Kernel Learning in the Huber ϵ-Contamination Model

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

Alex Gittens
RPI CS, gittea@rpi.edu

Abstract

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

1 Introduction

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

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

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

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

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

Explain the notation for pushforward measure. Explain Gaussian proxy.

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

We will now give one of the first results proving the effectiveness of Iterative Thresholding in Learning Problems.

Theorem 3 (Theorem 5 in [BJK15])**.** *Let X be a sub-Gaussian data matrix, and $\mathbf{y} = \mathrm{X}^{\mathrm{T}}\mathbf{w}^* + \mathbf{e}$ where $\mathbf{e}$ is the corruption. Then there exists an algorithm such that $\left\|\mathbf{w}^{(t)} - \mathbf{w}^*\right\|_2 \leq \varepsilon$ after $t = O\left(\left(\log\left(\frac{\|\mathbf{b}\|_2}{\sqrt{n}}\right)\right)\frac{1}{\varepsilon}\right)$ iterations.*

We will now give our results for the Kernelized GLM problem.

Theorem 4 (Informal of Theorem 13)**.** *Let the dataset be given as $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$ such that for $i \in \mathrm{P}$, $y_i = \omega(f^*(\mathbf{x}_i)) + \xi_i$. Then there exists an algorithm such $\left\|f^{(t)} - f^*\right\|_{\mathcal{H}} \leq \varepsilon + O\left(\|\Gamma\|\,\sigma\right) + O\left(\sigma\right)$ after $t = O\left(n\log\left(\frac{\|f^*\|_{\mathcal{H}}}{\varepsilon}\right)\right)$ iterations.*

1.1 Contributions

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

2 Preliminaries

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

2.1 Reproducing Kernel Hilbert Spaces

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

Definition 5 (Reproducing Property)**.** *Let $\mathbf{x} \in \mathcal{X}$, then for any $f \in \mathcal{H}$,*

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

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

2.2 Tensor Products

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

2.3 Sub-Gaussian Random Functions in the Hilbert Space

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

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

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

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

Then we say $X \sim \mathcal{SG}(\Gamma, \alpha)$, where if $\alpha = 1$, we will say $X \sim \mathcal{SG}(\Gamma)$. The Gaussian Design for the Feature Space has gained popularity in the study of kernel learning [CLKZ21]. The sub-Gaussian design is the standard assumed distribution in the robust statistics literature, [JLT20, ADKS22], and has been studied extensively in the context of iterative thresholding algorithms for linear regression.

2.4 Assumptions

We will first give our assumptions for robust kernelized regression.

Assumption 6 (Sub-Gaussian Design)**.** *We assume for $\mathbf{x}_i \sim \mathcal{X}$, then it follows for the function to the Hilbert Space, $\phi(\cdot) : \mathcal{X} \to \mathcal{H}$,*

$$\phi(\mathbf{x}) \triangleq X \sim \mathcal{P}_\sharp \phi \triangleq \mathcal{SG}\left(\Gamma, {}^1\!/_2\right)$$

where Γ is a possibly infinite dimensional covariance operator.

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

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

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

Assumption 8 (Normal Residuals)**.** *Let* $\inf_{f\in\mathcal{H}} \mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[\mathcal{R}(f;\mathbf{x},y)]$*. The residual is defined as* $\mu_i \triangleq f^*(\mathbf{x}_i) - y_i$*. Then we assume for some* $\sigma > 0$*, it follows*

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

2.5 Related Work

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

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

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

3 Subquantile Minimization

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

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

The p-subquantile of the risk is then given by

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

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

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

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

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

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

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

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

3.1 Reduction to Iterative Thresholding

To consider theoretical guarantees of Subquantile Minimization, we first analyze the inner and outer optimization problems. We first analyze kernel learning in the presence of corrupted data. Next, we provide error bounds for the two most important kernel learning problems, kernel ridge regression, and kernel classification. Now we will give our first result regarding kernel learning in the Huber ϵ-contamination model. Now we will analyze the two-step minimax optimization steps described in Section 3.

Lemma 9. *Let $\mathcal{R} : \mathcal{H} \times \mathbb{R} \to \mathbb{R}$ be a loss function (not necesarily convex). Let $\mathbf{x}_{\lfloor i \rfloor}$ represent the point with the i-th smallest loss w.r.t $\mathcal{R}$. If we denote $\hat{\nu}_i \triangleq \mathcal{R}(f; \mathbf{x}_{\lfloor i \rfloor}, y_{\lfloor i \rfloor})$, it then follows $\hat{\nu}_{n(1-\epsilon)} \in \arg\max_{t \in \mathbb{R}} g(t, f)$.*

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

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

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

From Lemma 9, 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 10. *Let $\hat{\nu}_i \triangleq \mathcal{R}(f; \mathbf{x}_{\lfloor i \rfloor}, y_{\lfloor i \rfloor})$, if we choose $t^{(k+1)} = \hat{\nu}_{n(1-\epsilon)}$ as by Lemma 9, it then follows $\nabla_f g(t^{(k)}, f^{(k)}) = \frac{1}{n(1-\epsilon)} \sum_{i=1}^{n(1-\epsilon)} \nabla_f \mathcal{R}(f^{(k)}; \mathbf{x}_{\lfloor i \rfloor}, y_{\lfloor i \rfloor})$.*

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

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

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

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

This concludes the proof. ■

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

4 Convergence

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

4.1 Kernelized Ridge Regression

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

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

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

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

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

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

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

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

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

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

   6. Perform Gradient Descent Iteration.

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

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

Theorem 11 (Subquantile Minimization for Overparameterized Regression)**.** *Algorithm 1 run on a dataset $\mathcal{D} \sim \hat{\mathcal{P}}$ such that $\mathrm{X} \in \mathbb{R}^{p\times n}$ where $p \geq n$ and return $\hat{\mathbf{w}}$. Then, with probability exceeding $1-\delta$, when $n \geq O(f(\Gamma))$, then*

$$\left\|\mathbf{w}^{(T)} - \mathbf{w}^*\right\|_2 \leq \varepsilon + O\left(\Xi\right)$$

after $T = O\left(\log\left(\frac{\|\mathbf{w}^*\|}{\varepsilon}\right)\right)$ iterations.

Theorem 12 (Subquantile Minimization for Kernelized Regression with Full Solves)**.** *Suppose $\phi(\mathbf{x}_i) \triangleq X_i \in \mathcal{H}$ such that $\mathcal{H}$ is of finite rank. Then there exists an algorithm such that with probability exceeding $1-\delta$ and when $n \geq \Xi$ and $\epsilon \leq \Xi$,*

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

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

4.2 Kernelized GLMs

The error function for the the Kernelized GLM problem is given by the following equation for a single training pair $(\mathbf{x}_i, y_i) \sim \mathcal{D}$.

$$\mathcal{R}\left(f; \mathbf{x}_i, y_i\right) = \left(\omega(f(\mathbf{x}_i)) - y_i\right)^2$$

Theorem 13 (Subquantile Minimization for Generalized Linear Models is Good with High Probability)**.** *Let Algorithm 1 be run on a dataset $\mathcal{D} \sim \hat{\mathcal{P}}$ with learning rate $\eta \triangleq \Omega(\ell^{-1})$ and link function $\omega : \mathbb{R} \to \mathbb{R}$, s.t. $C_1 \le \omega'(x) \le C_2$ for absolute constants $C_1, C_2 > 0$. Then after $O\left(n \log\left(\frac{\|f^*\|_{\mathcal{H}}}{\varepsilon}\right)\right)$ gradient descent iterations, with probability exceeding $1-\delta$ and a positive constant C,*

$$\left\|f^{(T)} - f^*\right\|_{\mathcal{H}} \le \varepsilon$$

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

Proof. The proof is deferred to § C.1. ■

4.3 Neural Networks

In this section we will consider Iterative Thresholding for a linear one-layer neural network and then a general two-layer neural network.

4.3.1 One Layer Linear Network

We start with the simple case of a linear one-layer neural network for multivariate regression. The error function for the linear one-layer Neural Network problem is given by the following equation for $\mathrm{X} \in \mathbb{R}^{d\times n}$ and $\mathrm{Y} \in \mathbb{R}^{k\times n}$.

$$\mathcal{R}(\mathrm{W}; \mathrm{X}, \mathrm{Y}) = \|\mathrm{WX} - \mathrm{Y}\|_{\mathrm{F}}^2$$

Assumption 14. *The true data is given by the following relation for a $\mathrm{W}^* \in \mathbb{R}^{k\times d}$ and $\mathbf{e} \in \mathbb{R}^k$,*

$$\mathbf{y} = \mathrm{W}^*\mathbf{x} + \mathbf{e}$$

where the elements of $\mathbf{e}$ are sampled from $\mathcal{N}(0, \sigma^2)$.

