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

author names withheld

Editor: Under Review for COLT 2024

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 find the function $f^* = \mathbb{E}_{\mathcal{D}\sim\mathcal{P}}[\ell(f;\mathcal{D})]$, from the noisy distribution, $\hat{\mathcal{P}}$. Superquantile objectives have been been studied extensively to reduce the risk of the tail Laguel et al. (2021); Rockafellar et al. (2014). We consider the contrasting case where we want to minimize the body of the risk. To our knowledge, we are the first to study the problem of general kernel learning in the Huber Contamination Model. We study a gradient-descent approach to solve a variational representation of the Subquantile Objective.

1. Introduction

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

Definition 1 (Huber ϵ-Contamination Model Huber and Ronchetti (2009)) *Given a corruption parameter $0 < \epsilon < 0.5$, a data matrix, $\mathbf{X}$ and labels $\mathbf{y}$. An adversary is allowed to inspect all samples and modify ϵn samples arbitrarily. The algorithm is then given the ϵ-corrupted data matrix $\mathbf{X}$ and 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, (Li et al., 2021). These robust objectives tend to not be convex and therefore do not have a strong analysis on the error bounds for general classes of models.

We similarly propose a robust objective which has a nonconvex-concave objective. This objective function has also been proposed recently in Hu et al. (2020) where there has been an analysis in

the Binary Classification Task. We show Subquantile Minimization reduces to the same objective function given in Hu et al. (2020).

The study of Kernel Learning in the Gaussian Design is quite popular, (Cui et al., 2021; Dicker, 2016). In (Cui et al., 2021), the feature space, $\phi(\mathbf{x}_i) \sim \mathcal{N}(0, \mathbf{\Sigma})$ where $\mathbf{\Sigma}$ is a diagonal matrix of dimension p, where p can be infinite. In this work, we adopt a more general framework, the sub-Gaussian Design where $\mathbf{E}_{X\sim\mathcal{P}}[X \otimes X] = \mathbf{\Gamma}$, and with the power of Mercer's Theorem (Mercer, 1909), we are able to say $\mathrm{Tr}(\mathbf{\Gamma}) < \infty$.

Theorem 2 *(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[1] corrupted by an adversary. Then in polynomial number of iterations we obtain the following approximation bounds.*
Kernelized Regression:

$$\|\hat{f} - f^*\|_{\mathcal{H}}^2 \leq \varepsilon + O\left(\frac{\sigma \log n \log(1/2\delta)\sqrt{\mathrm{Tr}(\mathbf{\Gamma})}}{\lambda_m(\mathbf{\Gamma})\left[n(1-\epsilon)\right]^{1/4}}\right) + O\left(\frac{\|\mathbf{y}\|_\infty^2 \left(\mathrm{Tr}(\mathbf{\Gamma})\epsilon \log \epsilon^{-1} + Q_k\right)\|f^*\|_{\mathcal{H}}}{\lambda_m(\mathbf{\Gamma})\ell\sqrt{n(1-\epsilon)}}\right) + \left(2 + O\left(\frac{\ell}{\lambda_m(\mathbf{\Gamma})}\right)\right)\|\mathrm{Proj}_{\Psi_m^\perp} f^*\|_{\mathcal{H}}$$

Kernel Binary Classification:

$$\|\hat{f} - f^*\|_{\mathcal{H}}^2 \leq \varepsilon + \|\mathrm{Proj}_{\Psi_m^\perp} f^*\|_{\mathcal{H}}\left(2 + \frac{\sqrt{(8/C)\,\mathrm{Tr}(\mathbf{\Gamma})\epsilon \log \epsilon^{-1}}}{\left[n(1-\epsilon)\right]^{1/4}}\right) + O\left(\frac{\ell}{[n(1-\epsilon)]}\right)$$

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 guarantees hold when ϵn elements of $\mathbf{y}$ are adversarially corrupted and ϵn vectors of $\mathbf{X}$ are adversarially corrupted such that $\left\|\sum_{i\in Q}\phi(\mathbf{x}_i)\right\|_{\mathcal{H}} = O(\sqrt{n})$, where $\epsilon < 1/2$ is dependent on $\sup_{x\in\mathcal{P}}\|\phi(\mathbf{x})\|_{\mathcal{H}}$ and the eigenvalues of the sub-Gaussian distribution of the covariates. Our proof techniques extend Bhatia et al. (2015); Awasthi et al. (2022) 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 $a \wedge b \triangleq \min(a, b)$.

1. Our only condition on the corrupted covariates is there exists a positive constant such that $\left\|\sum_{i\in Q}\phi(\mathbf{x}_i)\right\|_{\mathcal{H}} = O(n^{3/4})$

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 (Ryan and a Ryan, 2002). 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. Distribution

In this paper we sample $\mathbf{x} \sim \mathcal{X}$ such that $\phi(\mathbf{x}) \sim \mathcal{P}$ is sub-Gaussian in the Hilbert Space where $\mathbf{E}[\phi(\mathbf{x}_i)] = \mathbf{0}$ and $\mathbf{E}[\phi(\mathbf{x}_i) \otimes \phi(\mathbf{x}_i)] = \mathbf{\Gamma}$ where $\mathrm{Tr}(\mathbf{\Gamma}) < \infty$. We have X is a centered Hilbert Space sub-Gaussian random function if for all $\theta > 0$,

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

The Gaussian Design for the Feature Space has gained popularity in the study of kernel learning (Cui et al., 2021).

2.4. Related Work

The idea of iterative thresholding algorithms for robust learning tasks dates back to 1806 by Legendre (Legendre, 1806). Iterative thresholding have been studied theoretically and tested empirically in various machine learning domains (Hu et al., 2023; Mukhoty et al., 2019). Therefore, we will dedicate this subsection to reviewing such works and to make clear our contributions to the iterative thresholding literature.

Bhatia et al. (2015) 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 Awasthi et al. (2022), where the iterative trimmed maximum likelihood estimator is studied for General Linear Models. The algorithm studied by Awasthi et al. (2022) 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 (Woodruff et al., 2014) the input matrix to have well-conditioned covariates. Thus, we are left with $\mathbf{\Phi}$ where

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. Townsend and Trefethen (2015)).

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

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

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 (Razaviyayn et al., 2020). Let $\mathrm{Proj}_{\mathcal{K}}$ be the projection of a function on to the convex set $\mathcal{K} \triangleq \{f \in \mathcal{H} : \|f\|_{\mathcal{H}} \leq R\}$, then our update steps are

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

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

The proof of convergence for the above algorithm was given in Jin et al. (2020)[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.

Lemma 3 *Let $\mathcal{R} : \mathcal{H} \times \mathcal{D} \to \mathbb{R}$ be a loss function (not necesarily convex). Let $\mathbf{x}_1, \cdots, \mathbf{x}_n$ denote the n data points ordered WLOG such that $\mathcal{R}(f;\mathbf{x}_1,y_1) \leq \cdots \leq \mathcal{R}(f;\mathbf{x}_n,y_n)$. If we denote $\hat{\nu}_i \triangleq \mathcal{R}(f;\mathbf{x}_i,y_i)$, 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 3, 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 4 *Let $\hat{\nu}_i \triangleq \mathcal{R}(f;\mathbf{x}_i,y_i)$ s.t. $\hat{\nu}_{i-1} \leq \hat{\nu}_i \leq \hat{\nu}_{i+1}$, if we choose $t^{(k+1)} = \hat{\nu}_{n(1-\epsilon)}$ as by Lemma 3, it then follows $\nabla_{\mathbf{w}} g(t^{(k)}, f^{(k)}) = \frac{1}{n(1-\epsilon)}\sum_{i=1}^{n(1-\epsilon)}\nabla f(\mathbf{x}_i; f^{(k)}, y_i)$*

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

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

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

$$\nabla_f g(t^{(k+1)}, f^{(k)}) = \frac{1}{n(1-\epsilon)}\sum_{i=1}^{n(1-\epsilon)}\nabla_f\mathcal{R}(f^{(k)};\mathbf{x}_i,y_i)$$

This concludes the proof. ■

We have therefore shown that the two-step optimization of Subquantile Minimization is equivalent to solving 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

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

Input: Data Matrix: $\mathbf{X} \in \mathbb{R}^{n \times d}, n \gg d$; Labels: $\mathbf{y} \in \mathbb{R}^{n|}$, Closed and Convex set $\mathcal{K} \subset \mathcal{H}$
Output: Function in $\mathcal{H}$: $\hat{f}$

1. Calculate the ℓ-smoothness constant, $\ell = \frac{1}{n(1-\epsilon)} \|\mathbf{K}\|$
2. Set the step-size $\eta = 1/\ell$
3. Set the number of iterations

$$T = O\left(\log\left(\left(\|f^*\|_{\mathcal{H}}^2 + \|\mathbf{y}\|^2\right)\frac{2}{\varepsilon}\right)\right)$$

4. **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 Projected Gradient Descent Iteration with Lemma 20.

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

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

Algorithm 1: Subquantile Minimization for Kernelized Regression

We will now give the algorithm. 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. Instead, we will calculate the approximation bounds for the rank-m approximation of f^* and push $m \to \infty$.

