Iterative Thresholding for Non-Linear Learning in the Strong ϵ-Contamination Model

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

Alex Gittens
RPI CS, gittea@rpi.edu

Abstract

We study the problem of learning single neurons when both the labels and the covariates are possibly corrupted adversarially with a gradient-descent iterative thresholding algorithm. We assume the data is sampled from a ground truth distribution,

$$y = \sigma(\mathbf{x}^\top \mathbf{w}^*) + \xi$$

where ξ is Gaussian noise and σ is an activation function. We study sigmoid, leaky-ReLU, and ReLU activation functions.

We also study the linear regression problem when $\sigma(x) = x$. We improve upon previous approximation bounds for gradient based iterative thresholding algorithms [BJK15, SS19] and show with high probability a $O(\sigma\epsilon \log \epsilon^{-1})$ approximation upper bound, matching the best known approximation bound for iterative thresholding algorithms [ADKS22] while improving run-time from $O\left(\frac{1}{\epsilon^2}\right)$ to $O\left(\log\left(\frac{1}{\epsilon}\right)\right)$.

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

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

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

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

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

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

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

The aforementioned theorem has a log-dependence on $\|\mathbf{b}\|$ and is in the *realizable* setting, i.e. no variance of the optimal estimator. More recently, Awasthi et al. [ADKS22] studied the iterative trimmed maximum likelihood estimator. In their algorithm, at each step they find $\mathbf{w}^*$ which minimizes the elements in the trimmed set. 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 $y_i = \mathbf{w}^* \cdot \mathbf{x}_i + \eta_i$ where $\eta_i \sim \mathcal{N}(0, \sigma^2)$ and $\mathbf{x}_i$ are sampled from a sub-Gaussian Distribution with second-moment matrix 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, SS19, 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	Algorithm
[BJK15]	$O(\sigma)$	$O\left(N^2 d \log\left(\frac{1}{\sqrt{n}} \frac{\|\mathbf{b}\|_2}{\epsilon}\right)\right)$	Full Solve
[SS19]	$O(\sigma)$	$O\left(Nd \log\left(\frac{\|\mathbf{w}^*\|_2+\sigma^2}{\sigma}\right)\right)$	Gradient Descent
[ADKS22]	$O(\sigma\epsilon\log\epsilon^{-1})$	$O\left((Nd^2+d^3)\left(\frac{1}{\sigma\epsilon^2}\right)\right)$	Full Descent
Corollary 16	$O(\sigma\epsilon\log\epsilon^{-1})$	$O\left(Nd^2 \log\left(\frac{\|\mathbf{w}^*\|}{\sigma\epsilon\log\epsilon^{-1}}\right)\right)$	Gradient Descent

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, Σ, 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 $\max\{C_\psi x, x\}$.

Comparing to [BJK15], our bound does not depend on the norm of the noise, which can be made arbitrarily large by the adversary. We extend upon [SS19] by extending significantly past a linear convergence guarantee by also showing convergence to a near minimax-optimal error. We also offer a significant run-time improvement over the study in [ADKS22] from $O(1/\varepsilon^2)$ to $O(\log(1/\varepsilon^2))$.

Reference	Approximation	Neuron	Covariate Distribution
Theorem 15	$O\left(\epsilon\sqrt{\sigma C_Q}\log\epsilon^{-1}\right)$	Linear	Sub-Gaussian
Theorem ??	$O\left(\epsilon\sqrt{\sigma C_Q}\log\epsilon^{-1}\right)$	Sigmoid	Bounded, $\|\mathbf{x}\| \le B$
Theorem 18	$O\left(\epsilon\sqrt{\sigma C_Q}\log\epsilon^{-1}\right)$	Leaky-ReLU	Sub-Gaussian
Theorem 20	$O\left(\epsilon\sqrt{\sigma C_Q}\log\epsilon^{-1}\right)$	ReLU	$L_4 - L_2$ Hypercontractive

Table 2: Our results and Distributional assumptions for learning different neuron activation functions.

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 $x \ge x_0$ there exists C s.t. $y \le Cx$. We say $y = \Omega(x)$ if there exists x_0 s.t. for all $x \ge x_0$ there exists C s.t. $y \ge Cx$. 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 denote the Hadamard product between two vectors of the same size as $\mathbf{x} \circ \mathbf{y}$ such that for any vectors $(\mathbf{x} \circ \mathbf{y})_i = x_i y_i$.

Matrices. For a matrix A, let $\lambda_{\max}(A)$ and $\lambda_{\min}(A)$ represent the maximum and minimimum eigenvalues of A, respectively. We use the following matrix norms for a matrix $A \in \mathbb{R}^{m\times n}$,

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

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

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

Let $\mathrm{vec} : \mathbb{R}^{n\times k} \to \mathbb{R}^{nk}$ represent the vectorizaton of a matrix to a vector placing its columns one by one into a vector. We then have the useful facts,

Lemma 5. *Suppose $A, B \in \mathbb{R}^{m\times n}$, then*

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

Let $\otimes : \mathbb{R}^{N\times K} \times \mathbb{R}^{L\times M} \to \mathbb{R}^{NL\times KM}$ represent the Kronecker delta product between two matrices, this gives us the following relation,

Lemma 6. *Suppose A, B, C are conformal matrices, then*

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

Probability. We know discuss the probability theory concepts used throughout the paper. We consider the general sub-Gaussian design. The sub-Gaussian design is highly prevalent in the study of robust statistics [JLT20].

