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

Anonymous Authors[1]

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 (Laguel et al., 2021; Rockafellar et al., 2014). Our intuition is to learn a more favorable representation of the *majority* class, thus we propose to optimize over the p-subquantile of the loss in the dataset. In particular, we study the Huber Contamination Problem for Kernel Learning where the distribution is formed as, $\hat{\mathbb{P}} = (1 - \varepsilon)\mathbb{P} + \varepsilon\mathbb{Q}$, and we want to find the function $\inf_f \mathbb{E}_{\mathbf{x} \in \mathbb{P}}[\ell_f(\mathbf{x})]$, from the noisy distribution, $\hat{\mathbb{P}}$. We assume the adversary has knowledge of the true distribution of $\mathbb{P}$, and is able to corrupt the covariates and the labels of ε samples. To our knowledge, we are the first to study the problem of general kernel learning in the Huber Contamination Model. In our theoretical analysis, we analyze our non-convex concave objective function with the Moreau Envelope. We show (i) a stationary point with respect to the Moreau Envelope is a good point and (ii) we can reach a stationary point with gradient descent methods. Further, we analyze accelerated gradient methods for the non-convex concave minimax optimization problem. We empirically test Kernel Ridge Regression and Kernel Classification on various state of the art datasets and show Subquantile Minimization gives strong results. Furthermore, we run experiments on various datasets and compare with the state-of-the-art algorithms to show the superior performance of Subquantile Minimization.

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

1 Introduction

There has been extensive study of algorithms to learn the target distribution from a Huber ϵ-Contaminated Model for a Generalized Linear Model (GLM), (Diakonikolas et al., 2019; Awasthi et al., 2022; Li et al., 2021; Osama et al., 2020; Fischler & Bolles, 1981) as well as for linear regression (Bhatia et al., 2017; Mukhoty et al., 2019). Robust Statistics has been studied extensively (Diakonikolas & Kane, 2023) for problems such as high-dimensional mean estimation (Prasad et al., 2019; Cheng et al., 2020) and Robust Covariance Estimation (Cheng et al., 2019; Fan et al., 2018). Recently, there has been an interest in solving robust machine learning problems by gradient descent (Prasad et al., 2018; Diakonikolas et al., 2019). Subquantile minimization aims to address the shortcomings of standard ERM in applications of noisy/corrupted data (Khetan et al., 2018; Jiang et al., 2018). In many real-world applications, the covariates have a non-linear dependence on labels (Abu-Mostafa et al., 2012, Section 3.4). In which case it is suitable to transform the covariates to a different space utilizing kernels (Hofmann et al., 2008). Therefore, in this paper we consider the problem of Robust Learning for Kernel Learning.

Definition 1 (Huber ϵ-Contamination Model (Huber & Ronchetti, 2009))**.** Given a corruption parameter $0 < \epsilon < 0.5$, a data matrix, $\mathbf{X}$ and labels $\mathbf{y}$. An adversary is allowed to inspect all samples and modify ϵn samples arbitrarily. The algorithm is then given the ϵ-corrupted data matrix $\mathbf{X}$ and $\mathbf{y}$ as training data.

Current approaches for robust learning across various machine learning tasks often use gradient descent over a robust objective, (Li et al., 2021). 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 (Hu et al., 2020) where there has been an analysis in the Binary Classification Task. We show Subquantile Minimization reduces to the same objective in (Hu et al., 2020). 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, (Cui et al., 2021; Dicker, 2016). In (Cui et al., 2021), the feature space, $\phi(\mathbf{x}_i) \sim \mathcal{N}(0, \mathbf{\Sigma})$ where $\mathbf{\Sigma}$ 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 (Mercer, 1909), we are able to say $\mathrm{Tr}(\mathbf{\Sigma}) < \infty$. We use this fact extensively in our infinite-dimensional concentration inequalities.

Theorem 2. *(Informal). Let the dataset be given as $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$ such that the labels and features of ϵn samples are arbitrarily corrupted by an adversary. Assume Subquantile Minimization returns $f_{\widehat{\mathbf{w}}}$ for $n \geq \frac{(1-2\epsilon)(C_k\|\mathbf{\Sigma}\|_{\mathrm{op}} + \beta)}{(1-c_1)\lambda_{\min}(\mathbf{\Sigma})} + \sqrt{\beta}$ for a constant $c_1 \in (0,1)$ such that for*

Kernelized Regression:

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

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

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

Assume $n \geq \left(\frac{\mathrm{Tr}(\mathbf{\Sigma})}{\lambda_{\min}(\mathbf{\Sigma})(1-c)}\right)^2$ for a constant $0 < c < 1$.

Kernel Multi-Class Classification:

$$\underset{\mathcal{D} \sim \hat{\mathbb{P}}}{\mathbb{E}} \|f_{\widehat{\mathbf{W}}} - f_{\mathbf{W}}^*\|$$
$$\leq O\left(\frac{\sqrt{n(1-\epsilon)\mathrm{Tr}(\mathbf{\Sigma})} + \sqrt{n\epsilon Q_k}}{n(1-2\epsilon)\lambda_{\min}(\mathbf{\Sigma})}\right) \quad (3)$$

1.1 Related Work

The idea of iterative thresholding algorithms for robust learning tasks dates back to 1806 by Legendre (Legendre, 1806). 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 (Bhatia et al., 2015). 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 (Bhatia et al., 2015) (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 (Shen & Sanghavi, 2019b). 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 (Awasthi et al., 2022). 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 (Hu et al., 2020). 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, (Fan et al., 2017). 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 (Hu et al., 2023).

The iterative trimmed loss framework with batch Stochastic Gradient Descent (SGD) is analyzed in (Shen & Sanghavi, 2019a). 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, we use our experiments to demonstrate the practicality of our 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.

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

The p-subquantile of the risk is then given by

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

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

$$\min_{f_{\mathbf{w}} \in \mathcal{K}} \max_{t \in \mathbb{R}} \left\{ g(t, f_{\mathbf{w}}) \triangleq t - \sum_{i=1}^n \left(t - (f_{\mathbf{w}}(\mathbf{x}_i) - y_i)^2 \right)^+ \right\} \qquad (6)$$

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 $\mathbf{w}$. Thus, to solve this problem we use the iterations from Equation 11 in (Razaviyayn et al., 2020). Let $\mathrm{Proj}_{\mathcal{K}}$ be the projection of a function on to the convex set $\mathcal{K} \triangleq \{f \in \mathcal{H} : \|f\|_{\mathcal{H}} \leq R\}$, then our update steps are

$$t^{(k+1)} = \arg\max_{t \in \mathbb{R}} g(f_{\mathbf{w}}^{(k)}, t) \qquad (7)$$

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

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

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

Proof is given in Appendix B.1. From Lemma 3, we see that 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 4. *Let* $\hat{\nu}_i \triangleq f(\mathbf{x}_i; \mathbf{w}, y_i)$ *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* $\nabla_{\mathbf{w}} g(t^{(k)}, f_{\mathbf{w}}^{(k)}) = \frac{1}{np} \sum_{i=1}^{np} \nabla f(\mathbf{x}_i; f_{\mathbf{w}}^{(k)}, y_i)$

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,

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

We then have the approximation of g as

$$\tilde{g}_\lambda(t, f_{\mathbf{w}}) \triangleq t - \sum_{i=1}^n \zeta_\lambda\left(t - \ell(f_{\mathbf{w}}; \mathbf{x}_i, y_i)\right)$$
$$= t - \frac{1}{np} \sum_{i=1}^n \frac{1}{\lambda} \log\left(1 + \exp\left(\lambda\left(t - \ell(f_{\mathbf{w}}; \mathbf{x}_i, y_i)\right)\right)\right) \qquad (10)$$

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 $f_{\mathbf{w}}$.

Lemma 5 (Lipschitz continuous gradient). *Let* $f_{\mathbf{w}}, f_{\tilde{\mathbf{w}}} \in \mathcal{K}$, *then we have for any* $\lambda > 0$,

$$|\nabla_f \tilde{g}_\lambda(t, f_{\mathbf{w}}) - \nabla_f \tilde{g}_\lambda(t, f_{\tilde{\mathbf{w}}})| \leq \beta \|f_{\mathbf{w}} - f_{\tilde{\mathbf{w}}}\|_{\mathcal{H}} \qquad (11)$$

where

$$\beta = \frac{1}{np} \sum_{i=1}^n \left\|\nabla_f^2 \ell(f_{\mathbf{w}}; \mathbf{x}_i, y_i)\right\|_{\mathrm{op}} \qquad (12)$$

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.

Input: Iterations: T, Quantile: p; Data Matrix: $\mathbf{X} \in \mathbb{R}^{n \times d}, n \gg d$; Labels: $\mathbf{y} \in \mathbb{R}^{n \times 1}$; Learning Rate schedule: $\alpha^{(1)}, \cdots, \alpha^{(T)}$

Output: Function in RKHS: $\hat{f}_{\mathbf{w}}$

(1) Initialize the weights $w_i^{(0)} \sim \text{Unif} \left[-\sqrt{\frac{6}{n}}, \sqrt{\frac{6}{n}} \right]$ for all $i \in [n]$.

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

    (3) Find the Subquantile denoted as $S^{(k)}$ as the set of $(1 - \epsilon)n$ elements with the lowest error with respect to the loss function.

    (4) Update the gradient in accordance with the kernel learning problem.

$$\nabla_f g \left(t^{(k+1)}, f_{\mathbf{w}}^{(k)} \right) \leftarrow \sum_{i \in S^{(k)}} \left(f_{\mathbf{w}}^{(k)}(\mathbf{x}_i) - y_i \right) \cdot k(\mathbf{x}_i, \cdot) \qquad \text{(Regression)}$$

$$\nabla_f g \left(t^{(k+1)}, f_{\mathbf{w}}^{(k)} \right) \leftarrow \sum_{i \in S^{(k)}} \left(\sigma \left(f_{\mathbf{w}}^{(k)}(\mathbf{x}_i) \right) - y_i \right) \cdot k(\mathbf{x}_i, \cdot) \qquad \text{(Binary Classification)}$$

$$\nabla_f g \left(t^{(k+1)}, f_{\mathbf{W}}^{(k)} \right) \leftarrow \sum_{i \in S^{(k)}} \left(\text{softmax} \left(f_{\mathbf{W}}^{(k)}(\mathbf{x}_i) \right) - \mathbf{y}_i \right) \odot k(\mathbf{x}_i, \cdot) \qquad \text{(Multi-Class Classification)}$$

(5) Perform Projected Standard Gradient Descent to find the next iterate

$$f_{\mathbf{w}}^{(k+1)} \leftarrow \text{Proj}_{\mathcal{K}} \left(f_{\mathbf{w}}^{(k)} - \alpha^{(k)} \nabla_f g \left(t^{(k+1)}, f_{\mathbf{w}}^{(k)} \right) \right)$$

(6) Pick t uniformly at random from $[\![T]\!]$

(7) **Return:** $f_{\mathbf{w}}^{(t)}$

Figure 1: Subquantile Minimization for Kernel Learning

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 6. (**Moreau Envelope on closed, convex set**, (Moreau, 1965)). 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:

$$\mathsf{M}_{\lambda \ell}(f_{\mathbf{w}}) \triangleq \inf_{f_{\hat{\mathbf{w}}} \in \mathcal{K}} \left\{ \ell(f_{\hat{\mathbf{w}}}) + \frac{1}{2\rho} \| f_{\mathbf{w}} - f_{\hat{\mathbf{w}}} \|_{\mathcal{H}}^2 \right\} \quad (13)$$

Definition 7. Define the function $\Phi(f_{\mathbf{w}}) \triangleq \max_{t \in \mathbb{R}} g(t, f_{\mathbf{w}})$. This function is a L-weakly convex function in $\mathcal{K}$, i.e., $\Phi(f_{\mathbf{w}}) + \frac{L}{2} \| \mathbf{f}_{\mathbf{w}} \|_{\mathcal{H}}^2$ is a convex function over $\mathbf{w}$ in the convex and compact set $\mathcal{K}$.

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

$$\langle \nabla_f g(f_{\hat{\mathbf{w}}}), f_{\mathbf{w}} - f_{\hat{\mathbf{w}}} \rangle_{\mathcal{H}} \geq 0 \qquad \forall f_{\mathbf{w}} \in \mathcal{K} \qquad (14)$$

Definition 9 (Stationary Point of Moreau Envelope). A point $f_{\hat{\mathbf{w}}}$ is a stationary point of the Moreau Envelope defined in Definition 6 of Φ defined in Definition 7 if

$$f_{\hat{\mathbf{w}}} = \arg\inf_{f_{\mathbf{w}} \in \mathcal{K}} \left\{ \Phi_{\lambda}(f_{\mathbf{w}}) + \frac{1}{2\rho} \| f_{\mathbf{w}} - f_{\hat{\mathbf{w}}} \|_{\mathcal{H}}^2 \right\} \quad (15)$$

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

Lemma 10 (Lower bound on distance from stationary point and optimal point). *Let Φ_{λ} be defined as in Definition 7, then if $f_{\hat{\mathbf{w}}}$ is a stationary point as defined in Definition 9 and $g(t, f_{\mathbf{w}})$ has β-Lipschitz Gradient, then*

$$\lim_{\lambda \to \infty} (\Phi_{\lambda}(f_{\hat{\mathbf{w}}}) - \Phi_{\lambda}(f_{\mathbf{w}}^*)) \leq \beta \| f_{\hat{\mathbf{w}}} - f_{\mathbf{w}}^* \|_{\mathcal{H}}^2 \quad (16)$$

We can now upper bound $\| f_{\hat{\mathbf{w}}} - f_{\mathbf{w}}^* \|_{\mathcal{H}}$. 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 10. This bound is different for each of the loss functions, so we must upper bound $\| f_{\hat{\mathbf{w}}} - f_{\mathbf{w}}^* \|_{\mathcal{H}}$ 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 $(\mathbf{x}_i, y_i) \in \mathcal{D}$ is given by the following equation

$$\ell(f_{\mathbf{w}}; \mathbf{x}_i, y_i,) = (f_{\mathbf{w}}(\mathbf{x}_i) - y_i)^2 \tag{17}$$

It is important to note that β is upper bounded as

$$\beta = \frac{2}{np} \text{Tr}(\mathbf{K}) \leq \frac{2}{np}(n(1-\epsilon) \max_{i \in P} k(x_i, x_i)$$
$$+ n\epsilon \max_{j \in Q} k(x_j, x_j) \tag{18}$$
$$= 2p^{-1}((1-\epsilon)P_k + \epsilon Q_k)$$

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

Definition 11 (Strong Projection Property). Let $f_{\mathbf{w}}^*$ be the optimal function for the uncorrupted dataset, $\mathbb{P}$. Then, we have for a finite m and an absolute constant $c > 0$,

$$\|\text{Proj}_m f_{\mathbf{w}}^* - f_{\mathbf{w}}^*\|_{\mathcal{H}} = 0, \qquad \|\text{Proj}_m f_{\mathbf{w}} - f_{\mathbf{w}}^*\| > c \tag{19}$$

The *Strong Projection Property* is important as $\lambda_{\min}(\mathbf{\Sigma})$ 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 12 (Strong Projection Property Implication). *Assume the Strong Projection Property (Definition 11) holds for $f_{\mathbf{w}}^{(t)}$ for all $t \in [T]$, where $f_{\mathbf{w}}^{(t)}$ are iterates from Figure 1. Then, it follows for a $m \in \mathbb{N}$ and a constant $0 < C \leq 1$,*

$$\left\langle \mathbf{\Sigma}, \left(f_{\mathbf{w}}^{(t)} - f_{\mathbf{w}}^*\right) \otimes \left(f_{\mathbf{w}}^{(t)} - f_{\mathbf{w}}^*\right) \right\rangle_{\text{HS}} \tag{20}$$
$$\geq C\lambda_m \|f_{\mathbf{w}}^{(t)} - f_{\mathbf{w}}^*\|_{\mathcal{H}}^2,$$

where we define

$$C \triangleq \frac{\|f_{\mathbf{w}}^{(t)} - f_{\mathbf{w}}^*\|_{\mathcal{H}}^2 - \left\|\text{Proj}_{m,\perp} f_{\mathbf{w}}\right\|_{\mathcal{H}}^2}{\|f_{\mathbf{w}}^{(t)} - f_{\mathbf{w}}^*\|_{\mathcal{H}}^2} \tag{21}$$

Proof is given in Appendix C.2. We will discuss the implication of Lemma 12 after stating the following theorem.

Theorem 13 (Stationary Point for Kernelized Regression is Good on Average). *Let $f_{\hat{\mathbf{w}}}$ be a stationary point defined in Definition 9 for the function Φ defined in Definition 7. Then for a constant $c_1 \in (0,1)$, if $n \geq \frac{8 \text{Tr}(\mathbf{\Sigma})^2}{\lambda_{\min}(\mathbf{\Sigma})(1-c_1)^2(1-2\epsilon)} + \frac{8\beta}{(1-c_1)^2(1-2\epsilon)}$,*

$$\mathop{\mathbb{E}}_{\mathcal{D} \sim \hat{\mathbb{P}}} \|f_{\hat{\mathbf{w}}} - f_{\mathbf{w}^*}\|_{\mathcal{H}} \leq \sigma \sqrt{\frac{\gamma}{c_1 \lambda_{\min}(\mathbf{\Sigma})}}$$
$$+ \frac{O\left(\sigma \sqrt{\gamma \log\left(n(1-2\epsilon)\right) \text{Tr}(\mathbf{\Sigma})}\right)}{c_1 \sqrt{n(1-2\epsilon)} c \lambda_{\min}(\mathbf{\Sigma})} \tag{22}$$

where β is the Lipschitz Gradient Constant given in Lemma 33.

Theorem 14 (Stationary Point for Kernelized Regression is Good with High Probability). *Frame the same hypothesis as Theorem 13. Then for a constant $c_1 \in (0,1)$, if $n \geq \frac{8 \text{Tr}(\mathbf{\Sigma})^2}{\lambda_{\min}(\mathbf{\Sigma})(1-c_1)^2(1-2\epsilon)} + \frac{8\beta}{(1-c_1)^2(1-2\epsilon)}$,*

$$\mathop{\mathbb{E}}_{\mathcal{D} \sim \hat{\mathbb{P}}} \|f_{\hat{\mathbf{w}}} - f_{\mathbf{w}^*}\|_{\mathcal{H}} \leq \sqrt{\frac{\gamma\sigma}{c_1 \lambda_{\min}(\mathbf{\Sigma})}}$$
$$+ \frac{O\left(\sigma \sqrt{\gamma \log(\tilde{n}) \text{Tr}(\mathbf{\Sigma})}\right)}{c_1 \sqrt{\tilde{n}} c \lambda_{\min}(\mathbf{\Sigma})} \tag{23}$$

where β is the Lipschitz Gradient Constant given in Lemma 33.