Theorem 5 (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 $1-\delta$,*

$$\|\hat{f} - f^*\|_{\mathcal{H}}^2 \leq \varepsilon + O\left(\frac{\sigma \log n \log(1/2\delta)\sqrt{\mathrm{Tr}(\mathbf{\Gamma})}}{\lambda_m(\mathbf{\Gamma})\left[n(1-\epsilon)\right]^{1/4}}\right) + O\left(\frac{\|\mathbf{y}\|_\infty^2 \left(\mathrm{Tr}(\mathbf{\Gamma})\epsilon \log \epsilon^{-1} + Q_k\right)\|f^*\|_{\mathcal{H}}}{\lambda_m(\mathbf{\Gamma})\ell\sqrt{n(1-\epsilon)}}\right)$$
$$+ \left(2 + O\left(\frac{\ell}{\lambda_m(\mathbf{\Gamma})}\right)\right)\left\|\mathrm{Proj}_{\Psi_m^\perp} f^*\right\|_{\mathcal{H}}$$

when $n \geq (1-2\epsilon)^{-1}\left(256\|\mathbf{\Gamma}\|_{\mathrm{op}}^2 + 64P_k^2 \log(2/\delta)\right)$ *and* $\epsilon \leq 8^{-1/2}\left(\lambda_m(\mathbf{\Gamma})/P_k\right)^2$.

Proof Sketch. From the triangle inequality, we have

$$\|f^{(t)} - f^*\|_{\mathcal{H}}^2 \leq 2\|f^{(t)} - \text{Proj}_{\Psi_m} f^*\|_{\mathcal{H}}^2 + 2\|\text{Proj}_{\Psi_m^\perp} f^*\|_{\mathcal{H}}^2$$

It thus suffices to prove $\left\|f^{(t)} - \text{Proj}_{\Psi_m} f^*\right\|_{\mathcal{H}}^2 \leq (\varepsilon/2) + E$ for some $E \geq 0$. ■

Full proof with explicit constants is given in Appendix C.2. A direct application of Theorem 5 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 Schmidt et al. (2018).

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

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

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

Theorem 6 (Subquantile Minimization for Binary Classification is Good with High Probability) *Let Algorithm 2 be run on a dataset $\mathcal{D} \sim \hat{\mathcal{P}}$ with learning rate $\eta \triangleq \Omega(\ell^{-1})$. Then after $O\left(\log\left(\frac{\|f^*\|_{\mathcal{H}}}{\varepsilon}\right)\right)$ gradient descent iterations, with probability exceeding $1 - \delta$ and a positive constant C,*

$$\|f^{(T)} - f^*\|_{\mathcal{H}}^2 \leq \varepsilon + \|\text{Proj}_{\Psi_m^\perp} f^*\|_{\mathcal{H}} \left(2 + \frac{\sqrt{(8/C)\,\text{Tr}(\mathbf{\Gamma})\epsilon \log \epsilon^{-1}}}{[n(1-\epsilon)]^{1/4}}\right) + O\left(\frac{\ell}{[n(1-\epsilon)]}\right)$$

for $n \geq (1-\epsilon)^{-1} 64^4 C^{-4} \zeta^{-4} \lambda_m^{-4}(\mathbf{\Gamma}) \ell^{-4} 4 \left([\text{Tr}(\mathbf{\Gamma})]^2 \epsilon \log \epsilon^{-1} + Q_k^2\right)$.

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

From the sufficient data requirement given in Theorem 6, we can see Kernelized Binary Classification with Binary Cross-Entropy Loss is *consistent*. As $n \to \infty$, the corresponding m we can use also increases as the eigenvalues are decreasing.

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 do not require to work with projections over the orthonormal basis of functions in the Hilbert Space so many residual terms vanish.

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,

Input: Data Matrix: $\mathbf{X} \in \mathbb{R}^{n \times d}, n \gg d$; Labels: $\mathbf{y} \in \mathbb{R}^n$, Hilbert Space Norm Ball $\mathcal{K} \subset \mathcal{H}$
Output: Function in $\mathcal{H}$: $\hat{f}$

1. Calculate the ℓ-smoothness constant, $\ell = \frac{1}{n(1-\epsilon)} \|\mathbf{K}\|$
2. Set the step-size $\eta = 1/\ell$
3. Set the number of iterations

$$T = O\left(\log\left(\frac{2\|f^*\|_{\mathcal{H}}^2}{\varepsilon}\right)\right)$$

4. **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.

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

    5. Perform Projected Gradient Descent Iteration with Lemma 20.

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

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

Algorithm 2: Subquantile Minimization for Binary Classification

$\mathcal{P}$, 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

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

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.

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.

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.

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. doi: 10.3150/20-BEJ1251.

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.

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.

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

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.

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

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.

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

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.

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. ISSN 0001-0782. doi: 10.1145/358669.358692.

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

Arthur Gretton. Notes on mean embeddings and covariance operators, 2015.

Thomas Hofmann, Bernhard Schölkopf, and Alexander J. Smola. Kernel methods in machine learning. *The Annals of Statistics*, 36(3):1171 – 1220, 2008. doi: 10.1214/009053607000000677.

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

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.

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

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

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

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.

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.

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

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. ISSN 1877-0541. doi: 10.1007/s11228-021-00609-w.

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.

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

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

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

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.

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.

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.

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

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. doi: 10.1109/MSP.2020.3003851.

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. ISSN 0377-2217. doi: https://doi.org/10.1016/j.ejor.2013.10.046.

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

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.

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.

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.

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

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.

William Henry Young. On classes of summable functions and their fourier series. *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character*, 87(594):225–229, 1912.

Appendix A. Probability Theory

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

Assumption 7 (Gaussian Design) *We assume for $\mathbf{x}_i \sim \mathcal{P} \in \mathcal{X}$, then it follows for the feature map, $\phi(\cdot) : \mathcal{X} \to \mathcal{H}$,*

$$\mathbf{x}_i \sim \mathcal{P}$$

where $\mathbf{\Gamma}$ is a possibly infinite dimensional covariance operator.

Assumption 8 (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 9 (Normal Residuals) *The residual is defined as $\mu_i \triangleq f^*(\mathbf{x}_i) - y_i$. Then we assume for some $\sigma > 0$, it follows*

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

A.1. Finite Dimensional Concentrations of Measure

Proposition 10 *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 11 *Let $\mu_1, \ldots, \mu_n \sim \mathcal{N}(0, \sigma^2)$ for some $\sigma > 0$, then it follows for any $s \geq 1$,*

$$\mathbf{Pr}\left\{\sum_{i=1}^{n} \mu_i^2 \geq 8n\sigma^2 \cdot s\right\} \leq 4e^{-s}$$

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

$$\mathop{\mathbf{Pr}}_{\boldsymbol{\mu} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})}\left\{\|\boldsymbol{\mu}\|^2 \geq C \cdot s\right\} = \mathop{\mathbf{Pr}}_{\boldsymbol{\mu} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})}\left\{\|\boldsymbol{\mu}\| \geq \sqrt{C \cdot s}\right\} \leq 4\exp\left(-\frac{C \cdot s}{8\,\mathrm{Tr}\,(\mathbf{\Sigma})}\right)$$

where the last inequality follows from Proposition **??**. Now choosing $C \triangleq 8\,\mathrm{Tr}(\mathbf{\Sigma})$ completes the proof. ■

A.2. Hilbert Space Concentrations of Measure

Lemma 12 (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 $\theta > 0$,

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

where (i) follows from Hölder's Inequality for a product of functions (Hölder, 1889). We see the resultant variance proxy is $n\mathbf{\Gamma}$ and the proof is complete. ■

Theorem 13 (Hilbert Space Hanson Wright (Chen and Yang, 2021)) *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 $t > 0$,*

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

From Theorem 13, it follows that the LHS is less than $\delta \in (0,1)$ when

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

Furthermore, we have when

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

it follows

$$t \geq (n/C)\|\mathbf{\Gamma}\|_{\mathrm{op}} \log(2/\delta)$$

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 14 (Sum of Binomial Coefficients (Cormen et al., 2022)) *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 15 (Jensen's Inequality (Jensen, 1906)) *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 16 (RKHS Norm of Functions in the Reproducing Kernel Hilbert Space) *Let $\mathbf{x}_i \sim \mathcal{P}$ such that $\mathbf{x}_i \sim \mathcal{P}$ (Assumption 7). Denote $\mathcal{S}$ as all subsets of $[n(1-\epsilon)]$ with size $n(1-2\epsilon)$ for $\epsilon < 0.5$ and $P_B = \sum_{i=1}^{n(1-\epsilon)} \delta_{X_i}$ where δ is a Dirac measure at X_i. Then it follows with probability exceeding $1-\delta$,*

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

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 12, we have that for any $B \in \mathcal{S}$, $\sum_{i \in B} \phi(\mathbf{x})$ is sub-Gaussian with proxy $n\Gamma$. We will then apply a union bound over $\mathcal{S}$ to give the claimed bound.

$$\underset{Y \sim \mathcal{SG}(n\mathbf{\Gamma})}{\mathbf{Pr}} \left\{ \|Y\|_{\mathcal{H}}^2 \geq n \operatorname{Tr}(\mathbf{\Gamma}) + t \right\} \leq 2 \exp\left[-C \min\left\{ \frac{t^2}{n^2 \left\|\mathbf{\Gamma}\right\|_{\mathrm{HS}}^2}, \frac{t}{n \left\|\mathbf{\Gamma}\right\|_{\mathrm{op}}} \right\} \right]$$

Then for a sufficiently large δ, we have with probability $1-\delta$.

$$\|Y\|_{\mathcal{H}}^2 \geq n \operatorname{Tr}(\mathbf{\Gamma}) + n \left\|\mathbf{\Gamma}\right\|_{\mathrm{HS}} \sqrt{(1/C)\ln(2/\delta)}$$

Then with Fact 14, we have with a union bound,

$$\|Y\|_{\mathcal{H}}^2 \geq n \operatorname{Tr}(\mathbf{\Gamma}) + n \left\|\mathbf{\Gamma}\right\|_{\mathrm{HS}} \sqrt{(1/C)\ln(2/\delta) + (2n/C)\left(\epsilon \log \epsilon^{-1}\right)}$$

Simplifying the bound with the sufficient data requirement in the Proposition statement completes the proof. ■

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

Proposition 17 (McDiarmid's Inequality (McDiarmid et al., 1989)) *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 18 (Mean Estimation in the Hilbert Space (Tolstikhin et al., 2017)) *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 Tolstikhin et al. (2017) by using knowledge of the distribution to first derive the expectation.

