Kernel Learning in the Huber ϵ-Contamination Model

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

Alex Gittens
RPI CS, gittea@rpi.edu

Abstract

In this paper we study Subquantile Minimization for learning the Huber-ϵ Contamination Problem for Kernel Learning. We assume the adversary has knowledge of the true distribution of $\mathcal{P}$, and is able to corrupt the covariates and the labels of ϵn samples for $\epsilon \in [0, 0.5)$. The distribution is formed as $\hat{\mathcal{P}} = (1-\epsilon)\mathcal{P} + \epsilon\mathcal{Q}$, and we want to learn the function $f^* \triangleq \min_{f \in \mathcal{H}} \mathbb{E}_{\mathcal{D}\sim\mathcal{P}} [\mathcal{R}(f; \mathcal{D})]$, from the noisy distribution, $\hat{\mathcal{P}}$. Superquantile objectives have been been studied extensively to reduce the risk of the tail [LPMH21, RRM14]. We consider the contrasting case where we want to minimize the body of the risk. We study a gradient-descent approach to solve a variational representation of the Subquantile Objective. Our main algorithmic tool is the *ridge*, which allows us to give a near optimal approximation bound in kernelized ridge regression and kernelized binary cross entropy.

1 Introduction

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

Definition 1 (Huber ϵ-Contamination Model [HR09])**.** *Given a corruption parameter $0 < \epsilon < 0.5$, a data matrix, X and labels $\mathbf{y}$. An adversary is allowed to inspect all samples and modify ϵn samples arbitrarily. The algorithm is then given the ϵ-corrupted data matrix X and ϵ-corupted labels vector $\mathbf{y}$ as training data.*

Current approaches for robust learning across various machine learning tasks often use gradient descent over a robust objective, [LBSS21]. These robust objectives tend to not be convex and therefore do not have a strong analysis on the error bounds for general classes of models.

We similarly propose a robust objective which has a nonconvex-concave objective. This objective function has also been proposed recently in [HYwL20] where there has been an analysis in the Binary Classification Task. We show Subquantile Minimization reduces to the same objective function given in [HYwL20].

The study of Kernel Learning in the Gaussian Design is quite popular, [CLKZ21, Dic16]. In [CLKZ21], the feature space, $\phi(\mathbf{x}_i) \sim \mathcal{N}(0, \Sigma)$ where Σ is a diagonal matrix of dimension p, where p can be infinite. We will now give our formal definition of the dataset.

Definition 2 (Corruption Model)**.** *Let $\mathcal{P}$ be a distribution over $\mathbb{R}^d$ such that $\mathcal{P}_\sharp \phi$ is a centered distribution in the Hilbert Space $\mathcal{H}$ with trace-class covariance operator Σ and trace-class sub-Gaussian proxy Γ such that $\Sigma \preceq c\Gamma$. The original dataset is denoted as $\hat{P}$, the adversary is able to observe $\hat{P}$ and arbitrarily corrupts $n\epsilon$ samples denoted as Q such that $|Q| = n\epsilon$. The remaining uncorrupted samples are denoted as P such that $|P| = n(1-\epsilon)$. Together $X \triangleq P \cup Q$ represents the given dataset.*

Theorem 3. *(Informal). Let the dataset be given as $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$ such that the labels and covariates of ϵn samples arbitrarily corrupted by an adversary. Then in polynomial number of iterations we obtain the following approximation bounds.*
Kernelized Regression:

$$\left\|f^{(t+1)} - f^*\right\|_{\mathcal{H}} \leq \varepsilon + O\left(\left((Q_k \vee \|\Gamma\|_{\mathrm{op}}) \|\Gamma\|_{\mathrm{op}}\right)^{-1}\left(\sigma + \frac{\sigma}{\sqrt{n(1-\epsilon)}} + \frac{C\|f^*\|_{\mathcal{H}}}{n(1-\epsilon)}\right)\right)$$

Kernel Binary Classification:

$$\left\|f^{(T)} - f^*\right\|_{\mathcal{H}} \leq \varepsilon + O\left(\sqrt{\frac{\mathcal{E}_{\mathrm{OPT}}}{n(1-\epsilon)Q_m}}\right) + O\left(\frac{\sqrt{Q_k}\,\|f^*\|_{\mathcal{H}}}{\sqrt{Q_m}n(1-\epsilon)}\right)$$

1.1 Contributions

Our main contribution is the approximation bounds for Subquantile Minimization in kernelized ridge regression and kernelized binary classification with binary cross entropy loss described in Algorithms **??** and **??**, respectively. Our proof techniques extend [BJK15, ADKS22] as we do not assume the covariates follow the spherical Gaussian property, as such a property will not hold for any infinite-dimensional Hilbert Space.

2 Preliminaries

Notation. We denote $[T]$ as the set $\{1, 2, \dots, T\}$. We define $(x)^+ \triangleq \max(0, x)$ as the Recitificied Linear Unit (ReLU) function. We say $y = O(x)$ if there exists x_0 s.t. for all $x \geq x_0$ there exists C s.t. $y \leq Cx$. We denote $\tilde{O}$ to ignore log factors. We say $y = \Omega(x)$ if there exists x_0 s.t. for all $x \geq x_0$ there exists C s.t. $y \geq Cx$. We denote $a \vee b \triangleq \max(a, b)$ and $a \wedge b \triangleq \min(a, b)$. We define $\mathbb{S}^{d-1}$ as the sphere $\{\mathbf{x} \in \mathbb{R}^d : \|\mathbf{x}\| = 1\}$.

2.1 Reproducing Kernel Hilbert Spaces

Let the function $\phi : \mathbb{R}^d \to \mathcal{H}$ represent the Hilbert Space Representation or 'feature transform' from a vector in the original covariate space to the RKHS. We define $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ as $k(\mathbf{x}, \mathbf{x}) \triangleq \langle \phi(\mathbf{x}), \phi(\mathbf{x}) \rangle_{\mathcal{H}}$. For a function in a RKHS, $f \in \mathcal{H}$, it follows for a function f parameterized by weights $\mathbf{w} \in \mathbb{R}^n$, that the point evaluation function is given as $f : \mathbb{R}^d \to \mathbb{R}$ and defined $f(\cdot) \triangleq \sum_{i \in [n]} w_i k(\mathbf{x}_i, \cdot)$.

$$f(\mathbf{x}) = \langle f, k(\mathbf{x}, \cdot) \rangle_{\mathcal{H}} = \langle f, \phi(\mathbf{x}) \rangle_{\mathcal{H}}, \quad \forall f \in \mathcal{H}$$

The norm is given as $\|f\|_{\mathcal{H}} = \sqrt{\langle f, f \rangle_{\mathcal{H}}}$.

2.2 Tensor Products

Let $\mathcal{H}, \mathcal{K}$ be Hilbert Spaces, then $\mathcal{H} \otimes \mathcal{K}$ is the tensor product space and is also a Hilbert Space [RaR02]. For $\phi_1, \psi_1 \in \mathcal{H}$ and $\phi_2, \psi_2 \in \mathcal{K}$, the inner product is defined as $\langle \phi_1 \otimes \phi_2, \psi_1 \otimes \psi_2 \rangle_{\mathcal{H} \otimes \mathcal{K}} = \langle \phi_1, \psi_1 \rangle_{\mathcal{H}} \langle \phi_2, \psi_2 \rangle_{\mathcal{K}}$. We will utilize tensor products when we discuss infinite dimensional covariance estimation.

2.3 Sub-Gaussian Random Functions in the Hilbert Space

In this paper we sample the target covariates $\mathbf{x} \sim \mathcal{X}$ such that $\phi(\mathbf{x}) \triangleq X \sim \mathcal{P}_\sharp \phi$ is sub-Gaussian in the Hilbert Space where $\mathbf{E}[X] = \mathbf{0}$ and covariance $\mathbf{E}[X \otimes X] = \Sigma$ with proxy Γ, where $\Sigma \preceq 4 \|X\|_{\psi_2}^2 \Gamma$, where we denote $\preceq$ as the Löwner order. We have X is a centered Hilbert Space sub-Gaussian random function if for all $\theta > 0$,

$$\mathop{\mathbf{E}}_{X \sim \mathcal{P}} \left[\exp\left(\theta \langle X, v \rangle_{\mathcal{H}}\right)\right] \leq \exp\left(\frac{\alpha^2 \theta^2 \langle v, \Gamma v \rangle_{\mathcal{H}}}{2}\right) \tag{1}$$

where the sub-Gaussian Norm for a centered Hilbert Space Function is given as

$$\|X\|_{\psi_2} \triangleq \inf\left\{\alpha \geq 0 : \mathbf{E}\left[e^{\langle v, X \rangle_{\mathcal{H}}}\right] \leq e^{\alpha^2 \langle v, \Gamma v \rangle_{\mathcal{H}}/2} \ : \ \forall v \in \mathcal{H}\right\}$$

Then we say $X \sim \mathcal{SG}(\Gamma, \alpha)$, where if $\alpha = 1$, we will say $X \sim \mathcal{SG}(\Gamma)$. The Gaussian Design for the Feature Space has gained popularity in the study of kernel learning [CLKZ21].

2.4 Assumptions

We will first give our assumptions for robust kernelized regression.

Assumption 4 (Sub-Gaussian Design)**.** *We assume for $\mathbf{x}_i \sim \mathcal{X}$, then it follows for the function to the Hilbert Space, $\phi(\cdot) : \mathcal{X} \to \mathcal{H}$,*

$$\phi(\mathbf{x}) \triangleq X \sim \mathcal{P}_\sharp \phi \triangleq \mathcal{SG}\left(\Gamma, 1/2\right)$$

where Γ is a possibly infinite dimensional covariance operator.

Assumption 5 (Bounded Functions)**.** *We assume for $\mathbf{x}_i \sim \mathcal{P} \in \mathcal{X}$, then it follows for the feature map, $\phi(\cdot) : \mathcal{X} \to \mathcal{H}$,*

$$\sup_{\mathbf{x} \in \mathcal{X}} \|\phi(\mathbf{x})\|_{\mathcal{H}}^2 \leq P_k < \infty$$

where $\mathcal{H}$ is a Reproducing Kernel Hilbert Space.

Assumption 6 (Normal Residuals)**.** *Let $\inf_{f \in \mathcal{H}} \mathbf{E}_{(\mathbf{x}, y) \sim \mathcal{D}} [\mathcal{R}(f; \mathbf{x}, y)]$. The residual is defined as $\mu_i \triangleq f^*(\mathbf{x}_i) - y_i$. Then we assume for some $\sigma > 0$, it follows*

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

2.5 Related Work

The idea of iterative thresholding algorithms for robust learning tasks dates back to 1806 by Legendre [Leg06]. Iterative thresholding have been studied theoretically and tested empirically in various machine learning domains [HYW+23, MGJK19]. Therefore, we will dedicate this subsection to reviewing such works and to make clear our contributions to the iterative thresholding literature.

[BJK15] study iterative thresholding for least squares regression / sparse recovery. In particular, one part of their study is of a gradient descent algorithm when the data $\mathcal{P} = \mathcal{Q} = \mathcal{N}(\mathbf{0}, \mathrm{I})$ or multivariate sub-Gaussian with proxy I. Their approximation bounds relies on the fact that $\lambda_{\min}(\Sigma) = \lambda_{\max}(\Sigma)$ and with sufficiently large data and sufficiently small ϵ, $\lambda_{\max}(\mathrm{X})/\lambda_{\min}(\mathrm{X}) \searrow 1$. This is similar to the study by [ADKS22], where the iterative trimmed maximum likelihood estimator is studied for General Linear Models. The algorithm studied by [ADKS22] utilizes a filtering algorithm with the sketching matrix $\Sigma^{-1/2}$ so the columns of X are sampled from a multivariate sub-Gaussian Distribution with proxy I before running the iterative thresholding procedure. This 'whitening' procedure to decrease the conditioning number of the covariates is also done in recent work, [SBRJ19, BJKK17].

Conditioning covariates does not generalize to kernel learning where we are given a matrix K which is equivalent to inner product of the quasimatrix[1], Φ, with itself. In the infinite dimensional case, it is not possible to sketch the kernel matrix [W+14] in order to have the original covariates be well-conditioned. In the finite dimensional case, the feature maps can be quite large and it is very difficult to obtain in practice. Thus, we are left with Φ where the columns are sampled from a sub-Gaussian Distribution with proxy Γ is a trace-class operator, which implies the eigenvalues tend to zero, i.e. $\lambda_{\inf}(\Gamma) = 0$, and there is no longer a notion of $\lambda_{\min}(\Gamma)$.

3 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.

$$\mathcal{Q}_p(U) \leq t \iff \mathbf{Pr}\left\{U \leq t\right\} \geq p$$

The p-subquantile of the risk is then given by

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

Given an objective function, $\mathcal{R}$, the kernelized learning poblem becomes:

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

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 [RHL+20]. 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

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

$$f^{(k+1)} = \mathrm{Proj}_{\mathcal{K}}\left[f^{(k)} - \eta\nabla_f g(f^{(k)}, t^{(k+1)})\right]$$

The proof of convergence for the above algorithm was given in [JNJ20][Theorem 35]. The sufficient condition for convergence is $g(f,t)$ is concave with respect to t, which for the subquantile objective is simple to show.

[1] A quasimatrix is an infinite-dimensional analogue of a tall-skinny matrix that represents an ordered set of functions in ℓ_2 (see e.g. [TT15]).

3.1 Reduction to Iterative Thresholding

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 Section 3.