Proof is given in Appendix C.4. In Theorem 13, 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 (Schmidt et al., 2018). 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 12 can be seen in the denominator of both terms. Instead of $\lambda_{\min}(\mathbf{\Sigma})$ we have $c_4 \lambda_m$ for a finite m. This difference 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 Error Bound). *Consider Subquantile Minimization for Linear Regression on the data X with optimal parameters $\mathbf{w}^*$. Assume $\mathbf{x}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$ for $i \in [n]$. Then after T iterations of Figure 1, we have the following error bounds for robust kernelized linear regression. Given sufficient data*

$$\mathbb{E}\left\|\mathbf{w}^{(T)} - \mathbf{w}^*\right\|_2 \leq O\left(\frac{\gamma\sigma}{\sqrt{\lambda_{\min}(\mathbf{\Sigma})}}\right) \tag{24}$$

Proof given in Appendix C.5. Let us note for the case where p is finite, i.e. the feature mapping is finite-dimensional, e.g. linear or polynomial kernel. Then we have that $\text{Proj}_{\Psi_m^\perp}$ where $m = p$ is equal to zero as $\{\varphi_i\}_{i=1}^m$ spans the finite-dimensional space, in which case we have the absolute constant given in Definition 11 is equal to zero. It is important to note in all our bounds, $\gamma \leq \sqrt{\frac{\epsilon}{1-2\epsilon}}$ is

a theoretical worst case bound when the Subquantile contains the minimum possible number of uncorrupted points. In other words, we have $\gamma \triangleq \frac{|P \setminus S|}{|S \cap P|} \leq \frac{n\epsilon}{n(1-2\epsilon)} = \frac{\epsilon}{1-2\epsilon}$. 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 Kernelized Binary Classification

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

$$\ell(\mathbf{x}_i, y_i; f_{\mathbf{w}}) = -\left(y_i \log\left(\sigma\left(f_{\mathbf{w}}(\mathbf{x}_i)\right)\right) \right. \\ \left. - (1 - y_i)\log\left(1 - \sigma\left(f_{\mathbf{w}}(\mathbf{x}_i)\right)\right)\right) \tag{25}$$

Theorem 16. *[A stationary point is good for kernel binary classification] Let $f_{\hat{\mathbf{w}}}$ be a statoinary point defined in Definition 8 for the function Φ defined in Definition 7. Then for a constant $c_4 \in (0,1)$, if $n \geq \frac{4\,\mathrm{Tr}(\boldsymbol{\Sigma})}{\lambda_{\min}(\boldsymbol{\Sigma})(1-2\epsilon)(1-c_4)}$, then in expectation over the dataset distribution,*

$$\mathbb{E}_{\mathcal{D} \sim \hat{\mathbb{P}}} \|f_{\hat{\mathbf{w}}} - f_{\mathbf{w}}^*\|_{\mathcal{H}}$$
$$\leq O\left(\frac{\sqrt{\mathrm{Tr}(\boldsymbol{\Sigma})} + \sqrt{Q_k}}{\sqrt{n(1-2\epsilon)}\exp\left(-R\left(\mathrm{Tr}\left(\boldsymbol{\Sigma}\right) + \log n\right)\right)\lambda_{\min}(\boldsymbol{\Sigma})}\right) \tag{26}$$

Proof is given in Appendix D.2. This result although shows consistency, i.e. when $n \to \infty$, then we have in expectation $\|f_{\mathbf{w}} - f_{\mathbf{w}}^*\| \to 0$, 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. To mitigate the effect of a large Q_k, a filtering algorithm can be used to remove points, $\mathbf{x}_i$, such that $k(\mathbf{x}_i, \mathbf{x}_i)$ is far from the mean.

3.5 Kernelized Multi-Class Classification

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

$$\ell(\mathbf{x}_i, y_i; f_{\mathbf{W}}) = $$
$$-\sum_{j=1}^{|\mathcal{Y}|} \mathbb{I}\{j = y_i\}\log\left(\frac{\exp\left(f_{\mathbf{w}_j}(\mathbf{x}_i)\right)}{\sum_{k=1}^{|\mathcal{Y}|}\exp\left(f_{\mathbf{w}_k}(\mathbf{x}_i)\right)}\right) \tag{27}$$

Theorem 17 (Stationary Point for Kernelized Multi-Class Classification is Good). *Let $f_{\hat{\mathbf{w}}}$ be a stationary point defined in Definition 9 for the function Φ defined in Definition 7. Then for a constant $c_1 \in (0,1)$, if $n \geq$*

$$\frac{8\,\mathrm{Tr}(\boldsymbol{\Sigma})^2}{\lambda_{\min}(\boldsymbol{\Sigma})(1-c_1)^2(1-2\epsilon)} + \frac{8\beta}{(1-c_1)^2(1-2\epsilon)},$$

$$\mathbb{E}_{\mathcal{D} \sim \hat{\mathbb{P}}} \|f_{\hat{\mathbf{W}}} - f_{\mathbf{W}}^*\|_{\mathcal{H}}$$
$$\leq O\left(\frac{\sqrt{n(1-\epsilon)\,\mathrm{Tr}(\boldsymbol{\Sigma})} + \sqrt{n\epsilon Q_k}}{n(1-2\epsilon)\lambda_{\min}(\boldsymbol{\Sigma}) - \sqrt{n(1-2\epsilon)}\,\mathrm{Tr}(\boldsymbol{\Sigma})}\right) \tag{28}$$

where β is the Lipschitz Gradient Constant given in Lemma 37.

3.6 Optimization

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 18 ((Rockafellar, 1970; Davis & Drusvyatskiy, 2019)). *Assume the function Φ is β-weakly convex. Let $\lambda < \frac{1}{\beta}$, and let $f_{\hat{\mathbf{w}}} = \arg\min_{f_{\mathbf{w}} \in \mathcal{K}}(\Phi(f_{\mathbf{w}}) + \frac{1}{2\lambda}\|f_{\mathbf{w}} - f_{\hat{\mathbf{w}}}\|_{\mathcal{H}}^2)$, then $\|\nabla\Phi_\lambda(f_{\mathbf{w}})\|_{\mathcal{H}} \leq \epsilon$ implies:*

$$\|f_{\hat{\mathbf{w}}} - f_{\mathbf{w}}\|_{\mathcal{H}} = \lambda\epsilon \quad and \quad \min_{\mathbf{g} \in \partial\Phi(f_{\hat{\mathbf{w}}}) + \partial\mathcal{I}_{\mathcal{K}}(f_{\hat{\mathbf{w}}})} \|\mathbf{g}\|_{\mathcal{H}} \leq \epsilon \tag{29}$$

With Lemma 18 in hand, it suffices to show that $\|\nabla\Phi_\lambda(f_{\mathbf{w}})\|_{\mathcal{H}}$ is small, as it then follows that $f_{\mathbf{w}}$ is close to a stationary point of the Moreau Envelope. It has been shown in optimization theory that utilizing standard gradient descent, $\|\nabla\Phi_\lambda(f_{\mathbf{w}})\|_{\mathcal{H}}$ decreases at a rate of $O(T^{-1/2})$. The exact theorem and proof can be seen in (Davis & Drusvyatskiy, 2019) and a proof where the maximimum of the inner problem can be calculated to within $(1 + \epsilon)$ optimality can be seen in (Jin et al., 2020) and (Cheng et al., 2020).

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 $f_{\mathbf{w}}$ with the Glorot Initialization Scheme (Glorot & Bengio, 2010).

In Figure 2, 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 (Pedregosa et al., 2011),

Algorithms	Test RMSE							
	Concrete (Yeh, 2007)		Wine Quality (Cortez & Reis, 2009)		Boston Housing (Dua & Graff, 2017)		Drug (Olier et al., 2018)	
	$\epsilon = 0.2(\downarrow)$	$\epsilon = 0.4(\downarrow)$	$\epsilon = 0.2(\downarrow)$	$\epsilon = 0.4(\downarrow)$	$\epsilon = 0.2(\downarrow)$	$\epsilon = 0.4(\downarrow)$	$\epsilon = 0.2(\downarrow)$	$\epsilon = 0.4(\downarrow)$
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 $\mathbf{x}_{\text{noise}} = 100\mathbf{x}_{\text{original}}$. Polynomial Regression Synthetic Dataset. 1000 samples, $x \sim \mathcal{N}(0,1)$, $y \sim \mathcal{N}(\sum_{i=0} a_i x^i, 0.01)$ where $a_i \sim \mathcal{N}(0,1)$. The Radial Basis Function is used in first three experiments and polynomial kernel with degree 3 and $C = 1$ is used in the last experiment.
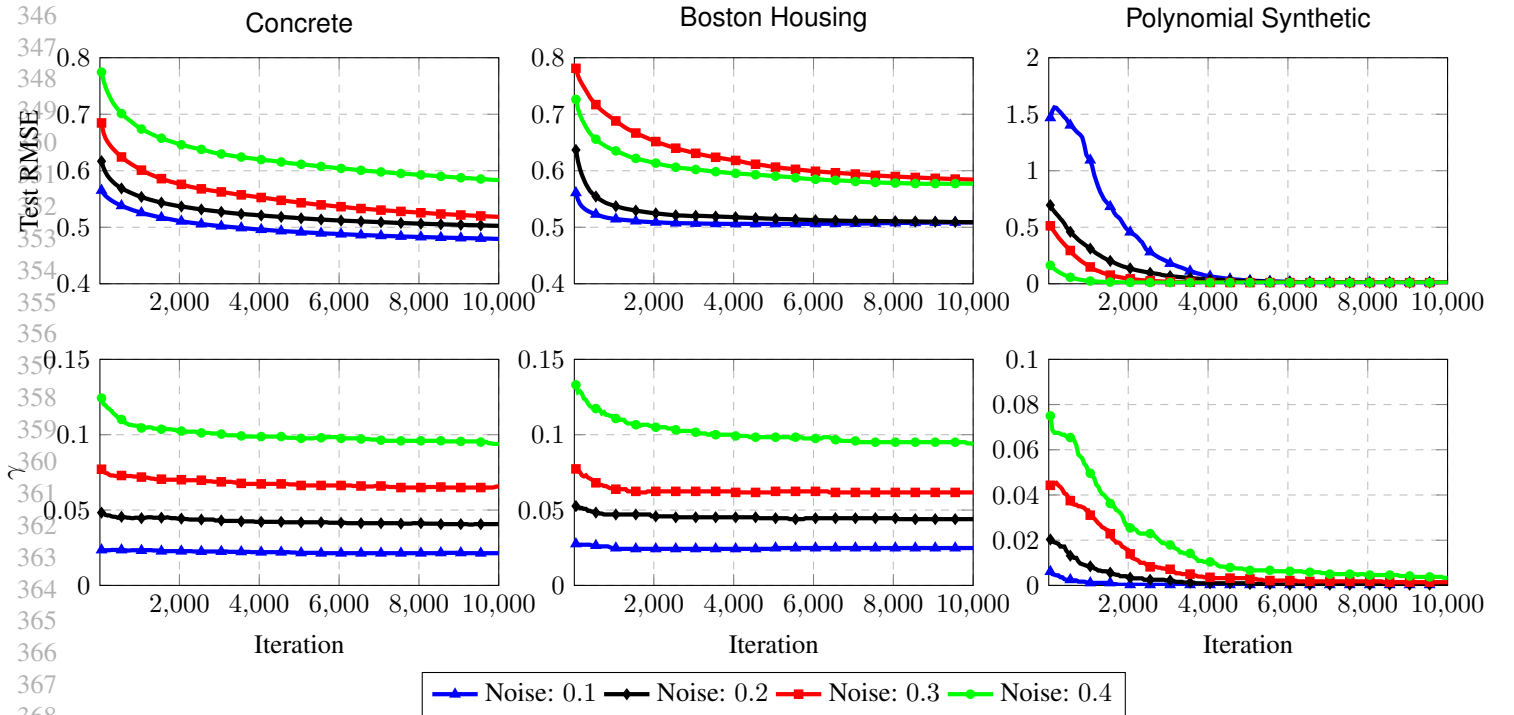

Figure 2: Test RMSE over the iterations in Concrete, Boston Housing, and Polynomial Datasets for SUBQUANTILE at different noise levels

Consistent Robust Regression (CRR) (Bhatia et al., 2017), Globally-convergent iteratively reweighted least squares (STIR) (Mukhoty et al., 2019). 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 (Diakonikolas et al., 2019) and Tilted Empirical Risk Minimization (Li et al., 2021).

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.

Algorithms	Test RMSE							
	Boston Housing (Dua & Graff, 2017)		Wine Quality (Cortez & Reis, 2009)		Concrete (Yeh, 2007)		Drug (Olier et al., 2018)	
	Label($\downarrow$)	Label+Feature($\downarrow$)	Label($\downarrow$)	Label+Feature($\downarrow$)	Label($\downarrow$)	Label+Feature($\downarrow$)	Label($\downarrow$)	Label+Feature($\downarrow$)
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	$\mathbf{0.503}_{(0.0470)}$	$\mathbf{0.548}^{*}_{(0.0286)}$	$\mathbf{0.813}_{(0.0357)}$	$0.821_{(0.0305)}$	$\mathbf{0.632}_{(0.0275)}$	$\mathbf{0.703}_{(0.0427)}$	$\mathbf{1.074}_{(0.1848)}$	$\underline{2.413}_{(0.6737)}$
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 $\mathbf{x}_{\text{noisy}} = 100\mathbf{x}_{\text{original}}$ and $y_{\text{noisy}} = 10000 y_{\text{original}}$. * 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 10^5 iterations, and the feature noise perturbed dataset for 10^6 iterations.

Algorithms	Test Accuracy							
	Heart Disease (Janosi & Detrano, 1988)				Breast Cancer (Wolberg & Street, 1995)			
	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	$\mathbf{0.803}_{(0.0293)}$	$\mathbf{0.790}_{(0.0350)}$	$\mathbf{0.833}_{(0.0318)}$	$\underline{0.807}_{(0.0468)}$	$0.928_{(0.0129)}$	$\mathbf{0.916}_{(0.0185)}$	$\underline{0.972}_{(0.0187)}$	$0.963_{(0.0170)}$
Oracle ERM	∞	∞	∞	∞	∞	∞	∞	∞

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

- **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 $\sqrt{\epsilon/(1-2\epsilon)}$. When $\epsilon \to 0.5$, 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.

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

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 $\mathbf{x}_{\text{noise}} = 100\mathbf{x}_{\text{original}}$. The Radial Basis Function is used in all experiments.

Algorithms	Test Accuracy							
	Iris (Fisher, 1988)		Glass (German, 1987)		Wine (Aeberhard & Forina, 1991)		Satimage (Srinivasan, 1993)	
	$\epsilon = 0.2(\uparrow)$	$\epsilon = 0.4(\uparrow)$	$\epsilon = 0.2(\uparrow)$	$\epsilon = 0.4(\uparrow)$	$\epsilon = 0.2(\uparrow)$	$\epsilon = 0.4(\uparrow)$	$\epsilon = 0.2(\uparrow)$	$\epsilon = 0.4(\uparrow)$
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	∞	∞	∞	∞	∞	∞	∞	∞

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

Abu-Mostafa, Y. S., Magdon-Ismail, M., and Lin, H.-T. *Learning from data*, volume 4. AMLBook New York, 2012. 1

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

Awasthi, P., Das, A., Kong, W., and Sen, R. Trimmed maximum likelihood estimation for robust generalized linear model. In Oh, A. H., Agarwal, A., Belgrave, D., and Cho, K. (eds.), *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=VHmdFPy4U_u. 1, 2

Bernstein, S. On a modification of chebyshev's inequality and of the error formula of laplace. *Ann. Sci. Inst. Sav. Ukraine, Sect. Math*, 1(4):38–49, 1924. 18

Bhatia, K., Jain, P., and Kar, P. Robust regression via hard thresholding. In Cortes, C., Lawrence, N., Lee, D., Sugiyama, M., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015. URL https://proceedings.neurips.cc/paper_files/paper/2015/file/1be3bc32e6564055d5ca3e5a354acbef-Paper.pdf. 2

Bhatia, K., Jain, P., Kamalaruban, P., and Kar, P. Consistent robust regression. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/e702e51da2c0f5be4dd354bb3e295d37-Paper.pdf. 1, 7

Bogachev, V. I. *Gaussian measures*. Number 62. American Mathematical Soc., 1998. 13

Cheng, Y., Diakonikolas, I., Ge, R., and Woodruff, D. P. Faster algorithms for high-dimensional robust covari-

ance estimation. In Beygelzimer, A. and Hsu, D. (eds.), *Proceedings of the Thirty-Second Conference on Learning Theory*, volume 99 of *Proceedings of Machine Learning Research*, pp. 727–757. PMLR, 25–28 Jun 2019. URL https://proceedings.mlr.press/v99/cheng19a.html. 1

Cheng, Y., Diakonikolas, I., Ge, R., and Soltanolkotabi, M. High-dimensional robust mean estimation via gradient descent. In III, H. D. and Singh, A. (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 1768–1778. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/cheng20a.html. 1, 6

Cortez, Paulo, C. A. A. F. M. T. and Reis, J. Wine Quality. UCI Machine Learning Repository, 2009. DOI: https://doi.org/10.24432/C56S3T. 7, 8

Cui, H., Loureiro, B., Krzakala, F., and Zdeborová, L. Generalization error rates in kernel regression: The crossover from the noiseless to noisy regime. *Advances in Neural Information Processing Systems*, 34:10131–10143, 2021. 2, 23

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

Diakonikolas, I. and Kane, D. M. *Algorithmic high-dimensional robust statistics*. Cambridge University Press, 2023. 1

Diakonikolas, I., Kamath, G., Kane, D. M., Li, J., Steinhardt, J., and Stewart, A. Sever: A robust meta-algorithm for stochastic optimization. In *Proceedings of the 36th International Conference on Machine Learning*, ICML '19, pp. 1596–1606. JMLR, Inc., 2019. 1, 7

Dicker, L. H. Ridge regression and asymptotic minimax estimation over spheres of growing dimension. 2016. 2

Dua, D. and Graff, C. UCI machine learning repository, 2017. URL http://archive.ics.uci.edu/ml. 7, 8

Fan, J., Wang, W., and Zhong, Y. An l eigenvector perturbation bound and its application to robust covariance estimation. *Journal of Machine Learning Research*, 18 (207):1–42, 2018. 1

Fan, Y., Lyu, S., Ying, Y., and Hu, B. Learning with average top-k loss. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/6c524f9d5d7027454a783c841250ba71-Paper.pdf. 2

Fischler, M. A. and Bolles, R. C. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Commun. ACM*, 24(6):381–395, jun 1981. ISSN 0001-0782. doi: 10.1145/358669.358692. URL https://doi.org/10.1145/358669.358692. 1

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

Gao, B. and Pavel, L. On the properties of the softmax function with application in game theory and reinforcement learning. *arXiv preprint arXiv:1704.00805*, 2017. 29

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

Glorot, X. and Bengio, Y. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pp. 249–256. JMLR Workshop and Conference Proceedings, 2010. 6

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

Gretton, A. Introduction to rkhs, and some simple kernel algorithms. *Adv. Top. Mach. Learn. Lecture Conducted from University College London*, 16(5-3):2, 2013b. 21

Hofmann, T., Schölkopf, B., and Smola, A. J. Kernel methods in machine learning. *The Annals of Statistics*, 36(3):1171 – 1220, 2008. doi: 10.1214/009053607000000677. URL https://doi.org/10.1214/009053607000000677. 1

Hölder, O. Ueber einen mittelwerthabsatz. *Nachrichten von der Königl. Gesellschaft der Wissenschaften und der Georg-Augusts-Universität zu Göttingen*, 1889:38–47, 1889. URL http://eudml.org/doc/180218. 29

Hu, S., Ying, Y., wang, x., and Lyu, S. Learning by minimizing the sum of ranked range. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 21013–21023. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/f0d7053396e765bf52de12133cf1afe8-Paper.pdf. 1, 2