Proposition 19 (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 7). 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 $1-\delta$,*

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

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

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.

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

In (ii) we apply a union bound. In (ii) we note that $\phi(\mathbf{x}_i) \otimes \phi(\mathbf{x}_i) - \mathbf{\Gamma}$ 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 $\mathbf{\Gamma}$. By expanding out the Hilbert-Schmidt Norm, we then have

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

$$\overset{(iv)}{\leq} \frac{2}{n(1-\epsilon)}\left(\mathop{\mathbf{E}}_{\phi(\mathbf{x}_i)\sim\mathcal{P}}\sum_{k=1}^{p}\sum_{i=1}^{n(1-\epsilon)}\langle\phi(\mathbf{x}_i)\otimes\phi(\mathbf{x}_i)e_k,\phi(\mathbf{x}_i)\otimes\phi(\mathbf{x}_i)e_k\rangle_{\text{HS}}\right)^{1/2}$$

$$= \frac{2}{n(1-\epsilon)}\left(\sum_{i=1}^{n(1-\epsilon)}\mathop{\mathbf{E}}_{\phi(\mathbf{x}_i)\sim\mathcal{P}}\|\phi(\mathbf{x}_i)\otimes\phi(\mathbf{x}_i)\|_{\text{HS}}^2\right)^{1/2}$$

$$\overset{(v)}{=} \frac{2}{\sqrt{n(1-\epsilon)}}\left(\mathop{\mathbf{E}}_{\phi(\mathbf{x}_i)\sim\mathcal{P}}\|\phi(\mathbf{x}_i)\|_{\mathcal{H}}^4\right)^{1/2}$$

(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

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

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

$$= \int_0^\infty \exp\left(-\frac{\sqrt{t}}{\|\mathbf{\Gamma}\|_{\text{op}}}\right)dt = 2\|\mathbf{\Gamma}\|_{\text{op}}^2$$

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

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

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

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

Then, we have from McDiarmid's inequality (Proposition 17),

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

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

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

Next, applying a union bound over $\mathcal{S}$ with Fact 14, we have

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

Simplifying the resultant bound completes the proof. ■

Appendix 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. Projection onto a Norm Ball

In this section we show normalizing on to a norm-ball in the RKHS can be implemented efficiently.

Lemma 20 *Let $\mathcal{K} \triangleq \{f : \|f\|_{\mathcal{H}} \leq R\}$. Then, for a $\hat{f} \notin \mathcal{K}$, it follows*

$$\mathrm{Proj}_{\mathcal{K}} \hat{f} = \left(\frac{R}{\|\hat{f}\|} \right) \hat{f}$$

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

$$\begin{aligned} \mathrm{Proj}_{\mathcal{K}} \hat{f} &= \arg\min_{f \in \mathcal{K}} \|f - \hat{f}\|_{\mathcal{H}}^2 = \arg\min_{f \in \mathcal{K}} \|f\|_{\mathcal{H}}^2 + \|\hat{f}\|_{\mathcal{H}}^2 - 2\langle f, \hat{f} \rangle_{\mathcal{H}} \\ &= \arg\min_{f \in \mathcal{K}} \|f\|_{\mathcal{H}}^2 - 2\langle f, \hat{f} \rangle_{\mathcal{H}} \end{aligned}$$

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

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

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

Appendix 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 21 *(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*

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

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

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

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

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

$$\begin{aligned}\left\|\nabla_f^2 g\left(t,f\right)\right\|_{\text{op}} &= \frac{2}{n(1-\epsilon)}\left\|\sum_{i=1}^{n}\mathbb{I}\left\{t \geq \ell\left(f;\mathbf{x}_i,y_i\right)\right\}\phi(\mathbf{x}_i)\otimes\phi(\mathbf{x}_i)\right\|_{\text{op}} \\ &\leq \frac{2}{n(1-\epsilon)}\left\|\sum_{i=1}^{n}\phi(\mathbf{x}_i)\otimes\phi(\mathbf{x}_i)\right\|_{\text{HS}} = \frac{2}{n(1-\epsilon)}\left\|\mathbf{\Phi}\otimes\mathbf{\Phi}\right\|_{\text{op}} = \frac{2}{n(1-\epsilon)}\left\|\mathbf{K}\right\|\end{aligned}$$

We will now give our probabilistic bounds using the first relation in our covariance estimation bound given in Proposition 19.

Lower Bound.

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

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

Upper Bound.

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

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

We now give the necessary results to prove our main approximation bound.

Proposition 22 (Young's Inequality (Young, 1912)) *For all $a, b \in \mathbb{R}$, it holds*

$$ab \leq \frac{a^2}{2} + \frac{b^2}{2}$$

Lemma 23 (Smooth Descent Lemma) *Suppose $\mu \leq \left\|\nabla^2 f(x)\right\|_{\text{op}} \leq \ell$ for all $x \in \mathcal{X}$, then for a stepsize $\eta \leq 1/\ell$, it follows for all $x \in \mathcal{X}$,*

$$f(x - \eta\nabla f(x)) \leq \left(1 - \frac{\mu}{\ell}\right)\left(f(x) - f^*(x)\right)$$

We are now ready to prove our main approximation bound.

C.2. Proof of Theorem 5

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

$$f^{(t+1)} = \text{Proj}_{\mathcal{K}}\left[f^{(t)} - \frac{2\eta}{n(1-\epsilon)}\sum_{i\in S^{(t)}}\left(f^{(t)}(\mathbf{x}_i) - y_i\right)\cdot\phi(\mathbf{x}_i) + Cf^{(t)}\right] \tag{2}$$

Next, we note that we can partition $S = (S \cap P) \cup (S \cap Q) \triangleq \text{TP} \cup \text{FP}$. Then we have

$$\begin{aligned}
\|f^{(t+1)} - f^*\|_{\mathcal{H}}^2 &= \|\text{Proj}_{\mathcal{K}}\left[f^{(t)} - \nabla_f g(f^{(t)}, t^*)\right] - f^*\|_{\mathcal{H}}^2 \\
&\overset{(i)}{\leq} \|f^{(t)} - \nabla_f g(f^{(t)}, t^*) - f^*\|_{\mathcal{H}}^2 \\
&\leq 2\|f^{(t)} - \nabla_f g(f^{(t)}, t^*) - \text{Proj}_{\Psi_m} f^*\|_{\mathcal{H}}^2 + 2\|\text{Proj}_{\Psi_m^\perp} f^*\|^2 \\
&= 2\|f^{(t)} - \text{Proj}_{\Psi_m} f^*\|_{\mathcal{H}}^2 - 4\eta \left\langle \nabla_f g(f^{(t)}, t^*), f^{(t)} - \text{Proj}_{\Psi_m} f^* \right\rangle_{\mathcal{H}} \\
&\qquad + 2\eta^2 \|\nabla_f g(f^{(t)}, t^*)\|_{\mathcal{H}}^2 + 2\|\text{Proj}_{\Psi_m^\perp} f^*\|^2
\end{aligned} \tag{3}$$

where (i) follows from noting the projection is a contraction. We will dedicate the rest of the proof to upper bounding the first three terms in Equation (3). We will first bound the second term in Equation (3) by splitting it into terms using the following relation,

$$\begin{aligned}
&2\eta \langle \nabla_f g(f^{(t)}, t^*), f^{(t)} - \text{Proj}_{\Psi_m} f^* \rangle_{\mathcal{H}} \\
&\overset{(2)}{=} 2\eta \left\langle f^{(t)} - \text{Proj}_{\Psi_m} f^*, \frac{2}{n(1-\epsilon)} \sum_{i \in S^{(t)}} (f^{(t)}(\mathbf{x}_i) - y_i) \cdot \phi(\mathbf{x}_i) + C f^{(t)} \right\rangle_{\mathcal{H}} \\
&\overset{(i)}{=} \frac{4\eta}{n(1-\epsilon)} \left\langle f^{(t)} - \text{Proj}_{\Psi_m} f^*, \sum_{i \in S^{(t)} \cap P} (f^{(t)}(\mathbf{x}_i) - f^*(\mathbf{x}_i) - \mu_i) \cdot \phi(\mathbf{x}_i) \right\rangle_{\mathcal{H}} \\
&+ \frac{4\eta}{n(1-\epsilon)} \left\langle f^{(t)} - \text{Proj}_{\Psi_m} f^*, \sum_{i \in S^{(t)} \cap Q} (f^{(t)}(\mathbf{x}_i) - y_i) \cdot \phi(\mathbf{x}_i) \right\rangle_{\mathcal{H}} \\
&\quad + \frac{4\eta}{n(1-\epsilon)} \left\langle f^{(t)} - \text{Proj}_{\Psi_m} f^*, C f^{(t)} \right\rangle_{\mathcal{H}}
\end{aligned} \tag{4}$$

where (i) follows from Theorem 7. We will now lower bound the first term of Equation (4).