Lemma 7. *Let $\mathcal{R} : \mathcal{H} \times \mathbb{R} \to \mathbb{R}$ be a loss function (not necesarily convex). Let $\mathbf{x}_{\lfloor i \rfloor}$ represent the point with the i-th smallest loss w.r.t $\mathcal{R}$. If we denote $\hat{\nu}_i \triangleq \mathcal{R}(f; \mathbf{x}_{\lfloor i \rfloor}, y_{\lfloor i \rfloor})$, it then follows $\hat{\nu}_{n(1-\epsilon)} \in \arg\max_{t \in \mathbb{R}} g(t, f)$.*

Proof. First we can note, the max value of t for g is equivalent to the min value of t for the convex w.r.t t function $-g$. We can now find the Fermat Optimality Conditions for g.

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

We observe when setting $t = \hat{\nu}_{n(1-\epsilon)}$, it follows that $0 \in \partial(-g(t,f))$. This is equivalent to the $(1-\epsilon)$-quantile of the Empirical Risk. ■

From Lemma 7, we see that t will be greater than or equal to the errors of exactly $n(1-\epsilon)$ points. Thus, we are continuously updating over the $n(1-\epsilon)$ minimum errors.

Lemma 8. *Let $\hat{\nu}_i \triangleq \mathcal{R}(f; \mathbf{x}_{\lfloor i \rfloor}, y_{\lfloor i \rfloor})$, if we choose $t^{(k+1)} = \hat{\nu}_{n(1-\epsilon)}$ as by Lemma 7, it then follows $\nabla_f g(t^{(k)}, f^{(k)}) = \frac{1}{n(1-\epsilon)}\sum_{i=1}^{n(1-\epsilon)} \nabla_f \mathcal{R}(f^{(k)}; \mathbf{x}_{\lfloor i \rfloor}, y_{\lfloor i \rfloor})$.*

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

$$\begin{aligned}
\partial_f g(t^{(k+1)}, f^{(k)}) &= \partial_f\left(t^{(k+1)} - \frac{1}{n(1-\epsilon)}\sum_{i=1}^{n}\left(t^{(k+1)} - \mathcal{R}(f^{(k)}; \mathbf{x}_{\lfloor i \rfloor}, y_{\lfloor i \rfloor})\right)^+\right) \\
&= -\frac{1}{n(1-\epsilon)}\sum_{i=1}^{n(1-\epsilon)}\partial_f\left(t^{(k+1)} - \mathcal{R}(f^{(k)}; \mathbf{x}_{\lfloor i \rfloor}, y_{\lfloor i \rfloor})\right)^+ \\
&= \frac{1}{n(1-\epsilon)}\sum_{i=1}^{n}\nabla_f \mathcal{R}(f^{(k)}; \mathbf{x}_{\lfloor i \rfloor}, y_{\lfloor i \rfloor})\begin{cases}1 & \text{if } t > \hat{\nu}_i \\ 0 & \text{if } t < \hat{\nu}_i \\ [0,1] & \text{if } t = \hat{\nu}_i\end{cases}
\end{aligned}$$

Now we note $\hat{\nu}_{n(1-\epsilon)} \le t^{(k+1)} \le \hat{\nu}_{n(1-\epsilon)+1}$. Then, we have

$$\partial_f g(t^{(k+1)}, f^{(k)}) \ni \frac{1}{n(1-\epsilon)}\sum_{i=1}^{n(1-\epsilon)}\nabla_f \mathcal{R}(f^{(k)}; \mathbf{x}_{\lfloor i \rfloor}, y_{\lfloor i \rfloor})$$

This concludes the proof. ■

We have therefore shown that the two-step optimization of Subquantile Minimization gives the iterative thresholding algorithm.

4 Convergence

In this section we give the algorithm for subquantile minimization for both kernelized ridge regression and kernelized binary classification. Then we give our convergence results.

4.1 Kernelized Ridge 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

$$\mathcal{R}\left(f; \mathbf{x}_i, y_i,\right) = \left(f(\mathbf{x}_i) - y_i\right)^2 + \tau \left\|f\right\|_{\mathcal{H}}^2$$

Our goals throughout the proofs will be to obtain approximation bounds for infinite-dimensional kernels. The key challenge is the obvious undetermined problem, i.e. considering an infinite eigenfunction basis, we require infinite samples to obtain an accurate approximation. We will now give the algorithm.

Algorithm 1 (Subquantile Minimization for Kernelized Ridge Regression and Binary Classification by Gradient Descent)**.**

Input: Data Matrix: $\mathrm{X} \in \mathbb{R}^{n\times d}, n \gg d$; Labels: $\mathbf{y} \in \mathbb{R}^n$

1. Calculate the Kernel Matrix, $\mathrm{K}_{ij} \triangleq k(\mathbf{x}_i, \mathbf{x}_j)$.
2. Set the number of iterations

$$T = O\left(n(1-\epsilon)\log\left(\frac{\|f^*\|_{\mathcal{H}}}{\varepsilon}\right)\right)$$

3. **for** $k = 1, 2, \ldots, T$ **do**
    4. Find the Subquantile denoted as $\mathrm{S}^{(k)}$ as the set of $(1-\epsilon)n$ elements with the lowest error with respect to the loss function.
    5. Calculate the gradient update.

$$\nabla_f g(t^{(k+1)}, f^{(k)}) \leftarrow \frac{2}{n(1-\epsilon)} \sum_{i\in\mathrm{S}^{(k)}} (f^{(k)}(\mathbf{x}_i) - y_i)\cdot\phi(\mathbf{x}_i) + \tau f^{(t)} \qquad \text{(Regression)}$$

$$\nabla_f g(t^{(k+1)}, f^{(k)}) \leftarrow \frac{1}{n(1-\epsilon)} \sum_{i\in\mathrm{S}^{(k)}} (\sigma(f^{(k)}(\mathbf{x}_i)) - y_i)\cdot\phi(\mathbf{x}_i) + \tau f^{(t)} \qquad \text{(Classification)}$$

    6. Perform Gradient Descent Iteration.

$$f^{(k+1)} \leftarrow f^{(k)} - \eta\nabla g(f^{(k)}, t^{(k+1)})$$

Return: Function in $\mathcal{H}$: $f^{(T)}$

Algorithm 2 (Subquantile Minimization for Kernelized Regression with Full Solves)**.**

Input: Data Matrix: $\mathrm{X} \in \mathbb{R}^{n\times d}$; Labels: $\mathbf{y} \in \mathbb{R}^n$

1. Calculate the kernel matrix, $\mathrm{K}_{ij} \triangleq k(\mathbf{x}_i, \mathbf{x}_j)$.
2. Set the number of iterations

$$T = O\left(n\log\left(\frac{\|f^*\|_{\mathcal{H}}}{\varepsilon}\right)\right)$$

3. **for** $k = 1, 2, \ldots, T$ **do**
    4. Find the Subquantile denoted as $\mathrm{S}^{(k)}$ as the set of $(1-\epsilon)n$ elements with the lowest error with respect to the loss function.
    5. Solve the sub-problem

$$f^{(k)} \leftarrow \Phi_{\mathrm{S}^{(k)}}\left(\mathrm{K}_{\mathrm{S}^{(k)}} + n(1-\epsilon)\lambda\mathrm{I}\right)^{-1}\mathbf{y}_{\mathrm{S}^{(k)}}$$

Return: Function in $\mathcal{H}$: $f^{(T)}$

Theorem 9 (Subquantile Minimization for Kernelized Regression)**.** *Algorithm 1 run on a dataset* $\mathcal{D} \sim \hat{\mathcal{P}}$ *and return* $\hat{f}$*. Then with probability exceeding* $1-\delta$ *and when* $n \geq (1-\epsilon)^{-1}\left(256 + 64\left(P_k/\lambda_{\min}(\Gamma)\right)^2\log(2/\delta)\right)$ *and* $\epsilon \leq \frac{1}{32}\left(\frac{\lambda_{\min}(\Gamma)}{P_k}\right)^2$*,*

$$\left\|f^{(T)} - f^*\right\|_{\mathcal{H}} \leq \varepsilon + \frac{2\sigma}{3\left\|\Gamma\right\|_{\text{op}}} + \frac{2\sqrt{(1-\epsilon)}\left\|f^*\right\|_{\mathcal{H}}}{3\epsilon\sqrt{n}Q_k} + \frac{20}{3}\left(\frac{\epsilon}{1-2\epsilon}\right)\frac{\sqrt{\lambda_{\max}(\Gamma)Q_k}}{\lambda_{\min}(\Gamma)}\left\|f^*\right\|_{\mathcal{H}}$$

after $T = O\left(\frac{\epsilon}{1-\epsilon}\log\left(\frac{\|f^*\|_{\mathcal{H}}}{\varepsilon}\right)\right)$ *iterations.*

Runtime. The calculation of the Kernel Matrix, K, can be done in $O(n^2)$. To find $t^{(k+1)}$, we first must calculate the errors of all the elements, which is given by $\boldsymbol{\xi}^{(t)} \triangleq \mathrm{K}\mathbf{w}^{(t)} - \mathbf{y}$, considering $\mathrm{K} \in \mathbb{R}^{n\times n}$ and $\mathbf{w}^{(t)} \in \mathbb{R}^n$, this is a $O(n^2)$ calculation. Then to find the $n(1-\epsilon)$-th largest element, we can run a selection algorithm in worst-case time $O(n\log n)$. Calculating the gradient and updating the function is a $O(n^2)$ time step due to the matrix-vector multiplication. Then considering the choice of T, we have the algorithm runs in time $O\left(n^2\log\left(\frac{\|f^*\|_{\mathcal{H}}}{\varepsilon}\right)\right)$.

Theorem 10 (Subquantile Minimization for Kernelized Regression with Full Solves)**.** *Algorithm 2 run on a dataset* $\mathcal{D} \sim \hat{\mathcal{P}}$ *and return* $\hat{f}$*. Then with probability exceeding* $1-\delta$ *and when* $n \geq \Xi$ *and* $\epsilon \leq \Xi$*,*

$$\left\|f^{(T)} - f^*\right\|_{\mathcal{H}} \leq \varepsilon + O(\Xi)$$

after $T = O\left(\log\left(\frac{\Xi}{\varepsilon}\right)\right)$ *iterations.*

The full proof of Theorem 9 with explicit constants is given in Appendix C.2. A direct application of Theorem 9 is that learning an infinite dimensional function f^* to within ε error in the Hilbert Space Norm requires infinite data. Furthermore, we see that given covariate noise and label noise, our bound requires more iterations dependent on the magnitude of the corruption. Such a result is corroborated in [SST+18].

4.2 Kernelized Binary Classification

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

$$\mathcal{R}\left(f; \mathbf{x}_i, y_i\right) = -\mathbb{I}\left\{y_i = 1\right\}\log\left(\sigma\left(f(\mathbf{x}_i)\right)\right) - \mathbb{I}\left\{y_i = 0\right\}\log\left(1 - \sigma\left(f(\mathbf{x}_i)\right)\right)$$

Theorem 11 (Subquantile Minimization for Binary Classification is Good with High Probability)**.** *Let Algorithm alg:subq-kernel be run on a dataset* $\mathcal{D} \sim \hat{\mathcal{P}}$ *with learning rate* $\eta \triangleq \Omega(\ell^{-1})$*. Then after* $O\left(n\log\left(\frac{\|f^*\|_{\mathcal{H}}}{\varepsilon}\right)\right)$ *gradient descent iterations, with probability exceeding* $1-\delta$ *and a positive constant* C*,*

$$\left\|f^{(T)} - f^*\right\|_{\mathcal{H}} \leq \varepsilon + O\left(\sqrt{\frac{\mathcal{E}_{\text{OPT}}}{n(1-\epsilon)Q_m}}\right) + O\left(\frac{\sqrt{Q_k}\left\|f^*\right\|_{\mathcal{H}}}{\sqrt{Q_m n(1-\epsilon)}}\right)$$

for $n \geq (1-\epsilon)^{-1}\left(16\left\|\Gamma\right\|_{\text{op}}^2 + 2P_k^2\log(2/\delta)\right)$.

Proof. The proof is deferred to Appendix D.1. ■

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.

Extension to Finite Dimensional Kernels. When considering finite dimensional kernels we no longer require the Ridge.

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, $\mathcal{P}$, in the sub-Gaussian Design. We are able to show if the number of inliers is sufficiently small, then the kernelized binary classification problem with binary cross-entropy loss is consistent.

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

References

[ADKS22] Pranjal Awasthi, Abhimanyu Das, Weihao Kong, and Rajat Sen. Trimmed maximum likelihood estimation for robust generalized linear model. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022.

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

[B$^+$15] Sébastien Bubeck et al. Convex optimization: Algorithms and complexity. *Foundations and Trends® in Machine Learning*, 8(3-4):231–357, 2015.

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

[BJKK17] Kush Bhatia, Prateek Jain, Parameswaran Kamalaruban, and Purushottam Kar. Consistent robust regression. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[CDGS20] Yu Cheng, Ilias Diakonikolas, Rong Ge, and Mahdi Soltanolkotabi. High-dimensional robust mean estimation via gradient descent. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 1768–1778. PMLR, 13–18 Jul 2020.

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

[CLKZ21] Hugo Cui, Bruno Loureiro, Florent Krzakala, and Lenka Zdeborová. Generalization error rates in kernel regression: The crossover from the noiseless to noisy regime. *Advances in Neural Information Processing Systems*, 34:10131–10143, 2021.

[CLRS22] Thomas H Cormen, Charles E Leiserson, Ronald L Rivest, and Clifford Stein. *Introduction to algorithms*. MIT press, 2022.

