Proceedings of Machine Learning Research vol 196:1–24, 2024 37th Annual Conference on Learning Theory

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 propose Subquantile Minimization for learning with adversarial corruption in the training set 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 $\varepsilon \in [0, 0.5)$. The distribution is formed as $\hat{\mathcal{P}} = (1-\varepsilon)\mathcal{P} + \varepsilon\mathcal{Q}$, 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. We study kernelized regression, kernelized binary classification, and kernelized one-vs-all multi-class classification.

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 $\mathbf{y}$ as training data.*

Current approaches for robust learning across various machine learning tasks often use gradient descent over a robust objective, (Li et al., 2021). These robust objectives tend to not be convex and therefore do not have a strong analysis on the error bounds for general classes of models.

We similarly propose a robust objective which has a nonconvex-concave objective. This objective has also been proposed recently in Hu et al. (2020) where there has been an analysis in the Binary Classification Task. We show Subquantile Minimization reduces to the same objective in Hu et al. (2020). We use theory from the weakly-convex concave optimization literature for our error bounds. We are able to levarage this theory by analyzing the asymptotic distribution of a softplus approximation of the Subquantile objective.

The study of Kernel Learning in the Gaussian Design is quite popular, (Cui et al., 2021; Dicker, 2016). In (Cui et al., 2021), the feature space, $\phi(\mathbf{x}_i) \sim \mathcal{N}(0, \mathbf{\Sigma})$ where $\mathbf{\Sigma}$ is a diagonal matrix of dimension p, where p can be infinite. In this work, we adopt a similar framework, and with the power of Mercer's Theorem (Mercer, 1909), we are able to say $\mathrm{Tr}(\mathbf{\Sigma}) < \infty$. We use this fact extensively in our infinite-dimensional concentration inequalities.

Theorem 2 *(Informal). Let the dataset be given as $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$ such that the labels and covariates of ϵn samples are arbitrarily corrupted by an adversary.*
Kernelized Regression:

$$\|\hat{f} - f^*\|_{\mathcal{H}} \leq \varepsilon + O(\sigma) \tag{1}$$

Kernel Binary Classification:

$$\|\hat{f} - f^*\|_{\mathcal{H}} \leq \varepsilon + \tilde{O}\left(\frac{\mathcal{E}_{\mathrm{OPT}}}{n(1-\epsilon)}\right) + \tilde{O}\left(\frac{1}{n^{\beta}(1-\epsilon)^{\beta}}\right) \tag{2}$$

Kernel Multi-Class Classification:

$$\|f - f^*\| \leq O(\Xi) \tag{3}$$

1.1. Related Work

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

Robust Regression via Hard Thresholding Bhatia et al. (2015). Bhatia et al. study iterative thresholding for least squares regression / sparse recovery. Their theoretical results for the standard gradient descent case cover for known covariance with no feature covariance or Gaussian Noise.

Learning with bad training data via iterative trimmed loss minimization (Shen and Sanghavi, 2019). This work considers optimizing over the bottom-k errors by choosing the αn points with smallest error and then updating the model from these αn. This general model is the same as ours. Theoretically, this work considers only general linear models.

Trimmed Maximum Likelihood Estimation for Robust Generalized Linear Model (Awasthi et al., 2022). This work studies a different class of generalized linear models. Interestingly, they show for Gaussian Regression the iterative trimmed maximum likelihood estimator is able to achieve near minimax optimal error. This work does not consider feature corruption and primarily focuses on the covariates sampled with Gaussian Design from Identity covariance.

1.2. Contributions

We will now state our main contributions clearly.

1. We provide a novel theoretical framework using the Moreau Envelope for analyzing the iterative trimmed estimator for machine learning tasks.
2. We provide rigorous error bounds for subquantile minimization in the kernel regression, kernel binary classification, and kernel multi-class classification. Furthermore, we provide our bounds for both label and feature corruption with a general Gaussian Design.

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. $y \geq Cx$.

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 $k(\mathbf{x}, \mathbf{x}) \triangleq \langle \phi(\mathbf{x}_i), \phi(\mathbf{x}_i) \rangle_{\mathcal{H}}$. For a function in a RKHS, $f \in \mathcal{H}$, it follows for some $\mathbf{w} \in \mathbb{R}^n$, 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. Distribution

In this paper we consider $\phi(\mathbf{x}_i) \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$ as a Gaussian Design where $\mathbf{E}[\phi(\mathbf{x}_i)] = \mathbf{0}$ and $\mathbf{E}[\phi(\mathbf{x}_i) \otimes \phi(\mathbf{x}_i)] = \mathbf{\Sigma}$ where $\mathrm{Tr}(\mathbf{\Sigma}) < \infty$. The Gaussian Design for the Feature Space has gained popularity in the study of kernel learning (Cui et al., 2021).

2.3. Mathematical Tools

Proposition 3 (Young's Inequality (Young, 1912)) *For all $a, b \in \mathbb{R}$, it holds*

$$ab \leq \frac{a^2}{2} + \frac{b^2}{2} \tag{4}$$

Proposition 4 (Jensen's Inequality (Jensen, 1906)) *Suppose φ is a convex function, then for a random variable X, it holds*

$$\varphi\left(\mathbf{E}\left[X\right]\right) \leq \mathbf{E}\left[\varphi\left(X\right)\right] \tag{5}$$

The inequality is reversed for φ concave.

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.

$$\mathcal{Q}_p(U) \leq t \iff \mathbb{P}\left\{U \leq t\right\} \geq p \tag{6}$$

The p-subquantile of the risk is then given by

$$\mathbb{L}_p\left(U\right) = \frac{1}{p}\int_0^p \mathcal{Q}_p\left(U\right) dq = \mathbb{E}\left[U | U \leq \mathcal{Q}_p\left(U\right)\right] = \max_{t \in \mathbb{R}} \left\{t - \frac{1}{p}\mathbb{E}\left(t - U\right)^+\right\} \tag{7}$$

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

$$\min_{f\in\mathcal{K}}\max_{t\in\mathbb{R}}\left\{g(t,f)\triangleq t-\sum_{i=1}^{n}\left(t-(f(\mathbf{x}_i)-y_i)^2\right)^+\right\} \tag{8}$$

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

$$t^{(k+1)}=\arg\max_{t\in\mathbb{R}} g(f^{(k)},t) \tag{9}$$

$$f^{(k+1)}=\mathrm{Proj}_{\mathcal{K}}\left(f^{(k)}-\alpha\nabla_f g(f^{(k)},t^{(k+1)})\right) \tag{10}$$

4. Theory

To consider theoretical guarantees of Subquantile Minimization, we first analyze the inner and outer optimization problems. We first analyze kernel learning in the presence of corrupted data. Next, we provide error bounds for the two most important kernel learning problems, kernel ridge regression, and kernel classification. Now we will give our first result regarding kernel learning in the Huber ϵ-contamination model. Now we will analyze the two-step minimax optimization steps described in Equations (9) and (10).

Lemma 5 *Let $f(\mathbf{x};\mathbf{w})$ be a convex loss function. Let $\mathbf{x}_1,\mathbf{x}_2,\cdots,\mathbf{x}_n$ denote the n data points ordered such that $f(\mathbf{x}_1;\mathbf{w},y_1)\leq f(\mathbf{x}_2;\mathbf{w},y_2)\leq\cdots\leq f(\mathbf{x}_n;\mathbf{w},y_n)$. If we denote $\hat{\nu}_i\triangleq f(\mathbf{x}_i;\mathbf{w},y_i)$, it then follows $\hat{\nu}_{n(1-\epsilon)}\in\arg\max_{t\in\mathbb{R}} g(t,\mathbf{w})$.*

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

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

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

From Theorem 5, 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 6 *Let $\hat{\nu}_i\triangleq f(\mathbf{x}_i;\mathbf{w},y_i)$ s.t. $\hat{\nu}_{i-1}\leq\hat{\nu}_i\leq\hat{\nu}_{i+1}$, if we choose $t^{(k+1)}=\hat{\nu}_{n(1-\epsilon)}$ as by Theorem 5, 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:

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

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

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

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

This concludes the proof. ■

4.1. Kernelized Regression

The loss for the Kernel Ridge Regression problem for a single training pair $(\mathbf{x}_i, y_i) \in \mathcal{D}$ is given by the following equation

$$\ell\left(f; \mathbf{x}_i, y_i,\right) = \left(f(\mathbf{x}_i) - y_i\right)^2 \tag{11}$$

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 7 (Subquantile Minimization for Kernelized Regression is Good with High Probability) *Let* **??** *be run on a dataset $\mathcal{D} \sim \hat{\mathcal{P}}$ with learning rate $\eta \triangleq \Omega\left((\lambda_{\max}(\mathbf{\Sigma}))^{-1}\right)$. Suppose $n = \max\{\Omega((Q_k/(3\lambda_{\max}^2(\mathbf{\Sigma})))^{1/(1-\beta)}), \Omega\left(\left(\frac{8}{3}u \operatorname{Tr}(\mathbf{\Sigma})\right)^{1/(\beta-\frac{1}{2})}(1-\epsilon)^{-1}\right)\}$. Then after $T = \tilde{O}\left(\log\left(\left(\frac{\lambda_{\max}(\mathbf{\Sigma})\|f^*\|_{\mathcal{H}}}{\sqrt{n}}\right)\frac{1}{\varepsilon}\right)\right)$ iterations, w.h.p*

$$\|f^{(T)} - \operatorname{Proj}_{\Psi_m} f^*\|_{\mathcal{H}}^2 \leq \varepsilon + \tilde{O}\left(\frac{\|\operatorname{Proj}_{\Psi_m^\perp} f^*\|_{\mathcal{H}}}{\lambda_{\max}(\mathbf{\Sigma})\sqrt{n}} + \sigma\sqrt{\frac{\operatorname{Tr}(\mathbf{\Sigma})\log n}{n\lambda_{\max}^2(\mathbf{\Sigma})}} + \lambda_m(\mathbf{\Sigma})R^2\right) \tag{12}$$

Full proof with explicit constants is given in Appendix C.1. A direct application of Theorem 7 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). For the linear and polynomial kernel, we then have β increases, therefore to obtain the same bound on η as with no feature noise, we simply need more data. The effect of **??** can be seen in the denominator of both terms. Instead of $\lambda_{\min}(\mathbf{\Sigma})$ we have $c_4\lambda_m$ for a finite m. This difference will be clear in the following corollary, where we utilize the theory develoepd for kernelized regression to imply a result for regularized linear regression.