$$\begin{aligned}
&\frac{4\eta}{n(1-\epsilon)} \left\langle f^{(t)} - \text{Proj}_{\Psi_m} f^*, \sum_{i \in S^{(t)} \cap P} (f^{(t)}(\mathbf{x}_i) - f^*(\mathbf{x}_i) - \mu_i) \cdot \phi(\mathbf{x}_i) \right\rangle_{\mathcal{H}} \\
&= \frac{4\eta}{n(1-\epsilon)} \left\langle f^{(t)} - \text{Proj}_{\Psi_m} f^*, \left[\sum_{i \in S^{(t)} \cap P} \phi(\mathbf{x}_i) \otimes \phi(\mathbf{x}_i) \right] \left(f^{(t)} - \text{Proj}_{\Psi_m} f^* \right) \right\rangle_{\mathcal{H}} \\
&\quad + \frac{4\eta}{n(1-\epsilon)} \left\langle f^{(t)} - \text{Proj}_{\Psi_m} f^*, \left[\sum_{i \in S^{(t)} \cap P} \phi(\mathbf{x}_i) \otimes \phi(\mathbf{x}_i) \right] \left(\text{Proj}_{\Psi_m^\perp} f^* \right) \right\rangle_{\mathcal{H}} \\
&\quad - \frac{4\eta}{n(1-\epsilon)} \left\langle f^{(t)} - \text{Proj}_{\Psi_m} f^*, \sum_{i \in S^{(t)} \cap P} \mu_i \phi(\mathbf{x}_i) \right\rangle_{\mathcal{H}} \\
&\overset{(ii)}{\geq} 8\eta \frac{(1-2\epsilon)}{(1-\epsilon)} \|f^{(t)} - \text{Proj}_{\Psi_m} f^*\|_{\mathcal{H}}^2 \left(\lambda_m(\mathbf{\Gamma}) - \left\| \frac{1}{n(1-\epsilon)} \sum_{i \in S^{(t)} \cap P} \phi(\mathbf{x}_i) \otimes \phi(\mathbf{x}_i) - \mathbf{\Gamma} \right\|_{\text{HS}} \right) \\
&\quad - \frac{4\eta^2}{n(1-\epsilon)} \|f^{(t)} - \text{Proj}_{\Psi_m} f^*\|_{\mathcal{H}}^2 \left\| \sum_{i \in S^{(t)} \cap P} \phi(\mathbf{x}_i) \otimes \phi(\mathbf{x}_i) \right\|_{\text{op}} \left\| \text{Proj}_{\Psi_m^\perp} f^* \right\|_{\mathcal{H}} \\
&\quad - \frac{1}{n(1-\epsilon)} \left\| \sum_{i \in S^{(t)} \cap P} \phi(\mathbf{x}_i) \otimes \phi(\mathbf{x}_i) \right\|_{\text{op}} \|\text{Proj}_{\Psi_m^\perp} f^*\|_{\mathcal{H}} - \frac{8\lambda_m(\mathbf{\Gamma})\eta(1-2\epsilon)}{(1-\epsilon)} \left\| \text{Proj}_{\Psi_m^\perp} f^{(t)} \right\|_{\mathcal{H}}^2
\end{aligned}$$

$$- \frac{4\eta}{n(1-\epsilon)}\|f^{(t)} - \text{Proj}_{\Psi_m} f^*\|_{\mathcal{H}}^2 \left\| \sum_{i \in S^{(t)} \cap P} \mu_i \phi(\mathbf{x}_i) \right\|_{\mathcal{H}} - \frac{1}{n(1-\epsilon)} \left\| \sum_{i \in S^{(t)} \cap P} \mu_i \phi(\mathbf{x}_i) \right\|_{\mathcal{H}} \tag{5}$$

where in (ii) we define $\tilde{n} \triangleq |S^{(t)} \cap P|$ and use Cauchy-Schwarz on the last two terms, we then have the simple inequality $\|\text{Proj}_{\Psi_m}[f^{(t)} - f^*]\|_{\mathcal{H}} = \|f^{(t)} - \text{Proj}_{\Psi_m} f^* - \text{Proj}_{\Psi_m^\perp} f^{(t)}\|_{\mathcal{H}}^2 \leq 2\|f^{(t)} - \text{Proj}_{\Psi_m} f^*\|_{\mathcal{H}}^2 + 2\|\text{Proj}_{\Psi_m^\perp} f^{(t)}\|_{\mathcal{H}}^2$ and split the final two terms with Young's Inequality (Proposition 22). We will now upper bound the second and third terms of Equation (4) together.

$$\begin{aligned}
&\frac{4\eta}{n(1-\epsilon)} \left\langle f^{(t)} - \text{Proj}_{\Psi_m} f^*, \sum_{i \in S^{(t)} \cap Q} (f^{(t)}(\mathbf{x}_i) - y_i) \cdot \phi(\mathbf{x}_i) \right\rangle_{\mathcal{H}} + \frac{4\eta}{n(1-\epsilon)} \left\langle f^{(t)} - \text{Proj}_{\Psi_m} f^*, C f^{(t)} \right\rangle_{\mathcal{H}} \\
&\leq \frac{4\eta}{n(1-\epsilon)} \|f^{(t)} - \text{Proj}_{\Psi_m} f^*\|_{\mathcal{H}} \left(\left\| \sum_{i \in S^{(t)} \cap Q} \phi(\mathbf{x}_i) \right\|_{\mathcal{H}} \sqrt{\sum_{i \in S^{(t)} \cap Q} (f^{(t)}(\mathbf{x}_i) - y_i)^2} + C\|f^{(t)}\|_{\mathcal{H}} \right) \\
&\overset{(iii)}{\leq} \frac{16C^2\eta}{[n(1-\epsilon)]^2} \|f^{(t)} - \text{Proj}_{\Psi_m} f^*\|_{\mathcal{H}}^2 \left\| \sum_{i \in S^{(t)} \cap Q} \phi(\mathbf{x}_i) \right\|_{\mathcal{H}}^2 + \eta \left(\sum_{i \in S^{(t)}} (f^{(t)}(\mathbf{x}_i) - y_i)^2 + C\|f^{(t)}\|_{\mathcal{H}}^2 \right)
\end{aligned} \tag{6}$$

where (iii) follows from Young's Inequality (Proposition 22) for a $\beta \in (0,1)$. We also assume $C \geq 1$. We will now upper bound the final term in Equation (3).

$$\begin{aligned}
\eta^2 \|\nabla_f g(f^{(t)}, t^*)\|_{\mathcal{H}}^2 &= \frac{4\eta^2}{n^2(1-\epsilon)^2} \left\| \sum_{i \in S^{(t)}} (f^{(t)}(\mathbf{x}_i) - y_i) \cdot \phi(\mathbf{x}_i) + C f^{(t)} \right\|_{\mathcal{H}}^2 \\
&\leq \frac{8C\eta^2}{n^2(1-\epsilon)^2} \left\| \sum_{i \in S^{(t)}} \phi(\mathbf{x}_i) \right\|_{\mathcal{H}}^2 \left(\sum_{i \in S^{(t)}} (f^{(t)}(\mathbf{x}_i) - y_i)^2 + C\|f^{(t)}\|_{\mathcal{H}}^2 \right)
\end{aligned} \tag{7}$$

We can now complete the upper bound for $\|f^{(t+1)} - \text{Proj}_{\Psi_m} f^*\|_{\mathcal{H}}^2$ combining (5)-(7).

$$\begin{aligned}
&\|f^{(t+1)} - \text{Proj}_{\Psi_m} f^*\|_{\mathcal{H}}^2 \leq \|f^{(t)} - \text{Proj}_{\Psi_m} f^*\|_{\mathcal{H}}^2 \\
&\cdot \left(1 - \frac{8\eta(1-2\epsilon)}{(1-\epsilon)} \left(\lambda_m(\mathbf{\Gamma}) - \left\| \frac{1}{n(1-\epsilon)} \sum_{i \in S^{(t+1)} \cap P} \phi(\mathbf{x}_i) \otimes \phi(\mathbf{x}_i) - \mathbf{\Gamma} \right\|_{\text{HS}} \right) + \frac{16\eta}{[n(1-\epsilon)]^2} \left\| \sum_{i \in S^{(t)} \cap Q} \phi(\mathbf{x}_i) \right\|_{\mathcal{H}}^2 \right. \\
&\quad \left. + \frac{4\eta^2}{n(1-\epsilon)} \left\| \sum_{i \in S^{(t)} \cap P} \phi(\mathbf{x}_i) \otimes \phi(\mathbf{x}_i) \right\|_{\text{op}} \|\text{Proj}_{\Psi_m^\perp} f^*\|_{\mathcal{H}} + \frac{4\eta}{n(1-\epsilon)} \left\| \sum_{i \in S^{(t)} \cap P} \mu_i \phi(\mathbf{x}_i) \right\|_{\mathcal{H}} \right) \\
&\quad + \left(\eta + \frac{8\eta^2}{n^2(1-\epsilon)^2} \left\| \sum_{i \in S^{(t)}} \phi(\mathbf{x}_i) \right\|_{\mathcal{H}}^2 \right) \left(\sum_{i \in S^{(t)}} (f^{(t)}(\mathbf{x}_i) - y_i)^2 + C\|f^{(t)}\|_{\mathcal{H}} \right) + \lambda_m(\mathbf{\Gamma}) \|\text{Proj}_{\Psi_m^\perp} f^{(t)}\|_{\mathcal{H}}^2 \\
&\quad + \frac{4\eta}{n(1-\epsilon)} \left\| \sum_{i \in S^{(t)} \cap P} \phi(\mathbf{x}_i) \otimes \phi(\mathbf{x}_i) \right\|_{\text{op}} \|\text{Proj}_{\Psi_m^\perp} f^*\|_{\mathcal{H}} + \frac{2\eta}{n(1-\epsilon)} \left\| \sum_{i \in S^{(t)} \cap P} \mu_i \phi(\mathbf{x}_i) \right\|_{\mathcal{H}} \\
&\triangleq \|f^{(t)} - \text{Proj}_{\Psi_m} f^*\|_{\mathcal{H}}^2 (1 - I_1 + I_2 + I_3 + I_4) \\
&\quad (\eta + II_1) \left(\sum_{i \in S^{(t)}} (f^{(t)}(\mathbf{x}_i) - y_i)^2 + C\|f^{(t)}\|_{\mathcal{H}}^2 \right) + III_1 + III_2 + III_3
\end{aligned} \tag{8}$$

