Non-Linear 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 Adversarial Huber-ϵ Contamination Problem for Kernel Learning. We first reduce the Subquantile Minimization Algorithm to Iterative Thresholding using ideas from convex optimization. Let the target data be distributed as $y = \sum_{k=1}^{K} \sigma(\mathbf{x}^{\mathrm{T}}\mathbf{w}_k^*) + \xi_i$ for a non-linear function, σ, s.t. range$(\sigma') \in [C_1, C_2]$ for positive constants C_1, C_2 and noise $\xi_i \sim \mathcal{N}(0, \sigma^2)$ and $\mathbf{x}_i$ be sampled from a centered sub-Gaussian distribution with multivariate proxy, $\mathbf{\Gamma}$. Our main result is for sufficiently large n, there exists an algorithm that returns $\mathbf{W}^{(T)}$ with high probability such that $\|\mathbf{W}^* - \mathbf{W}^{(T)}\|_{\mathrm{F}} \leq \varepsilon$ in time,

$$O(\mathsf{polylog}(n, \|\mathbf{W}^*\|_{\mathrm{F}}, \kappa(\mathbf{\Gamma}), K, (1/C_1), C_2, (1/\varepsilon), 1-\epsilon))$$

for a sufficiently small ϵ dependent on $\mathbf{\Gamma}$, N, and $\|\mathbf{X}_{\mathrm{Q}}\|$. Furthermore, we consider the noisy Kernelized Generalized Linear Model (GLM) where $y = \omega(f^*(\mathbf{x})) + \xi$ where $\xi \sim \mathcal{N}(0, \sigma^2)$ and prove the same algorithm returns $f^{(T)}$ with high probability such that $\|f^{(T)} - f^*\| \leq \varepsilon + O(\sigma)$ after T iterations, The iterative thresholding algorithm has been used in large neural network models in prior research [SS19a]. Our work provides the first steps for theoretical guarantees for neural networks and non-linear models for iterative thresholding algorithms in the Huber-ϵ contamination model.

1 Introduction

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

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

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

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

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

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

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

More recently, Awasthi et al. [ADKS22] studied the iterative trimmed maximum likelihood estimator. We will give their formal theorem result.

Theorem 4 (Theorem 4.2 in [ADKS22])**.** *Let $\mathrm{P} = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$ be the data generated by a Gaussian regression model defined as $y_i = \mathbf{x}_i^{\mathrm{T}}\mathbf{w}^* + \eta_i$ where $\eta_i \sim \mathcal{N}(0, \sigma^2)$ and $\mathbf{x}_i$ are sampled from a sub-Gaussian Distribution with second-moment matrix $\mathbf{I}$. Suppose the dataset has ϵ-fraction of label corruption and $n = \Omega\left(\frac{d+\log(1/\delta)}{\epsilon^2}\right)$. Then there exists an algorithm that returns $\widehat{\mathbf{w}}$ such that with probability $1-\delta$,*

$$\|\widehat{\mathbf{w}} - \mathbf{w}^*\|_2 = O(\sigma\epsilon\log(1/\epsilon))$$

Our first result recovers this result for vectorized-regression. We will now give our results for the Kernelized GLM problem.

1.1 Contributions

Our main contribution is the approximation bounds for Subquantile Minimization for various non-linear learning problems from the iterative thresholding algorithm given in Algorithm 1. Our proof techniques

extend [BJK15, SS19b, ADKS22] as we suppose the adversary also corrupts the covariates. To our knowledge, we are also the first to theoretically study iterative thresholding for non-linear learning algorithms beyond the generalized linear model.

Reference	Approximation	Runtime	Model
[ADKS22]	$O(\sigma\epsilon\log\epsilon^{-1})$	$O(\frac{1}{\epsilon^2}(Nd^2+d^3))$	Regression; Label
Corollary 13	$O(\sigma\epsilon\log\epsilon^{-1})$	$O\left(Nd\log\left(\frac{\|\mathbf{w}^*\|}{\epsilon^2}\right)\right)$	Regression; Label
Theorem 12	$O(\sigma\epsilon\sqrt{C_Q K\log\epsilon^{-1}})$	$O\left(NKd\log\left(\frac{\|\mathbf{W}^*\|_{\mathrm{F}}}{\epsilon^2}\right)\right)$	K-variate Regression
Theorem 16	$O\left(\sigma\kappa^2(\mathbf{\Sigma})C_\psi^{-1}C_K d\epsilon\log\epsilon^{-1}\right)$	$O\left(Nd\log\left(\frac{\|\mathbf{w}^*\|_2+(d/\delta)}{\epsilon}\right)\right)$	Leaky ReLU Neuron
Theorem 16	$O\left(\sigma\kappa^2(\mathbf{\Sigma})C_\psi^{-1}C_K d\epsilon\log\epsilon^{-1}\right)$	$O\left(Nd\log\left(\frac{\|\mathbf{w}^*\|_2+(d/\delta)}{\epsilon}\right)\right)$	Leaky ReLU Neuron

Table 1: Summary of related work on Iterative Thresholding Algorithms for Learning in the Huber-ϵ Contamination Model and our contributions. We assume the good data is sampled from a sub-Gaussian distribution with second-moment matrix, $\mathbf{\Sigma}$, and sub-Gaussian norm C_K and dimension d. We assume the variance of the optimal estimator is σ. The Leaky-ReLU function is given as $\max\{C_\psi x, x\}$.

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 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 $\{\mathbf{x} \in \mathbb{R}^d : \|\mathbf{x}\| = 1\}$. We will typically denote capital roman letters A, B, C as fixed constants, lower-case roman letters f, g, h as functions, bold-face lower-case letters $\mathbf{x}, \mathbf{y}, \mathbf{z}$ as vectors, and bold-face, upper-case letters $\mathbf{P}, \mathbf{Q}, \mathbf{R}$ as matrices. Throughout the paper, we will use the following matrix norms for a matrix $\mathbf{A} \in \mathbb{R}^{m\times n}$ and vector $\mathbf{x} \in \mathbb{R}^n$,

$$\text{Spectral Norm: } \|\mathbf{A}\| = \max_{\mathbf{x}\in\mathbb{S}^{m-1}} \|\mathbf{A}\mathbf{x}\| = \sigma_1(\mathbf{A})$$

$$\text{Trace Norm: } \mathrm{Tr}(\mathbf{A}) = \sum_{i\in[m\wedge n]} \sigma_i(\mathbf{A})$$

$$\text{Frobenius Norm: } \|\mathbf{A}\|_{\mathrm{F}}^2 = \mathrm{Tr}(\mathbf{A}^{\mathrm{T}}\mathbf{A}) = \sum_{i\in[m\wedge n]} \sigma_i^2(\mathbf{A})$$

$$\ell_2 \text{ Norm: } \|\mathbf{x}\|_2^2 = \mathbf{x}^{\mathrm{T}}\mathbf{x} = \sum_{i\in[n]} x_i^2$$

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 $f \in \mathcal{H}$,*

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

Definition 6 (Pushforward Measure)**.** *Let $\phi : \mathcal{X} \to \mathcal{H}$ represent the mapping from the input dimension to the Hilbert Space, and let $\mathcal{P}$ be the Probability Measure of the uncorrupted data over $\mathcal{X}$. Then $\mathcal{P}_\sharp\phi(X) = \mathcal{P}(\phi^{-1}(X))$ represents the measure over the Hilbert Space $\mathcal{H}$ using the measure of the good data defined over the original data space $\mathcal{X}$.*

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

$$\mathop{\mathbb{E}}_{X \sim \mathcal{P}} \left[\exp\left(\theta \langle X, v \rangle_{\mathcal{H}} \right) \right] \leq \exp\left(\frac{\alpha^2 \theta^2 \langle v, \mathbf{\Gamma} v \rangle_{\mathcal{H}}}{2} \right) \tag{2.1}$$

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

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

Then we say $X \sim \mathcal{SG}(\mathbf{\Gamma}, \alpha)$, where if $\alpha = 1$, we will say $X \sim \mathcal{SG}(\mathbf{\Gamma})$. The Gaussian Design for the Feature Space has gained popularity in the study of kernel learning [CLKZ21]. 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 7 (Sub-Gaussian Design)**.** *We assume for $\mathbf{x}_i \sim \mathcal{X}$, then it follows for the function to the Hilbert Space, $\phi(\cdot) : \mathcal{X} \to \mathcal{H}$,*

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

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

Assumption 8 (Bounded Functions)**.** *We assume for $\mathbf{x}_i \sim \mathcal{P} \in \mathcal{X}$, then it follows for the feature map, $\phi(\cdot) : \mathcal{X} \to \mathcal{H}$,*

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

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

Assumption 9 (Normal Residuals)**.** *Let $\inf_{f \in \mathcal{H}} \mathbb{E}_{(\mathbf{x}, y) \sim \mathcal{D}} \left[\mathcal{R}(f; \mathbf{x}, y) \right]$. The residual is defined as $\xi_i \triangleq f^*(\mathbf{x}_i) - y_i$. Then we assume for some $\sigma > 0$, it follows*

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

2.5 Related Work

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

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

sampled from a multivariate sub-Gaussian Distribution with proxy $\mathbf{I}$ before running the iterative thresholding procedure. This '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 $\mathbf{K}$ which is equivalent to inner product of the quasimatrix[1], $\mathbf{\Phi}$, 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 $\mathbf{\Phi}$ where the columns are sampled from a sub-Gaussian Distribution with proxy $\mathbf{\Gamma}$ is a trace-class operator, which implies the eigenvalues tend to zero, i.e. $\lambda_{\inf}(\mathbf{\Gamma}) = 0$, and there is no longer a notion of $\lambda_{\min}(\mathbf{\Gamma})$.

3 Subquantile Minimization

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

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

The p-subquantile of the risk is then given by

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

Given a function to minimize, $\mathcal{L}$, the variational problem is given as:

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

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

$$t^{(k+1)} = \arg\max_{t\in\mathbb{R}} g(f^{(k)}, t) \tag{3.1}$$

$$f^{(k+1)} = \text{Proj}_{\mathcal{K}}\left[f^{(k)} - \eta \nabla_f g(f^{(k)}, t^{(k+1)}) \right] \tag{3.2}$$

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

3.1 Reduction to Iterative Thresholding

To consider theoretical guarantees of Subquantile Minimization, we first analyze the inner and outer optimization problems. We first analyze kernel learning in the presence of corrupted data. Next, we provide error bounds for the two most important kernel learning problems, kernel ridge regression, and kernel classification. Now we will give our first result regarding kernel learning in the Huber ϵ-contamination model. Now we will analyze the two-step minimax optimization steps described in Equations (3.1) and (3.2).

Lemma 10. *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 $\hat{\nu}_{(1-\epsilon)N} \in \arg\max_{t\in\mathbb{R}} g(t,f)$.*

[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]).

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.

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

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

From Lemma 10, we see that t will be greater than or equal to the errors of exactly $(1-\epsilon)N$ points. Thus, we are continuously updating over the $(1-\epsilon)N$ minimum errors.

Lemma 11. *Let $\hat{\nu}_i \triangleq \mathcal{R}(f; \mathbf{x}_{\lfloor i \rfloor}, y_{\lfloor i \rfloor})$, if we choose $t^{(k+1)} = \hat{\nu}_{(1-\epsilon)N}$ as by Lemma 10, it then follows $\nabla_f g(t^{(k)}, f^{(k)}) = \frac{1}{(1-\epsilon)N}\sum_{i=1}^{(1-\epsilon)N}\nabla_f \mathcal{R}(f^{(k)}; \mathbf{x}_{\lfloor i \rfloor}, y_{\lfloor i \rfloor})$.*

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

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

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

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

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. We will start with the simple case of vectorized regression as a warm-up to our general proof technique. We then move to the GLM with kernel learning. Finally, we give our results for one-hidden layer neural networks. We will now give the algorithm for Subquantile Minimization with Gradient Descent.

4.1 Algorithm

4.2 Warm-up: Multivariate Linear Regression

We will first present our results for the well-studied problem of linear regression in the Huber-ϵ contamination model. Our results will extend the results in [BJKK17] Theorem 5 and [ADKS22] Lemma A.1 by including covariate corruption, variance in the optimal estimator, and non-identity second-moment matrix of the uncorrupted data. The loss function for the multivariate linear regression problem for $\mathbf{W} \in \mathbb{R}^{k\times d}$, $\mathbf{X} \in \mathbb{R}^{d\times n}$, and $\mathbf{Y} \in \mathbb{R}^{k\times n}$.

$$\mathcal{L}(\mathbf{W}; \mathbf{X}, \mathbf{Y}) = \left\|\mathbf{W}\mathbf{X} - \mathbf{Y}\right\|_{\mathrm{F}}^2$$

Algorithm 1 Subquantile Minimization for One Hidden-Layer Neural Networks

input: Possibly corrupted $\mathbf{X} \in \mathbb{R}^{d\times N}$ with outputs $\mathbf{y} \in \mathbb{R}^N$ and corruption parameter $\epsilon = O(\mathsf{poly}(C_3, \mathbf{\Gamma}))$.
output: Approximate solution $\mathbf{W} \in \mathbb{R}^{k\times d}$ to $\min\|\mathbf{W} - \mathbf{W}^*\|_{\mathrm{F}}$.

$\mathbf{W}^{(0)} \leftarrow \mathbf{0}$
for $t \in [T]$ **do**
  $\mathrm{S}^{(t)} \leftarrow \{i \in [n] : \mathcal{L}(\mathbf{W}^{(t)}; \mathbf{x}_i, y_i) \le \mathcal{L}(\mathbf{W}^{(t)}; \mathbf{x}_{\lfloor(1-\epsilon)N\rfloor}; y_{\lfloor(1-\epsilon)N\rfloor})\}$
  $\nabla\mathcal{R}(\mathbf{W}; \mathrm{S}^{(t)}) \leftarrow \frac{2}{(1-\epsilon)N} \cdot \sum_{i\in\mathrm{S}^{(t)}} (\sum_{k\in[K]} \sigma(\mathbf{x}_i^{\mathrm{T}}\mathbf{w}_k) - y_i) \cdot (\sigma \circ (\mathbf{W}^{(t)}\mathbf{x}_i)) \cdot \mathbf{x}_i^{\mathrm{T}}$
  $\mathbf{W}^{(t+1)} \leftarrow \mathbf{W}^{(t)} - \eta\nabla\mathcal{R}(\mathbf{W}^{(t)}; \mathrm{S}^{(t)})$
return: $\mathbf{W}^{(T)}$

Theorem 12. *Let $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n]^{\mathrm{T}} \in \mathbb{R}^{d\times N}$ be the data matrix and $\mathbf{Y} = [\mathbf{y}_1, \ldots, \mathbf{y}_n] \in \mathbb{R}^{K\times N}$ be the output, such that for $i \in P$, $\mathbf{x}_i$ are sampled from a sub-Gaussian distribution with sub-Gaussian norm L and second-moment matrix $\mathbf{\Sigma}$. Suppose for $i \in P$ the output is given as $\mathbf{y}_i = \mathbf{W}\mathbf{x}_i + \mathbf{e}_i$ where for $j \in [K]$, $[\mathbf{e}_i]_j \sim \mathcal{N}(0, \sigma^2)$. Then after $O\left(\kappa(\mathbf{\Sigma}) \log\left(\frac{\|\mathbf{W}^*\|_{\mathrm{F}}}{\varepsilon}\right)\right)$ gradient descent iterations, $N \ge \frac{(1/\delta)K\,\mathrm{Tr}(\mathbf{\Sigma})}{1600\epsilon^2 \log\epsilon^{-1}\lambda_{\max}(\mathbf{\Sigma})}$, and learning rate $\eta = 0.1\lambda_{\max}^{-2}(\mathbf{\Sigma})$, with probability exceeding $1-\delta$, Algorithm 1 returns $\mathbf{W}^{(T)}$ such that*

$$\left\|\mathbf{W}^{(T)} - \mathbf{W}^*\right\|_{\mathrm{F}} \le \varepsilon + O\left(\sigma\epsilon\sqrt{KC_3 \log\epsilon^{-1}}\right)$$

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

We are able to recover the result of Lemma 4.2 in [ADKS22] when $K = 1$ and the covariates (corrupted and un-corrupted) are sampled from a sub-Gaussian distribution with second-moment matrix $\mathbf{I}$. The full solve algorithm developed in [ADKS22] returns a $O(\sigma\epsilon\log\epsilon^{-1})$ in time $O\left(\frac{1}{\epsilon^2}\left(Nd^2 + d^3\right)\right)$, with the gradient descent based approach, we are able to improve the runtime to $O\left(\log\left(\frac{\|\mathbf{w}^*\|}{\sigma\epsilon^2}\right)Nd^2\right)$ for the same approximation bound. We will formalize our results into a corollary to give a more representative comparison in the literature.