Corollary 8 (Linear Regression Expected Error Bound) *Consider Subquantile Minimization for Linear Regression on the data X with optimal parameters $\mathbf{w}^*$. Assume $\mathbf{x}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$ for $i \in [n]$. Then after T iterations of Algorithm 1, we have the following error bounds for robust kernelized linear regression. Given sufficient data*

$$\|f^{(T)} - f^*\|_{\mathcal{H}} \leq \varepsilon + O(\sigma) \tag{13}$$

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. Set the step-size
$$\eta \leq O\left(\frac{\lambda_m\left(\mathbf{\Sigma}\right)}{\mathrm{Tr}\left(\mathbf{\Sigma}\right)}\right)$$
2. Set the number of iterations
$$T = \tilde{O}\left(\log\left(\left(\frac{\lambda_{\max}(\mathbf{\Sigma})\left\|f^*\right\|_{\mathcal{H}}}{\sqrt{n}}\right)\frac{1}{\varepsilon}\right)\right)$$
3. **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.
    $$\nabla_f g(t^{(k+1)}, f^{(k)}) \leftarrow \frac{2}{n(1-\epsilon)}\sum_{i \in S^{(k)}}(f^{(k)}(\mathbf{x}_i) - y_i) \cdot \phi(\mathbf{x}_i)$$
    5. Perform Projected Gradient Descent Iteration.
    $$f^{(k+1)} \leftarrow \mathrm{Proj}_{\mathcal{K}}\left[f^{(k)} - \eta\nabla g(f^{(k)}, t^{(k+1)})\right]$$

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

Algorithm 1: Subquantile Minimization for Kernelized Regression

Proof given in **??**. Let us note for the case where p is finite, i.e. the feature mapping is finite-dimensional, e.g. linear or polynomial kernel. Then we have that $\mathrm{Proj}_{\Psi_m^\perp}$ where $m = p$ is equal to zero as $\{\varphi_i\}_{i=1}^m$ spans the finite-dimensional space, in which we case we have the absolute constant given in **??** is equal to zero. It is important to note in all our bounds, $\gamma \leq \sqrt{\frac{\epsilon}{1-2\epsilon}}$ is a theoretical worst case bound when the Subquantile contains the minimum possible number of uncorrupted points. In other words, we have $\gamma \triangleq \frac{|P \setminus S|}{|S \cap P|} \leq \frac{n\epsilon}{n(1-2\epsilon)} = \frac{\epsilon}{1-2\epsilon}$. So, as $|S \cap P|$ increases, we have a better error bound as $|P \setminus S|$ decreases. As is typical in the robust statistics literature, we make no assumptions on the distribution of the corrupted data so we cannot say anything about $|S \cap P|$. We will have γ decreases if stationary points give high error for corrupt points as our optimization procedure moves toward a stationary point.

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 $(\mathbf{x}_i, y_i)$

$$\ell\left(\mathbf{x}_i, y_i; f\right) = -y_i \log\left(\sigma\left(f(\mathbf{x}_i)\right)\right) - (1 - y_i)\log\left(1 - \sigma\left(f(\mathbf{x}_i)\right)\right) \tag{14}$$

We will now give our algorithm.

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. Set the step-size
$$\eta \leq O\left(\frac{\lambda_{\min}(\mathbf{\Sigma})}{\mathrm{Tr}\left(\mathbf{\Sigma}\right)}\right)$$
2. Set the number of iterations
$$T = O\left(\log\left(\left(\frac{\lambda_{\max}(\mathbf{\Sigma})\, \|f^*\|_{\mathcal{H}}}{\sqrt{n}}\right)\frac{1}{\varepsilon}\right)\right)$$
3. **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.
    $$\nabla_f g(t^{(k+1)}, f^{(k)}) \leftarrow \frac{2}{n(1-\epsilon)} \sum_{i \in S^{(k)}} \left(\sigma(f^{(k)}(\mathbf{x}_i)) - y_i\right) \cdot \phi(\mathbf{x}_i)$$
    5. Perform Projected Gradient Descent Iteration.
    $$f^{(k+1)} \leftarrow \mathrm{Proj}_{\mathcal{K}}\left[f^{(k)} - \eta \nabla g(f^{(k)}, t^{(k+1)})\right]$$

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

Algorithm 2: Subquantile Minimization for Binary Classification

Theorem 9 (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 $\eta \triangleq \Omega(L^{-1})$. Then after $O\left(\log\left(\left(\frac{\lambda_{\max}(\mathbf{\Sigma})\|f^*\|_{\mathcal{H}}}{\sqrt{n}}\right)\frac{1}{\varepsilon}\right)\right)$ gradient descent iterations,*

$$\|f^{(T)} - f^*\|_2 \leq \varepsilon + \tilde{O}\left(\frac{\mathcal{E}_{\mathrm{OPT}}}{n(1-\epsilon)}\right) + \tilde{O}\left(\frac{1}{n^\beta (1-\epsilon)^\beta}\right) \tag{15}$$

where $\mathcal{E}_{\text{OPT}} \triangleq \sum_{i \in P} \left[\underset{(\mathbf{x}_i, y_i) \sim \mathcal{P}}{\mathbf{Pr}} \{y_i \mid \mathbf{x}_i\} - y_i \right]^2.$

Proof Sketch. We will show there is a linear decrease in the average squared error each iteration.

$$\frac{1}{n(1-\epsilon)} \sum_{i \in P} (\sigma(f^{(t)}(\mathbf{x}_i)) - y_i)^2 \leq \frac{1}{2n(1-\epsilon)} \sum_{i \in P} (\sigma(f^{(t)}(\mathbf{x}_i)) - y_i)^2 + O\left(\frac{\mathcal{E}_{\text{OPT}}}{n^\beta (1-\epsilon)^\beta}\right) \quad (16)$$

for any $\beta \in [0, 1]$, where n must be large for larger β. ■

The full proof is given in Appendix D.1. In Theorem 9, we introduce $\mathcal{E}_{\text{OPT}}$, which says we are only able to learn up to the intrinsic noise within the target function.

5. Discussion

The main contribution of this paper is the study of a nonconvex-concave formulation of Subquantile minimization for the robust learning problem for kernel ridge regression and kernel classification. We present an algorithm to solve the nonconvex-concave formulation and prove rigorous error bounds which show that the more good data that is given decreases the error bounds. We also present accelerated gradient methods for the two-step algorithm to solve the nonconvex-concave optimization problem and give novel theoretical bounds.

Theory. We develop strong theoretical bounds on the normed difference between the function returned by Subquantile Minimization and the optimal function for data in the target distribution, $\mathbb{P}$, in the Gaussian Design. In expectation and with high probability, given sufficient data dependent on the kernel, we obtain a near minimax optimal error bound for a general positive definite continuous kernel. Our theoretical analysis is novel in that it utilizes the Moreau Envelope from a min-max formulation of the iterative thresholding algorithm.

Experiments. From our experiments, we see Subquantile Minimization is competitive with algorithms developed solely for robust linear regression as well as other meta-algorithms. Our theoretical analysis is through the lens of kernel-learning, but the generalization to linear regression from a non-kernel perspective can be done. In kernelized regression, we see SUBQUANTILE is the strongest of the meta-algorithms. Furthemore, in binary and multi-class classification, SUBQUANTILE is very strong. Thus, we can see empirically SUBQUANTILE is the strongest meta-algorithm across all kernelized regression and classification tasks and also the strongest algorithm in linear regression.

Interpretability. One of the strengths in Subquantile Optimization is the high interpretability. Once training is finished, we can see the $n(1-p)$ points with highest error to find the outliers and the features follow Gaussian Design. Furthermore, there is only hyperparameter p, which should be chosen to be approximately the percentage of inliers in the data and thus is not very difficult to tune for practical purposes. Our theory suggests for a problem where the amount of corruptions is unknown,

General Assumptions. The general assumption is the majority of the data should inliers. This is not a very strong assumption, as by the definition of outlier it should be in the minority. Furthermore, we assume the feature maps have a Gaussian Design. Such a design in many prior works in kernel learning and we therefore find it suitable.

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

References

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. URL https://openreview.net/forum?id=VHmdFPy4U_u.

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. URL https://proceedings.neurips.cc/paper_files/paper/2015/file/1be3bc32e6564055d5ca3e5a354acbef-Paper.pdf.

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. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/e702e51da2c0f5be4dd354bb3e295d37-Paper.pdf.

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. URL https://proceedings.mlr.press/v99/cheng19a.html.

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. URL https://proceedings.mlr.press/v119/cheng20a.html.

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. URL https://doi.org/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.

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. URL https://doi.org/10.1214/009053607000000677.

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. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/f0d7053396e765bf52de12133cf1afe8-Paper.pdf.

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

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.

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

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. URL https://doi.org/10.1007/s11228-021-00609-w.

Michel Ledoux and Michel Talagrand. *Probability in Banach Spaces: isoperimetry and processes*. Springer Science & Business Media, 2013.

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. URL https://openreview.net/forum?id=K5YasWXZT3O.

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. URL https://proceedings.mlr.press/v89/mukhoty19a.html.