We will bound each of the terms seperately with the theory we develop in Section A. We will now look at the residual term. From noting that Subquantile Minimization is ℓ-smooth and μ-strongly convex for kernelized ridge regression (see § C.1), we have from the Descent Lemma (see Lemma 23).

$$\begin{aligned}\sum_{i\in S^{(t+1)}}(f^{(t+1)}(\mathbf{x}_i)-y_i)^2+C\|f^{(t+1)}\|_{\mathcal{H}}^2 &\le \sum_{i\in S^{(t)}}(f^{(t+1)}(\mathbf{x}_i)-y_i)^2+C\|f^{(t+1)}\|_{\mathcal{H}}^2\\ &\le \left(1-\frac{C}{\ell}\right)\left(\sum_{i\in S^{(t)}}(f^{(t)}(\mathbf{x}_i)-y_i)^2+C\|f^{(t)}\|_{\mathcal{H}}^2\right)\end{aligned}$$

where the first inequality follows from the optimality of the subquantile set, and the second inequality follows from the Descent Lemma (Lemma 23). We denote the term parameterized by $t\in[T]$,

$$\Lambda^{(t)} \triangleq \|f^{(t)}-\mathrm{Proj}_{\Psi_m}f^*\|_{\mathcal{H}}^2+\sum_{i\in S^{(t)}}(f^{(t)}(\mathbf{x}_i)-y_i)^2+C\|f^{(t)}\|_{\mathcal{H}}^2$$

We will show there exists n such that $\Lambda^{(t+1)}\le(1-\Omega(\ell^{-1}))\cdot\Lambda^{(t)}+E$ where E is some small error such that $E\to 0$ as $n\to\infty$. On a high-level, we have $I_1=O(\eta\lambda_m(\mathbf{\Gamma}))$. We analyze II_1 first. From the assumption that the corrupted covariates are centered, we have for a sufficienty small δ that with probability at least $1-\delta$ from Proposition 16,

$$\begin{aligned}\left\|\sum_{i\in S^{(t)}}\phi(\mathbf{x}_i)\right\|_{\mathcal{H}}^2 &\le 2\max_{A\in\mathcal{S}}\left\|\sum_{i\in A}\phi(\mathbf{x}_i)\right\|_{\mathcal{H}}^2+2\left\|\sum_{i\in S^{(t)}\cap Q}\phi(\mathbf{x}_i)\right\|_{\mathcal{H}}^2\\ &\le 2\left[n(1-\epsilon)\right]^{3/2}\left(\sqrt{(8/C)\epsilon\log\epsilon^{-1}}\,\mathrm{Tr}(\mathbf{\Gamma})+Q_k\right)\end{aligned}$$

where $Q_k=\max_{i\in Q}k(\mathbf{x}_i,\mathbf{x}_i)$. We then have with probability exceeding $1-\delta$,

$$II_1\le\frac{16\eta^2}{\sqrt{n(1-\epsilon)}}\left(\sqrt{(8/C)\epsilon\log\epsilon^{-1}}\,\mathrm{Tr}(\mathbf{\Gamma})+Q_k\right)\le\frac{1}{\ell}$$

when $n\ge 16^3\ell^{-2}(1-\epsilon)^{-1}\left((1/C)\,\mathrm{Tr}^2(\mathbf{\Gamma})\epsilon\log\epsilon^{-1}+Q_k^2\right)$. This gives us $II\le\frac{C-2}{\ell}$. Then, from our choice of C, we have $II\le(1/8\ell)$. We will analyze the terms in I individually. Define $\zeta\triangleq(1-\epsilon)/(1-2\epsilon)$. Then, we have

$$\begin{aligned}I_1 &\triangleq 8\eta\zeta\left(\lambda_m(\mathbf{\Gamma})-\left\|\sum_{i\in S^{(t)}\cap P}\phi(\mathbf{x}_i)\otimes\phi(\mathbf{x}_i)-\mathbf{\Gamma}\right\|_{\mathrm{HS}}\right)\\ &\ge 8\lambda_m(\mathbf{\Gamma})\zeta\left(\lambda_m(\mathbf{\Gamma})-\sqrt{\frac{8\lambda_{\max}^2(\mathbf{\Gamma})}{n(1-\epsilon)}}-\sqrt{\frac{2P_k^2\log(2/\delta)}{n(1-\epsilon)}+\frac{2P_k^2\epsilon\log\epsilon^{-1}}{(1-\epsilon)}}\right)\ge\frac{\lambda_m(\mathbf{\Gamma})}{2\ell}\end{aligned}\tag{9}$$

when $n\ge(1-2\epsilon)^{-1}(256\,\|\mathbf{\Gamma}\|_{\mathrm{op}}^2+64P_k^2\log(2/\delta)$ and $\epsilon\le 8^{-1/2}\,(\lambda_m(\mathbf{\Gamma})/P_k)^2$.

$$I_2\triangleq\frac{4\eta}{[n(1-\epsilon)]^2}\left\|\sum_{i\in S^{(t)}\cap Q}\phi(\mathbf{x}_i)\right\|_{\mathcal{H}}^2\le\frac{4Q_k\ln(1/\delta)}{\ell n(1-\epsilon)}\le\frac{\lambda_m(\mathbf{\Gamma})}{8\ell}\tag{10}$$

when $n \geq 32 Q_k \lambda_m^{-1}(\mathbf{\Gamma}) \ln(1/\delta)$. The bound for I_3 follows simply from our calculation of η in Lemma 21.

$$I_3 \triangleq \frac{4\eta^2}{n(1-\epsilon)} \left\| \sum_{i \in S^{(t)} \cap P} \phi(\mathbf{x}_i) \otimes \phi(\mathbf{x}_i) \right\|_{\text{op}} \left\| \text{Proj}_{\Psi_m^\perp} f^* \right\|_{\mathcal{H}} \leq \frac{4\eta}{n(1-\epsilon)} \left\| \text{Proj}_{\Psi_m^\perp} f^* \right\|_{\mathcal{H}} \leq \frac{\lambda_m(\mathbf{\Gamma})}{8\ell} \tag{11}$$

when $n \geq 32(1-\epsilon)^{-1} \lambda_m^{-1}(\mathbf{\Gamma}) R$. Next we analyze I_4, let $\mathcal{S}$ be the set of all combinations of subsets of size $[n(1-2\epsilon)]$ to $[n(1-\epsilon)]$. Then with probability exceeding $1-\delta$, we have

$$\begin{aligned} I_4 &\triangleq \frac{4\eta^2}{n(1-\epsilon)} \left\| \sum_{i \in S^{(t)} \cap P} \mu_i \phi(\mathbf{x}_i) \right\|_{\mathcal{H}} \\ &\leq \frac{4\sigma}{\ell^2 \left[n(1-\epsilon)\right]^{1/4}} \sqrt{2 \log n \log \frac{\log n}{\sqrt{2}\delta} \left(\sqrt{(8/C)\epsilon \log \epsilon^{-1}} \operatorname{Tr}(\mathbf{\Gamma}) \right)} \leq \frac{\lambda_m(\mathbf{\Gamma})}{8\ell} \end{aligned} \tag{12}$$

when $n \geq 32^5 (1-\epsilon)^{-1} \lambda_m^{-4}(\mathbf{\Gamma}) \ell^4 \log n \log^2 \frac{\log n}{\sqrt{2}\delta} \left((1/C)\epsilon \log \epsilon^{-1} \operatorname{Tr}(\mathbf{\Gamma}) + Q_k \right)^2$. Combining Equations 9, 10, 11, and 12, we have $I \leq \frac{\lambda_m(\mathbf{\Gamma})}{8\ell}$ given sufficient data and small epsilon. Now we will analyze III_1. For a general $t \in [T]$, we have

