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

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

Alex Gittens
RPI CS, gittea@rpi.edu

Abstract

In this paper we propose Subquantile Minimization for learning with adversarial corruption in the training set. Superquantile objectives have been formed in the past in the context of fairness where one wants to learn an underrepresented distribution equally [LPMH21, RRM14]. Our intuition is to learn a more favorable representation of the *majority* class, thus we propose to optimize over the p-subquantile of the loss in the dataset. In particular, we study the Huber-ϵ Contamination Problem for Kernel Learning where the distribution is formed as $\hat{\mathcal{P}} = (1-\varepsilon)\mathcal{P} + \varepsilon\mathcal{Q}$, and we want to find the function $\inf_{f_{\mathbf{w}} \in \mathcal{H}} \mathbb{E}_{\mathcal{D}\sim\mathcal{P}}\left[\ell(f_{\mathbf{w}}; \mathbf{X}, \mathbf{y})\right]$, from the noisy distribution, $\hat{\mathcal{P}}$. We assume the adversary has knowledge of the true distribution of $\mathcal{P}$, and is able to corrupt the covariates and the labels of ε samples. To our knowledge, we are the first to study the problem of general kernel learning in the Huber Contamination Model. In our theoretical analysis, we analyze our non-convex concave objective function with the Moreau Envelope. We show (i) a stationary point with respect to the Moreau Envelope is a good point and (ii) we can reach a stationary point with gradient descent methods. We empirically test Kernel Regression and Kernel Classification on various state of the art datasets and show Subquantile Minimization gives strong results in comparison to the state of the art robust algorithms.

*Preliminary Work

1 Introduction

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

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

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

We similarly propose a robust objective which has a nonconvex-concave objective. This objective has also been proposed recently in [HYwL20] where there has been an analysis in the Binary Classification Task. We show Subquantile Minimization reduces to the same objective in [HYwL20]. 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, [CLKZ21, Dic16]. In [CLKZ21], the feature space, $\phi(\mathbf{x}_i) \sim \mathcal{N}(0, \mathbf{\Sigma})$ where $\mathbf{\Sigma}$ is a diagonal matrix of dimension p, where p can be infinite. In this work, we adopt a similar framework, and with the power of Mercer's Theorem [Mer09], we are able to say $\mathrm{Tr}(\mathbf{\Sigma}) < \infty$. We use this fact extensively in our infinite-dimensional concentration inequalities.

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

$$\left\|f_{\widehat{\mathbf{w}}} - f^*_{\mathbf{w}}\right\|_{\mathcal{H}} \leq O\left(\sigma\right) \tag{1}$$

Kernel Binary Classification:

$$\left\|f_{\widehat{\mathbf{w}}} - f^*_{\mathbf{w}}\right\|_{\mathcal{H}} \leq O\left(\Xi\right) \tag{2}$$

Kernel Multi-Class Classification:

$$\left\|f_{\hat{\mathbf{W}}} - f^*_{\mathbf{W}}\right\| \leq O\left(\Xi\right) \tag{3}$$

1.1 Related Work

The idea of iterative thresholding algorithms for robust learning tasks dates back to 1806 by Legendre [Leg06]. 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 [BJK15]. Bhatia et al. study iterative thresholding for least squares regression / sparse recovery. Their theoretical results

Learning with bad training data via iterative trimmed loss minimization [SS19a]. This work considers optimizing over the bottom-k errors by choosing the αn points with smallest error and then updating the model from these αn. This general model is the same as ours. Theoretically, this work considers only general linear models. Experimentally, this work considers more general machine learning models such as GANS.

Trimmed Maximum Likelihood Estimation for Robust Generalized Linear Model [ADKS22]. This work studies a different class of generalized linear models. Interestingly, they show for Gaussian Regression the iterative trimmed maximum likelihood estimator is able to achieve near minimax optimal error. This work does not consider feature corruption and primarily focuses on the covariates sampled with Gaussian Design from Identity covariance.

Sum of Ranked Range Loss for Supervised Learning [HYwL20]. Hu et al. proposed learning over the bottom k losses, this is an alternative formulation of our algorithm. This is an extension of previous work studying the learning of the top k losses, [FLYH17]. They solve their optimization problem with difference of sums convex solvers. This work considers only the classification task and does not give rigorous error bounds. Subsequent work on analyzing the middle k losses is analyzed in [HYW+23].

The iterative trimmed loss framework with batch Stochastic Gradient Descent (SGD) is analyzed in [SS19b]. They experimentally test their design in deep learning applications such as image classification and Generative Adversarial Networks (GANs).

1.2 Contributions

We will now state our main contributions clearly.

1. We provide a novel theoretical framework using the Moreau Envelope for analyzing the iterative trimmed estimator for machine learning tasks.
2. We provide rigorous error bounds for subquantile minimization in the kernel regression, kernel binary classification, and kernel multi-class classification. Furthermore, we provide our bounds for both label and feature corruption with a general Gaussian Design.
3. We perform experiments on state-of-the-art matrices and show the effectiveness of our algorithm compared to other robust learning procedures. Furthermore, we use our experiments to demonstrate the practicality of our theory.

2 Subquantile Minimization

We propose to optimize over the subquantile of the risk. The p-quantile of a random variable, U, is given as $\mathcal{Q}_p(U)$, this is the largest number, t, such that the probability of $U \leq t$ is at least p.

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

The p-subquantile of the risk is then given by

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

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

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

where t is the p-quantile of the empirical risk. Note that for a fixed t therefore the objective is not concave with respect to $\mathbf{w}$. Thus, to solve this problem we use the iterations from Equation 11 in [RHL+20]. Let $\text{Proj}_{\mathcal{K}}$ be the projection of a function on to the convex set $\mathcal{K} \triangleq \{f \in \mathcal{H} : \|f\|_{\mathcal{H}} \leq R\}$, then our update steps are

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

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

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

3 Theory

To consider theoretical guarantees of Subquantile Minimization, we first analyze the inner and outer optimization problems. We first analyze kernel learning in the presence of corrupted data. Next, we provide error bounds for the two most important kernel learning problems, kernel ridge regression, and kernel classification. Now we will give our first result regarding kernel learning in the Huber ϵ-contamination model. Now we will analyze the two-step minimax optimization steps described in Equations (7) and (8).

Lemma 3. *Let $f(\mathbf{x};\mathbf{w})$ be a convex loss function. Let $\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n$ denote the n data points ordered such that $f(\mathbf{x}_1;\mathbf{w},y_1) \leq f(\mathbf{x}_2;\mathbf{w},y_2) \leq \cdots \leq f(\mathbf{x}_n;\mathbf{w},y_n)$. If we denote $\hat{\nu}_i \triangleq f(\mathbf{x}_i;\mathbf{w},y_i)$, it then follows $\hat{\nu}_{np} \in \arg\max_{t\in\mathbb{R}} g(t,\mathbf{w})$.*

Proof is given in Appendix B.1. From Lemma 3, we see that t will be greater than or equal to the errors of exactly np points. Thus, we are continuously updating over the np minimum errors.

Lemma 4. *Let $\hat{\nu}_i \triangleq f(\mathbf{x}_i;\mathbf{w},y_i)$ s.t. $\hat{\nu}_{i-1} \leq \hat{\nu}_i \leq \hat{\nu}_{i+1}$, if we choose $t^{(k+1)} = \hat{\nu}_{np}$ as by Lemma 3, it then follows $\nabla_{\mathbf{w}} g(t^{(k)}, f_{\mathbf{w}}^{(k)}) = \frac{1}{np}\sum_{i=1}^{np} \nabla f(\mathbf{x}_i; f_{\mathbf{w}}^{(k)}, y_i)$*

Proof is given in Appendix B.2.

3.1 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

$$\ell\left(f_{\mathbf{w}};\mathbf{x}_i,y_i,\right) = \left(f_{\mathbf{w}}(\mathbf{x}_i) - y_i\right)^2 \tag{9}$$

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 $f_{\mathbf{w}}^*$ and push $m \to \infty$.

Theorem 5 (Subquantile Minimization for Kernelized Regression is Good with High Probability)**.** *Let Algorithm 1 be run on a dataset $\mathcal{D} \sim \hat{\mathcal{P}}$ with learning rate $\eta \triangleq \beta^{-1}$ where β is the Lipschitz Gradient Constant given in* **??**. *Suppose*

$$n = O\left(\frac{\mathrm{Tr}(\mathbf{\Sigma})^2}{(1-2\epsilon)\lambda_m^2\left(\mathbf{\Sigma}\right)}\right) \tag{10}$$

Then $\|\mathrm{Proj}_{\Psi_m} f_{\mathbf{w}} - \mathrm{Proj}_{\Psi_m} f_{\mathbf{w}}^*\|_{\mathcal{H}} \leq \varepsilon$, after

$$T = O\left(\frac{\log\left(\|f_{\mathbf{w}}^*\|_{\mathcal{H}} + \sigma\sqrt{n\log n\,\mathrm{Tr}(\mathbf{\Sigma})} + \|\boldsymbol{\xi}\|\sqrt{\mathrm{Tr}(\mathbf{K}_Q)}\right)}{\varepsilon\log\left(\frac{1}{1-\lambda_m}\right)}\right) \tag{11}$$

iterations with high probability.

Full proof with explicit constants is given in **??**. A direct application of Theorem 5 is that learning an infinite dimensional function $f_{\mathbf{w}}^*$ to within ε error in the Hilbert Space Norm requires infinite data. Furthermore, we see that given covariate noise and label noise, our bound requires more iterations dependent on the magnitude of the corruption. Such a result is corroborated in [SST+18]. 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 6 (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*

$$\mathbb{E}\|\mathbf{w}^{(T)} - \mathbf{w}^*\|_2 \leq O\left(\frac{\gamma\sigma}{\sqrt{\lambda_{\min}(\mathbf{\Sigma})}}\right) \tag{12}$$

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 $\text{Proj}_{\Psi_m^\perp}$ where $m = p$ is equal to zero as $\{\varphi_i\}_{i=1}^m$ spans the finite-dimensional space, in which 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.