Hu, S., Yang, Z., Wang, X., Ying, Y., and Lyu, S. Outlier robust adversarial training. *arXiv preprint arXiv:2309.05145*, 2023. 2

Huber, P. J. and Ronchetti, E. *Robust statistics*. Wiley series in probability and statistics. Wiley, Hoboken, N.J., 2nd ed. edition, 2009. URL http://catdir.loc.gov/catdir/toc/ecip0824/2008033283.html. 1

Janosi, Andras, S. W. P. M. and Detrano, R. Heart Disease. UCI Machine Learning Repository, 1988. DOI: https://doi.org/10.24432/C52P4X. 8

Jiang, L., Zhou, Z., Leung, T., Li, L.-J., and Fei-Fei, L. Mentornet: Learning data-driven curriculum for very deep neural networks on corrupted labels. In *ICML*, 2018. 1

Jin, C., Netrapalli, P., and Jordan, M. What is local optimality in nonconvex-nonconcave minimax optimization? In III, H. D. and Singh, A. (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 4880–4889. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/jin20e.html. 6

Khetan, A., Lipton, Z. C., and Anandkumar, A. Learning from noisy singly-labeled data. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=H1sUHgb0Z. 1

Laguel, Y., Pillutla, K., Malick, J., and Harchaoui, Z. Superquantiles at work: Machine learning applications and efficient subgradient computation. *Set-Valued and Variational Analysis*, 29(4):967–996, Dec 2021. ISSN 1877-0541. doi: 10.1007/s11228-021-00609-w. URL https://doi.org/10.1007/s11228-021-00609-w. 1

Legendre, A. M. *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. 2

Li, T., Beirami, A., Sanjabi, M., and Smith, V. Tilted empirical risk minimization. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=K5YasWXZT3O. 1, 7

Mercer, J. 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. 2

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

Mukhoty, B., Gopakumar, G., Jain, P., and Kar, P. Globally-convergent iteratively reweighted least squares for robust regression problems. In Chaudhuri, K. and Sugiyama, M. (eds.), *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pp. 313–322. PMLR, 16–18 Apr 2019. URL https://proceedings.mlr.press/v89/mukhoty19a.html. 1, 7

Olier, I., Sadawi, N., Bickerton, G. R., Vanschoren, J., Grosan, C., Soldatova, L., and King, R. D. Meta-qsar: a large-scale application of meta-learning to drug design and discovery. *Machine Learning*, 107:285–311, 2018. 7, 8

Osama, M., Zachariah, D., and Stoica, P. Robust risk minimization for statistical learning from corrupted data. *IEEE Open Journal of Signal Processing*, 1:287–294, 2020. 1

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830, 2011. 6

Prasad, A., Suggala, A. S., Balakrishnan, S., and Ravikumar, P. Robust estimation via robust gradient estimation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 82, 2018. URL https://api.semanticscholar.org/CorpusID:3614648. 1

Prasad, A., Balakrishnan, S., and Ravikumar, P. A unified approach to robust mean estimation. *arXiv preprint arXiv:1907.00927*, 2019. 1

Razaviyayn, M., Huang, T., Lu, S., Nouiehed, M., Sanjabi, M., and Hong, M. Nonconvex min-max optimization: Applications, challenges, and recent theoretical advances. *IEEE Signal Processing Magazine*, 37(5):55–66, 2020. doi: 10.1109/MSP.2020.3003851. 3

Rockafellar, R., Royset, J., and Miranda, S. Superquantile regression with applications to buffered reliability, uncertainty quantification, and conditional value-at-risk. *European Journal of Operational Research*, 234(1):140–154, 2014. ISSN 0377-2217. doi: https://doi.org/10.1016/j.ejor.2013.10.046. URL https://www.sciencedirect.com/science/article/pii/S0377221713008692. 1

Rockafellar, R. T. *Convex Analysis*. Princeton University Press, Princeton, 1970. ISBN 9781400873173. doi: doi: 10.1515/9781400873173. URL https://doi.org/10.1515/9781400873173. 6

Santin, G. and Schaback, R. Approximation of eigenfunctions in kernel-based spaces. *Advances in Computational Mathematics*, 42:973–993, 2016. 23

Schmidt, L., Santurkar, S., Tsipras, D., Talwar, K., and Madry, A. Adversarially robust generalization requires more data. *Advances in neural information processing systems*, 31, 2018. 5

Shen, Y. and Sanghavi, S. Learning with bad training data via iterative trimmed loss minimization. In Chaudhuri, K. and Salakhutdinov, R. (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 5739–5748. PMLR, 09–15 Jun 2019a. URL https://proceedings.mlr.press/v97/shen19e.html. 2

Shen, Y. and Sanghavi, S. Learning with bad training data via iterative trimmed loss minimization. In *International Conference on Machine Learning*, pp. 5739–5748. PMLR, 2019b. 2

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

Vershynin, R. High-dimensional probability. *University of California, Irvine*, 2020. 18

Weyl, H. Das asymptotische verteilungsgesetz der eigenwerte linearer partieller differentialgleichungen (mit einer anwendung auf die theorie der hohlraumstrahlung). *Mathematische Annalen*, 71(4):441–479, 1912. 20, 29

Wolberg, William, M. O. S. N. and Street, W. Breast Cancer Wisconsin (Diagnostic). UCI Machine Learning Repository, 1995. DOI: https://doi.org/10.24432/C5DW2B. 8

Yedida, R., Saha, S., and Prashanth, T. Lipschitzlr: Using theoretically computed adaptive learning rates for fast convergence. *Applied Intelligence*, 51:1460–1478, 2021. 21

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

A Probability Theory

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

$$\phi(\mathbf{x}_i) \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}) \tag{30}$$

where $\boldsymbol{\Sigma}$ is a possibly infinite dimensional covariance operator.

Assumption 20 (Normal Residuals)**.** The residual is defined as $\mu_i \triangleq f_{\mathbf{w}}^*(\mathbf{x}_i) - y_i$. Then we assume for some $\sigma > 0$, it follows

$$\mu_i \sim \mathcal{N}(0, \sigma^2) \tag{31}$$

Proposition 21 (Concentration for functions of a Gaussian Vector ([Bogachev](), [1998]))**.** *Suppose h is a Lipschitz function on vectors, i.e.*

$$|h(\mathbf{x}) - h(\mathbf{y})| \le \|h\|_{\mathrm{lip}} \|\mathbf{x} - \mathbf{y}\| \tag{32}$$

for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$. Then,

$$\mathbb{P}\{|h(\mathbf{g}) - \mathbb{E}h(\mathbf{g})| \ge T\} \le 2\exp\left[-\frac{t^2}{2\|h\|_{\mathrm{lip}}^2}\right] \tag{33}$$

Lemma 22 (Maximum of Gaussians)**.** *Let $\mu_1, \ldots, \mu_n \sim \mathcal{N}(0, \sigma^2)$ for some $\sigma > 0$. Then it follows*

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

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

$$\mathbb{E}_{\mu_i \sim \mathcal{N}(0,\sigma^2)} \max_{i \in [n]} |\mu_i| = \int_0^\infty \mathbb{P}_{\mu_i \sim \mathcal{N}(0,\sigma^2)}\left\{\max_{i \in [n]} |\mu_i| > t\right\} dt \overset{(i)}{\le} c_1 + n\int_{c_1}^\infty \mathbb{P}_{\mu_i \sim \mathcal{N}(0,\sigma^2)}\{|\mu_i| \ge t\} dt \tag{35}$$

$$\overset{(ii)}{=} c_1 + 2n\int_{c_1}^\infty \mathbb{P}_{\mu_i \sim \mathcal{N}(0,\sigma^2)}\{\mu_i \ge t\} dt = c_1 + 2n\int_{c_1}^\infty \int_t^\infty \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x}{\sigma}\right)^2} dx\, dt \tag{36}$$

$$\le c_1 + \frac{n}{\sigma}\sqrt{\frac{2}{\pi}} \int_{c_1}^\infty \int_t^\infty \left(\frac{x}{t}\right) e^{-\frac{1}{2}\left(\frac{x}{\sigma}\right)^2} dx\, dt = c_1 + n\sigma\sqrt{\frac{2}{\pi}} \int_{c_1}^\infty \frac{e^{-\frac{1}{2}\left(\frac{t}{\sigma}\right)^2}}{t} dt \tag{37}$$

$$\le c_1 + n\sigma\sqrt{\frac{2}{\pi}} \int_{c_1}^\infty \left(\frac{t}{c_1}\right) e^{-\frac{1}{2}\left(\frac{t}{\sigma}\right)^2} dt = c_1 + n\sigma^3\sqrt{\frac{2}{\pi}} \frac{e^{-\frac{1}{2}\left(\frac{c_1}{\sigma}\right)^2}}{c_1} \tag{38}$$

(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 $\mathbb{P}\{\max_{i \in [n]} X_i \ge C\} = n\mathbb{P}\{X \ge C\}$ 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,

$$\mathbb{E}_{\mu_i \sim \mathcal{N}(0,\sigma^2)} \max_{i \in [n]} |\mu_i| \le \sigma\sqrt{2\log n} + \frac{\sigma^2}{\sqrt{\pi \log n}} \tag{39}$$

This completes the proof. ∎

Proposition 23. *Let $\mu_1, \ldots, \mu_n \sim \mathcal{N}(0, \sigma^2)$ for some $\sigma > 0$, then it follows for any $s \ge 1$*

$$\mathbb{P}\left\{\max_{i \in [\![n]\!]} |\mu_i| \ge \sigma\sqrt{2\log n} \cdot s\right\} \le \frac{\sqrt{2}}{\log n} e^{-s^2} \tag{40}$$

Proof. The proof follows simply using similar steps as in the proof of Lemma [22]. Let C be a positive constant to be determined.

$$\mathbb{P}_{\mu_i \sim \mathcal{N}(0,\sigma^2)}\left\{\max_{i \in [\![n]\!]} |\mu_i| \ge C \cdot s\right\} = 2n\,\mathbb{P}_{\mu \sim \mathcal{N}(0,\sigma^2)}\{\mu \ge C \cdot s\} = n\sqrt{\frac{2}{\pi}} \int_{C \cdot s}^\infty \left(\frac{1}{\sigma}\right) e^{-\frac{1}{2}\left(\frac{x}{\sigma}\right)^2} dx \tag{41}$$

13

$$\leq 2\sigma n \left(\frac{1}{C \cdot s}\right) e^{-\frac{1}{2}\left(\frac{C \cdot s}{\sigma}\right)^2} \leq \frac{\sqrt{2}n^{1-s^2}}{s \log n} \leq \frac{\sqrt{2}}{\log n} e^{-s^2} \tag{42}$$

In the second to last inequality, we plug in $C \triangleq \sigma\sqrt{2\log n}$. Our proof is now complete. $\blacksquare$

Proposition 24. *Let $\mu_1, \ldots, \mu_n \sim \mathcal{N}(0, \sigma^2)$ for some $\sigma > 0$, then it follows for any $s \geq 1$,*

$$\mathbb{P}\left\{\sum_{i=1}^{n} \mu_i^2 \geq n\sigma^2 \cdot s\right\} \leq e^{-s/2} \tag{43}$$

Proof. Concatenate all the samples μ_i into a vector $\boldsymbol{\mu} \in \mathbb{R}^n$. Our proof generalizes for a $\boldsymbol{\mu} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma})$ where $\boldsymbol{\Sigma} \triangleq \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^\top$ for a unitary $\mathbf{U}$ and positive diagonal $\boldsymbol{\Lambda}$. Let C be a positive to be determined constant, we then have

$$\mathbb{P}_{\boldsymbol{\mu} \sim \mathcal{N}(\boldsymbol{0}, \sigma^2\mathbf{I})}\left\{\|\boldsymbol{\mu}\|^2 \geq C \cdot s\right\} = \mathbb{P}_{\boldsymbol{\mu} \sim \mathcal{N}(\boldsymbol{0}, \sigma^2\mathbf{I})}\left\{\|\boldsymbol{\mu}\| \geq \sqrt{C \cdot s}\right\} = \mathbb{P}_{\mathbf{g} \sim \mathcal{N}(\boldsymbol{0}, \mathbf{I})}\left\{\left\|\mathbf{U}\boldsymbol{\Lambda}^{1/2}\mathbf{g}\right\| \geq \sqrt{C \cdot s}\right\} \tag{44}$$

$$= \mathbb{P}_{\mathbf{g} \sim \mathcal{N}(\boldsymbol{0}, \mathbf{I})}\left\{\sqrt{\sum_{i=1}^{n} \lambda_i g_i^2} \geq \sqrt{C \cdot s}\right\} \leq \mathbb{P}_{\mathbf{g} \sim \mathcal{N}(\boldsymbol{0}, \mathbf{I})}\left\{\sum_{i=1}^{n} \sqrt{\lambda_i} g_i \geq \sqrt{C \cdot s}\right\} \tag{45}$$

$$\leq \inf_{\theta > 0} \mathbb{E}_{g_i \sim \mathcal{N}(0,1)}\left[\prod_{i=1}^{n} \exp\left(\theta\sqrt{\lambda_i} g_i\right)\right] \exp\left[-\theta\sqrt{C \cdot s}\right] \tag{46}$$

$$= \inf_{\theta > 0} \exp\left[\frac{\text{Tr}(\boldsymbol{\Lambda})}{2}\theta^2 - \theta\sqrt{C \cdot s}\right] = \exp\left[-\frac{(C \cdot s)}{2\text{Tr}(\boldsymbol{\Lambda})}\right] \tag{47}$$

The second to last equality follows from the MGF of a Gaussian. Then, plugging in $C \triangleq \text{Tr}(\boldsymbol{\Lambda})$ completes the proof. $\blacksquare$

Lemma 25 (Maximum of Squared Gaussians). *Let $\mu_1, \ldots, \mu_n \sim \mathcal{N}(0, \sigma^2)$ for $\sigma > 0$, $n > 1$. Then it follows*

$$\mathbb{E}_{\mu_i \sim \mathcal{N}(0, \sigma^2)} \max_{i \in [n]} \mu_i^2 \leq 2\sigma^2 \log(n) + \left(\sigma^3\sqrt{\frac{8}{\pi}}\right)\left(1 + \frac{1}{\log(n)}\right) \tag{48}$$

Proof. Our proof follows similarly to the proof for Lemma 22.

$$\mathbb{E}_{\mu_i \sim \mathcal{N}(0, \sigma^2)} \max_{i \in \mathbb{N}} \mu_i^2 = \int_0^\infty \mathbb{P}_{\mu_i \sim \mathcal{N}(0, \sigma^2)}\left\{\max_{i \in [n]} \mu_i^2 \geq t\right\} dt \leq c_2 + n\int_{c_2}^\infty \mathbb{P}_{\mu_i \sim \mathcal{N}(0, \sigma^2)}\left\{|\mu_i| \geq \sqrt{t}\right\} dt \tag{49}$$

$$= c_2 + 2n\int_{c_2}^\infty \mathbb{P}_{\mu_i \sim \mathcal{N}(0, \sigma^2)}\left\{\mu_i \geq \sqrt{t}\right\} dt = c_2 + 2n\int_{c_2}^\infty \int_{\sqrt{t}}^\infty \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x}{\sigma}\right)^2} dx\, dt \tag{50}$$

$$= c_2 + n\sigma\sqrt{\frac{2}{\pi}}\int_{c_2}^\infty \frac{e^{-\frac{1}{2}\left(\frac{t}{\sigma^2}\right)}}{\sqrt{t}} dt \overset{(i)}{\leq} c_2 + n\sigma\sqrt{\frac{2}{\pi}}\int_{c_2}^\infty \left(\frac{t}{c_2}\right) e^{-\frac{1}{2}\left(\frac{t}{\sigma^2}\right)} dt \tag{51}$$

$$\leq c_2 + \left(\sqrt{\frac{2}{\pi}}\right)\frac{n\sigma\left(4\sigma^4 + 2c_2\sigma^2\right) e^{-\frac{c_2}{2\sigma^2}}}{c_2} \tag{52}$$

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

$$\mathbb{E}_{\mu_i \sim \mathcal{N}(0, \sigma^2)} \max_{i \in [n]} \mu_i^2 \leq 2\sigma^2 \log(n) + \left(2\sigma^3\sqrt{\frac{2}{\pi}}\right)\left(1 + \frac{1}{\log(n)}\right) \tag{53}$$

This completes the proof. $\blacksquare$

Lemma 26 (Expected Maximum P_k). *Let $\mathbf{x}_i \sim \mathbb{P}$ such that $\phi(\mathbf{x}_i) \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma})$ from Assumption 19. Then it follows for any $s \geq 1$*

$$\mathbb{E}_{\phi(\mathbf{x}_i) \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma})}\left[\max_{i \in [n]} k(\mathbf{x}_i, \mathbf{x}_i)\right] \leq 2\text{Tr}(\boldsymbol{\Sigma}) \log\left(\frac{sn}{2\text{Tr}(\boldsymbol{\Sigma})}\right) + \frac{1}{s} \tag{54}$$

Proof. We will use the integral identity of the expectation of a random variable to make our claim. Throughout the proof, let C be a positive to be determined constant.

$$\mathop{\mathbb{E}}_{\phi(\mathbf{x}_i)\sim\mathcal{N}(\mathbf{0},\mathbf{\Sigma})}\left[\max_{i\in[n]}k(\mathbf{x}_i,\mathbf{x}_i)\right] \leq C + \int_C^\infty \mathop{\mathbb{P}}_{\phi(\mathbf{x}_i)\sim\mathcal{N}(\mathbf{0},\mathbf{\Sigma})}\left\{\max_{i\in[n]}k(\mathbf{x}_i,\mathbf{x}_i)\geq t\right\}dt \tag{55}$$

$$\stackrel{(i)}{\leq} C + n\int_C^\infty \mathop{\mathbb{P}}_{\phi(\mathbf{x})\sim\mathcal{N}(\mathbf{0},\mathbf{\Sigma})}\{k(\mathbf{x},\mathbf{x})\geq t\}dt \stackrel{(ii)}{=} C + n\int_C^\infty \mathop{\mathbb{P}}_{\phi(\mathbf{x})\sim\mathcal{N}(\mathbf{0},\mathbf{\Sigma})}\left\{\|\phi(\mathbf{x})\|_{\mathcal{H}}\geq\sqrt{t}\right\}dt \tag{56}$$

$$= C + n\int_C^\infty \mathop{\mathbb{P}}_{\psi(\mathbf{x})\sim\mathcal{N}(\mathbf{0},\mathbf{I})}\left\{\left\|\mathbf{\Psi}\mathbf{\Lambda}^{1/2}\psi(\mathbf{x})\right\|_{\mathcal{H}}\geq\sqrt{t}\right\}dt \tag{57}$$

$$\leq C + n\int_C^\infty \mathop{\mathbb{P}}_{\psi_i(\mathbf{x})\sim\mathcal{N}(0,1)}\left\{\sum_{i=1}^p\sqrt{\lambda_i}\psi_i(\mathbf{x})\geq\sqrt{t}\right\}dt \tag{58}$$

$$\leq C + n\int_C^\infty \inf_{\theta>0}\mathop{\mathbb{E}}_{\psi(\mathbf{x})\sim\mathcal{N}(0,1)}\left[\prod_{i=1}^p\exp\left(\theta\sqrt{\lambda_i}\psi_i(\mathbf{x})\right)\right]\exp\left(-\theta\sqrt{t}\right)dt \tag{59}$$

$$= C + n\int_C^\infty \inf_{\theta>0}\exp\left[\theta^2\frac{\operatorname{Tr}(\mathbf{\Sigma})}{2}-\theta\sqrt{t}\right]dt = C + n\int_C^\infty \exp\left[-\frac{t}{2\operatorname{Tr}(\mathbf{\Sigma})}\right]dt \tag{60}$$

$$= C + \frac{n}{2\operatorname{Tr}(\mathbf{\Sigma})}\exp\left[-\frac{C}{2\operatorname{Tr}(\mathbf{\Sigma})}\right] \tag{61}$$