Definition 7 (Sub-Gaussian Distribution)**.** *We say a vector $\mathbf{x}$ is sampled from a sub-Gaussian distribution with second-moment matrix Σ and sub-Gaussian norm K, if $\mathbf{E}[\mathbf{x}\mathbf{x}^\top] = \Sigma$ and*

$$\mathbf{E}[\exp(t\mathbf{x}^\top\mathbf{v})] \le \exp\left(\frac{t^2K^2\mathbf{v}^\top\Sigma\mathbf{v}}{2}\right) \text{ for } \mathbf{v} \in \mathbb{S}^{d-1},\ t \in \mathbb{R}$$

A scalar random variable X is sub-Gaussian with sub-Gaussian norm ν if for all $p \in \mathbb{N}$,

$$\|X\|_{L_p} = (\mathbf{E}|X|^p)^{1/p} \le \nu\sqrt{p}$$

We often work with the products of sub-Gaussian random variables, which by the following indicate they are sub-Exponential.

Lemma 8 (Lemma 2.7.7 in [Ver20])**.** *Let X, Y be sub-Gaussian random variables, then XY is sub-Exponential, furthermore,*

$$\|XY\|_{\psi_1} \le \|X\|_{\psi_2}\|Y\|_{\psi_2}$$

The sub-exponential norm of $X \in \mathbb{R}$ is given by the following,

$$\|X\|_{\psi_1} = \inf\{t \ge 0 : \mathbf{E}\exp(|X|/t) \le 2\}$$

To give probabilistic bounds on the concentration of sub-Exponential random variables, we often utilize Bernstein's Theorem.

Lemma 9 (Proposition 5.16 in [Ver10])**.** *Let $X_1, \ldots, X_N$ be independent centered sub-exponential random variables, and $K = \max_i\|X_i\|_{\psi_1}$. Then for every $\mathbf{a} \in \mathbb{R}^n$ and $t \ge 0$,*

$$\mathbf{Pr}\left\{\left|\sum_{i\in[N]} a_iX_i\right| \ge t\right\} \le 2\exp\left[-c\left(\frac{t^2}{K^2\|\mathbf{a}\|_2^2} \wedge \frac{t}{K\|\mathbf{a}\|_\infty}\right)\right]$$

From elementary algebraic manipulations, we are able to achieve the following upper bound with high probability.

Lemma 10. *Let $X_1, \ldots, X_N$ be independent centered sub-exponential random variables, and $K = \max_i\|X_i\|_{\psi_1}$. Then for every $\mathbf{a} \in \mathbb{R}^n$ and $\delta \in (0, 1)$, with probability exceeding $1 - \delta$,*

$$\left|\sum_{i\in[N]} a_iX_i\right| \le \left(\frac{1}{c}\cdot K^2\|\mathbf{a}\|_2^2\log(2/\delta)\right)^{1/2} \vee \left(\frac{1}{c}\cdot K\|\mathbf{a}\|_\infty\log(2/\delta)\right)$$

2.1 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].

[BJK15] study iterative thresholding for least squares regression / sparse recovery. In particular, one of their contributions is a gradient descent algorithm, Torrent, when the covariates are sampled from a sub-Gaussian distribution. Their approximation bound in Theorem 5 relies on the fact that $\lambda_{\min}(\Sigma) = \lambda_{\max}(\Sigma)$ and with sufficiently large data and sufficiently small ϵ, $\kappa(X) \searrow 1$. Bhatia et al. also study a full solve algorithm, where after each thresholding step and obtaining $(1-\epsilon)N$ samples, they set $\mathbf{w}^{(t)}$ to the minimizer of the squared loss over the $(1-\epsilon)N$ points and refer to this algorithm as Torrent-FC. They study this algorithm in the prescence of both adversarial and intrinsic noise in the optimal estimator. Their analysis in Corollary 11 gives $O(\sigma)$ error when the intrinsic noise is sub-Gaussian with sub-Gaussian norm $O(\sigma^2)$.

[SS19] study iterative thresholding for learning Generalized Linear Models (GLMs). In both the linear and non-linear case, they present results on linear convergence. Their results imply a bound of $O(\sigma)$ in the linear case. They furthermore provide experimental evidence of the success of iterative thresholding when applied to neural networks.

More recently, [ADKS22] studied the iterative trimmed maximum likelihood estimator for General Linear Models. Similar to Torrent-FC, their algorithm solves the MLE problem over the data kept after each thresholding step. They prove the best known bounds for iterative thresholding algorithms in the linear regression case, $O(\sigma\epsilon \log \epsilon^{-1})$. When the good data is sampled from a sub-Gaussian distribution non-identity covariance, they first run a near-linear filtering algorithm from [DHL19] to obtain covariates that are sub-Gaussian with close to identity covariance.

All the previous works described above have breakdown point $O(1)$. This breakdown point is typically a consequence of the poor conditioning of the good covariates remaining after the adversary has removed $N\epsilon$ good points. Hence, typically in the *oblivious* case, i.e. randomized corruption, there exists research [BJKK17, SBRJ19], with breakdown point $\Omega(1)$.

3 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.

3.1 Activation Functions

We first give the non-linear functions we will be learning.

Property 11. *σ is a continuous, monotically increasing, and differentiable almost everywhere.*

Property 12. *σ is Lipschitz, i.e. $|\sigma(x) - \sigma(y)| \le \|\sigma\|_{\text{lip}}|x - y|$.*

Property 13. *For any $x \ge 0$, there exists $\gamma > 0$ such that $\inf_{|z|\le x} \sigma'(z) \ge \gamma > 0$.*

Sigmoid functions such as tanh and sigmoid and the leaky-ReLU function satisfy Properties 11, 12, and 13. Property 13 does not hold for the ReLU function, and therefore we require stronger conditions for our approximation bounds to hold.

3.2 Algorithm

We will first define the thresholding operator to simplify the notation in our formal algorithm.