3.2 Kernelized Binary Classification

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

$$\ell(\mathbf{x}_i, y_i; \mathbf{w}) = -y_i \log\left(\sigma\left(\mathbf{w}^\top \mathbf{x}_i\right)\right) - (1 - y_i)\log\left(1 - \sigma\left(\mathbf{w}^\top \mathbf{x}_i\right)\right) \tag{13}$$

Algorithm 1 (Subquantile Minimization for Binary Classification)**.**

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

Output: Function in $\mathcal{H}$: $\hat{f}$

1. Set the step-size
$$\eta \geq \Omega\left(\frac{\lambda_{\min}\left(\mathbf{X}^\top\mathbf{X}\right)}{2\left\|\sum_{i\in X}\mathbf{x}_i\right\|_2^2}\right)$$
2. Set the number of iterations
$$T = O\left(\log\left(\frac{\sqrt{n}\lambda_{\max}(\mathbf{\Sigma})\left\|f^*\right\|_{\mathcal{H}} + n}{\epsilon}\right)\right)$$
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.
   $$\nabla_f g(t^{(k+1)}, f^{(k)}) \leftarrow \frac{1}{n(1-\epsilon)}\sum_{i\in S^{(k)}}(\sigma(f^{(k)}(\mathbf{x}_i) - y_i)\cdot\mathbf{x}_i^\top$$
   5. Perform Gradient Descent Iteration
   $$f^{(k+1)} \leftarrow f^{(k)} - \eta\nabla_{\mathbf{w}} g(f^{(k)}, t^{(k+1)})$$

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

Theorem 7 (Subquantile Minimization for Binary Classification is Good with High Probability)**.** *Let Algorithm 1 be run on a dataset $\mathcal{D} \sim \hat{\mathcal{P}}$ with learning rate $\eta \triangleq \Omega(L^{-1})$. Then after $O(\log(\frac{\sqrt{n}\lambda_{\max}(\mathbf{\Sigma})\|\mathbf{w}^*\|}{\varepsilon} + n))$*

gradient descent iterations,

$$\|\mathbf{w}^{(T)} - \mathbf{w}^*\|_2 \leq \epsilon + O\left(\mathcal{E}_{\text{OPT}}\right) \tag{14}$$

where $\mathcal{E}_{\text{OPT}} \triangleq \sum_{i \in P} (\sigma(\mathbf{x}_i^\top \mathbf{w}^*) - y_i)^2$.

In Theorem 7, we introduce $\mathcal{E}_{\text{OPT}}$, this can be interpreted as the seperability of the data. Well seperated data will have very small $\mathcal{E}_{\text{OPT}}$ as the points are far from the optimal classification boundary.

3.3 Kernelized Multi-Class Classification

The Negative Log-Likelihood Loss for the the Kernel Multi-Class Classification problem is given by the following equation for a single training pair $(\mathbf{x}_i, y_i)$, note $\mathbf{W} \in \mathbb{R}^{n \times |\mathcal{Y}|}$.

$$\ell\left(\mathbf{x}_i, y_i; f_{\mathbf{W}}\right) = -\sum_{j=1}^{|\mathcal{Y}|} \mathbb{I}\left\{j = y_i\right\} \log\left(\frac{\exp\left(f_{\mathbf{w}_j}(\mathbf{x}_i)\right)}{\sum_{k=1}^{|\mathcal{Y}|} \exp\left(f_{\mathbf{w}_k}(\mathbf{x}_i)\right)}\right) \tag{15}$$

Theorem 8 (Subquantile Minimization for Kernelized Multi-Class Classification is Good with High Probability)**.** *Let Algorithm 1 be run on a dataset* $\mathcal{D} \sim \hat{\mathcal{P}}$ *with learning rate* $\eta \triangleq \beta^{-1}$*. Then,*

$$\left\|f_{\hat{\mathbf{W}}} - f^*_{\mathbf{W}}\right\|_{\mathcal{H}} \leq O\left(\Xi\right) \tag{16}$$

4 Experiments

We perform numerical experiments on state of the art datasets comparing with other state of the art methods.

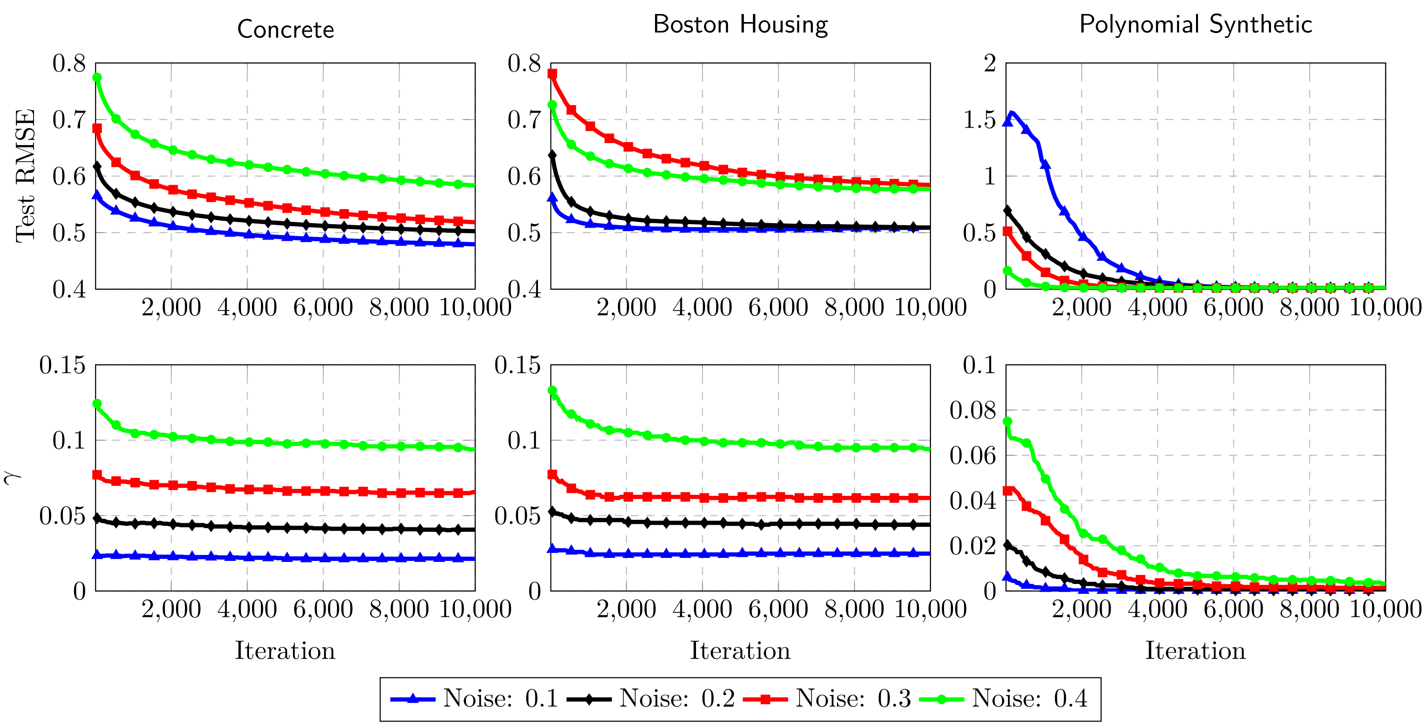

Figure 1: Test RMSE over the iterations in Concrete, Boston Housing, and Polynomial Datasets for Subquantile at different noise levels

In Figure 1, we see the final subquantile has significantly less outliers than the original corruption in the data set. Furthermore, we see there is a greater decrease in the higher outlier settings.

Results. We will clearly state our main findings.

Table 1: Iris ($R = 1$), Glass ($R = 10$), Wine ($R = 100$), and Satimage ($R = 10000$) Datasets. Label Noise is a randomly chosen incorrect label. Feature Noise: $y_{\text{noise}} = 10000y_{\text{original}}$ and $\mathbf{x}_{\text{noise}} = 100\mathbf{x}_{\text{original}}$. The Radial Basis Function is used in all experiments.

Algorithms	Test Accuracy							
	Iris [Fis88]		Glass [Ger87]		Wine [AF91]		Satimage [Sri93]	
	$\epsilon = 0.2(\uparrow)$	$\epsilon = 0.4(\uparrow)$	$\epsilon = 0.2(\uparrow)$	$\epsilon = 0.4(\uparrow)$	$\epsilon = 0.2(\uparrow)$	$\epsilon = 0.4(\uparrow)$	$\epsilon = 0.2(\uparrow)$	$\epsilon = 0.4(\uparrow)$
SVC	$0.977_{(0.0300)}$	$0.757_{(0.1155)}$	$0.553_{(0.0969)}$	$0.435_{(0.0721)}$	$0.928_{(0.0484)}$	$0.678_{(0.1368)}$	$0.882_{(0.0056)}$	$0.732_{(0.0168)}$
TERM	∞	∞	∞	∞	∞	∞	∞	∞
SEVER	∞	∞	∞	∞	∞	∞	∞	∞
SUBQUANTILE	$\mathbf{0.987}_{(0.0163)}$	$\mathbf{0.820}_{(0.1720)}$	$\mathbf{0.656}_{(0.0804)}$	$\mathbf{0.598}_{(0.0889)}$	$\mathbf{0.975}_{(0.0262)}$	$\mathbf{0.867}_{(0.1971)}$	$\mathbf{0.899}_{(0.0076)}$	$\mathbf{0.861}_{(0.0297)}$
Oracle ERM	∞	∞	∞	∞	∞	∞	∞	∞

- **Label Noise vs. Label and Feature Noise**. As suggested by our developed theory, for linear regression or using unbounded kernels, a large multiplicative term increases β and therefore requires more gradient descent iterations to achieve the same distance from a Moreau stationary point. Therefore, from simply increasing the number of gradient descent iterations, we are able to achieve similar RMSE in practice. This happens because the distance from a stationary point and the optimal is not affected by feature noise. This is one of the strengths of our theoretical analysis.
- **Error vs.** ϵ. We find approximately linear increase in the error with increasing ϵ. This can be seen in the γ term, which is upper bounded $\sqrt{\epsilon/(1-2\epsilon)}$. When $\epsilon \to 0.5$, the denominator approaches 0 and therefore our worst case bound increases.
- **Kernel.** Our error bounds are stronger when the dimension of the kernel is lower, i.e. we need more data to obtain the same error bounds. However, in practice, we find many datasets are better approximated by polynomial or RBF kernels, and therefore the γ term is significantly lower.