Theorem 15 (Subquantile Minimization for a One-layer Linear Network is Good with High Probability)**.** *Let Algorithm 1 be run on a dataset $\mathcal{D} \sim \hat{\mathcal{P}}$ such that $\mathrm{X} \in \mathbb{R}^{d\times n}$ and $\mathrm{Y} \in \mathbb{R}^{k\times n}$ with learning rate $\eta \triangleq \Omega(\|\Gamma\|^{-1})$. Then after $O\left(n \log\left(\frac{\|\mathrm{W}^*\|_{\mathrm{F}}}{\varepsilon}\right)\right)$ gradient descent iterations, with probability exceeding $1-\delta$ and a positive constant C,*

$$\left\|\mathrm{W}^{(T)} - \mathrm{W}^*\right\|_{\mathrm{F}} \le \varepsilon + O(k\sigma) + O(k\sigma\|\Gamma\|)$$

for $n = \Omega(\Xi)$.

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

4.3.2 Two Layer Neural Network

We will first consider the case where the output is a scalar, in which case we have the following function for a two-layer neural network given $\mathbf{a} \in \mathbb{R}^{\ell}$ and $\mathrm{W} \in \mathbb{R}^{\ell\times k}$.

$$f_{\mathrm{W},\mathbf{a}}(\mathbf{x}) = \sum_{k=1}^{\ell} a_k \sigma(\mathbf{x}^{\mathrm{T}}\mathbf{w}_k)$$

The squared error is then given by,

$$\mathcal{R}(f; \mathbf{x}, y) = (f(\mathbf{x}) - y)^2$$

We now give our assumption of the data.

Assumption 16. *The true data is given by the following relation for a $\mathrm{W}^* \in \mathbb{R}^{k\times d}$ and $\xi \sim \mathcal{N}(0, \sigma^2)$. ,*

$$y = \sum_{k=1}^{\ell} a_k^* \sigma(\mathbf{x}^{\mathrm{T}}\mathbf{w}_k^*) + \xi = f^*(\mathbf{x}) + \xi$$

where $\sigma : \mathbb{R} \to \mathbb{R}$.

Theorem 17 (Subquantile Minimization for Learning a Single Neuron is Good with High Probability)**.** *Let Algorithm 1 be run on a dataset $\mathcal{D} \sim \hat{\mathcal{P}}$ such that $\mathrm{X} \in \mathbb{R}^{d\times n}$ and $\mathbf{y} \in \mathbb{R}^n$ with learning rate $\eta \triangleq \Omega(\|\Gamma\|^{-1})$. Then after $O\left(n \log\left(\frac{\|\mathrm{W}^*\|_{\mathrm{F}}}{\varepsilon}\right)\right)$ gradient descent iterations, with probability exceeding $1-\delta$ and a positive constant C,*

$$\left\|\mathrm{W}^{(T)} - \mathrm{W}^*\right\|_{\mathrm{F}} \leq \varepsilon + O(k\sigma) + O(k\sigma \|\Gamma\|)$$

for $n = \Omega(\Xi)$.

Proof. The proof is deferred to § D.2 ∎

Theorem 18 (Subquantile Minimization for a One-layer Linear Network is Good with High Probability)**.** *Let Algorithm 1 be run on a dataset $\mathcal{D} \sim \hat{\mathcal{P}}$ such that $\mathrm{X} \in \mathbb{R}^{d\times n}$ and $\mathrm{Y} \in \mathbb{R}^{k\times n}$ with learning rate $\eta \triangleq \Omega(\|\Gamma\|^{-1})$. Then after $O\left(n \log\left(\frac{\|\mathrm{W}^*\|_{\mathrm{F}}}{\varepsilon}\right)\right)$ gradient descent iterations, with probability exceeding $1-\delta$ and a positive constant C,*

$$\left\|\mathrm{W}^{(T)} - \mathrm{W}^*\right\|_{\mathrm{F}} \leq \varepsilon + O(k\sigma) + O(k\sigma \|\Gamma\|)$$

for $n = \Omega(\Xi)$.

5 Discussion

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

Extension to Infinite Dimensional Kernels.

Theory. We develop strong theoretical bounds on the normed difference between the function returned by Subquantile Minimization and the optimal function for data in the target distribution, $\mathcal{P}$, in the sub-Gaussian Design. We are able to show if the number of inliers is sufficiently small, then the kernelized binary classification problem with binary cross-entropy loss is consistent.

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

References

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

[JLT20] Arun Jambulapati, Jerry Li, and Kevin Tian. Robust sub-gaussian principal component analysis and width-independent schatten packing. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 15689–15701. Curran Associates, Inc., 2020.

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

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

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

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

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

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

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

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

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

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

[PP+08] Kaare Brandt Petersen, Michael Syskind Pedersen, et al. The matrix cookbook. *Technical University of Denmark*, 7(15):510, 2008.

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

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

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

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

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

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

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

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

[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.

A Probability Theory

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

A.1 Finite Dimensional Concentrations of Measure

Proposition 19. *Let $\mu_1, \ldots, \mu_n \sim \mathcal{N}(0, \sigma^2)$ for some $\sigma > 0$, then it follows for any $C \geq 1$,*

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

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

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

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

A.2 Hilbert Space Concentrations of Measure

Fact 20 (Sum of Binomial Coefficients [CLRS22])**.** *Let $k, n \in \mathbb{N}$ such that $k \leq n$, then*

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

Proposition 21 (Jensen's Inequality [Jen06])**.** *Suppose φ is a convex function, then for a random variable X, it holds*

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

The inequality is reversed for φ concave.

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

Proposition 22 (McDiarmid's Inequality [M+89])**.** *Suppose $f : \mathcal{X}_1 \times \cdots \mathcal{X}_n \to \mathbb{R}$. Consider i.i.d $X_1, \ldots, X_n$ where $X_i \in \mathcal{X}_i$ for all $i \in [n]$. If there exists constants $c_1, \ldots, c_n$, such that for all $x_i \in \mathcal{X}_i$ for all $i \in [n]$, it holds*

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

Then for any $t > 0$, it holds

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

Theorem 23 (Mean Estimation in the Hilbert Space [TSM+17])**.** *Define $P_n \triangleq \frac{1}{n}\sum_{i=1}^{n} \delta_{X_i}$ and P be the distribution of the covariates in $\mathcal{X}$. Suppose $r : \mathcal{X} \to \mathcal{H}$ is a continuous function such that $\sup_{X\in\mathcal{X}} \|r(X)\|_{\mathcal{H}}^2 \leq C_k < \infty$. Then with probability at least $1 - \delta$,*

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

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

Proposition 24 (Probabilistic Bound on Infinite Dimensional Covariance Estimation)**.** *Let $\mathbf{x}_1, \ldots, \mathbf{x}_n$ be i.i.d sampled from $\mathcal{P}$ such that $\phi(\mathbf{x}_i) \triangleq X_i \sim \mathcal{P}_\sharp\phi$ (Assumption 6). Denote $\mathcal{S}$ as all subsets of $[n]$ with size from $(1-2\epsilon)n$ to $(1-\epsilon)n$. We then have simultaneously with probability exceeding $1-\delta$,*

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

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

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

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

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

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

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

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

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

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

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

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

Then, define the function $r(\mathbf{x}) : \mathcal{X} \to \mathcal{H}\otimes\mathcal{H}$, $\mathbf{x} \to \phi(\mathbf{x})\otimes\phi(\mathbf{x})$. From Assumption 7, we have $r(\mathbf{x}) = \|\phi(\mathbf{x})\otimes\phi(\mathbf{x})\|_{\text{HS}} \leq \|\phi(\mathbf{x})\|_{\mathcal{H}}^2 \leq P_k$. We will use McDiamard's Inequality, consider $\tilde{P} \triangleq \delta_{X_i}$ with one modified element. Then consider the equation $f(x_1,\ldots,x_n) : \mathcal{X}\times\cdots\times\mathcal{X} \to \mathcal{H}\otimes\mathcal{H}\times\cdots\times\mathcal{H}\otimes\mathcal{H}$, $x_1,\ldots,x_n \to \left\|\int_{\mathcal{X}} r(x)dP_B(x) - \int_{\mathcal{X}} r(x)P(x)\right\|_{\text{HS}}$.

