb{E}}_{\mu_i \sim \mathcal{N}(0,\sigma^2)} \left\|\sum_{i=1}^n \mu_i \phi(x_i)\right\|_{\mathcal{H}} &\leq \mathop{\mathbb{E}}_{\phi(x_i) \sim \mathcal{N}(0,\Sigma)} \mathop{\mathbb{E}}_{\mu_i \sim \mathcal{N}(0,\sigma^2)} \max_{i \in [n]} |\mu_i| \left\|\sum_{i=1}^n \phi(x_i)\right\|_{\mathcal{H}} && (52)\\
&\overset{\text{lem. 20}}{\leq} O\left(\sigma \sqrt{\log n}\right) \mathop{\mathbb{E}}_{\phi(x_i) \sim \mathcal{N}(0,\Sigma)} \mathop{\mathbb{E}}_{\xi_i \sim \mathcal{R}} \left\|\sum_{i=1}^n \xi_i \phi(x_i)\right\|_{\mathcal{H}} && (53)
\end{aligned}$$

[1]In Progress

$$\overset{(i)}{\leq} O\left(\sigma\sqrt{\log n}\right)\left(\mathop{\mathbb{E}}_{\phi(x_i)\sim\mathcal{N}(0,\Sigma)}\mathop{\mathbb{E}}_{\xi_i\sim\mathcal{R}}\left\|\sum_{i=1}^n \xi_i\phi(x_i)\right\|_{\mathcal{H}}^2\right)^{1/2} \tag{54}$$

$$= O\left(\sigma\sqrt{\log n}\right)\left(\mathop{\mathbb{E}}_{\phi(x_i)\sim\mathcal{N}(0,\Sigma)}\mathop{\mathbb{E}}_{\xi_i\sim\mathcal{R}}\left\langle\sum_{i=1}^n \xi_i\phi(x_i), \sum_{j=1}^n \xi_j\phi(x_j)\right\rangle_{\mathcal{H}}\right)^{1/2} \tag{55}$$

$$\overset{(ii)}{=} O\left(\sigma\sqrt{\log n}\right)\left(\mathop{\mathbb{E}}_{\phi(x_i)\sim\mathcal{N}(0,\Sigma)}\mathop{\mathbb{E}}_{\xi_i\sim\mathcal{R}}\sum_{i=1}^n\sum_{j=1}^n \xi_i\xi_j k(x_i,x_j)\right)^{1/2} \tag{56}$$

$$\overset{(iii)}{=} O\left(\sigma\sqrt{\log n}\right)\left(\mathop{\mathbb{E}}_{\phi(x_i)\sim\mathcal{N}(0,\Sigma)}\sum_{i=1}^n k(x_i,x_i)\right)^{1/2} = O\left(\sigma\sqrt{n\log n}\right)\left(\mathop{\mathbb{E}}_{\phi(x_i)\sim\mathcal{N}(0,\Sigma)}[k(x_i,x_i)]\right)^{1/2} \tag{57}$$

$$= O\left(\sigma\sqrt{n\log n\operatorname{Tr}(\Sigma)}\right) \tag{58}$$

(i) follows from Jensen's Inequality. (ii) follows from the definition of the kernel [Gre13a]. (iii) hols as we have $\mathbb{E}[\xi_i\xi_j] = \delta_{i,j}$, where δ is the Kronecker Delta function. ■

Lemma 24 (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* $\phi(x_i)\sim\mathcal{N}(0,\Sigma)$ *from Assumption 18, we then have*

$$\mathop{\mathbb{E}}_{\phi(x_i)\sim\mathcal{N}(0,\Sigma)}\left\|\frac{1}{n}\sum_{i=1}^n\phi(x_i)\otimes\phi(x_i)-\Sigma\right\|_{\text{HS}} \leq O\left(n^{-1/2}\operatorname{Tr}(\Sigma)\right) \tag{59}$$

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,

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

$$\overset{(i)}{\leq}\mathop{\mathbb{E}}_{\phi(x_i)\sim\mathcal{N}(0,\Sigma)}\mathop{\mathbb{E}}_{\tilde\phi(x_i)\sim\mathcal{N}(0,\Sigma)}\left\|\frac{1}{n}\sum_{i=1}^n\left(\phi(x_i)\otimes\phi(x_i)-\tilde\phi(x_i)\otimes\tilde\phi(x_i)\right)\right\|_{\text{HS}} \tag{60}$$

$$=\mathop{\mathbb{E}}_{\phi(x_i)\sim\mathcal{N}(0,\Sigma)}\mathop{\mathbb{E}}_{\tilde\phi(x_i)\sim\mathcal{N}(0,\Sigma)}\mathop{\mathbb{E}}_{\xi_i\sim\mathcal{R}}\left\|\frac{1}{n}\sum_{i=1}^n\xi_i\left(\phi(x_i)\otimes\phi(x_i)-\tilde\phi(x_i)\otimes\tilde\phi(x_i)\right)\right\|_{\text{HS}} \tag{61}$$

$$\overset{(ii)}{\leq}\frac{2}{n}\mathop{\mathbb{E}}_{\phi(x_i)\sim\mathcal{N}(0,\Sigma)}\mathop{\mathbb{E}}_{\xi_i\sim\mathcal{R}}\left\|\sum_{i=1}^n\xi_i\phi(x_i)\otimes\phi(x_i)\right\|_{\text{HS}} \tag{62}$$

$$\overset{(iii)}{\leq}\frac{2}{n}\left(\mathop{\mathbb{E}}_{\phi(x_i)\sim\mathcal{N}(0,\Sigma)}\mathop{\mathbb{E}}_{\xi_i\sim\mathcal{R}}\left\|\sum_{i=1}^n\xi_i\phi(x_i)\otimes\phi(x_i)\right\|_{\text{HS}}^2\right)^{1/2} \tag{63}$$

(i) follows from noticing $\phi(x_i)\otimes\phi(x_i)-\Sigma$ is a mean 0 operator in $\mathcal{H}\otimes\mathcal{H}$, then for $X,Y\in\mathcal{H}\otimes\mathcal{H}$ s.t. $\mathbb{E}[Y]=0$ it follows $\|X\|_{\text{HS}}=\|X-\mathbb{E}[Y]\|_{\text{HS}}=\|\mathbb{E}_Y[X-Y]\|_{\text{HS}}$ and finally applying Jensen's Inequality. (ii) follows from the triangle inequality. (iii) follows from Jensen's Inequality. Let e_k for $k\in[p]$ represent an orthonormal basis for the Hilbert Space $\mathcal{H}$. By expanding out the Hilbert-Schmidt Norm, we then have

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

$$=\frac{2}{n}\left(\mathop{\mathbb{E}}_{\phi(x_i)\sim\mathcal{N}(0,\Sigma)}\mathop{\mathbb{E}}_{\xi_i\sim\mathcal{R}}\sum_{k=1}^p\left\langle\sum_{i=1}^n\xi_i\phi(x_i)\otimes\phi(x_i)e_k, \sum_{j=1}^n\xi_j\phi(x_j)\otimes\phi(x_j)e_k\right\rangle\right)^{1/2} \tag{64}$$

$$= \frac{2}{n}\left(\mathop{\mathbb{E}}_{\phi(x_i)\sim\mathcal{N}(0,\Sigma)}\mathop{\mathbb{E}}_{\xi_i\sim\mathcal{R}}\sum_{k=1}^{p}\sum_{i=1}^{n}\sum_{j=1}^{n}\xi_i\xi_j\left\langle\phi(x_i)\otimes\phi(x_i)e_k,\phi(x_j)\otimes\phi(x_j)e_k\right\rangle\right)^{1/2} \tag{65}$$

$$\overset{(iv)}{\leq} \frac{2}{n}\left(\mathop{\mathbb{E}}_{\phi(x_i)\sim\mathcal{N}(0,\Sigma)}\sum_{k=1}^{p}\sum_{i=1}^{n}\left\langle\phi(x_i)\otimes\phi(x_i)e_k,\phi(x_i)\otimes\phi(x_i)e_k\right\rangle\right)^{1/2} \tag{66}$$

$$= \frac{2}{n}\left(\sum_{i=1}^{n}\mathop{\mathbb{E}}_{\phi(x_i)\sim\mathcal{N}(0,\Sigma)}\|\phi(x_i)\otimes\phi(x_i)\|_{\mathrm{HS}}^2\right)^{1/2} \overset{(v)}{=} \frac{2}{n}\left(\sum_{i=1}^{n}\mathop{\mathbb{E}}_{\phi(x_i)\sim\mathcal{N}(0,\Sigma)}\|\phi(x_i)\|_{\mathcal{H}}^4\right)^{1/2} \tag{67}$$

$$= \frac{2}{n}\left(\sum_{i=1}^{n}\mathop{\mathbb{E}}_{\phi(x_i)\sim\mathcal{N}(0,\Sigma)}\left[k^2(x_i,x_i)\right]\right)^{1/2} = \frac{2}{\sqrt{n}}\left(2\,\mathrm{Tr}\left(\Sigma^2\right)+\mathrm{Tr}\left(\Sigma\right)^2\right)^{1/2} \leq 2\sqrt{3}n^{-1/2}\,\mathrm{Tr}\left(\Sigma\right) \tag{68}$$

(iv) follows from noticing $\mathbb{E}_{\xi_i,\xi_j\sim\mathcal{R}}[\xi_i\xi_j] = \delta_{i,j}$. (v) follows from expanding the Hilbert-Schmidt Norm and applying Parseval's Identity. We note $\mathrm{Tr}(\Sigma) < \infty$ and therefore even though the covariance operator is infinite-dimensional we are able to get a finite bound on the covariance approximation. This completes the proof. ■