Definition 14 (Hard Thresholding Operator in [BJK15])**.** *For any vector $\mathbf{v} \in \mathbb{R}^n$, let $\sigma_{\mathbf{v}}$ be the permutation that orders elements in ascending order, i.e. $\mathbf{v}_{\sigma_{\mathbf{v}}(1)} \le \mathbf{v}_{\sigma_{\mathbf{v}}(2)} \le \cdots \le \mathbf{v}_{\sigma_{\mathbf{v}}(n)}$. Then for any $k \le n$, the hard thresholding oeprator is defined as,*

$$\mathrm{HT}(\mathbf{v}; k) = \left\{ i \in [n] : \sigma_{\mathbf{v}}^{-1}(i) \le k \right\}$$

We now give the gradient descent algorithm which we will study for the remainder of this paper.

Algorithm 1 Gradient Descent Iterative Thresholding for Learning a Non-linear Neuron

input: Possibly corrupted $X \in \mathbb{R}^{d\times N}$ with outputs $\mathbf{y} \in \mathbb{R}^N$, activation function σ, corruption parameter $\epsilon = O(1)$, and small constant α.

output: $\varepsilon-$Approximate solution $\mathbf{w} \in \mathbb{R}^d$ to minimize $\|\mathbf{w} - \mathbf{w}^*\|_2$.

$\mathbf{w}^{(0)} \sim \mathcal{B}_d(\alpha\|\mathbf{w}^*\|)$
$\eta \leftarrow 0.1\kappa^{-2}(\Sigma)$
$T \leftarrow O\left(\kappa^2(\Sigma)\log\left(\frac{\|\mathbf{w}^*\|_2}{\varepsilon}\right)\right)$
for $t \in [T]$ **do**
    $\nu_i^{(t)} = (\sigma(\mathbf{x}_i^\top \mathbf{w}^{(t)}) - y_i)^2 \quad \forall i \in [N]$
    $\mathrm{S}^{(t)} \leftarrow \mathrm{HT}(\boldsymbol{\nu}^{(t)}, (1-\epsilon)N)$
    $\mathbf{w}^{(t+1)} \leftarrow \mathbf{w}^{(t)} - \eta\nabla\mathcal{R}(\mathbf{w}^{(t)}; \mathrm{S}^{(t)})$

return: $\mathbf{w}^{(T)}$

Runtime. In each iteration we calculate the ℓ_2 error for N points, in total $O(Nd)$. For the Hard Thresholding step, it suffices to find the $n(1-\epsilon)$-th largest element, we can run a selection algorithm in worst-case time $O(N\log N)$, then partition the data in $O(N)$. The run-time for calculating the gradient and updating $\mathbf{w}^{(t)}$ is dominated by the matrix multiplication in $X_{\mathrm{S}^{(t)}}X_{\mathrm{S}^{(t)}}^\top$ which can be done in $O(Nd^2)$. Then considering the choice of T, we have the algorithm runs in time $O\left(Nd^2\log\left(\frac{\|\mathbf{w}^*\|_2}{\sigma\epsilon\log\epsilon^{-1}}\right)\right)$ to obtain $O(\sigma\epsilon\log\epsilon^{-1})$ ℓ_2-approximation error.

3.3 Proof Sketch

In this section we will give a general sketch of our proofs, from which all the individual theorems will be based upon. Let $t \in [T]$, we then have,

$$\begin{aligned}\|\mathbf{w}^{(t+1)} - \mathbf{w}^*\| &\leq \|\mathbf{w}^{(t)} - \eta\nabla\mathcal{R}(\mathbf{w}^{(t)}; \mathrm{S}^{(t)}) - \mathbf{w}^*\|_2 \\ &\leq \|\mathbf{w}^{(t)} - \mathbf{w}^* - \eta\nabla\mathcal{R}(\mathbf{w}^{(t)}; \mathrm{S}^{(t)}\cap \mathrm{P})\| + \|\eta\nabla\mathcal{R}(\mathbf{w}^{(t)}; \mathrm{S}^{(t)}\cap\mathrm{Q})\|_2 \qquad (1)\end{aligned}$$

We analyze the first term of Equation (1) through its square,

$$\begin{aligned}&\|\mathbf{w}^{(t)} - \mathbf{w}^* - \nabla\mathcal{R}(\mathbf{w}^{(t)}; \mathrm{S}^{(t)}\cap\mathrm{P})\|^2 \\ &\qquad\leq \|\mathbf{w}^{(t)} - \mathbf{w}^*\|_2^2 - 2\eta\cdot\langle\mathbf{w}^{(t)} - \mathbf{w}^*, \nabla\mathcal{R}(\mathbf{w}^{(t)}; \mathrm{S}^{(t)}\cap\mathrm{P})\rangle + \eta^2\cdot\|\nabla\mathcal{R}(\mathbf{w}^{(t)}; \mathrm{S}^{(t)}\cap\mathrm{P})\|^2\end{aligned}$$

In this step the finer details of the particular proof will differ, however the structure remains the same. We will prove there exists a constant $C_1 > 0$ such that,

$$\langle\mathbf{w}^{(t)} - \mathbf{w}^*, \nabla\mathcal{R}(\mathbf{w}^{(t)}; \mathrm{S}^{(t)}\cap\mathrm{P})\rangle \geq (1-2\epsilon)C_1\|\mathbf{w}^{(t)} - \mathbf{w}^*\|_2^2 - C_3\|\mathbf{w}^{(t)} - \mathbf{w}^*\|_2$$

where C_3 is a term that is dependent on the variance of the Gaussian noise, one of our contributions is that $C_3 = O(\sigma\epsilon\log\epsilon^{-1})$ when N sufficiently large for the activation functions studied in the text. Next, an application of Young's inequality gives us,

$$\langle\mathbf{w}^{(t)} - \mathbf{w}^*, \nabla\mathcal{R}(\mathbf{w}^{(t)}; \mathrm{S}^{(t)}\cap\mathrm{P})\rangle \geq ((1-2\epsilon)C_1 - C_3C_1)\|\mathbf{w}^{(t)} - \mathbf{w}^*\|_2^2 - C_3C_1^{-1}$$