Corollary 13. *Let $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n]^{\mathrm{T}} \in \mathbb{R}^{d\times N}$ be the data matrix and $\mathbf{y} = [y_1, \ldots, y_n] \in \mathbb{R}^N$ be the output, such that for $i \in P$, $\mathbf{x}_i$ are sampled from a sub-Gaussian distribution with sub-Gaussian norm L and second-moment matrix $\mathbf{I}$. Suppose for $i \in P$ the output is given as $\mathbf{y}_i = \mathbf{x}_i^{\mathrm{T}}\mathbf{w}^* + \xi_i$ where $\xi_i \sim \mathcal{N}(0, \sigma^2)$. Then after $O\left(\log\left(\frac{\|\mathbf{w}^*\|_2}{\varepsilon}\right)\right)$ gradient descent iterations, $N \ge \frac{(d/\delta)}{800\epsilon^2}$, and learning rate $\eta = 0.1$, with probability exceeding $1-\delta$, Algorithm 1 returns $\mathbf{w}^{(T)}$ such that*

$$\|\mathbf{w}^{(T)} - \mathbf{w}^*\|_2 \le \varepsilon + O\left(\sigma\epsilon\sqrt{KC_3\log\epsilon^{-1}}\right)$$

Suppose for $i \in Q$, $\mathbf{x}_i$ are sampled from a sub-Gaussian distribution with sub-Gaussian Norm K and second-moment matrix $\mathbf{I}$. Then,

$$\|\mathbf{w}^{(T)} - \mathbf{w}^*\|_2 \le \varepsilon + O\left(\sigma\epsilon\log\epsilon^{-1}\right)$$

4.3 Learning Generalized Linear Models

We next study the problem of learning GLMs following the model given in § 5 of [SS19b]. The error function for the the Kernelized GLM problem is given by the following equation for a single training pair $(\mathbf{x}_i, y_i) \sim \mathcal{D}$ in the kernelized case.

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

Theorem 14. *Let $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n]^{\mathrm{T}} \in \mathbb{R}^{d\times N}$ be the data matrix and $\mathbf{y} = [y_1, \ldots, y_n]$ be the output, such that for $i \in P$, $\mathbf{x}_i$ are sampled from a sub-Gaussian distribution s.t. $\phi(\mathbf{x}_i) = X_i$ has sub-Gaussian norm L and second-moment matrix $\mathbf{\Sigma}$, and $y_i = \omega(\mathbf{x}_i^{\mathrm{T}}\mathbf{w}^* + \xi_i)$ for ξ_i sampled from a centered sub-Gaussian distribution with sub-Gaussian norm ν. Suppose the link function $\omega : \mathbb{R} \to \mathbb{R}$, has bounded gradient s.t. $C_1 \le \omega'(x) \le C_2$ for absolute constants $C_1, C_2 > 0$ for all $x \in \mathbb{R}$. Then after $O\left(\kappa(\mathbf{\Sigma})\log\left(\frac{\|f^*\|_{\mathcal{H}}}{\varepsilon}\right)\right)$ gradient descent iterations, then with probability exceeding $1-\delta$,*

$$\|f^{(T)} - f^*\|_{\mathcal{H}} \le \varepsilon + O\left(C_1^{-1}C_2^2 \cdot \epsilon\sqrt{\lambda_{\max}(\mathbf{\Sigma})C_3\log\epsilon^{-1}}\right) + O\left(\lambda_{\max}(\mathbf{\Sigma})\lambda_{\min}^{-2}(\mathbf{\Sigma}) \cdot \sigma \cdot \left(C_1^{-2}C_2^2\right)\right)$$

when $N \geq \frac{1}{\lambda^2_{\min}(\mathbf{\Sigma})} \cdot \Big(8C_K \cdot d + \frac{2}{c_K} \cdot \log(2/\delta)\Big)$.

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

Theorem 15. *Let* $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n]^{\mathrm{T}} \in \mathbb{R}^{d \times N}$ *be the data matrix and* $\mathbf{y} = [y_1, \ldots, y_n]$ *be the output, such that for* $i \in P$, $\mathbf{x}_i$ *are sampled from a sub-Gaussian distribution s.t.* $\phi(\mathbf{x}_i) = X_i$ *has sub-Gaussian norm* L *and second-moment matrix* $\mathbf{\Sigma}$, *and* $y_i = \omega(f^*(\mathbf{x}_i)) + \xi_i$ *for* ξ_i *sampled from a centered sub-Gaussian distribution with sub-Gaussian norm* ν. *Suppose the link function* $\omega : \mathbb{R} \to \mathbb{R}$, *has bounded gradient s.t.* $C_1 \leq \omega'(x) \leq C_2$ *for absolute constants* $C_1, C_2 > 0$ *for all* $x \in \mathbb{R}$. *Then after* $O\left(\kappa(\mathbf{\Sigma}) \log\left(\frac{\|f^*\|_{\mathcal{H}}}{\varepsilon}\right)\right)$ *gradient descent iterations, then with probability exceeding* $1 - \delta$,

$$\|f^{(T)} - f^*\|_{\mathcal{H}} \leq \varepsilon + O\Big(C_1^{-1} C_2^2 \cdot \epsilon \sqrt{\lambda_{\max}(\mathbf{\Sigma}) C_3 \log \epsilon^{-1}}\Big) + O\big(\lambda_{\max}(\mathbf{\Sigma}) \lambda^{-2}_{\min}(\mathbf{\Sigma}) \cdot \sigma \cdot \big(C_1^{-2} C_2^2\big)\big)$$

when $N \geq \frac{1}{\lambda^2_{\min}(\mathbf{\Sigma})} \cdot \Big(8C_K \cdot d + \frac{2}{c_K} \cdot \log(2/\delta)\Big)$.

Proof. The proof is deferred to § B.1. ■
When considerin the linear regression case

4.4 Learning Leaky-ReLU 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. We start with the simple case of a linear regression for multi-variate outputs.

Theorem 16. *Let* $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n]^{\mathrm{T}} \in \mathbb{R}^{N \times d}$ *be the data matrix and* $\mathbf{y} = [y_1 \ldots, y_n]^{\mathrm{T}}$ *be the output, such that for* $i \in P$, $\mathbf{x}_i \sim \mathcal{P}$ *are sampled from a sub-Gaussian distribution with sub-Gaussian norm* K *and second-moment matrix* $\mathbf{\Sigma}$, *and* $y_i = \psi(\mathbf{x}^{\mathrm{T}} \mathbf{w}^*) + \xi_i$ *for* $\xi_i \sim \mathcal{N}(0, \sigma^2)$ *where* $\psi(x) = \max\{C_\psi x, x\}$. *Then after* $O\left(C_\psi^{-2} \kappa(\mathbf{\Sigma}) \log\left(\frac{\|\mathbf{W}^*\|_{\mathrm{F}}}{\varepsilon}\right)\right)$ *gradient descent iterations and* $\epsilon \leq \frac{C_\psi^2 \lambda_{\min}(\mathbf{\Sigma})}{\sqrt{32 C_Q \lambda_{\max}(\mathbf{\Sigma})}}$, *with probability exceeding* $1 - \delta$, *Algorithm 1 with learning rate* $\eta = O(\kappa^{-2}(\mathbf{\Sigma}))$ *returns* $\mathbf{w}^{(T)}$ *such that*

$$\|\mathbf{w}^{(T)} - \mathbf{w}^*\|_2 \leq \varepsilon + O\Big(\kappa^2(\mathbf{\Sigma}) C_\psi^{-1} C_K C_\sigma d \epsilon \log \epsilon^{-1}\Big)$$

Proof. The proof is deferred to § C.1.1. ■

Theorem 17. *Let* $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n]^{\mathrm{T}} \in \mathbb{R}^{N \times d}$ *be the data matrix and* $\mathbf{y} = [y_1 \ldots, y_n]^{\mathrm{T}}$ *be the output, such that for* $i \in P$, $\mathbf{x}_i \sim \mathcal{P}$ *are sampled from a sub-Gaussian distribution with sub-Gaussian norm* K *and second-moment matrix* $\mathbf{\Sigma}$, *and* $y_i = \sum_{k \in [K]} \psi(\mathbf{x}^{\mathrm{T}} \mathbf{w}_k^*) + \xi_i$ *for* $\xi_i \sim \mathcal{N}(0, \sigma^2)$ *where* $\psi(x) = \max\{C_\psi x, x\}$. *Then after* $O(\Xi)$ *gradient descent iterations and* $\epsilon \leq O(\Xi)$, *with probability exceeding* $1 - \delta$, *Algorithm 1 with learning rate* $\eta = O(\Xi)$ *returns* $\mathbf{W}^{(T)}$ *such that*

$$\|\mathbf{W}^{(T)} - \mathbf{W}^*\|_{\mathrm{F}} \leq \varepsilon + O(\Xi)$$

Proof. The proof is deferred to § C.1.1. ■

4.5 Learning ReLU Neural Networks

We will now consider the problem of learning ReLU neural networks. We will start with the simple case of learning a single neuron, before moving on to the K-hidden neuron case.

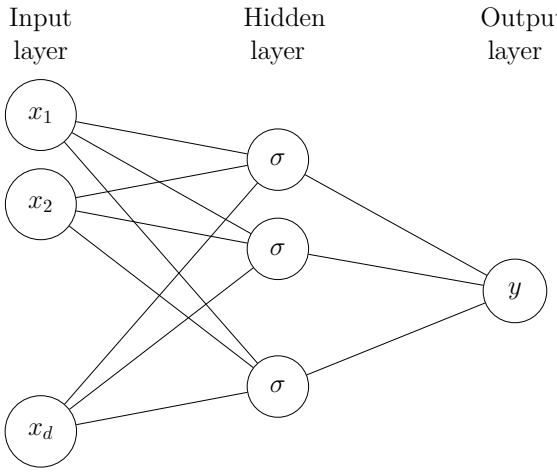

Figure 1: One Hidden Layer Neural Network.

4.5.1 Single Neuron

Theorem 18. *Let* $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n]^{\mathrm{T}} \in \mathbb{R}^{n \times d}$ *be the data matrix and* $\mathbf{y} = [y_1, \ldots, y_n]^{\mathrm{T}}$ *be the output, such that for* $i \in \mathrm{P}$, $\mathbf{x}_i$ *are sampled from a sub-Gaussian distribution with sub-Gaussian norm* K *and second-moment matrix* $\mathbf{\Sigma}$. *Suppose for* $i \in \mathrm{P}$, *the output is given as* $y_i = \psi(\mathbf{x}^{\mathrm{T}}\mathbf{w}^*) + \xi_i$ *for* $\xi_i \sim \mathcal{N}(0, \sigma^2)$ *and* $\psi = \max\{0, x\}$ *is the ReLU function. Then after* $O(\Xi))$ *gradient descent iterations and* $n = \Omega(\Xi)$, *then with probability exceeding* $1 - \delta$, *Algorithm 1 with learning rate* $\eta = O(\Xi)$ *returns* $\mathbf{w}^{(T)}$ *such that* $\|\mathbf{w}^{(T)} - \mathbf{w}^*\|_2 \leq \varepsilon + O(\Xi)$.

Proof. The proof is deferred to § C.2 ■

5 Discussion

In this paper, we study the theoretical convergence properties of iterative thresholding for non-linear learning problems in the Huber-ϵ contamination model. Our warm-up result for linear regression reduces the runtime while achieving the same approximation bound. Many papers have experimentally studied the iterative thresholding estimator in large scale neural networks. We are the first paper to give statistical learning guarantees in the two-layer case with the ReLU activation function.

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, 1, 4, 1.1, ??, 2.3, 2.5, 4.2, 4.2

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

[B+15] Sébastien Bubeck et al. Convex optimization: Algorithms and complexity. *Foundations and Trends® in Machine Learning*, 8(3-4):231–357, 2015. 32

[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. 3, 1.1, 2.5

[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, 2.5, 4.2

[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, 2.3

[CLRS22] Thomas H Cormen, Charles E Leiserson, Ronald L Rivest, and Clifford Stein. *Introduction to algorithms*. MIT press, 2022. 38

[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

[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

[HMT11] Nathan Halko, Per-Gunnar Martinsson, and Joel A Tropp. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM review*, 53(2):217–288, 2011. 34

[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. 2.5

[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

[Jen06] Johan Ludwig William Valdemar Jensen. Sur les fonctions convexes et les inégalités entre les valeurs moyennes. *Acta mathematica*, 30(1):175–193, 1906. 26

[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. 2.3

[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. 3

[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, 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. 2.5

[LM00] Beatrice Laurent and Pascal Massart. Adaptive estimation of a quadratic functional by model selection. *Annals of statistics*, pages 1302–1338, 2000. 21

[M+89] Colin McDiarmid et al. On the method of bounded differences. *Surveys in combinatorics*, 141(1):148–188, 1989. 27

[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, 2.5

[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

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

[RaR02] Raymond A Ryan and R a Ryan. *Introduction to tensor products of Banach spaces*, volume 73. Springer, 2002. 2.2

[RH23] Philippe Rigollet and Jan-Christian Hütter. High-dimensional statistics. *arXiv preprint arXiv:2310.19244*, 2023. C.2

[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. 3

[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. 2.5

[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. (document)

[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. 1.1, 4.3

[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. 28, D.2

[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. 1

[Ver10] Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*, 2010. 23

[Ver20] Roman Vershynin. High-dimensional probability. *University of California, Irvine*, 2020. C.2

[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. 1

[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. 33

[ZYWG19] Xiao Zhang, Yaodong Yu, Lingxiao Wang, and Quanquan Gu. Learning one-hidden-layer relu networks via gradient descent. In *The 22nd international conference on artificial intelligence and statistics*, pages 1524–1534. PMLR, 2019. C.2, C.2

A Proofs for Linear Regression

Notation. We will first give some notational preliminaries. Let $\mathrm{X} = \mathrm{P} \cup \mathrm{Q}$ for $|\mathrm{P}| = (1-\epsilon)N$ and $|\mathrm{Q}| = \epsilon N$ represent the sets such that for $i \in \mathrm{P}$, $(\mathbf{x}_i, y_i)$ is the good data and for $j \in \mathrm{Q}$, $(\mathbf{x}_j, y_j)$ has been given by the adversary. For $t \in [T]$, we denote $\mathrm{S}^{(t)}$ as the Subquantile set at iteration t and represents the points. We decompose $\mathrm{S}^{(t)} = \mathrm{S}^{(t)} \cap \mathrm{P} \cup \mathrm{S}^{(t)} \cap \mathrm{Q} = \mathrm{TP} \cup \mathrm{FP}$ to represent the *True Positives* and *False Positives*. We also decompose $\mathrm{X} \setminus \mathrm{S}^{(t)} = (\mathrm{X} \setminus \mathrm{S}^{(t)}) \cap \mathrm{P} \cup (\mathrm{X} \setminus \mathrm{S}^{(t)}) \cap \mathrm{Q} = \mathrm{FN} \cup \mathrm{TN}$ to represent the *False Negatives* and the *True Negatives*.