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. URL https://api.semanticscholar.org/CorpusID:3614648.

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. URL https://www.sciencedirect.com/science/article/pii/S0377221713008692.

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.

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

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.

Nicholas Young. *An introduction to Hilbert space*. Cambridge university press, 1988.

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

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

where $\mathbf{\Sigma}$ is a possibly infinite dimensional covariance operator. [Normal Residuals] The residual is defined as $\mu_i \triangleq f^*_{\mathbf{w}}(\mathbf{x}_i) - y_i$. Then we assume for some $\sigma > 0$, it follows

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

Proposition 10 (Gaussian Concentration (Ledoux and Talagrand, 2013)) *Suppose* $\mathbf{x}$ *is a Gaussian Variable in a Banach Space. Then,*

$$\mathbf{Pr}\left\{\|\mathbf{x}\| > t\right\} \leq 4\exp\left(-\frac{t^2}{8\,\mathbf{E}\,\|\mathbf{x}\|^2}\right)$$

Noting that all Hilbert Spaces are Banach Spaces (Young, 1988), we will use this proposition throughout the section.

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.

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

(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 folllows $\mathbf{Pr}\{\max_{i\in[n]} X_i \geq C\} = n\,\mathbf{Pr}\{X \geq C\}$. 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* $s \geq 1$*,*

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

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

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

where the last inequality follows from Theorem 10. Now choosing $C \triangleq 8\,\mathrm{Tr}(\mathbf{\Sigma})$ completes the proof. ■

Proposition 13 (Probabilistic Maximum P_k) *Let $\mathbf{x}_i \sim \mathcal{P}$ such that $\phi(\mathbf{x}_i) \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$ (Appendix A). Then it follows for any $s \geq 1$*

$$\underset{\phi(\mathbf{x}_i)\sim\mathcal{N}(\mathbf{0},\mathbf{\Sigma})}{\mathbf{Pr}} \left\{ \max_{i\in[n]} \|\phi(\mathbf{x}_i)\|_{\mathcal{H}} \geq \sqrt{8\operatorname{Tr}(\mathbf{\Sigma})\log n} \cdot s \right\} \leq 4e^{1-s^2}$$

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

$$\underset{\phi(\mathbf{x}_i)\sim\mathcal{N}(\mathbf{0},\mathbf{\Sigma})}{\mathbf{Pr}} \left\{ \max_{i\in[n]} \|\phi(\mathbf{x}_i)\|_{\mathcal{H}} \geq C \cdot s \right\} \overset{(i)}{\leq} n \underset{\phi(\mathbf{x})\sim\mathcal{N}(\mathbf{0},\mathbf{\Sigma})}{\mathbf{Pr}} \left\{ \|\phi(\mathbf{x})\|_{\mathcal{H}} \geq C \cdot s \right\} = 4n \exp\left(-\frac{C^2 \cdot s^2}{8\operatorname{Tr}(\mathbf{\Sigma})} \right)$$

See (i) from the proof of Theorem 11. Setting $C \triangleq \sqrt{8\operatorname{Tr}(\mathbf{\Sigma})\log n}$ completes the proof. ■

Proposition 14 (Probabilistic bound on Norm of Functions in the Reproducing Kernel Hilbert Space) *Let $\mathbf{x}_i \sim \mathcal{P}$ such that $\phi(\mathbf{x}_i) \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$ (Appendix A). Then it follows*

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

Proof. Let C be a positive constant to be determined and $u \geq 1$. We then have

$$\begin{aligned}
\underset{\phi(\mathbf{x}_i)\sim\mathcal{N}(\mathbf{0},\mathbf{\Sigma})}{\mathbf{Pr}} \left\{ \left\| \sum_{i=1}^{n} \phi(\mathbf{x}_i) \right\|_{\mathcal{H}} \geq C \cdot u \right\} &= \underset{\phi(\mathbf{x}_i)\sim\mathcal{N}(\mathbf{0},\mathbf{\Sigma})}{\mathbf{Pr}} \left\{ \underset{\xi_i\sim\mathcal{R}}{\mathbf{E}} \left\| \sum_{i\in A} \xi_i \phi(\mathbf{x}_i) \right\|_{\mathcal{H}} \geq C \cdot u \right\} \\
&\overset{(i)}{\leq} \underset{\phi(\mathbf{x}_i)\sim\mathcal{N}(\mathbf{0},\mathbf{\Sigma})}{\mathbf{Pr}} \left\{ \underset{\xi_i\sim\mathcal{R}}{\mathbf{E}} \left\| \sum_{i\in A} \xi_i \phi(\mathbf{x}_i) \right\|_{\mathcal{H}}^2 \geq C^2 \cdot u^2 \right\} \\
&= \underset{\phi(\mathbf{x}_i)\sim\mathcal{N}(\mathbf{0},\mathbf{\Sigma})}{\mathbf{Pr}} \left\{ \underset{\xi_i\sim\mathcal{R}}{\mathbf{E}} \sum_{i=1}^{n(1-\epsilon)} \sum_{j=1}^{n(1-\epsilon)} \xi_i \xi_j k\left(\mathbf{x}_i, \mathbf{x}_j\right) \geq C^2 \cdot u^2 \right\} \\
&= \underset{\phi(\mathbf{x}_i)\sim\mathcal{N}(\mathbf{0},\mathbf{\Sigma})}{\mathbf{Pr}} \left\{ \sum_{i=1}^{n(1-\epsilon)} k\left(\mathbf{x}_i, \mathbf{x}_i\right) \geq C^2 \cdot u^2 \right\} \leq \underset{\phi(\mathbf{x}_i)\sim\mathcal{N}(\mathbf{0},\mathbf{\Sigma})}{\mathbf{Pr}} \left\{ \sum_{i=1}^{n} \|\phi(\mathbf{x}_i)\|_{\mathcal{H}} \geq C \cdot u \right\} \\
&\overset{(ii)}{\leq} \inf_{\theta>0} \prod_{i=1}^{n} \prod_{j=1}^{p} \exp\left[\frac{\theta^2 \lambda_j}{2} \right] \exp\left[-\theta C \cdot u \right] = \exp\left[-\frac{C^2 \cdot u^2}{2n\operatorname{Tr}(\mathbf{\Sigma})} \right]
\end{aligned}$$

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

Proposition 15 (Probabilistic Bound on Infinite Dimensional Covariance Estimation in the Hilbert-Schmidt Norm *Let $\mathbf{x}_1, \ldots, \mathbf{x}_n$ be i.i.d sampled from $\mathcal{P}$ such that $\phi(\mathbf{x}_i) \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$ (Appendix A). Denote $\mathcal{S}$ as all subsets of $[n]$ with size $n(1-\epsilon)$. We then have for any $u \geq 1$*

$$\underset{\phi(\mathbf{x}_i)\sim\mathcal{N}(\mathbf{0},\mathbf{\Sigma})}{\mathbf{Pr}} \left\{ \max_{A\in\mathcal{S}} \left\| \frac{1}{n} \sum_{i\in A} \phi(\mathbf{x}_i) \otimes \phi(\mathbf{x}_i) - \mathbf{\Sigma} \right\|_{\text{HS}} \geq \frac{\sqrt{3}\operatorname{Tr}(\mathbf{\Sigma})}{\sqrt{n(1-\epsilon)}} \cdot u \right\} \leq e^{-u^2 \operatorname{Tr}(\mathbf{\Sigma})/2}$$

Proof. [1] Let C be a to be determined positive constant and u be a positive constant.

$$\underset{\phi(\mathbf{x}_i)\sim\mathcal{N}(\mathbf{0},\mathbf{\Sigma})}{\mathbf{Pr}} \left\{ \max_{A\in\mathcal{S}} \left\| \frac{1}{n(1-\epsilon)} \sum_{i\in A} \phi(\mathbf{x}_i) \otimes \phi(\mathbf{x}_i) - \mathbf{\Sigma} \right\|_{\text{HS}} \geq C \cdot u \right\}$$