We next show there exists a positive constant C_2, such that

$$\|\nabla\mathcal{R}(\mathbf{w}^{(t)}; \mathrm{S}^{(t)}\cap\mathrm{P})\| \leq (1-\epsilon)C_2\|\mathbf{w}^{(t)} - \mathbf{w}^*\|_2^2$$

Then, solving a simple quadratic equation, we have that $\eta^2C_2 \leq \eta C_1C_3$ for a $C_3 \in (0,1)$ when we choose $\eta \leq \frac{C_1C_3}{C_2}$ and we are able to eliminate the norm of the gradient squared term. We must now control the corrupted gradient term. The key idea is to note that from the optimality of the sub-quantile set,

$$\sum_{i\in\mathrm{S}^{(t)}\cap\mathrm{Q}}\mathcal{L}(\mathbf{w}^{(t)}; \mathbf{x}_i, y_i) \leq \sum_{i\in\mathrm{P}\setminus\mathrm{S}^{(t)}}\mathcal{L}(\mathbf{w}^{(t)}; \mathbf{x}_i, y_i)$$

and $|\mathrm{S}^{(t)} \cap \mathrm{Q}| = |\mathrm{P} \setminus \mathrm{S}^{(t)}|$. We then prove the existence of a constant C_4 such that,

$$\|\nabla \mathcal{R}(\mathbf{w}^{(t)}; \mathrm{S}^{(t)} \cap \mathrm{Q})\|_2 \leq \epsilon C_4 \|\mathbf{w}^{(t)} - \mathbf{w}^*\|_2^2$$

Then, combining our results, we end up with a linear convergence of the form,

$$\|\mathbf{w}^{(t+1)} - \mathbf{w}^*\|_2 \leq \|\mathbf{w}^{(t)} - \mathbf{w}^*\|_2 (1 - \eta(1-2\epsilon)C_1 + \eta\epsilon C_4) + C_3 C_1^{-1}$$

We obtain a bound that is of the form,

$$\|\mathbf{w}^{(t+1)} - \mathbf{w}^*\|_2 \leq \|\mathbf{w}^{(t)} - \mathbf{w}^*\|_2 (1 - \lambda) + E$$

Then, we find that

$$\|\mathbf{w}^{(t)} - \mathbf{w}^*\|_2 \leq \|\mathbf{w}^{(0)} - \mathbf{w}^*\|_2 (1 - \lambda)^t + \sum_{k \in [t]} (1-\lambda)^k E$$

We then find asymptotically, the second term converges to $\lambda^{-1} E$. Then, it suffices to find T such that,

$$\|\mathbf{w}^{(T)} - \mathbf{w}^*\|_2 \leq \lambda^{-1} E$$

We can note that $1 - \lambda \leq e^{-\lambda}$, then bounding T, we obtain a $\lambda^{-1}E$ approximation bound when

$$T \geq \lambda^{-1} \cdot \log\left(\frac{\|\mathbf{w}^* - \mathbf{w}^{(0)}\|_2}{\lambda^{-1} E}\right)$$

3.4 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 without requiring a filtering algorithm, 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 $W \in \mathbb{R}^{K \times d}$, $X \in \mathbb{R}^{d \times n}$, and $Y \in \mathbb{R}^{K \times n}$.

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

Theorem 15. *Let $X = [\mathbf{x}_1, \ldots, \mathbf{x}_n]^\top \in \mathbb{R}^{d \times N}$ be the data matrix and $Y = [\mathbf{y}_1, \ldots, \mathbf{y}_n] \in \mathbb{R}^{K \times N}$ be the output, such that for $i \in \mathrm{P}$, $\mathbf{x}_i$ are sampled from a sub-Gaussian distribution with sub-Gaussian norm L and second-moment matrix Σ, and for $j \in \mathrm{Q}$, $\mathbf{x}_i$ are given by the adversary where $\|\mathbf{x}_j\| \leq C_3$ for some positive constant. Suppose for $i \in \mathrm{P}$ the output is given as $\mathbf{y}_i = W\mathbf{x}_i + \mathbf{e}_i$ where for $j \in [K]$, $[\mathbf{e}_i]_j \sim \mathcal{N}(0, \sigma^2)$. Then after $O\left(\kappa(\Sigma) \log\left(\frac{\|W^*\|_{\mathrm{F}}}{\varepsilon}\right)\right)$ gradient descent iterations, $N \geq \frac{(1/\delta) K \operatorname{Tr}(\Sigma)}{1600\epsilon^2 \log \epsilon^{-1} \lambda_{\max}(\Sigma)}$, and learning rate $\eta = 0.1\lambda_{\max}^{-2}(\Sigma)$, with probability exceeding $1 - \delta$, Algorithm 1 returns $W^{(T)}$ such that*

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

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 I. The full solve algorithm studied in [ADKS22] returns a $O\big(\sigma\epsilon \log \epsilon^{-1}\big)$ in time $O\big(\frac{1}{\epsilon^2}\big(Nd^2 + d^3\big)\big)$ and the same algorithm studied in [BJK15], Torrent-FC otains $O(\sigma)$ approximation error in run-time $O\Big(\log\Big(\frac{1}{\sqrt{n}} \frac{\|\mathbf{b}\|}{\sigma\epsilon \log \epsilon^{-1}}\Big)(Nd^2 + d^3)\Big)$, with the gradient descent based approach, we are able to improve the runtime to $O\Big(\log\Big(\frac{\|\mathbf{w}^*\|}{\sigma\epsilon \log \epsilon^{-1}}\Big) Nd^2\Big)$ for the same approximation bound. By no longer requiring the full-solve, we are able to remove super-linear relation to d. In comparison to [BJK15], we do not have dependence on the noise vector $\mathbf{b}$, which can have very large norm in relation to the norm of $\mathbf{w}^*$. Our proof is also a significant improvement over the presentation given in Lemma 5 of [SS19] as under the same conditions, we give more than the linear convergence, but we show linear convergence is possible on any second-moment matrix of the good covariates and covariate corruption, and then develop concentration inequality bounds to match the best known result for iteratived trimmed estimators. We will formalize our results into a corollary to give a more representative comparison in the literature.