A.1 Proof of Theorem 12

Proof. Recall that for any $\mathbf{W} \in \mathbb{R}^{k \times d}$, $\mathbf{X} \in \mathbb{R}^{d \times n}$, and $\mathbf{Y} \in \mathbb{R}^{k \times n}$,

$$\begin{aligned}\mathcal{L}(\mathbf{W}; \mathbf{X}, \mathbf{Y}) &= \left\|\mathbf{W}\mathbf{X} - \mathbf{Y}\right\|_{\mathrm{F}}^2 = \mathrm{Tr}(\mathbf{X}^{\mathrm{T}}\mathbf{W}^{\mathrm{T}}\mathbf{W}\mathbf{X} - \mathbf{X}^{\mathrm{T}}\mathbf{W}^{\mathrm{T}}\mathbf{Y} - \mathbf{Y}^{\mathrm{T}}\mathbf{W}\mathbf{X} + \mathbf{Y}^{\mathrm{T}}\mathbf{Y}) \\ &= \mathrm{Tr}(\mathbf{X}^{\mathrm{T}}\mathbf{W}^{\mathrm{T}}\mathbf{W}\mathbf{X}) + \mathrm{Tr}(\mathbf{Y}^{\mathrm{T}}\mathbf{Y}) - 2\,\mathrm{Tr}(\mathbf{X}^{\mathrm{T}}\mathbf{W}^{\mathrm{T}}\mathbf{Y})\end{aligned}$$

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

$$\nabla\mathcal{L}(\mathbf{W}) = 2(\mathbf{W}\mathbf{X} - \mathbf{Y})\mathbf{X}^{\mathrm{T}}$$

Step 1: Linear Convergence. We will first show iterative thresholding has linear convergence to optimal. Let $\widehat{\mathbf{W}} = \arg\min_{\mathbf{W} \in \mathbb{R}^{K \times d}} \mathcal{R}(\mathbf{W}; \mathrm{P})$ be the minimizer over the good data. Then, we have

$$\begin{aligned}&\left\|\mathbf{W}^{(t+1)} - \mathbf{W}^*\right\|_{\mathrm{F}} = \left\|\mathbf{W}^{(t)} - \mathbf{W}^* - \eta\nabla\mathcal{R}(\mathbf{W}^{(t)}; \mathrm{S}^{(t)})\right\|_{\mathrm{F}} \\ &= \left\|\mathbf{W}^{(t)} - \mathbf{W}^* - \eta\nabla\mathcal{R}(\mathbf{W}^{(t)}; \mathrm{TP}) + \eta\nabla\mathcal{R}(\mathbf{W}^*; \mathrm{P}) - \eta\nabla\mathcal{R}(\mathbf{W}^*; \mathrm{P}) + \eta\nabla\mathcal{R}(\widehat{\mathbf{W}}; \mathrm{P}) - \eta\nabla\mathcal{R}(\mathbf{W}^{(t)}; \mathrm{FP})\right\|_{\mathrm{F}} \\ &\le \left\|\mathbf{W}^{(t)} - \mathbf{W}^* + \eta\nabla\mathcal{R}(\mathbf{W}^{(t)}; \mathrm{TP}) + \eta\nabla\mathcal{R}(\mathbf{W}^*; \mathrm{TP})\right\|_{\mathrm{F}} + \left\|\eta\nabla\mathcal{R}(\mathbf{W}^{(t)}; \mathrm{FP})\right\|_{\mathrm{F}} + \left\|\eta\nabla\mathcal{R}(\mathbf{W}^*; \mathrm{FN})\right\|_{\mathrm{F}} \\ &\quad + \left\|\eta\nabla\mathcal{R}(\mathbf{W}^*; \mathrm{P}) - \eta\nabla\mathcal{R}(\widehat{\mathbf{W}}; \mathrm{P})\right\|_{\mathrm{F}} \qquad (A.1)\end{aligned}$$

We first will upper bound the first term in Equation (A.1) through its square.

$$\begin{aligned}&\left\|\mathbf{W}^{(t)} - \mathbf{W}^* - \eta\nabla\mathcal{R}(\mathbf{W}^{(t)}; \mathrm{TP}) + \eta\nabla\mathcal{R}(\mathbf{W}^*; \mathrm{TP})\right\|_{\mathrm{F}}^2 = \left\|\mathbf{W}^{(t)} - \mathbf{W}^*\right\|_{\mathrm{F}}^2 \\ &\quad - 2\eta \cdot \mathrm{Tr}\big((\mathbf{W}^{(t)} - \mathbf{W}^*)^{\mathrm{T}}(\nabla\mathcal{R}(\mathbf{W}^{(t)}; \mathrm{TP}) - \nabla\mathcal{R}(\mathbf{W}^*; \mathrm{TP}))\big) + \left\|\eta\nabla\mathcal{R}(\mathbf{W}^{(t)}; \mathrm{TP}) - \eta\nabla\mathcal{R}(\mathbf{W}^{(t)}; \mathrm{TP})\right\|_{\mathrm{F}}^2 \qquad (A.2)\end{aligned}$$

We then lower bound the second term in Equation (A.2),

$$\begin{aligned}&2\eta \cdot \mathrm{Tr}\big((\mathbf{W}^{(t)} - \mathbf{W}^*)^{\mathrm{T}}(\nabla\mathcal{R}(\mathbf{W}^{(t)}; \mathrm{TP}) - \nabla\mathcal{R}(\mathbf{W}^*; \mathrm{TP}))\big) \\ &\quad \stackrel{\text{def}}{=} \frac{4\eta}{(1-\epsilon)N} \cdot \mathrm{Tr}\big((\mathbf{W}^{(t)} - \mathbf{W}^*)^{\mathrm{T}}\big((\mathbf{W}^{(t)}\mathbf{X}_{\mathrm{TP}} - \mathbf{Y}_{\mathrm{TP}})\mathbf{X}_{\mathrm{TP}}^{\mathrm{T}} - \mathbf{E}_{\mathrm{TP}}\mathbf{X}_{\mathrm{TP}}^{\mathrm{T}}\big)\big) \\ &\quad = \frac{4\eta}{(1-\epsilon)N} \cdot \mathrm{Tr}\Big((\mathbf{W}^{(t)} - \mathbf{W}^*)^{\mathrm{T}}(\mathbf{W}^{(t)} - \mathbf{W}^*)\mathbf{X}_{\mathrm{TP}}\mathbf{X}_{\mathrm{TP}}^{\mathrm{T}}\Big) \qquad (A.3)\end{aligned}$$

In the above, the second equation follows from noting that $\mathbf{Y}_{\mathrm{TP}} = \mathbf{W}^*\mathbf{X}_{\mathrm{TP}} - \mathbf{E}_{\mathrm{TP}}$. We can now lower bound the term in Equation (A.3).

$$\begin{aligned}&\frac{4\eta}{(1-\epsilon)N} \cdot \mathrm{Tr}\Big((\mathbf{W}^{(t)} - \mathbf{W}^*)^{\mathrm{T}}(\mathbf{W}^{(t)} - \mathbf{W}^*)\mathbf{X}_{\mathrm{TP}}\mathbf{X}_{\mathrm{TP}}^{\mathrm{T}}\Big) \\ &= \frac{4\eta}{(1-\epsilon)N} \cdot \mathrm{Tr}\Big((\mathbf{W}^{(t)} - \mathbf{W}^*)\mathbf{X}_{\mathrm{TP}}\mathbf{X}_{\mathrm{TP}}^{\mathrm{T}}(\mathbf{W}^{(t)} - \mathbf{W}^*)^{\mathrm{T}}\Big) \\ &= \frac{4\eta}{(1-\epsilon)N} \cdot \langle \mathrm{vec}((\mathbf{W}^{(t)} - \mathbf{W}^*)^{\mathrm{T}}), \mathrm{vec}(\mathbf{X}_{\mathrm{TP}}\mathbf{X}_{\mathrm{TP}}^{\mathrm{T}}(\mathbf{W}^{(t)} - \mathbf{W}^*)^{\mathrm{T}})\rangle \\ &= \frac{4\eta}{(1-\epsilon)N} \cdot \langle \mathrm{vec}((\mathbf{W}^{(t)} - \mathbf{W}^*)^{\mathrm{T}}), (\mathbf{I} \otimes \mathbf{X}_{\mathrm{TP}}\mathbf{X}_{\mathrm{TP}}^{\mathrm{T}})\mathrm{vec}((\mathbf{W}^{(t)} - \mathbf{W}^*)^{\mathrm{T}})\rangle\end{aligned}$$

$$= \frac{4\eta}{(1-\epsilon)N} \cdot \sum_{k\in[K]} \left\langle \mathbf{w}_k^{(t)} - \mathbf{w}_k^*, \mathbf{X}_{\mathrm{TP}}\mathbf{X}_{\mathrm{TP}}^{\mathrm{T}}(\mathbf{w}_k^{(t)} - \mathbf{w}_k^*) \right\rangle$$

$$\geq \frac{2\eta}{(1-\epsilon)N} \cdot \sum_{k\in[K]} \left(\lambda_{\min}(\mathbf{X}_{\mathrm{TP}}\mathbf{X}_{\mathrm{TP}}^{\mathrm{T}})\left\|\mathbf{w}_k^{(t)} - \mathbf{w}_k^*\right\|_2^2 + \left\|\mathbf{X}_{\mathrm{TP}}\mathbf{X}_{\mathrm{TP}}^{\mathrm{T}}\right\|_2^{-1}\left\|\mathbf{X}_{\mathrm{TP}}\mathbf{X}_{\mathrm{TP}}^{\mathrm{T}}(\mathbf{w}_k^{(t)} - \mathbf{w}_k^*)\right\|_2^2\right)$$

$$= \frac{2\eta}{(1-\epsilon)N} \cdot \lambda_{\min}(\mathbf{X}_{\mathrm{TP}}\mathbf{X}_{\mathrm{TP}}^{\mathrm{T}})\left\|\mathbf{W}^{(t)} - \mathbf{W}^*\right\|_{\mathrm{F}}^2 + \frac{2\eta}{(1-\epsilon)N} \cdot \left\|\mathbf{X}_{\mathrm{TP}}\mathbf{X}_{\mathrm{TP}}^{\mathrm{T}}\right\|_2^{-1}\left\|(\mathbf{W}^{(t)} - \mathbf{W}^*)\mathbf{X}_{\mathrm{TP}}\mathbf{X}_{\mathrm{TP}}^{\mathrm{T}}\right\|_{\mathrm{F}}^2 \quad \text{(A.4)}$$

In the above, in the first relation we use the cyclic property of the trace, in the second relation we us the relation given in Lemma 36, in the third relation we apply the relation given in Lemma 37, in the fifth relation we apply Lemma 32, and in the fourth relation we apply Lemma 37, which gives the following equality,

$$(\mathbf{I} \otimes \mathbf{X}_{\mathrm{TP}}\mathbf{X}_{\mathrm{TP}}^{\mathrm{T}})\mathrm{vec}((\mathbf{W}^{(t)} - \mathbf{W}^*)^{\mathrm{T}}) = \begin{bmatrix} \mathbf{X}_{\mathrm{TP}}\mathbf{X}_{\mathrm{TP}}^{\mathrm{T}} & & \\ & \ddots & \\ & & \mathbf{X}_{\mathrm{TP}}\mathbf{X}_{\mathrm{TP}}^{\mathrm{T}} \end{bmatrix} \begin{bmatrix} \mathbf{w}_1 - \mathbf{w}_1^* \\ \vdots \\ \mathbf{w}_K - \mathbf{w}_K^* \end{bmatrix} = \begin{bmatrix} \mathbf{X}_{\mathrm{TP}}\mathbf{X}_{\mathrm{TP}}^{\mathrm{T}}(\mathbf{w}_1 - \mathbf{w}_1^*) \\ \vdots \\ \mathbf{X}_{\mathrm{TP}}\mathbf{X}_{\mathrm{TP}}^{\mathrm{T}}(\mathbf{w}_K - \mathbf{w}_K^*) \end{bmatrix}$$

We now upper bound the second term in Equation (A.1),

$$\begin{aligned}
\left\|\eta\nabla\mathcal{R}(\mathbf{W}^{(t)}; \mathrm{FP})\right\|_{\mathrm{F}} &\stackrel{\mathrm{def}}{=} \frac{2\eta}{(1-\epsilon)N} \cdot \left\|(\mathbf{W}^{(t)}\mathbf{X}_{\mathrm{FP}} - \mathbf{Y}_{\mathrm{FP}})\mathbf{X}_{\mathrm{FP}}^{\mathrm{T}}\right\|_{\mathrm{F}} \\
&\leq \frac{2\eta}{(1-\epsilon)N} \cdot \left\|\mathbf{X}_{\mathrm{FP}}\right\|_2\left\|\mathbf{W}^{(t)}\mathbf{X}_{\mathrm{FP}} - \mathbf{Y}_{\mathrm{FP}}\right\|_{\mathrm{F}} \\
&\leq \frac{2\eta}{(1-\epsilon)N} \cdot \left\|\mathbf{X}_{\mathrm{FP}}\right\|_2\left\|\mathbf{W}^{(t)}\mathbf{X}_{\mathrm{FN}} - \mathbf{Y}_{\mathrm{FN}}\right\|_{\mathrm{F}} \\
&\leq \frac{2\eta}{(1-\epsilon)N} \cdot \left\|\mathbf{X}_{\mathrm{FP}}\right\|_2\left(\left\|\mathbf{W}^{(t)}\mathbf{X}_{\mathrm{FN}} - \mathbf{W}^*\mathbf{X}_{\mathrm{FN}}\right\|_{\mathrm{F}} + \left\|\mathbf{E}_{\mathrm{FN}}\right\|_{\mathrm{F}}\right) \\
&\leq \frac{2\eta}{(1-\epsilon)N} \cdot \left\|\mathbf{X}_{\mathrm{FP}}\right\|_2\left\|\mathbf{X}_{\mathrm{FN}}\right\|_2\left\|\mathbf{W}^{(t)} - \mathbf{W}^*\right\|_{\mathrm{F}} + \frac{2\eta}{(1-\epsilon)N} \cdot \left\|\mathbf{X}_{\mathrm{FP}}\right\|_2\left\|\mathbf{E}_{\mathrm{FN}}\right\|_{\mathrm{F}}
\end{aligned}$$

In the above, the first and fourth inequalities from the fact that for any two size compatible matrices, $\mathbf{A}, \mathbf{B}$, it holds that $\|\mathbf{AB}\|_{\mathrm{F}} \leq \|\mathbf{A}\|_{\mathrm{F}}\|\mathbf{B}\|_2$, 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 (A.1),

$$\left\|\eta\nabla\mathcal{R}(\mathbf{W}^*; \mathrm{FN})\right\|_{\mathrm{F}} \stackrel{\mathrm{def}}{=} \frac{2\eta}{(1-\epsilon)N} \cdot \left\|(\mathbf{W}^*\mathbf{X}_{\mathrm{FN}} - \mathbf{Y}_{\mathrm{FN}})\mathbf{X}_{\mathrm{FN}}^{\mathrm{T}}\right\|_{\mathrm{F}} \leq \frac{2\eta}{(1-\epsilon)N} \cdot \left\|\mathbf{E}_{\mathrm{FN}}\mathbf{X}_{\mathrm{FN}}^{\mathrm{T}}\right\|_{\mathrm{F}} \quad \text{(A.5)}$$

In the above, we use the fact that for any two size compatible matrices, $\mathbf{A}, \mathbf{B}$, it holds that $\|\mathbf{AB}\|_{\mathrm{F}} \leq \|\mathbf{A}\|_{\mathrm{F}}\|\mathbf{B}\|_2$.

Step 2: Consistency. We will now upper bound the consistency estimate of the fourth term of Equation (A.1).

$$\left\|\eta\nabla\mathcal{R}(\mathbf{W}^*; \mathrm{P}) - \eta\nabla\mathcal{R}(\widehat{\mathbf{W}}; \mathrm{P})\right\|_{\mathrm{F}} \stackrel{\mathrm{def}}{=} \frac{2\eta}{(1-\epsilon)N} \cdot \left\|(\widehat{\mathbf{W}} - \mathbf{W}^*)\mathbf{X}_{\mathrm{P}}\mathbf{X}_{\mathrm{P}}^{\mathrm{T}}\right\|_{\mathrm{F}} \leq \frac{2\eta}{(1-\epsilon)N} \cdot \left\|\mathbf{E}_{\mathrm{P}}\mathbf{X}_{\mathrm{P}}^{\mathrm{T}}\right\|_{\mathrm{F}}$$

We can then have from Proposition 34,

$$\mathbb{E}\left\|\mathbf{E}_{\mathrm{P}}\mathbf{X}_{\mathrm{P}}^{\mathrm{T}}\right\|_{\mathrm{F}}^2 = \sigma^2\left(K\,\mathbb{E}\left\|\mathbf{X}_{\mathrm{P}}\right\|_{\mathrm{F}}^2\right) = \sigma^2\left(K\,\mathbb{E}\,\mathrm{Tr}\left(\mathbf{X}_{\mathrm{P}}\mathbf{X}_{\mathrm{P}}^{\mathrm{T}}\right)\right) = \sigma^2((1-\epsilon)NK\,\mathrm{Tr}(\mathbf{\Sigma}))$$

Then from a simple application of Markov's Inequality, we have with probability exceeding $1-\delta$,

$$\frac{2\eta}{(1-\epsilon)N} \cdot \left\|\mathbf{E}_{\mathrm{P}}\mathbf{X}_{\mathrm{P}}^{\mathrm{T}}\right\|_{\mathrm{F}} \leq 2\eta\sigma \cdot \sqrt{\frac{(1/\delta)K\,\mathrm{Tr}(\mathbf{\Sigma})}{(1-\epsilon)N}}$$

We then have from our choice of $\eta = 0.1\lambda_{\max}^{-2}(\mathbf{\Sigma})$, the third term in Equation (A.1) will be less than the second term in Equation (A.4). We thus obtain from noting that $\sqrt{1-2x} \leq 1-x$ for any $x \leq 1/2$,

$$\left\|\mathbf{W}^{(t+1)}-\mathbf{W}^*\right\|_{\mathrm{F}} \leq \left\|\mathbf{W}^{(t)}-\mathbf{W}^*\right\|_{\mathrm{F}}\left(1-\frac{2\eta}{(1-\epsilon)N}\cdot\lambda_{\min}(\mathbf{X}_{\mathrm{TP}}\mathbf{X}_{\mathrm{TP}}^{\mathrm{T}})+\frac{2\eta}{(1-\epsilon)N}\cdot\left\|\mathbf{X}_{\mathrm{FP}}\right\|_2\left\|\mathbf{X}_{\mathrm{FN}}\right\|_2\right)$$
$$+\frac{2\eta}{(1-\epsilon)N}\cdot\left\|\mathbf{X}_{\mathrm{FP}}\right\|_2\left\|\mathbf{E}_{\mathrm{FN}}\right\|_{\mathrm{F}}+\frac{2\eta}{(1-\epsilon)N}\cdot\left\|\mathbf{E}_{\mathrm{FN}}\mathbf{X}_{\mathrm{FN}}^{\mathrm{T}}\right\|_{\mathrm{F}}+2\sigma\eta\cdot\sqrt{\frac{(1/\delta)K\,\mathrm{Tr}(\boldsymbol{\Sigma})}{(1-\epsilon)N}}$$