See (i) from the proof of Lemma 22. (ii) follows from the reproducing property. Setting $C \triangleq 2\operatorname{Tr}(\mathbf{\Sigma})\log(sn/(2\operatorname{Tr}(\mathbf{\Sigma})))$ completes the proof. ∎

Lemma 27 (Norm of Functions with Gaussian Design in the Reproducing Kernel Hilbert Space). *Let $\mathbf{x}_i\sim\mathbb{P}$ such that $\phi(\mathbf{x}_i)\sim\mathcal{N}(\mathbf{0},\mathbf{\Sigma})$ from Assumption 19 and Assumption 20. Then, it follows*

$$\mathop{\mathbb{E}}_{\phi(\mathbf{x}_i)\sim\mathcal{N}(\mathbf{0},\mathbf{\Sigma})}\mathop{\mathbb{E}}_{\mu_i\sim\mathcal{N}(0,\sigma^2)}\left\|\sum_{i=1}^n\mu_i\phi(\mathbf{x}_i)\right\|_{\mathcal{H}} \leq O\left(\sigma\sqrt{n\log n\operatorname{Tr}(\mathbf{\Sigma})}\right) \tag{62}$$

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,

$$\mathop{\mathbb{E}}_{\phi(\mathbf{x}_i)\sim\mathcal{N}(\mathbf{0},\mathbf{\Sigma})}\mathop{\mathbb{E}}_{\mu_i\sim\mathcal{N}(0,\sigma^2)}\left\|\sum_{i=1}^n\mu_i\phi(\mathbf{x}_i)\right\|_{\mathcal{H}} \leq \mathop{\mathbb{E}}_{\phi(\mathbf{x}_i)\sim\mathcal{N}(\mathbf{0},\mathbf{\Sigma})}\mathop{\mathbb{E}}_{\mu_i\sim\mathcal{N}(0,\sigma^2)}\max_{i\in[n]}|\mu_i|\left\|\sum_{i=1}^n\phi(\mathbf{x}_i)\right\|_{\mathcal{H}} \tag{63}$$

$$\stackrel{(i)}{\leq} O\left(\sigma\sqrt{\log n}\right)\mathop{\mathbb{E}}_{\phi(\mathbf{x}_i)\sim\mathcal{N}(\mathbf{0},\mathbf{\Sigma})}\mathop{\mathbb{E}}_{\xi_i\sim\mathcal{R}}\left\|\sum_{i=1}^n\xi_i\phi(\mathbf{x}_i)\right\|_{\mathcal{H}} \tag{64}$$

$$\stackrel{(ii)}{\leq} O\left(\sigma\sqrt{\log n}\right)\left(\mathop{\mathbb{E}}_{\phi(\mathbf{x}_i)\sim\mathcal{N}(\mathbf{0},\mathbf{\Sigma})}\mathop{\mathbb{E}}_{\xi_i\sim\mathcal{R}}\left\|\sum_{i=1}^n\xi_i\phi(\mathbf{x}_i)\right\|_{\mathcal{H}}^2\right)^{1/2} \tag{65}$$

$$= O\left(\sigma\sqrt{\log n}\right)\left(\mathop{\mathbb{E}}_{\phi(\mathbf{x}_i)\sim\mathcal{N}(\mathbf{0},\mathbf{\Sigma})}\mathop{\mathbb{E}}_{\xi_i\sim\mathcal{R}}\left\langle\sum_{i=1}^n\xi_i\phi(\mathbf{x}_i),\sum_{j=1}^n\xi_j\phi(\mathbf{x}_j)\right\rangle_{\mathcal{H}}\right)^{1/2} \tag{66}$$

$$\stackrel{(iii)}{=} O\left(\sigma\sqrt{\log n}\right)\left(\mathop{\mathbb{E}}_{\phi(\mathbf{x}_i)\sim\mathcal{N}(\mathbf{0},\mathbf{\Sigma})}\mathop{\mathbb{E}}_{\xi_i\sim\mathcal{R}}\sum_{i=1}^n\sum_{j=1}^n\xi_i\xi_j k(\mathbf{x}_i,\mathbf{x}_j)\right)^{1/2} \tag{67}$$

$$\stackrel{(iv)}{=} O\left(\sigma\sqrt{\log n}\right)\left(\mathop{\mathbb{E}}_{\phi(\mathbf{x}_i)\sim\mathcal{N}(\mathbf{0},\mathbf{\Sigma})}\sum_{i=1}^n k(x_i,x_i)\right)^{1/2} \tag{68}$$

$$= O\left(\sigma\sqrt{n\log n\operatorname{Tr}(\mathbf{\Sigma})}\right) \tag{69}$$

15

(i) follows from applying Lemma 22. (ii) follows from Jensen's Inequality. (iii) follows from the definition of the kernel (Gretton, 2013a). (iv) holds as we have $\mathbb{E}[\xi_i \xi_j] = \delta_{i,j}$, where δ is the Kronecker Delta function. $\blacksquare$

Proposition 28 (Probabilistic bound on Norm of Functions with Gaussian Design in the Reproducing Kernel Hilbert Space)**.** *Let $\mathbf{x}_i \sim \mathbb{P}$ such that $\phi(\mathbf{x}_i) \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$ from Assumption 19. Then it follows*

$$\mathbb{P}_{\phi(\mathbf{x}_i) \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})} \left\{ \left\| \sum_{i=1}^{n} \phi(\mathbf{x}_i) \right\|_{\mathcal{H}} \geq \sqrt{n \operatorname{Tr}(\boldsymbol{\Sigma})} \cdot u \right\} \leq e^{-u^2/2} \tag{70}$$

Proof. Our proof will utilize a symmetrization argument similar to our previous expected covariance approximation proof, the proof then follows similarly to Lemma 26. on the Let C be a positive constant to be determined and $u \geq 1$. We then have

$$\mathbb{P}_{\phi(\mathbf{x}_i) \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})} \left\{ \left\| \sum_{i=1}^{n} \phi(\mathbf{x}_i) \right\|_{\mathcal{H}} \geq C \cdot u \right\} = \mathbb{P}_{\phi(\mathbf{x}_i) \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})} \left\{ \mathbb{E}_{\xi_i \sim \mathcal{R}} \left\| \sum_{i=1}^{n} \xi_i \phi(\mathbf{x}_i) \right\|_{\mathcal{H}} \geq C \cdot u \right\} \tag{71}$$

$$\overset{(i)}{\leq} \mathbb{P}_{\phi(\mathbf{x}_i) \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})} \left\{ \sqrt{\mathbb{E}_{\xi_i \sim \mathcal{R}} \left[\left\| \sum_{i=1}^{n} \xi_i \phi(\mathbf{x}_i) \right\|_{\mathcal{H}}^2 \right]} \geq C \cdot u \right\} \tag{72}$$

$$= \mathbb{P}_{\phi(\mathbf{x}_i) \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})} \left\{ \sqrt{\mathbb{E}_{\xi_i \sim \mathcal{R}} \sum_{i=1}^{n} \sum_{j=1}^{n} \xi_i \xi_j k(\mathbf{x}_i, \mathbf{x}_j)} \geq C \cdot u \right\} \tag{73}$$

$$= \mathbb{P}_{\phi(\mathbf{x}_i) \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})} \left\{ \sqrt{\sum_{i=1}^{n} k(\mathbf{x}_i, \mathbf{x}_i)} \geq C \cdot u \right\} \leq \mathbb{P}_{\phi(\mathbf{x}_i) \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})} \left\{ \sum_{i=1}^{n} \|\phi(\mathbf{x}_i)\|_{\mathcal{H}} \geq C \cdot u \right\} \tag{74}$$

$$\overset{(ii)}{\leq} \inf_{\theta > 0} \prod_{i=1}^{n} \prod_{j=1}^{p} \exp\left[\frac{1}{2} \theta^2 \lambda_j \right] \exp\left[-\theta C \cdot u \right] = \exp\left[-\frac{C^2 \cdot u^2}{2 \operatorname{Tr}(\boldsymbol{\Sigma})} \right] \tag{75}$$

In (i) we use Jensen's Inequality. For (ii) see the proof for Lemma 26 with an additional product term over n. Finally, setting $C = \sqrt{n \operatorname{Tr}(\boldsymbol{\Sigma})}$ completes the proof.

Lemma 29 (Infinite Dimensional Covariance Estimation in the Hilbert-Schmidt Norm)**.** *Let $\boldsymbol{\Sigma} \triangleq \mathbb{E}_{\phi(\mathbf{x}_i) \sim \mathbb{P}}[\phi(\mathbf{x}_i) \otimes \phi(\mathbf{x}_i)]$. Then let $\mathbf{x}_1, \ldots, \mathbf{x}_n$ be i.i.d sampled from $\mathbb{P}$ such that $\phi(\mathbf{x}_i) \sim \mathcal{N}(0, \boldsymbol{\Sigma})$ from Assumption 19, we then have*

$$\mathbb{E}_{\phi(\mathbf{x}_i) \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})} \left\| \frac{1}{n} \sum_{i=1}^{n} \phi(\mathbf{x}_i) \otimes \phi(\mathbf{x}_i) - \boldsymbol{\Sigma} \right\|_{\mathrm{HS}} \leq C \left(\frac{\operatorname{Tr}(\boldsymbol{\Sigma})}{\sqrt{n}} \right) \tag{76}$$

where $C \leq 2\sqrt{3}$.

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

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

$$\overset{(i)}{\leq} \mathbb{E}_{\phi(\mathbf{x}_i) \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})} \mathbb{E}_{\tilde{\phi}(\mathbf{x}_i) \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})} \left\| \frac{1}{n} \sum_{i=1}^{n} \left(\phi(\mathbf{x}_i) \otimes \phi(\mathbf{x}_i) - \tilde{\phi}(\mathbf{x}_i) \otimes \tilde{\phi}(\mathbf{x}_i) \right) \right\|_{\mathrm{HS}} \tag{77}$$

$$= \mathbb{E}_{\phi(\mathbf{x}_i) \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})} \mathbb{E}_{\tilde{\phi}(\mathbf{x}_i) \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})} \mathbb{E}_{\xi_i \sim \mathcal{R}} \left\| \frac{1}{n} \sum_{i=1}^{n} \xi_i \left(\phi(\mathbf{x}_i) \otimes \phi(\mathbf{x}_i) - \tilde{\phi}(\mathbf{x}_i) \otimes \tilde{\phi}(\mathbf{x}_i) \right) \right\|_{\mathrm{HS}} \tag{78}$$

$$\overset{(ii)}{\leq} \frac{2}{n} \mathbb{E}_{\phi(\mathbf{x}_i) \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})} \mathbb{E}_{\xi_i \sim \mathcal{R}} \left\| \sum_{i=1}^{n} \xi_i \phi(\mathbf{x}_i) \otimes \phi(\mathbf{x}_i) \right\|_{\mathrm{HS}} \tag{79}$$

$$\stackrel{(iii)}{\leq} \frac{2}{n} \left(\mathop{\mathbb{E}}_{\phi(\mathbf{x}_i) \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma})} \mathop{\mathbb{E}}_{\xi_i \sim \mathcal{R}} \left\| \sum_{i=1}^{n} \xi_i \phi(\mathbf{x}_i) \otimes \phi(\mathbf{x}_i) \right\|_{\mathrm{HS}}^2 \right)^{1/2} \tag{80}$$

(i) follows from noticing $\phi(\mathbf{x}_i) \otimes \phi(\mathbf{x}_i) - \mathbf{\Sigma}$ is a mean $\mathbf{0}$ operator in $\mathcal{H} \otimes \mathcal{H}$, then for $X, Y \in \mathcal{H} \otimes \mathcal{H}$ s.t. $\mathbb{E}[Y] = \mathbf{0}$ it follows $\|X\|_{\mathrm{HS}} = \|X - \mathbb{E}[Y]\|_{\mathrm{HS}} = \|\mathbb{E}_Y[X - Y]\|_{\mathrm{HS}}$ 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

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

$$= \frac{2}{n} \left(\mathop{\mathbb{E}}_{\phi(\mathbf{x}_i) \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma})} \mathop{\mathbb{E}}_{\xi_i \sim \mathcal{R}} \sum_{k=1}^{p} \left\langle \sum_{i=1}^{n} \xi_i \phi(\mathbf{x}_i) \otimes \phi(\mathbf{x}_i) e_k, \sum_{j=1}^{n} \xi_j \phi(\mathbf{x}_j) \otimes \phi(\mathbf{x}_j) e_k \right\rangle_{\mathcal{H}} \right)^{1/2} \tag{81}$$

$$= \frac{2}{n} \left(\mathop{\mathbb{E}}_{\phi(\mathbf{x}_i) \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma})} \mathop{\mathbb{E}}_{\xi_i \sim \mathcal{R}} \sum_{k=1}^{p} \sum_{i=1}^{n} \sum_{j=1}^{n} \xi_i \xi_j \left\langle \phi(\mathbf{x}_i) \otimes \phi(\mathbf{x}_i) e_k, \phi(\mathbf{x}_j) \otimes \phi(\mathbf{x}_j) e_k \right\rangle_{\mathcal{H}} \right)^{1/2} \tag{82}$$

$$\stackrel{(iv)}{\leq} \frac{2}{n} \left(\mathop{\mathbb{E}}_{\phi(\mathbf{x}_i) \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma})} \sum_{k=1}^{p} \sum_{i=1}^{n} \left\langle \phi(\mathbf{x}_i) \otimes \phi(\mathbf{x}_i) e_k, \phi(\mathbf{x}_i) \otimes \phi(\mathbf{x}_i) e_k \right\rangle_{\mathcal{H}} \right)^{1/2} \tag{83}$$

$$= \frac{2}{n} \left(\sum_{i=1}^{n} \mathop{\mathbb{E}}_{\phi(\mathbf{x}_i) \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma})} \|\phi(\mathbf{x}_i) \otimes \phi(\mathbf{x}_i)\|_{\mathrm{HS}}^2 \right)^{1/2} \stackrel{(v)}{=} \frac{2}{n} \left(\sum_{i=1}^{n} \mathop{\mathbb{E}}_{\phi(\mathbf{x}_i) \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma})} \|\phi(\mathbf{x}_i)\|_{\mathcal{H}}^4 \right)^{1/2} \tag{84}$$

$$= \frac{2}{n} \left(\sum_{i=1}^{n} \mathop{\mathbb{E}}_{\phi(\mathbf{x}_i) \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma})} \left[k^2(x_i, x_i) \right] \right)^{1/2} = \frac{2}{\sqrt{n}} \left(2 \operatorname{Tr}\left(\mathbf{\Sigma}^2 \right) + \operatorname{Tr}\left(\mathbf{\Sigma} \right)^2 \right)^{1/2} \leq 2\sqrt{3} n^{-1/2} \operatorname{Tr}\left(\mathbf{\Sigma} \right) \tag{85}$$

(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 $\operatorname{Tr}(\mathbf{\Sigma}) < \infty$ 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. ∎

Proposition 30 (Probabilistic Bound on Infinite Dimensional Covariance Estimation in the Hilbert-Schmidt Norm)**.** *Let* $\mathbf{x}_1, \ldots, \mathbf{x}_n$ *be i.i.d sampled from* $\mathbb{P}$ *such that* $\phi(\mathbf{x}_i) \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$ *from Assumption 19, we then have*

$$\mathop{\mathbb{P}}_{\phi(\mathbf{x}_i) \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma})} \left\{ \left\| \frac{1}{n} \sum_{i=1}^{n} \phi(\mathbf{x}_i) \otimes \phi(\mathbf{x}_i) - \mathbf{\Sigma} \right\|_{\mathrm{HS}} \geq C \cdot u \right\} \leq f(u) \tag{86}$$

Proof. *(In Progress)* Let C be a to be determined positive constant and $u \geq 1$.

$$\mathop{\mathbb{P}}_{\phi(\mathbf{x}_i) \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma})} \left\{ \left\| \frac{1}{n} \sum_{i=1}^{n} \phi(\mathbf{x}_i) \otimes \phi(\mathbf{x}_i) - \mathbf{\Sigma} \right\|_{\mathrm{HS}} \geq C \cdot u \right\} \tag{87}$$

$$\leq \mathop{\mathbb{P}}_{\phi(\mathbf{x}_i) \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma})} \left\{ \mathop{\mathbb{E}}_{\xi_i \sim \mathcal{R}} \left\| \frac{1}{n} \sum_{i=1}^{n} \xi_i \phi(\mathbf{x}_i) \otimes \phi(\mathbf{x}_i) \right\|_{\mathrm{HS}} + \mathop{\mathbb{E}}_{\tilde{\phi}(\mathbf{x}_i) \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma})} \mathop{\mathbb{E}}_{\xi_i \sim \mathcal{R}} \left\| \frac{1}{n} \sum_{i=1}^{n} \xi_i \tilde{\phi}(\mathbf{x}_i) \otimes \tilde{\phi}(\mathbf{x}_i) \right\|_{\mathrm{HS}} \geq C \cdot u \right\}$$

$$\leq \mathop{\mathbb{P}}_{\phi(\mathbf{x}_i) \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma})} \left\{ \sqrt{\sum_{i=1}^{n} k(\mathbf{x}_i, \mathbf{x}_i)} + \frac{\operatorname{Tr}(\mathbf{\Sigma})}{\sqrt{n}} \geq C \cdot u \right\} \tag{88}$$

Lemma 31 (Finite Dimensional Covariate Estimation in the Spectral Norm)**.** *Let* $\mathbf{x}_1, \ldots, \mathbf{x}_n \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$. *It then follows,*

$$\mathop{\mathbb{E}}_{\mathbf{x}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma})} \left\| \frac{1}{n} \sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^\top - \mathbf{\Sigma} \right\|_2 \leq C \|\mathbf{\Sigma}\| \left(\sqrt{\frac{d}{n}} + \frac{1}{\sqrt{dn}} \right) \tag{89}$$

where $C \leq 62.82$.