Corollary 16. *Let $X = [\mathbf{x}_1, \ldots, \mathbf{x}_n]^\top \in \mathbb{R}^{d\times N}$ be the data matrix and $\mathbf{y} = [y_1, \ldots, y_n] \in \mathbb{R}^N$ be the output, such that for $i \in \mathrm{P}$, $\mathbf{x}_i$ are sampled from a sub-Gaussian distribution with sub-Gaussian norm L and second-moment matrix I. Suppose for $i \in \mathrm{P}$ the output is given as $\mathbf{y}_i = \mathbf{x}_i^\top \mathbf{w}^* + \xi_i$ 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, $N \geq \frac{(d/\delta)}{800\epsilon^2}$, and learning rate $\eta = 0.1$, with probability exceeding $1-\delta$, Algorithm 1 returns $\mathbf{w}^{(T)}$ such that*

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

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

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

The second relation given in Corollary 16 matches the best known bound for robust linear regression with iterative thresholding. The first relation given in Corollary 16 is the extension to handle second-moment matrices which do not have unitary condition number as well as corrupted covariates.

3.5 Learning Sigmoid Neurons

We next study the problem of learning GLMs following the model given in §5 of [SS19]. 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.

$$\mathcal{L}(\mathbf{w}; \mathbf{x}_i, y_i) = (\sigma(\mathbf{w}\cdot\mathbf{x}_i) - y_i)^2$$

Theorem 17. *Let $X = [\mathbf{x}_1, \ldots, \mathbf{x}_N]^\top \in \mathbb{R}^{d\times N}$ be the data matrix and $\mathbf{y} = [y_1, \ldots, y_n]$ be the output, such that for $i \in \mathrm{P}$, $\mathbf{x}_i$ are sampled from a sub-Gaussian distribution with second-moment matrix Σ and sub-Gaussian norm C_Σ, and $y_i = \sigma(\mathbf{w}^* \cdot \mathbf{x}_i) + \xi_i$ for ξ_i sampled from a sub-Gaussian distribution with sub-Gaussian norm ν. Suppose the activation function, satisfies Properties 11, 12, and 13. Then after $O\left(\kappa(\Sigma)\log\left(\frac{\|\mathbf{w}^*\|}{\varepsilon}\right)\right)$ gradient descent iterations, then with probability exceeding $1-\delta$,*

$$\|\mathbf{w}^{(T)} - \mathbf{w}^*\|_2 \leq$$

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

Proof. The proof is deferred to § B.1.1. ∎

3.6 Learning Leaky-ReLU Neural Networks

We will now consider learning a neuron with the Leaky-ReLU function. We can note that Properties 11, 12, and 13 all hold for the Leaky-ReLU. More conveniently, we have γ in Property 13 is constant over $\mathbb{R}$. In our proof we are able to leverage the fact that the second derivative is zero almost surely.

Theorem 18. *Let $X = [\mathbf{x}_1, \ldots, \mathbf{x}_n]^\top \in \mathbb{R}^{N\times d}$ be the data matrix and $\mathbf{y} = [y_1 \ldots, y_n]^\top$ be the output, such that for $i \in \mathrm{P}$, $\mathbf{x}_i \sim \mathcal{P}$ are sampled from a sub-Gaussian distribution with sub-Gaussian norm K and second-moment matrix Σ, and $y_i = \psi(\mathbf{x}^\top\mathbf{w}^*) + \xi_i$ for $\xi_i \sim \mathcal{N}(0,\sigma^2)$ where $\psi(x) = \max\{C_\psi x, x\}$. Then after $O\left(C_\psi^{-2}\kappa(\Sigma)\log\left(\frac{\|W^*\|_\mathrm{F}}{\varepsilon}\right)\right)$ gradient descent iterations and $\epsilon \leq \frac{C_\psi^2\lambda_{\min}(\Sigma)}{\sqrt{32C_Q\lambda_{\max}(\Sigma)}}$, with probability exceeding $1-\delta$, Algorithm 1 with learning rate $\eta = O(\kappa^{-2}(\Sigma))$ returns $\mathbf{w}^{(T)}$ such that*

$$\|\mathbf{w}^{(T)} - \mathbf{w}^*\|_2 \leq$$

Proof. The proof is deferred to § B.2.1. ∎

3.7 Learning ReLU Neural Networks

We will now consider the problem of learning ReLU neural networks. We first give a preliminary result for randomized initialization.

Lemma 19 (Theorem 3.4 in [DLT+18]). *Suppose* $\mathbf{w}^{(0)}$ *is sampled uniformly from a p-dimensional ball with radius* $\alpha\|\mathbf{w}^*\|$ *such that* $\alpha \leq \sqrt{\frac{1}{2\pi p}}$, *then with probability at least* $\frac{1}{2} - \alpha\sqrt{\frac{\pi p}{2}}$

$$\|\mathbf{w}^{(0)} - \mathbf{w}^*\|_2 \leq \sqrt{1-\alpha^2}\|\mathbf{w}^*\|_2$$

From this result we are able to derive porbabilistic guarantees on the convergence of learning ReLU neuron.