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

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

[DKK+19] Ilias Diakonikolas, Gautam Kamath, Daniel M. Kane, Jerry Li, Jacob Steinhardt, and Alistair Stewart. Sever: A robust meta-algorithm for stochastic optimization. In *Proceedings of the 36th International Conference on Machine Learning*, ICML '19, pages 1596–1606. JMLR, Inc., 2019.

[FB81] Martin A. Fischler and Robert C. Bolles. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Commun. ACM*, 24(6):381–395, jun 1981.

[Fis22] Ronald A Fisher. On the mathematical foundations of theoretical statistics. *Philosophical transactions of the Royal Society of London. Series A, containing papers of a mathematical or physical character*, 222(594-604):309–368, 1922.

[FWZ18] Jianqing Fan, Weichen Wang, and Yiqiao Zhong. An l eigenvector perturbation bound and its application to robust covariance estimation. *Journal of Machine Learning Research*, 18(207):1–42, 2018.

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

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

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

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

[Jen06] Johan Ludwig William Valdemar Jensen. Sur les fonctions convexes et les inégalités entre les valeurs moyennes. *Acta mathematica*, 30(1):175–193, 1906.

[JNJ20] Chi Jin, Praneeth Netrapalli, and Michael Jordan. What is local optimality in nonconvex-nonconcave minimax optimization? In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 4880–4889. PMLR, 13–18 Jul 2020.

[JZL+18] Lu Jiang, Zhengyuan Zhou, Thomas Leung, Li-Jia Li, and Li Fei-Fei. Mentornet: Learning data-driven curriculum for very deep neural networks on corrupted labels. In *ICML*, 2018.

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

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

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

[LPMH21] Yassine Laguel, Krishna Pillutla, Jérôme Malick, and Zaid Harchaoui. Superquantiles at work: Machine learning applications and efficient subgradient computation. *Set-Valued and Variational Analysis*, 29(4):967–996, Dec 2021.

[M+89] Colin McDiarmid et al. On the method of bounded differences. *Surveys in combinatorics*, 141(1):148–188, 1989.

[MGJK19] Bhaskar Mukhoty, Govind Gopakumar, Prateek Jain, and Purushottam Kar. Globally-convergent iteratively reweighted least squares for robust regression problems. In Kamalika Chaudhuri and Masashi Sugiyama, editors, *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pages 313–322. PMLR, 16–18 Apr 2019.

[OZS20] Muhammad Osama, Dave Zachariah, and Petre Stoica. Robust risk minimization for statistical learning from corrupted data. *IEEE Open Journal of Signal Processing*, 1:287–294, 2020.

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

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

[RaR02] Raymond A Ryan and R a Ryan. *Introduction to tensor products of Banach spaces*, volume 73. Springer, 2002.

[RHL+20] Meisam Razaviyayn, Tianjian Huang, Songtao Lu, Maher Nouiehed, Maziar Sanjabi, and Mingyi Hong. Nonconvex min-max optimization: Applications, challenges, and recent theoretical advances. *IEEE Signal Processing Magazine*, 37(5):55–66, 2020.

[RRM14] R.T. Rockafellar, J.O. Royset, and S.I. Miranda. Superquantile regression with applications to buffered reliability, uncertainty quantification, and conditional value-at-risk. *European Journal of Operational Research*, 234(1):140–154, 2014.

[SBRJ19] Arun Sai Suggala, Kush Bhatia, Pradeep Ravikumar, and Prateek Jain. Adaptive hard thresholding for near-optimal consistent robust regression. In Alina Beygelzimer and Daniel Hsu, editors, *Proceedings of the Thirty-Second Conference on Learning Theory*, volume 99 of *Proceedings of Machine Learning Research*, pages 2892–2897. PMLR, 25–28 Jun 2019.

[Slu25] Eugen Slutsky. Über stochastische asymptoten und grenzwerte. *(No Title)*, 1925.

[SST+18] Ludwig Schmidt, Shibani Santurkar, Dimitris Tsipras, Kunal Talwar, and Aleksander Madry. Adversarially robust generalization requires more data. *Advances in neural information processing systems*, 31, 2018.

[TSM+17] Ilya Tolstikhin, Bharath K Sriperumbudur, Krikamol Mu, et al. Minimax estimation of kernel mean embeddings. *Journal of Machine Learning Research*, 18(86):1–47, 2017.

[TT15] Alex Townsend and Lloyd N Trefethen. Continuous analogues of matrix factorizations. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 471(2173):20140585, 2015.

[W+14] David P Woodruff et al. Sketching as a tool for numerical linear algebra. *Foundations and Trends® in Theoretical Computer Science*, 10(1–2):1–157, 2014.

A Probability Theory

In this section we will give various concentration inequalities on the inlier data for functions in the Reproducing Kernel Hilbert Space.

A.1 Finite Dimensional Concentrations of Measure

Proposition 12. *Let $\mu_1, \dots, \mu_n \sim \mathcal{N}(0, \sigma^2)$ for some $\sigma > 0$, then it follows for any $C \geq 1$,*

$$\mathbf{Pr}\left\{\sum_{i=1}^{n} \mu_i^2 \geq Cn\sigma^2\right\} \leq \exp\left(-(n/2)\left(C - 1 + \ln(1/C)\right)\right)$$

Proof. Concatenate all the samples μ_i into a vector $\boldsymbol{\mu} \in \mathbb{R}^n$.

$$\begin{aligned}
\mathop{\mathbf{Pr}}_{\boldsymbol{\mu}\sim\mathcal{N}(\mathbf{0},\sigma^2\mathrm{I})}\left\{\|\boldsymbol{\mu}\|^2 \geq t\right\} &\leq \inf_{\theta>0} \mathop{\mathbf{E}}_{\boldsymbol{\mu}\sim\mathcal{N}(\mathbf{0},\sigma^2\mathrm{I})}\left[\exp\left(\theta\sum_{i=1}^{n}\mu_i^2\right)\right]\exp(-\theta t) \\
&= \inf_{\theta>0}\prod_{i=1}^{n}\mathop{\mathbf{E}}_{\mu_i\sim\mathcal{N}(0,\sigma^2)}\left[\exp\left(\theta\mu_i^2\right)\right]\exp(-\theta t) \leq \inf_{0<\theta<(1/2)\sigma^{-2}}\prod_{i=1}^{n}\frac{1}{\sqrt{1-2\theta\sigma^2}}\exp(-\theta t) \\
&= \inf_{0<\theta<(1/2)\sigma^{-2}}\exp\left(-\left(\theta t + (n/2)\ln(1-2\theta\sigma^2)\right)\right) \\
&= \exp\left(-\left((t/2\sigma^2) - (n/2) + (n/2)\ln(n\sigma^2/t)\right)\right) \\
&= \exp\left(-(n/2)\left(C - 1 + \ln(1/C)\right)\right)
\end{aligned}$$

In the second inequality we utilize the MGF for a non-standard χ^2 variable. In the final equality we substitute in $t \triangleq Cn\sigma^2$. ∎

A.2 Hilbert Space Concentrations of Measure

Fact 13 (Sum of Binomial Coefficients [CLRS22])**.** *Let $k, n \in \mathbb{N}$ such that $k \leq n$, then*

$$\sum_{i=0}^{k}\binom{n}{i} \leq \left(\frac{en}{k}\right)^k$$

Proposition 14 (Jensen's Inequality [Jen06])**.** *Suppose φ is a convex function, then for a random variable X, it holds*

$$\varphi\left(\mathbf{E}\left[X\right]\right) \leq \mathbf{E}\left[\varphi\left(X\right)\right]$$

The inequality is reversed for φ concave.

We will now study the covariance approximation problem. Our main probabilistic tool will be McDiarmid's Inequality.

Proposition 15 (McDiarmid's Inequality [M+89])**.** *Suppose $f : \mathcal{X}_1 \times \cdots \mathcal{X}_n \to \mathbb{R}$. Consider i.i.d $X_1, \dots, X_n$ where $X_i \in \mathcal{X}_i$ for all $i \in [n]$. If there exists constants $c_1, \dots, c_n$, such that for all $x_i \in \mathcal{X}_i$ for all $i \in [n]$, it holds*

$$\sup_{\tilde{X}_i\in\mathcal{X}_i}\left|f(X_1,\dots,X_{i-1},X_i,X_{i+1},\dots,X_n) - f(X_1,\dots,X_{i-1},\tilde{X}_i,X_{i+1},\dots,X_n)\right| \leq c_i$$

Then for any $t > 0$, it holds

$$\mathbf{Pr}\left\{f(X_1,\dots,X_n) - \mathbf{E}\left[f(X_1,\dots,X_n)\right] \geq t\right\} \leq \exp\left(-\frac{2t^2}{\sum_{i=1}^{n}c_i^2}\right)$$

Theorem 16 (Mean Estimation in the Hilbert Space [TSM+17])**.** *Define $P_n \triangleq \frac{1}{n}\sum_{i=1}^{n}\delta_{X_i}$ and P be the distribution of the covariates in $\mathcal{X}$. Suppose $r : \mathcal{X} \to \mathcal{H}$ is a continuous function such that $\sup_{X\in\mathcal{X}}\|r(X)\|_{\mathcal{H}}^2 \leq C_k < \infty$. Then with probability at least $1-\delta$,*

$$\left\|\int_{\mathcal{X}} r(x)dP_n(x) - \int_{\mathcal{X}} r(x)dP(x)\right\| \leq \sqrt{\frac{C_k}{n}} + \sqrt{\frac{2C_k\log(1/\delta)}{n}}$$

We will strengthen upon the result by [TSM+17] by using knowledge of the distribution to first derive the expectation.

Proposition 17 (Probabilistic Bound on Infinite Dimensional Covariance Estimation)**.** *Let $\mathbf{x}_1, \ldots, \mathbf{x}_n$ be i.i.d sampled from $\mathcal{P}$ such that $\phi(\mathbf{x}_i) \triangleq X_i \sim \mathcal{P}_\sharp\phi$ (Assumption 4). Denote $\mathcal{S}$ as all subsets of $[n]$ with size from $n(1-2\epsilon)$ to $n(1-\epsilon)$. We then have simultaneously with probability exceeding $1-\delta$,*

$$\left\| \frac{1}{n}\sum_{i=1}^{n} X_i \otimes X_i - \Sigma \right\|_{\mathrm{HS}} \leq \sqrt{\frac{8}{n}}\, \|\Gamma\|_{\mathrm{op}} + \sqrt{\frac{2\log(2/\delta)}{n}} P_k$$

$$\max_{A\in\mathcal{S}} \left\| \frac{1}{n(1-\epsilon)} \sum_{i\in A} X_i \otimes X_i - \Sigma \right\|_{\mathrm{HS}} \leq \sqrt{\frac{8}{n(1-\epsilon)}}\, \|\Gamma\|_{\mathrm{op}} + \sqrt{\frac{2P_k^2\log(2/\delta)}{n(1-\epsilon)}} + P_k\sqrt{\frac{\epsilon\log\epsilon^{-1}}{(1-\epsilon)}}$$

Proof. We will calculate the mean operator in the Hilbert Space $\mathcal{H}\otimes\mathcal{H}$ and use the $\sqrt{n}$-consistency of estimating the mean-element in a Hilbert Space to obtain the probability bounds.

$$\begin{aligned}
&\mathop{\mathbf{E}}_{X_i\sim\mathcal{P}_\sharp\phi} \left\| \frac{1}{n(1-\epsilon)} \sum_{i=1}^{n(1-\epsilon)} X_i \otimes X_i - \Sigma \right\|_{\mathrm{HS}} \\
&\overset{(ii)}{\leq} \mathop{\mathbf{E}}_{X_i\sim\mathcal{P}_\sharp\phi} \mathop{\mathbf{E}}_{\tilde{X}_i\sim\mathcal{P}_\sharp\phi} \left\| \frac{1}{n(1-\epsilon)} \sum_{i=1}^{n(1-\epsilon)} X_i \otimes X_i - \tilde{X}_i \otimes \tilde{X}_i \right\|_{\mathrm{HS}} \\
&\overset{(iii)}{=} \mathop{\mathbf{E}}_{X_i\sim\mathcal{P}_\sharp\phi} \mathop{\mathbf{E}}_{\tilde{X}_i\sim\mathcal{P}_\sharp\phi} \mathop{\mathbf{E}}_{\xi_i\sim\mathcal{R}} \left\| \frac{1}{n(1-\epsilon)} \sum_{i=1}^{n(1-\epsilon)} \xi_i \left(X_i \otimes X_i - \tilde{X}_i \otimes \tilde{X}_i \right) \right\|_{\mathrm{HS}} \\
&\leq \mathop{\mathbf{E}}_{X_i\sim\mathcal{P}_\sharp\phi} \mathop{\mathbf{E}}_{\xi_i\sim\mathcal{R}} \left\| \frac{2}{n(1-\epsilon)} \sum_{i=1}^{n(1-\epsilon)} \xi_i X_i \otimes X_i \right\|_{\mathrm{HS}} \\
&\leq \frac{2}{n(1-\epsilon)} \mathop{\mathbf{E}}_{X_i\sim\mathcal{P}_\sharp\phi} \left(\mathop{\mathbf{E}}_{\xi_i\sim\mathcal{R}} \left\| \sum_{i=1}^{n(1-\epsilon)} \xi_i \phi(\mathbf{x}_i) \otimes \phi(\mathbf{x}_i) \right\|_{\mathrm{HS}}^2 \right)^{1/2}
\end{aligned}$$