Lemma 25 (Finite Dimensional Covariate Estimation in the Spectral Norm)**.** *Let $x_1,\dots,x_n\sim\mathcal{N}(0,\Sigma)$. It then follows,*

$$\mathop{\mathbb{E}}_{x_i\sim\mathcal{N}(0,\Sigma)}\left\|\frac{1}{n}\sum_{i=1}^{n}x_ix_i^\top-\Sigma\right\|_2 \leq \frac{2\sqrt{3}\,\mathrm{Tr}\left(\Sigma\right)}{\sqrt{n}} \tag{69}$$

Proof. Our proof combines multiple results in High-Dimensional Probability for Sub-Gaussian vectors and adapting it for Gaussian-Design. We have,

$$\mathop{\mathbb{E}}_{x_i\sim\mathcal{N}(0,\Sigma)}\left\|\frac{1}{n}\sum_{i=1}^{n}x_ix_i^\top-\Sigma\right\|_2 \leq \left(\frac{\|\Sigma\|}{n}\right)\mathop{\mathbb{E}}_{\tilde{x}_i\sim\mathcal{N}(0,\mathrm{I})}\left\|\sum_{i=1}^{n}\tilde{x}_i\tilde{x}_i^\top-\mathrm{I}\right\|_2 \tag{70}$$

$$\overset{(i)}{=} \left(\frac{\|\Sigma\|}{n}\right)\mathop{\mathbb{E}}_{\tilde{x}_i\sim\mathcal{N}(0,\mathrm{I})}\sup_{\substack{u,v\in\mathbb{R}^d\\\|u\|=\|v\|=1}}u^\top\left(\sum_{i=1}^{n}\tilde{x}_i\tilde{x}_i^\top-\mathrm{I}\right)v \tag{71}$$

$$= \left(\frac{\|\Sigma\|_2}{n}\right)\mathop{\mathbb{E}}_{\tilde{x}_i\sim\mathcal{N}(0,\mathrm{I})}\sup_{\substack{u,v\in\mathbb{R}^d\\\|u\|=\|v\|=1}}\sum_{i=1}^{n}(u^\top\tilde{x}_i)(\tilde{x}_i^\top v)-u^\top v \tag{72}$$

(i) follows from the definition of the spectral norm. Note $\mathbb{E}[(u^\top\tilde{x}_i)(\tilde{x}_i^\top v)] = u^\top v$. Furthermore, we have for any $a\in\mathbb{R}^d$ s.t. $\|a\|=1$, it follows $u^\top\tilde{x}_i\sim\mathsf{subG}(\sqrt{\frac{8}{3}})$, then from [RH23] Lemma 1.12 and Theorem 4.16 we have $(u^\top x_i)(x_i^\top v)\sim\mathsf{subE}(16\sqrt{\frac{8}{3}})$. Then from Bernstein's Inequality in [RH23] Theorem 1.13 and (4.7), we have

$$\mathop{\mathbb{P}}_{\tilde{x}_i\sim\mathcal{N}(0,\mathrm{I})}\left\{\left\|\frac{1}{n}\sum_{i=1}^{n}\tilde{x}_i\tilde{x}_i^\top-\mathrm{I}\right\|_2 > t\right\} \leq 144^d\exp\left(-\frac{n}{2}\left(\left(\frac{t}{32\sqrt{8/3}}\right)^2\wedge\frac{t}{32\sqrt{8/3}}\right)\right) \tag{73}$$

Then, we can integrate to find the expectation.

$$\mathop{\mathbb{E}}_{\tilde{x}_i\sim\mathcal{N}(0,\mathrm{I})}\left[\left\|\frac{1}{n}\sum_{i=1}^{n}\tilde{x}_i\tilde{x}_i^\top-\mathrm{I}\right\|_2\right] \leq \int_0^\infty 144^d\exp\left(-\frac{n}{2}\left(\left(\frac{t}{32\sqrt{8/3}}\right)^2\wedge\frac{t}{32\sqrt{8/3}}\right)\right)dt \tag{74}$$

$$= \int_0^{32\sqrt{8/3}}144^d\exp\left(-\frac{n}{2}\left(\frac{t}{32\sqrt{8/3}}\right)^2\right)dt + \int_{32\sqrt{8/3}}^\infty 144^d\exp\left(-\frac{n}{2}\left(\frac{t}{32\sqrt{8/3}}\right)\right)dt \tag{75}$$

$$\leq \frac{\sqrt{\pi}144^d 32\sqrt{\frac{8}{3}}}{\sqrt{2n}} + \frac{64\sqrt{\frac{8}{3}}144^d e^{-n/2}}{n} < 32\sqrt{\frac{8\pi}{3}}144^d n^{-1/2} \tag{76}$$

This completes our proof. ■

B Proofs for Structural Results

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

B.1 Proof of Lemma 3

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.

$$\partial(-g(t, f_w)) = \partial \left(-t + \frac{1}{np} \sum_{i=1}^{n} (t - \hat{\nu}_i)^+ \right) = -1 + \frac{1}{np} \sum_{i=1}^{np} \begin{cases} 1 & \text{if } t > \hat{\nu}_i \\ 0 & \text{if } t < \hat{\nu}_i \\ [0,1] & \text{if } t = \hat{\nu}_i \end{cases} \tag{77}$$

Equation (77) is equal to 0 when $t = \hat{\nu}_{np}$. This is equivalent to the p-quantile of the Risk. ■

B.2 Proof of Lemma 5

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

$$\nabla_f g(t^{(k+1)}, f_w^{(k)}) = \nabla_f \left(\hat{\nu}_{np} - \frac{1}{np} \sum_{i=1}^{n} \left(\hat{\nu}_{np} - \ell(x_i; f_w^{(k)}, y_i) \right)^+ \right) \tag{78}$$

$$= -\frac{1}{np} \sum_{i=1}^{np} \nabla_f \left(\hat{\nu}_{np} - \ell(x_i; f_w^{(k)}, y_i) \right)^+ = \frac{1}{np} \sum_{i=1}^{n} \nabla_f \ell(x_i; f_w^{(k)}, y_i) \begin{cases} 1 & \text{if } t > \hat{\nu}_i \\ 0 & \text{if } t < \hat{\nu}_i \\ [0,1] & \text{if } t = \hat{\nu}_i \end{cases} \tag{79}$$

Now we note $\nu_{np} \leq t^{(k+1)} \leq \nu_{np+1}$. Then, plugging this into Equation (79), we have

$$\nabla_f g(t^{(k+1)}, f_w^{(k)}) = \frac{1}{np} \sum_{i=1}^{np} \nabla_f \ell(x_i; f_w^{(k)}, y_i) \tag{80}$$

This concludes the proof. ■

B.3 Some details on the Softplus Approximation

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

$$\nabla_t \tilde{g}_\lambda(t, f_w) = \nabla_t \left(t - \frac{1}{np} \sum_{i=1}^{n} \frac{1}{\lambda} \ln\left(1 + \exp\left(\lambda\left(t - \ell(f_w; x_i, y_i)\right)\right)\right) \right) \tag{81}$$

$$= 1 - \frac{1}{np} \sum_{i=1}^{n} \sigma\left(\lambda\left(t - \ell(f_w; x_i, y_i)\right)\right) \tag{82}$$

where $\sigma(\cdot)$ is the sigmoid function. We then note as $\lambda \to \infty$, the sigmoid function approaches the indicator function. It therefore follows the derivative of g with respect to t is given as,

$$\lim_{\lambda \to \infty} \nabla_t \tilde{g}_\lambda(t, f_w) = 1 - \frac{1}{np} \sum_{i=1}^{n} \mathbb{I}\left\{t - \ell\left(f_w; x_i, y_i\right)\right\} \tag{83}$$

$$\nabla_f \tilde{g}_\lambda(t, f_w) = \nabla_f \left(t - \frac{1}{np} \sum_{i=1}^{n} \frac{1}{\lambda} \ln\left(1 + \exp\left(\lambda\left(t - \ell(f_w; x_i, y_i)\right)\right)\right) \right) \tag{84}$$

$$= \frac{1}{np} \sum_{i=1}^{n} \nabla_f \ell(f_w; x_i, y_i)\, \sigma(\lambda(t - \ell(f_w; x_i, y_i))) \tag{85}$$

We therefore similarly have the derivative of g with respect to f,

$$\lim_{\lambda\to\infty} \nabla_f \tilde{g}_\lambda(t, f_w) = \frac{1}{np} \sum_{i=1}^{n} \mathbb{I}\{t - \ell(f_w; x_i, y_i)\} \nabla_f \ell(f_w; x_i, y_i) \tag{86}$$

B.4 Proof of Lemma 6

Proof. We will upper bound the operator norm of the Hessian. Let $\vartheta \triangleq \sigma(\lambda(t - \ell(f_w; x_i, y_i)))$, we then have

$$\nabla_f^2 \tilde{g}_\lambda(t, f_w) = \nabla_f \left(\frac{1}{np} \sum_{i=1}^{n} \vartheta \nabla_f \ell(f_w; x_i, y_i) \right) \tag{87}$$

$$= \frac{1}{np} \sum_{i=1}^{n} \Big(\vartheta \nabla_f^2 \ell(f_w; x_i, y_i) - \vartheta(1-\vartheta) (\nabla_f \ell(f_w; x_i, y_i) \otimes \nabla_f \ell(f_w; x_i, y_i)) \Big) \tag{88}$$

Now we will upper bound the operator norm of the Hessian.