$$\begin{aligned} &\left\| \text{Proj}_{\Psi_m^\perp} f^{(t+1)} \right\|_{\mathcal{H}}^2 \\ &\leq 2 \left\| \text{Proj}_{\Psi_m^\perp} f^{(t)} \right\|^2 + \frac{2\eta^2}{\left[n(1-\epsilon)\right]^2} \left\| \sum_{i \in S^{(t)}} (f^{(t)}(\mathbf{x}_i) - y_i)^2 \cdot \text{Proj}_{\Psi_m^\perp} \phi(\mathbf{x}_i) \right\|_{\mathcal{H}}^2 \\ &\overset{(iv)}{\leq} 2 \left\| \text{Proj}_{\Psi_m^\perp} f^{(t)} \right\|^2 + \frac{8\eta^2 \left\| \mathbf{y} \right\|_\infty^2}{\left[n(1-\epsilon)\right]^2} \left(\left\| \sum_{i \in S^{(t)} \cap Q} \text{Proj}_{\Psi_m^\perp} \phi(\mathbf{x}_i) \right\|_{\mathcal{H}}^2 + \left\| \sum_{i \in S^{(t)} \cap P} \text{Proj}_{\Psi_m^\perp} \phi(\mathbf{x}_i) \right\|_{\mathcal{H}}^2 \right) \\ &\leq 2 \left\| \text{Proj}_{\Psi_m^\perp} f^{(t)} \right\|^2 + \frac{8\eta^2 \left\| \mathbf{y} \right\|_\infty^2 \left(\sqrt{(8/C)\epsilon \log \epsilon^{-1}} \operatorname{Tr}(\mathbf{\Gamma}_{22}) + Q_k \right)}{\sqrt{n(1-\epsilon)}} \end{aligned}$$

Solving the recursion for t steps, we then have

$$III_1 \leq \frac{2^{t+3} \left\| \mathbf{y} \right\|_\infty^2 \left(\sqrt{(8/C)\epsilon \log \epsilon^{-1}} \operatorname{Tr}(\mathbf{\Gamma}_{22}) + Q_k \right)}{\ell^2 \sqrt{n(1-\epsilon)}}$$

Our bound for III_2 is simple considering the choice of ℓ.

$$III_2 \triangleq \frac{4\eta}{n(1-\epsilon)} \left\| \sum_{i \in S^{(t)} \cap P} \phi(\mathbf{x}_i) \otimes \phi(\mathbf{x}_i) \right\|_{\text{op}} \left\| \text{Proj}_{\Psi_m^\perp} f^* \right\|_{\mathcal{H}} \leq 4 \left\| \text{Proj}_{\Psi_m^\perp} f^* \right\|_{\mathcal{H}}$$

We finally give a bound for III_3

$$\begin{aligned} III_3 &\triangleq \frac{2\eta}{n(1-\epsilon)} \left\| \sum_{i \in S^{(t)} \cap P} \mu_i \phi(\mathbf{x}_i) \right\|_{\mathcal{H}} \\ &\leq \frac{2\sigma}{\ell \left[n(1-\epsilon)\right]^{1/4}} \sqrt{2 \log n \log \frac{\log n}{\sqrt{2}\delta} \left(\sqrt{(8/C)\epsilon \log \epsilon^{-1}} \operatorname{Tr}(\mathbf{\Gamma}) \right)} \end{aligned}$$

Combining our bounds, we have

$$\begin{aligned}\Lambda^{(t+1)} \leq & \left(1 - \frac{\lambda_m(\mathbf{\Gamma})}{8\ell}\right) \cdot \Lambda^{(t)} \\ & + \frac{2\sigma}{\ell \left[n(1-\epsilon)\right]^{1/4}} \sqrt{2 \log n \log \frac{\log n}{\sqrt{2}\delta} \left(\sqrt{(8/C)\epsilon \log \epsilon^{-1}} \operatorname{Tr}(\mathbf{\Gamma})\right)} \\ & + \frac{2^{t+3} \|\mathbf{y}\|_\infty^2 \left(\sqrt{(8/C)\epsilon \log \epsilon^{-1}} \operatorname{Tr}(\mathbf{\Gamma}_{22}) + Q_k\right)}{\ell^2 \sqrt{n(1-\epsilon)}} + 4 \left\| \operatorname{Proj}_{\Psi_m^\perp} f^* \right\|_{\mathcal{H}}\end{aligned}$$

We have,

$$\Lambda^{(0)} \leq \|f^*\|_{\mathcal{H}}^2 + \|\mathbf{y}\|^2$$

Finally, from solving the recursion, we have after $O\left(\frac{2\left(\|f^*\|_{\mathcal{H}}^2 + \|\mathbf{y}\|^2\right)}{\varepsilon}\right)$ iterations the claimed bound. ■

Appendix 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 Lipschitzness

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

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

where $\ell = \frac{1}{n(1-\epsilon)} \sum_{i \in X} \|\phi(\mathbf{x}_i)\|_{\mathcal{H}}$

Proof. We use the $\mathcal{H}$ norm of the gradient to bound ℓ from above. Let S be denoted as the subquantile set. Define the sigmoid function as $\sigma(x) = \frac{1}{1+e^{-x}}$.

$$\begin{aligned}\|\nabla_f g(t, f)\|_{\mathcal{H}} &= \left\| \frac{1}{n(1-\epsilon)} \sum_{i=1}^{n} \mathbb{I}\left\{t \geq (1 - y_i) \log(f^{(t)}(\mathbf{x}_i))\right\} \left(y_i - \sigma(f^{(t)}(\mathbf{x}_i))\right) \cdot \phi(\mathbf{x}_i) \right\|_{\mathcal{H}} \\ &\overset{(i)}{\leq} \frac{1}{n(1-\epsilon)} \left\| \sum_{i \in S^{(t)}} \left(y_i - \sigma(f^{(t)}(\mathbf{x}_i))\right) \cdot \phi(\mathbf{x}_i) \right\|_{\mathcal{H}} \\ &\overset{(ii)}{\leq} \frac{1}{n(1-\epsilon)} \max_{i \in [n]} \left| y_i - \sigma(f^{(t)}(\mathbf{x}_i)) \right| \left\| \sum_{i \in S^{(t)}} \phi(\mathbf{x}_i) \right\|_{\mathcal{H}} \\ &\overset{(iii)}{\leq} \frac{1}{n(1-\epsilon)} \left\| \sum_{i \in X} \phi(\mathbf{x}_i) \right\|_{\mathcal{H}}\end{aligned}$$

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

D.2. Proof of Theorem 6

From Algorithm 2, we have for kernelized binary classification,

$$f^{(t+1)} = \text{Proj}_{\mathcal{K}}\left[f^{(t)} - \frac{\eta}{n(1-\epsilon)}\sum_{i\in S^{(t)}}\Big(\sigma(f^{(t)}(\mathbf{x}_i)) - y_i\Big)\cdot\phi(\mathbf{x}_i)\right] \tag{13}$$

From which it follows,

$$\begin{aligned}
\|f^{(t+1)} - f^*\|_{\mathcal{H}}^2 &= \left\|\text{Proj}_{\mathcal{K}}\left[f^{(t)} - \frac{\eta}{n(1-\epsilon)}\nabla g(f^{(t)}, t^*)\right] - f^*\right\|_{\mathcal{H}}^2 \\
&\overset{(i)}{\leq} \left\|f^{(t)} - \frac{\eta}{n(1-\epsilon)}\nabla g(f^{(t)}, t^*) - f^*\right\|_{\mathcal{H}}^2 \\
&\leq 2\|f^{(t)} - \text{Proj}_{\Psi_m} f^*\|_{\mathcal{H}}^2 - \frac{4\eta}{n(1-\epsilon)}\left\langle \nabla_f g(f^{(t)}, t^*), f^{(t)} - \text{Proj}_{\Psi_m} f^*\right\rangle_{\mathcal{H}} \\
&\qquad + \frac{2\eta^2}{n^2(1-\epsilon)^2}\|\nabla_f g(f^{(t)}, t^*)\|_{\mathcal{H}}^2 + 2\left\|\text{Proj}_{\Psi_m^{\perp}} f^*\right\|_{\mathcal{H}}^2
\end{aligned} \tag{14}$$

where (i) follows from the contraction property of the projection operator onto norm ball $\mathcal{K}$ and assuming $f^* \in \mathcal{K}$. We will expand the second term in Equation 14.

$$\begin{aligned}
&\frac{2\eta}{n(1-\epsilon)}\left\langle \nabla_f g(f^{(t)}, t^*), f^{(t)} - \text{Proj}_{\Psi_m} f^*\right\rangle_{\mathcal{H}} \\
&\qquad \overset{(13)}{=} \left\langle f^{(t)} - \text{Proj}_{\Psi_m} f^*, \frac{2\eta}{n(1-\epsilon)}\sum_{i\in S^{(t)}}\Big(\sigma(f^{(t)}(\mathbf{x}_i)) - y_i\Big)\cdot\phi(\mathbf{x}_i)\right\rangle_{\mathcal{H}} \\
&\qquad = \left\langle f^{(t)} - \text{Proj}_{\Psi_m} f^*, \frac{2\eta}{n(1-\epsilon)}\sum_{i\in S^{(t)}}\Big(\sigma(f^{(t)}(\mathbf{x}_i)) - \sigma(f^*(\mathbf{x}_i))\Big)\cdot\phi(\mathbf{x}_i)\right\rangle_{\mathcal{H}} \\
&\qquad\qquad + \left\langle f^{(t)} - \text{Proj}_{\Psi_m} f^*, \frac{2\eta}{n(1-\epsilon)}\sum_{i\in S^{(t)}}(\sigma(f^*(\mathbf{x}_i)) - y_i)\cdot\phi(\mathbf{x}_i)\right\rangle_{\mathcal{H}}
\end{aligned} \tag{15}$$