Step 3: Concentration Bounds. From Proposition 22, we obtain with probability exceeding $1-\delta$ that $\|\mathbf{E}_{\mathrm{FN}}\|_{\mathrm{F}} \leq \sigma\sqrt{30NK\epsilon\log\epsilon^{-1}}$. From our assumption on bounded covariate corruptions, we have $\|\mathbf{E}_{\mathrm{FN}}\|_{\mathrm{F}}\|\mathbf{X}_{\mathrm{FP}}\|_{\mathrm{F}} \leq \sigma\epsilon\sqrt{30NKC_3\log\epsilon^{-1}}$. From Lemma 25, we have that $\|\mathbf{E}_{\mathrm{FN}}\mathbf{X}_{\mathrm{FN}}^{\mathrm{T}}\|_{\mathrm{F}} \leq \sqrt{6K\log N}\|\mathbf{X}_{\mathrm{FN}}\|$ when $n \geq (1/\delta)$ with probability exceeding $1-\delta$. From Lemma 24, we have for $\epsilon \leq \frac{1}{240}\kappa^{-1}(\boldsymbol{\Sigma})$, the minimum eigenvalue satisfies $\lambda_{\min}(\mathbf{X}_{\mathrm{TP}}\mathbf{X}_{\mathrm{TP}}^{\mathrm{T}}) \geq (N/2)\cdot(1-2\epsilon)\cdot\lambda_{\min}(\boldsymbol{\Sigma}) \geq (N/4)\cdot\lambda_{\min}(\boldsymbol{\Sigma})$. From our assumption of corrupted covariates, we have that $\|\mathbf{X}_{\mathrm{FP}}\| \leq \sqrt{\epsilon NC_3}$. We also have from Lemma 24, we have $\|\mathbf{X}_{\mathrm{FN}}\| \leq \sqrt{\lambda_{\max}(\boldsymbol{\Sigma})\cdot(10N\epsilon\log\epsilon^{-1})}$ with high probability. Then when $\epsilon \leq \sqrt{\frac{1}{960C_3}\cdot\kappa^{-1}(\boldsymbol{\Sigma})\lambda_{\min}(\boldsymbol{\Sigma})}$, we have that $\|\mathbf{X}_{\mathrm{FP}}\|_2\|\mathbf{X}_{\mathrm{FN}}\|_2 \leq (N/8)\lambda_{\min}(\boldsymbol{\Sigma})$. Then, solving for the induction with an infinite sum, we have after $O\Big(\kappa(\boldsymbol{\Sigma})\cdot\log\Big(\frac{\|\mathbf{W}^*\|_{\mathrm{F}}}{\varepsilon}\Big)\Big)$ iterations,

$$\left\|\mathbf{W}^{(t+1)}-\mathbf{W}^*\right\|_{\mathrm{F}} \leq \varepsilon+\sigma\epsilon\sqrt{4800KC_3\log\epsilon^{-1}}\cdot\frac{\sqrt{\lambda_{\max}(\boldsymbol{\Sigma})}}{\lambda_{\min}(\boldsymbol{\Sigma})}+\frac{\sqrt{60\sigma K\log N\cdot\epsilon\log\epsilon^{-1}\lambda_{\max}(\boldsymbol{\Sigma})}}{\sqrt{N}\lambda_{\min}(\boldsymbol{\Sigma})}$$
$$+\frac{\sigma}{\lambda_{\min}(\boldsymbol{\Sigma})}\cdot\sqrt{\frac{(8/\delta)K\,\mathrm{Tr}(\boldsymbol{\Sigma})}{N}} \leq \varepsilon+\sigma\epsilon\sqrt{43200KC_3\log\epsilon^{-1}}$$

In the above, the final equality follows when $N \geq \frac{(1/\delta)K\,\mathrm{Tr}(\boldsymbol{\Sigma})}{1600\epsilon^2\log\epsilon^{-1}\lambda_{\max}(\boldsymbol{\Sigma})} \vee \frac{1}{6400C_3^2\epsilon^2}$. Our proof is complete. ■

B Proofs for Learning Generalized Linear Models

B.1 Proof of Theorem 14

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

$$\mathbf{w}^{(t+1)} = \mathbf{w}^{(t)} - \frac{2\eta}{(1-\epsilon)N}\cdot\sum_{i\in S^{(t)}}(\omega(\mathbf{x}_i^{\mathrm{T}}\mathbf{w}^{(t)})-y_i)\cdot\omega'(f^{(t)}(\mathbf{x}_i))\cdot\mathbf{x}_i$$

Step 1: Linear Convergence.

$$\begin{aligned}\|\mathbf{w}^{(t+1)}-\mathbf{w}^*\|_2 &= \|\mathbf{w}^{(t)}-\eta\nabla\mathcal{R}(\mathbf{w}^{(t)};\mathrm{S}^{(t)})-\mathbf{w}^*\|_2\\ &= \|\mathbf{w}^{(t)}-\mathbf{w}^*-\eta\nabla\mathcal{R}(\mathbf{w}^{(t)};\mathrm{TP})-\eta\nabla\mathcal{R}(\mathbf{w}^{(t)};\mathrm{FP})\|_2\\ &\leq \|\mathbf{w}^{(t)}-\mathbf{w}^*-\eta\nabla\mathcal{R}(\mathbf{w}^{(t)};\mathrm{TP})\|_2+\|\eta\nabla\mathcal{R}(\mathbf{w}^{(t)};\mathrm{FP})\|_2 \qquad \text{(B.1)}\end{aligned}$$

We upper bound the first term of Equation (B.1) through its square.

$$\|\mathbf{w}^{(t)}-\mathbf{w}^*-\eta\nabla\mathcal{R}(\mathbf{w}^{(t)};\mathrm{TP})\|_2^2 = \|\mathbf{w}^{(t)}-\mathbf{w}^*\|_2^2-2\eta\cdot\langle\mathbf{w}^{(t)}-\mathbf{w}^*,\nabla\mathcal{R}(\mathbf{w}^{(t)};\mathrm{TP})\rangle+\eta^2\cdot\|\nabla\mathcal{R}(\mathbf{w}^{(t)};\mathrm{TP})\|_2^2 \qquad \text{(B.2)}$$

We now lower bound the second term of Equation (B.2).

$$\begin{aligned}&2\eta\cdot\langle\mathbf{w}^{(t)}-\mathbf{w}^*,\nabla\mathcal{R}(\mathbf{w}^{(t)};\mathrm{TP})\rangle = \frac{4\eta}{(1-\epsilon)N}\cdot\langle\mathbf{w}^{(t)}-\mathbf{w}^*,\sum_{i\in\mathrm{TP}}(\omega(\mathbf{x}_i^{\mathrm{T}}\mathbf{w}^{(t)})-y_i)\cdot\omega'(\mathbf{x}_i^{\mathrm{T}}\mathbf{w}^{(t)})\cdot\mathbf{x}_i\rangle\\ &= \frac{4\eta}{(1-\epsilon)N}\cdot\langle\mathbf{w}^{(t)}-\mathbf{w}^*,\sum_{i\in\mathrm{TP}}(\omega(\mathbf{x}_i^{\mathrm{T}}\mathbf{w}^{(t)})-\omega(\mathbf{x}_i^{\mathrm{T}}\mathbf{w}^*+\xi_i))\cdot\omega'(\mathbf{x}_i^{\mathrm{T}}\mathbf{w}^{(t)})\cdot\mathbf{x}_i\rangle\\ &\geq \frac{4\eta}{(1-\epsilon)N}\cdot C_{[\omega]}^2\lambda_{\min}(\mathbf{X}_{\mathrm{TP}}\mathbf{X}_{\mathrm{TP}}^{\mathrm{T}})\|\mathbf{w}^{(t)}-\mathbf{w}^*\|_2^2-\frac{4\eta}{(1-\epsilon)N}\cdot\|\mathbf{w}^{(t)}-\mathbf{w}^*\|_2\Big\|\sum_{i\in\mathrm{TP}}\xi_i\omega'(c_i)\omega'(\mathbf{x}_i^{\mathrm{T}}\mathbf{w}^{(t)})\cdot\mathbf{x}_i\Big\|_2\end{aligned} \qquad \text{(B.3)}$$

In the above, in the final relation, we apply Cauchy-Schwarz inequality after using the mean-value theorem, which gives us the existance of some $c_i \in \mathbb{R}$ such that $\omega'(c_i)(\mathbf{x}_i^{\mathrm{T}}\mathbf{w}^{(t)} - \mathbf{x}_i^{\mathrm{T}}\mathbf{w}^* - \xi_i) = \omega(\mathbf{x}_i^{\mathrm{T}}\mathbf{w}^{(t)}) - \omega(\mathbf{x}_i^{\mathrm{T}}\mathbf{w}^* + \xi_i)$. Then from an application of Young's Inequality (see Proposition 33), we obtain

$$\begin{aligned}
&\frac{4\eta}{(1-\epsilon)N} \cdot \|\mathbf{w}^{(t)} - \mathbf{w}^*\|_2 \Big\| \sum_{i \in \mathrm{TP}} \xi_i \omega'(c_i)\omega'(\mathbf{x}_i^{\mathrm{T}}\mathbf{w}^{(t)}) \cdot \mathbf{x}_i \Big\|_2 \\
&\leq \frac{\eta}{(1-\epsilon)N} \cdot C_{\lfloor \omega \rfloor}^2 \lambda_{\min}(\mathbf{X}_{\mathrm{TP}}\mathbf{X}_{\mathrm{TP}}^{\mathrm{T}}) \|\mathbf{w}^{(t)} - \mathbf{w}^*\|_2^2 + \frac{4\eta}{(1-\epsilon)N} \cdot \lambda_{\min}^{-1}(\mathbf{X}_{\mathrm{TP}}\mathbf{X}_{\mathrm{TP}}^{\mathrm{T}}) \Big\| \sum_{i \in \mathrm{TP}} \xi_i \omega'(c_i)\omega'(\mathbf{x}_i^{\mathrm{T}}\mathbf{w}^{(t)}) \cdot \mathbf{x}_i \Big\|_2^2
\end{aligned}$$

We next upper bound the third term in Equation (B.2).

$$\begin{aligned}
\eta^2 \cdot \|\nabla \mathcal{R}(\mathbf{w}^{(t)}; \mathrm{TP})\|_2^2 &= \frac{4\eta^2}{[(1-\epsilon)N]^2} \cdot \Big\| \sum_{i \in \mathrm{TP}} (\omega(\mathbf{x}_i^{\mathrm{T}}\mathbf{w}^{(t)}) - \omega(\mathbf{x}_i^{\mathrm{T}}\mathbf{w}^* + \xi_i)) \cdot \omega'(\mathbf{x}_i^{\mathrm{T}}\mathbf{w}^{(t)}) \cdot \mathbf{x}_i \Big\|_2^2 \\
&= \frac{4\eta^2}{[(1-\epsilon)N]^2} \cdot \Big\| \sum_{i \in \mathrm{TP}} (\mathbf{x}_i^{\mathrm{T}}\mathbf{w}^{(t)} - \mathbf{x}_i^{\mathrm{T}}\mathbf{w}^* + \xi_i) \cdot \omega'(c_i) \cdot \omega'(\mathbf{x}_i^{\mathrm{T}}\mathbf{w}^{(t)}) \cdot \mathbf{x}_i \Big\|_2^2 \\
&\leq \frac{8\eta^2}{[(1-\epsilon)N]^2} \cdot \left(C_{\lceil \omega \rceil}^4 \lambda_{\max}^2(\mathbf{X}_{\mathrm{TP}}\mathbf{X}_{\mathrm{TP}}^{\mathrm{T}}) \|\mathbf{w}^{(t)} - \mathbf{w}^*\|_2^2 + \Big\| \sum_{i \in \mathrm{TP}} \xi_i \cdot \omega'(c_i) \cdot \omega'(\mathbf{x}_i^{\mathrm{T}}\mathbf{w}^{(t)}) \cdot \mathbf{x}_i \Big\|_2^2 \right) \qquad \text{(B.4)}
\end{aligned}$$

Then, from choosing $\eta \leq \frac{C_{\lfloor \omega \rfloor}^2 (1-\epsilon) N \lambda_{\min}(\mathbf{X}_{\mathrm{TP}}\mathbf{X}_{\mathrm{TP}}^{\mathrm{T}})}{4 C_{\lceil \omega \rceil}^4 \lambda_{\max}^2(\mathbf{X}_{\mathrm{TP}}\mathbf{X}_{\mathrm{TP}}^{\mathrm{T}})}$. We see the first term of Equation (B.4) is less than half the first term in Equation (B.3). We now bound the second term in Equation (B.1).

$$\begin{aligned}
\eta^2 \cdot \|\nabla \mathcal{R}(\mathbf{w}^{(t)}; \mathrm{FP})\|_2^2 &= \frac{4\eta^2}{[(1-\epsilon)N]^2} \cdot \Big\| \sum_{i \in \mathrm{FP}} (\omega(\mathbf{x}_i^{\mathrm{T}}\mathbf{w}^{(t)}) - y_i) \cdot \omega'(\mathbf{x}_i^{\mathrm{T}}\mathbf{w}) \cdot \mathbf{x}_i \Big\|_2^2 \\
&\leq \frac{4\eta^2}{[(1-\epsilon)N]^2} \cdot C_{\lceil \omega \rceil}^2 \big\| \mathbf{X}_{\mathrm{FP}}\mathbf{X}_{\mathrm{FP}}^{\mathrm{T}} \big\|_2 \sum_{i \in \mathrm{FP}} (\omega(\mathbf{x}_i^{\mathrm{T}}\mathbf{w}^{(t)}) - y_i)^2 \\
&\leq \frac{4\eta^2}{[(1-\epsilon)N]^2} \cdot C_{\lceil \omega \rceil}^2 \big\| \mathbf{X}_{\mathrm{FP}}\mathbf{X}_{\mathrm{FP}}^{\mathrm{T}} \big\|_2 \sum_{i \in \mathrm{FN}} (\omega(\mathbf{x}_i^{\mathrm{T}}\mathbf{w}^{(t)}) - y_i)^2 \\
&= \frac{4\eta^2}{[(1-\epsilon)N]^2} \cdot C_{\lceil \omega \rceil}^2 \big\| \mathbf{X}_{\mathrm{FP}}\mathbf{X}_{\mathrm{FP}}^{\mathrm{T}} \big\|_2 \sum_{i \in \mathrm{FN}} (\omega(\mathbf{x}_i^{\mathrm{T}}\mathbf{w}^{(t)}) - \omega(\mathbf{x}_i^{\mathrm{T}}\mathbf{w}^*) + \xi_i)^2 \\
&\leq \frac{8\eta^2}{[(1-\epsilon)N]^2} \cdot C_{\lceil \omega \rceil}^2 \big\| \mathbf{X}_{\mathrm{FP}}\mathbf{X}_{\mathrm{FP}}^{\mathrm{T}} \big\|_2 \Big(C_{\lceil \omega \rceil}^2 \cdot \big\| \mathbf{X}_{\mathrm{FN}}\mathbf{X}_{\mathrm{FN}}^{\mathrm{T}} \big\|_2 \|\mathbf{w}^{(t)} - \mathbf{w}^*\|_2^2 + \|\boldsymbol{\xi}_{\mathrm{FN}}\|_2^2 \Big)
\end{aligned}$$

Concluding the step, we have

$$\begin{aligned}
\|\mathbf{w}^{(t+1)} - \mathbf{w}^*\|_2 \leq \|\mathbf{w}^{(t)} - \mathbf{w}^*\|_2 &\left(1 - \frac{\eta}{2(1-\epsilon)N} \cdot C_{\lfloor \omega \rfloor}^2 \lambda_{\min}(\mathbf{X}_{\mathrm{TP}}\mathbf{X}_{\mathrm{TP}}^{\mathrm{T}}) + \frac{\sqrt{8}\eta}{(1-\epsilon)N} \cdot C_{\lceil \omega \rceil}^2 \big\| \mathbf{X}_{\mathrm{FP}} \big\|_2 \big\| \mathbf{X}_{\mathrm{FN}} \big\|_2 \right) \\
+ &\left(\frac{\sqrt{8}\eta}{(1-\epsilon)N} + \frac{2\sqrt{\eta}}{\sqrt{(1-\epsilon)N}} \cdot \sigma_{\min}^{-1}(\mathbf{X}_{\mathrm{TP}}) \right) \Big\| \sum_{i \in \mathrm{TP}} \xi_i \omega'(c_i)\omega'(\mathbf{x}_i^{\mathrm{T}}\mathbf{w}^{(t)}) \mathbf{x}_i \Big\|_2 + \frac{\sqrt{8}\eta}{(1-\epsilon)N} \cdot \|\mathbf{X}_{\mathrm{FP}}\|_2 \|\boldsymbol{\xi}_{\mathrm{FN}}\|_2
\end{aligned}$$

