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 results are a near-optimal approximation bound for learning the rank-m projection of the target function for kernelized ridge regression and a consistent optimal[1] approximation bound for learning the rank-m projection of the target function for kernelized binary classification.

[1] We define consistent optimal as going to optimal as $n \to \infty$.

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, $\mathbf{X}$ 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 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, $\phi(\mathbf{x}_i) \sim \mathcal{N}(0, \mathbf{\Sigma})$ where $\mathbf{\Sigma}$ 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 $\mathbf{\Sigma}$ and trace-class sub-Gaussian proxy $\mathbf{\Gamma}$ such that $\mathbf{\Sigma} \preceq c\mathbf{\Gamma}$. 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:

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

Kernel Binary Classification:

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

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 1 and 2, 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 $\{1, 2, \ldots, T\}$. 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 $f(\cdot) \triangleq \sum_{i \in [n]} w_i k(\mathbf{x}_i, \cdot)$.

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 $\mathbf{E}[X \otimes X] = \mathbf{\Sigma}$ with proxy $\mathbf{\Gamma}$, where $\mathbf{\Sigma} \preceq 4 \|X\|_{\psi_2}^2 \mathbf{\Gamma}$, where we denote $\preceq$ as the Löwner order. We have X is a centered Hilbert Space sub-Gaussian random function if for all $\theta > 0$,

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

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

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

Then we say $X \sim \mathcal{SG}(\mathbf{\Gamma}, \alpha)$, where if $\alpha = 1$, we will say $X \sim \mathcal{SG}(\mathbf{\Gamma})$. 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, $\phi(\cdot) : \mathcal{X} \to \mathcal{H}$,*

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

where $\mathbf{\Gamma}$ 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 $\mathcal{P} = \mathcal{Q} = \mathcal{N}(\mathbf{0}, \mathbf{I})$ or multivariate sub-Gaussian with proxy $\mathbf{I}$. Their proof of optimality relies on the fact that $\lambda_{\min}(\mathbf{\Sigma}) = \lambda_{\max}(\mathbf{\Sigma})$ and with sufficient data, $\lambda_{\max}(\mathbf{X})/\lambda_{\min}(\mathbf{X}) \searrow 1$. 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 $\mathbf{\Sigma}^{-1/2}$ so the columns of $\mathbf{X}$ are sampled from a multivariate sub-Gaussian Distribution with proxy $\mathbf{I}$ before running the iterative thresholding procedure.

This does not generalize to kernel learning where we are given a matrix $\mathbf{K}$ which is equivalent to inner product of the quasimatrix[2], $\mathbf{\Phi}$, with itself. In this case is not possible to sketch [W+14] the input matrix to have well-conditioned covariates. Thus, we are left with $\mathbf{\Phi}$ where the columns are sampled from a sub-Gaussian Distribution with proxy $\mathbf{\Gamma}$ is a trace-class operator, which implies the eigenvalues tend to zero, i.e. $\lambda_{\inf}(\mathbf{\Gamma}) = 0$, and there is no longer a notion of $\lambda_{\min}(\mathbf{\Gamma})$.

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.

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

The p-subquantile of the risk is then given by

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

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

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

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 $\text{Proj}_{\mathcal{K}}$ 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

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

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

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.

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.

[2] 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]).

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 $\hat{\nu}_{n(1-\epsilon)} \leq t^{(k+1)} \leq \hat{\nu}_{n(1-\epsilon)+1}$. 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

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

We will now give the algorithm.

Algorithm 1 (Subquantile Minimization for Kernelized Ridge Regression)**.**

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

1. Calculate the Kernel Matrix, $\mathbf{K}_{ij} \triangleq k(\mathbf{x}_i, \mathbf{x}_j)$.
2. Calculate the ℓ-smoothness constant, $\ell = \frac{1}{n(1-\epsilon)} \|\mathbf{K}\|$
3. Set the step-size $\eta = 1/\ell$
4. Set the number of iterations

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

5. **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. Calculate the gradient update.

$$\nabla_f g(t^{(k+1)}, f^{(k)}) \leftarrow \frac{2}{n(1-\epsilon)}\left(\sum_{i \in S^{(k)}} (f^{(k)}(\mathbf{x}_i) - y_i) \cdot \phi(\mathbf{x}_i) + C f^{(t)}\right)$$

   5. Perform Gradient Descent Iteration.

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

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

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.

Theorem 9 (Subquantile Minimization for Kernelized Regression)**.** *Algorithm 2 run on a dataset $\mathcal{D} \sim \hat{\mathcal{P}}$ and return $\hat{f}$. Then with probability exceeding $1-\delta$,*

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

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

Runtime. The calculation of the Kernel Matrix, $\mathbf{K}$, 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 $\boldsymbol{\xi}^{(t)} \triangleq \mathbf{K}\mathbf{w}^{(t)} - \mathbf{y}$, considering $\mathbf{K} \in \mathbb{R}^{n \times n}$ 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 $O\left(n^3(1-\epsilon)\log\left(\frac{\|f^*\|_{\mathcal{H}}}{\varepsilon}\right)\right)$.

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 $(\mathbf{x}_i, y_i) \sim \mathcal{D}$.

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

We will now give our algorithm for subquantile minimization in kernelized binary classification.