$$\lim_{\lambda\to\infty} \sup_{f_w\in\mathcal{K}} \left\| \nabla_f^2 \tilde{g}_\lambda(t, f_w) \right\|_{\text{op}} \|f_w - f_{\tilde{w}}\|_{\mathcal{H}}$$

$$\overset{(88)}{=} \lim_{\lambda\to\infty} \sup_{f_w} \left\| \frac{1}{np} \sum_{i=1}^{n} \vartheta \nabla_f^2 \ell(f_w; x_i, y_i) - \vartheta(1-\vartheta) \nabla_f \ell(f_w; x_i, y_i) \otimes \nabla_f \ell(f_w; x_i, y_i) \right\|_{\text{op}} \|f_w - f_{\tilde{w}}\|_{\mathcal{H}} \tag{89}$$

$$\overset{(i)}{\leq} \lim_{\lambda\to\infty} \sup_{f_w\in\mathcal{K}} \frac{1}{np} \sum_{i=1}^{n} \left\| \vartheta \nabla_f^2 \ell(f_w; x_i, y_i) \right\|_{\text{op}} \|f_w - f_{\tilde{w}}\|_{\mathcal{H}} \overset{(ii)}{\leq} \sup_{f_w\in\mathcal{K}} \frac{1}{np} \sum_{i=1}^{n} \left\| \nabla_f^2 \ell(f_w; x_i, y_i) \right\|_{\text{op}} \|f_w - f_{\tilde{w}}\|_{\mathcal{H}}$$

(i) follows from applying the Triangle Inequality and then Weyl's Inequality [Wey12]. (ii) follows from noting $\vartheta \in (0, 1)$. We now note that removing ϑ also removes the dependence on λ which allows us to take the limit out of the expression. ■

B.5 Proof of Lemma 11

Proof. By the definition of stationary point, we have

$$f_{\hat{w}} = \lim_{\lambda\to\infty} \arg\inf_{f_w\in\mathcal{K}} \left\{ \Phi_\lambda(f_w) + \frac{1}{2\rho} \|f_w - f_{\hat{w}}\|_{\mathcal{H}}^2 \right\} \tag{90}$$

$$\overset{(i)}{=} \arg\inf_{f_w\in\mathcal{K}} \left\{ \lim_{\lambda\to\infty} \Phi_\lambda(f_w) + \frac{1}{2\rho} \|f_w - f_{\hat{w}}\|_{\mathcal{H}}^2 \right\} \tag{91}$$

(i) holds as we ρ is independent of λ as shown in the proof of Lemma 6. This implies then for any $f_w \in \mathcal{K}$ and noting $\rho \leq \beta^{-1}$, it follows

$$\lim_{\lambda\to\infty} \Phi_\lambda(f_{\hat{w}}) \leq \lim_{\lambda\to\infty} \Phi_\lambda(f_w) + \beta \|f_w - f_{\hat{w}}\|_{\mathcal{H}}^2 \tag{92}$$

where we choose $\rho \triangleq 1/(2\beta)$. We can then plug in the optimal, f_w^* for f_w and rearrange and we have the desired result. ■

C Proofs for Kernelized Regression

C.1 L-Lipschitz Constant and β-Smoothness

Lemma 26 (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:*

$$|g(t, f_w) - g(t, f_{\hat{w}})| \leq L \|f_w - f_{\hat{w}}\|_{\mathcal{H}} \tag{93}$$

where

$$L = \frac{2R}{np} \left(\sum_{i=1}^{n} \sqrt{k(x_i, x_i)} \right)^2 + \frac{2\|y\|_2}{p\sqrt{n}} \left(\sum_{i=1}^{n} \sqrt{k(x_i, x_i)} \right) \tag{94}$$

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

$$\begin{aligned}
|g(t, f_{w_1}) - g(t, f_{w_2})| &= \left| \int_0^1 \nabla_f g(t, (1-\lambda) f_{w_1} + \lambda f_{w_2})(f_{w_1} - f_{w_2}) d\lambda \right| & (95) \\
&\leq \|f_{w_1} - f_{w_2}\|_{\mathcal{H}} \left| \int_0^1 \nabla_f g(t, (1-\lambda) f_{w_1} + \lambda f_{w_2}) d\lambda \right| & (96) \\
&\overset{(a)}{\leq} \|f_{w_1} - f_{w_2}\|_{\mathcal{H}} \max_{f_w \in \mathcal{K}} \|\nabla_f g(t, f_w)\|_{\mathcal{H}} & (97)
\end{aligned}$$

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 $\mathcal{K}$.

$$\|\nabla g(t, f_w)\|_{\mathcal{H}} = \left\| \frac{2}{np} \sum_{i=1}^n \mathbb{I}\left\{ t \geq (f_w(x_i) - y_i)^2 \right\} (f_w(x_i) - y_i) \cdot k(x_i, \cdot) \right\|_{\mathcal{H}} \tag{98}$$

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

$$\begin{aligned}
&= \left\| \frac{2}{np} \sum_{i=1}^m (f_w(x_i) - y_i) \cdot k(x_i, \cdot) \right\|_{\mathcal{H}} \leq \frac{2}{np} \left(\left\| \sum_{i=1}^m f_w(x_i) \cdot k(x_i, \cdot) \right\|_{\mathcal{H}} + \left\| \sum_{i=1}^m y_i k(x_i, \cdot) \right\|_{\mathcal{H}} \right) & (99) \\
&\overset{(a)}{\leq} \frac{2}{np} \left(\left\| \sum_{i=1}^m \left\langle \sum_{j=1}^n w_j k(x_j, \cdot), k(x_i, \cdot) \right\rangle_{\mathcal{H}} \cdot k(x_i, \cdot) \right\|_{\mathcal{H}} + \left\| \sum_{i=1}^m y_i \right\| \left\| \sum_{i=1}^m k(x_i, \cdot) \right\|_{\mathcal{H}} \right) & (100) \\
&\leq \frac{2}{np} \left(\left| \left\langle \sum_{j=1}^n w_j k(x_j, \cdot), \sum_{i=1}^m k(x_i, \cdot) \right\rangle_{\mathcal{H}} \right| \left\| \sum_{i=1}^m k(x_i, \cdot) \right\|_{\mathcal{H}} + \left\| \sum_{i=1}^m y_i \right\| \sum_{i=1}^m \sqrt{k(x_i, x_i)} \right) & (101) \\
&\leq \frac{2}{np} \left(\|f_w\|_{\mathcal{H}} \left(\sum_{i=1}^m \sqrt{k(x_i, x_i)} \right)^2 + \sqrt{n} \|y\|_2 \left(\sum_{i=1}^n \sqrt{k(x_i, x_i)} \right) \right) & (102) \\
&\leq \frac{2R}{np} \left(\sum_{i=1}^n \sqrt{k(x_i, x_i)} \right)^2 + \frac{2\|y\|_2}{p\sqrt{n}} \left(\sum_{i=1}^n \sqrt{k(x_i, x_i)} \right) & (103)
\end{aligned}$$

(a) follows form the reproducing property for RKHS [Gre13b]. 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. Similar results for the Lipschitz Constant for non-kernelized learning algorithms can be seen in [YSP21]. This concludes the proof. ■

Lemma 27. *(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:*

$$\|\nabla_f g(t, f_w) - \nabla_f g(t, f_{\hat{w}})\| \leq \beta \|f_w - f_{\hat{w}}\|_{\mathcal{H}} \tag{104}$$

where

$$\beta = \frac{2}{np} \sum_{i=1}^n k(x_i, x_i) = \frac{2}{np} \operatorname{Tr}(\mathrm{K}) \tag{105}$$

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

$$\begin{aligned}
\left\|\nabla_f^2 g(t, f_w)\right\|_{\mathrm{op}} &= \left\| \frac{2}{np} \sum_{i=1}^n \mathbb{I}\left\{ t \geq (f_w(x_i) - y_i)^2 \right\} \phi(x_i) \otimes \phi(x_i) \right\|_{\mathrm{op}} \\
&\leq \frac{2}{np} \sum_{i=1}^n \|\phi(x_i) \otimes \phi(x_i)\|_{\mathrm{HS}} = \frac{2}{np} \sum_{i=1}^n k(x_i, x_i) = \frac{2}{np} \operatorname{Tr}(\mathrm{K})
\end{aligned}$$

This concludes the proof.

C.2 Proof of Lemma 13

Proof.[2] We will first expand the expression in the Lemma statement. Let λ_i and φ_i for $i \in \mathbb{N}$ represent the eigenvalues and eigenfunctions for $\mathbb{E}_{x\sim\mathbb{P}}[\phi(x) \otimes \phi(x)] \triangleq \Sigma$.

$$\langle f_w \otimes f_w^*, \Sigma\left(f_w - f_w^*\right)\rangle = \lim_{p\to\infty} \sum_{i=1}^{p} \lambda_i \left\langle f_w - f_w^*, \varphi_i\right\rangle_{\mathcal{H}}^2 \tag{106}$$

Therefore, for some $m \in \mathbb{N}$, we want the projection of $f_w - f_w^*$ to be non-zero for m. We will show, we only need to make an assumption on f_w^*. The projection in the Reproducing Kernel Hilbert Space is given as the following,

$$\left\|\text{Proj}_{U_m} f_w^*\right\|_{\mathcal{H}} \triangleq \left\|\sum_{i=1}^{m} \left\langle \varphi_i, f_w^*\right\rangle \varphi_i\right\|_{\mathcal{H}} \tag{107}$$

Let $f_w^{(T)}$ be the T iterate from Subquantile Kernel Algorithm. We then have,

$$\left\|\text{Proj}_{U_m}\left(f_w^{(T)} - f_w^*\right)\right\|_{\mathcal{H}} \overset{(i)}{\geq} \left\|\text{Proj}_{U_m} f_{w^*}\right\|_{\mathcal{H}} - \left\|\text{Proj}_{U_m} f_w\right\|_{\mathcal{H}} \tag{108}$$

$$= \left\|\text{Proj}_{U_m} f_w^*\right\|_{\mathcal{H}} - \left\|\text{Proj}_{U_m}\left(\text{Proj}_{\mathcal{K}}\left(\text{Proj}_{\mathcal{K}}\left(f_w^{(0)} - \alpha\nabla g(t^{(1)}, f_w^{(0)})\right) - \alpha\nabla g(t^{(2)}, f_w^{(1)})\right) - \cdots\right)\right\|_{\mathcal{H}} \tag{109}$$

(i) follows from reverse triangle inequality and linearlity of the projection operator. We then require the following result on the Projection onto the norm ball in the Reproducing Kernel Hilbert Space.