Step 2: Concentration Bounds. First we will show that $\xi\omega'(c)\omega'(\mathbf{x}^{\mathrm{T}}\mathbf{w})$ is sub-Gaussian for any $c \in \mathbb{R}$, $\mathbf{w} \in \mathbb{R}^d$ and $\mathbf{x}$ sampled from a sub-Gaussian distribution. Note that ξ is sub-Gaussian s.t. $\|\xi\|_{\psi_2} = C_\sigma$. Then, we have

$$\left(\mathbb{E} \big| \xi\omega'(c)\omega'(\mathbf{x}^{\mathrm{T}}\mathbf{w}) \big|^p \right)^{1/p} \overset{(i)}{\leq} \left(\sup_{x \in \mathbb{R}} [\omega'(x)]^{6p} \, \mathbb{E}|\xi|^{3p} \right)^{1/3p} \leq C_{\lceil \psi \rceil}^2 C_\sigma$$

In (i) we apply Hölder's Inequality. We thus have $\|\xi\omega'(c)\omega'(\mathbf{x}^{\mathrm{T}}\mathbf{w})\|_{\psi_2} \leq C_{\lceil \psi \rceil}^2 C_\sigma$. Then applying Lemma 19, we have with probability at least $1 - \delta$,

$$\| \sum_{i \in \mathrm{TP}} \xi_i \cdot \omega'(c_i) \cdot \omega'(\mathbf{x}_i^{\mathrm{T}}\mathbf{w}^{(t)}) \cdot \mathbf{x}_i \|_2 \leq N C_K C_{\lceil \psi \rceil}^2 C_\sigma \left(\frac{2d}{N} \log(5) + \frac{2}{N} \log(1/\delta) + 6\epsilon \log \epsilon^{-1} \right)^{1/2}$$

■

C Proofs for Neural Networks

C.1 Leaky-ReLU Networks

C.1.1 Proof of Theorem 16

Proof. Recall the neural network function is given as follows,

$$f_{\mathbf{w}}(\mathbf{x}) = \psi(\mathbf{x}^{\mathrm{T}}\mathbf{w})$$

where $\psi(x) = \max\{C_\psi x, x\}$ for a $0 < C_\psi < 1$. Then we have the empirical risk as,

$$\mathcal{R}(\mathbf{w}; \mathrm{S}^{(t)}) = \frac{1}{(1-\epsilon)N} \cdot \sum_{i \in \mathrm{S}^{(t)}} (\psi(\mathbf{x}_i^{\mathrm{T}}\mathbf{w}) - y_i)^2$$

The gradient is then given as follows„

$$\frac{\partial \mathcal{R}(\mathbf{w}; \mathrm{S}^{(t)})}{\partial \mathbf{w}_k} = \frac{2}{(1-\epsilon)N} \cdot \sum_{i \in \mathrm{S}^{(t)}} (\psi(\mathbf{x}_i^{\mathrm{T}}\mathbf{w}) - y_i) \cdot \psi'(\mathbf{x}_i^{\mathrm{T}}\mathbf{w}) \cdot \mathbf{x}_i$$

For the Hessian, we have

$$\frac{\partial^2 \mathcal{R}(\mathbf{w}; \mathrm{S}^{(t)})}{\partial \mathbf{w}^2} = \frac{2}{(1-\epsilon)N} \cdot \left(\sum_{i \in \mathrm{S}^{(t)}} (\psi(\mathbf{x}_i^{\mathrm{T}}\mathbf{w}) - y_i) \cdot \psi''(\mathbf{x}_i^{\mathrm{T}}\mathbf{w}) \cdot \mathbf{x}_i\mathbf{x}_i^{\mathrm{T}} + \sum_{i \in \mathrm{S}^{(t)}} [\psi'(\mathbf{x}_i^{\mathrm{T}}\mathbf{w})]^2 \cdot \mathbf{x}_i\mathbf{x}_i^{\mathrm{T}} \right)$$

Then, noting that $\psi''(x) = 0$ at finitely many points, we have,

$$\frac{\partial^2 \mathcal{R}(\mathbf{w}; \mathrm{S}^{(t)})}{\partial \mathbf{w}^2} \overset{\text{a.s.}}{=} \frac{2}{(1-\epsilon)N} \cdot \sum_{i \in \mathrm{S}^{(t)}} [\psi'(\mathbf{x}_i^{\mathrm{T}}\mathbf{w})]^2 \cdot \mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}$$

Step 1: Upper bounding the Frobenius norm distance between $\mathbf{w}^{(t+1)}$ and $\mathbf{w}^*$. Let $\widehat{\mathbf{w}} = \arg\min_{\mathbf{w} \in \mathbb{R}^{K \times d}} \mathcal{R}(\mathbf{w}; \mathrm{P})$. We then have,

$$\begin{aligned}
&\|\mathbf{w}^{(t+1)} - \mathbf{w}^*\|_2 = \|\mathbf{w}^{(t)} - \mathbf{w}^* - \eta \nabla \mathcal{R}(\mathbf{w}^{(t)}; \mathrm{S}^{(t)})\|_2 \\
&= \|\mathbf{w}^{(t)} - \mathbf{w}^* - \eta \nabla \mathcal{R}(\mathbf{w}^{(t)}; \mathrm{TP}) + \eta \nabla \mathcal{R}(\mathbf{w}^*; \mathrm{P}) - \eta \nabla \mathcal{R}(\mathbf{w}^*; \mathrm{P}) - \nabla \mathcal{R}(\mathbf{w}^{(t)}; \mathrm{FP})\|_2 \\
&\leq \|\mathbf{w}^{(t)} - \mathbf{w}^* - \eta \nabla \mathcal{R}(\mathbf{w}^{(t)}; \mathrm{TP}) + \eta \nabla \mathcal{R}(\mathbf{w}^*; \mathrm{TP})\|_2 + \|\eta \nabla \mathcal{R}(\mathbf{w}^*; \mathrm{TP})\|_2 + \|\eta \nabla \mathcal{R}(\mathbf{w}^{(t)}; \mathrm{FP})\|_2 \qquad \text{(C.1)}
\end{aligned}$$

We will first upper bound the first term of Equation (C.1) through its square.

$$\begin{aligned}
&\|\mathbf{w}^{(t)} - \mathbf{w}^* - \eta \nabla \mathcal{R}(\mathbf{w}^{(t)}; \mathrm{TP})\|_2^2 \\
&\qquad = \|\mathbf{w}^{(t)} - \mathbf{w}^*\|_2^2 - 2\eta \cdot \langle \mathbf{w}^{(t)} - \mathbf{w}^*, \nabla \mathcal{R}(\mathbf{w}^{(t)}; \mathrm{TP}) \rangle + \eta^2 \cdot \|\nabla \mathcal{R}(\mathbf{w}^{(t)}; \mathrm{TP})\|^2 \qquad \text{(C.2)}
\end{aligned}$$

We will first lower bound the second term of Equation (C.2).

$$\begin{aligned}
&2\eta \cdot \langle \mathbf{w}^{(t)} - \mathbf{w}^*, \nabla \mathcal{R}(\mathbf{w}^{(t)}; \mathrm{TP}) - \nabla \mathcal{R}(\mathbf{w}^{(*)}; \mathrm{TP}) \rangle \\
&\qquad = \frac{4\eta}{(1-\epsilon)N} \cdot \langle \mathbf{w}^{(t)} - \mathbf{w}^*, \int_0^1 \nabla^2 \mathcal{R}(\mathbf{w}^* + \theta(\mathbf{w}^* - \mathbf{w}^{(t)}); \mathrm{TP}) d\theta \cdot (\mathbf{w}^{(t)} - \mathbf{w}^*) \rangle \\
&\qquad \geq \frac{4C_\psi^2 \eta}{(1-\epsilon)N} \cdot \lambda_{\min}(\mathbf{X}_{\mathrm{TP}} \mathbf{X}_{\mathrm{TP}}^{\mathrm{T}}) \|\mathbf{w}^{(t)} - \mathbf{w}^*\|_2^2 \qquad \text{(C.3)}
\end{aligned}$$

We now will upper bound the second term of Equation (C.2).

$$\begin{aligned}\eta^2 \cdot \|\nabla\mathcal{R}(\mathbf{w}^{(t)};\mathrm{TP}) - \nabla\mathcal{R}(\mathbf{w}^*;\mathrm{TP})\|_2^2 \\ &= \frac{4\eta^2}{[(1-\epsilon)N]^2} \cdot \Big\| \int_0^1 \nabla^2\mathcal{R}(\mathbf{w}^* + \theta(\mathbf{w}^* - \mathbf{w}^{(t)});\mathrm{TP}) d\theta \cdot (\mathbf{w}^{(t)} - \mathbf{w}^*) \Big\|_2^2 \\ &\leq \frac{4\eta^2}{[(1-\epsilon)N]^2} \cdot \lambda_{\max}^2(\mathbf{X}_{\mathrm{TP}}\mathbf{X}_{\mathrm{TP}}^{\mathrm{T}}) \|\mathbf{w}^{(t)} - \mathbf{w}^*\|_2^2 \end{aligned} \tag{C.4}$$

We then observe that the first term of Equation (C.3) is greater than half the first term in Equation (C.4) when $\eta \leq \frac{C_\psi^2(1-\epsilon)N\lambda_{\min}(\mathbf{X}_{\mathrm{TP}}\mathbf{X}_{\mathrm{TP}}^{\mathrm{T}})}{2\lambda_{\max}^2(\mathbf{X}_{\mathrm{TP}}\mathbf{X}_{\mathrm{TP}}^{\mathrm{T}})}$.

Step 2: Upper bounding the corrupted gradient. We now upper bound the third term in Equation (C.2).

$$\begin{aligned}\eta^2 \cdot \|\nabla\mathcal{R}(\mathbf{W}^{(t)};\mathrm{FP})\|_{\mathrm{F}}^2 &= \frac{4\eta^2}{[(1-\epsilon)N]^2} \cdot \Big\| \sum_{i\in\mathrm{FP}} (\psi(\mathbf{x}_i^{\mathrm{T}}\mathbf{w}) - \psi(\mathbf{x}_i^{\mathrm{T}}\mathbf{w}^*)) \cdot \psi'(\mathbf{x}_i^{\mathrm{T}}\mathbf{w}_k^{(t)}) \cdot \mathbf{x}_i \Big\|_2^2 \\ &\leq \frac{4\eta^2}{[(1-\epsilon)N]^2} \cdot \big\|\mathbf{X}_{\mathrm{FP}}\mathbf{X}_{\mathrm{FP}}^{\mathrm{T}}\big\|_2 \sum_{i\in\mathrm{FP}} \big(\psi(\mathbf{x}_i^{\mathrm{T}}\mathbf{w}) - \psi(\mathbf{x}_i^{\mathrm{T}}\mathbf{w}^*)\big)^2 \\ &\leq \frac{4\eta^2}{[(1-\epsilon)N]^2} \cdot \big\|\mathbf{X}_{\mathrm{FP}}\mathbf{X}_{\mathrm{FP}}^{\mathrm{T}}\big\|_2 \sum_{i\in\mathrm{FN}} \big(\psi(\mathbf{x}_i^{\mathrm{T}}\mathbf{w}) - \psi(\mathbf{x}_i^{\mathrm{T}}\mathbf{w}^*)\big)^2 \\ &\leq \frac{4\eta^2}{[(1-\epsilon)N]^2} \cdot \big\|\mathbf{X}_{\mathrm{FP}}\mathbf{X}_{\mathrm{FP}}^{\mathrm{T}}\big\|_2 \big\|\mathbf{X}_{\mathrm{FN}}\mathbf{X}_{\mathrm{FN}}^{\mathrm{T}}\big\|_2 \|\mathbf{w}^{(t)} - \mathbf{w}^*\|_2^2\end{aligned}$$

In the above, the first inequality follows from Lemma 30, the second inequality follows from the optimality of the Subquantile set, the third inequality follows from the AM-QM inequality (see Lemma 35), the final inequality follows from the C_2-Lipschitzness of ψ. We will now show the random variable, $\psi'(\mathbf{x}_i^{\mathrm{T}}\mathbf{w}^{(t)}) \cdot \mathbf{x}_i$ is sub-Gaussian. Let $\mathbf{v} \in \mathbb{S}^{d-1}$, then since $\mathbf{x}_i$ is sampled from sub-Gaussian distribution with sub-Gaussian norm K, we then have $\mathbf{v}^{\mathrm{T}}\mathbf{x}_i$ is sub-Gaussian with norm K. Therefore,

$$\big(\mathbb{E}|\psi'(\mathbf{x}_i^{\mathrm{T}}\mathbf{w})\mathbf{v}^{\mathrm{T}}\mathbf{x}_i|^p\big)^{1/p} \leq \sup_{x\in\mathbb{R}}|\psi'(x)|\big(\mathbb{E}|\mathbf{v}^{\mathrm{T}}\mathbf{x}_i|^{2p}\big)^{1/2p} \leq K$$

In the above, in the first inequality we used Hölder's Inequality. We therefore have, $\|\psi'(\mathbf{x}_i^{\mathrm{T}}\mathbf{w}^{(t)}) \cdot \mathbf{x}_i\|_{\psi_2} = K$. Then from Lemma 19, we have with probability at least $1-\delta$,

$$\Big\| \sum_{i\in\mathrm{TP}} \xi_i \cdot \psi'(\mathbf{x}_i^{\mathrm{T}}\mathbf{w}^{(t)}) \cdot \mathbf{x}_i \Big\|_2 \leq 9KC_\sigma d\big(\log 5 + 3N\epsilon\log\epsilon^{-1}\big)$$

Step 3: Concentration Bounds First we give the linear convergence result. Noting that $\sqrt{1-2x} \leq 1-x$ when $x \leq 1/2$, we have

$$\begin{aligned}\|\mathbf{w}^{(t+1)} - \mathbf{w}^*\|_2 \leq \|\mathbf{w}^{(t)} - \mathbf{w}^*\|_2 \bigg(1 - \frac{2C_\psi^2\eta}{(1-\epsilon)N} \cdot \lambda_{\min}(\mathbf{X}_{\mathrm{TP}}\mathbf{X}_{\mathrm{TP}}^{\mathrm{T}}) + \frac{2\eta}{(1-\epsilon)N} \cdot \|\mathbf{X}_{\mathrm{FP}}\|_2\|\mathbf{X}_{\mathrm{FN}}\|_2\bigg) \\ + \frac{18KC_\sigma d\log 5}{N} + 54KC_\sigma d\epsilon\log\epsilon^{-1}\end{aligned}$$

We will give the relevant probabilistic bounds for the random variables in Steps 1 and 2. From Lemma 24, we have $\|\mathbf{X}_{\mathrm{FN}}\|_2\|\mathbf{X}_{\mathrm{FP}}\|_2 \leq \epsilon\sqrt{\lambda_{\max}(\mathbf{\Sigma}) \cdot 10C_3N\log\epsilon^{-1}}$ with probability at least $1-\delta$ when $N \geq \frac{2}{\epsilon} \cdot \Big(dC_K^2 + \frac{\log(2/\delta)}{c_K}\Big)$ when $\epsilon \leq \frac{1}{60} \cdot \kappa^{-1}(\mathbf{\Sigma})$. From the same Lemma and under the same data conditions we have $\lambda_{\min}(\mathbf{X}_{\mathrm{TP}}\mathbf{X}_{\mathrm{TP}}^{\mathrm{T}}) \geq \frac{1}{4} \cdot \lambda_{\min}(\mathbf{\Sigma})$. Then for $\epsilon \leq \frac{C_\psi^2\lambda_{\min}(\mathbf{\Sigma})}{\sqrt{32C_3\lambda_{\max}(\mathbf{\Sigma})}}$, we have $\|\mathbf{X}_{\mathrm{FP}}\|_2\|\mathbf{X}_{\mathrm{FN}}\|_2 \leq \frac{1}{2} \cdot \lambda_{\min}(\mathbf{\Sigma})$. We then have, after $O\Big(\kappa^2(\mathbf{\Sigma})\log\Big(\frac{\|\mathbf{w}^*\|_2+10d}{\varepsilon}\Big)\Big)$ iterations with high probability,

$$\|\mathbf{w}^{(T)} - \mathbf{w}^*\|_2 \leq \varepsilon + \frac{1}{N} \cdot 144C_\psi^{-2}\kappa^2(\mathbf{\Sigma})KC_\sigma d\log 5 + 432\kappa^2(\mathbf{\Sigma})C_\psi^{-1}KC_\sigma d\epsilon\log\epsilon^{-1}$$

$$= O\Big(\kappa^2(\mathbf{\Sigma})C_\psi^{-2}KC_\sigma d\epsilon \log \epsilon^{-1}\Big)$$

In the final inequality above, we set $\varepsilon = O\Big(\kappa^2(\mathbf{\Sigma})C_\psi^{-1}KC_\sigma d\epsilon \log \epsilon^{-1}\Big)$ for $N \geq \epsilon^{-2}C_\psi^{-2}\kappa^2(\mathbf{\Sigma}) \log 5$. Our proof is complete. ■

C.1.2 Proof of Theorem 17

Proof.