Algorithm 2 (Subquantile Minimization for Binary Classification)**.**

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

1. Calculate the kernel matrix, $\mathbf{K}_{ij} \triangleq k(\mathbf{x}_i, \mathbf{x}_j)$.
2. Calculate the ℓ-smoothness constant, $\ell = \frac{1}{n(1-\epsilon)} \|\mathbf{K}\|$
3. Set the step-size $\eta = 1/\ell$
4. Set the number of iterations
$$T = O\left(n \log\left(\frac{\|f^*\|_{\mathcal{H}}}{\varepsilon}\right)\right)$$
5. **for** $k = 1, 2, \ldots, T$ **do**
   6. 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.
   7. Calculate the gradient update.
$$\nabla_f g(t^{(k+1)}, f^{(k)}) \leftarrow \frac{1}{n(1-\epsilon)} \sum_{i \in S^{(k)}} (\sigma(f^{(k)}(\mathbf{x}_i)) - y_i) \cdot \phi(\mathbf{x}_i) + C f^{(k)}$$
   8. Perform Gradient Descent Iteration.
$$f^{(k+1)} \leftarrow f^{(k)} - \eta \nabla g(f^{(k)}, t^{(k+1)})$$

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

Theorem 10 (Subquantile Minimization for Binary Classification is Good with High Probability)**.** *Let Algorithm alg:subq-kernel-classification 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,*

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

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

Proof. The proof is deferred to Appendix D.2. ■

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.

[CY21] Xiaohui Chen and Yun Yang. Hanson–Wright inequality in Hilbert spaces with application to K-means clustering for non-Euclidean data. *Bernoulli*, 27(1):586 – 614, 2021.

[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.

[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.

[HÖ89] O. Hölder. Ueber einen mittelwerthabsatz. *Nachrichten von der Königl. Gesellschaft der Wissenschaften und der Georg-Augusts-Universität zu Göttingen*, 1889:38–47, 1889.

[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.

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

$$\mathbf{Pr}\left\{\max_{i\in[n]} |\mu_i| \geq \sigma\sqrt{2\log n}\cdot s\right\} \leq \frac{\sqrt{2}}{\log n} e^{-s^2}$$

Proof. Let C be a positive constant to be determined.

$$\begin{aligned}\mathop{\mathbf{Pr}}_{\mu_i\sim\mathcal{N}(0,\sigma^2)}\left\{\max_{i\in[n]} |\mu_i| \geq C\cdot s\right\} &\overset{(i)}{=} 2n \mathop{\mathbf{Pr}}_{\mu\sim\mathcal{N}(0,\sigma^2)}\{\mu \geq C\cdot s\} = \frac{2n}{\sigma\sqrt{2\pi}}\int_{C\cdot s}^{\infty} e^{-\frac{1}{2}\left(\frac{x}{\sigma}\right)^2} dx \\ &\leq 2\sigma n\left(\frac{1}{C\cdot s}\right) e^{-\frac{1}{2}\left(\frac{C\cdot s}{\sigma}\right)^2} \leq \frac{\sqrt{2}n^{1-s^2}}{s\log n} \leq \frac{\sqrt{2}}{\log n} e^{-s^2}\end{aligned}$$

(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 $\mathbf{Pr}\{\max_{i\in[n]} X_i \geq C\} = n\,\mathbf{Pr}\{X \geq C\}$. In the second to last inequality, we plug in $C \triangleq \sigma\sqrt{2\log n}$. Our proof is now complete. ■

Proposition 12. *Let $\mu_1, \ldots, \mu_n \sim \mathcal{N}(0, \sigma^2)$ 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 $\boldsymbol{\mu} \in \mathbb{R}^n$.

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

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

Proposition 13 (Maximum of Squared Gaussians)**.** *Let $\mu_1, \ldots, \mu_n \sim \mathcal{N}(0, \sigma^2)$ for $\sigma > 0$, $n > 1$. Then it follows*

$$\mathop{\mathbf{Pr}}_{\mu_i\sim\mathcal{N}(0,\sigma^2)}\left\{\max_{i\in[n]}\mu_i^2 \geq 2\sigma^2\log(n)\cdot t\right\} \leq e^{-t}$$

Proof. Our proof follows similarly to the proof for Proposition 11.

$$\begin{aligned}\mathop{\mathbf{Pr}}_{\mu_i\sim\mathcal{N}(0,\sigma^2)}\left\{\max_{i\in[n]}\mu_i^2 \geq Ct\right\} &\leq 2n \mathop{\mathbf{Pr}}_{\mu_i\sim\mathcal{N}(0,\sigma^2)}\left\{\mu_i \geq \sqrt{Ct}\right\} = 2n\int_{\sqrt{Ct}}^{\infty}\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{x}{\sigma}\right)^2}dx \\ &\leq 2n\int_{\sqrt{Ct}}^{\infty}\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{x}{\sigma}\right)^2}dx \leq 2n\int_{\sqrt{Ct}}^{\infty}\left(\frac{x}{\sqrt{Ct}}\right)\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{x}{\sigma}\right)^2}dx \leq \frac{2n\sigma}{\sqrt{2\pi Ct}}e^{-\frac{Ct}{2\sigma^2}}\end{aligned}$$

Then, setting $C \triangleq 2\sigma^2\log(n)$, and simplifying the resultant bound to a convenient form completes the proof. ■