5 Discussion

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

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

Experiments. From our experiments, we see Subquantile Minimization is competitive with algorithms developed solely for robust linear regression as well as other meta-algorithms. Our theoretical analysis is through the lens of kernel-learning, but the generalization to linear regression 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

[ADKS22] Pranjal Awasthi, Abhimanyu Das, Weihao Kong, and Rajat Sen. Trimmed maximum likelihood estimation for robust generalized linear model. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022. 1

[AF91] Stefan Aeberhard and M. Forina. Wine. UCI Machine Learning Repository, 1991. DOI: https://doi.org/10.24432/C5PC7J. 6

[AMMIL12] Yaser S Abu-Mostafa, Malik Magdon-Ismail, and Hsuan-Tien Lin. *Learning from data*, volume 4. AMLBook New York, 2012. 1

[Ber24] Sergei Bernstein. On a modification of chebyshev's inequality and of the error formula of laplace. *Ann. Sci. Inst. Sav. Ukraine, Sect. Math*, 1(4):38–49, 1924. 15

[BJK15] Kush Bhatia, Prateek Jain, and Purushottam Kar. Robust regression via hard thresholding. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015. 1

[BJKK17] Kush Bhatia, Prateek Jain, Parameswaran Kamalaruban, and Purushottam Kar. Consistent robust regression. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. 1

[Bog98] Vladimir Igorevich Bogachev. *Gaussian measures*. Number 62. American Mathematical Soc., 1998. 10

[CDGS20] Yu Cheng, Ilias Diakonikolas, Rong Ge, and Mahdi Soltanolkotabi. High-dimensional robust mean estimation via gradient descent. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 1768–1778. PMLR, 13–18 Jul 2020. 1

[CDGW19] Yu Cheng, Ilias Diakonikolas, Rong Ge, and David P. Woodruff. Faster algorithms for high-dimensional robust covariance estimation. In Alina Beygelzimer and Daniel Hsu, editors, *Proceedings of the Thirty-Second Conference on Learning Theory*, volume 99 of *Proceedings of Machine Learning Research*, pages 727–757. PMLR, 25–28 Jun 2019. 1

[CLKZ21] Hugo Cui, Bruno Loureiro, Florent Krzakala, and Lenka Zdeborová. Generalization error rates in kernel regression: The crossover from the noiseless to noisy regime. *Advances in Neural Information Processing Systems*, 34:10131–10143, 2021. 1

[Dic16] Lee H Dicker. Ridge regression and asymptotic minimax estimation over spheres of growing dimension. 2016. 1

[DK23] Ilias Diakonikolas and Daniel M Kane. *Algorithmic high-dimensional robust statistics*. Cambridge University Press, 2023. 1

[DKK+19] Ilias Diakonikolas, Gautam Kamath, Daniel M. Kane, Jerry Li, Jacob Steinhardt, and Alistair Stewart. Sever: A robust meta-algorithm for stochastic optimization. In *Proceedings of the 36th International Conference on Machine Learning*, ICML '19, pages 1596–1606. JMLR, Inc., 2019. 1

[FB81] Martin A. Fischler and Robert C. Bolles. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Commun. ACM*, 24(6):381–395, jun 1981. 1

[Fis88] R. A. Fisher. Iris. UCI Machine Learning Repository, 1988. DOI: https://doi.org/10.24432/C56C76. 6

[FLYH17] Yanbo Fan, Siwei Lyu, Yiming Ying, and Baogang Hu. Learning with average top-k loss. 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. 2

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

[Ger87] B. German. Glass Identification. UCI Machine Learning Repository, 1987. DOI: https://doi.org/10.24432/C5WW2P. 6

[Gre13] 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. 13

[HR09] Peter J. Huber and Elvezio. Ronchetti. *Robust statistics.* Wiley series in probability and statistics. Wiley, Hoboken, N.J., 2nd ed. edition, 2009. 1

[HSS08] Thomas Hofmann, Bernhard Schölkopf, and Alexander J. Smola. Kernel methods in machine learning. *The Annals of Statistics*, 36(3):1171 – 1220, 2008. 1

[HYW+23] Shu Hu, Zhenhuan Yang, Xin Wang, Yiming Ying, and Siwei Lyu. Outlier robust adversarial training. *arXiv preprint arXiv:2309.05145*, 2023. 2

[HYwL20] Shu Hu, Yiming Ying, xin wang, and Siwei Lyu. Learning by minimizing the sum of ranked range. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 21013–21023. Curran Associates, Inc., 2020. 1, 2

[JZL+18] Lu Jiang, Zhengyuan Zhou, Thomas Leung, Li-Jia Li, and Li Fei-Fei. Mentornet: Learning data-driven curriculum for very deep neural networks on corrupted labels. In *ICML*, 2018. 1

[KLA18] Ashish Khetan, Zachary C. Lipton, and Anima Anandkumar. Learning from noisy singly-labeled data. In *International Conference on Learning Representations*, 2018. 1

[LBSS21] Tian Li, Ahmad Beirami, Maziar Sanjabi, and Virginia Smith. Tilted empirical risk minimization. In *International Conference on Learning Representations*, 2021. 1

[Leg06] Adrien M Legendre. *Nouvelles methodes pour la determination des orbites des comtes: avec un supplement contenant divers perfectionnemens de ces methodes et leur application aux deux cometes de 1805.* Courcier, 1806. 1

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

[Mer09] 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. 1

[MGJK19] Bhaskar Mukhoty, Govind Gopakumar, Prateek Jain, and Purushottam Kar. Globally-convergent iteratively reweighted least squares for robust regression problems. In Kamalika Chaudhuri and Masashi Sugiyama, editors, *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pages 313–322. PMLR, 16–18 Apr 2019. 1

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

[PBR19] Adarsh Prasad, Sivaraman Balakrishnan, and Pradeep Ravikumar. A unified approach to robust mean estimation. *arXiv preprint arXiv:1907.00927*, 2019. 1

[PSBR18] Adarsh Prasad, Arun Sai Suggala, Sivaraman Balakrishnan, and Pradeep Ravikumar. Robust estimation via robust gradient estimation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 82, 2018. 1

[RHL+20] Meisam Razaviyayn, Tianjian Huang, Songtao Lu, Maher Nouiehed, Maziar Sanjabi, and Mingyi Hong. Nonconvex min-max optimization: Applications, challenges, and recent theoretical advances. *IEEE Signal Processing Magazine*, 37(5):55–66, 2020. 2

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

[Sri93] Ashwin Srinivasan. Statlog (Landsat Satellite). UCI Machine Learning Repository, 1993. DOI: https://doi.org/10.24432/C55887. 6

[SS19a] 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. 1

[SS19b] Yanyao Shen and Sujay Sanghavi. Learning with bad training data via iterative trimmed loss minimization. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 5739–5748. PMLR, 09–15 Jun 2019. 2

[SST+18] Ludwig Schmidt, Shibani Santurkar, Dimitris Tsipras, Kunal Talwar, and Aleksander Madry. Adversarially robust generalization requires more data. *Advances in neural information processing systems*, 31, 2018. 3

[Ver20] Roman Vershynin. High-dimensional probability. *University of California, Irvine*, 2020. 15

[Wey12] 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. 18

[You12] 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. 18

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

$$\phi(\mathbf{x}_i) \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}) \tag{17}$$

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

Assumption 10 (Normal Residuals)**.** The residual is defined as $\mu_i \triangleq f^*_{\mathbf{w}}(\mathbf{x}_i) - y_i$. Then we assume for some $\sigma > 0$, it follows

$$\mu_i \sim \mathcal{N}(0, \sigma^2) \tag{18}$$

Proposition 11 (Concentration for functions of a Gaussian Vector [Bog98])**.** *Suppose h is a Lipschitz function on vectors, i.e.*

$$|h(\mathbf{x}) - h(\mathbf{y})| \leq \|h\|_{\text{lip}} \|\mathbf{x} - \mathbf{y}\| \tag{19}$$

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

$$\mathbb{P}\left\{|h(\mathbf{g}) - \mathbb{E}h(\mathbf{g})| \geq T\right\} \leq 2\exp\left[-\frac{t^2}{2\|h\|^2_{\text{lip}}}\right] \tag{20}$$

Lemma 12 (Maximum of Gaussians)**.** *Let $\mu_1, \ldots, \mu_n \sim \mathcal{N}(0, \sigma^2)$ for some $\sigma > 0$. Then it follows*

$$\mathop{\mathbb{E}}_{\mu_i \sim \mathcal{N}(0,\sigma^2)} \max_{i \in [n]} |\mu_i| \leq \sigma\sqrt{2\log n} + \frac{\sigma^2}{\sqrt{\pi \log n}} \tag{21}$$

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

$$\mathop{\mathbb{E}}_{\mu_i \sim \mathcal{N}(0,\sigma^2)} \max_{i \in [n]} |\mu_i| = \int_0^\infty \mathop{\mathbb{P}}_{\mu_i \sim \mathcal{N}(0,\sigma^2)} \left\{\max_{i \in [n]} |\mu_i| > t\right\} dt \overset{(i)}{\leq} c_1 + n \int_{c_1}^\infty \mathop{\mathbb{P}}_{\mu_i \sim \mathcal{N}(0,\sigma^2)} \{|\mu_i| \geq t\} \, dt \tag{22}$$

$$\overset{(ii)}{=} c_1 + 2n \int_{c_1}^\infty \mathop{\mathbb{P}}_{\mu_i \sim \mathcal{N}(0,\sigma^2)} \{\mu_i \geq t\} \, dt = c_1 + 2n \int_{c_1}^\infty \int_t^\infty \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x}{\sigma}\right)^2} dx dt \tag{23}$$

$$\leq c_1 + \frac{n}{\sigma}\sqrt{\frac{2}{\pi}} \int_{c_1}^\infty \int_t^\infty \left(\frac{x}{t}\right) e^{-\frac{1}{2}\left(\frac{x}{\sigma}\right)^2} dx dt = c_1 + n\sigma\sqrt{\frac{2}{\pi}} \int_{c_1}^\infty \frac{e^{-\frac{1}{2}\left(\frac{t}{\sigma}\right)^2}}{t} dt \tag{24}$$

$$\leq c_1 + n\sigma\sqrt{\frac{2}{\pi}} \int_{c_1}^\infty \left(\frac{t}{c_1}\right) e^{-\frac{1}{2}\left(\frac{t}{\sigma}\right)^2} dt = c_1 + n\sigma^3\sqrt{\frac{2}{\pi}} \frac{e^{-\frac{1}{2}\left(\frac{c_1}{\sigma}\right)^2}}{c_1} \tag{25}$$

(i) follows from a union bound and noting for a i.i.d sequence of random variables $\{X_i\}_{i \in [n]}$ and a constant C, it follows $\mathbb{P}\{\max_{i \in [n]} X_i \geq C\} = n\mathbb{P}\{X \geq C\}$ where X is sampled from the same distribution as each X_i. (ii) follows from the symmetricity of the Gaussian distribution about zero. From here, we choose $c_1 \triangleq \sigma\sqrt{2\log n}$. Then we have,

$$\mathop{\mathbb{E}}_{\mu_i \sim \mathcal{N}(0,\sigma^2)} \max_{i \in [n]} |\mu_i| \leq \sigma\sqrt{2\log n} + \frac{\sigma^2}{\sqrt{\pi \log n}} \tag{26}$$