■

C.2 Proof of Theorem 18

Proof. The neuron function is given as follows,

$$f_{\mathbf{w}}(\mathbf{x}) = \psi(\mathbf{x}^{\mathrm{T}}\mathbf{w}) = \mathbf{x}^{\mathrm{T}}\mathbf{w} \cdot \mathbb{I}\{\mathbf{x}^{\mathrm{T}}\mathbf{w} \geq 0\}$$

The empirical risk function is given as follows,

$$\mathcal{R}(\mathbf{w}; \mathrm{S}) = \frac{1}{(1-\epsilon)N} \cdot \sum_{i \in \mathrm{S}} (\psi(\mathbf{x}_i^{\mathrm{T}}\mathbf{w}) - y_i)^2$$

The gradient is then given as follows,

$$\frac{\partial \mathcal{R}(\mathbf{w}; \mathrm{S})}{\partial \mathbf{w}} = \frac{2}{(1-\epsilon)N} \cdot \sum_{i \in \mathrm{S}} (\psi(\mathbf{x}_i^{\mathrm{T}}\mathbf{w}) - y_i) \cdot \mathbf{x}_i \cdot \mathbb{I}\{\mathbf{x}_i^{\mathrm{T}}\mathbf{w} \geq 0\}$$

Then, from noting that the ReLU function is non-differentiable at one point, then almost surely the Hessian is given as

$$\frac{\partial^2 \mathcal{R}(\mathbf{w}; \mathrm{S})}{\partial \mathbf{w}^2} = \frac{2}{(1-\epsilon)N} \cdot \sum_{i \in \mathrm{S}} \mathbf{x}_i\mathbf{x}_i^{\mathrm{T}} \cdot \mathbb{I}\{\mathbf{x}_i^{\mathrm{T}}\mathbf{w} \geq 0\}$$

We will now begin our standard analysis.

$$\begin{aligned}
\|\mathbf{w}^{(t+1)} - \mathbf{w}^*\|_2 &= \|\mathbf{w}^{(t)} - \mathbf{w}^* - \eta \nabla \mathcal{R}(\mathbf{w}; \mathrm{S}^{(t)})\|_2 \\
&= \|\mathbf{w}^{(t)} - \mathbf{w}^* - \eta \nabla \mathcal{R}(\mathbf{w}^{(t)}; \mathrm{TP}) - \eta \nabla \mathcal{R}(\mathbf{w}^{(t)}; \mathrm{FP})\|_2 \\
&\leq \|\mathbf{w}^{(t)} - \mathbf{w}^* - \eta \nabla \mathcal{R}(\mathbf{w}^{(t)}; \mathrm{TP})\|_2 + \|\eta \mathcal{R}(\mathbf{w}^{(t)}; \mathrm{FP})\|_2 \qquad \text{(C.5)}
\end{aligned}$$

We will now upper bound the first term of Equation C.5 through its square,

$$\begin{aligned}
&\|\mathbf{w}^{(t)} - \mathbf{w}^* - \eta \nabla \mathcal{R}(\mathbf{w}^{(t)}; \mathrm{TP})\|_2^2 = \\
&\qquad \|\mathbf{w}^{(t)} - \mathbf{w}^*\|_2^2 - 2\eta \cdot \langle \mathbf{w}^{(t)} - \mathbf{w}^*, \nabla \mathcal{R}(\mathbf{w}^{(t)}; \mathrm{TP}) \rangle + \eta^2 \cdot \|\nabla \mathcal{R}(\mathbf{w}^{(t)}; \mathrm{TP})\|_2^2 \qquad \text{(C.6)}
\end{aligned}$$

We will lower bound the second term of Equation C.6. We will first adopt the notation from [ZYWG19], let $\mathbf{\Sigma}_{\mathrm{TP}}(\mathbf{w}, \hat{\mathbf{w}}) = \mathbf{X}_{\mathrm{TP}}^{\mathrm{T}}\mathbf{X}_{\mathrm{TP}}^{\mathrm{T}} \cdot \mathbb{I}\{\mathbf{X}_{\mathrm{TP}}^{\mathrm{T}}\mathbf{w} \geq \mathbf{0}\} \cdot \mathbb{I}\{\mathbf{X}_{\mathrm{TP}}^{\mathrm{T}}\hat{\mathbf{w}} \geq \mathbf{0}\}$, it then follows

$$\begin{aligned}
&2\eta \cdot \langle \mathbf{w}^{(t)} - \mathbf{w}^*, \nabla \mathcal{R}(\mathbf{w}^{(t)}; \mathrm{TP}) \rangle \\
&\quad \stackrel{\text{def}}{=} \frac{4\eta}{(1-\epsilon)N} \cdot \langle \mathbf{w}^{(t)} - \mathbf{w}^*, \sum_{i \in \mathrm{TP}} (\psi(\mathbf{x}_i^{\mathrm{T}}\mathbf{w}^{(t)}) - y_i) \cdot \mathbf{x}_i \cdot \mathbb{I}\{\mathbf{x}_i^{\mathrm{T}}\mathbf{w}^{(t)} \geq 0\} \rangle \\
&\quad = \frac{4\eta}{(1-\epsilon)N} \cdot \langle \mathbf{w}^{(t)} - \mathbf{w}^*, \mathbf{\Sigma}_{\mathrm{TP}}(\mathbf{w}^{(t)}, \mathbf{w}^{(t)})\mathbf{w}^{(t)} - \mathbf{\Sigma}_{\mathrm{TP}}(\mathbf{w}^{(t)}, \mathbf{w}^*)\mathbf{w}^* \rangle \\
&\qquad + \frac{4\eta}{(1-\epsilon)N} \cdot \langle \mathbf{w}^{(t)} - \mathbf{w}^*, \sum_{i \in \mathrm{TP}} \xi_i \mathbf{x}_i \cdot \mathbb{I}\{\mathbf{x}_i^{\mathrm{T}}\mathbf{w}^{(t)} \geq 0\} \rangle
\end{aligned}$$

$$\geq \frac{4\eta}{(1-\epsilon)N} \cdot \lambda_{\min}\Big(\mathbf{\Sigma}_{\mathrm{TP}}(\mathbf{w}^{(t)}, \mathbf{w}^*)\Big) \|\mathbf{w}^{(t)} - \mathbf{w}^*\|_2^2$$
$$- \frac{4\eta}{(1-\epsilon)N} \cdot \|\mathbf{w}^{(t)} - \mathbf{w}^*\|_2 \big\| \sum_{i \in \mathrm{TP}} \xi_i \mathbf{x}_i \cdot \mathbb{I}\{\mathbf{x}_i^{\mathrm{T}} \mathbf{w}^{(t)} \geq 0\} \big\|_2 \tag{C.7}$$

In the above, the final inequality holds because $\mathbf{x}_i^{\mathrm{T}} \mathbf{w}^{(t)} \geq \mathbf{x}_i^{\mathrm{T}} \mathbf{w}^*$, therefore we have that $(\mathbf{x}_i^{\mathrm{T}} \mathbf{w}^{(t)} - \mathbf{x}_i^{\mathrm{T}} \mathbf{w}^*)(\mathbf{x}_i^{\mathrm{T}} \mathbf{w}^{(t)}) \geq 0$. From Weyl's Inequality, we have

$$\lambda_{\min}\Big(\mathbf{\Sigma}_{\mathrm{TP}}(\mathbf{w}^{(t)}, \mathbf{w}^*)\Big) \geq \lambda_{\min}(\mathbf{\Sigma}_{\mathrm{TP}}(\mathbf{w}^*, \mathbf{w}^*)) - \big\|\mathbf{\Sigma}_{\mathrm{TP}}(\mathbf{w}^*, \mathbf{w}^{(t)}) - \mathbf{\Sigma}_{\mathrm{TP}}(\mathbf{w}^*, \mathbf{w}^*)\big\|_2$$

We bound the two terms individually. We bound the first term now. Since $\|\mathbf{w}^*\|$ is fixed and non-zero, and $\mathcal{P}$ is symmetric, we have that $\Pr\{\mathbf{x}^{\mathrm{T}} \mathbf{w}^* \geq 0\} = \frac{1}{2}$ for $\mathbf{x} \sim \mathcal{P}$. Let B be the sum of $(1-\epsilon)N$ i.i.d Bernoulli random variables. Then from Hoeffding's Inequality, we have that

$$\Pr\{B - \mathbb{E}\, B \leq t\} \leq 2e^{-\frac{t^2}{(1-\epsilon)N}}$$

From which by elementary manipulations, we obtain with probability exceeding $1-\delta$,

$$B \geq \frac{1}{2} \cdot N - \frac{1}{2}\sqrt{N \log(2/\delta)} \geq \frac{1}{4} \cdot N$$

In the above, the final inequality holds when $N \geq 64 \log(2/\delta)$ with probability at least $1-\delta$. We then obtain with sufficient data that,

$$\lambda_{\min}(\mathbf{\Sigma}_{\mathrm{TP}}(\mathbf{w}^*, \mathbf{w}^*)) \geq \frac{n}{8} \cdot \lambda_{\min}(\mathbf{\Sigma})$$

We now upper bound the second term.

$$\begin{aligned}
\big\|\mathbf{\Sigma}_{\mathrm{TP}}(\mathbf{w}^*, \mathbf{w}^*) - \mathbf{\Sigma}_{\mathrm{TP}}(\mathbf{w}^*, \mathbf{w}^{(t)})\big\|_2 &= \big\| \sum_{i \in \mathrm{TP}} \mathbf{x}_i \mathbf{x}_i^{\mathrm{T}} \cdot \mathbb{I}\{\mathbf{x}_i^{\mathrm{T}} \mathbf{w}^* \geq 0\} \cdot \mathbb{I}\{\mathbf{x}_i^{\mathrm{T}} \mathbf{w}^{(t)} \leq 0\} \big\|_2 \\
&\leq \big\| \sum_{i \in \mathrm{P}} \mathbf{x}_i \mathbf{x}_i^{\mathrm{T}} \cdot \mathbb{I}\{\mathbf{x}_i^{\mathrm{T}} \mathbf{w}^* \geq 0\} \cdot \mathbb{I}\{\mathbf{x}_i^{\mathrm{T}} \mathbf{w}^{(t)} \leq 0\} \big\|_2 \\
&\leq (1/\delta) \cdot \mathbb{E}\left[\sum_{i \in \mathrm{P}} \|\mathbf{x}_i \mathbf{x}_i^{\mathrm{T}}\|_2 \cdot \mathbb{I}\{\mathbf{x}_i^{\mathrm{T}} \mathbf{w}^* \geq 0\} \cdot \mathbb{I}\{\mathbf{x}_i^{\mathrm{T}} \mathbf{w}^{(t)} \leq 0\} \right] \\
&\leq (1/\delta) \cdot \sum_{i \in \mathrm{P}} \big(\mathbb{E}\big[\|\mathbf{x}_i \mathbf{x}_i^{\mathrm{T}}\|_2^2\big]\big)^{1/2} \cdot \mathbb{E}\Big[\mathbb{I}\{\mathbf{x}_i^{\mathrm{T}} \mathbf{w}^* \geq 0\} \cdot \mathbb{I}\{\mathbf{x}_i^{\mathrm{T}} \mathbf{w}^{(t)} \leq 0\}\Big] \\
&\leq \frac{(1/\delta)}{2\pi} \cdot \sqrt{d}\, N \Theta(\mathbf{w}^{(t)}, \mathbf{w}^*)
\end{aligned}$$

In the above, the first inequality follows from Weyl's Inequality, and the second inequality follows from Markov's inequality to a triangle inequality and finally applying Lemma 20. Then, from our assumption that $\Theta(\mathbf{w}^{(t)}, \mathbf{w}^*) \leq \frac{\pi\delta}{8\sqrt{d}}$, we have

$$\lambda_{\min}\Big(\mathbf{\Sigma}_{\mathrm{TP}}(\mathbf{w}^{(t)}; \mathbf{w}^*)\Big) \geq \frac{N}{16} \cdot \lambda_{\min}(\mathbf{\Sigma})$$

Noise Bound: From Lemma 19, we can bound the noise with probability exceeding $1-\delta$, we have,

$$\|\boldsymbol{\xi}_{\mathrm{TP}}^{\mathrm{T}} \widetilde{\mathbf{X}}_{\mathrm{TP}}\|_2 \leq \frac{5}{2} \cdot K C_\sigma d\big(\log 5 + 3N\epsilon \log \epsilon^{-1}\big)$$

Step 2: Upper bounding the norm gradient.

$$\eta^2 \cdot \|\nabla \mathcal{R}(\mathbf{w}^{(t)}; \mathrm{TP})\|_2^2 = \frac{4\eta^2}{[(1-\epsilon)N]^2} \cdot \big\| \sum_{i \in \mathrm{TP}} (\psi(\mathbf{x}_i^{\mathrm{T}} \mathbf{w}^{(t)}) - \psi(\mathbf{x}_i^{\mathrm{T}} \mathbf{w}^*) - \xi_i) \cdot \mathbf{x}_i \cdot \mathbb{I}\{\mathbf{x}_i^{\mathrm{T}} \mathbf{w}^{(t)} \geq 0\} \big\|_2^2$$

$$\leq \frac{8\eta^2}{[(1-\epsilon)N]^2} \cdot \left(\Big\| \sum_{i \in \mathrm{TP}} (\psi(\mathbf{x}_i^{\mathrm{T}}\mathbf{w}^{(t)}) - \psi(\mathbf{x}_i^{\mathrm{T}}\mathbf{w}^*)) \cdot \mathbf{x}_i \cdot \mathbb{I}\{\mathbf{x}_i^{\mathrm{T}}\mathbf{w}^{(t)} \geq 0\} \Big\|_2^2 + \Big\| \sum_{i \in \mathrm{TP}} \xi_i \mathbf{x}_i \cdot \mathbb{I}\{\mathbf{x}_i^{\mathrm{T}}\mathbf{w}^{(t)} \geq 0\} \Big\|_2^2 \right) \tag{C.8}$$

Recall we have from Lemma 19, we have an upper bound on the second term of Equation (C.8). We give a bound on the first term of Equation (C.8).

$$\begin{aligned}
&\frac{8\eta^2}{[(1-\epsilon)N]^2} \cdot \Big\| \sum_{i \in \mathrm{TP}} (\psi(\mathbf{x}_i^{\mathrm{T}}\mathbf{w}^{(t)}) - \psi(\mathbf{x}_i^{\mathrm{T}}\mathbf{w}^*)) \cdot \mathbf{x}_i \cdot \mathbb{I}\{\mathbf{x}_i^{\mathrm{T}}\mathbf{w}^{(t)} \geq 0\} \Big\|_2^2 \\
&\qquad \leq \frac{8\eta^2}{[(1-\epsilon)N]^2} \cdot \big\|\mathbf{X}_{\mathrm{TP}}\mathbf{X}_{\mathrm{TP}}^{\mathrm{T}}\big\|_2 \sum_{i \in \mathrm{TP}} (\psi(\mathbf{x}_i^{\mathrm{T}}\mathbf{w}^{(t)}) - \psi(\mathbf{x}_i^{\mathrm{T}}\mathbf{w}^*))^2 \\
&\qquad \leq \frac{8\eta^2}{[(1-\epsilon)N]^2} \cdot \big\|\mathbf{X}_{\mathrm{TP}}\mathbf{X}_{\mathrm{TP}}^{\mathrm{T}}\big\|_2^2 \big\|\mathbf{w}^{(t)} - \mathbf{w}^*\big\|_2^2
\end{aligned} \tag{C.9}$$

Then, by choosing $\eta \leq \frac{\lambda_{\min}(\mathbf{\Sigma})}{80N\lambda_{\max}^2(\mathbf{\Sigma})}$, we have that the RHS in Equation C.9 will be less than $\frac{\lambda_{\min}(\mathbf{\Sigma})}{8}$.