Lemma 28. *Let $\mathcal{K} \triangleq \{f_w : \|f_w\|_{\mathcal{H}} \leq R\}$. Then, for a $f_{\hat{w}} \notin \mathcal{K}$, it follows*

$$\text{Proj}_{\mathcal{K}} f_{\hat{w}} = \Omega(1) f_{\hat{w}} \tag{110}$$

Proof. We will formulate the dual problem and then find the corresponding f_w that solves the dual.

$$\text{Proj}_{\mathcal{K}} f_{\hat{w}} = \arg\min_{f_w\in\mathcal{K}} \|f_w - f_{\hat{w}}\|_{\mathcal{H}}^2 = \arg\min_{f_w\in\mathcal{K}} \|f_w\|_{\mathcal{H}}^2 + \|f_{\hat{w}}\|_{\mathcal{H}}^2 - 2\left\langle f_w, f_{\hat{w}}\right\rangle_{\mathcal{H}} \tag{111}$$

$$= \arg\min_{f_w\in\mathcal{K}} \|f_w\|_{\mathcal{H}}^2 - 2\left\langle f_w, f_{\hat{w}}\right\rangle_{\mathcal{H}} \tag{112}$$

From here we can solve the dual problem. The Lagrangian is given by,

$$\mathcal{L}(f_w, u) \triangleq \|f_w\|_{\mathcal{H}}^2 - 2\left\langle f_w, f_{\hat{w}}\right\rangle + u\left(\|f_w\|_{\mathcal{H}}^2 - R^2\right) \tag{113}$$

Then, we have dual problem as $\theta(u) = \min_{w\in\mathcal{H}} \mathcal{L}(f_w, u)$. Taking the derivative of the Lagrangian and setting it to zero, we obtain $\arg\min_{f_w\in\mathcal{H}} \mathcal{L}(f_w, u) = (1+u)^{-1} f_{\hat{w}}$. With some more work, we obtain $\arg\max_{u>0} \theta(u) = R^{-1}\|f_{\hat{w}}\| - 1$. We then have f_w at u^* as $f_w = R\|f_{\hat{w}}\|_{\mathcal{H}}^{-1} f_{\hat{w}}$. Since $\|f_{\hat{w}}\| > R$ as $f_{\hat{w}} \notin \mathcal{K}$ by assumption, our proof is complete. ■

Now we can utilize the result in Lemma 28 to continue our proof of Lemma 13. Recall α is the fixed learning rate.

$$\left\|\text{Proj}_{U_m}\left(f_w^{(T)} - f_w^*\right)\right\|_{\mathcal{H}} \overset{\text{lem. 28}}{\geq} \left\|\text{Proj}_{U_m} f_w^*\right\|_{\mathcal{H}} - \left\|\text{Proj}_{U_m}\left(f_w^{(0)} - \alpha\sum_{k=1}^{T} \nabla g(t^{(k)}, f_w^{(k-1)})\right)\right\|_{\mathcal{H}} \tag{114}$$

$$\geq \left\|\text{Proj}_{U_m} f_w^*\right\|_{\mathcal{H}} - \left\|\text{Proj}_{U_m} f_w^{(0)}\right\|_{\mathcal{H}} - \alpha\left\|\sum_{k=1}^{T} \text{Proj}_{U_m} \nabla g(t^{(k)}, f_w^{(k-1)})\right\|_{\mathcal{H}} \tag{115}$$

$$\geq \left\|\text{Proj}_{U_m} f_w^*\right\|_{\mathcal{H}} - \left\|\text{Proj}_{U_m} f_w^{(0)}\right\|_{\mathcal{H}} - \alpha T \underbrace{\left\|\text{Proj}_{U_m} \nabla g(t^{(k)}, f_w^{(k-1)})\right\|_{\mathcal{H}}}_{C_6} \tag{116}$$

[2] In Progress

We will bound C_6, let $\phi_m(x_i) \triangleq \mathrm{Proj}_{U_m}\phi(x_i)$ for a $x_i \in \mathbb{R}^d$, $k_m(x_i, x_i) \triangleq \langle \phi_m(x_i), \phi_m(x_i)\rangle$, and $\Sigma_m \triangleq \Sigma_{1:m,1:m}$.

$$
\begin{aligned}
C_6 &= \left\| \mathrm{Proj}\left(\sum_{i\in S\cap P} 2\left\langle f_w^{(k)} - f_w^*, k(x_i,\cdot)\right\rangle_{\mathcal{H}} k(x_i,\cdot) - \eta_i k(x_i,\cdot) + 2\sum_{i\in S\cap Q} \nabla_f \ell\left(f_w; x_i, y_i\right)\right)\right\|_{\mathcal{H}} && (117)\\
&\leq \left\| \sum_{i\in S\cap P} \mathrm{Proj}_{U_m}\left[\phi(x_i)\otimes\phi(x_i)\right]\left(f_w^{(k)} - f_w^*\right)\right\|_{\mathcal{H}} + \left\| \sum_{i\in S\cap P} \eta_i \mathrm{Proj}_{U_m}\phi(x_i)\right\|_{\mathcal{H}} \\
&\quad + \max_{k\in[T]} \left\| f_w^{(k)}\right\|_{\mathcal{H}} \left(\sum_{i\in Q} \langle \phi_m(x_i), \phi_m(x_i)\rangle_{\mathcal{H}}\right) + \left(\sum_{i\in Q} |y_i|\right)\left(\sum_{i\in Q} \|\phi_m(x_i)\|_{\mathcal{H}}\right) && (118)\\
&\overset{(i)}{\leq} \left(\sup_{\substack{u,v\in\mathrm{range}(U_m)\\ \|u\|_{\mathcal{H}}=\|v\|_{\mathcal{H}}=1}} \sum_{i\in S\cap P} \langle u, \phi(x_i)\rangle_{\mathcal{H}} \langle \phi(x_i), v\rangle_{\mathcal{H}}\right)\left\| f_w^{(T)} - f_w^*\right\|_{\mathcal{H}} + \overbrace{\left\| \sum_{i\in S\cap P} \eta_i \phi_m(x_i)\right\|_{\mathcal{H}}}^{C_1} + c_7 R\left(\sum_{i\in Q} k_m(x_i,x_i)\right) \\
&\quad + \|y_Q\|_1 \sqrt{c_7}\left(\sum_{i\in Q}\sqrt{k(x_i,x_i)}\right) && (119)\\
&\overset{(ii)}{\leq} c_8\left(\sum_{i\in P} k_m(x_i,x_i)\right)\left(\alpha L T + \left\| f_w^{(0)} - f_w^*\right\|_{\mathcal{H}}\right) + O\left(\sigma\sqrt{n(1-\varepsilon)\log(n(1-\varepsilon))\,\mathrm{Tr}(\Sigma)}\right) + O\left(\sqrt{\left(\frac{1}{\delta}\right)\sigma^3 n\,\mathrm{Tr}(\Sigma)}\right) \\
&\quad + c_7 R\left(\sum_{i\in Q} k_m(x_i,x_i)\right) + c_7\|y_Q\|_1\left(\sum_{i\in Q}\sqrt{k_m(x_i,x_i)}\right) && (120)\\
&\overset{(iii)}{\leq} c_8\left(n(1-\varepsilon)\,\mathrm{Tr}(\Sigma_m) + \left(\sqrt{\left(\frac{1}{\delta}\right) 2n(1-\varepsilon)\,\mathrm{Tr}\left(\Sigma_m^2\right)}\right)\right)\left(\alpha L T + \left\| f_w^{(0)} - f_w^*\right\|_{\mathcal{H}}\right) + O\left(\sigma\sqrt{n(1-\varepsilon)\log(n(1-\varepsilon))\,\mathrm{Tr}(\Sigma)}\right) \\
&\quad + O\left(\sqrt{\left(\frac{1}{\delta}\right)\sigma^3 n\,\mathrm{Tr}(\Sigma)}\right) + c_7 R\left(\sum_{i\in Q} k_m(x_i,x_i)\right) + c_7\|y_Q\|_1\left(\sum_{i\in Q}\sqrt{k_m(x_i,x_i)}\right) && (121)
\end{aligned}
$$

In (ii) we bound C_1 with the Chebyshev Inequality with probability at least $(1-\delta)$ for $\delta \in (0,1)$. In (iii) we bound C_2 with the Chebyshev Inequality with probability with at least $(1-\delta)$ for $\delta \in (0,1)$. From our assumption of $\|\mathrm{Proj}_{\mathcal{K}} f_{\hat{w}}\|_{\mathcal{H}}$, we then have

$$
\langle \Sigma, (f_w - f_w^*)\otimes(f_w - f_w^*)\rangle_{\mathrm{HS}} \geq C\lambda_m \qquad (122)
$$

Lemma 29. *If $\|f_w - f_{w^*}\| \geq \eta$, then it follows*

$$
\begin{aligned}
\lim_{\lambda\to\infty}\left(\Phi_\lambda\left(f_w\right) - \Phi_\lambda\left(f_{w^*}\right)\right) &\geq \eta^2 n(1-2\varepsilon)\lambda_{\min}\left(\mathop{\mathbb{E}}_{x\sim\mathbb{P}}\left[\phi(x)\otimes\phi(x)\right]\right) \\
&\quad - O\left(\sigma\sqrt{n(1-2\varepsilon)\log(n(1-2\varepsilon))\,\|\Sigma\|_{\mathrm{HS}}}\right) - 2\eta\left\|\sum_{i\in S\cap P}\eta_i\phi\left(x_i\right)\right\| - \sum_{j\in P\setminus S}\eta_j^2
\end{aligned} \qquad (123)
$$