A.2 Hilbert Space Concentrations of Measure

Lemma 14 (Sum of Sub-Gaussian Hilbert Space Functions)**.** *Suppose $X_1, \dots, X_n \sim \mathcal{P}$ where $\mathcal{P}$ is sub-Gaussian with proxy trace class operator, $\mathbf{\Gamma}$. Let $a_1, \dots, a_n$ be a fixed set of numbers in $\mathbb{R}$. Then $\sum_{i=1}^n a_i X_i$ is sub-Gaussian with proxy $\left(\sum_{i=1}^n a_i^2\right)\mathbf{\Gamma}$.*

Proof. Let $v \in \mathcal{H}$ such that $\|v\|_{\mathcal{H}} = 1$. Then, we have for a $\theta > 0$,

$$\begin{aligned}\mathbf{E}\left[\exp\left(\theta\left\langle\sum_{i=1}^n a_i X_i, v\right\rangle_{\mathcal{H}}\right)\right] &= \mathbf{E}\left[\prod_{i=1}^n \exp\left(\theta\left\langle a_i X_i, v\right\rangle_{\mathcal{H}}\right)\right] \overset{(i)}{\leq} \prod_{i=1}^n \mathbf{E}\left[\exp\left(n a_i \theta\left\langle X_i, v\right\rangle_{\mathcal{H}}\right)\right]^{1/n} \\ &\leq \prod_{i=1}^{n(1-\epsilon)} \exp\left(\frac{\theta^2 a_i^2\left\langle v, \mathbf{\Gamma} v\right\rangle_{\mathcal{H}}}{2}\right) \leq \exp\left(\frac{\theta^2\left(\sum_{i=1}^n a_i^2\right)\left\langle v, (\mathbf{\Gamma})\, v\right\rangle_{\mathcal{H}}}{2}\right)\end{aligned}$$

where (i) follows from Hölder's Inequality for a product of functions [H$\ddot{\text{o}}$9]. We see the resultant variance proxy is $n\mathbf{\Gamma}$ and the proof is complete. ■

Theorem 15 (Hilbert Space Hanson Wright [CY21])**.** *Let X_i be a i.i.d sequence of sub-Gaussian random variables in $\mathcal{H}$ such that $\mathbf{E}[X_i] = 0$ and $\mathbf{E}[X_i \otimes X_i] = \mathbf{\Gamma}$. Then there exists a universal constant $C > 0$ s.t. for any $t > 0$,*

$$\mathbf{Pr}\left\{\left\|\sum_{i=1}^n X_i\right\|_{\mathcal{H}}^2 \geq n \operatorname{Tr}(\mathbf{\Gamma}) + t\right\} \leq 2\exp\left[-C\min\left(\frac{t^2}{n^2\|\mathbf{\Gamma}\|_{\text{HS}}^2}, \frac{t}{n\|\mathbf{\Gamma}\|_{\text{op}}}\right)\right]$$

From Theorem 15, it follows that the LHS is less than $\delta \in (0,1)$ when

$$t \geq \frac{1}{C} n\|\mathbf{\Gamma}\|_{\text{op}} \log(2/\delta) \vee \sqrt{\frac{1}{C} n^2\|\mathbf{\Gamma}\|_{\text{HS}}^2 \log(2/\delta)}$$

Furthermore, we have when

$$\delta \geq 2\exp\left[-C\left(\frac{\|\mathbf{\Gamma}\|_{\text{HS}}}{\|\mathbf{\Gamma}\|_{\text{op}}}\right)^2\right]$$

it follows

$$t \geq n\|\mathbf{\Gamma}\|_{\text{HS}} \sqrt{(1/C)\log(2/\delta)}$$

In other words, when the failure probability is sufficiently small we can use the above bound. We will reference this idea throughout this section.

Fact 16 (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 17 (Jensen's Inequality [Jen06])**.** *Suppose φ is a convex function, then for a random variable X, it holds*

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

The inequality is reversed for φ concave.

Proposition 18 (RKHS Norm of Functions in the Reproducing Kernel Hilbert Space)**.** *Let $\mathbf{x}_i \sim \mathcal{X}$ such that $\phi(\mathbf{x}_i) \triangleq X_i \sim \mathcal{P}_\sharp\phi$ (Assumption 4). Denote $\mathcal{S}$ as all subsets of $[n(1-\epsilon)]$ with size $n(1-2\epsilon)$ to $n(1-\epsilon)$ for $\epsilon < 0.5$, $P_B = \sum_{i=1}^{n(1-\epsilon)} \delta_{X_i}$ where δ is a Dirac measure at X_i, and $r(\mathbf{x}) : \mathbb{R}^d \to \mathcal{H}$ such that $\mathbf{x} \to \phi(\mathbf{x}_i)$. Then it follows with probability exceeding $1-\delta$,*

$$\max_{B \in S}\left\|\int_{\mathcal{X}} r(x) dP_B(x)\right\|_{\mathcal{H}} \leq n^{3/2}\sqrt{(8/C)\epsilon\log\epsilon^{-1}}\,\|\mathbf{\Gamma}\|_{\text{Tr}}$$

when $n \geq \left(1 + \sqrt{(1/C)\log(2/\delta)}\right)\left(\sqrt{(2/C)\epsilon\log\epsilon^{-1}}\right)^{-1}$.