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

Then, we have from McDiarmid's inequality (Proposition 22),

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

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

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

Next, applying a union bound over $\mathcal{S}$ with Fact 20, we have

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

Simplifying the resultant bound completes the proof. ∎

A.3 Kernel Matrix Eigenvalue Concentration

Lemma 25. *Let $X_i,\ldots,X_n \sim \mathcal{P}_\sharp\phi$. Let $\mathcal{S}$ represent all permutations of $[n]$ from size $[(1-2\epsilon)n]$ to $[(1-\epsilon)n]$. Form the kernel matrix $\mathrm{K}\in\mathbb{R}^{n\times n}$ s.t. $\mathrm{K}_{ij} \triangleq k(\mathbf{x}_i,\mathbf{x}_j)$. Then with probability exceeding $1-\delta$*

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

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

Proof. We will give our probabilistic bounds using the first and second relation in our covariance estimation bound given in Proposition 24.
Lower Bound.

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

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

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

Upper Bound.

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

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

A.4 Matrix Eigenvalue Concentration

First we will give our bounded covariate corruption assumption.

Assumption 26 (Bounded Covariate Corruption)**.** *There exists a constants C_1, C_2, such that*

$$
\|\Pi X_{\text{FP}}\|_2 \leqslant \sqrt{\epsilon n C_1} \quad \textit{and} \quad \|(I-\Pi)X_{\text{FP}}\|_2 \leqslant \sqrt{\epsilon n C_2}
$$

Lemma 27 (Norm of Subset of Good Covariates)**.** *Let X_* represent a subset of the good data, then it follows,*

$$
\|X_*\|_2 \leqslant \sqrt{n\|\Gamma\|}
$$

B Proofs for Structural Results

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

Lemma 28. *Consider a determinate set of numbers $(a_i)_{i=1}^n$, and determinate set of functions in the Hilbert Space, $(X_i)_{i=1}^n$. It then follows,*

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

Proof. The proof is a calculation.

$$
\begin{aligned}
\left\|\sum_{i=1}^n \alpha_i X_i\right\|_{\mathcal{H}}^2 \overset{(i)}{\leqslant} \|\boldsymbol{\alpha}\|_2^2 \max_{\mathbf{v}\in\mathbb{S}^{n-1}}\left\|\sum_{i=1}^n v_i X_i\right\|_{\mathcal{H}}^2 &= \|\boldsymbol{\alpha}\|_2^2 \max_{\mathbf{v}\in\mathbb{S}^{n-1}}\left\langle\sum_{i=1}^n v_i X_i, \sum_{j=1}^n v_j X_j\right\rangle_{\mathcal{H}} \\
&= \|\boldsymbol{\alpha}\|_2^2 \max_{\mathbf{v}\in\mathbb{S}^{n-1}}\sum_{i=1}^n\sum_{j=1}^n v_i v_j k(x_i, x_j) = \|\boldsymbol{\alpha}\|_2^2 \max_{\mathbf{v}\in\mathbb{S}^{n-1}}\mathbf{v}^\top K\mathbf{v} = \|\boldsymbol{\alpha}\|_2^2\|K\|
\end{aligned}
$$

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

Lemma 29. *Let $\boldsymbol{\alpha}, \boldsymbol{\beta} \in \mathbb{R}^n$ and $X \in \mathbb{R}^{p\times n}$, then the following holds,*

$$
\|\sum_{i=1}^n \alpha_i\beta_i\mathbf{x}_i\|_2^2 \leqslant n\|\boldsymbol{\alpha}\|_\infty\|\sum_{i=1}^n \beta_i\mathbf{x}_i\|_2^2
$$

Proof. The proof is a simple calculation.

$$\Big\|\sum_{i=1}^{n}\alpha_i\beta_i\mathbf{x}_i\Big\|_2^2=\Big\|\sum_{i=1}^{n}\beta_i(\alpha_i\mathbf{x}_i)\Big\|_2^2\leqslant\|\boldsymbol{\alpha}\|^2\Big\|\sum_{i=1}^{n}\beta_i\mathbf{x}_i\Big\|^2\leqslant n\,\|\boldsymbol{\alpha}\|_\infty^2\Big\|\sum_{i=1}^{n}\beta_i\mathbf{x}_i\Big\|_2^2$$

■

C Proofs for Kernelized GLMs

In this section, we will prove error bounds for Subquantile Minimization in the Kernelized GLM Problem. First we give the necessary results for our analysis.

Lemma 30 (Lemma 3.11 [B+15]). *Let f be β-smooth and α-strongly convex over $\mathbb{R}^n$, then for all $\mathbf{x},\mathbf{y}\in\mathbb{R}^n$, it follows,*

$$\langle\nabla f(\mathbf{x})-\nabla f(\mathbf{y}),\mathbf{x}-\mathbf{y}\rangle\geq\frac{\alpha\beta}{\alpha+\beta}\|\mathbf{x}-\mathbf{y}\|^2+\frac{1}{\alpha+\beta}\|\nabla f(\mathbf{x})-\nabla f(\mathbf{y})\|^2$$

Proposition 31 (Young's Inequality [You12]). *Suppose $a,b\in\mathbb{R}_+$, then for $p,q>1$ such that $\frac{1}{p}+\frac{1}{q}=1$, it follows*

$$ab\leq\frac{a^p}{p}+\frac{b^q}{q}$$

We are now ready to prove our main approximation bound.

C.1 Proof of Theorem 13

Proof. From Algorithm 1, we have for the generalized linear model.

$$f^{(t+1)}=f^{(t)}-\frac{\eta}{(1-\epsilon)n}\cdot\sum_{i\in S^{(t)}}(\omega(f^{(t)}(\mathbf{x}_i))-y_i)\cdot\omega'(f^{(t)}(\mathbf{x}_i))\cdot X_i$$

We then have,

$$\begin{aligned}\big\|f^{(t+1)}-f^*\big\|_{\mathcal{H}}&=\big\|f^{(t)}-\eta\nabla_f\mathcal{R}_{\mathrm{S}^{(t)}}(f^{(t)})-f^*\big\|_{\mathcal{H}}\\&=\big\|f^{(t)}-f^*-\eta\nabla\mathcal{R}_{\mathrm{TP}}(f^{(t)})-\eta\nabla\mathcal{R}_{\mathrm{FP}}(f^{(t)})\big\|_{\mathcal{H}}\\&\leq\big\|f^{(t)}-f^*-\eta\nabla\mathcal{R}_{\mathrm{TP}}(f^{(t)})\big\|_{\mathcal{H}}+\big\|\eta\nabla\mathcal{R}_{\mathrm{FP}}(f^{(t)})\big\|_{\mathcal{H}}\end{aligned}\tag{2}$$

We will now analyze the first term of Equation 2 through its square,

$$\big\|f^{(t)}-f^*-\eta\nabla\mathcal{R}_{\mathrm{TP}}(f^{(t)})\big\|_{\mathcal{H}}^2=\big\|f^{(t)}-f^*\big\|_{\mathcal{H}}^2-2\eta\cdot\big\langle f^{(t)}-f^*,\nabla\mathcal{R}_{\mathrm{TP}}(f^{(t)})\big\rangle_{\mathcal{H}}+\eta^2\cdot\big\|\nabla\mathcal{R}_{\mathrm{TP}}(f^{(t)})\big\|_{\mathcal{H}}^2\tag{3}$$

We then have,

$$\begin{aligned}&\big\langle f^{(t)}-f^*,\nabla\mathcal{R}_{\mathrm{TP}}(f^{(t)})\big\rangle_{\mathcal{H}}\overset{\text{def}}{=}\frac{1}{(1-\epsilon)n}\cdot\Big\langle f^{(t)}-f^*,\sum_{i\in\mathrm{TP}}((\omega(f^{(t)}(\mathbf{x}_i))-\omega(f^*(\mathbf{x}_i))-\xi_i)\cdot\omega'(f^{(t)}(\mathbf{x}_i))\cdot X_i\Big\rangle_{\mathcal{H}}\\&\geq\frac{C_1}{(1-\epsilon)n}\Big\langle f^{(t)}-f^*,\sum_{i\in\mathrm{TP}}((\omega(f^{(t)}(\mathbf{x}_i))-\omega(f^*(\mathbf{x}_i)))\cdot X_i\Big\rangle_{\mathcal{H}}-\frac{1}{(1-\epsilon)n}\Big\langle f^{(t)}-f^*,\sum_{i\in\mathrm{TP}}\xi_i\omega'(f^{(t)}(\mathbf{x}_i))\cdot X_i\Big\rangle_{\mathcal{H}}\end{aligned}\tag{4}$$