This completes the proof. ■

Proposition 13. *Let $\mu_1, \ldots, \mu_n \sim \mathcal{N}(0, \sigma^2)$ for some $\sigma > 0$, then it follows for any $s \geq 1$*

$$\mathbb{P}\left\{\max_{i \in [n]} |\mu_i| \geq \sigma\sqrt{2\log n} \cdot s\right\} \leq \frac{\sqrt{2}}{\log n} e^{-s^2} \tag{27}$$

Proof. The proof follows simply using similar steps as in the proof of Lemma 12. Let C be a positive constant to be determined.

$$\underset{\mu_i\sim\mathcal{N}(0,\sigma^2)}{\mathbb{P}}\left\{\max_{i\in[n]}|\mu_i|\geq C\cdot s\right\} = 2n\underset{\mu\sim\mathcal{N}(0,\sigma^2)}{\mathbb{P}}\{\mu\geq C\cdot s\} = n\sqrt{\frac{2}{\pi}}\int_{C\cdot s}^{\infty}\left(\frac{1}{\sigma}\right)e^{-\frac{1}{2}\left(\frac{x}{\sigma}\right)^2}dx \tag{28}$$

$$\leq 2\sigma n\left(\frac{1}{C\cdot s}\right)e^{-\frac{1}{2}\left(\frac{C\cdot s}{\sigma}\right)^2} \leq \frac{\sqrt{2}n^{1-s^2}}{s\log n} \leq \frac{\sqrt{2}}{\log n}e^{-s^2} \tag{29}$$

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

Proposition 14. *Let $\mu_1,\ldots,\mu_n \sim \mathcal{N}(0,\sigma^2)$ for some $\sigma>0$, then it follows for any $s\geq 1$,*

$$\mathbb{P}\left\{\sum_{i=1}^{n}\mu_i^2 \geq n\sigma^2\cdot s\right\} \leq e^{-s/2} \tag{30}$$

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

$$\underset{\boldsymbol{\mu}\sim\mathcal{N}(\mathbf{0},\sigma^2\mathbf{I})}{\mathbb{P}}\left\{\|\boldsymbol{\mu}\|^2\geq C\cdot s\right\} = \underset{\boldsymbol{\mu}\sim\mathcal{N}(\mathbf{0},\sigma^2\mathbf{I})}{\mathbb{P}}\left\{\|\boldsymbol{\mu}\|\geq\sqrt{C\cdot s}\right\} = \underset{\mathbf{g}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}{\mathbb{P}}\left\{\left\|\mathbf{U}\boldsymbol{\Lambda}^{1/2}\mathbf{g}\right\|\geq\sqrt{C\cdot s}\right\} \tag{31}$$

$$= \underset{\mathbf{g}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}{\mathbb{P}}\left\{\sqrt{\sum_{i=1}^{n}\lambda_i g_i^2}\geq\sqrt{C\cdot s}\right\} \leq \underset{\mathbf{g}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}{\mathbb{P}}\left\{\sum_{i=1}^{n}\sqrt{\lambda_i}g_i\geq\sqrt{C\cdot s}\right\} \tag{32}$$

$$\leq \inf_{\theta>0}\underset{g_i\sim\mathcal{N}(0,1)}{\mathbb{E}}\left[\prod_{i=1}^{n}\exp\left(\theta\sqrt{\lambda_i}g_i\right)\right]\exp\left[-\theta\sqrt{C\cdot s}\right] \tag{33}$$

$$= \inf_{\theta>0}\exp\left[\frac{\mathrm{Tr}(\boldsymbol{\Lambda})}{2}\theta^2-\theta\sqrt{C\cdot s}\right] = \exp\left[-\frac{(C\cdot s)}{2\,\mathrm{Tr}(\boldsymbol{\Lambda})}\right] \tag{34}$$

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

Lemma 15 (Maximum of Squared Gaussians)**.** *Let $\mu_1,\ldots,\mu_n\sim\mathcal{N}(0,\sigma^2)$ for $\sigma>0$, $n>1$. Then it follows*

$$\underset{\mu_i\sim\mathcal{N}(0,\sigma^2)}{\mathbb{E}}\max_{i\in[n]}\mu_i^2 \leq 2\sigma^2\log(n)+\left(\sigma^3\sqrt{\frac{8}{\pi}}\right)\left(1+\frac{1}{\log(n)}\right) \tag{35}$$

Proof. Our proof follows similarly to the proof for Lemma 12.

$$\underset{\mu_i\sim\mathcal{N}(0,\sigma^2)}{\mathbb{E}}\max_{i\in\mathbb{N}}\mu_i^2 = \int_0^\infty\underset{\mu_i\sim\mathcal{N}(0,\sigma^2)}{\mathbb{P}}\left\{\max_{i\in[n]}\mu_i^2\geq t\right\}dt \leq c_2+n\int_{c_2}^{\infty}\underset{\mu_i\sim\mathcal{N}(0,\sigma^2)}{\mathbb{P}}\left\{|\mu_i|\geq\sqrt{t}\right\}dt \tag{36}$$

$$= c_2+2n\int_{c_2}^{\infty}\underset{\mu_i\sim\mathcal{N}(0,\sigma^2)}{\mathbb{P}}\left\{\mu_i\geq\sqrt{t}\right\}dt = c_2+2n\int_{c_2}^{\infty}\int_{\sqrt{t}}^{\infty}\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{x}{\sigma}\right)^2}dxdt \tag{37}$$

$$= c_2+n\sigma\sqrt{\frac{2}{\pi}}\int_{c_2}^{\infty}\frac{e^{-\frac{1}{2}\left(\frac{t}{\sigma^2}\right)}}{\sqrt{t}}dt \overset{(i)}{\leq} c_2+n\sigma\sqrt{\frac{2}{\pi}}\int_{c_2}^{\infty}\left(\frac{t}{c_2}\right)e^{-\frac{1}{2}\left(\frac{t}{\sigma^2}\right)}dt \tag{38}$$

$$\leq c_2+\left(\sqrt{\frac{2}{\pi}}\right)\frac{n\sigma\left(4\sigma^4+2c_2\sigma^2\right)e^{-\frac{c_2}{2\sigma^2}}}{c_2} \tag{39}$$

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

$$\underset{\mu_i\sim\mathcal{N}(0,\sigma^2)}{\mathbb{E}}\max_{i\in[n]}\mu_i^2 \leq 2\sigma^2\log(n)+\left(2\sigma^3\sqrt{\frac{2}{\pi}}\right)\left(1+\frac{1}{\log(n)}\right) \tag{40}$$

This completes the proof. ∎

Proposition 16 (Expected Maximum P_k)**.** *Let $\mathbf{x}_i \sim \mathbb{P}$ such that $\phi(\mathbf{x}_i) \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$ from Assumption 9. Then it follows for any $s \geq 1$*

$$\mathop{\mathbb{E}}_{\phi(\mathbf{x}_i)\sim\mathcal{N}(\mathbf{0},\mathbf{\Sigma})}\left[\max_{i\in[n]} k(\mathbf{x}_i, \mathbf{x}_i)\right] \leq 2\operatorname{Tr}(\mathbf{\Sigma})\log\left(\frac{n\cdot s}{2\operatorname{Tr}(\mathbf{\Sigma})}\right) + \frac{1}{s} \tag{41}$$

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

$$\begin{aligned}
\mathop{\mathbb{E}}_{\phi(\mathbf{x}_i)\sim\mathcal{N}(\mathbf{0},\mathbf{\Sigma})}\left[\max_{i\in[n]} k(\mathbf{x}_i, \mathbf{x}_i)\right] &\leq C + \int_C^\infty \mathop{\mathbb{P}}_{\phi(\mathbf{x}_i)\sim\mathcal{N}(\mathbf{0},\mathbf{\Sigma})}\left\{\max_{i\in[n]} k(\mathbf{x}_i, \mathbf{x}_i) \geq t\right\} dt &(42)\\
&\overset{(i)}{\leq} C + n\int_C^\infty \mathop{\mathbb{P}}_{\phi(\mathbf{x})\sim\mathcal{N}(\mathbf{0},\mathbf{\Sigma})}\{k(\mathbf{x}, \mathbf{x}) \geq t\}\, dt &(43)\\
&\overset{(ii)}{=} C + n\int_C^\infty \mathop{\mathbb{P}}_{\phi(\mathbf{x})\sim\mathcal{N}(0,\mathbf{\Sigma})}\left\{\|\phi(\mathbf{x})\|_{\mathcal{H}} \geq \sqrt{t}\right\} dt &(44)\\
&\overset{(iii)}{=} C + n\int_C^\infty \mathop{\mathbb{P}}_{\psi(\mathbf{x})\sim\mathcal{N}(\mathbf{0},\mathbf{I})}\left\{\left\|\mathbf{\Psi}\mathbf{\Lambda}^{1/2}\psi(\mathbf{x})\right\|_{\mathcal{H}} \geq \sqrt{t}\right\} dt &(45)\\
&\leq C + n\int_C^\infty \mathop{\mathbb{P}}_{\psi_i(\mathbf{x})\sim\mathcal{N}(0,1)}\left\{\sum_{i=1}^p \sqrt{\lambda_i}\psi_i(\mathbf{x}) \geq \sqrt{t}\right\} dt &(46)\\
&\leq C + n\int_C^\infty \inf_{\theta>0}\mathop{\mathbb{E}}_{\psi(\mathbf{x})\sim\mathcal{N}(0,1)}\left[\prod_{i=1}^p \exp\left(\theta\sqrt{\lambda_i}\psi_i(\mathbf{x})\right)\right]\exp\left(-\theta\sqrt{t}\right) dt &(47)\\
&= C + n\int_C^\infty \inf_{\theta>0}\exp\left[\theta^2\frac{\operatorname{Tr}(\mathbf{\Sigma})}{2} - \theta\sqrt{t}\right] dt = C + n\int_C^\infty \exp\left[-\frac{t}{2\operatorname{Tr}(\mathbf{\Sigma})}\right] dt &(48)\\
&= C + \frac{n}{2\operatorname{Tr}(\mathbf{\Sigma})}\exp\left[-\frac{C}{2\operatorname{Tr}(\mathbf{\Sigma})}\right] &(49)
\end{aligned}$$

See (i) from the proof of Lemma 12. (ii) follows from the reproducing property. In (iii) we define $\psi(\mathbf{x})$ as the whitened RKHS function. Setting $C \triangleq 2\operatorname{Tr}(\mathbf{\Sigma})\log(s\cdot n/(2\operatorname{Tr}(\mathbf{\Sigma})))$ completes the proof. ■