17

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

$$\mathbb{E}_{\mathbf{x}_i \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})} \left\| \frac{1}{n} \sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^\top - \boldsymbol{\Sigma} \right\|_2 \leq \|\boldsymbol{\Sigma}\| \, \mathbb{E}_{\tilde{\mathbf{x}}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I})} \left\| \frac{1}{n} \tilde{\mathbf{X}}^\top \tilde{\mathbf{X}} - \mathbf{I} \right\|_2 \tag{90}$$

$$= \|\boldsymbol{\Sigma}\| \int_0^\infty \mathbb{P}_{\tilde{\mathbf{x}}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I})} \left\{ \left\| \frac{1}{n} \tilde{\mathbf{X}}^\top \tilde{\mathbf{X}} - \mathbf{I} \right\|_2 \geq t \right\} dt \tag{91}$$

Let $\mathcal{M}$ be an ε-net of $\mathbb{S}^{d-1}$ for $\varepsilon = \frac{1}{4}$, then $|\mathcal{M}| \leq 9^d$. It then follows from (Vershynin, 2020) Corollary 4.2.13,

$$\mathbb{P}_{\tilde{\mathbf{x}}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I})} \left\{ \left\| \frac{1}{n} \tilde{\mathbf{X}}^\top \tilde{\mathbf{X}} - \mathbf{I} \right\|_2 \geq t \right\} \leq \mathbb{P}_{\tilde{\mathbf{x}}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I})} \left\{ \max_{\mathbf{y} \in \mathcal{M}} \left| \frac{1}{n} \|\tilde{\mathbf{X}} \mathbf{y}\|_2^2 - 1 \right| \geq \frac{t}{2} \right\} \tag{92}$$

Denote $K \triangleq 16\sqrt{\frac{8}{3}}$, then we have from a union bound and Bernstein's Inequality (Bernstein, 1924).

$$\mathbb{P}_{\tilde{\mathbf{x}}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I})} \left\{ \max_{\mathbf{y} \in \mathcal{M}} \left| \frac{1}{n} \|\tilde{\mathbf{X}} \mathbf{y}\|_2^2 - 1 \right| \geq \frac{t}{2} \right\} \leq 9^d \exp\left[-\frac{n}{2} \left(\frac{t^2}{K^2} \wedge \frac{t}{K} \right) \right] \tag{93}$$

Let $\delta \in (0, 1)$, then we find RHS Equation (93) is less than δ when

$$t \geq K \left(\frac{2d \log(9) + 2\log(2/\delta)}{n} \vee \left(\frac{2d \log(9) + 2\log(2/\delta)}{n} \right)^{1/2} \right) \tag{94}$$

Furthermore, we note we have equality with one, when $t = K\sqrt{(2d\log(9) + \log(4))/n} \triangleq C$ all $t \leq C$ occur with probability also equal to one. Therefore, plugging this back into the RHS of Equation (91).

$$\int_0^\infty \mathbb{P}_{\tilde{\mathbf{x}}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I})} \left\{ \left\| \frac{1}{n} \tilde{\mathbf{X}}^\top \tilde{\mathbf{X}} - \mathbf{I} \right\|_2 \geq t \right\} dt \leq K\sqrt{\frac{2d\log(9)}{n} + \frac{2\log(2)}{n}} + \int_C^\infty 9^d \exp\left[-\frac{n}{2} \left(\frac{t^2}{K^2} \right) \right] dt \tag{95}$$

$$\leq K\sqrt{\frac{2d\log(9)}{n} + \frac{2\log(2)}{n}} + \frac{K 9^d \exp\left[-\left(K\sqrt{\frac{2d\log(9)}{n} + \frac{2\log(2)}{n}} \right)^2 \left(\frac{n}{2K^2} \right) \right]}{2n\sqrt{\frac{2d\log(9)}{n} + \frac{2\log(2)}{n}}} \tag{96}$$

$$\leq K\sqrt{\frac{2d\log(9) + \log(4)}{n}} + \frac{K}{4\sqrt{n\left(2d\log(9) + \log(4)\right)}} \tag{97}$$

$$\leq \sqrt{\log(324)} K \left(\sqrt{\frac{d}{n}} + \frac{1}{\sqrt{nd}} \right) \tag{98}$$

In the second inequality, we use the integral inequality $\int_c^\infty e^{-x^2} dx \leq \int_c^\infty (\frac{x}{c}) e^{-x^2} dx = e^{-c^2}/(2c)$. ∎

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.

$$\partial(-g(t, f_{\mathbf{w}})) = \partial\left(-t + \frac{1}{np} \sum_{i=1}^{n} (t - \hat{\nu}_i)^+ \right) = -1 + \frac{1}{np} \sum_{i=1}^{np} \begin{cases} 1 & \text{if } t > \hat{\nu}_i \\ 0 & \text{if } t < \hat{\nu}_i \\ [0, 1] & \text{if } t = \hat{\nu}_i \end{cases} \tag{99}$$

We observe when setting $t = \hat{\nu}_{np}$, it follows that $0 \in \partial(-g(t, f_{\mathbf{w}}))$. This is equivalent to the p-quantile of the Risk. ∎

B.2   Proof of Lemma 4

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

$$\nabla_f g(t^{(k+1)}, f_{\mathbf{w}}^{(k)}) = \nabla_f \left(t^{(k+1)} - \frac{1}{np} \sum_{i=1}^{n} \left(t^{(k+1)} - \ell(\mathbf{x}_i; f_{\mathbf{w}}^{(k)}, y_i) \right)^+ \right) \tag{100}$$

$$= -\frac{1}{np} \sum_{i=1}^{np} \nabla_f \left(t^{(k+1)} - \ell(\mathbf{x}_i; f_{\mathbf{w}}^{(k)}, y_i) \right)^+ = \frac{1}{np} \sum_{i=1}^{n} \nabla_f \ell(\mathbf{x}_i; f_{\mathbf{w}}^{(k)}, y_i) \begin{cases} 1 & \text{if } t > \hat{\nu}_i \\ 0 & \text{if } t < \hat{\nu}_i \\ [0,1] & \text{if } t = \hat{\nu}_i \end{cases} \tag{101}$$

Now we note $\nu_{np} \leq t^{(k+1)} \leq \nu_{np+1}$. Then, plugging this into Equation (101), we have

$$\nabla_f g(t^{(k+1)}, f_{\mathbf{w}}^{(k)}) = \frac{1}{np} \sum_{i=1}^{np} \nabla_f \ell(\mathbf{x}_i; f_{\mathbf{w}}^{(k)}, y_i) \tag{102}$$

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 $\lambda \to \infty$.

$$\nabla_t \tilde{g}_\lambda(t, f_{\mathbf{w}}) = \nabla_t \left(t - \frac{1}{np} \sum_{i=1}^{n} \frac{1}{\lambda} \ln\left(1 + \exp\left(\lambda\left(t - \ell(f_{\mathbf{w}}; \mathbf{x}_i, y_i)\right)\right)\right) \right) \tag{103}$$

$$= 1 - \frac{1}{np} \sum_{i=1}^{n} \sigma\left(\lambda\left(t - \ell(f_{\mathbf{w}}; \mathbf{x}_i, y_i)\right)\right) \tag{104}$$

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,

$$\lim_{\lambda \to \infty} \nabla_t \tilde{g}_\lambda(t, f_{\mathbf{w}}) = 1 - \frac{1}{np} \sum_{i=1}^{n} \mathbb{I}\left\{t - \ell\left(f_{\mathbf{w}}; \mathbf{x}_i, y_i\right)\right\} \tag{105}$$

Next, we will calculate the derivative of $g(t, f_{\mathbf{w}})$ with respect to f.

$$\nabla_f \tilde{g}_\lambda(t, f_{\mathbf{w}}) = \nabla_f \left(t - \frac{1}{np} \sum_{i=1}^{n} \frac{1}{\lambda} \ln\left(1 + \exp\left(\lambda\left(t - \ell(f_{\mathbf{w}}; \mathbf{x}_i, y_i)\right)\right)\right) \right) \tag{106}$$

$$= \frac{1}{np} \sum_{i=1}^{n} \nabla_f \ell\left(f_{\mathbf{w}}; \mathbf{x}_i, y_i\right) \sigma\left(\lambda\left(t - \ell(f_{\mathbf{w}}; \mathbf{x}_i, y_i)\right)\right) \tag{107}$$

We therefore similarly have the derivative of g with respect to f as λ is pushed to infinity.

$$\lim_{\lambda \to \infty} \nabla_f \tilde{g}_\lambda(t, f_{\mathbf{w}}) = \frac{1}{np} \sum_{i=1}^{n} \mathbb{I}\left\{t - \ell\left(f_{\mathbf{w}}; \mathbf{x}_i, y_i\right)\right\} \nabla_f \ell\left(f_{\mathbf{w}}; \mathbf{x}_i, y_i\right) \tag{108}$$

B.4   Proof of Lemma 5

Proof. We will upper bound the operator norm of the Hessian. Let $\upsilon \triangleq \sigma(\lambda(t - \ell(f_{\mathbf{w}}; \mathbf{x}_i, y_i)))$, we then have

$$\nabla_f^2 \tilde{g}_\lambda(t, f_{\mathbf{w}}) = \nabla_f \left(\frac{1}{np} \sum_{i=1}^{n} \upsilon \nabla_f \ell\left(f_{\mathbf{w}}; \mathbf{x}_i, y_i\right) \right) \tag{109}$$

$$= \frac{1}{np} \sum_{i=1}^{n} \left(\upsilon \nabla_f^2 \ell \left(f_{\mathbf{w}}; \mathbf{x}_i, y_i \right) - \upsilon(1 - \upsilon) \left(\nabla_f \ell \left(f_{\mathbf{w}}; \mathbf{x}_i, y_i \right) \otimes \nabla_f \ell \left(f_{\mathbf{w}}; \mathbf{x}_i, y_i \right) \right) \right) \tag{110}$$

$$\overset{(i)}{\preceq} \frac{1}{np} \sum_{i=1}^{n} \left(\upsilon \nabla_f^2 \ell \left(f_{\mathbf{w}}; \mathbf{x}_i, y_i \right) \right) \tag{111}$$

In (i), $\preceq$ denotes the Löwner Ordering. Now we will upper bound the operator norm of the Hessian.

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

$$\overset{(110)}{=} \lim_{\lambda \to \infty} \sup_{f_{\mathbf{w}}} \left\| \frac{1}{np} \sum_{i=1}^{n} \upsilon \nabla_f^2 \ell \left(f_{\mathbf{w}}; \mathbf{x}_i, y_i \right) - \upsilon(1 - \upsilon) \nabla_f \ell \left(f_{\mathbf{w}}; \mathbf{x}_i, y_i \right) \otimes \nabla_f \ell \left(f_{\mathbf{w}}; \mathbf{x}_i, y_i \right) \right\|_{\text{op}} \| f_{\mathbf{w}} - f_{\tilde{\mathbf{w}}} \|_{\mathcal{H}} \tag{112}$$

$$\overset{(111)}{\leq} \lim_{\lambda \to \infty} \sup_{f_{\mathbf{w}} \in \mathcal{K}} \frac{1}{np} \sum_{i=1}^{n} \left\| \upsilon \nabla_f^2 \ell \left(f_{\mathbf{w}}; \mathbf{x}_i, y_i \right) \right\|_{\text{op}} \| f_{\mathbf{w}} - f_{\tilde{\mathbf{w}}} \|_{\mathcal{H}} \tag{113}$$

$$\overset{(ii)}{\leq} \sup_{f_{\mathbf{w}} \in \mathcal{K}} \frac{1}{np} \sum_{i=1}^{n} \left\| \nabla_f^2 \ell \left(f_{\mathbf{w}}; \mathbf{x}_i, y_i \right) \right\|_{\text{op}} \| f_{\mathbf{w}} - f_{\tilde{\mathbf{w}}} \|_{\mathcal{H}} \tag{114}$$

(i) follows from applying Weyl's Inequality (Weyl, 1912). (ii) follows from noting $\upsilon \in (0, 1)$. We now note that removing υ also removes the dependence on λ which allows us to take the limit out of the expression. ∎

B.5 Proof of Lemma 10

Proof. By the definition of stationary point, we have

$$f_{\hat{\mathbf{w}}} = \lim_{\lambda \to \infty} \arg\inf_{f_{\mathbf{w}} \in \mathcal{K}} \left\{ \Phi_\lambda \left(f_{\mathbf{w}} \right) + \frac{1}{2\rho} \| f_{\mathbf{w}} - f_{\hat{\mathbf{w}}} \|_{\mathcal{H}}^2 \right\} \tag{115}$$

$$\overset{(i)}{=} \arg\inf_{f_{\mathbf{w}} \in \mathcal{K}} \left\{ \lim_{\lambda \to \infty} \Phi_\lambda \left(f_{\mathbf{w}} \right) + \frac{1}{2\rho} \| f_{\mathbf{w}} - f_{\hat{\mathbf{w}}} \|_{\mathcal{H}}^2 \right\} \tag{116}$$

(i) holds as we ρ is independent of λ as shown in the proof of Lemma 5. This implies then for any $f_{\mathbf{w}} \in \mathcal{K}$ and noting $\rho \leq \beta^{-1}$, it follows

$$\lim_{\lambda \to \infty} \Phi_\lambda \left(f_{\hat{\mathbf{w}}} \right) \leq \lim_{\lambda \to \infty} \Phi_\lambda \left(f_{\mathbf{w}} \right) + \beta \| f_{\mathbf{w}} - f_{\hat{\mathbf{w}}} \|_{\mathcal{H}}^2 \tag{117}$$

where we choose $\rho \triangleq 1/(2\beta)$. We can then plug in the optimal, $f_{\mathbf{w}}^*$ for $f_{\mathbf{w}}$ and rearrange and we have the desired result. ∎

C Proofs for Kernelized Regression

C.1 L-Lipschitz Constant and β-Smoothness

Lemma 32 (L-Lipschitz of $g(t, \mathbf{w})$ w.r.t $\mathbf{w}$). *Let* $\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n$, *represent the data vectors. It then follows:*

$$|g(t, f_{\mathbf{w}}) - g(t, f_{\hat{\mathbf{w}}})| \leq L \| f_{\mathbf{w}} - f_{\hat{\mathbf{w}}} \|_{\mathcal{H}} \tag{118}$$

where

$$L = \frac{2R}{np} \left(\sum_{i=1}^{n} \sqrt{k(\mathbf{x}_i, \mathbf{x}_i)} \right)^2 + \frac{2 \|\mathbf{y}\|_2}{p\sqrt{n}} \left(\sum_{i=1}^{n} \sqrt{k(\mathbf{x}_i, \mathbf{x}_i)} \right) \tag{119}$$

Proof. For any $f_{\mathbf{w}_1}, f_{\mathbf{w}_2} \in \mathcal{K}$, we will first show the gradient is bounded.

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

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

$$\overset{(i)}{\leq} \|f_{\mathbf{w}_1} - f_{\mathbf{w}_2}\|_{\mathcal{H}} \max_{f_{\mathbf{w}} \in \mathcal{K}} \|\nabla_f g(t, f_{\mathbf{w}})\|_{\mathcal{H}} \tag{122}$$

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

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

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

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

$$\overset{(i)}{\leq} \frac{2}{np} \left(\left\| \sum_{i=1}^m \left\langle \sum_{j=1}^n w_j k(\mathbf{x}_j, \cdot), k(\mathbf{x}_i, \cdot) \right\rangle_{\mathcal{H}} \cdot k(\mathbf{x}_i, \cdot) \right\|_{\mathcal{H}} + \left\| \sum_{i=1}^m y_i \right\| \left\| \sum_{i=1}^m k(\mathbf{x}_i, \cdot) \right\|_{\mathcal{H}} \right) \tag{125}$$

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

$$\leq \frac{2}{np} \left(\|f_{\mathbf{w}}\|_{\mathcal{H}} \left(\sum_{i=1}^m \sqrt{k(\mathbf{x}_i, \mathbf{x}_i)} \right)^2 + \sqrt{n} \|\mathbf{y}\|_2 \left(\sum_{i=1}^n \sqrt{k(\mathbf{x}_i, \mathbf{x}_i)} \right) \right) \tag{127}$$

$$\leq \frac{2R}{np} \left(\sum_{i=1}^n \sqrt{k(\mathbf{x}_i, \mathbf{x}_i)} \right)^2 + \frac{2\|\mathbf{y}\|_2}{p\sqrt{n}} \left(\sum_{i=1}^n \sqrt{k(\mathbf{x}_i, \mathbf{x}_i)} \right) \tag{128}$$

(i) follows form the reproducing property for RKHS (Gretton, 2013b). 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 $f_{\mathbf{w}}$ has bounded norm, so it will have infinite error. Similar results for the Lipschitz Constant for non-kernelized learning algorithms can be seen in (Yedida et al., 2021). This concludes the proof. ∎

Lemma 33. *(β-Smoothness of $g(t, \mathbf{w})$ w.r.t $\mathbf{w}$). Let $\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n$ represent the rows of the data matrix $\mathbf{X}$. It then follows:*

$$\|\nabla_f g(t, f_{\mathbf{w}}) - \nabla_f g(t, f_{\hat{\mathbf{w}}})\| \leq \beta \|f_{\mathbf{w}} - f_{\hat{\mathbf{w}}}\|_{\mathcal{H}} \tag{129}$$

where

$$\beta = \frac{2}{np} \sum_{i=1}^n k(x_i, x_i) = \frac{2}{np} \operatorname{Tr}(\mathbf{K}) \tag{130}$$

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

$$\left\| \nabla_f^2 g(t, f_{\mathbf{w}}) \right\|_{\text{op}} = \left\| \frac{2}{np} \sum_{i=1}^n \mathbb{I}\left\{ t \geq (f_{\mathbf{w}}(\mathbf{x}_i) - y_i)^2 \right\} \phi(\mathbf{x}_i) \otimes \phi(\mathbf{x}_i) \right\|_{\text{op}} \tag{131}$$

$$\leq \frac{2}{np} \sum_{i=1}^n \|\phi(\mathbf{x}_i) \otimes \phi(\mathbf{x}_i)\|_{\text{HS}} = \frac{2}{np} \sum_{i=1}^n k(\mathbf{x}_i, \mathbf{x}_i) = \frac{2}{np} \operatorname{Tr}(\mathbf{K}) \tag{132}$$

This concludes the proof.