Proof. Let S be the set containing the points with the minimum error from X w.r.t to the weights vector w. Define $\eta_i \triangleq \left(f_w^*\left(x_i\right) - y_i\right)$ where $i \in P$.

$$
\begin{aligned}
\lim_{\lambda\to\infty}\left(\Phi_\lambda\left(f_w\right) - \Phi_\lambda\left(f_w^*\right)\right) &= \sum_{i\in S}\left(f_w\left(x_i\right) - y_i\right)^2 - \sum_{j\in P}\left(f_w^*\left(x_j\right) - y_j\right)^2 && (124)\\
&= \sum_{i\in S\cap P}\left(f_w\left(x_i\right) - y_i\right)^2 + \sum_{i\in S\cap Q}\left(f_w\left(x_i\right) - y_i\right)^2 - \sum_{j\in P}\left(f_w^*\left(x_j\right) - y_j\right)^2 && (125)\\
&\geq \sum_{i\in S\cap P}\left(f_w\left(x_i\right) - y_i\right)^2 - \sum_{j\in P}\left(f_w^*\left(x_j\right) - y_j\right)^2 = \sum_{i\in S\cap P}\left(f_w\left(x_i\right) - f_w^*\left(x_i\right) - \eta_i\right)^2 - \sum_{j\in P}\eta_j^2 && (126)\\
&\geq \sum_{i\in S\cap P}\underbrace{\left(\left(f_w - f_w^*\right)\left(x_i\right)\right)^2}_{A_1} \underbrace{-2\sum_{i\in S\cap P}\eta_i\left(f_w - f_w^*\right)\left(x_i\right)}_{A_2} - \underbrace{\sum_{j\in P\setminus S}\eta_j^2}_{A_3} && (127)
\end{aligned}
$$

Now we will upper bound A_1. Similar to [CLKZ21] 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 $\Sigma = \mathbb{E}_{x\sim\mathbb{P}}[\phi(x)\otimes\phi(x)] = \mathrm{diag}(\xi_1,\ldots,\xi_n)$. This is the eigenfunction basis described in [SS16].

$$A_1 \triangleq \sum_{i\in S\cap P}\left(\left(f_w - f_{w^*}\right)(x_i)\right)^2 \overset{(a)}{=} \sum_{i\in S\cap P}\left\langle \sum_{j\in X}\left(w_j - w_j^*\right)k(x_j,\cdot), k(x_i,\cdot)\right\rangle_{\mathcal{H}}^2 \tag{128}$$

$$= \sum_{i\in S\cap P}\left\langle \sum_{j\in X}\left(w_j - w_j^*\right)\phi(x_j), \phi(x_i)\right\rangle_{\mathcal{H}}\left\langle \phi(x_i), \sum_{j\in X}\left(w_j - w_j^*\right)\phi(x_j)\right\rangle_{\mathcal{H}} \tag{129}$$

$$= \sum_{i\in S\cap P}\left\langle \sum_{j\in X}\left(w_j - w_j^*\right)\phi(x_j), \phi(x_i)\otimes\phi(x_i)\sum_{j\in X}\left(w_j - w_j^*\right)\phi(x_j)\right\rangle_{\mathcal{H}} \tag{130}$$

$$= \sum_{i\in S\cap P}\left\langle \phi(x)\otimes\phi(x), \left(f_w - f_w^*\right)\otimes\left(f_w - f_w^*\right)\right\rangle_{\mathrm{HS}} \tag{131}$$

$$= \sum_{i\in S\cap P}\left\langle \Sigma + \phi(x)\otimes\phi(x) - \Sigma, \left(f_w - f_w^*\right)\otimes\left(f_w - f_w^*\right)\right\rangle_{\mathrm{HS}} \tag{132}$$

$$\overset{\text{lem. 24}}{\geq} \left(n(1-2\varepsilon)\lambda_{\min}\left(\Sigma\right) - \left\|\sum_{i\in S\cap P}\phi(x)\otimes\phi(x) - \Sigma\right\|_{\mathrm{HS}}\right)\|f_w - f_w^*\|_{\mathcal{H}}^2 \tag{133}$$

Next we will upper bound A_2,

$$A_2 \triangleq \sum_{i\in S\cap P}\eta_i\left(f_w - f_w^*\right)(x_i) = \sum_{i\in S\cap P}\left\langle \sum_{j\in X}(w_j - w_j^*)k(x_j,\cdot), \eta_i k(x_i,\cdot)\right\rangle_{\mathcal{H}} \tag{134}$$

$$= \left\langle \sum_{j\in X}(w_j - w_j^*)k(x_j,\cdot), \sum_{i\in S\cap P}\eta_i k(x_i,\cdot)\right\rangle_{\mathcal{H}} \tag{135}$$

$$\leq \|f_w - f_w^*\|_{\mathcal{H}}\left\|\sum_{i\in S\cap P}\eta_i k(x_i,\cdot)\right\|_{\mathcal{H}} = \|f_w - f_w^*\|_{\mathcal{H}}\left\|\sum_{i\in S\cap P}\eta_i\phi(x_i)\right\|_{\mathcal{H}} \tag{136}$$

Then, combining our bounds, we have

$$\begin{aligned}\lim_{\lambda\to\infty}\left(\Phi_\lambda\left(f_w\right) - \Phi_\lambda\left(f_w^*\right)\right) &\overset{(133)\text{ and }(136)}{\geq} \eta^2\left(n(1-2\varepsilon)\lambda_{\min}\left(\mathop{\mathbb{E}}_{x\sim\mathbb{P}}[\phi(x)\otimes\phi(x)]\right)\right. \\ &\left.- \left\|\sum_{i\in S\cap P}\phi(x)\otimes\phi(x) - \Sigma\right\|_{\mathrm{HS}}\right) - 2\eta\left\|\sum_{i\in S\cap P}\eta_i\phi\left(x_i\right)\right\| - \sum_{j\in P\setminus S}\eta_j^2\end{aligned} \tag{137}$$

This completes the proof. ∎

C.3 Proof of Theorem 14

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

$$\|\nabla\mathsf{M}_{\Phi_\lambda,\rho}\left(f_w\right)\|_{\mathcal{H}} = \left\|\frac{1}{\rho}\left(f_w - \mathop{\arg\min}_{f_{\hat{w}}\in\mathcal{K}}\left(\Phi\left(f_{\hat{w}}\right) + \frac{1}{2\rho}\|f_w - f_{\hat{w}}\|_{\mathcal{H}}^2\right)\right)\right\|_{\mathcal{H}} = 0 \tag{138}$$

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

$$\lim_{\lambda\to\infty}\left(\Phi_\lambda\left(f_{\hat{w}}\right)\right) < \lim_{\lambda\to\infty}\left(\Phi_\lambda\left(f_{\tilde{w}}\right)\right) + \frac{1}{2\rho}\|f_{\tilde{w}} - f_{\hat{w}}\|_{\mathcal{H}}^2 \tag{139}$$

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

$$\lim_{\lambda\to\infty}\left(\Phi_\lambda\left(f_{\hat{w}}\right) - \Phi_\lambda\left(f_w^*\right)\right) \leq \frac{1}{2\rho}\|f_{\hat{w}} - f_w^*\|_{\mathcal{H}}^2 \tag{140}$$

We proceed by proof by contradiction. Assume $\|f_{\hat{w}} - f^*_w\| > \eta$, then if $\Phi(f_{\hat{w}}) - \Phi(f^*_w) > \frac{\eta^2}{2\rho}$, then we will have $f_{\hat{w}}$ is not a stationary point, which will imply $\|f_{\hat{w}} - f^*_w\|_{\mathcal{H}} \leq \eta$. Therefore, we attempt to find the minimum value for η. From Lemma 29, we have the expected distance from a stationary point of the Moreau Envelope from the optimal point over the distribution of uncorrupted datasets.

$$\begin{aligned}\mathop{\mathbb{E}}_{\mathcal{D}\sim\hat{\mathbb{P}}} \lim_{\lambda\to\infty} \left(\Phi\left(f_w\right) - \Phi\left(f^*_w\right)\right) &\overset{\text{lem. 29}}{\geq} \eta^2 \Bigg(n(1-2\varepsilon)\lambda_{\min}\left(\mathop{\mathbb{E}}_{x\sim\mathbb{P}}\left[\phi(x)\otimes\phi(x)\right]\right) \\ &- \mathop{\mathbb{E}}_{x_i\sim\mathbb{P}} \left\| \sum_{i\in S\cap P} \phi(x_i)\otimes\phi(x_i) - \Sigma \right\|_{\text{HS}} \Bigg) - 2\eta \mathop{\mathbb{E}}_{\mu_i, x_i\sim\mathbb{P}} \left\| \sum_{i\in S\cap P} \eta_i \phi\left(x_i\right)\right\| - \mathop{\mathbb{E}}_{\mu_j\sim\mathbb{P}} \sum_{j\in P\setminus S} \eta_j^2 \end{aligned} \tag{141}$$

$$\overset{\text{lems. 23 and 24}}{\geq} \eta^2\left(n(1-2\varepsilon)\lambda_{\min}\left(\Sigma\right) - 2\operatorname{Tr}\left(\Sigma\right)\sqrt{n(1-2\varepsilon)}\right) - \eta O\left(\sigma\sqrt{n(1-2\varepsilon)\log\left(n(1-2\varepsilon)\right)\operatorname{Tr}\left(\Sigma\right)}\right) - \sigma\varepsilon n \tag{142}$$