Lemma 17 (Norm of Functions with Gaussian Design in the Reproducing Kernel Hilbert Space)**.** *Let $\mathbf{x}_i \sim \mathbb{P}$ such that $\phi(\mathbf{x}_i) \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$ from Assumption 9 and Assumption 10. Then, it follows*

$$\mathop{\mathbb{E}}_{\phi(\mathbf{x}_i)\sim\mathcal{N}(\mathbf{0},\mathbf{\Sigma})}\mathop{\mathbb{E}}_{\mu_i\sim\mathcal{N}(0,\sigma^2)}\left\|\sum_{i=1}^n \mu_i\phi(\mathbf{x}_i)\right\|_{\mathcal{H}} \leq O\left(\sigma\sqrt{n\log n\operatorname{Tr}(\mathbf{\Sigma})}\right) \tag{50}$$

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

$$\begin{aligned}
&\mathop{\mathbb{E}}_{\phi(\mathbf{x}_i)\sim\mathcal{N}(\mathbf{0},\mathbf{\Sigma})}\mathop{\mathbb{E}}_{\mu_i\sim\mathcal{N}(0,\sigma^2)}\left\|\sum_{i=1}^n \mu_i\phi(\mathbf{x}_i)\right\|_{\mathcal{H}} &(51)\\
&\qquad\qquad\leq \mathop{\mathbb{E}}_{\phi(\mathbf{x}_i)\sim\mathcal{N}(\mathbf{0},\mathbf{\Sigma})}\mathop{\mathbb{E}}_{\mu_i\sim\mathcal{N}(0,\sigma^2)}\max_{i\in[n]}|\mu_i|\left\|\sum_{i=1}^n \phi(\mathbf{x}_i)\right\|_{\mathcal{H}}\\
&\overset{(i)}{\leq} O\left(\sigma\sqrt{\log n}\right)\mathop{\mathbb{E}}_{\phi(\mathbf{x}_i)\sim\mathcal{N}(\mathbf{0},\mathbf{\Sigma})}\mathop{\mathbb{E}}_{\xi_i\sim\mathcal{R}}\left\|\sum_{i=1}^n \xi_i\phi(\mathbf{x}_i)\right\|_{\mathcal{H}} &(52)\\
&\overset{(ii)}{\leq} O\left(\sigma\sqrt{\log n}\right)\left(\mathop{\mathbb{E}}_{\phi(\mathbf{x}_i)\sim\mathcal{N}(\mathbf{0},\mathbf{\Sigma})}\mathop{\mathbb{E}}_{\xi_i\sim\mathcal{R}}\left\|\sum_{i=1}^n \xi_i\phi(\mathbf{x}_i)\right\|_{\mathcal{H}}^2\right)^{1/2} &(53)\\
&= O\left(\sigma\sqrt{\log n}\right)\left(\mathop{\mathbb{E}}_{\phi(\mathbf{x}_i)\sim\mathcal{N}(\mathbf{0},\mathbf{\Sigma})}\mathop{\mathbb{E}}_{\xi_i\sim\mathcal{R}}\left\langle\sum_{i=1}^n \xi_i\phi(\mathbf{x}_i), \sum_{j=1}^n \xi_j\phi(\mathbf{x}_j)\right\rangle_{\mathcal{H}}\right)^{1/2} &(54)
\end{aligned}$$

$$\overset{(iii)}{=} O\left(\sigma\sqrt{\log n}\right)\left(\underset{\phi(\mathbf{x}_i)\sim\mathcal{N}(\mathbf{0},\mathbf{\Sigma})}{\mathbb{E}}\ \underset{\xi_i\sim\mathcal{R}}{\mathbb{E}}\sum_{i=1}^{n}\sum_{j=1}^{n}\xi_i\xi_j k(\mathbf{x}_i,\mathbf{x}_j)\right)^{1/2} \tag{55}$$

$$\overset{(iv)}{=} O\left(\sigma\sqrt{\log n}\right)\left(\underset{\phi(\mathbf{x}_i)\sim\mathcal{N}(\mathbf{0},\mathbf{\Sigma})}{\mathbb{E}}\sum_{i=1}^{n}k(x_i,x_i)\right)^{1/2} \tag{56}$$

$$= O\left(\sigma\sqrt{n\log n\operatorname{Tr}(\mathbf{\Sigma})}\right) \tag{57}$$

(i) follows from applying Lemma 12. (ii) follows from Jensen's Inequality. (iii) follows from the definition of the kernel [Gre13]. (iv) holds as we have $\mathbb{E}[\xi_i\xi_j] = \delta_{i,j}$, where δ is the Kronecker Delta function. ■

Proposition 18 (Probabilistic bound on Norm of Functions with Gaussian Design in the Reproducing Kernel Hilbert Space)**.** *Let $\mathbf{x}_i \sim \mathcal{P}$ such that $\phi(\mathbf{x}_i) \sim \mathcal{N}(\mathbf{0},\mathbf{\Sigma})$ (Assumption 9). Then it follows*

$$\underset{\phi(\mathbf{x}_i)\sim\mathcal{N}(\mathbf{0},\mathbf{\Sigma})}{\mathbb{P}}\left\{\left\|\sum_{i=1}^{n}\phi(\mathbf{x}_i)\right\|_{\mathcal{H}} \geq \sqrt{n\operatorname{Tr}(\mathbf{\Sigma})}\cdot u\right\} \leq e^{-u^2/2} \tag{58}$$

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

$$\underset{\phi(\mathbf{x}_i)\sim\mathcal{N}(\mathbf{0},\mathbf{\Sigma})}{\mathbb{P}}\left\{\left\|\sum_{i=1}^{n}\phi(\mathbf{x}_i)\right\|_{\mathcal{H}} \geq C\cdot u\right\} = \underset{\phi(\mathbf{x}_i)\sim\mathcal{N}(\mathbf{0},\mathbf{\Sigma})}{\mathbb{P}}\left\{\underset{\xi_i\sim\mathcal{R}}{\mathbb{E}}\left\|\sum_{i=1}^{n}\xi_i\phi(\mathbf{x}_i)\right\|_{\mathcal{H}} \geq C\cdot u\right\} \tag{59}$$

$$\overset{(i)}{\leq} \underset{\phi(\mathbf{x}_i)\sim\mathcal{N}(\mathbf{0},\mathbf{\Sigma})}{\mathbb{P}}\left\{\underset{\xi_i\sim\mathcal{R}}{\mathbb{E}}\left\|\sum_{i=1}^{n}\xi_i\phi(\mathbf{x}_i)\right\|_{\mathcal{H}}^2 \geq C^2\cdot u^2\right\} \tag{60}$$

$$= \underset{\phi(\mathbf{x}_i)\sim\mathcal{N}(\mathbf{0},\mathbf{\Sigma})}{\mathbb{P}}\left\{\underset{\xi_i\sim\mathcal{R}}{\mathbb{E}}\sum_{i=1}^{n}\sum_{j=1}^{n}\xi_i\xi_j k\left(\mathbf{x}_i,\mathbf{x}_j\right) \geq C^2\cdot u^2\right\} \tag{61}$$

$$= \underset{\phi(\mathbf{x}_i)\sim\mathcal{N}(\mathbf{0},\mathbf{\Sigma})}{\mathbb{P}}\left\{\sum_{i=1}^{n}k\left(\mathbf{x}_i,\mathbf{x}_i\right) \geq C^2\cdot u^2\right\} \leq \underset{\phi(\mathbf{x}_i)\sim\mathcal{N}(\mathbf{0},\mathbf{\Sigma})}{\mathbb{P}}\left\{\sum_{i=1}^{n}\|\phi(\mathbf{x}_i)\|_{\mathcal{H}} \geq C\cdot u\right\} \tag{62}$$

$$\overset{(ii)}{\leq} \inf_{\theta>0}\prod_{i=1}^{n}\prod_{j=1}^{p}\exp\left[\frac{\theta^2\lambda_j}{2}\right]\exp\left[-\theta C\cdot u\right] = \exp\left[-\frac{C^2\cdot u^2}{2n\operatorname{Tr}(\mathbf{\Sigma})}\right] \tag{63}$$

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.

Lemma 19 (Infinite Dimensional Covariance Estimation in the Hilbert-Schmidt Norm)**.** *Let $\mathbf{\Sigma} \triangleq \mathbb{E}_{\phi(\mathbf{x}_i)\sim\mathbb{P}}[\phi(\mathbf{x}_i)\otimes\phi(\mathbf{x}_i)]$. Then let $\mathbf{x}_1,\ldots,\mathbf{x}_n$ be i.i.d sampled from $\mathbb{P}$ such that $\phi(\mathbf{x}_i) \sim \mathcal{N}(0,\mathbf{\Sigma})$ from Assumption 9, we then have*

$$\underset{\phi(\mathbf{x}_i)\sim\mathcal{N}(\mathbf{0},\mathbf{\Sigma})}{\mathbb{E}}\left\|\frac{1}{n}\sum_{i=1}^{n}\phi(\mathbf{x}_i)\otimes\phi(\mathbf{x}_i) - \mathbf{\Sigma}\right\|_{\mathrm{HS}} \leq C\left(\frac{\operatorname{Tr}(\mathbf{\Sigma})}{\sqrt{n}}\right) \tag{64}$$

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

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

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

$$\overset{(i)}{\leq} \underset{\phi(\mathbf{x}_i)\sim\mathcal{N}(\mathbf{0},\mathbf{\Sigma})}{\mathbb{E}}\ \underset{\tilde{\phi}(\mathbf{x}_i)\sim\mathcal{N}(\mathbf{0},\mathbf{\Sigma})}{\mathbb{E}}\left\|\frac{1}{n}\sum_{i=1}^{n}\left(\phi(\mathbf{x}_i)\otimes\phi(\mathbf{x}_i) - \tilde{\phi}(\mathbf{x}_i)\otimes\tilde{\phi}(\mathbf{x}_i)\right)\right\|_{\mathrm{HS}} \tag{65}$$

$$= \mathop{\mathbb{E}}_{\phi(\mathbf{x}_i)\sim\mathcal{N}(\mathbf{0},\mathbf{\Sigma})} \mathop{\mathbb{E}}_{\tilde{\phi}(\mathbf{x}_i)\sim\mathcal{N}(\mathbf{0},\mathbf{\Sigma})} \mathop{\mathbb{E}}_{\xi_i\sim\mathcal{R}} \left\| \frac{1}{n}\sum_{i=1}^{n} \xi_i \left(\phi(\mathbf{x}_i)\otimes\phi(\mathbf{x}_i) - \tilde{\phi}(\mathbf{x}_i)\otimes\tilde{\phi}(\mathbf{x}_i)\right) \right\|_{\mathrm{HS}} \tag{66}$$

$$\overset{(ii)}{\leq} \frac{2}{n} \mathop{\mathbb{E}}_{\phi(\mathbf{x}_i)\sim\mathcal{N}(\mathbf{0},\mathbf{\Sigma})} \mathop{\mathbb{E}}_{\xi_i\sim\mathcal{R}} \left\| \sum_{i=1}^{n} \xi_i \phi(\mathbf{x}_i)\otimes\phi(\mathbf{x}_i) \right\|_{\mathrm{HS}} \tag{67}$$

$$\overset{(iii)}{\leq} \frac{2}{n} \left(\mathop{\mathbb{E}}_{\phi(\mathbf{x}_i)\sim\mathcal{N}(\mathbf{0},\mathbf{\Sigma})} \mathop{\mathbb{E}}_{\xi_i\sim\mathcal{R}} \left\| \sum_{i=1}^{n} \xi_i \phi(\mathbf{x}_i)\otimes\phi(\mathbf{x}_i) \right\|_{\mathrm{HS}}^2 \right)^{1/2} \tag{68}$$