In the above, in the first inequality we can note that as the link function is monotonic increasing, we have $(f^{(t)}(\mathbf{x})-f^*(\mathbf{x}))(\omega(f^{(t)}(\mathbf{x}))-\omega(f^*(\mathbf{x})))\geq0$ for any $\mathbf{x}\in\mathcal{X}$. We will first lower bound the first termin Equation 4. Then, consider the function, $h:\mathcal{H}\to\mathbb{R}$, for a Dirac measure $P(X)=\frac{1}{|\mathrm{TP}|}\delta_{\mathrm{TP}}(X)$.

$$h(f)\triangleq\int_{\mathcal{H}}\left(\int\omega(y)dy\right)(f(X))dP(X)$$

The first derivative gives us the following,

$$\nabla h(f) \triangleq \int_{\mathcal{H}} \omega(f(X)) \cdot X \, dP(X)$$

From which we have the following,

$$\begin{aligned}\nabla^2 h(f) &\triangleq \int_{\mathcal{H}} \omega'(f(X)) \cdot X \otimes X \, dP(X) \succcurlyeq \min_{y \in \mathbb{R}} \omega'(y) \cdot \lambda_{\min} \left(\int_{\mathcal{H}} X \otimes X dP(X) \right) \cdot \mathrm{Id} \\ &\stackrel{\text{def}}{=} C_1 \lambda_{\min} \left(\int_{\mathcal{H}} X \otimes X dP(X) \right) \cdot \mathrm{Id} \triangleq \alpha \cdot \mathrm{Id}\end{aligned}$$

Finally, we can note that

$$\nabla^2 h(f) \preccurlyeq \max_{y \in \mathbb{R}} \omega'(y) \cdot \lambda_{\max} \left(\int_{\mathcal{H}} X \otimes X \, dP(X) \right) \cdot \mathrm{Id} \stackrel{\text{def}}{=} C_2 \cdot \lambda_{\max} \left(\int_{\mathcal{H}} X \otimes X \, dP(X) \right) \cdot \mathrm{Id} \triangleq \beta \cdot \mathrm{Id}$$

Then, with Lemma 30, we have

$$\begin{aligned}2\eta \cdot \left\langle f^{(t)} - f^*, \nabla \mathcal{R}_{\mathrm{TP}}(f^{(t)}) \right\rangle_{\mathcal{H}} \geq &\frac{\eta C_1 \lambda_{\min}(\Phi_{\mathrm{TP}} \otimes \Phi_{\mathrm{TP}})}{(1-\epsilon)n} \left\| f^{(t)} - f^* \right\|_{\mathcal{H}}^2 \\ &+ \frac{\eta}{C_2(1-\epsilon)n\lambda_{\max}(\Phi_{\mathrm{TP}} \otimes \Phi_{\mathrm{TP}})} \Big\| \sum_{i \in \mathrm{TP}} (\omega(f^{(t)}(\mathbf{x}_i)) - \omega(f^*(\mathbf{x}_i))) \cdot X_i \Big\|_{\mathcal{H}}^2 \quad (5)\end{aligned}$$

We will now upper bound the second term in Equation 4.

$$\begin{aligned}\frac{\eta}{(1-\epsilon)n} \cdot \langle f^{(t)} - f^*, \sum_{i \in \mathrm{TP}} \xi_i \omega'(f^{(t)}(\mathbf{x}_i)) \cdot X_i \rangle_{\mathcal{H}} &\leq \frac{\eta}{(1-\epsilon)n} \cdot \left\| f^{(t)} - f^* \right\|_{\mathcal{H}} \Big\| \sum_{i \in \mathrm{TP}} \xi_i \omega'(f^{(t)}(\mathbf{x}_i)) \cdot X_i \Big\|_{\mathcal{H}} \\ &\leq \frac{\eta^2}{2} \cdot \left\| f^{(t)} - f^* \right\|_{\mathcal{H}}^2 + \frac{1}{2\left[(1-\epsilon)n\right]^2} \cdot \Big\| \sum_{i \in \mathrm{TP}} \xi_i \omega'(f^{(t)}(\mathbf{x}_i)) \cdot X_i \Big\|_{\mathcal{H}}^2 \\ &\leq \frac{\eta^2}{2} \cdot \left\| f^{(t)} - f^* \right\|_{\mathcal{H}}^2 + \frac{C_2^2}{2(1-\epsilon)n} \cdot \Big\| \sum_{i \in \mathrm{TP}} \xi_i X_i \Big\|_{\mathcal{H}}^2 \quad (6)\end{aligned}$$

In the above, in the first inequality we applied the Cauchy-Schwarz Inequality, in the second inequality we utilize Young's Inequality (see Proposition 31), in the third inequality we applied Lemma 29. Then for the third term of Equation 3, we have

$$\begin{aligned}&\left\| \eta \nabla \mathcal{R}_{\mathrm{TP}}(f^{(t)}) \right\|_{\mathcal{H}}^2 \stackrel{\text{def}}{=} \frac{\eta^2}{\left[(1-\epsilon)n\right]^2} \cdot \Big\| \sum_{i \in \mathrm{TP}} (\omega(f^{(t)}(\mathbf{x}_i)) - y_i) \cdot \omega'(f^{(t)}(\mathbf{x}_i)) \cdot X_i \Big\|_{\mathcal{H}}^2 \\ &= \frac{\eta^2}{\left[(1-\epsilon)n\right]^2} \cdot \Big\| \sum_{i \in \mathrm{TP}} (\omega(f^{(t)}(\mathbf{x}_i)) - \omega(f^*(\mathbf{x}_i)) + \xi_i) \cdot \omega'(f^{(t)}(\mathbf{x}_i)) \cdot X_i \Big\|_{\mathcal{H}}^2 \\ &\leq \frac{2\eta^2}{\left[(1-\epsilon)n\right]^2} \cdot \Big(\Big\| \sum_{i \in \mathrm{TP}} (\omega(f^{(t)}(\mathbf{x}_i)) - \omega(f^*(\mathbf{x}_i)) \cdot \omega'(f^{(t)}(\mathbf{x}_i)) \cdot X_i \Big\|_{\mathcal{H}}^2 + \Big\| \sum_{i \in \mathrm{TP}} \xi_i \omega'(f^{(t)}(\mathbf{x}_i)) \cdot X_i \Big\|_{\mathcal{H}}^2 \Big) \\ &\leq \frac{2C_2^2 \eta^2}{(1-\epsilon)n} \cdot \Big\| \sum_{i \in \mathrm{TP}} (\omega(f^{(t)}(\mathbf{x}_i)) - \omega(f^*(\mathbf{x}_i)) \cdot X_i \Big\|_{\mathcal{H}}^2 + \frac{2C_2^2 \eta^2}{(1-\epsilon)n} \cdot \Big\| \sum_{i \in \mathrm{TP}} \xi_i X_i \Big\|_{\mathcal{H}}^2 \quad (7)\end{aligned}$$

In the above, in the first inequality we utilize the elementary inequality $(a+b)^2 \leq 2a^2 + 2b^2$, in the second inequality we utilize Lemma 29. We then see the second term of Equation 5 is greater than the first term of Equation 7 when $\eta \leq (2C_2 \left\| \Phi_{\mathrm{TP}} \otimes \Phi_{\mathrm{TP}} \right\|)^{-1}$. We now will upper bound second term of Equation 2 through its square.