1. In Progress

$$
\begin{aligned}
&\overset{(i)}{\leq} \binom{n}{\lceil n(1-\epsilon)\rceil} \Pr_{\phi(\mathbf{x}_i)\sim\mathcal{N}(\mathbf{0},\mathbf{\Sigma})}\left\{\left\|\frac{1}{n(1-\epsilon)}\sum_{i=1}^{n(1-\epsilon)}\phi(\mathbf{x}_i)\otimes\phi(\mathbf{x}_i)-\mathbf{\Sigma}\right\|_{\text{HS}}\geq C\cdot u\right\}\\
&\overset{(ii)}{\leq} \left(\frac{e}{1-\epsilon}\right)^{n(1-\epsilon)} \Pr_{\phi(\mathbf{x}_i)\sim\mathcal{N}(\mathbf{0},\mathbf{\Sigma})}\left\{\mathop{\mathbf{E}}_{\tilde{\phi}(\mathbf{x}_i)\sim\mathcal{N}(\mathbf{0},\mathbf{\Sigma})}\left\|\frac{1}{n(1-\epsilon)}\sum_{i=1}^{n(1-\epsilon)}\phi(\mathbf{x}_i)\otimes\phi(\mathbf{x}_i)-\tilde{\phi}(\mathbf{x}_i)\otimes\tilde{\phi}(\mathbf{x}_i)\right\|_{\text{HS}}\geq C\cdot u\right\}\\
&\overset{(iii)}{=} \left(\frac{e}{1-\epsilon}\right)^{n(1-\epsilon)} \Pr_{\phi(\mathbf{x}_i)\sim\mathcal{N}(\mathbf{0},\mathbf{\Sigma})}\left\{\mathop{\mathbf{E}}_{\tilde{\phi}(\mathbf{x}_i)\sim\mathcal{N}(\mathbf{0},\mathbf{\Sigma})}\mathop{\mathbf{E}}_{\xi_i\sim\mathcal{R}}\right.\\
&\qquad\qquad\left.\left\|\frac{1}{n(1-\epsilon)}\sum_{i=1}^{n(1-\epsilon)}\xi_i\left(\phi(\mathbf{x}_i)\otimes\phi(\mathbf{x}_i)-\tilde{\phi}(\mathbf{x}_i)\otimes\tilde{\phi}(\mathbf{x}_i)\right)\right\|_{\text{HS}}\geq C\cdot u\right\}\\
&\leq \left(\frac{e}{1-\epsilon}\right)^{n(1-\epsilon)} \Pr_{\phi(\mathbf{x}_i)\sim\mathcal{N}(\mathbf{0},\mathbf{\Sigma})}\left\{\mathop{\mathbf{E}}_{\xi_i\sim\mathcal{R}}\left\|\frac{1}{n(1-\epsilon)}\sum_{i=1}^{n(1-\epsilon)}\xi_i\phi(\mathbf{x}_i)\otimes\phi(\mathbf{x}_i)\right\|_{\text{HS}}\right.\\
&\qquad\qquad\left.+\mathop{\mathbf{E}}_{\tilde{\phi}(\mathbf{x}_i)\sim\mathcal{N}(\mathbf{0},\mathbf{\Sigma})}\mathop{\mathbf{E}}_{\xi_i\sim\mathcal{R}}\left\|\frac{1}{n(1-\epsilon)}\sum_{i=1}^{n(1-\epsilon)}\xi_i\tilde{\phi}(\mathbf{x}_i)\otimes\tilde{\phi}(\mathbf{x}_i)\right\|_{\text{HS}}\geq C\cdot u\right\}\\
&\leq \left(\frac{e}{1-\epsilon}\right)^{n(1-\epsilon)} \Pr_{\phi(\mathbf{x}_i)\sim\mathcal{N}(\mathbf{0},\mathbf{\Sigma})}\left\{\left(\mathop{\mathbf{E}}_{\xi_i\sim\mathcal{R}}\left\|\frac{1}{n(1-\epsilon)}\sum_{i=1}^{n(1-\epsilon)}\xi_i\phi(\mathbf{x}_i)\otimes\phi(\mathbf{x}_i)\right\|_{\text{HS}}^2\right)^{1/2}\right. \qquad (17)\\
&\qquad\qquad\left.+\left(\mathop{\mathbf{E}}_{\tilde{\phi}(\mathbf{x}_i)\sim\mathcal{N}(\mathbf{0},\mathbf{\Sigma})}\mathop{\mathbf{E}}_{\xi_i\sim\mathcal{R}}\left\|\frac{1}{n(1-\epsilon)}\sum_{i=1}^{n(1-\epsilon)}\xi_i\tilde{\phi}(\mathbf{x}_i)\otimes\tilde{\phi}(\mathbf{x}_i)\right\|_{\text{HS}}^2\right)^{1/2}\geq C\cdot u\right\}
\end{aligned}
$$

In (ii) we apply a union bound. We will expand the second term in Equation (17). In (ii) we note that $\phi(\mathbf{x}_i)\otimes\phi(\mathbf{x}_i)-\mathbf{\Sigma}$ is a mean $\mathbf{0}$ operator in the tensor outer-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 an orthonormal basis for the Hilbert Space $\mathcal{H}$. By expanding out the Hilbert-Schmidt Norm, we then have

$$
\begin{aligned}
&\frac{1}{n(1-\epsilon)}\left(\mathop{\mathbf{E}}_{\phi(\mathbf{x}_i)\sim\mathcal{N}(\mathbf{0},\mathbf{\Sigma})}\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{1}{n(1-\epsilon)}\left(\mathop{\mathbf{E}}_{\phi(\mathbf{x}_i)\sim\mathcal{N}(\mathbf{0},\mathbf{\Sigma})}\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_{\mathcal{H}}\right)^{1/2}\\
&=\frac{1}{n(1-\epsilon)}\left(\mathop{\mathbf{E}}_{\phi(\mathbf{x}_i)\sim\mathcal{N}(\mathbf{0},\mathbf{\Sigma})}\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_{\mathcal{H}}\right)^{1/2}\\
&\overset{(iv)}{\leq}\frac{1}{n(1-\epsilon)}\left(\mathop{\mathbf{E}}_{\phi(\mathbf{x}_i)\sim\mathcal{N}(\mathbf{0},\mathbf{\Sigma})}\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_{\mathcal{H}}\right)^{1/2}\\
&=\frac{1}{n(1-\epsilon)}\left(\sum_{i=1}^{n}\mathop{\mathbf{E}}_{\phi(\mathbf{x}_i)\sim\mathcal{N}(\mathbf{0},\mathbf{\Sigma})}\|\phi(\mathbf{x}_i)\otimes\phi(\mathbf{x}_i)\|_{\text{HS}}^2\right)^{1/2}\overset{(v)}{=}\frac{1}{\sqrt{n(1-\epsilon)}}\left(\mathop{\mathbf{E}}_{\phi(\mathbf{x}_i)\sim\mathcal{N}(\mathbf{0},\mathbf{\Sigma})}\|\phi(\mathbf{x}_i)\|_{\mathcal{H}}^4\right)^{1/2}\\
&=\frac{1}{\sqrt{n(1-\epsilon)}}\left(\mathop{\mathbf{E}}_{\phi(\mathbf{x}_i)\sim\mathcal{N}(\mathbf{0},\mathbf{\Sigma})}\left[k^2(\mathbf{x}_i,\mathbf{x}_i)\right]\right)^{1/2}=\frac{1}{\sqrt{n(1-\epsilon)}}\left(2\operatorname{Tr}\left(\mathbf{\Sigma}^2\right)+\operatorname{Tr}\left(\mathbf{\Sigma}\right)^2\right)^{1/2}
\end{aligned}
$$

$$\leq \sqrt{\frac{3}{n(1-\epsilon)}}\operatorname{Tr}(\boldsymbol{\Sigma}) \tag{18}$$

(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 note $\operatorname{Tr}(\boldsymbol{\Sigma}) < \infty$ and therefore even though the covariance operator is infinite-dimensional we are able to get a finite bound on the covariance approximation.

$$\begin{aligned}
(17)\text{ RHS} &\leq \left(\frac{e}{1-\epsilon}\right)^{n(1-\epsilon)} \Pr_{\phi(\mathbf{x}_i)\sim\mathcal{N}(\mathbf{0},\boldsymbol{\Sigma})}\left\{\frac{\sqrt{\operatorname{Tr}(\mathbf{K})}}{n(1-\epsilon)} + \frac{\sqrt{3}\operatorname{Tr}(\boldsymbol{\Sigma})}{\sqrt{n(1-\epsilon)}} \geq C\cdot u\right\} \\
&= \left(\frac{e}{1-\epsilon}\right)^{n(1-\epsilon)} \Pr_{\phi(\mathbf{x}_i)\sim\mathcal{N}(\mathbf{0},\boldsymbol{\Sigma})}\left\{\|\boldsymbol{\Phi}\|_{\text{HS}} \geq n(1-\epsilon)C\cdot u - \sqrt{3n(1-\epsilon)}\operatorname{Tr}(\boldsymbol{\Sigma})\right\} \\
&\leq \left(\frac{e}{1-\epsilon}\right)^{n(1-\epsilon)} \exp\left[-\frac{\left(n(1-\epsilon)C\cdot u + \sqrt{3n(1-\epsilon)}\operatorname{Tr}(\boldsymbol{\Sigma})\right)^2}{n(1-\epsilon)\operatorname{Tr}(\boldsymbol{\Sigma})}\right] \overset{(vi)}{\leq} e^{-u^2\operatorname{Tr}(\boldsymbol{\Sigma})/2}
\end{aligned}$$

In (vi) we chose $C \triangleq \sqrt{3}\operatorname{Tr}(\boldsymbol{\Sigma})/\sqrt{n(1-\epsilon)}$ and then simplify the resultant probability bound. ■

Proposition 16 (Probabilistic Bound on Operator Norm of Outer Products) *Let* $\mathbf{x}_1,\ldots,\mathbf{x}_n$ *be i.i.d sampled from* $\mathcal{P}$ *such that* $\phi(\mathbf{x}_i)\sim\mathcal{N}(\mathbf{0},\boldsymbol{\Sigma})$ *(Appendix A). Denote* $\mathcal{S}$ *as all subsets of* $[n]$ *with size* $n(1-\epsilon)$*. We then have for any* $u \geq 1$

$$\Pr_{\phi(\mathbf{x}_i)\sim\mathcal{N}(\mathbf{0},\boldsymbol{\Sigma})}\left\{\left\|\sum_{i\in A}\phi(\mathbf{x}_i)\otimes\phi(\mathbf{x}_i)\right\|_{\text{op}} \geq \frac{\|\boldsymbol{\Sigma}\|}{\sqrt{n}}\cdot u\right\} \leq e^{-u^2\|\boldsymbol{\Sigma}\|/2}$$

Proof in Progress. Let C be a to be determined positive constant, and u be a positive constant.