C.2 Proof of Lemma 12

Proof. 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 $\mathbb{E}_{\mathbf{x} \sim \mathbb{P}}[\phi(\mathbf{x}) \otimes \phi(\mathbf{x})] \triangleq \boldsymbol{\Sigma}$.

$$\langle f_{\mathbf{w}} - f_{\mathbf{w}}^*, \boldsymbol{\Sigma} (f_{\mathbf{w}} - f_{\mathbf{w}}^*) \rangle = \lim_{p \to \infty} \sum_{i=1}^{p} \lambda_i \langle f_{\mathbf{w}} - f_{\mathbf{w}}^*, \varphi_i \rangle_{\mathcal{H}}^2 \tag{133}$$

Therefore, for some $m \in \mathbb{N}$, we want the projection of $f_{\mathbf{w}} - f_{\mathbf{w}}^*$ to be non-zero for m. We will show, we only need to make an assumption on $f_{\mathbf{w}}^*$. Let Ψ represent the concatenation of the eigenfunctions φ_i for $i \in 1, \dots, p$ as $p \to \infty$. Furthermore, let Ψ_m represent the m eigenfunctions corresponding to $\lambda_1, \dots, \lambda_m$ and $\Psi_m^\perp$ represents the eigenfunctions corresponding to $\lambda_{m+1}, \dots,$. The projection in the Reproducing Kernel Hilbert Space is given as the following,

$$\left\| \mathrm{Proj}_{\Psi_m} f_{\mathbf{w}}^* \right\|_{\mathcal{H}} \triangleq \left\| \sum_{i=1}^{m} \langle \varphi_i, f_{\mathbf{w}}^* \rangle \varphi_i \right\|_{\mathcal{H}} = \sum_{i=1}^{m} |\langle \varphi_i, f_{\mathbf{w}}^* \rangle_{\mathcal{H}}| \tag{134}$$

Let $f_{\mathbf{w}}^{(T)}$ be the T iterate from Subquantile Kernel Algorithm. From our assumption that $f_{\mathbf{w}}^* \in \mathrm{Span}(\{\varphi_i\}_{i=1}^{m}) \triangleq \mathrm{Span}(\Psi_m)$, it suffices to prove $\mathrm{Proj}_{\Psi_m} f_{\mathbf{w}}^{(t)} \neq f_{\mathbf{w}}^*$ for all $t \in [\![T]\!]$. In this case, we can note that $f_{\mathbf{w}} - f_{\mathbf{w}}^*$ will in some part be in the Span of Ψ_m.

$$\left\langle f_{\mathbf{w}}^{(t)} - f_{\mathbf{w}}^*, \boldsymbol{\Sigma} \left(f_{\mathbf{w}}^{(t)} - f_{\mathbf{w}}^* \right) \right\rangle_{\mathcal{H}} \overset{(133)}{=} \lim_{p \to \infty} \sum_{i=1}^{p} \lambda_i \left\langle f_{\mathbf{w}}^{(t)} - f_{\mathbf{w}}^*, \varphi_i \right\rangle_{\mathcal{H}}^2 \tag{135}$$

$$\overset{(i)}{=} \sum_{i=1}^{m} \lambda_i \left\langle f_{\mathbf{w}}^{(t)} - f_{\mathbf{w}}^*, \varphi_i \right\rangle_{\mathcal{H}}^2 + \lim_{p \to \infty} \sum_{i=m+1}^{p} \lambda_i \left\langle f_{\mathbf{w}}^{(t)}, \varphi_i \right\rangle_{\mathcal{H}}^2 \tag{136}$$

$$= \sum_{i=1}^{m} \lambda_i \left\langle f_{\mathbf{w}}^{(t)} - f_{\mathbf{w}}^*, \varphi_i \right\rangle_{\mathcal{H}}^2 + \left\| \mathrm{Proj}_{\Psi_m^\perp} f_{\mathbf{w}}^{(t)} \right\|_{\mathcal{H}}^2 \tag{137}$$

$$\overset{(ii)}{\geq} \lambda_m C \| f_{\mathbf{w}}^{(t)} - f_{\mathbf{w}}^* \|_{\mathcal{H}}^2 \tag{138}$$

Note in (i), since $\left\| \mathrm{Proj}_{m,\perp} f_{\mathbf{w}}^* \right\|_{\mathcal{H}} = 0$, then it holds for all $i \in \mathbb{N}$ s.t. $i > m$ that $\langle f_{\mathbf{w}}^*, \varphi_i \rangle_{\mathcal{H}} = 0$. In (ii), we define $C \triangleq (\| f_{\mathbf{w}}^{(t)} - f_{\mathbf{w}}^* \|_{\mathcal{H}}^2 - \| \mathrm{Proj}_{\Psi_m^\perp} f_{\mathbf{w}}^{(t)} \|_{\mathcal{H}}^2) / \| f_{\mathbf{w}}^{(t)} - f_{\mathbf{w}}^* \|_{\mathcal{H}}^2 > 0$ from our assumption that $f_{\mathbf{w}}^* \neq \mathrm{Proj}_{\Psi_m} f_{\mathbf{w}}^{(t)}$ for all $t \in [\![T]\!]$. This concludes the proof. ∎

Lemma 34. *Let* $\mathcal{K} \triangleq \{f_{\mathbf{w}} : \| f_{\mathbf{w}} \|_{\mathcal{H}} \leq R\}$. *Then, for a* $f_{\hat{\mathbf{w}}} \notin \mathcal{K}$, *it follows*

$$\mathrm{Proj}_{\mathcal{K}} f_{\hat{\mathbf{w}}} = \Omega(1) f_{\hat{\mathbf{w}}} \tag{139}$$

Proof. We will formulate the dual problem and then find the corresponding $f_{\mathbf{w}}$ that solves the dual.

$$\mathrm{Proj}_{\mathcal{K}} f_{\hat{\mathbf{w}}} = \arg\min_{f_{\mathbf{w}} \in \mathcal{K}} \| f_{\mathbf{w}} - f_{\hat{\mathbf{w}}} \|_{\mathcal{H}}^2 = \arg\min_{f_{\mathbf{w}} \in \mathcal{K}} \| f_{\mathbf{w}} \|_{\mathcal{H}}^2 + \| f_{\hat{\mathbf{w}}} \|_{\mathcal{H}}^2 - 2 \langle f_{\mathbf{w}}, f_{\hat{\mathbf{w}}} \rangle_{\mathcal{H}} \tag{140}$$

$$= \arg\min_{f_{\mathbf{w}} \in \mathcal{K}} \| f_{\mathbf{w}} \|_{\mathcal{H}}^2 - 2 \langle f_{\mathbf{w}}, f_{\hat{\mathbf{w}}} \rangle_{\mathcal{H}} \tag{141}$$

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

$$\mathcal{L}(f_{\mathbf{w}}, u) \triangleq \| f_{\mathbf{w}} \|_{\mathcal{H}}^2 - 2 \langle f_{\mathbf{w}}, f_{\hat{\mathbf{w}}} \rangle + u \left(\| f_{\mathbf{w}} \|_{\mathcal{H}}^2 - R^2 \right) \tag{142}$$

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

Lemma 35. *If* $\|f_{\mathbf{w}} - f_{\mathbf{w}^*}\| \geq \eta$, *then it follows*

$$
\lim_{\lambda \to \infty} \left(\Phi_\lambda \left(f_{\mathbf{w}} \right) - \Phi_\lambda \left(f_{\mathbf{w}^*} \right) \right) \geq \eta^2 \left(\tilde{n} \lambda_{\min} \left(\mathbb{E}_{\mathbf{x} \sim \mathbb{P}} [\phi(\mathbf{x}) \otimes \phi(\mathbf{x})] \right) \right)
$$
$$
- O \left(\sigma \sqrt{\tilde{n} \log(\tilde{n}) \|\mathbf{\Sigma}\|_{\mathrm{HS}}} \right) - 2\eta \left\| \sum_{i \in S \cap P} \mu_i \phi\left(\mathbf{x}_i \right) \right\| - \sum_{j \in P \setminus S} \mu_j^2 \tag{143}
$$

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

$$
\lim_{\lambda \to \infty} \left(\Phi_\lambda \left(f_{\mathbf{w}} \right) - \Phi_\lambda \left(f_{\mathbf{w}}^* \right) \right) = \sum_{i \in S} \left(f_{\mathbf{w}} \left(\mathbf{x}_i \right) - y_i \right)^2 - \sum_{j \in P} \left(f_{\mathbf{w}}^* \left(\mathbf{x}_j \right) - y_j \right)^2 \tag{144}
$$

$$
= \sum_{i \in S \cap P} \left(f_{\mathbf{w}} \left(\mathbf{x}_i \right) - y_i \right)^2 + \sum_{i \in S \cap Q} \left(f_{\mathbf{w}} \left(\mathbf{x}_i \right) - y_i \right)^2 - \sum_{j \in P} \left(f_{\mathbf{w}}^* \left(\mathbf{x}_j \right) - y_j \right)^2 \tag{145}
$$

$$
\geq \sum_{i \in S \cap P} \left(f_{\mathbf{w}} \left(\mathbf{x}_i \right) - y_i \right)^2 - \sum_{j \in P} \left(f_{\mathbf{w}}^* \left(\mathbf{x}_j \right) - y_j \right)^2 = \sum_{i \in S \cap P} \left(f_{\mathbf{w}} \left(\mathbf{x}_i \right) - f_{\mathbf{w}}^* \left(\mathbf{x}_i \right) - \mu_i \right)^2 - \sum_{j \in P} \mu_j^2 \tag{146}
$$

$$
= \sum_{i \in S \cap P} \underbrace{\left(\left(f_{\mathbf{w}} - f_{\mathbf{w}}^* \right) \left(\mathbf{x}_i \right) \right)^2}_{A_1} - 2 \underbrace{\sum_{i \in S \cap P} \mu_i \left(f_{\mathbf{w}} - f_{\mathbf{w}}^* \right) \left(\mathbf{x}_i \right)}_{A_2} - \underbrace{\sum_{j \in P \setminus S} \mu_j^2}_{A_3} \tag{147}
$$

Now we will upper bound A_1. Similar to (Cui et al., 2021) Let $\mathbb{E}_{\mathbf{x} \sim \mathbb{P}}[\varphi(\mathbf{x}) \otimes \varphi(\mathbf{x})] = \mathbb{I}_m$ where $\varphi(\mathbf{x}) = \{\varphi(\mathbf{x})\}_{k=1}^m$ and m is possibly infinite. We can then rescale the basis features. Then let $\phi(\mathbf{x}) = \mathbf{\Sigma}^{1/2} \varphi(\mathbf{x})$. We therefore have $\mathbf{\Sigma} = \mathbb{E}_{\mathbf{x} \sim \mathbb{P}}[\phi(\mathbf{x}) \otimes \phi(\mathbf{x})] = \mathrm{diag}(\xi_1, \ldots, \xi_n)$. This is the eigenfunction basis described in (Santin & Schaback, 2016).

$$
A_1 \triangleq \sum_{i \in S \cap P} \left(\left(f_{\mathbf{w}} - f_{\mathbf{w}^*} \right) \left(\mathbf{x}_i \right) \right)^2 \overset{(a)}{=} \sum_{i \in S \cap P} \left\langle \sum_{j \in X} \left(w_j - w_j^* \right) k(\mathbf{x}_j, \cdot), k(\mathbf{x}_i, \cdot) \right\rangle_{\mathcal{H}}^2 \tag{148}
$$

$$
= \sum_{i \in S \cap P} \left\langle \sum_{j \in X} \left(w_j - w_j^* \right) \phi(\mathbf{x}_j), \phi(\mathbf{x}_i) \right\rangle_{\mathcal{H}} \left\langle \phi(\mathbf{x}_i), \sum_{j \in X} \left(w_j - w_j^* \right) \phi(\mathbf{x}_j) \right\rangle_{\mathcal{H}} \tag{149}
$$

$$
= \sum_{i \in S \cap P} \left\langle \sum_{j \in X} \left(w_j - w_j^* \right) \phi(\mathbf{x}_j), \phi(\mathbf{x}_i) \otimes \phi(\mathbf{x}_i) \sum_{j \in X} \left(w_j - w_j^* \right) \phi(\mathbf{x}_j) \right\rangle_{\mathcal{H}} \tag{150}
$$

$$
= \sum_{i \in S \cap P} \left\langle \phi(\mathbf{x}) \otimes \phi(\mathbf{x}), \left(f_{\mathbf{w}} - f_{\mathbf{w}}^* \right) \otimes \left(f_{\mathbf{w}} - f_{\mathbf{w}}^* \right) \right\rangle_{\mathrm{HS}} \tag{151}
$$

$$
= \sum_{i \in S \cap P} \left\langle \mathbf{\Sigma} + \phi(\mathbf{x}) \otimes \phi(\mathbf{x}) - \mathbf{\Sigma}, \left(f_{\mathbf{w}} - f_{\mathbf{w}}^* \right) \otimes \left(f_{\mathbf{w}} - f_{\mathbf{w}}^* \right) \right\rangle_{\mathrm{HS}} \tag{152}
$$

$$
\geq \left(\tilde{n} \lambda_{\min} \left(\mathbf{\Sigma} \right) - \left\| \sum_{i \in S \cap P} \phi(\mathbf{x}) \otimes \phi(\mathbf{x}) - \mathbf{\Sigma} \right\|_{\mathrm{HS}} \right) \| f_{\mathbf{w}} - f_{\mathbf{w}}^* \|_{\mathcal{H}}^2 \tag{153}
$$

Next we will upper bound A_2,

$$
A_2 \triangleq \sum_{i \in S \cap P} \mu_i \left(f_{\mathbf{w}} - f_{\mathbf{w}}^* \right) \left(\mathbf{x}_i \right) = \sum_{i \in S \cap P} \left\langle \sum_{j \in X} (w_j - w_j^*) k(\mathbf{x}_j, \cdot), \mu_i k(\mathbf{x}_i, \cdot) \right\rangle_{\mathcal{H}} \tag{154}
$$

$$
= \left\langle \sum_{j \in X} (w_j - w_j^*) k(\mathbf{x}_j, \cdot), \sum_{i \in S \cap P} \mu_i k(\mathbf{x}_i, \cdot) \right\rangle_{\mathcal{H}} \tag{155}
$$

$$
\leq \left\| f_{\mathbf{w}} - f_{\mathbf{w}}^* \right\|_{\mathcal{H}} \left\| \sum_{i \in S \cap P} \mu_i k(\mathbf{x}_i, \cdot) \right\|_{\mathcal{H}} = \| f_{\mathbf{w}} - f_{\mathbf{w}}^* \|_{\mathcal{H}} \left\| \sum_{i \in S \cap P} \mu_i \phi(\mathbf{x}_i) \right\|_{\mathcal{H}} \tag{156}
$$

23

Then, combining our bounds, we have

$$
\lim_{\lambda \to \infty} (\Phi_\lambda(f_{\mathbf{w}}) - \Phi_\lambda(f_{\mathbf{w}}^*)) \overset{(i)}{\geq} \eta^2 \Bigg(\tilde{n}\lambda_{\min} \Big(\underset{\mathbf{x} \sim \mathbb{P}}{\mathbb{E}}[\phi(\mathbf{x}) \otimes \phi(\mathbf{x})] \Big)
$$
$$
- \Bigg\| \sum_{i \in S \cap P} \phi(\mathbf{x}) \otimes \phi(\mathbf{x}) - \mathbf{\Sigma} \Bigg\|_{\mathrm{HS}} \Bigg) - 2\eta \Bigg\| \sum_{i \in S \cap P} \mu_i \phi(\mathbf{x}_i) \Bigg\| - \sum_{j \in P \backslash S} \mu_j^2 \tag{157}
$$

In (i), we combine Equation (153) and Equation (156). This completes the proof. ∎

C.3 Proof of Theorem 13

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

$$
\|\nabla \mathsf{M}_{\Phi_\lambda, \rho}(f_{\mathbf{w}})\|_{\mathcal{H}} = \left\| \frac{1}{\rho} \left(f_{\mathbf{w}} - \underset{f_{\hat{\mathbf{w}}} \in \mathcal{K}}{\arg\min} \left(\Phi(f_{\hat{\mathbf{w}}}) + \frac{1}{2\rho} \|f_{\mathbf{w}} - f_{\hat{\mathbf{w}}}\|_{\mathcal{H}}^2 \right) \right) \right\|_{\mathcal{H}} = 0 \tag{158}
$$

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

$$
\lim_{\lambda \to \infty} (\Phi_\lambda(f_{\hat{\mathbf{w}}})) < \lim_{\lambda \to \infty} (\Phi_\lambda(f_{\tilde{\mathbf{w}}})) + \frac{1}{2\rho} \|f_{\tilde{\mathbf{w}}} - f_{\hat{\mathbf{w}}}\|_{\mathcal{H}}^2 \tag{159}
$$

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

$$
\lim_{\lambda \to \infty} (\Phi_\lambda(f_{\hat{\mathbf{w}}}) - \Phi_\lambda(f_{\mathbf{w}}^*)) \leq \frac{1}{2\rho} \|f_{\hat{\mathbf{w}}} - f_{\mathbf{w}}^*\|_{\mathcal{H}}^2 \tag{160}
$$

We proceed by proof by contradiction. Assume $\|f_{\hat{\mathbf{w}}} - f_{\mathbf{w}}^*\| > \eta$, then if $\Phi(f_{\hat{\mathbf{w}}}) - \Phi(f_{\mathbf{w}}^*) > \frac{\eta^2}{2\rho}$, then we will have $f_{\hat{\mathbf{w}}}$ is not a stationary point, which will imply $\|f_{\hat{\mathbf{w}}} - f_{\mathbf{w}}^*\|_{\mathcal{H}} \leq \eta$. Therefore, we attempt to find the minimum value for η. From Lemma 35, we have the expected distance from a stationary point of the Moreau Envelope from the optimal point over the distribution of uncorrupted datasets.

$$
\underset{\mathcal{D} \sim \hat{\mathbb{P}}}{\mathbb{E}} \lim_{\lambda \to \infty} (\Phi(f_{\mathbf{w}}) - \Phi(f_{\mathbf{w}}^*)) \geq \eta^2 \Bigg(\tilde{n}\lambda_{\min} \Big(\underset{\mathbf{x} \sim \mathbb{P}}{\mathbb{E}}[\phi(\mathbf{x}) \otimes \phi(\mathbf{x})] \Big)
$$
$$
- \underset{\mathbf{x}_i \sim \mathbb{P}}{\mathbb{E}} \Bigg\| \sum_{i \in S \cap P} \phi(\mathbf{x}_i) \otimes \phi(\mathbf{x}_i) - \mathbf{\Sigma} \Bigg\|_{\mathrm{HS}} \Bigg) - 2\eta \underset{\mu_i, \mathbf{x}_i \sim \mathbb{P}}{\mathbb{E}} \Bigg\| \sum_{i \in S \cap P} \mu_i \phi(\mathbf{x}_i) \Bigg\| - \underset{\mu_j \sim \mathbb{P}}{\mathbb{E}} \sum_{j \in P \backslash S} \mu_j^2 \tag{161}
$$
$$
\overset{(i)}{\geq} \eta^2 \Big(\tilde{n}\lambda_{\min}(\mathbf{\Sigma}) - 2\sqrt{\tilde{n}} \operatorname{Tr}(\mathbf{\Sigma}) \Big) - \eta O \Big(\sigma \sqrt{\tilde{n} \log(\tilde{n}) \operatorname{Tr}(\mathbf{\Sigma})} \Big) - \sigma\gamma\tilde{n} \tag{162}
$$

In (i), we bound first two expectation terms with Lemmas 27 and 29. From the definition of stationary point, we have

$$
\eta^2 \Big(\tilde{n}\lambda_{\min}(\mathbf{\Sigma}) - 2\operatorname{Tr}(\mathbf{\Sigma})\sqrt{\tilde{n}} - \beta \Big) - \eta O \Big(\sigma \sqrt{\tilde{n} \log(\tilde{n}) \operatorname{Tr}(\mathbf{\Sigma})} \Big) - \sigma\gamma\tilde{n} \leq 0 \tag{163}
$$

Therefore, when Equation (163) does not hold, we have a contradiction. It thus follows from upper bounding the positive solution of the quadratic equation,

$$
\eta \leq (\sigma\gamma\tilde{n})^{1/2} \left(\tilde{n} \left(\lambda_{\min}(\mathbf{\Sigma}) - \frac{2\operatorname{Tr}(\mathbf{\Sigma})}{\sqrt{\tilde{n}}} \right) - \beta \right)^{-1/2} + O \Big(\sigma \sqrt{\tilde{n} \log(\tilde{n}) \operatorname{Tr}(\mathbf{\Sigma})} \Big) \left(\tilde{n} \left(\lambda_{\min}(\mathbf{\Sigma}) - \frac{2\operatorname{Tr}(\mathbf{\Sigma})}{\sqrt{\tilde{n}}} \right) - \beta \right)^{-1} \tag{164}
$$