(i) follows from noticing $\phi(\mathbf{x}_i)\otimes\phi(\mathbf{x}_i) - \mathbf{\Sigma}$ is a mean $\mathbf{0}$ operator in $\mathcal{H}\otimes\mathcal{H}$, then for $X, Y \in \mathcal{H}\otimes\mathcal{H}$ s.t. $\mathbb{E}[Y] = \mathbf{0}$ it follows $\|X\|_{\mathrm{HS}} = \|X - \mathbb{E}[Y]\|_{\mathrm{HS}} = \|\mathbb{E}_Y[X - Y]\|_{\mathrm{HS}}$ and finally applying Jensen's Inequality. (ii) follows from the triangle inequality. (iii) follows from Jensen's Inequality. Let e_k for $k \in [p]$ represent an orthonormal basis for the Hilbert Space $\mathcal{H}$. By expanding out the Hilbert-Schmidt Norm, we then have

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

$$= \frac{2}{n} \left(\mathop{\mathbb{E}}_{\phi(\mathbf{x}_i)\sim\mathcal{N}(\mathbf{0},\mathbf{\Sigma})} \mathop{\mathbb{E}}_{\xi_i\sim\mathcal{R}} \sum_{k=1}^{p} \left\langle \sum_{i=1}^{n} \xi_i \phi(\mathbf{x}_i)\otimes\phi(\mathbf{x}_i) e_k, \sum_{j=1}^{n} \xi_j \phi(\mathbf{x}_j)\otimes\phi(\mathbf{x}_j) e_k \right\rangle_{\mathcal{H}} \right)^{1/2} \tag{69}$$

$$= \frac{2}{n} \left(\mathop{\mathbb{E}}_{\phi(\mathbf{x}_i)\sim\mathcal{N}(\mathbf{0},\mathbf{\Sigma})} \mathop{\mathbb{E}}_{\xi_i\sim\mathcal{R}} \sum_{k=1}^{p}\sum_{i=1}^{n}\sum_{j=1}^{n} \xi_i\xi_j \left\langle \phi(\mathbf{x}_i)\otimes\phi(\mathbf{x}_i) e_k, \phi(\mathbf{x}_j)\otimes\phi(\mathbf{x}_j) e_k \right\rangle_{\mathcal{H}} \right)^{1/2} \tag{70}$$

$$\overset{(iv)}{\leq} \frac{2}{n} \left(\mathop{\mathbb{E}}_{\phi(\mathbf{x}_i)\sim\mathcal{N}(\mathbf{0},\mathbf{\Sigma})} \sum_{k=1}^{p}\sum_{i=1}^{n} \left\langle \phi(\mathbf{x}_i)\otimes\phi(\mathbf{x}_i) e_k, \phi(\mathbf{x}_i)\otimes\phi(\mathbf{x}_i) e_k \right\rangle_{\mathcal{H}} \right)^{1/2} \tag{71}$$

$$= \frac{2}{n} \left(\sum_{i=1}^{n} \mathop{\mathbb{E}}_{\phi(\mathbf{x}_i)\sim\mathcal{N}(\mathbf{0},\mathbf{\Sigma})} \|\phi(\mathbf{x}_i)\otimes\phi(\mathbf{x}_i)\|_{\mathrm{HS}}^2 \right)^{1/2} \overset{(v)}{=} \frac{2}{n} \left(\sum_{i=1}^{n} \mathop{\mathbb{E}}_{\phi(\mathbf{x}_i)\sim\mathcal{N}(\mathbf{0},\mathbf{\Sigma})} \|\phi(\mathbf{x}_i)\|_{\mathcal{H}}^4 \right)^{1/2} \tag{72}$$

$$= \frac{2}{n} \left(\sum_{i=1}^{n} \mathop{\mathbb{E}}_{\phi(\mathbf{x}_i)\sim\mathcal{N}(\mathbf{0},\mathbf{\Sigma})} \left[k^2(x_i, x_i)\right] \right)^{1/2} = \frac{2}{\sqrt{n}} \left(2\,\mathrm{Tr}\left(\mathbf{\Sigma}^2\right) + \mathrm{Tr}\left(\mathbf{\Sigma}\right)^2\right)^{1/2} \leq 2\sqrt{3} n^{-1/2}\,\mathrm{Tr}\left(\mathbf{\Sigma}\right) \tag{73}$$

(iv) follows from noticing $\mathbb{E}_{\xi_i,\xi_j\sim\mathcal{R}}[\xi_i\xi_j] = \delta_{ij}$. (v) follows from expanding the Hilbert-Schmidt Norm and applying Parseval's Identity. We note $\mathrm{Tr}(\mathbf{\Sigma}) < \infty$ and therefore even though the covariance operator is infinite-dimensional we are able to get a finite bound on the covariance approximation. This completes the proof. ∎

Proposition 20 (Probabilistic Bound on Infinite Dimensional Covariance Estimation in the Hilbert-Schmidt Norm)**.** *Let $\mathbf{x}_1, \ldots, \mathbf{x}_n$ be i.i.d sampled from $\mathbb{P}$ such that $\phi(\mathbf{x}_i) \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$ from Assumption 9, we then have for any $u \geq 1$*

$$\mathop{\mathbb{P}}_{\phi(\mathbf{x}_i)\sim\mathcal{N}(\mathbf{0},\mathbf{\Sigma})} \left\{ \left\| \frac{1}{n}\sum_{i=1}^{n} \phi(\mathbf{x}_i)\otimes\phi(\mathbf{x}_i) - \mathbf{\Sigma} \right\|_{\mathrm{HS}} \geq \frac{\mathrm{Tr}(\mathbf{\Sigma})}{\sqrt{n}} \cdot u \right\} \leq e^{-u^2\,\mathrm{Tr}(\mathbf{\Sigma})/2} \tag{74}$$

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

$$\mathop{\mathbb{P}}_{\phi(\mathbf{x}_i)\sim\mathcal{N}(\mathbf{0},\mathbf{\Sigma})} \left\{ \left\| \frac{1}{n}\sum_{i=1}^{n} \phi(\mathbf{x}_i)\otimes\phi(\mathbf{x}_i) - \mathbf{\Sigma} \right\|_{\mathrm{HS}} \geq C \cdot u \right\} \tag{75}$$

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

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

$$\overset{(i)}{\leq} \underset{\phi(\mathbf{x}_i)\sim\mathcal{N}(\mathbf{0},\mathbf{\Sigma})}{\mathbb{P}} \left\{ \frac{\sqrt{\operatorname{Tr}(\mathbf{K})}}{n} + \frac{\operatorname{Tr}(\mathbf{\Sigma})}{\sqrt{n}} \geq C \cdot u \right\} \tag{76}$$

$$\leq \underset{\phi(\mathbf{x}_i)\sim\mathcal{N}(\mathbf{0},\mathbf{\Sigma})}{\mathbb{P}} \left\{ \sum_{i=1}^{n} \|\phi(\mathbf{x}_i)\|_{\mathcal{H}} \geq nC \cdot u - \sqrt{n}\operatorname{Tr}(\mathbf{\Sigma}) \right\} \tag{77}$$

$$\leq \exp\left[-\frac{(nC\cdot u + \sqrt{n}\operatorname{Tr}(\mathbf{\Sigma}))^2}{n\operatorname{Tr}(\mathbf{\Sigma})} \right] \overset{(ii)}{\leq} e^{-u^2\operatorname{Tr}(\mathbf{\Sigma})/2} \tag{78}$$

In (i) we apply Jensen's Inequality to both expectation terms and denote $\mathbf{K} \triangleq \mathbf{\Phi}^\top\mathbf{\Phi}$. In (ii) we chose $C \triangleq \operatorname{Tr}(\mathbf{\Sigma})/\sqrt{n}$ and then simplify the resultant probability bound. ■

Proposition 21 (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 $\phi(\mathbf{x}_i) \sim \mathcal{N}(\mathbf{0},\mathbf{\Sigma})$ from Assumption 9, we then have for any $u \geq 1$*

$$\underset{\phi(\mathbf{x}_i)\sim\mathcal{N}(\mathbf{0},\mathbf{\Sigma})}{\mathbb{P}} \left\{ \left\| \frac{1}{n}\sum_{i=1}^{n} \phi(\mathbf{x}_i)\otimes\phi(\mathbf{x}_i) - \mathbf{\Sigma} \right\|_{\mathrm{op}} \geq \frac{\|\mathbf{\Sigma}\|}{\sqrt{n}} \cdot u \right\} \leq e^{-u^2\|\mathbf{\Sigma}\|/2} \tag{79}$$

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

$$\begin{aligned} \underset{\phi(\mathbf{x}_i)\sim\mathcal{N}(\mathbf{0},\mathbf{\Sigma})}{\mathbb{P}} \left\{ \left\| \frac{1}{n}\sum_{i=1}^{n} \phi(\mathbf{x}_i)\otimes\phi(\mathbf{x}_i) - \mathbf{\Sigma} \right\|_{\mathrm{op}} \geq C\cdot u \right\} &\leq \underset{\phi(\mathbf{x}_i)\sim\mathcal{N}(\mathbf{0},\mathbf{\Sigma})}{\mathbb{P}} \Bigg\{ \underset{\xi\sim\mathcal{R}}{\mathbb{E}} \left\| \sum_{i=1}^{n} \xi_i\phi(\mathbf{x}_i)\otimes\phi(\mathbf{x}_i) \right\|_{\mathrm{op}} \\ &+ \underset{\tilde{\phi}(\mathbf{x}_i)\sim\mathcal{N}(\mathbf{0},\mathbf{\Sigma})}{\mathbb{E}} \underset{\xi_i\sim\mathcal{R}}{\mathbb{E}} \left\| \frac{1}{n}\sum_{i=1}^{n} \xi_i\tilde{\phi}(\mathbf{x}_i)\otimes\tilde{\phi}(\mathbf{x}_i) \right\|_{\mathrm{HS}} \geq C\cdot u \Bigg\} \end{aligned} \tag{80}$$

Lemma 22 (Finite Dimensional Covariate Estimation in the Spectral Norm)**.** *Let $\mathbf{x}_1, \ldots, \mathbf{x}_n \sim \mathcal{N}(\mathbf{0},\mathbf{\Sigma})$. It then follows,*

$$\underset{\mathbf{x}_i\sim\mathcal{N}(\mathbf{0},\mathbf{\Sigma})}{\mathbb{E}} \left\| \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i\mathbf{x}_i^\top - \mathbf{\Sigma} \right\|_2 \leq C\|\mathbf{\Sigma}\| \left(\sqrt{\frac{d}{n}} + \frac{1}{\sqrt{dn}} \right) \tag{81}$$

where $C \leq 62.82$.