Theorem 20. *Let* $X = [\mathbf{x}_1, \ldots, \mathbf{x}_n]^\top \in \mathbb{R}^{n\times d}$ *be the data matrix and* $\mathbf{y} = [y_1, \ldots, y_n]^\top$ *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* Σ. *Suppose for* $i \in \mathrm{P}$, *the output is given as* $y_i = \psi(\mathbf{x}^\top\mathbf{w}^*) + \xi_i$ *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*

$$\|\mathbf{w}^{(T)} - \mathbf{w}^*\|_2 \leq$$

Proof. The proof is deferred to § B.3.1 ∎

4 Discussion

In this paper, we study the theoretical convergence properties of iterative thresholding for non-linear learning problems in the Strong ϵ-contamination model. Our warm-up result for linear regression reduces the runtime while achieving the best known approximation for iterative thresholding algorithms. Many papers have experimentally studied the iterative thresholding estimator in large scale neural networks [SS19, HYW+23] and to our knowledge, we are the first paper to make advancements in the theory of iterative thresholding for a general class of activation functions.

References

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

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

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

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

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

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

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

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

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

[DHL19] Yihe Dong, Samuel Hopkins, and Jerry Li. Quantum entropy scoring for fast robust mean estimation and improved outlier detection. *Advances in Neural Information Processing Systems*, 32, 2019.

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

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

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

[DLT+18] Simon Du, Jason Lee, Yuandong Tian, Aarti Singh, and Barnabas Poczos. Gradient descent learns one-hidden-layer cnn: Don't be afraid of spurious local minima. In *International Conference on Machine Learning*, pages 1339–1348. PMLR, 2018.

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

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

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

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

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

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

[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.

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

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

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

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

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

[MBM16] Song Mei, Yu Bai, and Andrea Montanari. The landscape of empirical risk for non-convex losses. *arXiv preprint arXiv:1607.06534*, 2016.

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

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

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

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

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

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

[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.

[SS19] 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.

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

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

[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.

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 15

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

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

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

$$\nabla \mathcal{L}(W) = 2(WX - Y)X^\top$$

We first show we can obtain Linear Convergence plus a noise term and a consistency term, which we define as the difference between the optimal estimator in expectation and the optimal estimator over a population. We then show the consistency term goes to zero with high probability as $N \to \infty$. Finally, we use the concentration inequalities developed in § C to give a clean approximation bound.

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

$$\begin{aligned}&\|W^{(t+1)} - W^*\|_{\mathrm{F}} = \|W^{(t)} - W^* - \eta \nabla \mathcal{R}(W^{(t)}; \mathrm{S}^{(t)})\|_{\mathrm{F}} \\ &= \|W^{(t)} - W^* - \eta \nabla \mathcal{R}(W^{(t)}; \mathrm{TP}) + \eta \nabla \mathcal{R}(W^*; \mathrm{P}) - \eta \nabla \mathcal{R}(W^*; \mathrm{P}) + \eta \nabla \mathcal{R}(\widehat{W}; \mathrm{P}) - \eta \nabla \mathcal{R}(W^{(t)}; \mathrm{FP})\|_{\mathrm{F}} \\ &\le \|W^{(t)} - W^* + \eta \nabla \mathcal{R}(W^{(t)}; \mathrm{TP}) + \eta \nabla \mathcal{R}(W^*; \mathrm{TP})\|_{\mathrm{F}} + \|\eta \nabla \mathcal{R}(W^{(t)}; \mathrm{FP})\|_{\mathrm{F}} + \|\eta \nabla \mathcal{R}(W^*; \mathrm{FN})\|_{\mathrm{F}} \\ &\quad + \|\eta \nabla \mathcal{R}(W^*; \mathrm{P}) - \eta \nabla \mathcal{R}(\widehat{W}; \mathrm{P})\|_{\mathrm{F}}\end{aligned} \tag{2}$$

We first will upper bound the first term in Equation (2) through its square.

$$\begin{aligned}&\|W^{(t)} - W^* - \eta \nabla \mathcal{R}(W^{(t)}; \mathrm{TP}) + \eta \nabla \mathcal{R}(W^*; \mathrm{TP})\|_{\mathrm{F}}^2 = \|W^{(t)} - W^*\|_{\mathrm{F}}^2 \\ &- 2\eta \cdot \mathrm{Tr}\big((W^{(t)} - W^*)^\top (\nabla \mathcal{R}(W^{(t)}; \mathrm{TP}) - \nabla \mathcal{R}(W^*; \mathrm{TP}))\big) + \|\eta \nabla \mathcal{R}(W^{(t)}; \mathrm{TP}) - \eta \nabla \mathcal{R}(W^{(t)}; \mathrm{TP})\|_{\mathrm{F}}^2\end{aligned} \tag{3}$$

We then lower bound the second term in Equation (3),

$$\begin{aligned}&2\eta \cdot \mathrm{Tr}\big((W^{(t)} - W^*)^\top (\nabla \mathcal{R}(W^{(t)}; \mathrm{TP}) - \nabla \mathcal{R}(W^*; \mathrm{TP}))\big) \\ &\stackrel{\text{def}}{=} \frac{4\eta}{(1-\epsilon)N} \cdot \mathrm{Tr}\big((W^{(t)} - W^*)^\top \big((W^{(t)} X_{\mathrm{TP}} - Y_{\mathrm{TP}}) X_{\mathrm{TP}}^\top - E_{\mathrm{TP}} X_{\mathrm{TP}}^\top\big)\big) \\ &\stackrel{(i)}{=} \frac{4\eta}{(1-\epsilon)N} \cdot \mathrm{Tr}\Big((W^{(t)} - W^*)^\top (W^{(t)} - W^*) X_{\mathrm{TP}} X_{\mathrm{TP}}^\top\Big)\end{aligned}$$