In (ii) we apply a union bound. In (ii) we note that $X_i \otimes X_i - \Gamma$ is a mean $\mathbf{0}$ operator in the tensor product space $\mathcal{H}\otimes\mathcal{H}$. Then for $X, Y \in \mathcal{H}\otimes\mathcal{H}$ s.t. $\mathbf{E}[Y] = \mathbf{0}$ it follows $\|X\|_{\mathrm{HS}} = \|X - \mathbf{E}[Y]\|_{\mathrm{HS}} = \|\mathbf{E}[X-Y]\|_{\mathrm{HS}}$ and finally we apply Jensen's Inequality. Let e_k for $k\in[p]$ (p possibly infinite) represent a complete orthonormal basis for the image of Γ. By expanding out the Hilbert-Schmidt Norm, we then have

$$\begin{aligned}
&\frac{2}{n(1-\epsilon)} \left(\mathop{\mathbf{E}}_{X_i\sim\mathcal{P}_\sharp\phi} \mathop{\mathbf{E}}_{\xi_i\sim\mathcal{R}} \left\| \sum_{i=1}^{n(1-\epsilon)} \xi_i X_i \otimes X_i \right\|_{\mathrm{HS}}^2 \right)^{1/2} \\
&= \frac{2}{n(1-\epsilon)} \left(\mathop{\mathbf{E}}_{X_i\sim\mathcal{P}_\sharp\phi} \mathop{\mathbf{E}}_{\xi_i\sim\mathcal{R}} \sum_{k=1}^{p} \left\langle \sum_{i=1}^{n(1-\epsilon)} \xi_i (X_i\otimes X_i) e_k, \sum_{j=1}^{n(1-\epsilon)} \xi_j (X_j\otimes X_j) e_k \right\rangle_{\mathcal{H}} \right)^{1/2} \\
&= \frac{2}{n(1-\epsilon)} \left(\mathop{\mathbf{E}}_{X_i\sim\mathcal{P}_\sharp\phi} \mathop{\mathbf{E}}_{\xi_i\sim\mathcal{R}} \sum_{k=1}^{p} \sum_{i=1}^{n(1-\epsilon)} \sum_{j=1}^{n(1-\epsilon)} \xi_i\xi_j \left\langle (X_i\otimes X_i)e_k, (X_j\otimes X_j)e_k \right\rangle_{\mathcal{H}} \right)^{1/2} \\
&\overset{(iv)}{=} \frac{2}{n(1-\epsilon)} \left(\mathop{\mathbf{E}}_{X_i\sim\mathcal{P}_\sharp\phi} \sum_{k=1}^{p} \sum_{i=1}^{n(1-\epsilon)} \left\langle (X_i\otimes X_i)e_k, (X_i\otimes X_i)e_k \right\rangle_{\mathcal{H}} \right)^{1/2} \\
&= \frac{2}{n(1-\epsilon)} \left(\sum_{i=1}^{n(1-\epsilon)} \mathop{\mathbf{E}}_{X_i\sim\mathcal{P}_\sharp\phi} \|X_i\otimes X_i\|_{\mathrm{HS}}^2 \right)^{1/2} \\
&\overset{(v)}{=} \frac{2}{\sqrt{n(1-\epsilon)}} \left(\mathop{\mathbf{E}}_{X_i\sim\mathcal{P}_\sharp\phi} \|X_i\|_{\mathcal{H}}^4 \right)^{1/2}
\end{aligned}$$

(iv) follows from noticing $\mathbf{E}_{\xi_i,\xi_j\sim\mathcal{R}}[\xi_i\xi_j] = \delta_{ij}$. (v) follows from expanding the Hilbert-Schmidt Norm and applying Parseval's Identity. We will now calculate the fourth moment of a norm of sub-Gaussia function in the Hilbert Space.

$$\begin{aligned}
&\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{X}}\left[\|\phi(\mathbf{x})\|_{\mathcal{H}}^4\right] = \int_0^\infty \mathbf{Pr}\left\{\|\phi(\mathbf{x})\|_{\mathcal{H}}^4 \geq t\right\} dt = \int_0^\infty \mathbf{Pr}\left\{\|\phi(\mathbf{x})\|_{\mathcal{H}} \geq t^{1/4}\right\} dt \\
&\overset{(vi)}{\leq} \int_0^\infty \inf_{\theta>0} \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{X}}\left[\exp\left(\theta\|\phi(\mathbf{x})\|_{\mathcal{H}}\right)\right]\exp\left(-\theta t^{1/4}\right) dt \leq \int_0^\infty \inf_{\theta>0}\exp\left(\frac{\theta^2\|\Gamma\|_{\mathrm{op}}}{2} - \theta t^{1/4}\right) dt \\
&= \int_0^\infty \exp\left(-\frac{\sqrt{t}}{\|\Gamma\|_{\mathrm{op}}}\right) dt = 2\|\Gamma\|_{\mathrm{op}}^2
\end{aligned}$$

In (vi) we apply Markov's Inequality. From which we obtain,

$$\mathop{\mathbf{E}}_{X_i\sim\mathcal{P}_\sharp\phi}\left\|\frac{1}{n(1-\epsilon)}\sum_{i=1}^{n(1-\epsilon)} X_i\otimes X_i - \Sigma\right\|_{\mathrm{HS}} \leq \sqrt{\frac{8}{n(1-\epsilon)}}\|\Gamma\|_{\mathrm{op}}$$

Then, define the function $r(\mathbf{x}) : \mathcal{X} \to \mathcal{H}\otimes\mathcal{H}$, $\mathbf{x} \to \phi(\mathbf{x})\otimes\phi(\mathbf{x})$. From Assumption 5, we have $r(\mathbf{x}) = \|\phi(\mathbf{x})\otimes\phi(\mathbf{x})\|_{\mathrm{HS}} \leq \|\phi(\mathbf{x})\|_{\mathcal{H}}^2 \leq P_k$. We will use McDiamard's Inequality, consider $\tilde{P} \triangleq \delta_{X_i}$ with one modified element. Then consider the equation $f(x_1,\ldots,x_n) : \mathcal{X}\times\cdots\times\mathcal{X} \to \mathcal{H}\otimes\mathcal{H}\times\cdots\times\mathcal{H}\otimes\mathcal{H}$, $x_1,\ldots,x_n \to \left\|\int_{\mathcal{X}} r(x)dP_B(x) - \int_{\mathcal{X}} r(x)P(x)\right\|_{\mathrm{HS}}$.

$$\begin{aligned}
&\left\|\int_{\mathcal{X}} r(x)dP_B(x)dx - \int_{\mathcal{X}} r(x)dP(x)dx\right\|_{\mathrm{HS}} - \left\|\int_{\mathcal{X}} r(x)dP_{\tilde{B}}(x)dx - \int_{\mathcal{X}} r(x)dP(x)dx\right\|_{\mathrm{HS}} \\
&\qquad\leq \frac{1}{n(1-\epsilon)}\left(\|r(x_i)\|_{\mathrm{HS}} + \|r(\tilde{x}_i)\|_{\mathrm{HS}}\right) \leq \frac{2P_k}{n(1-\epsilon)}
\end{aligned}$$

Then, we have from McDiarmid's inequality (Proposition 15),

$$\mathbf{Pr}\left\{\left\|\int_{\mathcal{X}} r(x)dP_B(x) - \int_{\mathcal{X}} r(x)dP(x)\right\|_{\mathrm{HS}} - \sqrt{\frac{8}{n(1-\epsilon)}}\|\Gamma\|_{\mathrm{op}} \geq t\right\} \leq \exp\left(-\frac{t^2 n(1-\epsilon)}{P_k^2}\right)$$

We then have our first claim with probability exceeding $1-\delta$,

$$\left\|\int_{\mathcal{X}} r(x)dP_B(x) - \int_{\mathcal{X}} r(x)dP(x)\right\|_{\mathrm{HS}} \leq \sqrt{\frac{8}{n(1-\epsilon)}}\|\Gamma\|_{\mathrm{op}} + \sqrt{\frac{P_k^2\log(2/\delta)}{n(1-\epsilon)}}$$

Next, applying a union bound over $\mathcal{S}$ with Fact 13, we have

$$\max_{B\in S}\left\|\int_{\mathcal{X}} r(x)dP_B(x) - \int_{\mathcal{X}} r(x)dP(x)\right\|_{\mathrm{HS}} \leq \sqrt{\frac{8}{n(1-\epsilon)}}\|\Gamma\|_{\mathrm{op}} + \sqrt{\frac{P_k^2\log(2/\delta)}{n(1-\epsilon)} + \frac{P_k^2\epsilon\log\epsilon^{-1}}{(1-\epsilon)}}$$

Simplifying the resultant bound completes the proof. ■

B Proofs for Structural Results

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

B.1 Collaboration as a Criterion for Independence

In this section we discuss collaboration in the corrupted covariates. We consider the problems of fastly finding collaborating covariates and the probability the good points are collaborative.

Lemma 18. *Consider a determinate set of numbers* $(a_i)_{i=1}^n$*, and determinate set of functions in the Hilbert Space,* $(X_i)_{i=1}^n$*. It then follows,*

$$\left\|\sum_{i=1}^n a_i X_i\right\|_{\mathcal{H}}^2 \leq \|\boldsymbol{\alpha}\|_2^2 \|\mathrm{K}\|$$

Proof. The proof is a calculation.

$$\left\|\sum_{i=1}^n \alpha_i X_i\right\|_{\mathcal{H}}^2 \overset{(i)}{\leq} \|\boldsymbol{\alpha}\|_2^2 \max_{\mathbf{v}\in\mathbb{S}^{n-1}} \left\|\sum_{i=1}^n v_i X_i\right\|_{\mathcal{H}}^2 = \|\boldsymbol{\alpha}\|_2^2 \max_{\mathbf{v}\in\mathbb{S}^{n-1}} \left\langle \sum_{i=1}^n v_i X_i, \sum_{j=1}^n v_j X_j \right\rangle_{\mathcal{H}}$$

$$= \|\boldsymbol{\alpha}\|_2^2 \max_{\mathbf{v}\in\mathbb{S}^{n-1}} \sum_{i=1}^n \sum_{j=1}^n v_i v_j k(x_i, x_j) = \|\boldsymbol{\alpha}\|_2^2 \max_{\mathbf{v}\in\mathbb{S}^{n-1}} \mathbf{v}^\top \mathrm{K} \mathbf{v} = \|\boldsymbol{\alpha}\|_2^2 \|\mathrm{K}\|$$

where $\mathrm{K} \triangleq [K]_{ij} = k(x_i, x_j)$. The inequality in (i) is the most important step, $\mathbf{v}$ can be considered a unit weighting vector and we then multiply by the total weight. This inequality is sharp when $\alpha_i = \alpha_j$ for all $i, j \in [n]$. ■

Lemma 19. *Let* $(x_i)_{i=1}^n$ *such that* $X_i \triangleq \phi(x_i) \in \mathcal{H}$*. Let* $\mathrm{K} \in \mathbb{R}^{n\times n}$ *s.t.* $\mathrm{K}_{ij} \triangleq k(x_i, x_j)$*. Then it follows*

$$\left\|\sum_{i=1}^n X_i \otimes X_i\right\|_{\mathcal{H}\to\mathcal{H}}^2 = \|\mathrm{K}\|^2$$

Proof.

$$\left\|\sum_{i=1}^n X_i \otimes X_i\right\|_{\mathcal{H}\to\mathcal{H}}^2 = \max_{Y\in\mathcal{H}} \left\|\left[\sum_{i=1}^n X_i \otimes X_i\right] Y\right\|_{\mathcal{H}}^2 = \max_{Y\in\mathcal{H}} \left\|\sum_{i=1}^n X_i \langle X_i, Y\rangle_{\mathcal{H}}\right\|_{\mathcal{H}}^2$$

$$= \max_{Y\in\mathcal{H}} \sum_{i=1}^n \sum_{j=1}^n \mathrm{K}_{ij} \langle X_i, Y\rangle_{\mathcal{H}} \langle X_j, Y\rangle_{\mathcal{H}}$$

Let $\mathbf{v} \in \mathbb{R}^n$ s.t. $\mathrm{K}\mathbf{v} = \lambda_{\max}(\mathrm{K})\mathbf{v}$. We thus have $\Phi(\mathrm{K}\mathbf{v}) = \lambda_{\max}(\mathrm{K})\Phi\mathbf{v}$. Expanding out K, we have $\Phi \otimes \Phi(\Phi\mathbf{v}) = \lambda_{\max}(\mathrm{K})\Phi\mathbf{v}$. Alternatively, let $Y \in \mathcal{H}$ s.t. $[\Phi\otimes\Phi]\, Y = \lambda_{\max}(\Phi\otimes\Phi)Y$. Then, we have $\mathrm{K}(\Phi^* Y) = \lambda_{\max}(\Phi\otimes\Phi)\Phi^* Y$. Thus the maximum eigenvector of K is equal to $\Phi^*\mathrm{Y}$. ■

C Proofs for Kernelized Regression

We will first give a simple calculation of the β-smoothness parameter of the subquantile objective. We then will give proofs for our approximation error bounds.