$$\left\| \eta \nabla \mathcal{R}_{\mathrm{FP}}(f^{(t)}) \right\|_{\mathcal{H}}^2 \stackrel{\text{def}}{=} \frac{\eta^2}{\left[(1-\epsilon)n\right]^2} \cdot \Big\| \sum_{i \in \mathrm{FP}} (\omega(f^{(t)}(\mathbf{x}_i)) - y_i) \cdot \omega'(f^{(t)}(\mathbf{x}_i)) \cdot X_i \Big\|_{\mathcal{H}}^2$$

$$\le \frac{\eta^2}{[(1-\epsilon)n]^2} \cdot \Big\| \sum_{i\in\mathrm{FP}} X_i \otimes X_i \Big\|_{\mathrm{op}} \sum_{i\in\mathrm{FP}} \left[(\omega(f^{(t)}(\mathbf{x}_i)) - y_i) \cdot \omega'(f^{(t)}(\mathbf{x}_i))\right]^2$$
$$\le \frac{C_2^2\eta^2}{[(1-\epsilon)n]^2} \cdot \Big\| \sum_{i\in\mathrm{FP}} X_i \otimes X_i \Big\|_{\mathrm{op}} \sum_{i\in\mathrm{FN}} (\omega(f^{(t)}(\mathbf{x}_i)) - y_i)^2$$
$$= \frac{C_2^2\eta^2}{[(1-\epsilon)n]^2} \cdot \Big\| \sum_{i\in\mathrm{FP}} X_i \otimes X_i \Big\|_{\mathrm{op}} \sum_{i\in\mathrm{FN}} (\omega(f^{(t)}(\mathbf{x}_i)) - \omega(f^*(\mathbf{x}_i)) + \xi_i)^2$$
$$\le \frac{C_2^4\eta^2}{[(1-\epsilon)n]^2} \cdot \Big\| \sum_{i\in\mathrm{FP}} X_i \otimes X_i \Big\|_{\mathrm{op}} \Big(\sum_{i\in\mathrm{FN}} (f^{(t)}(\mathbf{x}_i) - f^*(\mathbf{x}_i))^2 + \xi_i^2\Big)$$
$$\le \frac{2C_2^4\eta^2}{[(1-\epsilon)n]^2} \cdot \Big\| \sum_{i\in\mathrm{FP}} X_i \otimes X_i \Big\|_{\mathrm{op}} \Big(\Big\| \sum_{i\in\mathrm{FN}} X_i \otimes X_i \Big\|_{\mathrm{op}} \big\| f^{(t)} - f^* \big\|_{\mathcal{H}}^2 + \|\boldsymbol{\xi}_{\mathrm{FN}}\|^2 \Big) \tag{8}$$

In the above, the first inequality follows from Lemma 28, the second inequality follows from the optimality of the Subquantile set and the bounded link function gradient, the third inequality follows from noting for any $x, y \in \mathbb{R}$, the bounded gradient implies Lipschitzness, i.e. $|\omega(x) - \omega(y)| \le C_2|x-y|$. The final equality follows from the following,

$$\sum_{i\in\mathrm{FN}} (f^{(t)}(\mathbf{x}_i) - f^*(\mathbf{x}_i))^2 = \Big\langle f^{(t)} - f^*, \Big[\sum_{i\in\mathrm{FN}} X_i \otimes X_i\Big](f^{(t)} - f^*) \Big\rangle_{\mathcal{H}} \overset{\mathrm{def}}{=} \big\| f^{(t)} - f^* \big\|_{\Sigma_{\mathrm{FN}},\mathcal{H}}^2$$
$$\le \big\| f^{(t)} - f^* \big\|_{\mathcal{H}}^2 \Big\| \sum_{i\in\mathrm{FN}} X_i \otimes X_i \Big\|_{\mathrm{op}}$$

Then from Equations 5, 6, and 8, we have

$$\big\| f^{(t+1)} - f^* \big\|_{\mathcal{H}} \le \frac{\eta C_2 \|\boldsymbol{\xi}_{\mathrm{FN}}\| \sqrt{\big\|\sum_{i\in\mathrm{FP}} X_i \otimes X_i\big\|_{\mathrm{op}}}}{(1-\epsilon)n} + \Big(\frac{C_2^2((1/2) + 2\eta^2)}{(1-\epsilon)n}\Big) \Big\| \sum_{i\in\mathrm{TP}} \xi_i X_i \Big\|_{\mathcal{H}}$$
$$+ \big\| f^{(t)} - f^* \big\|_{\mathcal{H}} \Big(1 - \eta\Big(\frac{1}{\sqrt{2}} + \frac{\lambda_{\min}\big(\sum_{i\in\mathrm{TP}} X_i \otimes X_i\big)}{2(1-\epsilon)n} - \frac{C_2\sqrt{2C_3\epsilon \cdot \big\|\sum_{i\in\mathrm{FN}} X_i \otimes X_i\big\|_{\mathrm{op}}}}{(1-\epsilon)\sqrt{n}}\Big)\Big)$$

■

D Proofs for Neural Networks

D.1 Proof of Theorem 15

Proof. Recall that for any $\mathrm{W} \in \mathbb{R}^{k\times n}$,

$$\mathcal{R}(\mathrm{W}) = \|\mathrm{WX} - \mathrm{Y}\|_{\mathrm{F}}^2 = \mathrm{Tr}(\mathrm{X}^{\mathrm{T}}\mathrm{W}^{\mathrm{T}}\mathrm{WX} - \mathrm{X}^{\mathrm{T}}\mathrm{W}^{\mathrm{T}}\mathrm{Y} - \mathrm{Y}^{\mathrm{T}}\mathrm{WX} + \mathrm{Y}^{\mathrm{T}}\mathrm{Y})$$
$$= \mathrm{Tr}(\mathrm{X}^{\mathrm{T}}\mathrm{W}^{\mathrm{T}}\mathrm{WX}) + \mathrm{Tr}(\mathrm{Y}^{\mathrm{T}}\mathrm{Y}) - 2\,\mathrm{Tr}(\mathrm{X}^{\mathrm{T}}\mathrm{W}^{\mathrm{T}}\mathrm{Y})$$

Then, from [PP+08] Equations (102) and (119) (where we set $\mathrm{B} = \mathrm{Id}$ and $\mathrm{C} = 0$). We have,

$$\nabla_{\mathrm{W}}\mathcal{R}(\mathrm{W}) = 2(\mathrm{WX} - \mathrm{Y})\mathrm{X}^{\mathrm{T}}$$

Our proof will begin similarly to the proof of Theorem 13. We then have,

$$\big\|\mathrm{W}^{(t+1)} - \mathrm{W}^*\big\|_{\mathrm{F}} = \big\|\mathrm{W}^{(t)} - \mathrm{W}^* - \eta\nabla_{\mathrm{W}}\mathcal{R}_{\mathrm{S}^{(t)}}(\mathrm{W}^{(t)})\big\|_{\mathrm{F}}$$
$$= \big\|\mathrm{W}^{(t)} - \mathrm{W}^* - \eta\nabla_{\mathrm{W}}\mathcal{R}_{\mathrm{TP}}(\mathrm{W}^{(t)}) - \eta\nabla_{\mathrm{W}}\mathcal{R}_{\mathrm{FP}}(\mathrm{W}^{(t)})\big\|_{\mathrm{F}}$$
$$\le \big\|\mathrm{W}^{(t)} - \mathrm{W}^* - \eta\nabla_{\mathrm{W}}\mathcal{R}_{\mathrm{TP}}(\mathrm{W}^{(t)})\big\|_{\mathrm{F}} + \big\|\eta\nabla_{\mathrm{W}}\mathcal{R}_{\mathrm{FP}}(\mathrm{W}^{(t)})\big\|_{\mathrm{F}} \tag{9}$$

We first will upper bound the first term in Equation 9.

$$\begin{aligned}
&\left\|\mathrm{W}^{(t)}-\mathrm{W}^{*}-\nabla_{\mathrm{W}}\mathcal{R}_{\mathrm{TP}}(\mathrm{W}^{(t)})\right\|_{\mathrm{F}}^{2} \\
&\quad=\left\|\mathrm{W}^{(t)}-\mathrm{W}^{*}\right\|_{\mathrm{F}}^{2}-\eta\cdot\mathrm{Tr}\big((\mathrm{W}^{(t)}-\mathrm{W}^{*})^{\mathrm{T}}(\nabla_{\mathrm{W}}\mathcal{R}_{\mathrm{TP}}(\mathrm{W}^{(t)}))\big)+\left\|\eta\nabla_{\mathrm{W}}\mathcal{R}_{\mathrm{TP}}(\mathrm{W}^{(t)})\right\|_{\mathrm{F}}^{2}
\end{aligned}\tag{10}$$