Proof. From Lemma 14, we have that for any $B \in \mathcal{S}$, $\sum_{i \in B} \phi(\mathbf{x})$ is sub-Gaussian with proxy $n(1-\epsilon)\mathbf{\Gamma}$. We will then apply a union bound over $\mathcal{S}$ to give the claimed bound. First, we calculate the expectation with the sub-Gaussian property.

$$\begin{aligned}\mathop{\mathbf{E}}_{Y \sim \mathcal{SG}(n\mathbf{\Gamma})} \|Y\|_{\mathcal{H}}^2 &= \int_0^\infty \mathop{\mathbf{Pr}}_{Y \sim \mathcal{SG}(n\mathbf{\Gamma})} \left\{ \|Y\| \geq \sqrt{t} \right\} dt \leq \int_0^\infty \inf_{\theta > 0} \mathop{\mathbf{E}}_{Y \sim \mathcal{SG}(n\mathbf{\Gamma})} \left[\exp\left(\theta \|Y\|_{\mathcal{H}}\right) \right] \exp\left(-\theta\sqrt{t}\right) dt \\ &\leq \int_0^\infty \inf_{\theta > 0} \exp\left((n/2)\theta^2 \|\mathbf{\Gamma}\|_{\text{op}} - \theta\sqrt{t} \right) dt = \int_0^\infty \exp\left(-\frac{t}{2n\|\mathbf{\Gamma}\|_{\text{op}}} \right) dt = 2n\|\mathbf{\Gamma}\|_{\text{op}}\end{aligned}$$

Next, we calculate the c_i for McDiarmid's Inequality. Define $h(x_1, \ldots, x_{n(1-\epsilon)} : \mathcal{X} \times \cdots \times \cdots \mathcal{X} \to \mathcal{H} \times \cdots \times \mathcal{H}$, then define the Dirac Measure $P \triangleq \sum_{i=1}^{n(1-\epsilon)} \delta_{x_i}$. Then $x_1 \times \cdots \times x_n \to \left\| \int_{\mathcal{X}} \phi(x) P(x) \right\|_{\mathcal{H}}$. Then consider x_i modified as $\tilde{x}_i$ with correspondingly modified Dirac measure, $\tilde{P}$.

$$\left\| \int_{\mathcal{X}} \phi(x) dP(x) dx \right\|_{\mathcal{H}} - \left\| \int_X \phi(x) d\tilde{P}(x) \right\|_{\mathcal{H}} \leq \|\phi(x_i)\|_{\mathcal{H}} + \|\phi(\tilde{x}_i)\|_{\mathcal{H}} \leq 2P_k$$

Then from McDiarmid's Inequality (Proposition 19), we have

$$\mathbf{Pr}\left\{ \|Y\|_{\mathcal{H}}^2 - 2n\|\mathbf{\Gamma}\|_{\text{op}} \geq t \right\} \leq \exp\left(-\frac{2t^2}{nP_k^2} \right)$$

This gives us a probabilistic bound for a single sample of Y, to find the maximum over $\mathcal{S}$, we will use a union bound and Fact 16. Then, we have

$$\mathbf{Pr}\left\{ \max_{Y \in \mathcal{S}} \|Y\|_{\mathcal{H}}^2 \geq 2n\|\mathbf{\Gamma}\|_{\text{op}} + t \right\} \leq \frac{e^{-\frac{2t^2}{nP_k^2}}}{\binom{n}{n\epsilon}} \leq e^{-\frac{2t^2}{nP_k^2}} \left(\frac{\epsilon}{e}\right)^{n\epsilon}$$

Then, we have with probability exceeding $1-\delta$,

$$\max_{Y \in \mathcal{S}} \|Y\|_{\mathcal{H}}^2 \leq 2n\|\mathbf{\Gamma}\|_{\text{op}} + (1/\sqrt{2})P_k \sqrt{n^2 \epsilon \log \epsilon^{-1} + n \log(1/\delta)}$$

After using the sufficient data condition, the proof is complete. ■

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

Proposition 19 (McDiarmid's Inequality [M+89])**.** *Suppose $f : \mathcal{X}_1 \times \cdots \mathcal{X}_n \to \mathbb{R}$. Consider i.i.d $X_1, \ldots, X_n$ where $X_i \in \mathcal{X}_i$ for all $i \in [n]$. If there exists constants $c_1, \ldots, c_n$, such that for all $x_i \in \mathcal{X}_i$ for all $i \in [n]$, it holds*

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

Then for any $t > 0$, it holds

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

Theorem 20 (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 21 (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 $\phi(\mathbf{x}_i) \sim \mathcal{P}$ (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 $1-\delta$,*

$$\left\| \frac{1}{n} \sum_{i=1}^{n} \phi(\mathbf{x}_i) \otimes \phi(\mathbf{x}_i) - \boldsymbol{\Sigma} \right\|_{\text{HS}} \leq \sqrt{\frac{8}{n}} \|\boldsymbol{\Gamma}\|_{\text{op}} + \sqrt{\frac{2\log(2/\delta)}{n}} P_k$$

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

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.