C.1 Kernel Matrix Eigenvalue Concentration

Lemma 20. *Let* $X_i, \ldots, X_n \sim \mathcal{P}_\sharp \phi$*. Let* $\mathcal{S}$ *represent all permutations of* $[n]$ *from size* $[n(1-2\epsilon)]$ *to* $[n(1-\epsilon)]$*. Form the kernel matrix* $\mathrm{K} \in \mathbb{R}^{n\times n}$ *s.t.* $\mathrm{K}_{ij} \triangleq k(\mathbf{x}_i, \mathbf{x}_j)$*. Then with probability exceeding* $1-\delta$

$$\min_{\mathrm{A}\in\mathrm{S}} \lambda_{\max}(\mathrm{K}_\mathrm{A}) \geq 0.5n(1-2\epsilon)\lambda_{\min}(\Gamma)$$
$$\max_{\mathrm{A}\in\mathrm{S}} \lambda_{\min}(\mathrm{K}_\mathrm{A}) \leq 2n(1-\epsilon)\lambda_{\max}(\Gamma)$$

$n \geq (1-\epsilon)^{-1}\left(256 + 64\left(P_k/\lambda_{\min}(\Gamma)\right)^2 \log(2/\delta)\right)$ *and* $\epsilon \leq \frac{1}{32}\left(\lambda_{\min}(\Gamma)/P_k\right)^2$.

Proof. We will give our probabilistic bounds using the first and second relation in our covariance estimation bound given in Proposition 17.
Lower Bound.

$$\|\mathrm{K}_\mathrm{A}\| = \|\mathrm{X}_\mathrm{A} \otimes \mathrm{X}_\mathrm{A}\|_{\mathrm{op}} = \|n(1-2\epsilon)\Gamma + \mathrm{X}_\mathrm{A}\otimes\mathrm{X}_\mathrm{A} - n(1-2\epsilon)\Gamma\|_{\mathrm{op}}$$

$$
\begin{aligned}
&\geq n(1-\epsilon)\lambda_{\min}(\Gamma) - \|\mathrm{X_A} \otimes \mathrm{X_A} - n(1-\epsilon)\Gamma\|_{\mathrm{op}} \\
&\geq n(1-\epsilon)\left(\lambda_{\min}(\Gamma) - P_k\sqrt{\frac{\epsilon\log\epsilon^{-1}}{(1-\epsilon)}}\right) - \sqrt{n(1-\epsilon)}\left(\sqrt{8}\lambda_{\min}(\Gamma) + \sqrt{2P_k^2\log(2/\delta)}\right) \\
&\geq (1/2)n(1-\epsilon)\lambda_{\min}(\Gamma)
\end{aligned}
$$

when $n \geq (1-\epsilon)^{-1}\left(256 + 64\left(P_k/\lambda_{\min}(\Gamma)\right)^2\log(2/\delta)\right)$ and $\epsilon \leq \frac{1}{32}\left(\frac{\lambda_{\min}(\Gamma)}{P_k}\right)^2$ with probability exceeding $1-\delta$.

Upper Bound.

$$
\begin{aligned}
\|\mathrm{K_A}\| \leq \|\mathrm{K_P}\| &= \|\mathrm{X_P} \otimes \mathrm{X_P}\|_{\mathrm{op}} \\
&\leq n(1-\epsilon)\lambda_{\max}(\Gamma) + \sqrt{n(1-\epsilon)}\left(\sqrt{8}\lambda_{\max}(\Gamma) + \sqrt{2P_k^2\log(2/\delta)}\right) \\
&\leq 2n(1-\epsilon)\lambda_{\max}(\Gamma)
\end{aligned}
$$

when $n \geq (1-\epsilon)^{-1}\left(16 + 4\left(P_k/\|\Gamma\|_{\mathrm{op}}\right)^2\log(2/\delta)\right)$. This completes the proof. ■

We are now ready to prove our main approximation bound.

C.2 Proof of Theorem 9

Proof. From Algorithm 1, we have for kernelized linear regression the following update,

$$
f^{(t+1)} = f^{(t)} - \frac{2\eta}{n(1-\epsilon)}\sum_{i\in\mathrm{S}^{(t)}}(f^{(t)}(\mathbf{x}_i) - y_i)\cdot\phi(\mathbf{x}_i) + \lambda f^{(t)} \tag{2}
$$

Next, we note that we can partition $\mathrm{S}^{(t)} = (\mathrm{S}^{(t)} \cap \mathrm{P}) \cup (\mathrm{S}^{(t)} \cap \mathrm{Q}) \triangleq \mathrm{TP} \cup \mathrm{FP}$ as the *True Positives* and *False Positives*. Similarly, $X \setminus \mathrm{S}^{(t)} = ((X \setminus \mathrm{S}^{(t)}) \cap P) \cup ((X \setminus \mathrm{S}^{(t)}) \cap Q) \triangleq \mathrm{FN} \cup \mathrm{TN}$ represent the *False Negatives* and *True Negatives*. Define the following two functions that sum to $g(f^{(t)}, t^*)$,

$$
G(f) \triangleq \frac{1}{n(1-\epsilon)}\left[\sum_{i\in\mathrm{S}^{(t)}\cap\mathrm{P}}(f(\mathbf{x}_i) - y_i)^2\right] + \lambda\|f^{(t)}\|_{\mathcal{H}}^2
$$

$$
B(f) \triangleq \frac{1}{n(1-\epsilon)}\sum_{i\in\mathrm{S}^{(t)}\cap\mathrm{Q}}(f(\mathbf{x}_i) - y_i)^2
$$

Let f^*_{TP} represent the minimizer of G, i.e. $f^*_{\mathrm{TP}} \triangleq \mathrm{X_{TP}}\left(\mathrm{K_{TP}} + \lambda\mathrm{I}\right)^{-1}\mathbf{y}_{\mathrm{TP}}$. It then follows by the strong convexity of the KRR problem that $\nabla G(f^*_{\mathrm{TP}}) = \mathbf{0}$. Then, we have

$$
\begin{aligned}
\|f^{(t+1)} - f^*\|_{\mathcal{H}} &= \|f^{(t)} - \eta\nabla g(f^{(t)}, \hat{t}) - f^*\|_{\mathcal{H}} \\
&= \|f^{(t)} - f^* - \eta\nabla G(f^{(t)}) + \eta\nabla G(f^*) - \eta\nabla B(f^{(t)}) - \eta\nabla G(f^*) + \eta\nabla G(f^*_{\mathrm{TP}})\|_{\mathcal{H}} \\
&\leq \|f^{(t)} - f^* - \eta\nabla G(f^{(t)}) + \eta\nabla G(f^*)\|_{\mathcal{H}} + \|\eta\nabla G(f^*) - \eta\nabla G(f^*_{\mathrm{TP}})\|_{\mathcal{H}} + \|\eta\nabla B(f^{(t)})\|_{\mathcal{H}}
\end{aligned} \tag{3}
$$

We will expand the first term in Equation 3 through its square,

$$
\begin{aligned}
&\|f^{(t)} - f^* - \eta\nabla G(f^{(t)}) + \eta\nabla G(f^*)\|_{\mathcal{H}}^2 \\
&\qquad = \|f^{(t)} - f^*\|_{\mathcal{H}}^2 - 2\eta\left\langle f^{(t)} - f^*, \nabla G(f^{(t)}) - \nabla G(f^*)\right\rangle_{\mathcal{H}} + \eta^2\|\nabla G(f^{(t)}) - \nabla G(f^*)\|_{\mathcal{H}}^2
\end{aligned} \tag{4}
$$

Expanding out ∇G, we have

$$
\nabla G(f^{(t)}) - \nabla G(f^*) = \frac{2}{n(1-\epsilon)}\left[\sum_{i\in\mathrm{S}^{(t)}\cap\mathrm{P}}\phi(\mathbf{x}_i)\otimes\phi(\mathbf{x}_i)\right](f^{(t)} - f^*) + 2\lambda(f^{(t)} - f^*)
$$

We then have from the convexity of G and noting G is β-smooth with $\beta = 2\lambda + \frac{2\|\mathrm{K_{TP}}\|}{n(1-\epsilon)}$ (see [B+15] Lemma 3.5), furthermore we have with probability exceeding $1-\delta$, when $n \geq (1-\epsilon)^{-1}\left(16\|\Gamma\|_{\mathrm{op}}^2 + 2P_k^2\log(2/\delta)\right)$. Then $\beta \leq 2\lambda + 4\|\Gamma\|_{\mathrm{op}}$ with probability exceeding $1-\delta$,

$$2\eta\left\langle f^{(t)} - f^*, \nabla G(f^{(t)}) - \nabla G(f^*)\right\rangle_{\mathcal{H}} \geq 4\lambda\eta\left\|f^{(t)} - f^*\right\|_{\mathcal{H}}^2 + \frac{\eta}{2\lambda + 4\|\Gamma\|_{\mathrm{op}}}\left\|\nabla G(f^{(t)}) - \nabla G(f^*)\right\|_{\mathcal{H}}^2 \tag{5}$$

Solving the quadratic, we choose $\eta \leq (2\lambda + 4\|\Gamma\|_{\mathrm{op}})^{-1}$, we then have from Equations 4 and 5,

$$\left\|f^{(t)} - f^* - \eta\nabla G(f^{(t)}) + \eta\nabla G(f^*)\right\|_{\mathcal{H}} \leq \left\|f^{(t)} - f^*\right\|_{\mathcal{H}}\sqrt{1-4\eta\lambda}$$

We bound the two residual terms in the second term in Equation 3 individually.

$$\begin{aligned}
\left\|\nabla B(f^{(t)})\right\|_{\mathcal{H}}^2 &\stackrel{\text{def}}{=} \frac{4\eta^2}{[n(1-\epsilon)]^2}\left\|\sum_{i\in \mathrm{S}^{(t)}\cap \mathrm{Q}}(f^{(t)}(\mathbf{x}_i) - y_i)\cdot\phi(\mathbf{x}_i)\right\|_{\mathcal{H}}^2 \\
&\leq \frac{4\eta^2}{[n(1-\epsilon)]^2}\left(\|\mathrm{K_{FP}}\|\sum_{i\in \mathrm{S}^{(t)}\cap \mathrm{Q}}(f^{(t)}(\mathbf{x}_i) - y_i)^2\right) \leq \frac{4\eta^2}{[n(1-\epsilon)]^2}\left(\|\mathrm{K_{FP}}\|\sum_{i\in \mathrm{P}\setminus \mathrm{S}^{(t)}}(f^{(t)}(\mathbf{x}_i) - y_i)^2\right) \\
&\leq \frac{8\eta^2}{[n(1-\epsilon)]^2}\left(\|\mathrm{K_{FP}}\|\,\|\mathrm{K_{FN}}\|\left\|f^{(t)} - f^*\right\|_{\mathcal{H}}^2 + \|\mathrm{K_{FP}}\|\,\|\boldsymbol{\xi}_{\mathrm{FN}}\|^2\right)
\end{aligned}$$

where in the first inequality we applied Lemma 18, the second inequality follows from the optimality of the subquantile set, in the third inequality we applied Lemma 18 and the simple inequality $(a+b)^2 \leq a^2 + b^2$. Concatenate the Gaussian noise random variables, ξ_i into a vector $\boldsymbol{\xi}$, that we further partition into $\boldsymbol{\xi} \triangleq [\boldsymbol{\xi}_{\mathrm{TP}}, \boldsymbol{\xi}_{\mathrm{FN}}]$. For the next term, we have

$$\begin{aligned}
\|\nabla G(f^*) - \nabla G(f^*_{\mathrm{TP}})\|_{\mathcal{H}}^2 &\stackrel{\text{def}}{=} \left\|\frac{2}{n(1-\epsilon)}\sum_{i\in \mathrm{S}^{(t)}\cap \mathrm{P}}(f^* - f^*_{\mathrm{TP}})(\mathbf{x}_i)\cdot\phi(\mathbf{x}_i) + \lambda(f^* - f^*_{\mathrm{TP}})\right\|_{\mathcal{H}}^2 \\
&\leq \frac{8\|\mathrm{K_{TP}}\|\,\|f^* - f^*_{\mathrm{TP}}\|_{\mathcal{H}}^2}{[n(1-\epsilon)]^2} + 8\lambda^2\|f^* - f^*_{\mathrm{TP}}\|_{\mathcal{H}}^2
\end{aligned}$$

from applying the simple inequality $(a+b)^2 \leq a^2 + b^2$. We then obtain,

$$\begin{aligned}
\left\|f^{(t+1)} - f^*\right\|_{\mathcal{H}} \leq \left\|f^{(t)} - f^*\right\|_{\mathcal{H}}&\left(1 - \eta\left(2\lambda - \frac{\sqrt{8\|\mathrm{K_{FP}}\|\,\|\mathrm{K_{FN}}\|}}{n(1-\epsilon)}\right)\right) \\
&+ \frac{\eta}{n(1-\epsilon)}\left(\sqrt{8\|\mathrm{K_{FP}}\|}\,\|\boldsymbol{\xi}_{\mathrm{FN}}\| + \sqrt{8\|\mathrm{K_{TP}}\|}\,\|f^* - f^*_{\mathrm{TP}}\|_{\mathcal{H}}\right) + \sqrt{8}\eta\lambda\|f^* - f^*_{\mathrm{TP}}\|_{\mathcal{H}}
\end{aligned}$$

where we apply the inequality for any $x \leq 1$, it holds $\sqrt{1-x} \leq 1 - (x/2)$. We now bound $\|f^* - f^*_{\mathrm{TP}}\|$. Consider the $\mathbb{R}^n$ space first for simplicity. Let $\mathrm{X} \in \mathbb{R}^{\infty\times n}$ be a quasimatrix whose columns are in $\mathcal{H}$, then we denote $\mathrm{X}^*f \in \mathbb{R}^n$ as a vector where each entry is equal to $f(x_i)$.