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

$$
\eta \leq \left(\frac{\sigma\gamma\tilde{n}}{c_1\tilde{n}\lambda_{\min}(\mathbf{\Sigma})} \right)^{1/2} + \frac{O \Big(\sigma\sqrt{\log(\tilde{n})\operatorname{Tr}(\mathbf{\Sigma})} \Big)}{c_1\sqrt{\tilde{n}}\lambda_{\min}(\mathbf{\Sigma})} \tag{165}
$$

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

$$
\underset{\mathcal{D} \sim \hat{\mathbb{P}}}{\mathbb{E}} \|f_{\hat{\mathbf{w}}} - f_{\mathbf{w}}^*\|_{\mathcal{H}} \leq O \left(\sigma \sqrt{\frac{\gamma}{C\lambda_m(\mathbf{\Sigma})}} \right) \tag{166}
$$

This completes our proof. ∎

C.4 Proof of Theorem 14 *(In Progress)*

Proof. We start from Lemma 35, which gives us the following,

$$
(\Phi(f_{\mathbf{w}}) - \Phi(f_{\mathbf{w}}^*)) \geq \eta^2 \Bigg(\tilde{n}\lambda_{\min} \Big(\mathop{\mathbb{E}}_{\mathbf{x}\sim\mathbb{P}} [\phi(\mathbf{x}) \otimes \phi(\mathbf{x})] \Big)
$$
$$
- \Big\| \sum_{i\in S\cap P} \phi(\mathbf{x}_i) \otimes \phi(\mathbf{x}_i) - \mathbf{\Sigma} \Big\|_{\mathrm{HS}} \Bigg) - 2\eta \max_{i\in S\cap P} |\mu_i| \Big\| \sum_{i\in S\cap P} \phi(\mathbf{x}_i) \Big\|_{\mathcal{H}} - \sum_{j\in P\backslash S} \mu_j^2 \tag{167}
$$

Concatenate all the samples of μ_i for $i \in P$ in to a vector $\boldsymbol{\mu} \in \mathbb{R}^{(1-\epsilon)n}$ where $\boldsymbol{\mu}^+ \in \mathbb{R}^{\tilde{n}}$ denotes the samples of μ_i for $i \in S \cap P$ and $\boldsymbol{\mu}^- \in \mathbb{R}^{\gamma\tilde{n}}$ denotes the samples of μ_i for $i \in P \cap S$. We will create a parameterized event over $s \geq 1$.

$$
E_s = \Big\{ \boldsymbol{\mu} : \big\| \boldsymbol{\mu}^+ \big\|_2^2 \leq \sigma^2 \gamma\tilde{n} \cdot s \text{ and } \big\| \boldsymbol{\mu}^- \big\|_\infty \leq \sigma\sqrt{2\log(\gamma\tilde{n})} \cdot s \Big\} \tag{168}
$$

First note that $\boldsymbol{\mu} \sim \mathcal{N}(\mathbf{0}_n, \sigma^2\mathbf{I}_n)$, therefore for any subset of the indices $S \subset [\![n]\!]$, we have $\boldsymbol{\mu}_S \sim \mathcal{N}(\mathbf{0}_{|S|}, \sigma^2\mathbf{I}_{|S|})$. We then invoke Proposition 23 and Proposition 24 together and obtain $\mathbb{P}\{E_s^c\} \leq e^{-s/2} + (\sqrt{2}/\log(\gamma\tilde{n}))e^{-s^2} \leq 2.05e^{-s/2}$ for all $s \geq 1$ and assuming $\gamma\tilde{n} \geq 2$. Concatenate the samples $\mathbf{x}_i$ into a matrix $\mathbf{X} \in \mathbb{R}^{n(1-\epsilon)\times d}$. We next create a parameterized event over $u \geq 1$.

$$
E_u = \Bigg\{ \mathbf{X} : \Big\| \sum_{i\in S\cap P} \phi(\mathbf{x}_i) \otimes \phi(\mathbf{x}_i) - \mathbf{\Sigma} \Big\|_{\mathrm{HS}} \leq \tilde{n}\,\mathrm{Tr}(\mathbf{\Sigma}) \cdot u \text{ and } \Big\| \sum_{i\in S\cap P} \phi(\mathbf{x}_i) \Big\|_{\mathcal{H}} \leq \sqrt{\tilde{n}\,\mathrm{Tr}(\mathbf{\Sigma})} \cdot u \Bigg\} \tag{169}
$$

Since $\mathbf{x}_i$ are sampled i.i.d from $\hat{\mathbb{P}}$, it implies that $\hat{P}$ is mutually independent. From which it follows that all subsets of $\hat{P}$ are independent. We find $\mathbb{P}\{E_u^c\} \leq +e^{-u^2/2}$ for all $u \geq 1$.

C.5 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, $\mathbf{x}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$. To simplify notation, let us define $\tilde{n} \triangleq \tilde{n}$ to represent the absolute minimum number of uncorrupted points in the Subquantile. We then have,

$$
\mathop{\mathbb{E}}_{\mathcal{D}\sim\hat{\mathbb{P}}} \lim_{\lambda\to\infty} (\Phi_\lambda(\mathbf{w}) - \Phi_\lambda(\mathbf{w}^*)) \overset{(i)}{\geq} \eta^2 \Big(\tilde{n}\lambda_{\min}(\mathbf{\Sigma}) - \mathop{\mathbb{E}}_{\mathbf{x}_i\sim\mathbb{P}} \Big\| \sum_{i\in S\cap P} \mathbf{x}_i\mathbf{x}_i^\top - \mathbf{\Sigma} \Big\|_2 \Big) - \mathop{\mathbb{E}}_{\mu_i\sim\mathbb{P}} \Big\| \sum_{i\in S\cap P} \mu_i\mathbf{x}_i \Big\|_2 - \mathop{\mathbb{E}}_{\mu_i\sim\mathbb{P}} \sum_{i\in P\backslash S} \mu_i^2 \tag{170}
$$

$$
\overset{(ii)}{\geq} \eta^2 \Big(\tilde{n}\lambda_{\min}(\mathbf{\Sigma}) - \sqrt{\tilde{n}}\Big(2\sqrt{3}\,\mathrm{Tr}(\mathbf{\Sigma})\Big) \Big) - \eta O\Big(\sigma\sqrt{\tilde{n}\log(\tilde{n})\,\mathrm{Tr}(\mathbf{\Sigma})} \Big) - \gamma\tilde{n}\sigma^2 \tag{171}
$$

In (i), we use Lemma 35. In (ii) we bound the three expectation terms with Lemmas 22, 27 and 31. Then from a similar contradiction idea and upper bounding the quadratic, we have in expectation

$$
\eta \leq O\Big(\sigma\sqrt{\tilde{n}\log(\tilde{n})\,\mathrm{Tr}(\mathbf{\Sigma})} \Big) \Big(\tilde{n}\lambda_{\min}(\mathbf{\Sigma}) - \sqrt{\tilde{n}}\Big(2\sqrt{3}\,\mathrm{Tr}(\mathbf{\Sigma})\Big) - \beta \Big)^{-1} + \sigma\sqrt{\gamma\tilde{n}}\Big(\tilde{n}\lambda_{\min}(\mathbf{\Sigma}) - \sqrt{\tilde{n}}\Big(2\sqrt{3}\,\mathrm{Tr}(\mathbf{\Sigma})\Big) - \beta \Big)^{-1/2} \tag{172}
$$

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

$$
\eta \leq \sigma\sqrt{\frac{\gamma}{c_2\lambda_{\min}(\mathbf{\Sigma})}} + \frac{O\Big(\sigma\sqrt{\log(\tilde{n})\,\mathrm{Tr}(\mathbf{\Sigma})} \Big)}{\sqrt{\tilde{n}}c_2\lambda_{\min}(\mathbf{\Sigma})} \tag{173}
$$

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

$$
\mathop{\mathbb{E}}_{\mathcal{D}\sim\hat{\mathbb{P}}} \|\hat{\mathbf{w}} - \mathbf{w}^*\|_2 \leq O\Big(\sigma\sqrt{\frac{\gamma}{\lambda_{\min}(\mathbf{\Sigma})}} \Big) \tag{174}
$$

We obtain the same bound as in the kernelized regression case up to constant factors. This completes the proof. ∎

D Kernlized Binary Classification

In this section, we will prove error bounds for Subquantile Minimization in the Kernelized Binary Classification Problem.

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

Lemma 36. (L-Lipschitz of $g(t, \mathbf{w})$ w.r.t $\mathbf{w}$). Let $\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n$, represent the data vectors. It then follows:

$$|g(t, f_{\mathbf{w}}) - g(t, f_{\hat{\mathbf{w}}})| \leq L \|f_{\mathbf{w}} - f_{\hat{\mathbf{w}}}\|_{\mathcal{H}} \tag{175}$$

where

$$L = \frac{1}{np} \sum_{i \in X} \sqrt{k(\mathbf{x}_i, \mathbf{x}_i)} = \frac{1}{np} \operatorname{Tr}(\mathbf{K}) \tag{176}$$

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

$$\|\nabla_f g(t, f_{\mathbf{w}})\|_{\mathcal{H}} = \left\| \frac{1}{np} \sum_{i=1}^{n} \mathbb{I}\{t \geq (1 - y_i) \log(f_{\mathbf{w}}(\mathbf{x}_i))\} (y_i - \sigma(f_{\mathbf{w}}(\mathbf{x}_i))) \cdot k(\mathbf{x}_i, \cdot) \right\|_{\mathcal{H}} \tag{177}$$

$$\overset{(i)}{\leq} \frac{1}{np} \sum_{i \in S} \|(y_i - \sigma(f_{\mathbf{w}}(\mathbf{x}_i))) \cdot k(\mathbf{x}_i, \cdot)\|_{\mathcal{H}} \overset{(ii)}{\leq} \frac{1}{np} \sum_{i \in S} |y_i - \sigma(f_{\mathbf{w}}(\mathbf{x}_i))| \|k(\mathbf{x}_i, \cdot)\|_{\mathcal{H}} \overset{(iii)}{\leq} \frac{1}{np} \sum_{i=1}^{n} \sqrt{k(\mathbf{x}_i, \mathbf{x}_i)} \tag{178}$$

(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. $\blacksquare$

Lemma 37. (β-Smoothness of $g(t, \mathbf{w})$ w.r.t $\mathbf{w}$). Let $\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n$ represent the rows of the data matrix $\mathbf{X}$. It then follows:

$$\|\nabla_f g(t, f_{\mathbf{w}}) - \nabla_f g(t, f_{\hat{\mathbf{w}}})\| \leq \beta \|f_{\mathbf{w}} - f_{\hat{\mathbf{w}}}\|_{\mathcal{H}} \tag{179}$$

where

$$\beta = \frac{1}{4p} \sum_{i=1}^{n} k(x_i, x_i) = \frac{1}{4p} \operatorname{Tr}(\mathbf{K}) \tag{180}$$

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

$$\left\| \nabla_f^2 g(t, f_{\mathbf{w}}) \right\|_{\text{op}} = \frac{1}{np} \sum_{i=1}^{n} \mathbb{I}\{t \geq (1 - y_i) \log(f_{\mathbf{w}}(\mathbf{x}_i))\} \sigma(f_{\mathbf{w}}(\mathbf{x}_i)) (1 - \sigma(f_{\mathbf{w}}(\mathbf{x}_i))) \|\phi(\mathbf{x}_i) \otimes \phi(\mathbf{x}_i)\|_{\text{op}} \tag{181}$$

$$\leq \frac{1}{np} \sum_{i=1}^{n} |\sigma(f_{\mathbf{w}}(\mathbf{x}_i)) (1 - \sigma(f_{\mathbf{w}}(\mathbf{x}_i)))| \|\phi(\mathbf{x}_i)\|_{\text{op}}^2 \overset{(i)}{\leq} \frac{1}{4np} \sum_{i=1}^{n} k(x_i, x_i) = \frac{1}{4np} \operatorname{Tr}(\mathbf{K}) \tag{182}$$

(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. $\blacksquare$

Lemma 38. Assume $f_{\hat{\mathbf{w}}}$ is a first-order stationary point as defined in Definition 8. If $\|f_{\hat{\mathbf{w}}} - f_{\mathbf{w}^*}\|_{\mathcal{H}} \geq \eta$, then it follows

$$\underset{\mathcal{D} \sim \hat{\mathbb{P}}}{\mathbb{E}} \|f_{\hat{\mathbf{w}}} - f_{\mathbf{w}}^*\|_{\mathcal{H}} \leq O\left(\frac{\sqrt{\operatorname{Tr}(\mathbf{\Sigma})} + \sqrt{\gamma Q_k}}{c_4 \sqrt{\tilde{n}} \lambda_{\min}(\mathbf{\Sigma})} \right) \tag{183}$$

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

$$\langle \nabla_f g(f_{\hat{\mathbf{w}}}, t), f_{\hat{\mathbf{w}}} - f_{\mathbf{w}}^* \rangle_{\mathcal{H}} = (f_{\hat{\mathbf{w}}} - f_{\mathbf{w}}^*) \left(\sum_{i \in S} (\sigma(f_{\hat{\mathbf{w}}}(\mathbf{x}_i)) - y_i) \phi(\mathbf{x}_i) \right) \tag{184}$$

$$= (f_{\hat{\mathbf{w}}} - f_{\mathbf{w}}^*) \left(\sum_{i \in S} \left(\sigma \left(f_{\hat{\mathbf{w}}}(\mathbf{x}_i) \right) - \sigma \left(f_{\mathbf{w}}^*(\mathbf{x}_i) \right) + \sigma \left(f_{\mathbf{w}}^*(\mathbf{x}_i) \right) - y_i \right) \phi(\mathbf{x}_i) \right) \tag{185}$$

$$\overset{(i)}{\geq} (f_{\hat{\mathbf{w}}} - f_{\mathbf{w}}^*) \underbrace{\left(\sum_{S \cap P} \left(\sigma \left(f_{\hat{\mathbf{w}}}(\mathbf{x}_i) \right) - \sigma \left(f_{\mathbf{w}}^*(\mathbf{x}_i) \right) \right) \phi\left(\mathbf{x}_i \right) \right)}_{B_1} + (f_{\hat{\mathbf{w}}} - f_{\mathbf{w}}^*) \underbrace{\left(\sum_{i \in S} \left(\sigma \left(f_{\mathbf{w}}^*(\mathbf{x}_i) \right) - y_i \right) \phi\left(\mathbf{x}_i \right) \right)}_{B_2} \tag{186}$$

(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 $h(f_{\mathbf{w}}) = \sum_{i \in S \cap P} \log(1 + \exp(f_{\mathbf{w}}(\mathbf{x}_i)))$. We then have $h'(f_{\mathbf{w}}) = \sum_{i \in S \cap P} \sigma(f_{\mathbf{w}}(\mathbf{x}_i))\phi(\mathbf{x}_i)$, from which we have $h''(f_{\mathbf{w}}) = \sum_{i \in S \cap P} \sigma(f_{\mathbf{w}}(\mathbf{x}_i))(1 - \sigma(f_{\mathbf{w}}(\mathbf{x}_i)))(\phi(\mathbf{x}_i) \otimes \phi(\mathbf{x}_i))$. We can then note h is strongly convex with $\mu = \Omega(\lambda_{\min}(\sum_{i \in S \cap P} \phi(\mathbf{x}_i) \otimes \phi(\mathbf{x}_i)))$. Then from the properties of strongly convex functions, we have

$$\sum_{i \in S \cap P} (f_{\hat{\mathbf{w}}}(\mathbf{x}_i) - f_{\mathbf{w}}^*(\mathbf{x}_i)) \left(\sigma \left(f_{\hat{\mathbf{w}}}(\mathbf{x}_i) \right) - \sigma \left(f_{\mathbf{w}}^*(\mathbf{x}_i) \right) \right) \gtrsim \lambda_{\min} \left(\sum_{i \in S \cap P} \phi(\mathbf{x}_i) \otimes \phi(\mathbf{x}_i) \right) \| f_{\mathbf{w}}^* - f_{\hat{\mathbf{w}}} \|_{\mathcal{H}}^2 \tag{187}$$

Then from the Cauchy-Schwarz Inequality, we have

$$\sum_{i \in S} (f_{\mathbf{w}}^*(\mathbf{x}_i) - f_{\hat{\mathbf{w}}}(\mathbf{x}_i)) (y_i - \sigma(f_{\mathbf{w}}^*(\mathbf{x}_i))) \leq \max_{i \in S} |y_i - \sigma(f_{\mathbf{w}}^*(\mathbf{x}_i))| \left\langle \sum_{j \in X} (w_j^* - \hat{w}_j)\phi(\mathbf{x}_j), \sum_{i \in S} \phi(\mathbf{x}_i) \right\rangle \tag{188}$$

$$\leq \| f_{\mathbf{w}}^* - f_{\hat{\mathbf{w}}} \|_{\mathcal{H}} \left\| \sum_{i \in S} \phi(\mathbf{x}_i) \right\|_{\mathcal{H}} \leq \| f_{\mathbf{w}}^* - f_{\hat{\mathbf{w}}} \|_{\mathcal{H}} \left(\left\| \sum_{i \in S \cap P} \phi(\mathbf{x}_i) \right\|_{\mathcal{H}} + \left\| \sum_{i \in S \cap Q} \phi(\mathbf{x}_i) \right\|_{\mathcal{H}} \right) \tag{189}$$

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

D.2 Proof of Theorem 16

Proof. From Lemma 38, we have in expectation

$$\mathbb{E}_{\mathcal{D} \sim \hat{\mathbb{P}}} \langle \nabla_f g \left(f_{\hat{\mathbf{w}}}, t \right), f_{\mathbf{w}}^* - f_{\hat{\mathbf{w}}} \rangle_{\mathcal{H}} \geq c_3 \left(\tilde{n} \lambda_{\min} (\boldsymbol{\Sigma}) - \mathbb{E}_{\mathbf{x}_i \sim \mathbb{P}} \left\| \sum_{i \in S \cap P} \phi(\mathbf{x}_i) \otimes \phi(\mathbf{x}_i) - \boldsymbol{\Sigma} \right\| \right) \| f_{\hat{\mathbf{w}}} - f_{\mathbf{w}}^* \|_{\mathcal{H}}^2$$
$$- \left(\sqrt{\tilde{n} \operatorname{Tr}(\boldsymbol{\Sigma})} + \sqrt{\gamma \tilde{n} Q_k} \right) \| f_{\hat{\mathbf{w}}} - f_{\mathbf{w}}^* \|_{\mathcal{H}} \tag{190}$$

We will lower bound the constant we introduced in Equation (187) and call it c_3, recall for $f \in \mathcal{K}$, we have $\|f\|_{\mathcal{H}} \leq R$ and $P_k \triangleq \max_{i \in P} k(x_i, x_i)$.