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

In (ii) we apply a union bound. In (ii) we note that $\phi(\mathbf{x}_i) \otimes \phi(\mathbf{x}_i) - \boldsymbol{\Gamma}$ 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 $\|X\|_{\text{HS}} = \|X - \mathbf{E}[Y]\|_{\text{HS}} = \|\mathbf{E}[X - Y]\|_{\text{HS}}$ 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 $\boldsymbol{\Gamma}$. By expanding out the Hilbert-Schmidt Norm, we then have

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

(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 have

$$\mathop{\mathbf{E}}_{\mathbf{x} \sim \mathcal{X}} \left[\|\phi(\mathbf{x})\|_{\mathcal{H}}^4 \right] = \int_0^\infty \mathbf{Pr} \left\{ \|\phi(\mathbf{x})\|_{\mathcal{H}}^4 \geq t \right\} dt = \int_0^\infty \mathbf{Pr} \left\{ \|\phi(\mathbf{x})\|_{\mathcal{H}} \geq t^{1/4} \right\} dt$$

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

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

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

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

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

Then, we have from McDiarmid's inequality (Proposition 19),

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

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

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

Next, applying a union bound over $\mathcal{S}$ with Fact 16, we have

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

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 22. *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,*

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

Proof. The proof is a calculation.

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

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

where $\mathbf{K} \triangleq [K]_{ij} = k(x_i,x_j)$. 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 $i,j \in [n]$. ■

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 Subquantile Smoothness

Lemma 23. *(ℓ-Smoothness of $g(t,f)$ w.r.t f). Let $\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_n \sim \hat{\mathcal{P}}$. It then follows*

$$\left\|\nabla_f g(t,f) - \nabla_f g(t,\hat{f})\right\|_{\mathcal{H}} \leq \ell\left\|f-\hat{f}\right\|_{\mathcal{H}}$$

where $\ell = \frac{2}{n(1-\epsilon)}\operatorname{Tr}(\mathbf{K})$ and with probability exceeding the following hold simulataneously.

$$\lambda_{\max}(\boldsymbol{\Gamma}) \leq \ell \leq 4\lambda_{\max}(\boldsymbol{\Gamma}) + \frac{\epsilon}{1-\epsilon}Q_k$$

if $n \geq 128 + 32\left(P_k/\lambda_{\max}(\boldsymbol{\Gamma})\right)^2\log(2/\delta)$.

Proof. We will upper bound the operator norm of the Hessian Operator. We have from Section 3,

$$\left\|\nabla_f^2 g(t,f)\right\|_{\text{op}} = \frac{2}{n(1-\epsilon)}\left\|\sum_{i=1}^{n}\mathbb{I}\left\{t \geq \ell(f;\mathbf{x}_i,y_i)\right\}\phi(\mathbf{x}_i)\otimes\phi(\mathbf{x}_i) + C\mathbf{I}\right\|_{\text{op}}$$

$$\leq \frac{2}{n(1-\epsilon)}\left\|\sum_{i=1}^{n}\phi(\mathbf{x}_i)\otimes\phi(\mathbf{x}_i) + C\mathbf{I}\right\|_{\text{op}} = \frac{2}{n(1-\epsilon)}\left\|\boldsymbol{\Phi}\otimes\boldsymbol{\Phi} + nC\mathbf{I}\right\|_{\text{op}} = \frac{2}{n(1-\epsilon)}\|\mathbf{K}\| + C$$

We will now give our probabilistic bounds using the first relation in our covariance estimation bound given in Proposition 21.

Lower Bound.

$$\begin{aligned}\|\mathbf{K}\| = \|\boldsymbol{\Phi}\otimes\boldsymbol{\Phi}\| &\geq \|\boldsymbol{\Phi}_P\otimes\boldsymbol{\Phi}_P\|_{\text{op}} = \|n(1-\epsilon)\boldsymbol{\Gamma} + \boldsymbol{\Phi}_P\otimes\boldsymbol{\Phi}_P - n(1-\epsilon)\boldsymbol{\Gamma}\|_{\text{op}}\\ &\geq n(1-\epsilon)\lambda_{\max}(\boldsymbol{\Gamma}) - \|\boldsymbol{\Phi}_P\otimes\boldsymbol{\Phi}_P - n(1-\epsilon)\boldsymbol{\Gamma}\|_{\text{op}}\\ &\geq n(1-\epsilon)\lambda_{\max}(\boldsymbol{\Gamma}) - \sqrt{n(1-\epsilon)}\left(\sqrt{8}\lambda_{\max}(\boldsymbol{\Gamma}) + \sqrt{2P_k^2\log(2/\delta)}\right)\\ &\geq (1/2)n(1-\epsilon)\lambda_{\max}(\boldsymbol{\Gamma})\end{aligned}$$

when $n \geq (1-\epsilon)^{-1}\left(64 + 16\left(P_k/\lambda_{\max}(\boldsymbol{\Gamma})\right)^2\log(2/\delta)\right)$ with probability exceeding $1-\delta$.