Step 3: Upper bounding the corrupted gradient. We now upper bound the third term in Equation (C.6).

$$\begin{aligned}
\eta^2 \cdot \|\nabla \mathcal{R}(\mathbf{w}^{(t)}; \mathrm{FP})\|_{\mathrm{F}}^2 &= \frac{4\eta^2}{[(1-\epsilon)N]^2} \cdot \Big\| \sum_{i \in \mathrm{FP}} (\psi(\mathbf{x}_i^{\mathrm{T}}\mathbf{w}^{(t)}) - \psi(\mathbf{x}_i^{\mathrm{T}}\mathbf{w}^*)) \cdot \mathbf{x}_i \cdot \mathbb{I}\{\mathbf{x}_i^{\mathrm{T}}\mathbf{w}^{(t)} \geq 0\} \Big\|_2^2 \\
&\leq \frac{4\eta^2}{[(1-\epsilon)N]^2} \cdot \|\mathbf{\Sigma}_{\mathrm{FP}}(\mathbf{w}^{(t)}, \mathbf{w}^{(t)})\|_2 \sum_{i \in \mathrm{FP}} (\psi(\mathbf{x}_i^{\mathrm{T}}\mathbf{w}^{(t)}) - \psi(\mathbf{x}_i^{\mathrm{T}}\mathbf{w}^*))^2 \\
&\leq \frac{4\eta^2}{[(1-\epsilon)N]^2} \cdot \|\mathbf{\Sigma}_{\mathrm{FP}}(\mathbf{w}^{(t)}, \mathbf{w}^{(t)})\|_2 \sum_{i \in \mathrm{FN}} (\psi(\mathbf{x}_i^{\mathrm{T}}\mathbf{w}^{(t)}) - \psi(\mathbf{x}_i^{\mathrm{T}}\mathbf{w}^*))^2 \\
&\leq \frac{4\eta^2}{[(1-\epsilon)N]^2} \cdot \big\|\mathbf{X}_{\mathrm{FP}}\mathbf{X}_{\mathrm{FP}}^{\mathrm{T}}\big\|_2 \big\|\mathbf{X}_{\mathrm{FN}}\mathbf{X}_{\mathrm{FN}}^{\mathrm{T}}\big\|_2 \|\mathbf{w}^{(t)} - \mathbf{w}^*\|_2^2
\end{aligned}$$

In the above, the first inequality follows from the same argument as Lemma 31, the second inequality follows from the optimality of the Subquantile set, and the final inequality follows from noting that ψ is 1-Lipschitz. ■

Lemma 19 (Noise Bound). *Fix $\mathbf{w} \in \mathbb{R}^d$. Let $\widetilde{\mathbf{X}} = [\mathbf{x}_1 \cdot \mathbb{I}\{\mathbf{x}_1^{\mathrm{T}}\mathbf{w} \geq 0\}, \ldots, \mathbf{x}_{(1-\epsilon)N} \cdot \mathbb{I}\{\mathbf{x}_n^{\mathrm{T}}\mathbf{w} \geq 0\}]$ be the data matrix such that for $i \in [(1-\epsilon)N]$, $\mathbf{x}_i$ are sampled from a sub-Gaussian distribution with second-moment matrix $\mathbf{\Sigma}$ with sub-Gaussian norm K. Suppose for $i \in [(1-\epsilon)N]$, ξ_i are sampled from sub-Gaussian distribution with sub-Gaussian norm C_σ. Let $\mathcal{S}$ represent all subset of $[N]$ of size $(1-2\epsilon)N$ to $(1-\epsilon)N$. Then, with probability at least $1-\delta$,*

$$\max_{S \in \mathcal{S}} \cdot \|\boldsymbol{\xi}_S^{\mathrm{T}}\widetilde{\mathbf{X}}_S\|_2 \leq \frac{4}{3} \cdot NKC_\sigma \cdot \left(\frac{2d}{N} \log(5) + \frac{2}{N} \log(1/\delta) + 6\epsilon \log \epsilon^{-1} \right)^{1/2}$$

Proof. We will use the following characterization of the spectral norm.

$$\Big\| \sum_{i \in \mathrm{TP}} \xi_i \mathbf{x}_i \cdot \mathbb{I}\{\mathbf{x}_i^{\mathrm{T}}\mathbf{w}^{(t)} \geq 0\} \Big\|_2 = \max_{\mathbf{v} \in \mathbb{S}^{d-1}} \sum_{i \in \mathrm{TP}} \xi_i \mathbf{x}_i^{\mathrm{T}}\mathbf{v} \cdot \mathbb{I}\{\mathbf{x}_i^{\mathrm{T}}\mathbf{w}^{(t)} \geq 0\}$$

Our proof uses standard ideas in High-Dimensional Probability and will follow similarly to [ZYWG19]. Let $\widetilde{\mathbf{x}} = \mathbf{x} \cdot \mathbb{I}\{\mathbf{x}^{\mathrm{T}}\mathbf{w}\}$, then we can note that $\widetilde{\mathbf{x}}$ is also sub-Gaussian with second-moment matrix $\mathbf{\Sigma}$. It thus follows for any $\mathbf{v} \in \mathbb{S}^{d-1}$, the random variable, $\widetilde{\mathbf{x}}_i^{\mathrm{T}}\mathbf{v}$ is sub-Gaussian by definition s.t. $\|\widetilde{\mathbf{x}}_i^{\mathrm{T}}\mathbf{v}\|_{\psi_2} = K$. Let $C_\sigma \triangleq \|\xi_i\|_{\psi_2}$, then from Lemma 2.7.6 in [Ver20], the random variable $\xi_i\widetilde{\mathbf{x}}_i^{\mathrm{T}}\mathbf{v}$ is sub-exponential s.t. $\|\xi_i\widetilde{\mathbf{x}}_i^{\mathrm{T}}\mathbf{v}\|_{\psi_1} \leq KC_\sigma$. We will now use a ε-covering argument. Let $\mathcal{C}$ be a ε-net of $\mathbb{S}^{d-1}$ such that for any $\mathbf{v} \in$

$\mathbb{S}^{d-1}$, there exists $\mathbf{u} \in \mathcal{C}$ such that $\|\mathbf{u}-\mathbf{v}\|_2 \leq \varepsilon$. Let $\mathbf{u}^* = \arg\max_{\mathbf{u}\in\mathcal{C}} \boldsymbol{\xi}^{\mathrm{T}}\widetilde{\mathbf{X}}\mathbf{u}$ and $\mathbf{v}^* = \arg\max_{\mathbf{v}\in\mathbb{S}^{d-1}} \boldsymbol{\xi}^{\mathrm{T}}\widetilde{\mathbf{X}}\mathbf{v}$. We then have,

$$|\boldsymbol{\xi}^{\mathrm{T}}\widetilde{\mathbf{X}}\mathbf{v}^* - \boldsymbol{\xi}^{\mathrm{T}}\widetilde{\mathbf{X}}\mathbf{u}^*| \leq \left\|\boldsymbol{\xi}^{\mathrm{T}}\widetilde{\mathbf{X}}\right\|_2 \|\mathbf{u}^* - \mathbf{v}^*\|_2 \leq \varepsilon \cdot \|\boldsymbol{\xi}^{\mathrm{T}}\widetilde{\mathbf{X}}\|_2$$

where in the final inequality we use the definition of a ε-net. We then have,

$$|\boldsymbol{\xi}^{\mathrm{T}}\widetilde{\mathbf{X}}\mathbf{u}^*| \geq |\boldsymbol{\xi}^{\mathrm{T}}\mathbf{X}\mathbf{v}^*| - |\boldsymbol{\xi}\widetilde{\mathbf{X}}\mathbf{u}^* - \boldsymbol{\xi}\widetilde{\mathbf{X}}\mathbf{v}^*| \geq (1-\varepsilon)|\boldsymbol{\xi}^{\mathrm{T}}\widetilde{\mathbf{X}}\mathbf{v}^*|$$

Then, this implies that

$$|\boldsymbol{\xi}^{\mathrm{T}}\widetilde{\mathbf{X}}\mathbf{v}^*| \leq \frac{1}{1-\varepsilon} \cdot |\boldsymbol{\xi}^{\mathrm{T}}\widetilde{\mathbf{X}}\mathbf{v}^*|$$

With this result we are ready to make the probabilistic bounds,

$$\Pr\left\{\frac{1}{N}\cdot\|\boldsymbol{\xi}^{\mathrm{T}}\widetilde{\mathbf{X}}\| \geq t\right\} \leq \frac{1}{1-\varepsilon}\cdot\Pr\left\{\frac{1}{N}\cdot\max_{\mathbf{u}\in\mathcal{C}}|\boldsymbol{\xi}^{\mathrm{T}}\widetilde{\mathbf{X}}\mathbf{u}| \geq t\right\} \leq \frac{1}{1-\varepsilon}\cdot\sum_{i\in|\mathcal{C}|}\Pr\left\{\frac{1}{N}\cdot|\boldsymbol{\xi}^{\mathrm{T}}\widetilde{\mathbf{X}}\mathbf{u}_i| \geq t\right\}$$

Suppose $\mathcal{C}$ is a 1/4-net. We can now apply Bernstein's Inequality from [RH23] Lemma 1.13.

$$\frac{1}{1-\varepsilon}\cdot\sum_{i\in|\mathcal{C}|}\Pr\left\{\frac{1}{N}\cdot|\boldsymbol{\xi}^{\mathrm{T}}\widetilde{\mathbf{X}}\mathbf{u}_i| \geq t\right\} \leq 5^d\cdot\Pr\left\{\frac{1}{N}\cdot|\boldsymbol{\xi}^{\mathrm{T}}\widetilde{\mathbf{X}}\mathbf{u}| \geq \frac{3}{4}\cdot t\right\} \leq 5^d\cdot\exp\left[-\frac{3N}{4}\left(\frac{t^2}{\frac{4}{3}\cdot K^2C_\sigma^2}\wedge\frac{t}{\frac{4}{3}\cdot KC_\sigma}\right)\right] \leq \delta$$

The probability holds when $N \geq 2d\log 5 + 2\log(1/\delta)$ for

$$t \geq \frac{4}{3}\cdot KC_\sigma\cdot\left(\frac{2d}{N}\log(5)+\frac{2}{N}\log(1/\delta)\right)^{1/2}$$

Let $\mathcal{S}$ represent all subset of $[(1-\epsilon)N]$ of size $(1-2\epsilon)N$ to $(1-\epsilon)N$. We then have from a union bound over $\mathcal{S}$, with probability exceeding $1-\delta$,

$$\max_{S\in\mathcal{S}}\cdot\|\boldsymbol{\xi}_S^{\mathrm{T}}\widetilde{\mathbf{X}}_S\|_2 \leq \frac{4}{3}\cdot NKC_\sigma\cdot\left(\frac{2d}{N}\log(5)+\frac{2}{N}\log(1/\delta)+6\epsilon\log\epsilon^{-1}\right)^{1/2}$$

where in the above we use the fact that $\log\binom{N}{(1-\epsilon)N} = \log\binom{N}{\epsilon N}$. Our proof is complete. ■

Lemma 20. *Fix $\mathbf{w}_1, \mathbf{w}_2 \in \mathbb{S}^d$. Let $\mathbf{x}$ be sampled from $\mathcal{N}(\mathbf{0}, \mathbf{I})$. Then,*

$$\Pr\{\mathbf{x}^{\mathrm{T}}\mathbf{w}_1 \geq 0, \mathbf{x}^{\mathrm{T}}\mathbf{w}_2 \geq 0\} = \frac{\Theta(\mathbf{w}_1,\mathbf{w}_2)}{2\pi}$$

Proof. We can note that $\mathbf{x}^{\mathsf{T}}\mathbf{w}_1$ and $\mathbf{x}^{\mathsf{T}}\mathbf{w}_2$ are Gaussian variables with distribution $\mathcal{N}(0,1)$. The covariance of the variables can be calculated simply,

$$\mathrm{Var}(\mathbf{x}^{\mathsf{T}}\mathbf{w}_1, \mathbf{x}^{\mathsf{T}}\mathbf{w}_2) = \mathbb{E}[\mathbf{w}_1^{\mathsf{T}}\mathbf{x}\mathbf{x}^{\mathsf{T}}\mathbf{w}_2] = \mathrm{Tr}(\mathbb{E}[\mathbf{w}_1^{\mathsf{T}}\mathbf{x}\mathbf{x}^{\mathsf{T}}\mathbf{w}_2]) = \mathbf{w}_1^{\mathrm{T}}\mathbf{w}_2$$

Define $\boldsymbol{\Gamma} \triangleq \begin{bmatrix} 1 & \cos\Theta(\mathbf{w}_1,\mathbf{w}_2) \\ \cos\Theta(\mathbf{w}_1,\mathbf{w}_2) & 1 \end{bmatrix}$. Let $\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Gamma})$, then we have

$$\begin{aligned}\Pr\{x_1 \geq 0, x_2 \geq 0\} &= \int_0^\infty\int_0^\infty \frac{1}{2\pi}\csc\Theta(\mathbf{w}_1,\mathbf{w}_2)\exp\left[-\frac{1}{2}\cdot\mathbf{x}^{\mathrm{T}}\boldsymbol{\Gamma}^{-1}\mathbf{x}\right]dx_1dx_2 \\ &= \frac{1}{2\pi}\cdot\arccos(\cos\Theta(\mathbf{w}_1,\mathbf{w}_2)) = \frac{\Theta(\mathbf{w}_1,\mathbf{w}_2)}{2\pi}\end{aligned}$$

The proof is complete. ■

D Probability Theory

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

D.1 Finite Dimensional Concentrations of Measure

Lemma 21 (Upper Bound on Sum of Chi-Squared Variables [LM00]). *Suppose $\xi_i \sim \mathcal{N}(0, \sigma^2)$ for $i \in [n]$, then*

$$\mathbb{P}\{\|\boldsymbol{\xi}\|_2^2 \geq \sigma\big(n + 2\sqrt{nx} + 2x\big)\} \leq e^{-x}$$

Proposition 22 (Probabilistic Upper Bound on Sum of Chi-Squared Variables). *Suppose $\xi_i \sim \mathcal{N}(0, \sigma^2)$ for $i \in [n]$. Let $S \subset [n]$ such that $|S| = \epsilon n$ for $\epsilon \in (0, 1]$ and let $\mathcal{C}$ represent all such subsets. Given a failure probability $\delta \in (0, 1)$, when $n \geq \log(1/\delta)$, with probability exceeding $1 - \delta$,*

$$\max_{S \in \mathcal{C}} \|\boldsymbol{\xi}_S\|_2^2 \leq \sigma\big(30 n\epsilon \log \epsilon^{-1}\big)$$

Proof. Directly from Lemma 21, we have with probability exceeding $1 - \delta$.

$$\|\boldsymbol{\xi}\|_2^2 \leq \sigma\Big(n + 2\sqrt{n \log(1/\delta)} + 2\log(1/\delta)\Big)$$

We now can prove the claimed bound using the layer-cake representation,

$$\mathbb{P}\left\{\max_{S \in \mathcal{C}} \|\boldsymbol{\xi}\|_2^2 \geq \sigma\big(\epsilon n + 2\sqrt{\epsilon n x} + 2x\big)\right\} \leq \binom{e}{\epsilon}^{\epsilon n} \mathbb{P}\{\|\boldsymbol{\xi}\|_2^2 \geq \sigma\big(\epsilon n + 2\sqrt{\epsilon n x} + 2x\big)\} \leq \binom{e}{\epsilon}^{\epsilon n} e^{-x}$$

In the first inequality we apply a union bound over $\mathcal{C}$ with Lemma 38, and in the second inequality we use Lemma 21. We then obtain with probability exceeding $1 - \delta$,

$$\begin{aligned}
\max_{S \in \mathcal{C}} \|\boldsymbol{\xi}_S\|_2^2 &\leq \sigma\Big(\epsilon n + 2\sqrt{n\epsilon \log(1/\delta) + 3n^2\epsilon^2 \log \epsilon^{-1}} + 2\log(1/\delta) + 6n\epsilon \log \epsilon^{-1}\Big) \\
&\leq \sigma\Big(9n\epsilon \log \epsilon^{-1} + 2\sqrt{n\epsilon \log(1/\delta)} + 2\sqrt{3}\, n\epsilon \sqrt{\log \epsilon^{-1}} + 2\log(1/\delta)\Big) \\
&\leq \sigma\Big(15 n\epsilon \log \epsilon^{-1} + 2\sqrt{n\epsilon \log(1/\delta)} + 2\log(1/\delta)\Big) \\
&\leq \sigma\big(30 n\epsilon \log \epsilon^{-1}\big)
\end{aligned}$$

In the above, in the first inequality, we note that $\log \binom{n}{\epsilon n} \leq 3n\epsilon \log \epsilon^{-1}$ as $\epsilon < 0.5$, in the second inequality we note that $\sqrt{\log \epsilon^{-1}} \leq (\log(2))^{-1/2} \log \epsilon^{-1} \leq \sqrt{3} \log \epsilon^{-1}$ when $\epsilon < 0.5$, the final inequality holds when $n \geq \log(1/\delta)$ by solving for the quadratic equation. The proof is complete. ■

Lemma 23 (Sub-Gaussian Covariance Matrix Estimation [Ver10] Theorem 5.40). *Let $\mathbf{X} \in \mathbb{R}^{d \times n}$ have columns sampled from a sub-Gaussian distribution with sub-Gaussian norm K and second-moment matrix $\boldsymbol{\Sigma}$, then there exists positive constants c_k, C_K, dependent on the sub-Gaussian norm such that with probability at least $1 - 2e^{-c_K t^2}$,*

$$\lambda_{\max}(\mathbf{X}\mathbf{X}^{\mathrm{T}}) \leq n \cdot \lambda_{\max}(\boldsymbol{\Sigma}) + \lambda_{\max}(\boldsymbol{\Sigma}) \cdot \Big(C_K \cdot \sqrt{dn} + t \cdot \sqrt{n}\Big)$$

Lemma 24. *Let $\mathbf{X} \in \mathbb{R}^{d \times n}$ have columns sampled from a sub-Gaussian distribution with sub-Gaussian norm K and second-moment matrix $\boldsymbol{\Sigma}$. Let $S \subset [n]$ such that $|S| = \epsilon n$ for $\epsilon \in (0, 0.5)$ and let $\mathcal{C}$ represent all such subsets. Fix a constant $\zeta \in (0, 1)$, then with probability at least $1 - \delta$,*

$$\begin{aligned}
\max_{S \in \mathcal{C}} \lambda_{\max}(\mathbf{X}_S \mathbf{X}_S^{\mathrm{T}}) &\leq \lambda_{\max}(\boldsymbol{\Sigma}) \cdot \big(10 n\epsilon \log \epsilon^{-1}\big) \\
\min_{S \in \mathcal{C}} \lambda_{\min}(\mathbf{X}_{X \setminus S} \mathbf{X}_{X \setminus S}^{\mathrm{T}}) &\geq \frac{n}{4} \cdot \lambda_{\min}(\boldsymbol{\Sigma})
\end{aligned}$$

when

$$n \geq \frac{2}{\epsilon} \cdot \left(C_K^2 \cdot d + \frac{\log(2/\delta)}{c_K} \right)$$

and $\epsilon \leq \frac{1}{60} \cdot \kappa^{-1}(\mathbf{\Sigma})$.