From the definition of stationary point, we have

$$\eta^2\left(n(1-2\varepsilon)\lambda_{\min}\left(\Sigma\right) - 2\operatorname{Tr}\left(\Sigma\right)\sqrt{n(1-2\varepsilon)} - \beta\right) - \eta O\left(\sigma\sqrt{n(1-2\varepsilon)\log\left(n(1-2\varepsilon)\right)\operatorname{Tr}\left(\Sigma\right)}\right) - \sigma\varepsilon n \leq 0 \tag{143}$$

Therefore, when Equation (143) does not hold, we have a contradiction. It thus follows from upper bounding the positive solution of the quadratic equation,

$$\begin{aligned}\eta &\leq (\sigma\varepsilon n)^{1/2}\left(n(1-2\varepsilon)\left(\lambda_{\min}\left(\Sigma\right) - \frac{2\operatorname{Tr}\left(\Sigma\right)}{\sqrt{n\left(1-2\varepsilon\right)}}\right) - \beta\right)^{-1/2} \\ &+ O\left(\sigma\sqrt{n(1-2\varepsilon)\log\left(n(1-2\varepsilon)\right)\operatorname{Tr}\left(\Sigma\right)}\right)\left(n(1-2\varepsilon)\left(\lambda_{\min}\left(\Sigma\right) - \frac{2\operatorname{Tr}\left(\Sigma\right)}{\sqrt{n\left(1-2\varepsilon\right)}}\right) - \beta\right)^{-1}\end{aligned} \tag{144}$$

Then for some constant $c_1 \in (0,1)$, if $n \geq \frac{8\operatorname{Tr}(\Sigma)^2}{\lambda_{\min}(\Sigma)(1-c_1)^2(1-2\varepsilon)} + \frac{8\beta}{(1-c_1)^2(1-2\varepsilon)}$, we have

$$\eta \leq \left(\frac{\sigma\varepsilon n}{c_1 n(1-2\varepsilon)\lambda_{\min}\left(\Sigma\right)}\right)^{1/2} + \frac{O\left(\sigma\sqrt{\log\left(n(1-2\varepsilon)\right)\operatorname{Tr}\left(\Sigma\right)}\right)}{c_1\sqrt{n(1-2\varepsilon)}\lambda_{\min}\left(\Sigma\right)} \tag{145}$$

we therefore see as n goes large, $c_1 \to 1$, and we have in the worst case

$$\mathop{\mathbb{E}}_{\mathcal{D}\sim\hat{\mathbb{P}}} \|f_{\hat{w}} - f^*_w\|_{\mathcal{H}} \leq O\left(\sqrt{\frac{\varepsilon}{1-2\varepsilon}}\frac{\sigma}{\lambda_{\min}\left(\Sigma\right)}\right) \tag{146}$$

This completes the proof. ■

C.4 Proof of Corollary 15

We follow the same framework as our proof for kernelized linear regression, we will simply give the new constants. Assuming the uncorrupted covariates, $x_i \sim \mathcal{N}(f_{\mathrm{W}}(x_i)ero, \Sigma)$. To simplify notation, let us define $\tilde{n} \triangleq n(1-2\varepsilon)$ to represent the absolute minimum number of uncorrupted points in the Subquantile. We then have,

$$\begin{aligned}\mathop{\mathbb{E}}_{\mathcal{D}\sim\hat{\mathbb{P}}} \lim_{\lambda\to\infty} \left(\Phi_\lambda(w) - \Phi_\lambda(w^*)\right) &\overset{\text{lem. 29}}{\geq} \eta^2\left(\tilde{n}\lambda_{\min}\left(\Sigma\right) - \mathbb{E}\left\|\sum_{i\in S\cap P} x_i x_i^\top - \Sigma\right\|_2\right) - \mathop{\mathbb{E}}_{\xi,\mu_i\sim\mathbb{P}}\left\|\sum_{i\in S\cap P}\mu_i x_i\right\|_2 - \mathop{\mathbb{E}}_{\mu_i\sim\mathbb{P}}\sum_{i\in P\setminus S}\mu_i^2 \\ &\overset{\text{lems. 20, 23 and 25}}{\geq} \eta^2\left(\tilde{n}\lambda_{\min}\left(\Sigma\right) - \sqrt{\tilde{n}}\left(2\sqrt{3}\operatorname{Tr}\left(\Sigma\right)\right)\right) - \eta O\left(\sigma\sqrt{\tilde{n}\log(\tilde{n})\operatorname{Tr}\left(\Sigma\right)}\right) - \varepsilon n\sigma^2\end{aligned} \tag{147}$$

Then from a similar contradiction idea and upper bounding the quadratic, we have in expectation

$$\eta \overset{\text{thm. 14}}{\leq} O\left(\sigma\sqrt{\tilde{n}\log(\tilde{n})\operatorname{Tr}\left(\Sigma\right)}\right)\left(\tilde{n}\lambda_{\min}\left(\Sigma\right) - \sqrt{\tilde{n}}\left(2\sqrt{3}\operatorname{Tr}\left(\Sigma\right)\right) - \beta\right)^{-1} + \sigma\sqrt{\tilde{n}\frac{\varepsilon}{1-2\varepsilon}}\left(\tilde{n}\lambda_{\min}\left(\Sigma\right) - \sqrt{\tilde{n}}\left(2\sqrt{3}\operatorname{Tr}\left(\Sigma\right)\right) - \beta\right)^{-1/2}$$

We then have for a constant $c_2 \in (0,1)$, if $n \geq \frac{54\operatorname{Tr}(\Sigma)}{(1-c_2)^2(1-2\varepsilon)\lambda_{\min}^2(\Sigma)} + 2\beta$, it follows

$$\eta \leq \sqrt{\frac{\sigma^2\varepsilon}{(1-2\varepsilon)c_2\lambda_{\min}(\Sigma)}} + \frac{O\left(\sigma\sqrt{\log\left(n(1-2\varepsilon)\right)\operatorname{Tr}\left(\Sigma\right)}\right)}{\sqrt{n(1-2\varepsilon)}c_2\lambda_{\min}(\Sigma)} \tag{148}$$

We thus see as n goes large, $c_2 \to 1$ and we will have in worst case,

$$\mathop{\mathbb{E}}_{\mathcal{D}\sim\hat{\mathbb{P}}} \|\hat{w} - w^*\|_2 \leq O\left(\frac{\gamma\sigma}{\sqrt{\lambda_{\min}(\Sigma)}}\right) \tag{149}$$

where $\gamma \triangleq \sqrt{\frac{|P\setminus S|}{|S\cap P|}}$. Obtaining the same asymptotic bound as in the kernelized regression case. This completes the proof. ∎

D Kernlized Binary Classification

D.1 L-Lipschitz Constant and β-Smoothness Constant

Lemma 30. *(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:*

$$|g\left(t, f_w\right) - g\left(t, f_{\hat{w}}\right)| \leq L\left\|f_w - f_{\hat{w}}\right\|_{\mathcal{H}} \tag{150}$$

where

$$L = \frac{1}{np}\sum_{i\in X}\sqrt{k(x_i,x_i)} = \frac{1}{np}\operatorname{Tr}\left(\mathrm{K}\right) \tag{151}$$

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

$$\|\nabla_{\mathrm{f}} g(t, f_w)\|_{\mathcal{H}} = \left\|\frac{1}{np}\sum_{i=1}^{n}\mathbb{I}\{t \geq (1-y_i)\log(f_w(x_i))\}\left(y_i - \sigma\left(f_w(x_i)\right)\right)\cdot k(x_i,\cdot)\right\|_{\mathcal{H}} \tag{152}$$

$$\stackrel{(i)}{\leq} \frac{1}{np}\sum_{i\in S}\|(y_i - \sigma\left(f_w(x_i)\right))\cdot k(x_i,\cdot)\|_{\mathcal{H}} \stackrel{(ii)}{\leq} \frac{1}{np}\sum_{i\in S}|y_i - \sigma\left(f_w(x_i)\right)|\,\|k(x_i,\cdot)\|_{\mathcal{H}} \stackrel{(iii)}{\leq} \frac{1}{np}\sum_{i=1}^{n}\sqrt{k(x_i,x_i)}$$

(i) follows from the triangle inequality. (ii) follows from the Cauchy-Schwarz inequality. (iii) follows from the fact that $y_i \in \{0,1\}$ and range$(\sigma) \in [0,1]$. This completes the proof. ∎

Lemma 31. *(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:*

$$\|\nabla_f g\left(t, f_w\right) - \nabla_f g\left(t, f_{\hat{w}}\right)\| \leq \beta\left\|f_w - f_{\hat{w}}\right\|_{\mathcal{H}} \tag{153}$$

where

$$\beta = \frac{1}{4p}\sum_{i=1}^{n}k(x_i,x_i) = \frac{1}{4p}\operatorname{Tr}\left(\mathrm{K}\right) \tag{154}$$

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

$$\left\|\nabla_f^2 g(t,f_w)\right\|_{\mathrm{op}} = \frac{1}{np}\sum_{i=1}^{n}\mathbb{I}\left\{t \geq (1-y_i)\log(f_w(x_i))\right\}\sigma\left(f_w(x_i)\right)\left(1-\sigma\left(f_w(x_i)\right)\right)\|\phi(x_i)\otimes\phi(x_i)\|_{\mathrm{op}} \tag{155}$$

$$\leq \frac{1}{np}\sum_{i=1}^{n}|\sigma\left(f_w(x_i)\right)\left(1-\sigma\left(f_w(x_i)\right)\right)|\,\|\phi(x_i)\|_{\mathrm{op}}^2 \stackrel{(i)}{\leq} \frac{1}{4np}\sum_{i=1}^{n}k(x_i,x_i) = \frac{1}{4np}\operatorname{Tr}\left(\mathrm{K}\right) \tag{156}$$