$$\begin{aligned}
\|f^* - f^*_{\mathrm{TP}}\|_{\mathcal{H}} &\stackrel{\text{def}}{=} \left\|f^* - (\mathrm{X_{TP}}\otimes\mathrm{X_{TP}} + n(1-\epsilon)\lambda\mathrm{I})^{-1}\mathbf{y}_{\mathrm{TP}}\right\|_{\mathcal{H}} \\
&= \left\|f^* - (\mathrm{X_{TP}}\otimes\mathrm{X_{TP}} + n(1-\epsilon)\lambda\mathrm{I})^{-1}(\mathrm{X}^*_{\mathrm{TP}}f^* + \boldsymbol{\xi}_{\mathrm{TP}})\right\|_{\mathcal{H}} \\
&\leq \left\|f^* - (\mathrm{X_{TP}}\otimes\mathrm{X_{TP}} + n(1-\epsilon)\lambda\mathrm{I})^{-1}\mathrm{X}^*_{\mathrm{TP}}f^*\right\|_{\mathcal{H}} + \left\|\mathrm{X_{TP}}(\mathrm{K_{TP}} + n(1-\epsilon)\lambda\mathrm{I})^{-1}\boldsymbol{\xi}_{\mathrm{TP}}\right\|_{\mathcal{H}} \\
&\leq \|f^* - \underbrace{(\mathrm{X_{TP}}\otimes\mathrm{X_{TP}} + n(1-\epsilon)\lambda\mathrm{I})^{-1}\mathbf{y}^*_{\mathrm{TP}}}_{\triangleq \hat{f}^*_{\mathrm{TP}}}\|_{\mathcal{H}} + n(1-\epsilon)^{-1}\lambda\sqrt{\|\mathrm{K_{TP}}\|}\,\|\boldsymbol{\xi}_{\mathrm{TP}}\|
\end{aligned}$$

We then have,

$$\hat{f}^*_{\mathrm{TP}} \triangleq \arg\min_{f\in\mathcal{H}}\frac{1}{n(1-\epsilon)}\sum_{i\in\mathrm{TP}}(f(\mathbf{x}_i) - f^*(\mathbf{x}_i))^2 + \lambda\|f\|_{\mathcal{H}}^2$$

This is equivalent to satisfying the following equation,

$$\frac{1}{n(1-\epsilon)}\left[\sum_{i\in \mathrm{TP}} \left(f(\mathbf{x}_i) - f^*(\mathbf{x}_i)\right)\cdot \phi(\mathbf{x}_i)\right] = -\lambda f$$

From which it follows,

$$\left[\frac{1}{n(1-\epsilon)}\sum_{i\in \mathrm{TP}} \phi(\mathbf{x}_i)\otimes\phi(\mathbf{x}_i)\right](f - f^*) = -\lambda f$$

Then subtracting λf^* from both sides, we have

$$\left[\frac{1}{n(1-\epsilon)}\sum_{i\in \mathrm{TP}} \phi(\mathbf{x}_i)\otimes\phi(\mathbf{x}_i) + \lambda \mathrm{I}\right](f - f^*) = -\lambda f^*$$

We then obtain,

$$f - f^* = -\lambda\left[\frac{1}{n(1-\epsilon)}\sum_{i\in \mathrm{TP}} \phi(\mathbf{x}_i)\otimes\phi(\mathbf{x}_i) + \lambda \mathrm{I}\right]^{-1} f^*$$

Define $\hat{\Sigma}_{\mathrm{TP}} \triangleq [n(1-\epsilon)]^{-1}\sum_{i\in\mathrm{TP}}\phi(\mathbf{x}_i)\otimes\phi(\mathbf{x}_i)$. Then, we have,

$$\left\|\hat{f}^*_{\mathrm{TP}} - f^*\right\|_{\mathcal{H}} \le \frac{\lambda}{\lambda + \lambda_{\min}(\hat{\Sigma}_{\mathrm{TP}})}\left\|f^*\right\|_{\mathcal{H}} \le \lambda^{-1}_{\min}(\hat{\Sigma}_{\mathrm{TP}})\lambda\left\|f^*\right\|_{\mathcal{H}} \tag{6}$$

A similar bound in Equation 6 is difficult to find for infinite dimensional kernels as $\hat{\Sigma}_{\mathrm{TP}}$ will not have a minimum eigenvalue. Our final error bound thus pays the price of the Gaussian noise and the sub-optimality of the Ridge required for overcoming the corrupted gradient. From our choice of $\eta \le (32\left\|\Gamma\right\|_{\mathrm{op}}(Q_k \vee \left\|\Gamma\right\|_{\mathrm{op}}))^{-1/2}$ and $\lambda \ge 2.5\sqrt{Q_k\left\|\Gamma\right\|_{\mathrm{op}}}$, we have

$$\left\|f^{(t+1)} - f^*\right\|_{\mathcal{H}} \le \frac{3}{4}\left\|f^{(t)} - f^*\right\|_{\mathcal{H}} + \frac{\epsilon\sigma}{2(1-\epsilon)\left\|\Gamma\right\|_{\mathrm{op}}} + \frac{\left\|f^*\right\|}{2\sqrt{n(1-\epsilon)Q_k}} + 5\left(\frac{\epsilon}{1-\epsilon}\right)^2\frac{\sqrt{\left\|\Gamma\right\|_{\mathrm{op}}Q_k}}{\frac{1-2\epsilon}{1-\epsilon}\lambda_{\min}(\Gamma)}\left\|f^*\right\|_{\mathcal{H}}$$

when $n \ge (1-\epsilon)^{-1}\left(256 + 64\left(P_k/\lambda_{\min}(\Gamma)\right)^2\log(2/\delta)\right)$ and $\epsilon \le \frac{1}{32}\left(\frac{\lambda_{\min}(\Gamma)}{P_k}\right)^2$ with probability exceeding $1-\delta$. Then after $T = O\left(\frac{\epsilon}{1-\epsilon}\log\left(\frac{\|f^*\|}{\varepsilon}\right)\right)$ iterations, from the infinite geometric series, we have,

$$\left\|f^{(T)} - f^*\right\|_{\mathcal{H}} \le \varepsilon + \frac{2\sigma}{3\left\|\Gamma\right\|_{\mathrm{op}}} + \frac{2\sqrt{(1-\epsilon)}\left\|f^*\right\|_{\mathcal{H}}}{3\epsilon\sqrt{n}Q_k} + \frac{20}{3}\left(\frac{\epsilon}{1-2\epsilon}\right)\frac{\sqrt{\lambda_{\max}(\Gamma)Q_k}}{\lambda_{\min}(\Gamma)}\left\|f^*\right\|_{\mathcal{H}}$$

and the proof is complete. ∎

C.3 Proof of Theorem 10

Proof. Let $\mathrm{A},\mathrm{B} \in \mathbb{R}^{p\times p}$, then we have from the Matrix Woodburry Identity,

$$(\mathrm{A}+\mathrm{B})^{-1} = \mathrm{A}^{-1} - \left(\mathrm{A} + \mathrm{A}\mathrm{B}^{-1}\mathrm{A}\right)^{-1} = \mathrm{A}^{-1} - \left(\mathrm{A}\left(\mathrm{I} + \mathrm{B}^{-1}\mathrm{A}\right)\right)^{-1} = \mathrm{A}^{-1} - \left(\mathrm{I} + \mathrm{B}^{-1}\mathrm{A}\right)^{-1}\mathrm{A}^{-1} \tag{7}$$

Let $\left[E_{\mathrm{S}^{(t)}}\right]_i \triangleq y_i - f^*(\mathbf{x}_i)$. If we consider vectors where the i-th entry represents the projection on to ψ_i, then from Algorithm 2, we have

$$\begin{aligned}
f^{(t+1)} &\overset{\mathrm{def}}{=} \left(\mathrm{X}_{\mathrm{S}^{(t)}}\otimes\mathrm{X}_{\mathrm{S}^{(t)}} + \lambda\mathrm{I}_p\right)^{-1}\mathrm{X}_{\mathrm{S}^{(t)}}\mathbf{y}_{\mathrm{S}^{(t)}} \\
&= \left(\mathrm{X}_{\mathrm{S}^{(t)}}\otimes\mathrm{X}_{\mathrm{S}^{(t)}} + \lambda\mathrm{I}_p\right)^{-1}\left(\left[\mathrm{X}_{\mathrm{S}^{(t)}}\otimes\mathrm{X}_{\mathrm{S}^{(t)}}\right]f^* + \mathrm{X}_{\mathrm{S}^{(t)}}E_{\mathrm{S}^{(t)}}\right) \\
&\overset{(7)}{=} \left(\left[\mathrm{X}_{\mathrm{S}^{(t)}}\otimes\mathrm{X}_{\mathrm{S}^{(t)}}\right]^{-1} - \left(\mathrm{I}_p + \lambda^{-1}\mathrm{X}_{\mathrm{S}^{(t)}}\otimes\mathrm{X}_{\mathrm{S}^{(t)}}\right)^{-1}\left[\mathrm{X}_{\mathrm{S}^{(t)}}\otimes\mathrm{X}_{\mathrm{S}^{(t)}}\right]^{-1}\right)\left(\left[\mathrm{X}_{\mathrm{S}^{(t)}}\otimes\mathrm{X}_{\mathrm{S}^{(t)}}\right]f^* + \mathrm{X}_{\mathrm{S}^{(t)}}E_{\mathrm{S}^{(t)}}\right)
\end{aligned}$$

$$= f^* - \left(\lambda^{-1}\mathrm{X}_{\mathrm{S}^{(t)}} \otimes \mathrm{X}_{\mathrm{S}^{(t)}} + \mathrm{I}_p\right)^{-1} f^* + \left(\mathrm{X}_{\mathrm{S}^{(t)}} \otimes \mathrm{X}_{\mathrm{S}^{(t)}} + \lambda \mathrm{I}_p\right)^{-1} \mathrm{X}_{\mathrm{S}^{(t)}} E_{\mathrm{S}^{(t)}}$$
$$= f^* - \left(\lambda^{-1}\mathrm{X}_{\mathrm{S}^{(t)}} \otimes \mathrm{X}_{\mathrm{S}^{(t)}} + \mathrm{I}_p\right)^{-1} \left(f^* - \lambda^{-1}\mathrm{X}_{\mathrm{S}^{(t)}} E_{\mathrm{S}^{(t)}}\right)$$

From which it follows,

$$f^* - f^{(t+1)} = \left(\lambda^{-1}\mathrm{X}_{\mathrm{S}^{(t)}} \otimes \mathrm{X}_{\mathrm{S}^{(t)}} + \mathrm{I}_p\right)^{-1} f^* - \left(\lambda^{-1}\mathrm{X}_{\mathrm{S}^{(t)}} \otimes \mathrm{X}_{\mathrm{S}^{(t)}} + \mathrm{I}_p\right)^{-1} \lambda^{-1}\mathrm{X}_{\mathrm{S}^{(t)}} E_{\mathrm{S}^{(t)}}$$

Similarly, it follows

$$\mathbf{y}_{\mathrm{S}^{(t+1)}} - \mathrm{X}^*_{\mathrm{S}^{(t+1)}} f^{(t+1)} = E_{\mathrm{S}^{(t+1)}} + \mathrm{X}^*_{\mathrm{S}^{(t+1)}} \left(\lambda^{-1}\mathrm{X}_{\mathrm{S}^{(t)}} \otimes \mathrm{X}_{\mathrm{S}^{(t)}} + \mathrm{I}_p\right)^{-1} \left(f^* - \lambda^{-1}\mathrm{X}_{\mathrm{S}^{(t)}} E_{\mathrm{S}^{(t)}}\right)$$