Proof. We will use the layer-cake representation to obtain our claimed error bound.

$$\begin{aligned}
&\mathbb{P}\Big\{ \max_{S \in \mathcal{C}} \lambda_{\max}(\mathbf{X}_S \mathbf{X}_S^{\mathrm{T}}) \geq n\epsilon \cdot \lambda_{\max}(\mathbf{\Sigma}) + \lambda_{\max}(\mathbf{\Sigma}) \cdot \left(C_K \cdot \sqrt{dn\epsilon} + t\sqrt{n\epsilon} \right) \Big\} \\
&\quad \leq \left(\frac{e}{\epsilon} \right)^{\epsilon n} \mathbb{P}\Big\{ \lambda_{\max}(\mathbf{X}_S \mathbf{X}_S^{\mathrm{T}}) \geq n\epsilon \cdot \lambda_{\max}(\mathbf{\Sigma}) + \lambda_{\max}(\mathbf{\Sigma}) \cdot \left(C_K \cdot \sqrt{dn\epsilon} + t\sqrt{n\epsilon} \right) \Big\} \leq 2 \cdot \left(\frac{e}{\epsilon} \right)^{\epsilon n} e^{-c_K t^2}
\end{aligned}$$

In the above, the first inequality follows from a union bound over $\mathcal{C}$ and Lemma 38, the second inequality follows from Lemma 23. Then from elementary inequalities, we obtain with probability $1 - \delta$,

$$\begin{aligned}
\max_{S \in \mathcal{C}} \lambda_{\max}(\mathbf{X}_S \mathbf{X}_S^{\mathrm{T}}) &\leq n\epsilon \cdot \lambda_{\max}(\mathbf{\Sigma}) + \lambda_{\max}(\mathbf{\Sigma}) \cdot \left(C_K \cdot \sqrt{dn\epsilon} + \sqrt{\frac{1}{c_K}(n\epsilon \cdot \log(2/\delta) + 3n^2\epsilon^2 \log \epsilon^{-1})} \right) \\
&\leq n \cdot \lambda_{\max}(\mathbf{\Sigma}) \cdot (\epsilon + 3^{3/4}\epsilon \log \epsilon^{-1}) + \lambda_{\max}(\mathbf{\Sigma}) \cdot \left(C_K \cdot \sqrt{dn\epsilon} + \sqrt{\frac{1}{c_K} n\epsilon \cdot \log(2/\delta)} \right) \\
&\leq \lambda_{\max}(\mathbf{\Sigma}) \cdot \left(6n\epsilon \log \epsilon^{-1} \right) + \lambda_{\max}(\mathbf{\Sigma}) \cdot \left(C_K \cdot \sqrt{dn\epsilon} + \sqrt{\frac{1}{c_K} n\epsilon \cdot \log(2/\delta)} \right) \\
&\leq \lambda_{\max}(\mathbf{\Sigma}) \cdot \left(10n\epsilon \log \epsilon^{-1} \right)
\end{aligned}$$

In the above, the last inequality holds when

$$n \geq \frac{2}{\epsilon} \cdot \left(C_k^2 \cdot d + \frac{\log(2/\delta)}{c_K} \right)$$

and our proof of the upper bound for the maximal eigenvalue is complete. We have from Weyl's Inequality for any $S \in \mathcal{C}$,

$$\lambda_{\min}(\mathbf{X}_{\mathrm{X} \setminus \mathrm{S}} \mathbf{X}_{\mathrm{X} \setminus \mathrm{S}}^{\mathrm{T}}) = \lambda_{\min}(\mathbf{X}\mathbf{X}^{\mathrm{T}} - \mathbf{X}_{\mathrm{S}} \mathbf{X}_{\mathrm{S}}^{\mathrm{T}}) \geq \lambda_{\min}(\mathbf{X}\mathbf{X}^{\mathrm{T}}) - \lambda_{\max}(\mathbf{X}_{\mathrm{S}} \mathbf{X}_{\mathrm{S}}^{\mathrm{T}})$$

We then have with probability at least $1 - \delta$,

$$\begin{aligned}
\lambda_{\min}(\mathbf{X}_{\mathrm{X} \setminus \mathrm{S}} \mathbf{X}_{\mathrm{X} \setminus \mathrm{S}}^{\mathrm{T}}) &\geq n \cdot \lambda_{\min}(\mathbf{\Sigma}) - C_K \cdot \sqrt{dn} - \sqrt{\frac{1}{c_K} \cdot n \cdot \log(2/\delta)} - \lambda_{\max}(\mathbf{\Sigma}) \cdot (10n\epsilon \log \epsilon^{-1}) \\
&\geq \frac{n}{2} \cdot \lambda_{\min}(\mathbf{\Sigma}) - \lambda_{\max}(\mathbf{\Sigma}) \cdot (10n\epsilon \log \epsilon^{-1}) \geq \frac{n}{4} \cdot \lambda_{\min}(\mathbf{\Sigma})
\end{aligned}$$

In the above, the first inequality follows when $n \geq \frac{1}{\lambda_{\min}^2(\mathbf{\Sigma})} \Big(8C_k \cdot d + \frac{2}{c_K} \cdot \log(2/\delta) \Big)$, and the last inequality follows when $\epsilon \leq \frac{1}{60} \cdot \kappa^{-1}(\mathbf{\Sigma})$. The proof is complete. ■

Lemma 25. *Fix* $\mathbf{S} \in \mathbb{R}^{K \times N\epsilon}, \mathbf{T} \in \mathbb{R}^{N\epsilon \times L}$*, then sample a matrix* $\mathbf{G} \in \mathbb{R}^{N\epsilon \times N\epsilon}$ *such that each column of* $\mathbf{G}$ *represents a* ϵ*-subset of a* n*-dimensional vector sampled from* $\mathcal{N}(\mathbf{0}, \mathbf{I})$*, then with probability exceeding* $1 - \delta$*,*

$$\|\mathbf{SGT}\|_{\mathrm{F}}^2 \leq \|\mathbf{S}\|_{\mathrm{F}}^2 \|\mathbf{T}\|_{\mathrm{F}}^2 \cdot (4 \log N + 2 \log(1/\delta))$$

Proof. The proof will be a calculation.

$$\|\mathbf{SGT}\|_{\mathrm{F}}^2 = \sum_{i \in [K]} \sum_{j \in [L]} \sum_{k_1, k_2 \in [N\epsilon] \times [N\epsilon]} (\mathbf{S}_{i,k_1} \mathbf{G}_{k_1,k_2} \mathbf{T}_{k_2,j})^2 \leq \|\mathbf{S}\|_{\mathrm{F}}^2 \|\mathbf{T}\|_{\mathrm{F}}^2 \max_{i,j \in [N\epsilon] \times [N\epsilon]} (\mathbf{G}_{i,j})^2$$

It then suffices to bound the maximum value of a Gaussian squared over N^2 samples.

$$\mathbb{P}\Big\{ \max_{i,j} G_{i,j}^2 \geq t \Big\} \leq N^2 \, \mathbb{P}\big\{ G_{i,j}^2 \geq t \big\} \leq 2N^2 \, \mathbb{P}\Big\{ G_{i,j} \geq \sqrt{t} \Big\} = 2N^2 \int_{\sqrt{t}}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx$$

$$\leq N^2\sqrt{\frac{2}{\pi}}\int_{\sqrt{t}}^{\infty}\frac{x}{\sqrt{t}}\cdot e^{-x^2/2}dx = \frac{1}{\sqrt{t}}\cdot N^2\sqrt{\frac{2}{\pi}}\cdot e^{-t/2} \leq N^2\sqrt{\frac{2}{\pi}}\cdot e^{-t/2}$$

where the last inequality follows $t \geq 1$. We thus obtain from elementary inequalities,

$$\max_{i,j} G_{i,j}^2 \leq 2\log\left(N^2\sqrt{\frac{2}{\pi}}\cdot\frac{1}{\delta}\right)$$

for $N \geq (2/\pi)^{1/4}$. Our proof is complete. ■

D.2 Hilbert Space Concentrations of Measure

Proposition 26 (Jensen's Inequality [Jen06])**.** *Suppose φ is a convex function, then for a random variable X, it holds*

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

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 27 (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

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

Theorem 28 (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 $1-\delta$,*

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

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

Proposition 29 (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 7). 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 $1-\delta$,*

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

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

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.

$$\mathop{\mathbb{E}}_{X_i\sim\mathcal{P}_\sharp\phi}\left\|\frac{1}{(1-\epsilon)N}\sum_{i=1}^{(1-\epsilon)N} X_i\otimes X_i - \boldsymbol{\Sigma}\right\|_{\text{HS}} \overset{(ii)}{\leq} \mathop{\mathbb{E}}_{X_i\sim\mathcal{P}_\sharp\phi}\mathop{\mathbb{E}}_{\tilde{X}_i\sim\mathcal{P}_\sharp\phi}\left\|\frac{1}{(1-\epsilon)N}\sum_{i=1}^{(1-\epsilon)N} X_i\otimes X_i - \tilde{X}_i\otimes\tilde{X}_i\right\|_{\text{HS}}$$

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

In (ii) we note that $X_i\otimes X_i-\mathbf{\Gamma}$ is a mean $\mathbf{0}$ operator in the tensor product space $\mathcal{H}\otimes\mathcal{H}$. Then for $X,Y\in\mathcal{H}\otimes\mathcal{H}$ s.t. $\mathbb{E}[Y]=\mathbf{0}$ it follows $\|X\|_{\mathrm{HS}}=\|X-\mathbb{E}[Y]\|_{\mathrm{HS}}=\|\mathbb{E}[X-Y]\|_{\mathrm{HS}}$ 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 $\mathbf{\Gamma}$. By expanding out the Hilbert-Schmidt Norm, we then have

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

(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 will now calculate the fourth moment of a norm of sub-Gaussia function in the Hilbert Space.

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

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

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

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

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

Then, we have from McDiarmid's inequality (Proposition 27),

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

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

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

Next, applying a union bound over $\mathcal{S}$ with lemma 38, we have

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

Simplifying the resultant bound completes the proof. ■

E Additional Lemmas

In this section, we state additional lemmas referenced throughout the text.

Lemma 30. *Let* $\mathbf{a}, \mathbf{b} \in \mathbb{R}^n$ *and* $\mathbf{X} \in \mathbb{R}^{p\times n}$*, then*

$$\Big\|\sum_{i=1}^n a_i b_i \mathbf{x}_i\Big\|^2 \leq \|\mathbf{a}\|_\infty^2\|\mathbf{b}\|_2^2\big\|\mathbf{X}\mathbf{X}^{\mathrm{T}}\big\|_2$$

Proof. The proof is a simple calculation. Expanding out the LHS, we have

$$\Big\|\sum_{i=1}^n a_i b_i \mathbf{x}_i\Big\|_2^2 = \sum_{i=1}^n\sum_{j=1}^n a_i a_j b_i b_j \mathbf{x}_i^{\mathrm{T}}\mathbf{x}_j = (\mathbf{a}\circ\mathbf{b})^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}(\mathbf{a}\circ\mathbf{b}) \leq \|\mathbf{a}\circ\mathbf{b}\|_2^2\big\|\mathbf{X}^{\mathrm{T}}\mathbf{X}\big\|_2 \leq \|\mathbf{a}\|_\infty^2\|\mathbf{b}\|_2^2\big\|\mathbf{X}^{\mathrm{T}}\mathbf{X}\big\|_2$$

where the final inequality comes from noting

$$\|\mathbf{a}\circ\mathbf{b}\|^2 = \sum_{i=1}^n a_i^2 b_i^2 \leq \max_{i\in[n]} a_i^2 \cdot \sum_{i=1}^n b_i^2$$

Our proof is complete. ■

Lemma 31. *Consider a determinate set of numbers* $(a_i)_{i=1}^n$*, and determinate set of functions in the Hilbert Space,* $(X_i)_{i=1}^n$*, then*

$$\Big\|\sum_{i=1}^n a_i X_i\Big\|_{\mathcal{H}}^2 \leq \|\mathbf{a}\|_2^2\|\mathbf{K}\|$$

Proof. The proof is a calculation.

$$\Big\|\sum_{i=1}^n a_i X_i\Big\|_{\mathcal{H}}^2 = \Big\langle\sum_{i=1}^n a_i X_i, \sum_{j=1}^n a_j X_j\Big\rangle_{\mathcal{H}} = \sum_{i=1}^n\sum_{j=1}^n a_i a_j k(x_i,x_j) = \mathbf{a}^{\mathrm{T}}\mathbf{K}\mathbf{a} \leq \|\mathbf{a}\|_2^2\|\mathbf{K}\|$$

where $[\mathbf{K}]_{i,j} = k(x_i,x_j)$ is the kernel Gram matrix. ■

Lemma 32 (Lemma 3.11 [B+15]). *Let f be β-smooth and α-strongly convex over $\mathbb{R}^n$, then for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$,*

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

Proposition 33 (Young's Inequality [You12]). *Suppose $a, b \in \mathbb{R}_+$, then for $p, q > 1$ such that $\frac{1}{p} + \frac{1}{q} = 1$, then*

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

Proposition 34 (Proposition 10.1 [HMT11]). *Fix matrices $\mathbf{S}, \mathbf{T}$, then sample matrix $\mathbf{G}$ such that the elements of $\mathbf{G}$ are sampled from $\mathcal{N}(0, 1)$, then*

$$\mathbb{E}\|\mathbf{SGT}\|_{\mathrm{F}}^2 = \|\mathbf{S}\|_{\mathrm{F}}^2 \|\mathbf{T}\|_{\mathrm{F}}^2$$

Lemma 35 (AM-QM Inequality). *Let $x_1, \ldots, x_n$ be scalars in $\mathbb{R}$, then*

$$\big(\sum_{i \in [n]} x_i\big)^2 \leq n \sum_{i \in [n]} x_i^2$$

Lemma 36. *Suppose $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{m \times n}$, then*

$$\langle \mathbf{A}, \mathbf{B} \rangle_{\mathrm{Tr}} = \langle \mathrm{vec}(\mathbf{A}), \mathrm{vec}(\mathbf{B}) \rangle$$

Lemma 37. *Suppose $\mathbf{A}, \mathbf{B}, \mathbf{C}$ are size compatible matrices, then*

$$\mathrm{vec}(\mathbf{ABC}) = (\mathbf{C}^{\mathrm{T}} \otimes \mathbf{A})\mathrm{vec}(\mathbf{B})$$

Lemma 38 (Sum of Binomial Coefficients [CLRS22]). *Let $k, n \in \mathbb{N}$ such that $k \leq n$, then*

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