(i) 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. ∎

Lemma 32. *Assume $f_{\hat{w}}$ is a first-order stationary point as defined in Definition 9. If $\|f_{\hat{w}} - f_{w^*}\|_{\mathcal{H}} \geq \eta$, then it follows*

$$\mathop{\mathbb{E}}_{\mathcal{D}\sim\hat{\mathbb{P}}} \|f_{\hat{w}} - f_w^*\|_{\mathcal{H}} \leq O\left(\frac{\sqrt{n(1-2\varepsilon)\operatorname{Tr}(\Sigma)} + \sqrt{n\varepsilon Q_k}}{n(1-2\varepsilon)c_4\lambda_{\min}(\Sigma)}\right) \tag{157}$$

Proof. By the Lemma statement, we have $f_{\hat{w}}$ is a stationary point, i.e. $0 \in \partial\Phi(f_{\hat{w}})$. This implies for all $\mathsf{f}_w \in \mathcal{K}$, we have $\Phi(f_{\hat{w}}) \leq \Phi(f_w)$. As Φ is differentiable, we have the first-order stationary condition, which is $\nabla\Phi(f_{\hat{w}})(f_{\hat{w}} - f_w) \leq 0$ or for all $w \in \mathcal{K}$. We assume $f_w^* \in \mathcal{K}$. Let S be the Subquantile set for $f_{\hat{w}}$. We will proceed by contradiction, assume $\|f_{\hat{w}} - f_w^*\|_{\mathcal{H}} \geq \eta$. Then, we have

$$(\nabla_f g(f_{\hat{w}}, t))(f_{\hat{w}} - f_w^*) = (f_{\hat{w}} - f_w^*)\left(\sum_{i\in S}(\sigma(f_{\hat{w}}(x_i)) - y_i)\phi(x_i)\right) \tag{158}$$

$$= (f_{\hat{w}} - f_w^*)\left(\sum_{i\in S}(\sigma(f_{\hat{w}}(x_i)) - \sigma(f_w^*(x_i)) + \sigma(f_w^*(x_i)) - y_i)\phi(x_i)\right) \tag{159}$$

$$\stackrel{(i)}{\geq} \underbrace{(f_{\hat{w}} - f_w^*)\left(\sum_{S\cap P}(\sigma(f_{\hat{w}}(x_i)) - \sigma(f_w^*(x_i)))\phi(x_i)\right)}_{B_1} + \underbrace{(f_{\hat{w}} - f_w^*)\left(\sum_{i\in S}(\sigma(f_w^*(x_i)) - y_i)\phi(x_i)\right)}_{B_2} \tag{160}$$

(i) follows from noting $\sigma(\cdot)$ is a monotonically increasing function. Let us now consider the function $h : \mathcal{H} \to \mathbb{R}$ defined as $h(f_w) = \sum_{i\in S\cap P}\log(1+\exp(f_w(x_i)))$. We then have $h'(f_w) = \sum_{i\in S\cap P}\sigma(f_w(x_i))\phi(x_i)$, from which we have $h''(f_w) = \sum_{i\in S\cap P}\sigma(f_w(x_i))(1-\sigma(f_w(x_i)))(\phi(x_i)\otimes\phi(x_i))$. We can then note h is strongly convex with $\mu = \Omega(\lambda_{\min}(\sum_{i\in S\cap P}\phi(x_i)\otimes\phi(x_i)))$. Then from the properties of strongly convex functions, we have

$$\sum_{i\in S\cap P}(f_{\hat{w}}(x_i) - f_w^*(x_i))(\sigma(f_{\hat{w}}(x_i)) - \sigma(f_w^*(x_i))) \gtrsim \lambda_{\min}\left(\sum_{i\in S\cap P}\phi(x_i)\otimes\phi(x_i)\right)\|f_w^* - f_{\hat{w}}\|_{\mathcal{H}}^2 \tag{161}$$

Then from the Cauchy-Schwarz Inequality, we have

$$\sum_{i\in S}(f_w^*(x_i) - f_{\hat{w}}(x_i))(y_i - \sigma(f_w^*(x_i))) \leq \max_{i\in S}|y_i - \sigma(f_w^*(x_i))|\left\langle\sum_{j\in X}(w_j^* - \hat{w}_j)\phi(x_j), \sum_{i\in S}\phi(x_i)\right\rangle \tag{162}$$

$$\leq \|f_w^* - f_{\hat{w}}\|_{\mathcal{H}}\left\|\sum_{i\in S}\phi(x_i)\right\|_{\mathcal{H}} \leq \|f_w^* - f_{\hat{w}}\|_{\mathcal{H}}\left(\left\|\sum_{i\in S\cap P}\phi(x_i)\right\|_{\mathcal{H}} + \left\|\sum_{i\in S\cap Q}\phi(x_i)\right\|_{\mathcal{H}}\right) \tag{163}$$

for a small positive constant we denote c_3. This completes the proof. ∎

D.2 Proof of Theorem 16

Proof. From Lemma 32, we have in expectation

$$\mathop{\mathbb{E}}_{\mathcal{D}\sim\hat{\mathbb{P}}}(\nabla_f g(f_{\hat{w}}, t))(f_w^* - f_{\hat{w}}) \stackrel{\text{lem. 32}}{\geq} c_3\left(n(1-2\varepsilon)\lambda_{\min}(\Sigma) - \mathop{\mathbb{E}}_{x_i\sim\mathbb{P}}\left\|\sum_{i\in S\cap P}\phi(x_i)\otimes\phi(x_i) - \Sigma\right\|\right)\|f_{\hat{w}} - f_w^*\|_{\mathcal{H}}^2$$
$$- \left(\sqrt{n(1-2\varepsilon)\operatorname{Tr}(\Sigma)} + \sqrt{n\varepsilon Q_k}\right)\|f_{\hat{w}} - f_w^*\|_{\mathcal{H}} \tag{164}$$

We will lower bound the constant we introduced in Equation (161) and call it c_3, recall for $f \in \mathcal{K}$, we have $\|f\|_{\mathcal{H}} \leq R$ and $P_k \triangleq \max_{i\in P} k(x_i, x_i)$.

$$\mathop{\mathbb{E}}_{\mathcal{D}\sim\hat{\mathbb{P}}} c_3 \stackrel{(161)}{=} \mathop{\mathbb{E}}_{\mathcal{D}\sim\hat{\mathbb{P}}}\min_{i\in S\cap P}(1-\sigma(f_{\hat{w}}(x_i)))\sigma(f_{\hat{w}}(x_i)) \tag{165}$$

$$\geq \mathop{\mathbb{E}}_{\mathcal{D}\sim\hat{\mathbb{P}}}\left(1 - \sigma(R\max_{i\in P}k(x_i, x_i))\right)\sigma(R\max_{i\in P}k(x_i, x_i)) \tag{166}$$

$$\geq \mathop{\mathbb{E}}_{\mathcal{D}\sim\hat{\mathbb{P}}} \frac{\sigma\left(-R \max_{i\in P} k\left(x_i, x_i\right)\right)}{2} \overset{(i)}{\gtrsim} \exp\left(-R \mathop{\mathbb{E}}_{\mathcal{D}\sim\hat{\mathbb{P}}}\left[\max_{i\in P} k\left(x_i, x_i\right)\right]\right) \tag{167}$$

$$\overset{\text{lem. 22}}{\geq} \exp\left(-RC_8\left(\operatorname{Tr}\left(\Sigma\right) + \log n\right)\right) \tag{168}$$

(i) follows from Jensen's Inequality as $\exp(-x)$ is a convex functon. Then we have from the definition of a stationary point, $\nabla_f g\left(f_{\hat{w}}, t\right)\left(f_{\hat{w}} - f^*_w\right) \leq 0$ when

$$\mathop{\mathbb{E}}_{\mathcal{D}\sim\hat{\mathbb{P}}} \|f_{\hat{w}} - f^*_w\|_{\mathcal{H}} \leq O\left(\frac{\sqrt{n(1-2\varepsilon)\operatorname{Tr}\left(\Sigma\right)} + \sqrt{n\varepsilon Q_k}}{\exp\left(-RC_8\left(\operatorname{Tr}\left(\Sigma\right) + \log n\right)\right)\left(n(1-2\varepsilon)\lambda_{\min}(\Sigma) - 2\sqrt{n(1-2\varepsilon)}\operatorname{Tr}\left(\Sigma\right)\right)}\right) \tag{169}$$

If $n \geq \frac{4\operatorname{Tr}(\Sigma)}{\lambda_{\min}(\Sigma)(1-2\varepsilon)(1-c_4)}$ for $c_4 \in (0,1)$, then we have

$$\mathop{\mathbb{E}}_{\mathcal{D}\sim\hat{\mathbb{P}}} \|f_{\hat{w}} - f^*_w\|_{\mathcal{H}} \leq O\left(\frac{\sqrt{\operatorname{Tr}(\Sigma)} + \sqrt{Q_k}}{\sqrt{n(1-2\varepsilon)}\exp\left(-R\left(\operatorname{Tr}\left(\Sigma\right) + \log n\right)\right)\lambda_{\min}(\Sigma)}\right) \tag{170}$$

This completes the proof as we see we have $O(1/\sqrt{n})$ convergence. ■

E Proofs for Kernelized Multi-Class Classification

E.1 L-Lipschitz Constant and β-Smoothness Constant

Lemma 33. *Let $x_1, x_2, \ldots, x_n \sim \hat{\mathbb{P}}$. It then follows for a $f_w \in \mathcal{K}$, then $g(t, f_w)$ is L-Lipschitz and β-Smooth for constants $L = \frac{1}{np}\sum_{i=1}^n \sqrt{k(x_i, x_i)}$ and $\beta = \frac{1}{np}\operatorname{Tr}\left(\mathrm{K}\right)$.*