$$\begin{aligned}
&\Pr_{\phi(\mathbf{x}_i)\sim\mathcal{N}(\mathbf{0},\boldsymbol{\Sigma})}\left\{\max_{A\in\mathcal{S}}\left\|\sum_{i\in A}\phi(\mathbf{x}_i)\otimes\phi(\mathbf{x}_i)\right\|_{\text{op}} \geq C\cdot u\right\} \\
&\overset{(i)}{\leq} \left(\frac{e}{1-\epsilon}\right)^{n(1-\epsilon)} \Pr_{\phi(\mathbf{x}_i)\sim\mathcal{N}(\mathbf{0},\boldsymbol{\Sigma})}\left\{\left\|\sum_{i=1}^{n(1-\epsilon)}\phi(\mathbf{x}_i)\otimes\phi(\mathbf{x}_i)\right\|_{\text{op}} \geq C\cdot u\right\} \\
&= \left(\frac{e}{1-\epsilon}\right)^{n(1-\epsilon)} \Pr_{\phi(\mathbf{x}_i)\sim\mathcal{N}(\mathbf{0},\boldsymbol{\Sigma})}\left\{\left\|n(1-\epsilon)\boldsymbol{\Sigma} + \sum_{i=1}^{n(1-\epsilon)}\phi(\mathbf{x}_i)\otimes\phi(\mathbf{x}_i) - n(1-\epsilon)\boldsymbol{\Sigma}\right\|_{\text{op}} \geq C\cdot u\right\} \\
&\leq \left(\frac{e}{1-\epsilon}\right)^{n(1-\epsilon)} \Pr_{\phi(\mathbf{x}_i)\sim\mathcal{N}(\mathbf{0},\boldsymbol{\Sigma})}\left\{\left\|\frac{1}{n(1-\epsilon)}\sum_{i=1}^{n(1-\epsilon)}\phi(\mathbf{x}_i)\otimes\phi(\mathbf{x}_i) - \boldsymbol{\Sigma}\right\|_{\text{op}} \geq \frac{C\cdot u}{n(1-\epsilon)} - \lambda_{\max}(\boldsymbol{\Sigma})\right\}
\end{aligned}$$

where (i) follows from a union bound over the subsets of $\mathcal{S}$. We can now follow similar steps to Theorem 15.

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

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

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

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

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. Proof of Theorem 7

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

$$f^{(t+1)} = \text{Proj}_{\mathcal{K}}\left[f^{(t)} - \frac{2\eta}{n(1-\epsilon)}\sum_{i\in S^{(t)}}(f^{(t)}(\mathbf{x}_i) - y_i)\cdot\phi(\mathbf{x}_i)\right] \tag{19}$$

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

$$\begin{aligned}
\|f^{(t+1)} - f^*\|_{\mathcal{H}}^2 &= \left\|\text{Proj}_{\mathcal{K}}\left[f^{(t)} - \nabla_f g(f^{(t)},t^*)\right] - f^*\right\|_{\mathcal{H}}^2 \\
&\leq \left\|f^{(t)} - \nabla_f g(f^{(t)},t^*) - f^*\right\|_{\mathcal{H}}^2 \\
&\leq 2\left\|f^{(t)} - \nabla_f g(f^{(t)},t^*) - \text{Proj}_{\Psi_m}f^*\right\|_{\mathcal{H}}^2 + 2\left\|\text{Proj}_{\Psi_m^\perp}f^*\right\|^2 \\
&= 2\left\|f^{(t)} - \text{Proj}_{\Psi_m}f^*\right\|_{\mathcal{H}}^2 - 4\eta\left\langle\nabla_f g(f^{(t)},t^*), f^{(t)} - \text{Proj}_{\Psi_m}f^*\right\rangle_{\mathcal{H}} \\
&\qquad + 2\eta^2\left\|\nabla_f g(f^{(t)},t^*)\right\|_{\mathcal{H}}^2 + 2\left\|\text{Proj}_{\Psi_m^\perp}f^*\right\|^2
\end{aligned} \tag{20}$$

We will dedicate the rest of the proof to upper bounding the first three terms in Equation (20). We will first bound the second term in Equation (20) by splitting it into terms using the following relation,

$$
\begin{aligned}
&2\eta\langle \nabla_f g(f^{(t)}, t^*), f^{(t)} - \mathrm{Proj}_{\Psi_m} f^* \rangle_{\mathcal{H}} \\
&\overset{(19)}{=} \frac{4\eta}{n(1-\epsilon)} \left\langle f^{(t)} - \mathrm{Proj}_{\Psi_m} f^*, \sum_{i \in S^{(t)}} (f^{(t)}(\mathbf{x}_i) - y_i) \cdot \phi(\mathbf{x}_i) \right\rangle_{\mathcal{H}} \\
&= \frac{4\eta}{n(1-\epsilon)} \left\langle f^{(t)} - \mathrm{Proj}_{\Psi_m} f^*, \sum_{i \in S^{(t)} \cap P} (f^{(t)}(\mathbf{x}_i) - f^*(\mathbf{x}_i) - \mu_i) \cdot \phi(\mathbf{x}_i) \right\rangle_{\mathcal{H}} \\
&\quad + \frac{4\eta}{n(1-\epsilon)} \left\langle f^{(t)} - \mathrm{Proj}_{\Psi_m} f^*, \sum_{i \in S^{(t)} \cap Q} (f^{(t)}(\mathbf{x}_i) - y_i) \cdot \phi(\mathbf{x}_i) \right\rangle_{\mathcal{H}}
\end{aligned} \tag{21}
$$

We will now lower bound the first term of Equation (21).

$$
\begin{aligned}
&\frac{4\eta}{n(1-\epsilon)} \left\langle f^{(t)} - \mathrm{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)} - \mathrm{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)} - \mathrm{Proj}_{\Psi_m} f^* \right) \right\rangle_{\mathcal{H}} \\
&\quad + \frac{4\eta}{n(1-\epsilon)} \left\langle f^{(t)} - \mathrm{Proj}_{\Psi_m} f^*, \left[\sum_{i \in S^{(t)} \cap P} \phi(\mathbf{x}_i) \otimes \phi(\mathbf{x}_i) \right] \left(\mathrm{Proj}_{\Psi_m^{\perp}} f^* \right) \right\rangle_{\mathcal{H}} \\
&\quad - \frac{4\eta}{n(1-\epsilon)} \left\langle f^{(t)} - \mathrm{Proj}_{\Psi_m} f^*, \sum_{i \in S^{(t)} \cap P} \mu_i \phi(\mathbf{x}_i) \right\rangle_{\mathcal{H}} \\
&\overset{(i)}{\geq} \frac{4\eta}{n(1-\epsilon)} \left\langle \tilde{n}\mathbf{\Sigma} + \sum_{i \in S^{(t)} \cap P} \phi(\mathbf{x}_i) \otimes \phi(\mathbf{x}_i) - \tilde{n}\mathbf{\Sigma}, \left(f^{(t)} - \mathrm{Proj}_{\Psi_m} f^* \right) \otimes \left(f^{(t)} - \mathrm{Proj}_{\Psi_m} f^* \right) \right\rangle_{\mathrm{HS}} \\
&\quad - \frac{4\eta}{n(1-\epsilon)} \| f^{(t)} - \mathrm{Proj}_{\Psi_m} f^* \|_{\mathcal{H}} \left\| \sum_{i \in S^{(t)} \cap P} \phi(\mathbf{x}_i) \otimes \phi(\mathbf{x}_i) \right\|_{\mathrm{op}} \| \mathrm{Proj}_{\Psi_m^{\perp}} f^* \|_{\mathcal{H}} \\
&\quad - \frac{4\eta}{n(1-\epsilon)} \| f^{(t)} - \mathrm{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}} \\
&\geq \frac{4\eta}{(1-\epsilon)} (1-2\epsilon) \lambda_m(\mathbf{\Sigma}) \| \mathrm{Proj}_{\Psi_m} [f^{(t)} - f^*] \|_{\mathcal{H}}^2 - \frac{8\eta}{n(1-\epsilon)} \left\| \sum_{i \in S^{(t)} \cap P} \phi(\mathbf{x}_i) \otimes \phi(\mathbf{x}_i) - \tilde{n}\mathbf{\Sigma} \right\|_{\mathrm{HS}} \| f^{(t)} - \mathrm{Proj}_{\Psi_m} f^* \|_{\mathcal{H}}^2 \\
&\quad - \frac{4\eta^2}{n(1-\epsilon)} \| f^{(t)} - \mathrm{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\|_{\mathrm{op}} \left\| \mathrm{Proj}_{\Psi_m^{\perp}} f^* \right\|_{\mathcal{H}} \\
&\qquad + \frac{1}{n(1-\epsilon)} \left\| \sum_{i \in S^{(t)} \cap P} \phi(\mathbf{x}_i) \otimes \phi(\mathbf{x}_i) \right\|_{\mathrm{op}} \| \mathrm{Proj}_{\Psi_m^{\perp}} f^* \|_{\mathcal{H}} \\
&\quad - \frac{4\eta}{n(1-\epsilon)} \| f^{(t)} - \mathrm{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}}
\end{aligned} \tag{22}
$$

where in (i) we define $\tilde{n} \triangleq |S^{(t)} \cap P|$. We will now lower bound the second term of Equation (21).

$$
\begin{aligned}
&\frac{8\eta}{n(1-\epsilon)}\left\langle f^{(t)} - \mathrm{Proj}_{\Psi_m} f^*, \sum_{i \in S^{(t)} \cap Q} (f^{(t)}(\mathbf{x}_i) - y_i) \cdot \phi(\mathbf{x}_i) \right\rangle_{\mathcal{H}} \\
&\quad \leq \frac{8\eta}{n(1-\epsilon)} \|f^{(t)} - \mathrm{Proj}_{\Psi_m} f^*\|_{\mathcal{H}} \left\| \sum_{i \in S^{(t)} \cap Q} \phi(\mathbf{x}_i) \right\|_{\mathcal{H}} \sqrt{\sum_{i \in S^{(t)} \cap Q} (f^{(t)}(\mathbf{x}_i) - y_i)^2} \\
&\quad \overset{(i)}{\leq} \frac{8\eta^{2-\beta/2}}{n^{2-\beta}(1-\epsilon)^{2-\beta}} \|f^{(t)} - \mathrm{Proj}_{\Psi_m} f^*\|_{\mathcal{H}}^2 \left\| \sum_{i \in S^{(t)} \cap Q} \phi(\mathbf{x}_i) \right\|_{\mathcal{H}}^2 + \frac{\eta^{\beta/2}}{n^{\beta}(1-\epsilon)^{\beta}} \sum_{i \in P \setminus S} (f^{(t)}(\mathbf{x}_i) - y_i)^2
\end{aligned} \tag{23}
$$

where (i) follows from Young's Inequality (Theorem 3) for a $\beta \in (0,1)$. We see the step size, η, must have a sub-linear inverse relation to n. We will now upper bound the final term in Equation (20).