Upper Bound.

$$\begin{aligned}\|\mathbf{K}\| &\leq n(1-\epsilon)\lambda_{\max}(\boldsymbol{\Gamma}) + \sqrt{n(1-\epsilon)}\left(\sqrt{8}\lambda_{\max}(\boldsymbol{\Gamma}) + \sqrt{2P_k^2\log(2/\delta)}\right) + n\epsilon Q_k\\ &\leq 2n(1-\epsilon)\lambda_{\max}(\boldsymbol{\Gamma}) + n\epsilon Q_k\end{aligned}$$

when $n \geq (1-\epsilon)^{-1}\left(16 + 4\left(P_k/\|\boldsymbol{\Gamma}\|_{\text{op}}\right)^2\log(2/\delta)\right)$. 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,

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

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

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

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

Let f^*_{TP} represent the minimizer of G, i.e. $f^*_{\mathrm{TP}} \triangleq \mathbf{\Phi}_{\mathrm{TP}} \left(\mathbf{K}_{\mathrm{TP}} + \frac{C}{2} \mathbf{I} \right)^{-1} \mathbf{y}$. It then follows by the strong convexity of the KRR problem that $\nabla G(f^*_{\mathrm{TP}}) = \mathbf{0}$. Then, we have

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

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

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

Expanding out ∇G, we have

$$\begin{aligned}
&\nabla G(f^{(t)}) - \nabla G(f^*) \\
&= \frac{2\eta}{n(1-\epsilon)} \left(\sum_{i \in S^{(t)} \cap P} (f^{(t)}(\mathbf{x}_i) - y_i) \cdot \phi(\mathbf{x}_i) \right) + C f^{(t)} - \frac{2\eta}{n(1-\epsilon)} \left(\sum_{i \in S^{(t)} \cap P} (f^*(\mathbf{x}_i) - y_i) \cdot \phi(\mathbf{x}_i) \right) + C f^* \\
&= \frac{2\eta}{n(1-\epsilon)} \left(\left[\sum_{i \in S^{(t)} \cap P} \phi(\mathbf{x}_i) \otimes \phi(\mathbf{x}_i) \right] (f^{(t)} - f^*) \right) + C(f^{(t)} - f^*)
\end{aligned}$$

We then have from the convexity of G and noting G is smooth with constant $\|\mathbf{K}_{\mathrm{TP}}\|$,

$$2\eta \left\langle f^{(t)} - f^*, \nabla G(f^{(t)}) - \nabla G(f^*) \right\rangle_{\mathcal{H}} \geq 2C\eta \left\| f^{(t)} - f^* \right\|_{\mathcal{H}}^2 + \frac{\eta}{2 \left\| \mathbf{K}_{\mathrm{TP}} \right\|} \left\| \nabla G(f^{(t)}) - G(f^*) \right\|_{\mathcal{H}}^2 \tag{5}$$

Then, by choosing $\eta \leq (1/2 \left\| \mathbf{K}_{\mathrm{TP}} \right\|)$, we have from Equations 4 and 5,

$$\left\| f^{(t)} - f^* - \eta \nabla G(f^{(t)}) + \eta \nabla G(f^*) \right\|_{\mathcal{H}} \leq \left\| f^{(t)} - f^* \right\|_{\mathcal{H}} \sqrt{1 - 2\eta C}$$

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

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

$$\leq \frac{8\eta^2}{[n(1-\epsilon)]^2}\left(\|\mathbf{K}_{\mathrm{FP}}\|\,\|\mathbf{K}_{\mathrm{FN}}\|\,\left\|f^{(t)}-f^*\right\|_{\mathcal{H}}^2+\|\boldsymbol{\mu}_{\mathrm{FN}}\|_2^2\right)$$

where in the first and third inequalities we utilize Lemma 22. Concatenate the Gaussian noise random variables, μ_i into a vector $\boldsymbol{\mu}$, that we further partition into $\boldsymbol{\mu} \triangleq [\boldsymbol{\mu}_{\mathrm{TP}}, \boldsymbol{\mu}_{\mathrm{FN}}]$. For the next term, we have

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

We now bound $\|f^*-f_{\mathrm{TP}}^*\|$.

$$\|f^*-f_{\mathrm{TP}}^*\|_{\mathcal{H}} \stackrel{\mathrm{def}}{=} \left\|f^*-\boldsymbol{\Phi}_{\mathrm{TP}}\left(\mathbf{K}_{\mathrm{TP}}+(C/2)\mathbf{I}\right)^{-1}\mathbf{y}_{\mathrm{TP}}\right\|_{\mathcal{H}} = \left\|f^*-\boldsymbol{\Phi}_{\mathrm{TP}}\left(\mathbf{K}_{\mathrm{TP}}+(C/2)\mathbf{I}\right)^{-1}\left(\boldsymbol{\Phi}_{\mathrm{TP}}f^*+\boldsymbol{\mu}\right)\right\|_{\mathcal{H}}$$

Since our choice of C is constant, if we are to increase n, then we should converge to optimal as $n\to\infty$ then I hypothesize $\boldsymbol{\Phi}_{\mathrm{TP}}(\mathbf{K}_{\mathrm{TP}}+(C/2)\mathbf{I})^{-1}\mathbf{y}_{\mathrm{TP}} \to \boldsymbol{\Phi}_{\mathrm{TP}}(\mathbf{K}_{\mathrm{TP}})^{-1}\mathbf{y}_{\mathrm{TP}}$, which is the 'pseudoinverse' in the Hilbert Space (Requires Proof). We then obtain,

$$\begin{aligned}\left\|f^{(t+1)}-f^*\right\|_{\mathcal{H}} \leq \left\|f^{(t)}-f^*\right\|_{\mathcal{H}}&\left(1-\eta\left(C-\frac{\sqrt{8\,\|\mathbf{K}_{\mathrm{FP}}\|\,\|\mathbf{K}_{\mathrm{FN}}\|}}{n(1-\epsilon)}\right)\right)\\ &+\frac{\eta}{n(1-\epsilon)}\left(4\sigma\sqrt{n\epsilon}+4\sigma\sqrt{n(1-\epsilon)\,\|\mathbf{K}_{\mathrm{TP}}\|}\right)+\sqrt{8}C\,\|f^*\|_{\mathcal{H}}\end{aligned}$$