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

$$\underset{\mathbf{x}_i\sim\mathcal{N}(\mathbf{0},\mathbf{\Sigma})}{\mathbb{E}} \left\| \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i\mathbf{x}_i^\top - \mathbf{\Sigma} \right\|_2 \leq \|\mathbf{\Sigma}\| \underset{\tilde{\mathbf{x}}_i\sim\mathcal{N}(\mathbf{0},\mathbf{I})}{\mathbb{E}} \left\| \frac{1}{n}\tilde{\mathbf{X}}^\top\tilde{\mathbf{X}} - \mathbf{I} \right\|_2 \tag{82}$$

$$= \|\mathbf{\Sigma}\| \int_0^\infty \underset{\tilde{\mathbf{x}}_i\sim\mathcal{N}(\mathbf{0},\mathbf{I})}{\mathbb{P}} \left\{ \left\| \frac{1}{n}\tilde{\mathbf{X}}^\top\tilde{\mathbf{X}} - \mathbf{I} \right\|_2 \geq t \right\} dt \tag{83}$$

Let $\mathcal{M}$ be an ε-net of $\mathbb{S}^{d-1}$ for $\varepsilon = \frac{1}{4}$, then $|\mathcal{M}| \leq 9^d$. It then follows from [Ver20] Corollary 4.2.13,

$$\underset{\tilde{\mathbf{x}}_i\sim\mathcal{N}(\mathbf{0},\mathbf{I})}{\mathbb{P}} \left\{ \left\| \frac{1}{n}\tilde{\mathbf{X}}^\top\tilde{\mathbf{X}} - \mathbf{I} \right\|_2 \geq t \right\} \leq \underset{\tilde{\mathbf{x}}_i\sim\mathcal{N}(\mathbf{0},\mathbf{I})}{\mathbb{P}} \left\{ \max_{\mathbf{y}\in\mathcal{M}} \left| \frac{1}{n}\|\tilde{\mathbf{X}}\mathbf{y}\|_2^2 - 1 \right| \geq \frac{t}{2} \right\} \tag{84}$$

Denote $K \triangleq 16\sqrt{\frac{8}{3}}$, then we have from a union bound and Bernstein's Inequality [Ber24].

$$\underset{\tilde{\mathbf{x}}_i\sim\mathcal{N}(\mathbf{0},\mathbf{I})}{\mathbb{P}} \left\{ \max_{\mathbf{y}\in\mathcal{M}} \left| \frac{1}{n}\|\tilde{\mathbf{X}}\mathbf{y}\|_2^2 - 1 \right| \geq \frac{t}{2} \right\} \leq 9^d \exp\left[-\frac{n}{2}\left(\frac{t^2}{K^2} \wedge \frac{t}{K} \right) \right] \tag{85}$$

Let $\delta \in (0,1)$, then we find RHS Equation (85) is less than δ when

$$t \geq K\left(\frac{2d\log(9) + 2\log(2/\delta)}{n} \vee \left(\frac{2d\log(9) + 2\log(2/\delta)}{n} \right)^{1/2} \right) \tag{86}$$

Furthermore, we note we have equality with one, when $t = K\sqrt{(2d\log(9)+\log(4))/n} \triangleq C$ all $t \le C$ occur with probability also equal to one. Therefore, plugging this back into the RHS of Equation (83).

$$\begin{aligned}
&\int_0^\infty \underset{\tilde{\mathbf{x}}_i \sim \mathcal{N}(\mathbf{0},\mathbf{I})}{\mathbb{P}} \left\{ \left\| \frac{1}{n}\tilde{\mathbf{X}}^\top \tilde{\mathbf{X}} - \mathbf{I} \right\|_2 \ge t \right\} dt \\
&\quad \le K\sqrt{\frac{2d\log(9)}{n} + \frac{2\log(2)}{n}} + \int_C^\infty 9^d \exp\left[-\frac{n}{2}\left(\frac{t^2}{K^2}\right)\right] dt
\end{aligned} \tag{87}$$

$$\le K\sqrt{\frac{2d\log(9)}{n} + \frac{2\log(2)}{n}} + \frac{K9^d \exp\left[-\left(K\sqrt{\frac{2d\log(9)}{n} + \frac{2\log(2)}{n}}\right)^2 \left(\frac{n}{2K^2}\right)\right]}{2n\sqrt{\frac{2d\log(9)}{n} + \frac{2\log(2)}{n}}} \tag{88}$$

$$\le K\sqrt{\frac{2d\log(9)+\log(4)}{n}} + \frac{K}{4\sqrt{n\left(2d\log(9)+\log(4)\right)}} \tag{89}$$

$$\le \sqrt{\log(324)}K\left(\sqrt{\frac{d}{n}} + \frac{1}{\sqrt{nd}}\right) \tag{90}$$

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

B Proofs for Structural Results

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

B.1 Proof of Lemma 3

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

$$\partial(-g(t,f_{\mathbf{w}})) = \partial\left(-t + \frac{1}{n(1-\epsilon)}\sum_{i=1}^{n}(t-\hat{\nu}_i)^+\right) = -1 + \frac{1}{n(1-\epsilon)}\sum_{i=1}^{n(1-\epsilon)}\begin{cases}1 & \text{if } t > \hat{\nu}_i\\ 0 & \text{if } t < \hat{\nu}_i\\ [0,1] & \text{if } t = \hat{\nu}_i\end{cases} \tag{91}$$

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 p-quantile of the Risk. ■

B.2 Proof of Lemma 4

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

$$\begin{aligned}
&\nabla_f g(t^{(k+1)}, f_{\mathbf{w}}^{(k)}) = \nabla_f\left(t^{(k+1)} - \frac{1}{n(1-\epsilon)}\sum_{i=1}^{n}\left(t^{(k+1)} - \ell(\mathbf{x}_i; f_{\mathbf{w}}^{(k)}, y_i)\right)^+\right)\\
&= -\frac{1}{n(1-\epsilon)}\sum_{i=1}^{n(1-\epsilon)}\nabla_f\left(t^{(k+1)} - \ell(\mathbf{x}_i; f_{\mathbf{w}}^{(k)}, y_i)\right)^+ = \frac{1}{n(1-\epsilon)}\sum_{i=1}^{n}\nabla_f \ell(\mathbf{x}_i; f_{\mathbf{w}}^{(k)}, y_i)\begin{cases}1 & \text{if } t > \hat{\nu}_i\\ 0 & \text{if } t < \hat{\nu}_i\\ [0,1] & \text{if } t = \hat{\nu}_i\end{cases}
\end{aligned} \tag{92}$$

Now we note $\nu_{n(1-\epsilon)} \le t^{(k+1)} \le \nu_{n(1-\epsilon)+1}$. Then, plugging this into Equation (92), we have

$$\nabla_f g(t^{(k+1)}, f_{\mathbf{w}}^{(k)}) = \frac{1}{n(1-\epsilon)}\sum_{i=1}^{n(1-\epsilon)}\nabla_f \ell(\mathbf{x}_i; f_{\mathbf{w}}^{(k)}, y_i) \tag{93}$$

This concludes the proof. ■

C Binary Classification

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

C.1 Subquantile Lipschitzness

Lemma 23. *(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:*

$$|g(t,f) - g(t,\hat{f})| \leq L\|f - \hat{f}\|_{\mathcal{H}} \tag{94}$$

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

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

$$\|\nabla_f g(t,f)\|_{\mathcal{H}} = \left\| \frac{1}{n(1-\epsilon)} \sum_{i=1}^{n} \mathbb{I}\left\{ t \geq (1-y_i)\log(f^{(t)}(\mathbf{x}_i)) \right\} \left(y_i - \sigma(f^{(t)}(\mathbf{x}_i)) \right) \cdot \mathbf{x}_i^\top \right\|_{\mathcal{H}} \tag{95}$$

$$\overset{(i)}{\leq} \frac{1}{n(1-\epsilon)} \left\| \sum_{i \in S^{(t)}} \left(y_i - \sigma(f^{(t)}(\mathbf{x}_i)) \right) \cdot \mathbf{x}_i^\top \right\|_{\mathcal{H}} \tag{96}$$

$$\overset{(ii)}{\leq} \frac{1}{n(1-\epsilon)} \max_{i \in [n]} \left| y_i - \sigma(f^{(t)}(\mathbf{x}_i)) \right| \left\| \sum_{i \in S^{(t)}} \mathbf{x}_i \right\|_{\mathcal{H}} \tag{97}$$

$$\overset{(iii)}{\leq} \frac{1}{n(1-\epsilon)} \sum_{i \in X} \|\mathbf{x}_i\|_2 \tag{98}$$

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

C.2 Proof of Theorem 7

From Algorithm 1, we have for kernelized binary classification with linear kernel,

$$f^{(t+1)} = f^{(t)} - \frac{\eta}{n(1-\epsilon)} \sum_{i \in S^{(t)}} \left(\sigma(f^{(t)}(\mathbf{x}_i)) - y_i \right) \cdot \mathbf{x}_i^\top \tag{99}$$

From which it follows,

$$\left\| f^{(t+1)} - f^* \right\|_{\mathcal{H}}^2 = \left\| f^{(t)} - f^* \right\|_{\mathcal{H}}^2 - 2\eta \left\langle \nabla_f g(f^{(t)}, t^*), f^{(t)} - f^* \right\rangle_{\mathcal{H}} + \eta^2 \left\| \nabla_f g(f^{(t)}, t^*) \right\|_{\mathcal{H}}^2 \tag{100}$$

We will expand the second term.