We then lower bound the second term in Equation 10,

$$\begin{aligned}
&\eta\cdot\mathrm{Tr}\big((\mathrm{W}^{(t)}-\mathrm{W}^{*})^{\mathrm{T}}(\nabla_{\mathrm{W}}\mathcal{R}_{\mathrm{TP}}(\mathrm{W}^{(t)}))\big)\stackrel{\text{def}}{=}2\eta\cdot\mathrm{Tr}\big((\mathrm{W}^{(t)}-\mathrm{W}^{*})^{\mathrm{T}}(\mathrm{W}\mathrm{X}_{\mathrm{TP}}-\mathrm{Y}_{\mathrm{TP}})\mathrm{X}_{\mathrm{TP}}^{\mathrm{T}}\big) \\
&=\frac{2\eta}{(1-\epsilon)n}\cdot\mathrm{Tr}\big((\mathrm{W}^{(t)}-\mathrm{W}^{*})^{\mathrm{T}}(\mathrm{W}^{(t)}-\mathrm{W}^{*})\mathrm{X}_{\mathrm{TP}}\mathrm{X}_{\mathrm{TP}}^{\mathrm{T}}-(\mathrm{W}^{(t)}-\mathrm{W}^{*})^{\mathrm{T}}\mathrm{E}_{\mathrm{TP}}\mathrm{X}_{\mathrm{TP}}^{\mathrm{T}}\big) \\
&=\frac{2\eta}{(1-\epsilon)n}\cdot\sum_{\ell=1}^{k}\Big\langle\mathbf{w}_{\ell}^{(t)}-\mathbf{w}_{\ell}^{*},\sum_{i\in\mathrm{TP}}(\mathbf{x}_{i}^{\mathrm{T}}\mathbf{w}^{(t)}-\mathbf{x}_{i}^{\mathrm{T}}\mathbf{w}_{\ell}^{*})\cdot\mathbf{x}_{i}\Big\rangle-\frac{2\eta}{(1-\epsilon)n}\cdot\left\|\mathrm{W}^{(t)}-\mathrm{W}^{*}\right\|_{\mathrm{F}}\left\|\mathrm{E}_{\mathrm{TP}}\mathrm{X}_{\mathrm{TP}}^{\mathrm{T}}\right\|_{\mathrm{F}} \\
&\geq\frac{\eta}{(1-\epsilon)n}\cdot\sum_{\ell=1}^{k}\Big(\lambda_{\min}(\mathrm{X}_{\mathrm{TP}}\mathrm{X}_{\mathrm{TP}}^{\mathrm{T}})\left\|\mathbf{w}_{\ell}^{(t)}-\mathbf{w}_{\ell}^{*}\right\|^{2}+\Big\|\sum_{i\in\mathrm{TP}}(\mathbf{x}_{i}^{\mathrm{T}}\mathbf{w}_{\ell}^{(t)}-\mathbf{x}_{i}^{\mathrm{T}}\mathbf{w}_{\ell}^{*})\cdot\mathbf{x}_{i}\Big\|_{2}^{2}\Big) \\
&\quad-\frac{\eta^{2}}{2}\cdot\left\|\mathrm{W}^{(t)}-\mathrm{W}^{*}\right\|_{\mathrm{F}}^{2}-\frac{4}{[(1-\epsilon)n]^{2}}\cdot\left\|\mathrm{X}_{\mathrm{TP}}\mathrm{X}_{\mathrm{TP}}^{\mathrm{T}}\right\|_{2}\left\|\mathrm{E}_{\mathrm{TP}}\right\|_{\mathrm{F}}^{2} \\
&=\frac{\eta}{(1-\epsilon)n}\cdot\lambda_{\min}(\mathrm{X}_{\mathrm{TP}}\mathrm{X}_{\mathrm{TP}}^{\mathrm{T}})\left\|\mathrm{W}^{(t)}-\mathrm{W}^{*}\right\|_{\mathrm{F}}^{2}+\frac{\eta}{(1-\epsilon)n}\cdot\left\|(\mathrm{W}^{(t)}-\mathrm{W}^{*})\mathrm{X}_{\mathrm{TP}}\mathrm{X}_{\mathrm{TP}}^{\mathrm{T}}\right\|_{\mathrm{F}}^{2} \\
&\quad-\frac{\eta^{2}}{2}\cdot\left\|\mathrm{W}^{(t)}-\mathrm{W}^{*}\right\|_{\mathrm{F}}^{2}-\frac{4}{[(1-\epsilon)n]^{2}}\cdot\left\|\mathrm{X}_{\mathrm{TP}}\mathrm{X}_{\mathrm{TP}}^{\mathrm{T}}\right\|_{2}\left\|\mathrm{E}_{\mathrm{TP}}\right\|_{\mathrm{F}}^{2}
\end{aligned}$$

In the above, int the first inequality we apply Lemma 30. We now upper bound the second term in Equation 9,

$$\begin{aligned}
&\left\|\eta\nabla_{\mathrm{W}}\mathcal{R}_{\mathrm{FP}}(\mathrm{W}^{(t)})\right\|_{\mathrm{F}}^{2}\stackrel{\text{def}}{=}\frac{4\eta^{2}}{[(1-\epsilon)n]^{2}}\cdot\left\|(\mathrm{W}^{(t)}\mathrm{X}_{\mathrm{FP}}-\mathrm{Y}_{\mathrm{FP}})\mathrm{X}_{\mathrm{FP}}^{\mathrm{T}}\right\|_{\mathrm{F}}^{2} \\
&\leq\frac{4\eta^{2}}{[(1-\epsilon)n]^{2}}\cdot\left\|\mathrm{X}_{\mathrm{FP}}\mathrm{X}_{\mathrm{FP}}^{\mathrm{T}}\right\|_{2}\left\|\mathrm{W}^{(t)}\mathrm{X}_{\mathrm{FP}}-\mathrm{Y}_{\mathrm{FP}}\right\|_{\mathrm{F}}^{2} \\
&\leq\frac{4\eta^{2}}{[(1-\epsilon)n]^{2}}\cdot\left\|\mathrm{X}_{\mathrm{FP}}\mathrm{X}_{\mathrm{FP}}^{\mathrm{T}}\right\|_{2}\left\|\mathrm{W}^{(t)}\mathrm{X}_{\mathrm{FN}}-\mathrm{Y}_{\mathrm{FN}}\right\|_{\mathrm{F}}^{2} \\
&\leq\frac{8\eta^{2}}{[(1-\epsilon)n]^{2}}\cdot\left\|\mathrm{X}_{\mathrm{FP}}\mathrm{X}_{\mathrm{FP}}^{\mathrm{T}}\right\|_{2}\big(\left\|\mathrm{W}^{(t)}\mathrm{X}_{\mathrm{FN}}-\mathrm{W}^{*}\mathrm{X}_{\mathrm{FN}}\right\|_{\mathrm{F}}^{2}+\left\|\mathrm{E}_{\mathrm{FN}}\right\|_{\mathrm{F}}^{2}\big) \\
&\leq\frac{8\eta^{2}}{[(1-\epsilon)n]^{2}}\cdot\left\|\mathrm{X}_{\mathrm{FP}}\mathrm{X}_{\mathrm{FP}}^{\mathrm{T}}\right\|_{2}\left\|\mathrm{X}_{\mathrm{FN}}\mathrm{X}_{\mathrm{FN}}^{\mathrm{T}}\right\|_{2}\left\|\mathrm{W}^{(t)}-\mathrm{W}^{*}\right\|_{\mathrm{F}}^{2}+\frac{8\eta^{2}}{[(1-\epsilon)n]^{2}}\cdot\left\|\mathrm{X}_{\mathrm{FP}}\mathrm{X}_{\mathrm{FP}}\right\|_{2}\left\|\mathrm{E}_{\mathrm{FN}}\right\|_{\mathrm{F}}^{2}
\end{aligned}$$

In the above, the first and fourth inequalities from the fact that for any two size compatible matrices, A,B, it holds that $\|\mathrm{AB}\|_{\mathrm{F}}\leq\|\mathrm{A}\|_{\mathrm{F}}\|\mathrm{B}\|_{2}$, the second inequality follows from the optimality of the Subquantile set, the third inequailty follows from the sub-additivity of the Frobenius norm. We will now upper bound the third term in Equation 10,

$$\begin{aligned}
\left\|\eta\nabla_{\mathrm{W}}\mathcal{R}_{\mathrm{FP}}(\mathrm{W}^{(t)})\right\|_{\mathrm{F}}^{2}&\stackrel{\text{def}}{=}\frac{4\eta^{2}}{[(1-\epsilon)n]^{2}}\cdot\left\|(\mathrm{W}^{(t)}\mathrm{X}_{\mathrm{TP}}-\mathrm{Y}_{\mathrm{FP}})\mathrm{X}_{\mathrm{TP}}^{\mathrm{T}}\right\|_{\mathrm{F}}^{2} \\
&\leq\frac{8\eta^{2}}{[(1-\epsilon)n]^{2}}\cdot\left\|(\mathrm{W}^{(t)}-\mathrm{W}^{*})\mathrm{X}_{\mathrm{TP}}\mathrm{X}_{\mathrm{TP}}^{\mathrm{T}}\right\|_{\mathrm{F}}^{2}+\frac{8\eta^{2}}{[(1-\epsilon)n]^{2}}\cdot\left\|\mathrm{X}_{\mathrm{TP}}\mathrm{X}_{\mathrm{TP}}^{\mathrm{T}}\right\|_{2}\left\|\mathrm{E}_{\mathrm{TP}}\right\|_{\mathrm{F}}^{2}
\end{aligned}$$