$$
\begin{aligned}
&2\eta^2 \|\nabla_f g(f^{(t)}, t^*)\|_{\mathcal{H}}^2 = \frac{8\eta^2}{n^2(1-\epsilon)^2} \left\| \sum_{i \in S^{(t)}} (f^{(t)}(\mathbf{x}_i) - y_i) \cdot \phi(\mathbf{x}_i) \right\|_{\mathcal{H}}^2 \\
&\leq \frac{8\eta^2}{n^2(1-\epsilon)^2} \left\| \sum_{i \in S^{(t)}} \phi(\mathbf{x}_i) \right\|_{\mathcal{H}}^2 \sum_{i \in S^{(t)}} (f^{(t)}(\mathbf{x}_i) - y_i)^2 \overset{(ii)}{\leq} \frac{8\eta^2}{n^2(1-\epsilon)^2} \left\| \sum_{i \in S^{(t)}} \phi(\mathbf{x}_i) \right\|_{\mathcal{H}}^2 \sum_{i \in P} (f^{(t)}(\mathbf{x}_i) - y_i)^2
\end{aligned} \tag{24}
$$

where (ii) follows from the optimality of $S^{(t)}$. We can now complete the upper bound for $\|f^{(t+1)} - \mathrm{Proj}_{\Psi_m} f^*\|_{\mathcal{H}}^2$ combining (22)-(24).

$$
\begin{aligned}
&\|f^{(t+1)} - \mathrm{Proj}_{\Psi_m} f^*\|_{\mathcal{H}}^2 \leq \|f^{(t)} - \mathrm{Proj}_{\Psi_m} f^*\|_{\mathcal{H}}^2 \Bigg(1 - \frac{4\eta}{n(1-\epsilon)} \Bigg(n(1-2\epsilon)\lambda_m(\mathbf{\Sigma}) - \Bigg\| \sum_{i \in S^{(t+1)} \cap P} \phi(\mathbf{x}_i) \otimes \phi(\mathbf{x}_i) - \mathbf{\Sigma} \Bigg\|_{\mathrm{HS}} \Bigg) \\
&+ \frac{4\eta^{2-\beta/2}}{n^{2-\beta}(1-\epsilon)^{2-\beta}} \left\| \sum_{i \in S^{(t)} \cap Q} \phi(\mathbf{x}_i) \right\|_{\mathcal{H}}^2 + \frac{4\eta^2}{n(1-\epsilon)} \left\| \sum_{i \in S^{(t)} \cap P} \phi(\mathbf{x}_i) \otimes \phi(\mathbf{x}_i) \right\|_{\mathrm{op}} \|\mathrm{Proj}_{\Psi_m^{\perp}} f^*\|_{\mathcal{H}} \Bigg) + \frac{\eta^{\beta/2}}{n^{\beta}(1-\epsilon)^{\beta}} \sum_{i \in P} (f^{(t)}(\mathbf{x}_i) - \\
&+ \frac{4\eta}{n(1-\epsilon)} \left\| \sum_{i \in S^{(t)} \cap P} \phi(\mathbf{x}_i) \otimes \phi(\mathbf{x}_i) \right\|_{\mathrm{op}} \|\mathrm{Proj}_{\Psi_m^{\perp}} f^*\|_{\mathcal{H}} + \frac{2\eta}{n(1-\epsilon)} \left\| \sum_{i \in S^{(t)} \cap P} \mu_i \phi(\mathbf{x}_i) \right\|_{\mathcal{H}} + R^2 \lambda_m(\mathbf{\Sigma})
\end{aligned} \tag{25}
$$

We will bound the final term in Equation (24) for $(t+1)$.

$$
\begin{aligned}
&\frac{1}{n(1-\epsilon)} \sum_{i \in P} \left(f^{(t+1)}(\mathbf{x}_i) - y_i \right)^2 = \frac{1}{n(1-\epsilon)} \sum_{i \in P} \left(f^{(t+1)}(\mathbf{x}_i) - f^*(\mathbf{x}_i) - \mu_i \right)^2 \\
&\leq \frac{2}{n(1-\epsilon)} \left(\|f^{(t+1)} - \mathrm{Proj}_{\Psi_m} f^*\|_{\mathcal{H}}^2 + \|\mathrm{Proj}_{\Psi_m^{\perp}} f^*\|_{\mathcal{H}}^2 \right) \left\| \sum_{i \in P} \phi(\mathbf{x}_i) \otimes \phi(\mathbf{x}_i) \right\|_{\mathrm{op}} + \frac{2}{n(1-\epsilon)} \|\boldsymbol{\mu}\|^2
\end{aligned} \tag{26}
$$

We denote the term parameterized by $t \in [T]$,

$$
\Lambda^{(t)} \triangleq \frac{1}{n(1-\epsilon)} \left\| \sum_{i \in P} \phi(\mathbf{x}_i) \otimes \phi(\mathbf{x}_i) \right\| \|f^{(t)} - \mathrm{Proj}_{\Psi_m} f^*\|_{\mathcal{H}}^2 + \frac{1}{n(1-\epsilon)} \sum_{i \in P} \left(f^{(t)}(\mathbf{x}_i) - y_i \right)^2
$$

We can now expand $\|f^{(t+1)} - \mathrm{Proj}_{\Psi_m} f^*\|_{\mathcal{H}}$ from (22)-(24) and set $\beta \triangleq 1/2$.

$$\frac{1}{n(1-\epsilon)} \sum_{i \in P} \left(f^{(t+1)}(\mathbf{x}_i) - y_i\right)^2 \leq \|f^{(t)} - \mathrm{Proj}_{\Psi_m} f^*\|_{\mathcal{H}}^2 \left\|\sum_{i\in P} \phi(\mathbf{x}_i) \otimes \phi(\mathbf{x}_i)\right\|_{\mathrm{op}}$$

$$\underbrace{\cdot 2\Bigg(1 - \frac{4\eta}{n(1-\epsilon)}\Bigg(n(1-\epsilon)\lambda_m(\mathbf{\Sigma}) - \left\|\sum_{i=1}^{n(1-\epsilon)} \phi(\mathbf{x}_i) \otimes \phi(\mathbf{x}_i) - \mathbf{\Sigma}\right\|_{\mathrm{HS}}\Bigg) \cdots}$$

$$\underbrace{\cdots + \frac{4\eta^{4/3}}{[n(1-\epsilon)]^{5/3}} \left\|\sum_{i \in S^{(t)} \cap Q} \phi(\mathbf{x}_i)\right\|_{\mathcal{H}}^2 + \frac{4\eta^2}{n(1-\epsilon)} \left\|\sum_{i\in S^{(t+1)} \cap P}^{n(1-\epsilon)} \phi(\mathbf{x}_i) \otimes \phi(\mathbf{x}_i)\right\|_{\mathrm{op}} \left\|\mathrm{Proj}_{\Psi_m^\perp}\right\|_{\mathcal{H}}\Bigg)}_{\Xi_1}$$

$$+ \frac{1}{n(1-\epsilon)} \underbrace{\left(2 \cdot \left\|\sum_{i\in P} \phi(\mathbf{x}_i) \otimes \phi(\mathbf{x}_i)\right\|_{\mathrm{op}} \frac{\eta^{2/3}}{[n(1-\epsilon)]^{1/3}}\right)}_{\Xi_2}\Bigg) \sum_{i \in P} (f^{(t)}(\mathbf{x}_i) - y_i)^2$$

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

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

We now have all the results to show convergence. First we note that η must be of the form $\eta \triangleq \frac{C}{n(1-\epsilon)}$ so Ξ_2 will be on the order of $\frac{1}{n(1-\epsilon)}$. Then, to obtain the probabilistic sample complexity we create two parameterized probabilistic events. Since $\mathbf{x}_i \in P$ are sampled i.i.d from $\mathcal{P}$, it implies that P is mutually independent. From which it follows that all subsets of P are independent.

$$E_u = \left\{\left\|\sum_{i \in S \cap P} \phi(\mathbf{x}_i) \otimes \phi(\mathbf{x}_i)\right\|_{\mathrm{op}} \leq \tilde{n} \lambda_{\max}(\mathbf{\Sigma}) \cdot u \text{ and } \left\|\sum_{i \in S\cap P} \phi(\mathbf{x}_i) \otimes \phi(\mathbf{x}_i) - \mathbf{\Sigma}\right\|_{\mathrm{HS}} \leq \sqrt{\tilde{n}}\, \mathrm{Tr}(\mathbf{\Sigma}) \cdot u\right\}$$

Invoking Theorem 14 and Theorem 15 together we find $\mathbb{P}\{E_u^c\} \leq e^{-\mathrm{Tr}(\mathbf{\Sigma})u^2/2} + e^{-u^2/2}$ for all $u \geq 1$. Concatenate all the samples of μ_i for $i \in P$ in to a vector $\boldsymbol{\mu} \in \mathbb{R}^{(1-\epsilon)n}$ where $\boldsymbol{\mu}^+ \in \mathbb{R}^{\tilde{n}}$ denotes the samples of μ_i for $i \in S \cap P$ and $\boldsymbol{\mu}^- \in \mathbb{R}^{\gamma \tilde{n}}$ denotes the samples of μ_i for $i \in P \cap S$. We will create a parameterized event over $s \geq 1$.