We first upper bound upper bound the second term in Equation 15. From the Cauchy-Schwarz Inequality and noting $y_i \in \{0, 1\}$ and $\text{range}(\sigma) \in (0, 1)$, we have the following,

$$\begin{aligned}
&\left\langle f^{(t)} - \text{Proj}_{\Psi_m} f^*, \frac{2\eta}{n(1-\epsilon)}\sum_{i\in S^{(t)}}(\sigma(f^*(\mathbf{x}_i)) - y_i)\cdot\phi(\mathbf{x}_i)\right\rangle_{\mathcal{H}} \\
&\qquad \leq \frac{2\eta}{n(1-\epsilon)}\|f^{(t)} - \text{Proj}_{\Psi_m} f^*\|_{\mathcal{H}}\left\|\sum_{i\in S^{(t)}}\phi(\mathbf{x}_i)\right\|_{\mathcal{H}}\max_{i\in S^{(t)}}|\sigma(f^*(\mathbf{x}_i)) - y_i| \\
&\qquad \overset{(ii)}{\leq} \frac{\eta^2}{n^{2-\beta}(1-\epsilon)^{2-\beta}}\|f^{(t)} - \text{Proj}_{\Psi_m} f^*\|_{\mathcal{H}}^2\left\|\sum_{i\in S^{(t)}}\phi(\mathbf{x}_i)\right\|_{\mathcal{H}}^2 + \frac{2}{n^{\beta}(1-\epsilon)^{\beta}}
\end{aligned} \tag{16}$$

where (ii) follows from Young's Inequality (Proposition 22) and noting for a vector $\mathbf{x} \in \mathbb{R}^d$ it holds $\|\mathbf{x}\|_\infty \leq \|\mathbf{x}\|_2$ and letting $\beta \in [0, 1]$ be an undetermined constant . 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 $\nabla^2 h(f) = \sum_{i\in S\cap P}\sigma(f(\mathbf{x}_i))(1 - \sigma(f(\mathbf{x}_i)))\cdot$

$\phi(\mathbf{x}_i) \otimes \phi(\mathbf{x}_i)$. From the properties of strong convexity, we have for any $f, \hat{f} \in \mathcal{H}$, there exists $\tilde{f} \in \mathcal{H}$ such that,

$$\begin{aligned}
&\left\langle f - \text{Proj}_{\Psi_m} \hat{f}, \nabla h(f) - \nabla h(\hat{f}) \right\rangle_{\mathcal{H}} = \left\langle f - \text{Proj}_{\Psi_m} \hat{f}, \nabla^2 h(\tilde{f})(f - \hat{f}) \right\rangle_{\mathcal{H}} \\
&\overset{(iii)}{=} \left\langle \nabla^2 h(\tilde{f}), (f - \hat{f}) \otimes (f - \hat{f}) \right\rangle_{\text{HS}} + \left\langle \nabla^2 h(\tilde{f}), (\text{Proj}_{\Psi_m^\perp} f^*) \otimes (f - \hat{f}) \right\rangle_{\mathcal{H}} \qquad (17)
\end{aligned}$$

where the equality in (iii) is given in (Gretton, 2015, Section 3.2). Then, from the strong convexity of h, there exists a constant C such that the following inequality holds,

$$\begin{aligned}
&\left\langle f^{(t)} - \text{Proj}_{\Psi_m} f^*, \frac{2\eta}{n(1-\epsilon)} \sum_{i \in S^{(t)} \cap P} \Big(\sigma(f^{(t)}(\mathbf{x}_i)) - \sigma(f^*(\mathbf{x}_i)) \Big) \cdot \phi(\mathbf{x}_i) \right\rangle_{\mathcal{H}} \\
&\overset{(17)}{\gtrsim} \frac{2\eta}{n(1-\epsilon)} \left\langle \sum_{i \in S^{(t)}} \phi(\mathbf{x}_i) \otimes \phi(\mathbf{x}_i), \text{Proj}_{\Psi_m} \left[f^{(t)} - f^* \right] \otimes \text{Proj}_{\Psi_m} \left[f^{(t)} - f^* \right] \right\rangle_{\text{HS}} \\
&\quad - \frac{2\eta}{n(1-\epsilon)} \left\| \text{Proj}_{\Psi_m^\perp} f^* \right\|_{\mathcal{H}} \left\| \sum_{i \in S^{(t)} \cap P} \phi(\mathbf{x}_i) \right\|_{\mathcal{H}} \\
&\overset{(iv)}{\gtrsim} 4\eta \frac{(1-2\epsilon)}{(1-\epsilon)} \left(\lambda_m(\mathbf{\Gamma}) - \left\| \frac{1}{n(1-2\epsilon)} \sum_{i \in S^{(t)} \cap P} \phi(\mathbf{x}_i) \otimes \phi(\mathbf{x}_i) - \mathbf{\Gamma} \right\|_{\text{HS}} \right) \| f^{(t)} - \text{Proj}_{\Psi_m} f^* \|_{\mathcal{H}}^2 \\
&\quad - \frac{2\eta}{n(1-\epsilon)} \left\| \text{Proj}_{\Psi_m^\perp} f^* \right\|_{\mathcal{H}} \left\| \sum_{i \in S^{(t)} \cap P} \phi(\mathbf{x}_i) \right\|_{\mathcal{H}} - 4\eta\zeta\lambda_m(\mathbf{\Gamma}) \| \text{Proj}_{\Psi_m^\perp} f^{(t)} \|_{\mathcal{H}}^2 \qquad (18)
\end{aligned}$$

where (iv) follows from Weyl's inequality (Weyl, 1912) and noting that $\left| S^{(t)} \cap P \right| \geq n(1-2\epsilon)$. We now briefly analyze the constant introduced in Equation 18.

$$\begin{aligned}
C_s \triangleq \inf_{\mathbf{x} \in P} \sigma(f(\mathbf{x}_i))(1 - \sigma(f(\mathbf{x}_i))) &\geq (1/2) \exp\left(- \max_{\mathbf{x} \in P} f(\mathbf{x}) \right) \\
&\geq (1/2) \exp\left(-2R \max_{\mathbf{x} \in P} \| \phi(\mathbf{x}) \|_{\mathcal{H}} \right) \geq (1/2) \exp\left(-2R\sqrt{P_k} \right) \qquad (19)
\end{aligned}$$

The second to last inequality follows from (Gretton, 2013, Theorem 17). We will now bound the third term in Equation 14.

$$\begin{aligned}
&\| \nabla_f g(f^{(t)}, t^*) \|_{\mathcal{H}}^2 = \left\| \frac{\eta}{n(1-\epsilon)} \sum_{i \in S^{(t)}} (\sigma(f^{(t)}(\mathbf{x}_i)) - y_i) \cdot \phi(\mathbf{x}_i) \right\|_{\mathcal{H}}^2 \\
&\leq \frac{\eta^2}{n^2(1-\epsilon)^2} \max_{i \in S^{(t)}} \left| \sigma(f^{(t)}(\mathbf{x}_i)) - y_i \right|^2 \cdot \left\| \sum_{i \in S^{(t)}} \phi(\mathbf{x}_i) \right\|_{\mathcal{H}}^2 \overset{(v)}{\leq} \frac{\eta^2}{n^2(1-\epsilon)^2} \left\| \sum_{i \in S^{(t)}} \phi(\mathbf{x}_i) \right\|_{\mathcal{H}}^2 \qquad (20)
\end{aligned}$$

where (v) follows from noting for any $\mathbf{x} \in \mathbb{R}^d$ it holds $\| \mathbf{x} \|_\infty \leq \| \mathbf{x} \|_2$. Furthermore, we note that if $-\log(\sigma(f(\mathbf{x}))) \leq -\log(\sigma(f(\hat{\mathbf{x}})))$, then $\sigma(f(\mathbf{x})) \geq \sigma(f(\hat{\mathbf{x}}))$. Now, combining (14)-(20), we obtain

$$\| f^{(t+1)} - \text{Proj}_{\Psi_m} f^* \|_{\mathcal{H}}^2 \leq \| f^{(t)} - \text{Proj}_{\Psi_m} f^* \|_{\mathcal{H}}^2$$

$$\cdot \left(1 - \frac{2C_s\eta(1-2\epsilon)}{(1-\epsilon)}\left(\lambda_m\left(\mathbf{\Gamma}\right) - \left\|\frac{1}{n(1-2\epsilon)}\sum_{i\in S^{(t)}\cap P}\phi(\mathbf{x}_i)\otimes\phi(\mathbf{x}_i) - \mathbf{\Gamma}\right\|_{\text{HS}}\right) + \frac{\eta^2}{[n(1-\epsilon)]^{7/4}}\left\|\sum_{i\in S^{(t)}}\phi(\mathbf{x}_i)\right\|_{\mathcal{H}}^2\right)$$
$$+ \frac{\eta^2}{n^2(1-\epsilon)^2}\left\|\sum_{i\in S^{(t)}}\phi(\mathbf{x}_i)\right\|_{\mathcal{H}}^2 + \frac{2}{[n(1-\epsilon)]^{1/4}} + \frac{2\eta}{n(1-\epsilon)}\|\text{Proj}_{\Psi_m^\perp} f^*\|_{\mathcal{H}}^2\left\|\sum_{i\in S^{(t)}\cap P}\phi(\mathbf{x}_i)\right\|_{\mathcal{H}}$$
$$+ 4\eta\zeta C\lambda_m(\mathbf{\Gamma})\left\|\text{Proj}_{\Psi_m^\perp} f^{(t)}\right\|_{\mathcal{H}}^2$$
$$\triangleq \|f^{(t)} - \text{Proj}_{\Psi_m} f^*\|_{\mathcal{H}}^2\left(1 - IV_1 + IV_2\right) + V_1 + V_2 + V_3 + V_4$$