$$\mathbb{E}_{\mathcal{D} \sim \hat{\mathbb{P}}} c_3 \overset{(187)}{=} \mathbb{E}_{\mathcal{D} \sim \hat{\mathbb{P}}} \min_{i \in S \cap P} (1 - \sigma(f_{\hat{\mathbf{w}}}(\mathbf{x}_i))) \sigma(f_{\hat{\mathbf{w}}}(\mathbf{x}_i)) \tag{191}$$

$$\geq \mathbb{E}_{\mathcal{D} \sim \hat{\mathbb{P}}} \left(1 - \sigma(R \max_{i \in P} k(\mathbf{x}_i, \mathbf{x}_i)) \right) \sigma(R \max_{i \in P} k(\mathbf{x}_i, \mathbf{x}_i)) \tag{192}$$

$$\geq \left(\frac{1}{2} \right) \mathbb{E}_{\mathcal{D} \sim \hat{\mathbb{P}}} \sigma \left(-R \max_{i \in P} k(\mathbf{x}_i, \mathbf{x}_i) \right) \overset{(i)}{\gtrsim} \left(\frac{1}{2} \right) \exp \left(-R \mathbb{E}_{\mathcal{D} \sim \hat{\mathbb{P}}} \left[\max_{i \in P} k(\mathbf{x}_i, \mathbf{x}_i) \right] \right) \tag{193}$$

$$\overset{(ii)}{\geq} \left(\frac{1}{2} \right) \exp \left(-R \left(2 \operatorname{Tr}(\boldsymbol{\Sigma}) \log \left(\frac{nR}{\operatorname{Tr}(\boldsymbol{\Sigma})} \right) + \frac{1}{R} \right) \right) \tag{194}$$

(i) follows from Jensen's Inequality as $\exp(-x)$ is a convex functon. The bound in (ii) follows from an application of Lemma 26 and choosing $s = R$. Then we have from the definition of a stationary point, $\nabla_f g \left(f_{\hat{\mathbf{w}}}, t \right) \left(f_{\hat{\mathbf{w}}} - f_{\mathbf{w}}^* \right) \leq 0$ when

$$\mathbb{E}_{\mathcal{D} \sim \hat{\mathbb{P}}} \| f_{\hat{\mathbf{w}}} - f_{\mathbf{w}}^* \|_{\mathcal{H}} \leq O \left(\frac{e \left(\sqrt{\tilde{n} \operatorname{Tr}(\boldsymbol{\Sigma})} + \sqrt{\gamma \tilde{n} Q_k} \right)}{\exp \left(-R \left(2 \operatorname{Tr}(\boldsymbol{\Sigma}) \log \left(\frac{nR}{\operatorname{Tr}(\boldsymbol{\Sigma})} \right) \right) \right) \left(\tilde{n} \lambda_{\min}(\boldsymbol{\Sigma}) - 2\sqrt{\tilde{n}} \operatorname{Tr}(\boldsymbol{\Sigma}) \right)} \right) \tag{195}$$

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

$$
\underset{\mathcal{D}\sim\hat{\mathbb{P}}}{\mathbb{E}}\|f_{\hat{\mathbf{w}}} - f_{\mathbf{w}}^*\|_{\mathcal{H}} \leq O\left(\frac{\sqrt{\text{Tr}(\boldsymbol{\Sigma})} + \sqrt{Q_k}}{\sqrt{\tilde{n}}\exp\left(-2R\,\text{Tr}(\boldsymbol{\Sigma})\log\left(\frac{nR}{\text{Tr}(\boldsymbol{\Sigma})}\right)\lambda_{\min}(\boldsymbol{\Sigma})\right)}\right) \tag{196}
$$

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 39. *Let* $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n \sim \hat{\mathbb{P}}$. *It then follows for a* $f_{\mathbf{w}} \in \mathcal{K}$, *then* $g(t, f_{\mathbf{w}})$ *is* L-*Lipschitz and* β-*Smooth for constants* $L = \frac{1}{np}\sum_{i=1}^n \sqrt{k(\mathbf{x}_i, \mathbf{x}_i)}$ *and* $\beta = \frac{1}{np}\text{Tr}(\mathbf{K})$.

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.

$$
\frac{\partial}{\partial \mathbf{w}_k}\left(\ell\left(\mathbf{x}_i, y_i; f_{\mathbf{W}}\right)\right) = \begin{cases} -\phi(\mathbf{x}_i)\text{softmax}(f_{\mathbf{W}}(\mathbf{x}_i))_k & \text{if } k = y_i \\ \phi(\mathbf{x}_i)\left(1 - \text{softmax}(f_{\mathbf{W}}(\mathbf{x}_i))_k\right) & \text{if } k \neq y_i \end{cases} \tag{197}
$$

Our proof then follows similarly to the proof for Lemma 36. We utilize $\odot$ to denote entry wise multiplication, i.e $\mathbf{x} \odot \mathbf{y}$ indicates $\mathbf{y}$ is multiplied to each element of $\mathbf{x}$ (not commutative).

$$
\|\nabla_f g(t, f_{\mathbf{W}})\|_{\mathcal{H}} = \left\|\frac{1}{np}\sum_{i=1}^n \mathbb{I}\left\{-\log\left(\text{softmax}\left(f_{\mathbf{W}}(\mathbf{x}_i)\right)_{y_i}\right) \geq t\right\}\left(\text{softmax}\left(f_{\mathbf{W}}(\mathbf{x}_i)\right) - y_i\right) \odot k(\mathbf{x}_i, \cdot)\right\| \tag{198}
$$

$$
\leq \frac{1}{np}\sum_{i=1}^n \|(\text{softmax}\left(f_{\mathbf{W}}(\mathbf{x}_i)\right) - y_i) \odot k(\mathbf{x}_i, \cdot)\| \leq \frac{1}{np}\sum_{i=1}^n \|k(\mathbf{x}_i, \cdot)\|_{\mathcal{H}} = \frac{1}{np}\sum_{i=1}^n \sqrt{k(\mathbf{x}_i, \mathbf{x}_i)} \tag{199}
$$

This gives the L-Lipschitz Constant.

We upper bound the operator norm of the Hessian to find the β-smoothness constant.

$$
\left\|\nabla_f^2 g\left(t, f_{\mathbf{W}}\right)\right\|_{\text{op}} \tag{200}
$$

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

$$
\leq \frac{1}{np}\sum_{i=1}^n \left\|\left(\text{diag}(\text{softmax}(f_{\mathbf{W}}(\mathbf{x}_i)) - \text{softmax}(f_{\mathbf{W}}(\mathbf{x}_i)\text{softmax}(f_{\mathbf{W}}(\mathbf{x}_i)^\top) \odot (\phi(\mathbf{x}_i) \otimes \phi(\mathbf{x}_i))\right)\right\|_{\text{op}} \tag{201}
$$

$$
\leq \frac{1}{np}\sum_{i=1}^n \left\|\text{diag}(\text{softmax}(f_{\mathbf{W}}(\mathbf{x}_i)) - \text{softmax}(f_{\mathbf{W}}(\mathbf{x}_i)\text{softmax}(f_{\mathbf{W}}(\mathbf{x}_i)^\top\right\|_{\text{op}} \|\phi(\mathbf{x}_i)\|_{\mathcal{H}}^2 \tag{202}
$$

$$
\leq \frac{1}{np}\sum_{i=1}^n k(\mathbf{x}_i, \mathbf{x}_i) = \frac{1}{np}\text{Tr}(\mathbf{K}) \tag{203}
$$

This gives the β-Smoothness Constant. ∎

Lemma 40. *Assume* $f_{\hat{\mathbf{w}}}$ *is a first-order stationary point as defined in Definition 8. If* $\|f_{\hat{\mathbf{w}}} - f_{\mathbf{w}^*}\|_{\mathcal{H}} \geq \eta$, *then it follows*

$$
\lim_{\lambda\to\infty}\left\langle\nabla_f \tilde{g}_\lambda\left(f_{\hat{\mathbf{w}}}, \hat{t}\right), f_{\hat{\mathbf{w}}} - f_{\mathbf{W}}^*\right\rangle_{\mathcal{H}} \gtrsim \lambda_{\min}\left(\sum_{i\in S\cap P}\phi(\mathbf{x}_i) \otimes \phi(\mathbf{x}_i)\right)\eta^2 + \eta\left(\left\|\sum_{i\in S\cap P}\phi(\mathbf{x}_i)\right\|_{\mathcal{H}} + \left\|\sum_{i\in S\cap Q}\phi(\mathbf{x}_i)\right\|_{\mathcal{H}}\right) \tag{204}
$$

[1] **Proof.** We follow the similar set up as in Lemma 38. We have $f_{\hat{\mathbf{W}}}$ is a stationary point, i.e. $\mathbf{0} \in \partial \Phi(f_{\hat{\mathbf{W}}})$. This implies for all $f_{\mathbf{W}} \in \mathcal{K}$, we have $\Phi(f_{\hat{\mathbf{W}}}) \leq \Phi(f_{\mathbf{W}})$. As Φ is differentiable, we have the first-order stationary condition, which is $\nabla \Phi(f_{\hat{\mathbf{W}}})(f_{\hat{\mathbf{W}}} - f_{\mathbf{W}}) \leq 0$ or for all $\mathbf{W} \in \mathcal{K}$. We assume $f_{\mathbf{W}}^* \in \mathcal{K}$. Let S be the Subquantile set for $f_{\hat{\mathbf{W}}}$. We will proceed by contradiction, assume $\left\| f_{\hat{\mathbf{W}}} - f_{\mathbf{W}}^* \right\|_{\mathcal{H}} \geq \eta$. Then, we have

$$\left\langle \nabla_f g\left(f_{\hat{\mathbf{W}}}, t\right), f_{\hat{\mathbf{W}}} - f_{\mathbf{W}}^* \right\rangle_{\mathcal{H}} = \left\langle f_{\hat{\mathbf{W}}} - f_{\mathbf{W}}^*, \sum_{i \in S} \left(\text{softmax}\left(f_{\hat{\mathbf{W}}}(\mathbf{x}_i)\right) - y_i\right) \odot k(\mathbf{x}_i, \cdot) \right\rangle_{\mathcal{H}} \tag{205}$$

$$= \underbrace{\left\langle f_{\hat{\mathbf{W}}} - f_{\mathbf{W}}^*, \sum_{i \in S} \left(\text{softmax}\left(f_{\hat{\mathbf{W}}}(\mathbf{x}_i)\right) - \text{softmax}\left(f_{\mathbf{W}}^*(\mathbf{x}_i)\right)\right) \odot k(\mathbf{x}_i, \cdot) \right\rangle_{\mathcal{H}}}_{C_1} + \underbrace{\left\langle f_{\hat{\mathbf{W}}} - f_{\mathbf{W}}^*, \sum_{i \in S} \left(\text{softmax}\left(f_{\mathbf{W}}^*(\mathbf{x}_i)\right) - y_i\right) \odot k(\mathbf{x}_i, \cdot) \right\rangle_{\mathcal{H}}}_{C_2}$$

Let us now consider the function $h : \mathcal{H} \times \cdots \times \mathcal{H} :\to \mathbb{R}^n$ defined as $h(f_{\mathbf{W}}) = \sum_{i \in P \cap S} \log(\sum_{j=1}^{|\mathcal{Y}|} \exp(f_{\mathbf{w}_j}(\mathbf{x}_i) - y_{i,j}))$. We then have $\nabla h(f_{\mathbf{W}}) = \sum_{i \in P \cap S} \text{softmax}(f_{\mathbf{W}}(\mathbf{x}_i) - \mathbf{y}_i) \odot \phi(\mathbf{x}_i)$. From which it follows $\nabla^2 h(\mathbf{x}_i) = \sum_{i \in P \cap S} (\text{diag}(\text{softmax}(f_{\mathbf{W}}(\mathbf{x}_i))) - \text{softmax}(f_{\mathbf{W}}(\mathbf{x}_i))\text{softmax}(f_{\mathbf{W}}(\mathbf{x}_i))^{\top}) \odot (\phi(\mathbf{x}_i) \otimes \phi(\mathbf{x}_i))$. This function is not strictly convex, as corroborated in (Gao & Pavel, 2017). If the softmax returns the same value for all inputs, then we note $\alpha^2 \mathbf{1}^{\top} \nabla^2 h(f_{\mathbf{W}}) \mathbf{1} = 0$ where $\alpha = \frac{1}{|\mathcal{Y}|}$. Since we assume random initialization of the functions in $f_{\mathbf{W}}$, this event does not occur almost surely. Therefore, we have the function is strictly convex over the domain.

$$(205) \text{ LHS} \geq \Omega\left(\lambda_{\min}\left(\sum_{i \in S} \phi(\mathbf{x}_i) \otimes \phi(\mathbf{x}_i)\right) \left\| f_{\hat{\mathbf{W}}} - f_{\mathbf{W}}^* \right\|_{\mathcal{H}}^2\right) + \left\langle f_{\hat{\mathbf{W}}} - f_{\mathbf{W}}^*, \sum_{i \in S} \left(\text{softmax}\left(f_{\mathbf{W}}^*(\mathbf{x}_i)\right) - \mathbf{y}_i\right) \odot k(\mathbf{x}_i, \cdot) \right\rangle_{\mathcal{H}} \tag{206}$$

$$\overset{(i)}{\geq} \Omega\left(\lambda_{\min}\left(\sum_{i \in S \cap P} \phi(\mathbf{x}_i) \otimes \phi(\mathbf{x}_i)\right) \left\| f_{\hat{\mathbf{W}}} - f_{\mathbf{W}}^* \right\|_{\mathcal{H}}^2\right) + \left\langle f_{\hat{\mathbf{W}}} - f_{\mathbf{W}}^*, \sum_{i \in S} \left(\text{softmax}\left(f_{\mathbf{W}}^*(\mathbf{x}_i)\right) - \mathbf{y}_i\right) \odot k(\mathbf{x}_i, \cdot) \right\rangle_{\mathcal{H}} \tag{207}$$

Where (i) follows from Weyl's Inequality (Weyl, 1912). Now we will upper bound C_2.

$$C_2 = \sum_{i \in S} \left\langle f_{\hat{\mathbf{W}}}(\mathbf{x}_i) - f_{\mathbf{W}}^*(\mathbf{x}_i), \text{softmax}\left(f_{\mathbf{W}}^*(\mathbf{x}_i)\right) - \mathbf{y}_i \right\rangle \tag{208}$$

$$\overset{(i)}{\leq} \sum_{i \in S} \sqrt{\sum_{j \in \mathcal{Y}} \left(f_{\hat{\mathbf{w}}_j}(\mathbf{x}_i) - f_{\mathbf{w}_j}^*(\mathbf{x}_i)\right)^2} \sqrt{\sum_{j \in \mathcal{Y}} \left(\text{softmax}\left(f_{\mathbf{w}_j}^*(\mathbf{x}_i)\right) - y_{ij}\right)^2} \tag{209}$$

$$\leq 2 \sum_{i \in S} \sum_{j \in \mathcal{Y}} \left(f_{\hat{\mathbf{w}}_j}(\mathbf{x}_i) - f_{\mathbf{w}_j}^*(\mathbf{x}_i)\right) \leq 2 \left\| f_{\hat{\mathbf{W}}} - f_{\mathbf{W}}^* \right\|_{\mathcal{H}} \left\| \sum_{i \in S} \phi(\mathbf{x}_i) \right\| \tag{210}$$

$$\leq 2 \left\| f_{\hat{\mathbf{W}}} - f_{\mathbf{W}}^* \right\|_{\mathcal{H}} \left(\left\| \sum_{i \in S \cap P} \phi(\mathbf{x}_i) \right\|_{\mathcal{H}} + \left\| \sum_{i \in S \cap Q} \phi(\mathbf{x}_i) \right\|_{\mathcal{H}}\right) \tag{211}$$

where (i) follows from Hölder's Inequality (Hölder, 1889). Substituting Equations (206) and (211) into Equation (205) completes the proof. ∎

E.2 Proof of Theorem 17

[2]From Lemma 40, we have in expectation over the dataset sampling,

$$\underset{\mathcal{D} \sim \hat{\mathbb{P}}}{\mathbb{E}} \lim_{\lambda \to \infty} \left\langle \nabla_f \tilde{g}\left(f_{\hat{\mathbf{W}}}, \hat{t}\right), f_{\hat{\mathbf{W}}} - f_{\mathbf{W}}^* \right\rangle_{\mathcal{H}} \geq \Omega\left(\lambda_{\min}\left(\sum_{i \in S \cap P} \phi(\mathbf{x}_i) \otimes \phi(\mathbf{x}_i)\right) \left\| f_{\hat{\mathbf{W}}} - f_{\mathbf{W}}^* \right\|_{\mathcal{H}}^2\right)$$

$$- \left\| f_{\hat{\mathbf{W}}} - f_{\mathbf{W}}^* \right\|_{\mathcal{H}} \underset{\mathcal{D} \sim \hat{\mathbb{P}}}{\mathbb{E}}\left[\left(\left\| \sum_{i \in S \cap P} \phi(\mathbf{x}_i) \right\|_{\mathcal{H}} + \left\| \sum_{i \in S \cap Q} \phi(\mathbf{x}_i) \right\|_{\mathcal{H}}\right)\right] \tag{212}$$

---

[1]In Progress

[2]In Progress

$$\overset{(i)}{\geq} \Omega\left(\left(\lambda_{\min}\left(\boldsymbol{\Sigma}\right) - \underset{\mathcal{D}\sim\hat{\mathbb{P}}}{\mathbb{E}} \left\|\sum_{i\in S\cap P} \phi(\mathbf{x}_i)\otimes\phi(\mathbf{x}_i) - \boldsymbol{\Sigma}\right\|_{\mathrm{HS}}\right) \left\|f_{\hat{\mathbf{W}}} - f_{\mathbf{W}}^*\right\|_{\mathcal{H}}^2\right)$$

$$- \left\|f_{\hat{\mathbf{W}}} - f_{\mathbf{W}}^*\right\|_{\mathcal{H}} \sqrt{\mathbb{E}\left[\left\|\sum_{i\in S\cap P}\phi(\mathbf{x}_i)\right\|^2\right]} + \sqrt{\gamma\tilde{n}Q_k} \tag{213}$$

$$\overset{(ii)}{\geq} \Omega\left(\tilde{n}\lambda_{\min}(\boldsymbol{\Sigma}) - \sqrt{\tilde{n}}\operatorname{Tr}(\boldsymbol{\Sigma})\right) \left\|f_{\hat{\mathbf{W}}} - f_{\mathbf{W}}^*\right\|_{\mathcal{H}}^2 - \left\|f_{\hat{\mathbf{W}}} - f_{\mathbf{W}}^*\right\|_{\mathcal{H}} \left(\sqrt{\tilde{n}}\operatorname{Tr}(\boldsymbol{\Sigma}) + \sqrt{\gamma\tilde{n}Q_k}\right) \tag{214}$$

(i) follows from an application of Jensen's Inequality as $(\cdot)^2$ is a convex function. (ii) follows from Lemma 29. Note, if $f_{\hat{\mathbf{W}}}$ is a stationary point, then we must have $\langle\nabla_f\tilde{g}\left(f_{\hat{\mathbf{W}}}, \hat{t}\right), f_{\hat{\mathbf{W}}} - f_{\mathbf{W}}^*\rangle_{\mathcal{H}} \leq 0$. To satisfy this inequality in expectation over the datasets, we must have

$$\underset{\mathcal{D}\sim\hat{\mathbb{P}}}{\mathbb{E}}\left\|f_{\hat{\mathbf{W}}} - f_{\mathbf{W}}^*\right\| \overset{(214)}{\leq} O\left(\frac{\sqrt{\tilde{n}}\operatorname{Tr}(\boldsymbol{\Sigma}) + \sqrt{\gamma\tilde{n}Q_k}}{\tilde{n}\lambda_{\min}(\boldsymbol{\Sigma}) - \sqrt{\tilde{n}}\operatorname{Tr}(\boldsymbol{\Sigma})}\right) \tag{215}$$

This completes the proof. $\blacksquare$