In the above, (i) follows from recalling that $Y_{\mathrm{TP}} = W^* X_{\mathrm{TP}} - E_{\mathrm{TP}}$. We then have,

$$\begin{aligned}&\frac{4\eta}{(1-\epsilon)N} \cdot \mathrm{Tr}\Big((W^{(t)} - W^*)^\top (W^{(t)} - W^*) X_{\mathrm{TP}} X_{\mathrm{TP}}^\top\Big) \\ &\stackrel{(ii)}{=} \frac{4\eta}{(1-\epsilon)N} \cdot \mathrm{Tr}\Big((W^{(t)} - W^*) X_{\mathrm{TP}} X_{\mathrm{TP}}^\top (W^{(t)} - W^*)^\top\Big) \\ &\stackrel{(iii)}{=} \frac{4\eta}{(1-\epsilon)N} \cdot \langle \mathrm{vec}((W^{(t)} - W^*)^\top), \mathrm{vec}(X_{\mathrm{TP}} X_{\mathrm{TP}}^\top (W^{(t)} - W^*)^\top) \rangle\end{aligned}$$

$$
\begin{aligned}
&\overset{(iv)}{=} \frac{4\eta}{(1-\epsilon)N} \cdot \langle \mathrm{vec}((W^{(t)} - W^*)^\top), (I \otimes X_{\mathrm{TP}} X_{\mathrm{TP}}^\top)\mathrm{vec}((W^{(t)} - W^*)^\top) \rangle \\
&\overset{(v)}{=} \frac{4\eta}{(1-\epsilon)N} \cdot \sum_{k \in [K]} \left\langle \mathbf{w}_k^{(t)} - \mathbf{w}_k^*, X_{\mathrm{TP}} X_{\mathrm{TP}}^\top (\mathbf{w}_k^{(t)} - \mathbf{w}_k^*) \right\rangle \\
&\overset{(vi)}{\geq} \frac{2\eta}{(1-\epsilon)N} \cdot \sum_{k \in [K]} \left(\lambda_{\min}(X_{\mathrm{TP}} X_{\mathrm{TP}}^\top) \|\mathbf{w}_k^{(t)} - \mathbf{w}_k^*\|_2^2 + \|X_{\mathrm{TP}} X_{\mathrm{TP}}^\top\|_2^{-1} \|X_{\mathrm{TP}} X_{\mathrm{TP}}^\top (\mathbf{w}_k^{(t)} - \mathbf{w}_k^*)\|_2^2 \right) \\
&= \frac{2\eta}{(1-\epsilon)N} \cdot \lambda_{\min}(X_{\mathrm{TP}} X_{\mathrm{TP}}^\top) \|W^{(t)} - W^*\|_{\mathrm{F}}^2 + \frac{2\eta}{(1-\epsilon)N} \cdot \|X_{\mathrm{TP}} X_{\mathrm{TP}}^\top\|_2^{-1} \|(W^{(t)} - W^*) X_{\mathrm{TP}} X_{\mathrm{TP}}^\top\|_{\mathrm{F}}^2 \qquad (4)
\end{aligned}
$$

In the above, (ii) follows from the cyclic property of the trace, (iii) follows from the relation given in Lemma 5, (iv) holds from the relation given in Lemma 6, the inequality in (vi) follows from Lemma 31, and in (v) we apply Lemma 6, which gives the following equality,

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

We now upper bound the second term in Equation (2),

$$
\begin{aligned}
\left\| \eta \nabla \mathcal{R}(W^{(t)}; \mathrm{FP}) \right\|_{\mathrm{F}} &\overset{\mathrm{def}}{=} \frac{2\eta}{(1-\epsilon)N} \cdot \|(W^{(t)} X_{\mathrm{FP}} - Y_{\mathrm{FP}}) X_{\mathrm{FP}}^\top\|_{\mathrm{F}} \\
&\overset{(vii)}{\leq} \frac{2\eta}{(1-\epsilon)N} \cdot \|X_{\mathrm{FP}}\|_2 \|W^{(t)} X_{\mathrm{FP}} - Y_{\mathrm{FP}}\|_{\mathrm{F}} \\
&\overset{(viii)}{\leq} \frac{2\eta}{(1-\epsilon)N} \cdot \|X_{\mathrm{FP}}\|_2 \|W^{(t)} X_{\mathrm{FN}} - Y_{\mathrm{FN}}\|_{\mathrm{F}} \\
&\overset{(ix)}{\leq} \frac{2\eta}{(1-\epsilon)N} \cdot \|X_{\mathrm{FP}}\|_2 \left(\|W^{(t)} X_{\mathrm{FN}} - W^* X_{\mathrm{FN}}\|_{\mathrm{F}} + \|E_{\mathrm{FN}}\|_{\mathrm{F}} \right) \\
&\overset{(x)}{\leq} \frac{2\eta}{(1-\epsilon)N} \cdot \|X_{\mathrm{FP}}\|_2 \|X_{\mathrm{FN}}\|_2 \|W^{(t)} - W^*\|_{\mathrm{F}} + \frac{2\eta}{(1-\epsilon)N} \cdot \|X_{\mathrm{FP}}\|_2 \|E_{\mathrm{FN}}\|_{\mathrm{F}}
\end{aligned}
$$

In the above, the equalities in (vii) and (x) from the fact that for any two size compatible matrices, A, B, it holds that $\|AB\|_{\mathrm{F}} \leq \|A\|_{\mathrm{F}} \|B\|_2$, $(viii)$ follows from the optimality of the Subquantile set, and (ix) follows from the sub-additivity of the Frobenius norm. We will now upper bound the third term in Equation (2),

$$
\|\eta \nabla \mathcal{R}(W^*; \mathrm{FN})\|_{\mathrm{F}} \overset{\mathrm{def}}{=} \frac{2\eta}{(1-\epsilon)N} \cdot \|(W^* X_{\mathrm{FN}} - Y_{\mathrm{FN}}) X_{\mathrm{FN}}^\top\|_{\mathrm{F}} \leq \frac{2\eta}{(1-\epsilon)N} \cdot \|E_{\mathrm{FN}} X_{\mathrm{FN}}^\top\|_{\mathrm{F}} \qquad (5)
$$