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

First note that $\boldsymbol{\mu} \sim \mathcal{N}(\mathbf{0}_{n(1-\epsilon)}, \sigma^2 \mathbf{I}_{n(1-\epsilon)})$, therefore for any subset of the indices $A \subseteq [n(1-\epsilon)]$,we have $\boldsymbol{\mu}_A \sim \mathcal{N}(\mathbf{0}_{|A|}, \sigma^2 \mathbf{I}_{|A|})$. We then invoke Theorem 11 and Theorem 12 together and obtain $\mathbb{P}\{E_s^c\} \leq e^{-s/2} + (\sqrt{2}/\log(\gamma\tilde{n}))e^{-s^2} \leq 2.05 e^{-s/2}$ for all $s \geq 1$ and assuming $\gamma \tilde{n} \geq 2$. To satisfy $\Xi_1 \leq 3/4$ we require $n = \Omega\left(\left(\frac{Q_k(1-2\epsilon)}{2(1-\epsilon)\lambda_m(\mathbf{\Sigma})}\right)^{3/2} (1-\epsilon)^{-1}\right)$ and $\eta < \frac{\lambda_m(\mathbf{\Sigma})}{2n(1-\epsilon)\lambda_{\max}(\mathbf{\Sigma})}$. Next, to satisfy $\Xi_2 \leq 3/(4n(1-\epsilon))$, we require $\eta \leq \frac{1}{4u\lambda_{\max}(\mathbf{\Sigma}) n (1-\epsilon)}$, then with probability at least $1 - e^{-u^2/2} - e^{-s/2}$ we have

$$\Lambda^{(t+1)} \leq \frac{3}{4} \cdot \Lambda^{(t)} + \frac{2}{\lambda_{\max}(\mathbf{\Sigma})\sqrt{n(1-\epsilon)}} \left(\|\mathrm{Proj}_{\Psi_m^\perp} f^*\|_{\mathcal{H}} + 2s\sigma \sqrt{\frac{\log(n(1-\epsilon))\, \mathrm{Tr}(\mathbf{\Sigma})}{n(1-\epsilon)\lambda_{\max}^2(\mathbf{\Sigma})}} + \lambda_m(\mathbf{\Sigma}) R^2\right)$$

Now we note that $\Lambda^{(0)} \leq O\left(\frac{\lambda_{\max}(\mathbf{\Sigma})\left\|\text{Proj}_{\Psi_m} f^*\right\|_{\mathcal{H}}^2}{\sqrt{n(1-\epsilon)}} + \frac{\|f^*\|_{\mathcal{H}}^2 \lambda_{\max}(\mathbf{\Sigma})}{\sqrt{n(1-\epsilon)}} + \sigma \cdot s\right)$ w.h.p, and thus after $T = O\left(\left(\frac{\lambda_{\max}(\mathbf{\Sigma})\|f^*\|_{\mathcal{H}}^2}{\sqrt{n}} + \sigma\right)\frac{1}{\varepsilon}\right)$ iterations we obtain $\left\|f^{(T)} - \text{Proj}_{\Psi_m}\right\|_{\mathcal{H}}^2 \leq \varepsilon + O\left(\frac{\|\text{Proj}_{\Psi_m^\perp} f^*\|_{\mathcal{H}}}{\lambda_{\max}(\mathbf{\Sigma})\sqrt{n}} + \sigma\sqrt{\frac{\text{Tr}(\mathbf{\Sigma})\log n}{n\lambda_{\max}^2(\mathbf{\Sigma})}} + \lambda_m(\mathbf{\Sigma})\left\|\text{Proj}_{\Psi_m^\perp} f^{(t)}\right\|_{\mathcal{H}}^2\right)$ w.h.p. Our proof is complete.

Appendix D. Kernelized Binary Classification

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

D.1. Proof of Theorem 9

From Algorithm 2, we have for kernelized binary classification,

$$f^{(t+1)} = \text{Proj}_{\mathcal{K}}\left[f^{(t)} - \frac{\eta}{n(1-\epsilon)}\sum_{i\in S^{(t)}}\left(\sigma(f^{(t)}(\mathbf{x}_i)) - y_i\right)\cdot\phi(\mathbf{x}_i)\right] \tag{27}$$

From which it follows,

$$\begin{aligned}
\|f^{(t+1)} - f^*\|_{\mathcal{H}}^2 &= \left\|\text{Proj}_{\mathcal{K}}\left[f^{(t)} - \frac{\eta}{n(1-\epsilon)}\nabla g(f^{(t)}, t^*)\right] - f^*\right\|_{\mathcal{H}}^2 \\
&\overset{(i)}{\leq} \left\|f^{(t)} - \frac{\eta}{n(1-\epsilon)}\nabla g(f^{(t)}, t^*) - f^*\right\|_{\mathcal{H}}^2 \\
&\leq 2\|f^{(t)} - \text{Proj}_{\Psi_m} f^*\|_{\mathcal{H}}^2 - \frac{4\eta}{n(1-\epsilon)}\left\langle\nabla_f g(f^{(t)}, t^*), f^{(t)} - \text{Proj}_{\Psi_m} f^*\right\rangle_{\mathcal{H}} \\
&\qquad + \frac{2\eta^2}{n^2(1-\epsilon)^2}\|\nabla_f g(f^{(t)}, t^*)\|_{\mathcal{H}}^2 + 2\left\|\text{Proj}_{\Psi_m^\perp} f^*\right\|_{\mathcal{H}}^2
\end{aligned} \tag{28}$$

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 (28).

$$\begin{aligned}
&\frac{2\eta}{n(1-\epsilon)}\left\langle\nabla_f g(f^{(t)}, t^*), f^{(t)} - \text{Proj}_{\Psi_m} f^*\right\rangle_{\mathcal{H}} \\
&\overset{(27)}{=} \left\langle f^{(t)} - \text{Proj}_{\Psi_m} f^*, \frac{2\eta}{n(1-\epsilon)}\sum_{i\in S^{(t)}}\left(\sigma(f^{(t)}(\mathbf{x}_i)) - y_i\right)\cdot\phi(\mathbf{x}_i)\right\rangle_{\mathcal{H}} \\
&= \left\langle f^{(t)} - \text{Proj}_{\Psi_m} f^*, \frac{2\eta}{n(1-\epsilon)}\sum_{i\in S^{(t)}}\left(\sigma(f^{(t)}(\mathbf{x}_i)) - \sigma(f^*(\mathbf{x}_i))\right)\cdot\phi(\mathbf{x}_i)\right\rangle_{\mathcal{H}} \\
&\qquad + \left\langle f^{(t)} - \text{Proj}_{\Psi_m} f^*, \frac{2\eta}{n(1-\epsilon)}\sum_{i\in S^{(t)}}(\sigma(f^*(\mathbf{x}_i)) - y_i)\cdot\phi(\mathbf{x}_i)\right\rangle_{\mathcal{H}}
\end{aligned} \tag{29}$$

We first upper bound upper bound the second term in Equation (29). From the Cauchy-Schwarz Inequality and noting $y_i \in \{0, 1\}$ and $\text{range}(\sigma) \in (0, 1)$, we have the following,

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

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

$$\overset{(ii)}{\leq} \frac{\eta^2}{n^{2-\beta}(1-\epsilon)^{2-\beta}} \|f^{(t)} - \text{Proj}_{\Psi_m} f^*\|_{\mathcal{H}}^2 \Big\| \sum_{i \in S^{(t)}} \phi(\mathbf{x}_i) \Big\|_{\mathcal{H}}^2 + \frac{2}{n^{\beta}(1-\epsilon)^{\beta}} \quad (30)$$

where (ii) follows from Young's Inequality (Theorem 3) 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 $f, \hat{f} \in \mathcal{H}$, if

$$\begin{aligned} \left\langle f - \hat{f}, \nabla h(f) - \nabla h(\hat{f}) \right\rangle_{\mathcal{H}} &\geq \left\langle \nabla^2 h(\tilde{f}), (f - \hat{f}) \otimes (f - \hat{f}) \right\rangle_{\text{HS}} \\ &\geq \left\langle \nabla^2 h(\tilde{f}), \text{Proj}_{\Psi_m} \left[f - \hat{f} \right] \otimes \text{Proj}_{\Psi_m} \left[f - \hat{f} \right] \right\rangle_{\text{HS}} \end{aligned}$$

Then, from the strong convexity of h, there exists a constant C such that the following inequality holds,

$$\begin{aligned} &\left\langle f^{(t)} - f^*, \frac{2\eta}{n(1-\epsilon)} \sum_{i \in S^{(t)} \cap P} \Big(\sigma(f^{(t)}(\mathbf{x}_i)) - \sigma(f^*(\mathbf{x}_i)) \Big) \cdot \phi(\mathbf{x}_i) \right\rangle_{\mathcal{H}} \\ &\gtrsim \frac{2\eta}{n(1-\epsilon)} \left\langle \sum_{i \in S^{(t)}} \phi(\mathbf{x}_i) \otimes \phi(\mathbf{x}_i), \text{Proj}_{\Psi_m} \left[f^{(t)} - f^* \right] \otimes \text{Proj}_{\Psi_m} \left[f^{(t)} - f^* \right] \right\rangle_{\text{HS}} \\ &\overset{(iii)}{\gtrsim} 2\eta \left(\frac{(1-2\epsilon)}{(1-\epsilon)} \lambda_m(\mathbf{\Sigma}) - \left\| \frac{1}{n(1-2\epsilon)} \sum_{i \in S^{(t)} \cap P} \phi(\mathbf{x}_i) \otimes \phi(\mathbf{x}_i) - \mathbf{\Sigma} \right\|_{\text{HS}} \right) \left\| \text{Proj}_{\Psi_m} \left[f^{(t)} - f^* \right] \right\|_{\mathcal{H}}^2 \end{aligned} \quad (31)$$

where (iii) follows from Weyl's inequality (Weyl, 1912). We now briefly analyze the constant introduced in Equation (31).

$$C \triangleq \inf_{\mathbf{x} \sim \mathcal{P}} \sigma(f(\mathbf{x}_i))(1 - \sigma(f(\mathbf{x}_i))) \geq (1/2) \exp\left(-\max_{\mathbf{x} \in P} f(\mathbf{x}) \right) \geq (1/2) \exp\left(-R \max_{\mathbf{x} \in P} \|\phi(\mathbf{x})\|_{\mathcal{H}} \right)$$

The final inequality follows from (Gretton, 2013, Theorem 17). Then, from the bijectivity of the exponential function, we can invoke Theorem 13, and with probability exceeding $1 - \delta$, we have $C \geq (1/2) \exp\left(-R\sqrt{8 \operatorname{Tr}(\mathbf{\Sigma}) \log n \log \frac{4e}{\delta}} \right)$. We will now bound the third term in Equation (28).