$$\left\langle \nabla_f g(f^{(t)}, t^*), f^{(t)} - f^* \right\rangle_{\mathcal{H}} \overset{(99)}{=} \left\langle f^{(t)} - f^*, \sum_{i \in S^{(t)}} \left(\sigma(f^{(t)}(\mathbf{x}_i)) - y_i \right) \cdot \mathbf{x}_i^\top \right\rangle_{\mathcal{H}} \tag{101}$$

$$= \left\langle f^{(t)} - f^*, \sum_{i \in S^{(t)}} \left(\sigma(f^{(t)}(\mathbf{x}_i)) - \sigma(f^*(\mathbf{x}_i)) \right) \cdot \mathbf{x}_i^\top \right\rangle_{\mathcal{H}} + \left\langle f^{(t)} - f^*, \sum_{i \in S^{(t)}} \left(\sigma(f^*(\mathbf{x}_i)) - y_i \right) \cdot \mathbf{x}_i^\top \right\rangle_{\mathcal{H}} \tag{102}$$

We first upper bound upper bound the second term in Equation (102). From the Cauchy-Schwarz Inequality and noting $y_i \in \{0,1\}$ and range$(\sigma) \in (0,1)$, we have the following,

$$\left\langle f^* - f^{(t)}, \sum_{i \in S^{(t)}} \left(\sigma(f^*(\mathbf{x}_i)) - y_i \right) \mathbf{x}_i^\top \right\rangle_{\mathcal{H}} \leq \left\| f^* - f^{(t)} \right\|_{\mathcal{H}} \left\| \sum_{i \in S^{(t)}} \mathbf{x}_i \right\|_2 \max_{i \in S^{(t)}} \left| \sigma(f^*(\mathbf{x}_i)) - y_i \right| \tag{103}$$

$$\overset{(i)}{\leq} \frac{1}{2}\left\|f^* - f^{(t)}\right\|_{\mathcal{H}}^2 \left\|\sum_{i \in S^{(t)}} \mathbf{x}_i\right\|_2^2 + \frac{1}{2}\sum_{i \in P}(\sigma(f^*(\mathbf{x}_i)) - y_i)^2 \tag{104}$$

where (i) follows from Young's Inequality [You12]. Let us now consider the function $h : \mathcal{H} \to \mathbb{R}$ defined as $h(f) \triangleq \sum_{i \in S \cap P} \log(1 + \exp(f^{(t)}(\mathbf{x}_i))$. We can then calculate the gradients by hand, $\nabla h(f) = \sum_{i \in S \cap P} \sigma(f^{(t)}(\mathbf{x}_i)) \cdot \mathbf{x}_i^\top$ and $\nabla^2 h(f) = \sum_{i \in S \cap P} \sigma(f^{(t)}(\mathbf{x}_i))(1 - \sigma(f^{(t)}(\mathbf{x}_i))) \cdot \mathbf{x}_i \mathbf{x}_i^\top$. From the Taylor Series expansion, we have for any $f, \hat{f} \in \mathbb{R}^d$, there exists $\tilde{f}$ such that

$$\left\langle f - \hat{f}, \nabla h(f) - \nabla h(\hat{f}) \right\rangle_{\mathcal{H}} = \left\langle \nabla^2 h(\tilde{f}), (f - \hat{f}) \otimes (f - \hat{f}) \right\rangle_{\mathrm{HS}} \tag{105}$$

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

$$\begin{aligned}
&\left\langle f^{(t)} - f^*, \sum_{i \in S^{(t)} \cap P} \left(\sigma(f^{(t)}(\mathbf{x}_i)) - \sigma(f^*(\mathbf{x}_i))\right) \cdot \mathbf{x}_i^\top \right\rangle_{\mathcal{H}} \\
&\qquad \gtrsim \lambda_{\min}\left(\sum_{i \in S^{(t)}} \mathbf{x}_i \mathbf{x}_i^\top\right) \left\|f^{(t)} - f^*\right\|_{\mathcal{H}}^2 \overset{(ii)}{\geq} \lambda_{\min}\left(\sum_{i \in S^{(t)} \cap P} \mathbf{x}_i \mathbf{x}_i^\top\right) \left\|f^{(t)} - f^*\right\|_{\mathcal{H}}^2
\end{aligned} \tag{106}$$

where (ii) follows from Weyl's inequality [Wey12]. It is possible to show $C = \Omega(\exp(-R\,\mathrm{Tr}(\mathbf{\Sigma}) \log n))$ where $R = \max_{t \in [T]} \left\|f^{(t)}\right\|_2$ from Proposition 16. We will now bound the final term in Equation (100).

$$\begin{aligned}
\left\|\nabla_f g(f^{(t)}, t^*)\right\|_{\mathcal{H}}^2 &= \left\|\sum_{i \in S^{(t)}} (\sigma(f^{(t)}(\mathbf{x}_i)) - y_i) \cdot \mathbf{x}_i^\top\right\|_{\mathcal{H}}^2 \leq \max_{i \in S^{(t)}} \left|\sigma(f^{(t)}(\mathbf{x}_i)) - y_i\right|^2 \cdot \left\|\sum_{i \in S^{(t)}} \mathbf{x}_i\right\|_2^2 \\
&\overset{(iii)}{\leq} \left\|\sum_{i \in S^{(t)}} \mathbf{x}_i\right\|_2^2 \cdot \sum_{i \in S^{(t)}} \left(\sigma(f^{(t)}(\mathbf{x}_i)) - y_i\right)^2 \leq \left\|\sum_{i \in S^{(t)}} \mathbf{x}_i\right\|_2^2 \cdot \sum_{i \in P} \left(\sigma(f^{(t)}(\mathbf{x}_i)) - y_i\right)^2
\end{aligned} \tag{107}$$

where (iii) follows from noting for any $\mathbf{x} \in \mathbb{R}^d$ it holds $\|\mathbf{x}\|_\infty \leq \|\mathbf{x}\|_{\mathcal{H}}$. Now, combining Equations (100), (104), (106) and (107), we obtain

$$\begin{aligned}
\left\|f^{(t+1)} - f^*\right\|_{\mathcal{H}}^2 &\leq \left\|f^{(t)} - f^*\right\|_{\mathcal{H}}^2 \left(1 - 2C\eta\lambda_{\min}\left(\sum_{i \in S^{(t)} \cap P} \mathbf{x}_i \mathbf{x}_i^\top\right) + \frac{\eta^2}{2}\left\|\sum_{i \in S^{(t)}} \mathbf{x}_i\right\|_2^2\right) \\
&\quad + \eta^2 \left\|\sum_{i \in S^{(t)}} \mathbf{x}_i\right\|_2^2 \cdot \sum_{i \in P}(\sigma(f^{(t)}(\mathbf{x}_i)) - y_i)^2
\end{aligned} \tag{108}$$

We will now expand out the final term in Equation (108).

$$\sum_{i \in P}(\sigma(f^{(t+1)}(\mathbf{x}_i)) - y_i)^2 = \sum_{i \in P}(\sigma(f^{(t+1)}(\mathbf{x}_i)) - \sigma(f^*(\mathbf{x}_i)) + \sigma(f^*(\mathbf{x}_i)) - y_i)^2 \tag{109}$$

$$\leq 2\left\|(f^{(t+1)} - f^*)\mathbf{X}_P^\top\right\|_{\mathcal{H}}^2 + 2\sum_{i \in P}(\sigma(f^*(\mathbf{x}_i)) - y_i)^2 \tag{110}$$

$$\leq 2\left\|f^{(t+1)} - f^*\right\|_{\mathcal{H}}^2 \left\|\mathbf{X}_P^\top \mathbf{X}_P\right\|_2 + 2\mathcal{E}_{\mathrm{OPT}} \tag{111}$$

Now, we can use Equation (108) to complete the bound.

$$\begin{aligned}
&\sum_{i \in P}(\sigma(f^{(t+1)}(\mathbf{x}_i)) - y_i)^2 \\
&\quad \leq 2\left\|f^{(t)} - f^*\right\|_{\mathcal{H}}^2 \left\|\sum_{i \in P} \mathbf{x}_i \mathbf{x}_i^\top\right\|_2 \left(1 - 2C\eta\lambda_{\min}\left(\mathbf{X}_{\mathrm{TP}}^\top \mathbf{X}_{\mathrm{TP}}\right) + \frac{\eta^2}{2}\left\|\sum_{i \in S^{(t)}} \mathbf{x}_i\right\|_{\mathcal{H}}^2\right) + 2.5\mathcal{E}_{\mathrm{OPT}}
\end{aligned} \tag{112}$$

Now let us the define for all $t \in [T]$,

$$\Lambda^{(t)} \triangleq \left\| \sum_{i \in P} \mathbf{x}_i \mathbf{x}_i^\top \right\| \left\| f^* - f^{(t)} \right\|_{\mathcal{H}}^2 + \sum_{i \in P} (\sigma(f^{(t)}(\mathbf{x}_i)) - y_i)^2 \tag{113}$$

We then have from Equations (107) and (112),

$$\Lambda^{(t+1)} \leq \max\left\{ 3\left(1 - 2C\eta\lambda_{\min}\left(\mathbf{X}_{\mathrm{TP}}^\top \mathbf{X}_{\mathrm{TP}}\right) + \frac{\eta^2}{2} \left\| \sum_{i \in S^{(t)}} \mathbf{x}_i \right\|_2^2 \right), \eta^2 \left\| \sum_{i \in S^{(t)}} \mathbf{x}_i \right\|_{\mathcal{H}}^2 \right\} \cdot \left(\Lambda^{(t)} - 2.5\mathcal{E}_{\mathrm{OPT}} \right) + 2.5\mathcal{E}_{\mathrm{OPT}} \tag{114}$$

Solving the quadratic equation, we observe for

$$\eta \leq \frac{2C\lambda_{\min}\left(\mathbf{X}_{\mathrm{TP}}^\top \mathbf{X}_{\mathrm{TP}}\right)}{\left\| \sum_{i \in S^{(t)}} \mathbf{x}_i \right\|_2^2} \tag{115}$$

We then observe there is a linear decrease in $\Lambda^{(t)} - 2.5\mathcal{E}_{\mathrm{OPT}}$ plus a constant,

$$\Lambda^{(t+1)} \leq \max\left\{ \frac{1}{2}, \frac{4C^2\lambda_{\min}^2\left(\mathbf{X}_{\mathrm{TP}}^\top \mathbf{X}_{\mathrm{TP}}\right)}{\left\| \sum_{i \in S^{(t)}} \mathbf{x}_i \right\|_2^2} \right\} \cdot \left(\Lambda^{(t)} - 2.5\mathcal{E}_{\mathrm{OPT}} \right) + 2.5\mathcal{E}_{\mathrm{OPT}} \tag{116}$$

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