We then have,

$$\left\|\mathrm{W}^{(t+1)}-\mathrm{W}^{*}\right\|_{\mathrm{F}}\leq\left\|\mathrm{W}^{(t)}-\mathrm{W}^{*}\right\|_{\mathrm{F}}\left(1-\eta\left(\frac{1}{\sqrt{2}}+\frac{\lambda_{\min}(\mathrm{X}_{\mathrm{TP}}\mathrm{X}_{\mathrm{TP}}^{\mathrm{T}})}{(1-\epsilon)n}-\frac{\sqrt{8\epsilon}C_{3}}{(1-\epsilon)\sqrt{n}}\cdot\|\mathrm{X}_{\mathrm{FP}}\|_{2}\right)\right)$$

$$+\frac{\sqrt{8}\eta}{(1-\epsilon)n}\cdot\|\mathrm{E_{FN}}\|_{\mathrm{F}}\|\mathrm{X_{FP}}\|_2+\frac{2}{(1-\epsilon)n}\cdot\|\mathrm{E_{TP}}\|_{\mathrm{F}}\|\mathrm{X_{TP}}\|_2$$

■

D.2 Proof of Theorem 17 (In Progress)

Proof. Recall the function for a single neuron is given as follows,

$$f_{\mathbf{w},a}(\mathbf{x})=a\cdot\sigma(\mathbf{x}^{(t)}\mathbf{w})$$

Then the gradients are given as follows,

$$\begin{aligned}\nabla_a\mathcal{R}(\mathbf{w},a)&=\frac{2}{n}\cdot\sum_{i=1}^{n}(f(\mathbf{x}_i)-y_i)\cdot\sigma(\mathbf{x}_i^{\mathrm{T}}\mathbf{w})\\ \nabla_{\mathbf{w}}\mathcal{R}(\mathbf{w},a)&=\frac{2}{n}\cdot\sum_{i=1}^{n}(f(\mathbf{x}_i)-y_i)\cdot a\cdot\sigma'(\mathbf{x}_i^{\mathrm{T}}\mathbf{w})\cdot\mathbf{x}_i\end{aligned}$$

Step 1: Upper bounding the norm of the difference between $\mathbf{w}^{(t+1)}$ and $\mathbf{w}^*$.

$$\begin{aligned}\left\|\mathbf{w}^{(t+1)}-\mathbf{w}^*\right\|&=\left\|\mathbf{w}^{(t)}-\mathbf{w}^*-\eta\nabla_{\mathbf{w}}\mathcal{R}(\mathbf{w}^{(t)},a^{(t)};\mathrm{S}^{(t)})\right\|_2\\&=\left\|\mathbf{w}^{(t)}-\mathbf{w}^*-\eta\nabla_{\mathbf{w}}\mathcal{R}(\mathbf{w}^{(t)},a^{(t)};\mathrm{TP})-\eta\nabla_{\mathbf{w}}\mathcal{R}(\mathbf{w}^{(t)},a^{(t)};\mathrm{FP})\right\|_2\\&\le\left\|\mathbf{w}^{(t)}-\mathbf{w}^*-\eta\nabla_{\mathbf{w}}\mathcal{R}(\mathbf{w}^{(t)},a^{(t)};\mathrm{TP})\right\|_2+\left\|\eta\nabla_{\mathbf{w}}\mathcal{R}(\mathbf{w}^{(t)},a^{(t)};\mathrm{FP})\right\|_2\end{aligned}\tag{11}$$

We will expand the first term of Equation 11 through its square.

$$\begin{aligned}&\left\|\mathbf{w}^{(t)}-\mathbf{w}^*-\eta\nabla_{\mathbf{w}}\mathcal{R}(\mathbf{w}^{(t)},a^{(t)};\mathrm{TP})\right\|_2^2\\&\quad=\left\|\mathbf{w}^{(t)}-\mathbf{w}^*\right\|_2^2-2\eta\cdot\left\langle\mathbf{w}^{(t)}-\mathbf{w}^*,\nabla_{\mathbf{w}}\mathcal{R}(\mathbf{w}^{(t)},a^{(t)};\mathrm{TP})\right\rangle+\eta^2\cdot\left\|\nabla_{\mathbf{w}}\mathcal{R}(\mathbf{w}^{(t)},a^{(t)};\mathrm{TP})\right\|_2^2\end{aligned}\tag{12}$$

For the second term of Equation 12, we have

$$\begin{aligned}2\eta\cdot\left\langle\mathbf{w}^{(t)}-\mathbf{w}^*,\nabla_{\mathbf{w}}\mathcal{R}(\mathbf{w}^{(t)},a^{(t)};\mathrm{TP})\right\rangle&=\frac{4\eta}{(1-\epsilon)n}\cdot\left\langle\mathbf{w}^{(t)}-\mathbf{w}^*,\sum_{i\in\mathrm{TP}}(f^{(t)}(\mathbf{x}_i)-y_i)\cdot a^{(t)}\sigma'(\mathbf{x}_i^{\mathrm{T}}\mathbf{w})\cdot\mathbf{x}_i\right\rangle\\&=\frac{4\eta}{(1-\epsilon)n}\cdot\left\langle\mathbf{w}^{(t)}-\mathbf{w}^*,\sum_{i\in\mathrm{TP}}(a^{(t)}\sigma(\mathbf{x}_i^{\mathrm{T}}\mathbf{w}^{(t)})-a^*\sigma(\mathbf{x}_i^{\mathrm{T}}\mathbf{w}^*))\cdot a^{(t)}\sigma'(\mathbf{x}_i^{\mathrm{T}}\mathbf{w})\cdot\mathbf{x}_i\right\rangle\\&\ge\frac{4C_1|a^{(t)}|^2\eta}{(1-\epsilon)n}\cdot\left\langle\mathbf{w}^{(t)}-\mathbf{w}^*,\sum_{i\in\mathrm{TP}}(\sigma(\mathbf{x}_i^{\mathrm{T}}\mathbf{w}^{(t)})-\sigma(\mathbf{x}_i^{\mathrm{T}}\mathbf{w}^*))\cdot\mathbf{x}_i\right\rangle\\&\quad+\frac{4\eta}{(1-\epsilon)n}\cdot\left\langle\mathbf{w}^{(t)}-\mathbf{w}^*,\sum_{i\in\mathrm{TP}}(a^{(t)}\sigma(\mathbf{x}_i^{\mathrm{T}}\mathbf{w}^*)-a^*\sigma(\mathbf{x}_i^{\mathrm{T}}\mathbf{w}^*))\cdot a^{(t)}\sigma'(\mathbf{x}_i^{\mathrm{T}}\mathbf{w})\cdot\mathbf{x}_i\right\rangle\end{aligned}\tag{13}$$

In the above, in the last relation we use the fact that σ is monotonically increasing. For the first term of Equation 13, we have from Lemma 30,

$$\begin{aligned}&\frac{4C_1|a^{(t)}|^2\eta}{(1-\epsilon)n}\cdot\left\langle\mathbf{w}^{(t)}-\mathbf{w}^*,\sum_{i\in\mathrm{TP}}(\sigma(\mathbf{x}_i^{\mathrm{T}}\mathbf{w}^{(t)})-\sigma(\mathbf{x}_i^{\mathrm{T}}\mathbf{w}^*))\cdot\mathbf{x}_i\right\rangle\\&\ge\frac{2C_1|a^{(t)}|^2\eta}{(1-\epsilon)n}\cdot\lambda_{\min}(\mathrm{X_{TP}}\mathrm{X_{TP}^T})\left\|\mathbf{w}^{(t)}-\mathbf{w}^*\right\|_2^2+\frac{2C_1|a^{(t)}|^2\eta}{\left\|\mathrm{X_{TP}}\mathrm{X_{TP}^T}\right\|_2(1-\epsilon)n}\cdot\left\|\sum_{i\in\mathrm{TP}}(\sigma(\mathbf{x}_i^{\mathrm{T}}\mathbf{w}^{(t)})-\sigma(\mathbf{x}_i^{\mathrm{T}}\mathbf{w}^*))\cdot\mathbf{x}_i\right\|_2^2\end{aligned}$$