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

$$\overset{(v)}{\leq} \frac{\eta^2}{n^2(1-\epsilon)^2}\left\|\sum_{i\in S^{(t)}}\phi(\mathbf{x}_i)\right\|_{\mathcal{H}}^2 \cdot \sum_{i\in P}\Big(\sigma(f^{(t)}(\mathbf{x}_i)) - y_i\Big)^2 \tag{32}$$

where (iv) follows from noting for any $\mathbf{x}\in\mathbb{R}^d$ it holds $\|\mathbf{x}\|_\infty \leq \|\mathbf{x}\|_2$, (v) follows from noting 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 (28)-(32), we obtain

$$\|f^{(t+1)}-f^*\|_{\mathcal{H}}^2 \leq \|f^{(t)}-f^*\|_{\mathcal{H}}^2\left(1-\frac{2C\eta}{n(1-\epsilon)}\lambda_{\min}\Bigg(\sum_{i\in S^{(t)}\cap P}\phi(\mathbf{x}_i)\otimes\phi(\mathbf{x}_i)\Bigg)+\frac{\eta^2}{n^{2-\beta}(1-\epsilon)^{2-\beta}}\left\|\sum_{i\in S^{(t)}}\phi(\mathbf{x}_i)\right\|_{\mathcal{H}}^2\right)$$
$$+\frac{\eta^2}{n^2(1-\epsilon)^2}\left\|\sum_{i\in S^{(t)}}\phi(\mathbf{x}_i)\right\|_{\mathcal{H}}^2\cdot\sum_{i\in P}(\sigma(f^{(t)}(\mathbf{x}_i))-y_i)^2+\frac{2}{n^\beta(1-\epsilon)^\beta}\sum_{i\in P}(\sigma(f^*(\mathbf{x}_i))-y_i)^2 \tag{33}$$

We will now expand out the final term in Equation (33) for $(t+1)$ as we will show we can linearly decrease this term through the iterations.

$$\begin{aligned}
\frac{1}{n(1-\epsilon)}\sum_{i\in P}(\sigma(f^{(t+1)}(\mathbf{x}_i))-y_i)^2 &= \frac{1}{n(1-\epsilon)}\sum_{i\in P}(\sigma(f^{(t+1)}(\mathbf{x}_i))-\sigma(f^*(\mathbf{x}_i))+\sigma(f^*(\mathbf{x}_i))-y_i)^2\\
&\leq \frac{2}{n(1-\epsilon)}\sum_{i\in P}(f^{(t+1)}-f^*)(\mathbf{x}_i)^2+\frac{2}{n(1-\epsilon)}\sum_{i\in P}(\sigma(f^*(\mathbf{x}_i))-y_i)^2\\
&= \frac{2}{n(1-\epsilon)}\left\langle\sum_{i\in P}\phi(\mathbf{x}_i)\otimes\phi(\mathbf{x}_i),(f^{(t+1)}-f^*)\otimes(f^{(t+1)}-f^*)\right\rangle_{\text{HS}}\\
&\qquad+\frac{2}{n(1-\epsilon)}\sum_{i\in P}\left[\mathop{\mathbf{Pr}}_{(\mathbf{x}_i,y_i)\sim\mathcal{P}}\{y_i=+1\mid\mathbf{x}_i\}-y_i\right]^2\\
&\leq \frac{2}{n(1-\epsilon)}\|f^{(t+1)}-f^*\|_{\mathcal{H}}^2\left\|\sum_{i\in P}\phi(\mathbf{x}_i)\otimes\phi(\mathbf{x}_i)\right\|_{\mathcal{H}}+\frac{2}{n(1-\epsilon)}\mathcal{E}_{\text{OPT}}
\end{aligned}$$

Now, we can use Equation (33) to complete the bound.

$$\frac{1}{n(1-\epsilon)}\sum_{i\in P}(\sigma(f^{(t+1)}(\mathbf{x}_i))-y_i)^2 \leq \frac{2}{n(1-\epsilon)}\|f^{(t)}-f^*\|_{\mathcal{H}}^2\left\|\sum_{i\in P}\phi(\mathbf{x}_i)\otimes\phi(\mathbf{x}_i)\right\|_{\mathcal{H}}$$
$$\cdot\left(1-\frac{2C\eta}{n(1-\epsilon)}\lambda_{\min}\left(\mathbf{\Phi}_{\text{TP}}\mathbf{\Phi}_{\text{TP}}^\top\right)+\frac{\eta^2}{n(1-\epsilon)}\left\|\sum_{i\in S^{(t)}}\phi(\mathbf{x}_i)\right\|_{\mathcal{H}}^2\right)+\frac{2}{n(1-\epsilon)}\mathcal{E}_{\text{OPT}} \tag{34}$$

Now we define for all $t\in[T]$,

$$\Lambda^{(t)} \triangleq \frac{1}{n(1-\epsilon)}\left\|\sum_{i\in P}\phi(\mathbf{x}_i)\otimes\phi(\mathbf{x}_i)\right\|\|f^*-f^{(t)}\|_{\mathcal{H}}^2+\frac{1}{n(1-\epsilon)}\sum_{i\in P}(\sigma(f^{(t)}(\mathbf{x}_i))-y_i)^2$$

We then have from Equations (32) and (34),

$$\Lambda^{(t+1)} \leq \max\left\{3\left(1-\frac{2C\eta}{n(1-\epsilon)}\lambda_{\min}\left(\mathbf{\Phi}_{\text{TP}}\mathbf{\Phi}_{\text{TP}}^\top\right)+\frac{\eta^2}{n^{2-\beta}(1-\epsilon)^{2-\beta}}\left\|\sum_{i\in S^{(t)}}\phi(\mathbf{x}_i)\right\|_2^2\right)\right.$$

$$, \frac{\eta^2}{n^2(1-\epsilon)^2}\left\| \sum_{i \in S^{(t)}} \phi(\mathbf{x}_i) \right\|_{\mathcal{H}}^2 \Bigg\} \cdot \Lambda^{(t)} + \frac{2}{n^\beta(1-\epsilon)^\beta} + \frac{2}{n(1-\epsilon)}\mathcal{E}_{\mathrm{OPT}}$$

Solving the quadratic equation, we observe for sufficiently large n such that $\frac{4C^2}{n^2(1-\epsilon)^2}\lambda^2_{\min}(\mathbf{\Phi}_{\mathrm{TP}}\mathbf{\Phi}_{\mathrm{TP}}^\top) \geq \frac{10}{3n(1-\epsilon)}\left\|\sum_{i\in S^{(t)}}\phi(\mathbf{x}_i)\right\|^2$ as the LHS scales in $O(\lambda_{\min}(\mathbf{\Sigma}))$ and the RHS scales in $O(\mathrm{Tr}(\mathbf{\Sigma})/\sqrt{n})$, and choosing

$$\eta \leq \frac{2C\lambda_{\min}\left(\mathbf{\Phi}_{\mathrm{TP}}\mathbf{\Phi}_{\mathrm{TP}}^\top\right)}{\left\|\sum_{i\in S^{(t)}}\phi(\mathbf{x}_i)\right\|_2^2} = O\left(\frac{\lambda_{\min}(\mathbf{\Sigma})}{\mathrm{Tr}(\mathbf{\Sigma})}\right)$$

We then observe there is a linear decrease in $\Lambda^{(t)}$ plus a constant,

$$\Lambda^{(t+1)} \leq \max\left\{\frac{1}{2}, \frac{4C^2\lambda^2_{\min}\left(\mathbf{\Phi}_{\mathrm{TP}}\mathbf{\Phi}_{\mathrm{TP}}^\top\right)}{n^2(1-\epsilon)^2\left\|\sum_{i\in S^{(t)}}\phi(\mathbf{x}_i)\right\|_2^2}\right\} \cdot \Lambda^{(t)} + \frac{2}{n^\beta(1-\epsilon)^\beta} + \frac{2}{n(1-\epsilon)}\mathcal{E}_{\mathrm{OPT}}$$

Then, noting $\sum_{i\in P}(\sigma(f^{(t)}(\mathbf{x}_i)) - y_i)^2 \leq n(1-\epsilon)$. We have $\left\|f^{(t+1)} - f^*\right\|_{\mathcal{H}} \leq \varepsilon + \tilde{O}\left(\frac{\mathcal{E}_{\mathrm{OPT}}}{n(1-\epsilon)}\right) + \tilde{O}\left(\frac{1}{n^\beta(1-\epsilon)^\beta}\right)$ after $T = \tilde{O}\left(\log\left(\left(\frac{\lambda_{\max}(\mathbf{\Sigma})\|f^*\|_{\mathcal{H}}}{\sqrt{n}}\right)\frac{1}{\varepsilon}\right)\right)$ iterations. Our proof is complete. ■

Appendix E. Proofs for Kernelized Multi-Class Classification