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

$$\frac{\partial}{\partial w_k}\left(\ell\left(x_i, y_i; f_{\mathrm{W}}\right)\right) = \begin{cases} -\phi(x_i)\mathrm{softmax}(f_{\mathrm{W}}(x_i))_k & \text{if } k = y_i \\ \phi(x_i)\left(1 - \mathrm{softmax}(f_{\mathrm{W}}(x_i))_k\right) & \text{if } k \neq y_i \end{cases} \tag{171}$$

Our proof then follows similarly to the proof for Lemma 30. We utilize $\odot$ to denote entry wise multiplication, i.e $x \cdot y$ indicates y is multiplied to each element of x.

$$\begin{aligned} \|\nabla_f g(t, f_{\mathrm{W}})\|_{\mathcal{H}} &= \left\|\frac{1}{np}\sum_{i=1}^n \mathbb{I}\left\{-\log\left(\mathrm{softmax}\left(f_{\mathrm{W}}(x_i)\right)_{y_i}\right) \geq t\right\}\left(\mathrm{softmax}\left(f_{\mathrm{W}}(x_i)\right) - y_i\right) \odot k(x_i, \cdot)\right\| \\ &\leq \frac{1}{np}\sum_{i=1}^n \|\left(\mathrm{softmax}\left(f_{\mathrm{W}}(x_i)\right) - y_i\right) \odot k(x_i, \cdot)\| \leq \frac{1}{np}\sum_{i=1}^n \|k(x_i, \cdot)\|_{\mathcal{H}} = \frac{1}{np}\sum_{i=1}^n \sqrt{k(x_i, x_i)} \end{aligned} \tag{172}$$

This gives the L-Lipschitz Constant.
We upper bound the operator norm of the Hessian to find the β-smoothness constant.

$$\|\nabla_f g\left(t, f_{\mathrm{W}}\right)\|_{\mathrm{op}} \tag{173}$$

$$= \left\|\frac{1}{np}\sum_{i=1}^n \mathbb{I}\left\{-\log\left(\mathrm{softmax}\left(f_{\mathrm{W}}(x_i)\right)_{y_i}\right) \geq t\right\}\left(\mathrm{diag}(\mathrm{softmax}(f_{\mathrm{W}}(x_i)) - \mathrm{softmax}(f_{\mathrm{W}}(x_i)\mathrm{softmax}(f_{\mathrm{W}}(x_i)^\top\right) \odot \left(\phi(x_i) \otimes \phi(x_i)\right)\right\|_{\mathrm{op}}$$

$$\leq \frac{1}{np}\sum_{i=1}^n \left\|\left(\mathrm{diag}(\mathrm{softmax}(f_{\mathrm{W}}(x_i)) - \mathrm{softmax}(f_{\mathrm{W}}(x_i)\mathrm{softmax}(f_{\mathrm{W}}(x_i)^\top\right) \odot \left(\phi(x_i) \otimes \phi(x_i)\right)\right\|_{\mathrm{op}} \tag{174}$$

$$\leq \frac{1}{np}\sum_{i=1}^n \left\|\mathrm{diag}(\mathrm{softmax}(f_{\mathrm{W}}(x_i)) - \mathrm{softmax}(f_{\mathrm{W}}(x_i)\mathrm{softmax}(f_{\mathrm{W}}(x_i)^\top\right\|_{\mathrm{op}} \|\phi(x_i)\|^2_{\mathcal{H}} \tag{175}$$

$$\leq \frac{1}{np}\sum_{i=1}^n k(x_i, x_i) = \frac{1}{np}\operatorname{Tr}\left(\mathrm{K}\right) \tag{176}$$

■

Lemma 34. [3] *Assume $f_{\hat{w}}$ is a first-order stationary point as defined in Definition 9. If $\|f_{\hat{w}} - f_{w^*}\|_{\mathcal{H}} \geq \eta$, then it follows*

$$\mathop{\mathbb{E}}_{\mathcal{D}\sim\hat{\mathbb{P}}} \|f_{\hat{w}} - f^*_w\|_{\mathcal{H}} \leq \tag{177}$$

Proof.[4] We follow the similar set up as in Lemma 32. We have $f_{\hat{\mathrm{W}}}$ is a stationary point, i.e. $0 \in \partial\Phi(f_{\hat{\mathrm{W}}})$. This implies for all $\mathsf{f}_{\mathrm{W}} \in \mathcal{K}$, we have $\Phi(f_{\hat{\mathrm{W}}}) \leq \Phi(f_{\mathrm{W}})$. As Φ is differentiable, we have the first-order stationary condition, which is $\nabla\Phi(f_{\hat{\mathrm{W}}})(f_{\hat{\mathrm{W}}} - f_{\mathrm{W}}) \leq 0$ or for all $\mathrm{W} \in \mathcal{K}$. We assume $f^*_{\mathrm{W}} \in \mathcal{K}$. Let S be the Subquantile set for $f_{\hat{\mathrm{W}}}$. We will proceed by contradiction, assume $\left\|f_{\hat{\mathrm{W}}} - f^*_{\mathrm{W}}\right\|_{\mathcal{H}} \geq \eta$. Then, we have

$$\left\langle \nabla_f g\left(f_{\hat{\mathrm{W}}}, t\right), f_{\hat{\mathrm{W}}} - f^*_{\mathrm{W}}\right\rangle_{\mathcal{H}} = \left\langle f_{\hat{\mathrm{W}}} - f^*_{\mathrm{W}}, \sum_{i\in S}\left(\mathrm{softmax}\left(f_{\hat{\mathrm{W}}}(x_i)\right) - y_i\right) \odot k(x_i,\cdot)\right\rangle_{\mathcal{H}} \tag{178}$$

$$= \left\langle f_{\hat{\mathrm{W}}} - f^*_{\mathrm{W}}, \sum_{i\in S}\left(\mathrm{softmax}\left(f_{\hat{\mathrm{W}}}(x_i)\right) - \mathrm{softmax}\left(f^*_{\mathrm{W}}(x_i)\right)\right) \odot k(x_i,\cdot)\right\rangle_{\mathcal{H}} + \left\langle f_{\hat{\mathrm{W}}} - f^*_{\mathrm{W}}, \sum_{i\in S}\left(\mathrm{softmax}\left(f^*_{\mathrm{W}}(x_i)\right) - y_i\right) \odot k(x_i,\cdot)\right\rangle_{\mathcal{H}}$$

Let us now consider the function $h : \mathcal{H}\times\cdots\times\mathcal{H} :\to \mathbb{R}^n$ defined as $h(f_{\mathrm{W}}) = \sum_{i\in P\cap S}\log(\sum_{j=1}^{|\mathcal{Y}|}\exp(f_{w_j}(x_i) - y_{i,j}))$. We then have $\nabla h(f_{\mathrm{W}}) = \sum_{i\in P\cap S}\mathrm{softmax}(f_{\mathrm{W}}(x_i) - y_i)\odot\phi(x_i)$. From which it follows $\nabla^2 h(x_i) = \sum_{i\in P\cap S}(\mathrm{diag}(\mathrm{softmax}(f_{\mathrm{W}}(x_i))) - \mathrm{softmax}(f_{\mathrm{W}}(x_i))\mathrm{softmax}(f_{\mathrm{W}}(x_i))^\top)\odot(\phi(x_i)\otimes\phi(x_i))$. This function is not strictly convex, as corroborated in [GP17]. If the softmax returns the same value for all inputs, then we note $\alpha^2 1^\top\nabla^2 h(f_{\mathrm{W}})1 = 0$ where $\alpha = \frac{1}{|\mathcal{Y}|}$. Since we assume random initialization of the functions in f_{W}, this event does not occur almost surely. Therefore, we have the function is strictly convex over the domain.

$$\text{Equation (178) LHS} \geq \Omega\left(\left\|f_{\hat{\mathrm{W}}} - f^*_{\mathrm{W}}\right\|^2_{\mathcal{H}}\right) + \left\langle f_{\hat{\mathrm{W}}} - f^*_{\mathrm{W}}, \sum_{i\in S}\left(\mathrm{softmax}\left(f^*_{\mathrm{W}}(x_i)\right) - y_i\right)\odot k(x_i,\cdot)\right\rangle_{\mathcal{H}} \tag{179}$$

Now we will upper bound C_2.

$$C_2 = \sum_{i\in S}\left\langle f_{\hat{\mathrm{W}}}(x_i) - f^*_{\mathrm{W}}(x_i), \mathrm{softmax}\left(f^*_{\mathrm{W}}(x_i)\right) - y_i\right\rangle \tag{180}$$

$$= \sum_{j\in\mathcal{Y}}\sum_{i\in S}\left(f_{\hat{w}_j}(x_i) - f^*_{w_j}(x_i)\right)\left(\mathrm{softmax}\left(f^*_{w_j}(x_i)\right) - y_{i,j}\right) \tag{181}$$

$$\overset{(i)}{\leq} \sqrt{\sum_{j\in\mathcal{Y}}\sum_{i\in S}\left(f_{\hat{w}_j}(x_i) - f^*_{w_j}(x_i)\right)^2}\sqrt{\sum_{j\in\mathcal{Y}}\sum_{i\in S}\left(\mathrm{softmax}\left(f^*_{w_j}(x_i)\right) - y_{i,j}\right)^2} \tag{182}$$

$$\leq \left\|f_{\hat{\mathrm{W}}} - f^*_{\mathrm{W}}\right\|_{\mathcal{H}}\sqrt{2n(1-\varepsilon)\lambda_{\max}\left(\sum_{i\in S}\phi(x_i)\otimes\phi(x_i)\right)} \tag{183}$$

Where (i) follows from Hölder's Inequality [H89̈].

[3] In Progress
[4] In Progress