Denote $\mathcal{S}$ as the set of combinations of $[n(1-2\epsilon)]$ to $[n(1-\epsilon)]$ probability at least $1-\delta$, we have

$$V_3 \leq \max_{\substack{\sigma\in\Pi\\|\sigma|=n(1-\epsilon)}}\left\|\sum_{i=1}^{n(1-\epsilon)}\phi(\mathbf{x}_{\sigma(i)})\right\|_{\mathcal{H}} \leq [n(1-\epsilon)]^{3/4}\sqrt{(8/C)\,\text{Tr}(\mathbf{\Gamma})\epsilon\log\epsilon^{-1}}$$

Then, from the assumption that the corrupted covariates are centered, we have for a sufficienty small δ that with probability at least $1-\delta$ from Proposition 16,

$$\left\|\sum_{i\in S^{(t)}}\phi(\mathbf{x}_i)\right\|_{\mathcal{H}}^2 \leq 2\max_{\substack{\sigma\in\Pi\\|\sigma|=n(1-\epsilon)}}\left\|\sum_{i=1}^{n(1-\epsilon)}\phi(\mathbf{x}_{\sigma(i)})\right\|_{\mathcal{H}}^2 + 2\left\|\sum_{i\in S^{(t)}\cap Q}\phi(\mathbf{x}_i)\right\|_{\mathcal{H}}^2$$
$$\leq 2\left[n(1-\epsilon)\right]^{3/2}\left((8/C)\epsilon\log\epsilon^{-1}\,\text{Tr}(\mathbf{\Gamma}) + Q_k\right)$$

where $Q_k = \max_{i\in Q} k(\mathbf{x}_i, \mathbf{x}_i)$. We thus have,

$$IV_2 \leq 2\eta^2\left[n(1-\epsilon)\right]^{-1/4}\left((8/C)\epsilon\log\epsilon^{-1}\,\text{Tr}(\mathbf{\Gamma}) + Q_k\right) \leq \frac{\zeta C_s\lambda_m(\mathbf{\Gamma})}{4\ell} \tag{21}$$

when $n \geq (1-\epsilon)^{-1}64^4C^{-4}\zeta^{-4}\lambda_m^{-4}(\mathbf{\Gamma})\ell^{-4}4\left(\left[\text{Tr}(\mathbf{\Gamma})\right]^2\epsilon\log\epsilon^{-1} + Q_k^2\right)$. Next, to obtain

$$\lambda_m(\mathbf{\Gamma}) - \left\|\frac{1}{n(1-2\epsilon)}\sum_{i\in S^{(t)}\cap P}\phi(\mathbf{x}_i)\otimes\phi(\mathbf{x}_i) - \mathbf{\Gamma}\right\|_{\text{HS}} \geq (1/2)\lambda_m(\mathbf{\Gamma})$$

with probability greater than $1-\delta$. We utilize the second relation in Proposition 19 and require $n \geq (1-2\epsilon)^{-1}\left(256\,\|\mathbf{\Gamma}\|^2 + 64P_k^2\log(2/\delta)\right)$ and $\epsilon \leq (1/\sqrt{8})\left(\lambda_m(\mathbf{\Gamma})/P_k\right)^2$. Thus when the data requirements are satisfied, with step size $\eta = \frac{1}{\ell}$, we obtain $IV \leq (\lambda_m(\mathbf{\Gamma})/4\ell)$. We will now show all the terms in V are small. With probability at least $1-\delta$ and supposing the data satisfied the requirement posed in Equation 21, it follows

$$\frac{\eta^2}{[n(1-\epsilon)]^2}\left\|\sum_{i\in S^{(t)}}\phi(\mathbf{x}_i)\right\|_{\mathcal{H}}^2 \leq \frac{2\eta^2\left((8/C)\epsilon\log\epsilon^{-1}\,\text{Tr}(\mathbf{\Gamma}) + Q_k\right)}{\sqrt{n(1-\epsilon)}} \leq \frac{\zeta C_s\lambda_m(\mathbf{\Gamma})}{8\ell\sqrt{n(1-\epsilon)}} \tag{22}$$

Next, we analyze $\|\text{Proj}_{\Psi_m^\perp} f^{(t)}\|_{\mathcal{H}}^2$.

$$\left\|\text{Proj}_{\Psi_m^\perp} f^{(t+1)}\right\|_{\mathcal{H}}^2 \leq 2\left\|\text{Proj}_{\Psi_m^\perp} f^{(t)}\right\|_{\mathcal{H}}^2 + 2\left\|\frac{\eta}{n(1-\epsilon)}\sum_{i\in S^{(t)}}(\sigma(f(\mathbf{x}_i)) - y_i)^2\cdot\text{Proj}_{\Psi_m^\perp}\phi(\mathbf{x}_i)\right\|_{\mathcal{H}}^2$$

$$\leq 2\left\|\mathrm{Proj}_{\Psi_m^\perp} f^{(t)}\right\|_{\mathcal{H}}^2 + \frac{4\eta^2}{[n(1-\epsilon)]^2}\left(\left\|\sum_{i\in S^{(t)}\cap Q}\mathrm{Proj}_{\Psi_m^\perp}\phi(\mathbf{x}_i)\right\|_{\mathcal{H}}^2 + \left\|\sum_{i\in S^{(t)}\cap P}\mathrm{Proj}_{\Psi_m^\perp}\phi(\mathbf{x}_i)\right\|_{\mathcal{H}}^2\right)$$

$$\overset{(v)}{\leq} 2\left\|\mathrm{Proj}_{\Psi_m^\perp} f^{(t)}\right\|_{\mathcal{H}}^2 + \frac{4\eta^2((8/C)\epsilon\log\epsilon^{-1}\,\mathrm{Tr}(\mathbf{\Gamma}_{22}) + Q_k)}{\sqrt{n(1-\epsilon)}} \tag{23}$$

when in (v) we apply Proposition 16 and partition $\mathbf{\Gamma}$ into

$$\mathbf{\Gamma} = \begin{matrix} \\ m \\ \infty \end{matrix}\begin{matrix} \begin{matrix} m & \infty \end{matrix} \\ \begin{bmatrix} \mathbf{\Gamma}_{11} & \mathbf{\Gamma}_{12} \\ \mathbf{\Gamma}_{12} & \mathbf{\Gamma}_{22} \end{bmatrix} \end{matrix}$$

From the recursion in Equation 23 and noting $\left\|\mathrm{Proj}_{\Psi_m^\perp} f^{(t)}\right\|_{\mathcal{H}} = 0$, we have

$$\left\|\mathrm{Proj}_{\Psi_m^\perp} f^{(t)}\right\|_{\mathcal{H}}^2 \leq \frac{2^{t+2}\left((8/C)\epsilon\log\epsilon^{-1}\,\mathrm{Tr}(\mathbf{\Gamma}_{22}) + Q_k\right)}{\ell^2\sqrt{n(1-\epsilon)}}$$

From the data requirement in Equation 21, we have

$$V_4 \triangleq 4\eta\zeta C_s\lambda_m(\mathbf{\Gamma})\left\|\mathrm{Proj}_{\Psi_m^\perp} f^{(t)}\right\|_{\mathcal{H}}^2 \leq \frac{2^{t+2}\zeta^2 C_s^2\lambda_m^2(\mathbf{\Gamma})}{\ell\left[n(1-\epsilon)\right]^{1/4}}\left(\frac{Q_m}{Q_k}\vee\frac{\|\mathbf{\Gamma}_{22}\|_{\mathrm{Tr}}}{\|\mathbf{\Gamma}\|_{\mathrm{Tr}}}\right) \tag{24}$$

In Equations 22 and 24, we see the $n^{1/4}$ term is in the denominator and goes to ∞ as $n\to\infty$. Thus we see $V\to 0$ as $n\to\infty$. We then have

$$\left\|f^{(T)} - \mathrm{Proj}_{\Psi_m} f^*\right\|_{\mathcal{H}} \leq \left(1 - \frac{\zeta C_s\lambda_m(\mathbf{\Gamma})}{\ell}\right)^T\left\|f^{(0)} - \mathrm{Proj}_{\Psi_m} f^*\right\|_{\mathcal{H}} + \sum_{k=0}^{T}\left(1 - \frac{\zeta C_s\lambda_m(\mathbf{\Gamma})}{\ell}\right)^k (V)$$

Then, from upper bounding the second term by the infinite sum, we have after $\log\left(\frac{2\|f^*\|_{\mathcal{H}}}{\varepsilon}\right)$ iterations

$$\left\|f^{(T)} - \mathrm{Proj}_{\Psi_m} f^*\right\|_{\mathcal{H}} \leq \frac{\varepsilon}{2} + \frac{\ell}{\zeta C_s\lambda_m(\mathbf{\Gamma})}(V)$$

after $\log\left(\frac{2\|f^*\|_{\mathcal{H}}}{\varepsilon}\right)$ iterations, our proof is complete ■