We then have,

$$\begin{aligned}
\|E_{\mathrm{S}^{(t+1)}}\|^2 &= \left\|E_{\mathrm{S}^{(t+1)}} + \mathrm{X}^*_{\mathrm{S}^{(t+1)}}(\lambda^{-1}\mathrm{X}_{\mathrm{S}^{(t)}} \otimes \mathrm{X}_{\mathrm{S}^{(t)}} + \mathrm{I}_p)^{-1}(f^* - \lambda^{-1}\mathrm{X}_{\mathrm{S}^{(t)}} E_{\mathrm{S}^{(t)}})\right\|^2 \\
&\quad - \left\|\mathrm{X}^*_{\mathrm{S}^{(t+1)}}(\lambda^{-1}\mathrm{X}_{\mathrm{S}^{(t)}} \otimes \mathrm{X}_{\mathrm{S}^{(t)}} + \mathrm{I}_p)^{-1}(f^* - \lambda^{-1}\mathrm{X}_{\mathrm{S}^{(t)}} E_{\mathrm{S}^{(t)}})\right\|^2 \\
&\quad - 2 \cdot E^\top_{\mathrm{S}^{(t+1)}} \mathrm{X}^*_{\mathrm{S}^{(t+1)}}(\lambda^{-1}\mathrm{X}_{\mathrm{S}^{(t)}} \otimes \mathrm{X}_{\mathrm{S}^{(t)}} + \mathrm{I}_p)^{-1}(f^* - \lambda^{-1}\mathrm{X}_{\mathrm{S}^{(t)}} E_{\mathrm{S}^{(t)}}) \\
&\overset{(i)}{\leq} 2 \cdot \left\|\mathrm{X}^*_{\mathrm{P}}(\lambda^{-1}\mathrm{X}_{\mathrm{S}^{(t)}} \otimes \mathrm{X}_{\mathrm{S}^{(t)}} + \mathrm{I}_p)^{-1}(f^* - \lambda^{-1}\mathrm{X}_{\mathrm{S}^{(t)}} E_{\mathrm{S}^{(t)}})\right\|^2 + 2\left\|\boldsymbol{\mu}_{\mathrm{P}}\right\|^2 \\
&\quad - \left\|\mathrm{X}^*_{\mathrm{S}^{(t+1)}}(\lambda^{-1}\mathrm{X}_{\mathrm{S}^{(t)}} \otimes \mathrm{X}_{\mathrm{S}^{(t)}} + \mathrm{I}_p)^{-1}(f^* - \lambda^{-1}\mathrm{X}_{\mathrm{S}^{(t)}} E_{\mathrm{S}^{(t)}})\right\|^2 \\
&\quad - 2 \cdot E^\top_{\mathrm{S}^{(t+1)}} \mathrm{X}^*_{\mathrm{S}^{(t+1)}}(\lambda^{-1}\mathrm{X}_{\mathrm{S}^{(t)}} \otimes \mathrm{X}_{\mathrm{S}^{(t)}} + \mathrm{I}_p)^{-1}(f^* - \lambda^{-1}\mathrm{X}_{\mathrm{S}^{(t)}} E_{\mathrm{S}^{(t)}}) \\
&\leq 4 \cdot \left\|\mathrm{X}^*_{\mathrm{P}}(\lambda^{-1}\mathrm{X}_{\mathrm{S}^{(t)}} \otimes \mathrm{X}_{\mathrm{S}^{(t)}} + \mathrm{I}_p)^{-1} f^*\right\|^2_{\mathcal{H}} + 4 \cdot \left\|\mathrm{X}^*_{\mathrm{P}}(\lambda^{-1}\mathrm{X}_{\mathrm{S}^{(t)}} \otimes \mathrm{X}_{\mathrm{S}^{(t)}} + \mathrm{I}_p)^{-1}\lambda^{-1}\mathrm{X}_{\mathrm{S}^{(t)}} E_{\mathrm{S}^{(t)}}\right\|^2 \\
&\quad - 2 \cdot E^\top_{\mathrm{S}^{(t+1)}} \mathrm{X}^*_{\mathrm{S}^{(t+1)}}(\lambda^{-1}\mathrm{X}_{\mathrm{S}^{(t)}} \otimes \mathrm{X}_{\mathrm{S}^{(t)}} + \mathrm{I}_p)^{-1}(f^* - \lambda^{-1}\mathrm{X}_{\mathrm{S}^{(t)}} E_{\mathrm{S}^{(t)}}) + 2\left\|\boldsymbol{\mu}\right\|^2 \\
&\leq 4 \cdot \lambda^{-2} \left\|\mathrm{K}_{\mathrm{P}}\right\|_{\mathrm{F}} \left\|\mathrm{K}_{\mathrm{S}^{(t)}}\right\|_{\mathrm{F}} \left\|(\lambda^{-1}\mathrm{X}_{\mathrm{S}^{(t)}} \otimes \mathrm{X}_{\mathrm{S}^{(t)}} + \mathrm{I}_p)^{-1}\right\|^2 \left\|E_{\mathrm{S}^{(t)}}\right\|^2 \\
&\quad + 2 \cdot \sqrt{\left\|\mathrm{K}_{\mathrm{S}^{(t+1)}}\right\|_{\mathrm{F}}} \left\|f^*\right\|_{\mathcal{H}} \left\|(\lambda^{-1}\mathrm{X}_{\mathrm{S}^{(t)}} \otimes \mathrm{X}_{\mathrm{S}^{(t)}} + \mathrm{I}_p)^{-1}\right\|_{\mathrm{op}} \left\|E_{\mathrm{S}^{(t+1)}}\right\| \\
&\quad + 2 \cdot \sqrt{\left\|\mathrm{K}_{\mathrm{S}^{(t+1)}}\right\|_{\mathrm{F}}} \left\|(\lambda^{-1}\mathrm{X}_{\mathrm{S}^{(t)}} \otimes \mathrm{X}_{\mathrm{S}^{(t)}} + \mathrm{I}_p)^{-1}\right\|_{\mathrm{op}} \left\|E_{\mathrm{S}^{(t)}}\right\| \left\|E_{\mathrm{S}^{(t+1)}}\right\| \\
&\quad + 4 \cdot \sqrt{\left\|\mathrm{K}_{\mathrm{P}}\right\|_{\mathrm{F}}} \left\|(\lambda^{-1}\mathrm{X}_{\mathrm{S}^{(t)}} \otimes \mathrm{X}_{\mathrm{S}^{(t)}} + \mathrm{I}_p)^{-1}\right\|^2_{\mathrm{op}} \left\|f^*\right\|^2_{\mathcal{H}} + 2\left\|\boldsymbol{\mu}\right\|^2
\end{aligned}$$

where (i) follows from the optimality of the Subquantile set and the elementary inequality $(a^2 + b^2) \leq 2a^2 + 2b^2$. ■

D Proofs for Kernelized Binary Classification

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

D.1 Proof of Theorem 11

Proof. From Algorithm 1, we have for kernelized binary classification,

$$f^{(t+1)} = f^{(t)} - \frac{\eta}{n(1-\epsilon)}\left(\sum_{i \in \mathrm{S}^{(t)}} \left(\sigma(f^{(t)}(\mathbf{x}_i)) - y_i\right) \cdot \phi(\mathbf{x}_i)\right) - 2\tau f^{(t)} \tag{8}$$

Define the following two functions whose sum give $g(f^{(t)}, t^*)$,

$$G(f) \triangleq \frac{1}{n(1-\epsilon)}\left(\sum_{i \in \mathrm{S}^{(t)} \cap \mathrm{P}} -\mathbb{I}\{y_i = 1\} \log\left(\sigma(f(\mathbf{x}_i))\right) + -\mathbb{I}\{y_i = 0\} \log\left(1 - \sigma(f(\mathbf{x}_i))\right)\right) + \tau\left\|f^{(t)}\right\|^2_{\mathcal{H}}$$

$$B(f) \triangleq \frac{1}{n(1-\epsilon)}\left(\sum_{i \in \mathrm{S}^{(t)} \cap \mathrm{Q}} -\mathbb{I}\{y_i = 1\} \log\left(\sigma(f(\mathbf{x}_i))\right) - \mathbb{I}\{y_i = 0\} \log\left(1 - \sigma(f(\mathbf{x}_i))\right)\right)$$

Then, we have from the previous section,

$$\left\|f^{(t+1)}-f^*\right\|_{\mathcal{H}} \leq \left\|f^{(t)}-f^*-\eta\nabla G(f^{(t)})+\eta\nabla G(f^*)\right\|_{\mathcal{H}} + \left\|\eta\nabla G(f^*)-\eta\nabla G(f^*_{\mathrm{TP}})\right\| + \left\|\eta\nabla B(f^{(t)})\right\|_{\mathcal{H}}$$

We then have,

$$\nabla G(f^{(t)})-\nabla G(f^*) = \frac{1}{n(1-\epsilon)}\sum_{i\in\mathrm{S}^{(t)}\cap\mathrm{P}}\Big(\sigma(f^{(t)}(\mathbf{x}_i))-\sigma(f^*(\mathbf{x}_i)\Big)\cdot\phi(\mathbf{x}_i)+2\tau(f^{(t)}-f^*)$$

Let us now consider the function $h:\mathcal{H}\to\mathbb{R}$ defined as $h(f)\triangleq\sum_{i\in S\cap\mathrm{P}}\log(1+\exp(f(\mathbf{x}_i))$. We can then calculate the gradients by hand, $\nabla h(f)=\sum_{i\in S\cap\mathrm{P}}\sigma(f(\mathbf{x}_i))\cdot\phi(\mathbf{x}_i)$ and $\nabla^2 h(f)=\sum_{i\in S\cap\mathrm{P}}\sigma(f(\mathbf{x}_i))(1-\sigma(f(\mathbf{x}_i)))\cdot\phi(\mathbf{x}_i)\otimes\phi(\mathbf{x}_i)$. Considering $\mathrm{S}^{(t)}\cap\mathrm{P}$ has a finite cardinality, $h(f)$ is convex and has smoothness constant $(1/4)\left\|\mathrm{K}_{\mathrm{TP}}\right\|$, then when $n\geq(1-\epsilon)^{-1}\left(16\left\|\Gamma\right\|_{\mathrm{op}}^2+2P_k^2\log(2/\delta)\right)$, then with probability exceeding $1-\delta$, we have $\left\|\mathrm{K}_{\mathrm{TP}}\right\|\leq 2n(1-\epsilon)\left\|\Gamma\right\|_{\mathrm{op}}$. Then we have from Lemma 3.5 in [B+15] and the fact that $\mathcal{H}$ is an inner-product space,

$$\left\langle f^{(t)}-f^*,\eta\nabla G(f^{(t)})-\eta\nabla G(f^*)\right\rangle_{\mathcal{H}} \geq 2\eta\tau\left\|f^{(t)}-f^*\right\|_{\mathcal{H}}^2+\frac{\eta}{(1/2)\left\|\mathrm{K}_{\mathrm{TP}}\right\|+2\tau}\left\|\nabla G(f^{(t)})-\nabla G(f^*)\right\|_{\mathcal{H}}^2$$

We thus have for $\eta\leq((1/2)\left\|\mathrm{K}_{\mathrm{TP}}\right\|+2\tau)^{-1}$,

$$\left\|f^{(t)}-f^*-\eta\nabla G(f^{(t)})+\eta\nabla G(f^*)\right\|_{\mathcal{H}}\leq\left\|f^{(t)}-f^*\right\|_{\mathcal{H}}\sqrt{1-4\eta\tau}$$

We now analyze the error term by term individually.

$$\begin{aligned}
\left\|\eta\nabla B(f^{(t)})\right\|_{\mathcal{H}}^2 &\overset{\text{def}}{=} \frac{\eta^2}{[n(1-\epsilon)]^2}\left\|\sum_{i\in\mathrm{S}^{(t)}\cap\mathrm{Q}}(\sigma(f^{(t)}(\mathbf{x}_i)-y_i)\cdot\phi(\mathbf{x}_i)\right\|_{\mathcal{H}}^2\\
&\leq\frac{\eta^2}{[n(1-\epsilon)]^2}\left\|\mathrm{K}_{\mathrm{FP}}\right\|\sum_{i\in\mathrm{P}\setminus\mathrm{S}^{(t)}}(\sigma(f^{(t)}(\mathbf{x}_i))-y_i)^2\\
&\leq\frac{2\eta^2}{[n(1-\epsilon)]^2}\left\|\mathrm{K}_{\mathrm{FP}}\right\|\left(\sum_{i\in\mathrm{P}\setminus\mathrm{S}^{(t)}}(\sigma(f^{(t)}(\mathbf{x}_i))-\sigma(f^*(\mathbf{x}_i)))^2+\mathcal{E}_{\mathrm{OPT}}\right)
\end{aligned}$$

where $\mathcal{E}_{\mathrm{OPT}}\triangleq\sum_{i\in\mathrm{P}\setminus S^{(t)}}(\sigma(f^*(\mathbf{x}_i))-y_i)^2=\sum_{i\in\mathrm{P}\setminus\mathrm{S}^{(t)}}(\mathbf{E}\left[y_i|\mathbf{x}_i\right]-y_i)^2$. In the first inequality we use Lemma 18 and in the last inequality we use the elementary inequality $(a+b)^2\leq 2a^2+2b^2$. The sigmoid function is 1-Lipschitz, thus we have for any $\mathbf{x}\in\mathcal{X}$,

$$(\sigma(f^{(t)}(\mathbf{x}))-\sigma(f^*(\mathbf{x})))^2\leq(f^{(t)}(\mathbf{x})-f^*(\mathbf{x}))^2$$

Then, we have

$$\begin{aligned}
\sum_{i\in\mathrm{P}\setminus\mathrm{S}^{(t)}}(f^{(t)}(\mathbf{x}_i)-f^*(\mathbf{x}_i))^2 &= \left\langle\sum_{i\in\mathrm{P}\setminus\mathrm{S}^{(t)}}\phi(\mathbf{x}_i)\otimes\phi(\mathbf{x}_i),(f^{(t)}-f^*)\otimes(f^{(t)}-f^*)\right\rangle_{\mathrm{HS}}\\
&\leq\left\|\sum_{i\in\mathrm{P}\setminus\mathrm{S}^{(t)}}\phi(\mathbf{x}_i)\otimes\phi(\mathbf{x}_i)\right\|_{\mathrm{HS}}\left\|(f^{(t)}-f^*)\otimes(f^{(t)}-f^*)\right\|_{\mathrm{HS}}\\
&=\left\|\mathrm{K}_{\mathrm{FN}}\right\|_{\mathrm{F}}\left\|f^{(t)}-f^*\right\|_{\mathcal{H}}^2
\end{aligned}$$

where in the last inequality we use the fact that the eigenvalues of K_{FN} are equivalent to the sum of tensor products and Parseval's Identity in the second term. Combining the previous two results, we have

$$\left\|\eta\nabla B(f^{(t)})\right\|_{\mathcal{H}}\leq\frac{\eta\sqrt{2\left\|\mathrm{K}_{\mathrm{FP}}\right\|\left\|\mathrm{K}_{\mathrm{FN}}\right\|_{\mathrm{F}}}\left\|f^{(t)}-f^*\right\|_{\mathcal{H}}}{n(1-\epsilon)}+\frac{\eta\sqrt{2\left\|\mathrm{K}_{\mathrm{FP}}\right\|\mathcal{E}_{\mathrm{OPT}}}}{n(1-\epsilon)}$$

Now we will analyze the second term in the error.