We now upper bound the second term of Equation 13.

$$\frac{4\eta}{(1-\epsilon)n}\cdot\left\langle\mathbf{w}^{(t)}-\mathbf{w}^*,\sum_{i\in\mathrm{TP}}(a^{(t)}\sigma(\mathbf{x}_i^{\mathrm{T}}\mathbf{w}^*)-a^*\sigma(\mathbf{x}_i^{\mathrm{T}}\mathbf{w}^*))\cdot a^{(t)}\sigma'(\mathbf{x}_i^{\mathrm{T}}\mathbf{w})\cdot\mathbf{x}_i\right\rangle$$

$$\leq \frac{4|a^{(t)}|\eta}{(1-\epsilon)n} \cdot \left\|\mathbf{w}^{(t)} - \mathbf{w}^*\right\|_2 \Big\| \sum_{i \in \mathrm{TP}} (a^{(t)} - a^*) \cdot \sigma(\mathbf{x}_i^{\mathrm{T}} \mathbf{w}^*) \mathbf{x}_i \Big\|_2$$

$$\leq 2\eta^2 \cdot \left\|\mathbf{w}^{(t)} - \mathbf{w}^*\right\|_2^2 + \frac{C_2^2 |a^{(t)}|^2}{2\left[(1-\epsilon)n\right]^2} \cdot \left\|\mathrm{X}_{\mathrm{TP}} \mathrm{X}_{\mathrm{TP}}^{\mathrm{T}}\right\|_2^2 |a^{(t)} - a^*|^2$$

We now will upper bound the third term of Equation 12.

$$\eta^2 \cdot \left\|\nabla_{\mathbf{w}} \mathcal{R}(\mathbf{w}^{(t)}, a^{(t)}; \mathrm{TP})\right\|_2^2 = \frac{4\eta^2}{\left[(1-\epsilon)n\right]^2} \cdot \Big\| \sum_{i \in \mathrm{TP}} (a^{(t)} \sigma(\mathbf{x}_i^{\mathrm{T}} \mathbf{w}^{(t)}) - a^* \sigma(\mathbf{x}_i^{\mathrm{T}} \mathbf{w}^*)) \cdot a^{(t)} \sigma'(\mathbf{x}_i^{\mathrm{T}} \mathbf{w}^{(t)}) \cdot \mathbf{x}_i \Big\|^2$$

$$\leq \frac{8|a^{(t)}|^2 \eta^2}{\left[(1-\epsilon)n\right]^2} \cdot \Big\| \sum_{i \in \mathrm{TP}} (\sigma(\mathbf{x}_i^{\mathrm{T}} \mathbf{w}^{(t)}) - \sigma(\mathbf{x}_i^{\mathrm{T}} \mathbf{w}^*)) \cdot \sigma'(\mathbf{x}_i^{\mathrm{T}} \mathbf{w}^{(t)}) \mathbf{x}_i \Big\|^2$$

$$+ \frac{8\eta^2}{\left[(1-\epsilon)n\right]^2} \Big\| \sum_{i \in \mathrm{TP}} (a^{(t)} \sigma(\mathbf{x}_i^{\mathrm{T}} \mathbf{w}^*) - a^* \sigma(\mathbf{x}_i^{\mathrm{T}} \mathbf{w}^*)) \cdot a^{(t)} \sigma'(\mathbf{x}_i^{\mathrm{T}} \mathbf{w}^{(t)}) \cdot \mathbf{x}_i) \Big\|^2$$

$$\leq \frac{8C_2^2 |a^{(t)}|^2 \eta^2}{(1-\epsilon)n} \cdot \Big\| \sum_{i \in \mathrm{TP}} (\sigma(\mathbf{x}_i^{\mathrm{T}} \mathbf{w}^{(t)}) - \sigma(\mathbf{x}_i^{\mathrm{T}} \mathbf{w}^*)) \cdot \mathbf{x}_i \Big\|^2 + \frac{8C_2^2 |a^{(t)}|^2 \eta^2}{(1-\epsilon)n} \cdot \Big\| \sum_{i \in \mathrm{TP}} (a^{(t)} - a^*) \cdot \sigma(\mathbf{x}_i^{\mathrm{T}} \mathbf{w}^*) \mathbf{x}_i \Big\|_2^2$$

$$\leq \frac{8C_2^2 |a^{(t)}|^2 \eta^2}{(1-\epsilon)n} \cdot \Big\| \sum_{i \in \mathrm{TP}} (\sigma(\mathbf{x}_i^{\mathrm{T}} \mathbf{w}^{(t)}) - \sigma(\mathbf{x}_i^{\mathrm{T}} \mathbf{w}^*)) \cdot \mathbf{x}_i \Big\|^2 + \frac{8C_2^4 |a^{(t)}|^2 \eta^2}{(1-\epsilon)n} \cdot |a^{(t)} - a^*|^2 \left\|\mathbf{w}^*\right\|_2^2 \left\|\mathrm{X}_{\mathrm{TP}} \mathrm{X}_{\mathrm{TP}}^{\mathrm{T}}\right\|_2^2$$

In the above, in the first inequality we use the elementary inequality $(a+b)^2 \leq 2a^2 + 2b^2$, in the final inequality we note that σ is C_2-Lipschitz. We will now bound the second term in Equation 11.

$$\left\|\eta \nabla_{\mathbf{w}} \mathcal{R}(\mathbf{w}^{(t)}, a^{(t)}; \mathrm{FP})\right\|_2^2 = \frac{4\eta^2}{\left[(1-\epsilon)n\right]^2} \cdot \Big\| \sum_{i \in \mathrm{FP}} (a^{(t)} \sigma(\mathbf{x}_i^{\mathrm{T}} \mathbf{w}^{(t)}) - a^* \sigma(\mathbf{x}_i^{\mathrm{T}} \mathbf{w}^*)) \cdot a^{(t)} \sigma'(\mathbf{x}_i^{\mathrm{T}} \mathbf{w}) \cdot \mathbf{x}_i \Big\|_2^2$$

$$\leq \frac{4|a^{(t)}|^2 C_2^2 \eta^2}{(1-\epsilon)n} \cdot \left\|\mathrm{X}_{\mathrm{FP}} \mathrm{X}_{\mathrm{FP}}^{\mathrm{T}}\right\|_2 \sum_{i \in \mathrm{FP}} (a^{(t)} \sigma(\mathbf{x}_i^{\mathrm{T}} \mathbf{w}^{(t)}) - a^* \sigma(\mathbf{x}_i^{\mathrm{T}} \mathbf{w}^*))^2$$

$$\leq \frac{4|a^{(t)}|^2 C_2^2 \eta^2}{(1-\epsilon)n} \cdot \left\|\mathrm{X}_{\mathrm{FP}} \mathrm{X}_{\mathrm{FP}}^{\mathrm{T}}\right\|_2 \sum_{i \in \mathrm{FN}} (a^{(t)} \sigma(\mathbf{x}_i^{\mathrm{T}} \mathbf{w}^{(t)}) - a^* \sigma(\mathbf{x}_i^{\mathrm{T}} \mathbf{w}^*))^2$$

In the above, the second inequality follows from the optimality of the Subquantile set. Then, choosing $\eta \leq \frac{C_1}{4C_2 \|\mathrm{X}_{\mathrm{TP}} \mathrm{X}_{\mathrm{TP}}^{\mathrm{T}}\|}$, we obtain

$$\left\|\mathbf{w}^{(t)} - \mathbf{w}^* - \eta \nabla_{\mathbf{w}} \mathcal{R}(\mathbf{w}^{(t)}, a^{(t)}; \mathrm{TP})\right\|_2^2$$

$$\leq \left\|\mathbf{w}^{(t)} - \mathbf{w}^*\right\|_2^2 \big(1 - \eta \big(2\eta - \frac{2C_1 |a^{(t)}|^2}{(1-\epsilon)n} \cdot \lambda_{\min}(\mathrm{X}_{\mathrm{TP}} \mathrm{X}_{\mathrm{TP}}^{\mathrm{T}})\big)\big)$$

Step 2: Upper bounding the difference between $a^{(t)}$ and a^*. ■