Then, noting that with probability exceeding $1-\delta$, when $n \geq (1-\epsilon)^{-1}\left(16\,\|\boldsymbol{\Gamma}\|_{\mathrm{op}}^2+2P_k^2\log(2/\delta)\right)$, we have $\|\mathbf{K}_{\mathrm{FN}}\| \leq 2\,\|\boldsymbol{\Gamma}\|_{\mathrm{op}}\,n(1-\epsilon)$, and assuming $\|\mathbf{K}_{\mathrm{FP}}\| \leq n\epsilon Q_k$. We choose $C \geq 5\sqrt{Q_k\,\|\boldsymbol{\Gamma}\|_{\mathrm{op}}}$, and $\eta \leq \left[n(1-\epsilon)\sqrt{\|\boldsymbol{\Gamma}_{\mathrm{op}}\|\,(Q_k\vee\|\boldsymbol{\Gamma}\|_{\mathrm{op}})}\right]^{-1}$.

$$\begin{aligned}\left\|f^{(t+1)}-f^*\right\|_{\mathcal{H}} \leq \left\|f^{(t)}-f^*\right\|_{\mathcal{H}}&\left(1-\frac{1}{n(1-\epsilon)}\right)+\frac{\sqrt{8}\sigma}{(Q_k\vee\|\boldsymbol{\Gamma}\|_{\mathrm{op}})n(1-\epsilon)}\\ &+\frac{4\sigma}{(Q_k\vee\|\boldsymbol{\Gamma}\|_{\mathrm{op}})\,\|\boldsymbol{\Gamma}\|_{\mathrm{op}}\,[n(1-\epsilon)]^{3/2}}+\frac{\sqrt{8}C\,\|f^*\|_{\mathcal{H}}}{(Q_k\vee\|\boldsymbol{\Gamma}\|_{\mathrm{op}})\,\|\boldsymbol{\Gamma}\|_{\mathrm{op}}\,[n(1-\epsilon)]^2}\end{aligned}$$

It then follows after T iterations,

$$\begin{aligned}\left\|f^{(t+1)}-f^*\right\|_{\mathcal{H}} \leq \left\|f^{(t)}-f^*\right\|_{\mathcal{H}}&\left(1-\frac{1}{n(1-\epsilon)}\right)^T+\frac{\sqrt{8\sigma}}{(Q_k\vee\|\boldsymbol{\Gamma}\|_{\mathrm{op}})\,\|\boldsymbol{\Gamma}\|_{\mathrm{op}}}\\ &+\frac{4\sigma}{(Q_k\vee\|\boldsymbol{\Gamma}\|_{\mathrm{op}})\,\|\boldsymbol{\Gamma}\|_{\mathrm{op}}\,\sqrt{n(1-\epsilon)}}+\frac{\sqrt{2}C\,\|f^*\|_{\mathcal{H}}}{(Q_k\vee\|\boldsymbol{\Gamma}\|_{\mathrm{op}})\,\|\boldsymbol{\Gamma}\|_{\mathrm{op}}\,n(1-\epsilon)}\end{aligned}$$

Then, we obtain the claimed bound after $T = O\left(n(1-\epsilon)\log\left(\frac{\|f^*\|_{\mathcal{H}}}{\varepsilon}\right)\right)$ iterations. ■

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 Subquantile Smoothness

Lemma 24. *(β-Smoothness of $g(t,f)$ w.r.t f). Let $\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n$ represent the rows of the data matrix $\mathbf{X}$. It then follows:*

$$\left\|\nabla_f g(t,f) - \nabla_f g(t,\hat{f})\right\|_{\mathcal{H}} \leq \beta \left\|f - \hat{f}\right\|_{\mathcal{H}}$$

where $\beta = \frac{1}{4n(1-\epsilon)} \|\mathbf{K}\|$

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

$$\begin{aligned}
&\left\|\nabla_f^2 g(t,f)\right\|_{\mathcal{H}} \\
&\quad= \frac{1}{n(1-\epsilon)} \left\| \sum_{i=1}^{n} \mathbb{I}\left\{t \geq (1-y_i)\log(f^{(t)}(\mathbf{x}_i))\right\} \sigma(f^{(t)}(\mathbf{x}_i)) \left(1 - \sigma(f^{(t)}(\mathbf{x}_i))\right) \phi(\mathbf{x}_i) \otimes \phi(\mathbf{x}_i) \right\|_{\mathcal{H}} \\
&\quad\leq \frac{1}{n(1-\epsilon)} \max_{i\in[n]} \left| \sigma(f^{(t)}(\mathbf{x}_i)) \left(1 - \sigma(f^{(t)}(\mathbf{x}_i))\right) \right| \left\| \sum_{i=1}^{n} \phi(\mathbf{x}_i) \otimes \phi(\mathbf{x}_i) \right\|_{\mathcal{H}} \\
&\quad\overset{(i)}{\leq} \frac{1}{4n(1-\epsilon)} \|\mathbf{\Phi} \otimes \mathbf{\Phi}\|_{\text{op}} = \frac{1}{4n(1-\epsilon)} \|\mathbf{K}\|
\end{aligned}$$