$$\|\eta\nabla G(f^*) - \eta\nabla G(f^*_{\mathrm{TP}})\|^2_{\mathcal{H}} = \eta^2\left\|\frac{1}{n(1-\epsilon)}\sum_{i\in \mathrm{S}^{(t)}\cap \mathrm{P}}(\sigma(f^*(\mathbf{x}_i)) - \sigma(f^*_{\mathrm{TP}}(\mathbf{x}_i)))\cdot\phi(\mathbf{x}_i) + \tau(f^*-f^*_{\mathrm{TP}})\right\|^2_{\mathcal{H}}$$
$$\leq \frac{\eta^2\,\|\mathrm{K}_{\mathrm{TP}}\|\,\|f^*-f^*_{\mathrm{TP}}\|^2_{\mathcal{H}}}{[n(1-\epsilon)]^2} + \eta^2\tau^2\,\|f^*-f^*_{\mathrm{TP}}\|^2_{\mathcal{H}}$$

It now suffices to bound $\|f^*-f^*_{\mathrm{TP}}\|_{\mathcal{H}}$. The analysis will rely on probability theory for maximum likelihood estimation. First we decompse the normed differene with the triangle inequality.

$$\|f^*-f^*_{\mathrm{TP}}\|_{\mathcal{H}} = \left\|f^*-\hat{f}^*_{\mathrm{TP}}+\hat{f}^*_{\mathrm{TP}}-f^*_{\mathrm{TP}}\right\|_{\mathcal{H}} \leq \underbrace{\left\|f^*-\hat{f}^*_{\mathrm{TP}}\right\|_{\mathcal{H}}}_{\text{SMALL DATA}} + \underbrace{\left\|\hat{f}^*_{\mathrm{TP}}-f^*_{\mathrm{TP}}\right\|_{\mathcal{H}}}_{\text{RIDGE}}$$

We will show the SMALL DATA term goes to zero. First by defintion of $\hat{f}^*_{\mathrm{TP}}$, we have

$$\hat{f}^*_{\mathrm{TP}} \overset{\mathrm{def}}{=} \arg\max_{f\in\mathcal{H}} \frac{1}{n(1-\epsilon)}\sum_{i\in\mathrm{TP}}\log\mathcal{L}\left(X_i \mid f^*\right)$$

The Fisher Information Matrix [Fis22] is given as

$$I(f^*) \overset{\mathrm{def}}{=} -\mathbf{E}_Y\left[\nabla^2_f\mathcal{L}(f^*;\mathrm{X}_{\mathrm{TP}},\mathbf{y}_{\mathrm{TP}})\right] = \frac{1}{n(1-\epsilon)}\sum_{i\in\mathrm{TP}}\sigma(f^*(\mathbf{x}_i))(1-\sigma(f^*(\mathbf{x}_i))\cdot\phi(\mathbf{x}_i)\otimes\phi(\mathbf{x}_i)$$

Then from the Slutsky Theorem [Slu25], we have $f^*-\hat{f}^*_{\mathrm{TP}} \sim \frac{1}{\sqrt{n}}\mathcal{N}\left(\mathbf{0}, I^{-1}\right)$. From standard concentration inequalities, we have with probability exceeding $1-\delta$, for a vector $X\sim\mathcal{N}(\mathbf{0},I^{-1})$. If we consider the eigenvalue decomposition $I^{-1} = \mathrm{U}\Lambda\mathrm{U}^*$, it then follows $\|X\|_{\mathcal{H}} \leq \sqrt{(1/\delta)\,\|\Lambda\|^2_{\mathrm{HS}}}$ with probability exceeding $1-\delta$. We then have $\left\|f^*-\hat{f}^*_{\mathrm{TP}}\right\|_{\mathcal{H}} \lesssim \frac{\sqrt{(1/\delta)}\|\Gamma^{-1}\|_{\mathrm{HS}}}{\sqrt{n}}$. Now we analyze the effect of the RIDGE. By definition of f^*_{TP}, we have

$$f^*_{\mathrm{TP}} \overset{\mathrm{def}}{=} \arg\max_{f\in\mathcal{H}} \frac{1}{n(1-\epsilon)}\sum_{i\in\mathrm{TP}}\log\mathcal{L}(X_i,f) - \tau\,\|f\|^2_{\mathcal{H}}$$

Then, from the strong convexity of the regularlized maximum likelihood estimation problem, we have

$$\left\|\hat{f}^*_{\mathrm{TP}}-f^*_{\mathrm{TP}}\right\|_{\mathcal{H}} \leq \mu^{-1}\left\|\frac{1}{n(1-\epsilon)}\sum_{i\in\mathrm{TP}}\nabla_f\log\mathcal{L}(X_i\mid\hat{f}^*_{\mathrm{TP}}) - \nabla_f\log\mathcal{L}(X_i\mid f^*_{\mathrm{TP}}) - \tau(\hat{f}^*_{\mathrm{TP}}-f^*_{\mathrm{TP}})\right\|_{\mathcal{H}}$$
$$= \mu^{-1}\left\|\sum_{i=1}^{n}\nabla_f\log\mathcal{L}(X_i\mid\hat{f}^*_{\mathrm{TP}}) - \tau\hat{f}^*_{\mathrm{TP}}\right\|_{\mathcal{H}} = \frac{\tau}{\tau+C_1\lambda_{\min}(\hat{\Sigma})}_{\mathcal{H}}\left\|\hat{f}^*_{\mathrm{TP}}\right\|_{\mathcal{H}} \leq \tau C_1^{-1}\lambda^{-1}_{\min}(\hat{\Sigma})\left\|\hat{f}^*_{\mathrm{TP}}\right\|_{\mathcal{H}}$$

where in the first equality, we used the optimality of $\hat{f}^*_{\mathrm{TP}}$ over the regularized log-likehood and in the second inequality we used the optimality of f^*_{TP} over the standard log-likelihood. Then from the triangle inequality, we have

$$\left\|f^*-f_{\mathrm{TP}}\right\|_{\mathcal{H}} \leq \tau C_1^{-1}\lambda^{-1}_{\min}(\hat{\Sigma})\left(\|f^*\|_{\mathcal{H}} + \frac{C_1^{-1}\lambda^{-1}_{\min}(\Gamma)}{\sqrt{n}}\right)$$

■

Table 1: Notations for Kernelized Regression

Symbol	**Definition**	**Takes Values In**
$g(f,t)$	Subquantile Function	$\mathcal{H} \times \mathbb{R} \to \mathbb{R}$
$\mathcal{P}$	Distribution of Good Data in the input space $\mathcal{X}$	$(\mathcal{X}, \mathcal{A}, \mathbf{Pr})$
$\phi(\mathbf{x}_i)$	Feature Transform Function	$\mathcal{X} \to \mathcal{H}$
ξ_i	Gaussian Noise, $\xi_i \sim \mathcal{N}(0, \sigma^2)$	$\mathbb{R}$
Φ_{A}	Quasimatrix (columns represent Hilbert Space Functions in A)	$\mathcal{H} \times \cdots \times \mathcal{H}$
Σ	Covariance of Good Data, $\mathbf{E}_{X \sim \mathcal{P}_\sharp \phi}[X \otimes X]$	$\mathcal{H} \otimes \mathcal{H}$
Γ	Sub-Gaussian Proxy for good data (see § 2.3)	$\mathcal{H} \otimes \mathcal{H}$
K_{A}	Kernel Matrix s.t. $\mathrm{K}_{ij} = k(\mathbf{x}_i, \mathbf{x}_j)$ for $i, j \in \mathrm{A}$	$\mathbb{R}^{n \times n}$
P	Set of vectors for good data	$\{\mathcal{X}\}$
Q	Set of vectors for bad data	$\{\mathcal{X}\}$
X	All Data, $\mathrm{P} \cup \mathrm{Q}$	$\{\mathcal{X}\}$
TP	*True Positives*, $\mathrm{S}^{(t)} \cap \mathrm{P}$	$\{\mathcal{X}\}$
FP	*False Positives*, $\mathrm{S}^{(t)} \cap \mathrm{Q}$	$\{\mathcal{X}\}$
TN	*True Negatives*, $(\mathrm{X} \setminus \mathrm{S}^{(t)}) \cap \mathrm{Q}$	$\{\mathcal{X}\}$
FN	*False Negatives*, $(\mathrm{X} \setminus \mathrm{S}^{(t)}) \cap \mathrm{P}$	$\{\mathcal{X}\}$
P_k	$\max_{\mathbf{x} \in P} \sqrt{k(\mathbf{x}_i, \mathbf{x}_i)}$	$\mathbb{R}_+$
Q_k	$\max_{\mathbf{x} \in Q} \sqrt{k(\mathbf{x}_i, \mathbf{x}_i)}$	$\mathbb{R}_+$
f^*_{TP}	$\arg\min_{f \in \mathcal{H}} \frac{1}{n(1-\epsilon)} \sum_{i \in \mathrm{TP}} (f(\mathbf{x}_i) - f^*(\mathbf{x}_i) - \xi_i)^2 + \tau \left\|f\right\|_{\mathcal{H}}^2$	$\mathcal{H}$
$\hat{f}^*_{\mathrm{TP}}$	$\arg\min_{f \in \mathcal{H}} \frac{1}{n(1-\epsilon)} \sum_{i \in \mathrm{TP}} (f(\mathbf{x}_i) - f^*(\mathbf{x}_i))^2 + \tau \left\|f\right\|_{\mathcal{H}}^2$	$\mathcal{H}$
$\hat{\Sigma}_{\mathrm{TP}}$	$\frac{1}{n(1-\epsilon)} \sum_{i \in \mathrm{TP}} \phi(\mathbf{x}_i) \otimes \phi(\mathbf{x}_i)$	$\mathcal{H} \otimes \mathcal{H}$

Table 2: Notations for Kernelized Regression

Symbol	Definition	Takes Values In
$g(f, t)$	Subquantile Function	$\mathcal{H} \times \mathbb{R} \to \mathbb{R}$
$\mathcal{P}$	Distribution of Good Data in the input space $\mathcal{X}$	$(\mathcal{X}, \mathcal{A}, \mathbf{Pr})$
$\phi(\mathbf{x}_i)$	Feature Transform Function	$\mathcal{X} \to \mathcal{H}$
p	Rank of the Finite-Dimensional Kernel	$\mathbb{N}$
$\mathcal{L}(f; \mathrm{X}, \mathbf{y})$	$\sum_{i=1}^n -\mathbb{I}\{y_i = 1\} \log(\sigma(f(\mathbf{x}_i))) - \mathbb{I}\{y_i = 0\} \log(1 - \sigma(f(\mathbf{x}_i)))$	$\mathbb{R}$
I	Fisher Information Matrix	$\mathbb{R}^{p \times p}$
Φ	Quasimatrix (columns represent Hilbert Space Functions)	$\mathcal{H} \times \cdots \times \mathcal{H}$
Σ	Covariance of Good Data, $\mathbf{E}_{X \sim \mathcal{P}_\sharp \phi}[X \otimes X]$	$\mathcal{H} \otimes \mathcal{H}$
Γ	Sub-Gaussian Proxy for good data (see § 2.3)	$\mathcal{H} \otimes \mathcal{H}$
K_A	Kernel Matrix s.t. $\mathrm{K}_{ij} = k(\mathbf{x}_i, \mathbf{x}_j)$ for $i, j \in \mathrm{A}$	$\mathbb{R}^{n \times n}$
P	Set of vectors for good data	$\{\mathcal{X}\}$
Q	Set of vectors for bad data	$\{\mathcal{X}\}$
X	All Data, $\mathrm{P} \cup \mathrm{Q}$	$\{\mathcal{X}\}$
TP	*True Positives*, $\mathrm{S}^{(t)} \cap \mathrm{P}$	$\{\mathcal{X}\}$
FP	*False Positives*, $\mathrm{S}^{(t)} \cap \mathrm{Q}$	$\{\mathcal{X}\}$
TN	*True Negatives*, $(\mathrm{X} \setminus \mathrm{S}^{(t)}) \cap \mathrm{Q}$	$\{\mathcal{X}\}$
FN	*False Negatives*, $(\mathrm{X} \setminus \mathrm{S}^{(t)}) \cap \mathrm{P}$	$\{\mathcal{X}\}$
P_k	$\max_{\mathbf{x} \in P} \sqrt{k(\mathbf{x}_i, \mathbf{x}_i)}$	$\mathbb{R}_+$
Q_k	$\max_{\mathbf{x} \in Q} \sqrt{k(\mathbf{x}_i, \mathbf{x}_i)}$	$\mathbb{R}_+$
f^*_{TP}	$\arg\min_{f \in \mathcal{H}} \frac{1}{n(1-\epsilon)} \sum_{i \in \mathrm{TP}} (f(\mathbf{x}_i) - f^*(\mathbf{x}_i) - \xi_i)^2 + \tau \Vert f \Vert_{\mathcal{H}}^2$	$\mathcal{H}$
$\hat{f}^*_{\mathrm{TP}}$	$\arg\min_{f \in \mathcal{H}} \frac{1}{n(1-\epsilon)} \sum_{i \in \mathrm{TP}} (f(\mathbf{x}_i) - f^*(\mathbf{x}_i))^2 + \tau \Vert f \Vert_{\mathcal{H}}^2$	$\mathcal{H}$
$\hat{\Sigma}_{\mathrm{TP}}$	$\frac{1}{n(1-\epsilon)} \sum_{i \in \mathrm{TP}} \phi(\mathbf{x}_i) \otimes \phi(\mathbf{x}_i)$	$\mathcal{H} \otimes \mathcal{H}$