(i) follows as for a scaler $\alpha \in [0,1]$, the maximum value of $\alpha(1-\alpha)$ is obtained at $\frac{1}{4}$, we also let $\mathbf{\Phi} \in \mathbb{R}^{\infty \times n}$ be the quasimatrix representing the concatentation of the RKHS functions. This completes the proof. ■

D.2 Proof of Theorem 10

Proof. From Algorithm 2, we have for kernelized binary classification,

$$f^{(t+1)} = \text{Proj}_{\mathcal{K}}\left[f^{(t)} - \frac{\eta}{n(1-\epsilon)} \left(\sum_{i \in S^{(t)}} \left(\sigma(f^{(t)}(\mathbf{x}_i)) - y_i \right) \cdot \phi(\mathbf{x}_i) + C f^{(t)} \right) \right] \tag{6}$$

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

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

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

Then, we have from the previous section,

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

We then have,

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

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

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

We thus have for $\eta \leq 2/\|\mathbf{K}_{\text{TP}}\|$,

$$\left\|f^{(t)} - f^* - \eta\nabla G(f^{(t)}) + \eta\nabla G(f^*)\right\|_{\mathcal{H}} \leq \left\|f^{(t)} - f^*\right\|_{\mathcal{H}}\sqrt{1-2\eta C}$$

We now analyze the error term by term individually.

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

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

$$\sum_{i\in S^{(t)}\cap P}(f^{(t)}(\mathbf{x}_i) - f^*(\mathbf{x}_i))^2 = \left\langle\sum_{i\in S^{(t)}\cap P}\phi(\mathbf{x}_i)\otimes\phi(\mathbf{x}_i), (f^{(t)} - f^*)\otimes(f^{(t)} - f^*)\right\rangle_{\text{HS}} \leq \|\mathbf{K}_{\text{P}}\|\left\|f^{(t)} - f^*\right\|_{\mathcal{H}}^2$$

Combining the previous two results, we have

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

Now we will analyze the second term in the error.

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

Then, we have

$$\begin{aligned}
\left\|f^{(t+1)} - f^*\right\|_{\mathcal{H}} \leq \left\|f^{(t)} - f^*\right\|_{\mathcal{H}}&\left(1 - \eta C + \frac{\sqrt{2}\eta}{n(1-\epsilon)}\sqrt{\|\mathbf{K}_{\text{FP}}\|\,\|\mathbf{K}_{\text{P}}\|}\right) \\
&+ \frac{\eta}{n(1-\epsilon)}\left(\sqrt{8\,\|\mathbf{K}_{\text{TP}}\|\,\mathcal{E}_{\text{OPT}}} + C\left\|f^*\right\|_{\mathcal{H}}\right)
\end{aligned}$$

With sufficient data we can obtain $(1/2)n(1-\epsilon)\|\mathbf{\Gamma}\|_{\text{op}} \leq \|\mathbf{K}_{\text{TP}}\| \leq 2n(1-\epsilon)\|\mathbf{\Gamma}\|_{\text{op}}$. We then assume $n\epsilon Q_m \leq \|\mathbf{K}_{\text{FP}}\| \leq n\epsilon Q_k$, choosing $C = 3\sqrt{\|\mathbf{\Gamma}\|_{\text{op}}\,Q_k}$, and our choice for η, we have

$$\left\|f^{(t+1)} - f^*\right\|_{\mathcal{H}} \leq \left\|f^{(t)} - f^*\right\|_{\mathcal{H}}\left(1 - \frac{1}{n(1-\epsilon)}\right) + \frac{\sqrt{18\mathcal{E}_{\text{OPT}}}}{n(1-\epsilon)\sqrt{\|\mathbf{K}_{\text{FP}}\|}} + \frac{2C\left\|f^*\right\|_{\mathcal{H}}}{n(1-\epsilon)\sqrt{\|\mathbf{K}_{\text{FP}}\|\,\|\mathbf{K}_{\text{P}}\|}}$$

Solving the recursion, we obtain after T steps,

$$\left\|f^{(T)} - f^*\right\|_{\mathcal{H}} \leq \left\|f^{(0)} - f^*\right\|_{\mathcal{H}}\left(1 - \frac{1}{n(1-\epsilon)}\right)^T + \sqrt{\frac{18\mathcal{E}_{\text{OPT}}}{n(1-\epsilon)Q_k}} + \frac{\sqrt{2}\left\|f^*\right\|_{\mathcal{H}}}{\|\mathbf{\Gamma}\|_{\text{op}}\sqrt{Q_k n(1-\epsilon)}}$$

Then, after $T = O\left(n(1-\epsilon)\log\frac{\|f^*\|_{\mathcal{H}}}{\varepsilon}\right)$ iterations we obtain the claimed bound. ■