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

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

$$
\|\eta \nabla \mathcal{R}(W^*; \mathrm{P}) - \eta \nabla \mathcal{R}(\widehat{W}; \mathrm{P})\|_{\mathrm{F}} \overset{\mathrm{def}}{=} \frac{2\eta}{(1-\epsilon)N} \cdot \|(\widehat{W} - W^*) X_{\mathrm{P}} X_{\mathrm{P}}^\top\|_{\mathrm{F}} \leq \frac{2\eta}{(1-\epsilon)N} \cdot \|E_{\mathrm{P}} X_{\mathrm{P}}^\top\|_{\mathrm{F}}
$$

We can then have from Lemma 29,

$$
\|E_{\mathrm{P}} X_{\mathrm{P}}^\top\|_{\mathrm{F}}^2 \leq \frac{16}{3} \cdot \sigma\left(\|X_{\mathrm{P}}\|_{\mathrm{F}}^2 \log\left((2/\delta) N^2 \right) \right) \leq \frac{32}{3} \cdot \sigma\left(N(1-\epsilon) d \lambda_{\max}(\Sigma) \log\left((2/\delta) N^2 \right) \right)
$$

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

$$
\frac{2\eta}{(1-\epsilon)N} \cdot \|E_{\mathrm{P}} X_{\mathrm{P}}^\top\|_{\mathrm{F}} \leq \eta \cdot \sqrt{\frac{32\sigma^2 d \lambda_{\max}(\Sigma) \log((2/\delta)N^2)}{3(1-\epsilon)N}}
$$

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

$$\|W^{(t+1)} - W^*\|_{\mathrm{F}} \le \|W^{(t)} - W^*\|_{\mathrm{F}} \left(1 - \frac{2\eta}{(1-\epsilon)N} \cdot \lambda_{\min}(X_{\mathrm{TP}} X_{\mathrm{TP}}^\top) + \frac{2\eta}{(1-\epsilon)N} \cdot \|X_{\mathrm{FP}}\|_2 \|X_{\mathrm{FN}}\|_2\right) + \frac{2\eta}{(1-\epsilon)N} \cdot \|X_{\mathrm{FP}}\|_2 \|E_{\mathrm{FN}}\|_{\mathrm{F}} + \frac{2\eta}{(1-\epsilon)N} \cdot \|E_{\mathrm{FN}} X_{\mathrm{FN}}^\top\|_{\mathrm{F}} + \eta \cdot \sqrt{\frac{32\sigma^2 d\lambda_{\max}(\Sigma)\log((2/\delta)N^2)}{3(1-\epsilon)N}}$$

Step 3: Concentration Bounds. From Proposition 22, we obtain with probability exceeding $1-\delta$ that $\|E_{\mathrm{FN}}\|_{\mathrm{F}} \le \sigma\sqrt{30NK\epsilon\log\epsilon^{-1}}$. From our assumption on bounded covariate corruptions, we have $\|E_{\mathrm{FN}}\|_{\mathrm{F}}\|X_{\mathrm{FP}}\|_{\mathrm{F}} \le \sigma\epsilon\sqrt{30NKC_3\log\epsilon^{-1}}$. From Lemma 29, we have that $\|E_{\mathrm{FN}} X_{\mathrm{FN}}^\top\|_{\mathrm{F}} \le \sqrt{6K\log N}\|X_{\mathrm{FN}}\|$ when $n \ge (1/\delta)$ with probability exceeding $1-\delta$. From Lemma 25, we have for $\epsilon \le \frac{1}{240}\kappa^{-1}(\Sigma)$, the minimum eigenvalue satisfies $\lambda_{\min}(X_{\mathrm{TP}} X_{\mathrm{TP}}^\top) \ge (N/2)\cdot(1-2\epsilon)\cdot\lambda_{\min}(\Sigma) \ge (N/4)\cdot\lambda_{\min}(\Sigma)$. From our assumption of corrupted covariates, we have that $\|X_{\mathrm{FP}}\| \le \sqrt{\epsilon N C_3}$. We also have from Lemma 25, we have $\|X_{\mathrm{FN}}\| \le \sqrt{\lambda_{\max}(\Sigma)\cdot(10N\epsilon\log\epsilon^{-1})}$ with high probability. Then when $\epsilon \le \sqrt{\frac{1}{960C_3}\cdot\kappa^{-1}(\Sigma)\lambda_{\min}(\Sigma)}$, we have that $\|X_{\mathrm{FP}}\|_2\|X_{\mathrm{FN}}\|_2 \le (N/8)\lambda_{\min}(\Sigma)$. Then, solving for the induction with an infinite sum, we have after $O\left(\kappa(\Sigma)\cdot\log\left(\frac{\|W^*\|_{\mathrm{F}}}{\varepsilon}\right)\right)$ iterations,

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

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

B Proofs for Learning Nonlinear Neurons

B.1 Sigmoid Neurons

B.1.1 Proof of Theorem 17

Proof. From Algorithm 1, we have the gradient update for learning a Sigmoid neuron for the ℓ_2 loss.

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

Our proof will follow a similar structure to the proof for linear regression. We first show we can obtain linear convergence of $\mathbf{w}^{(t)}$ to $\mathbf{w}^*$ with some error. Then we rigorously analyze the concentration inequalities to give crisp bounds on the upper bound for ϵ and show the noise term is $O(\epsilon\log\epsilon^{-1})$.
Step 1: Linear Convergence.

$$\begin{aligned}\|\mathbf{w}^{(t+1)} - \mathbf{w}^*\|_2 &= \|\mathbf{w}^{(t)} - \eta\nabla\mathcal{R}(\mathbf{w}^{(t)}; \mathrm{S}^{(t)}) - \mathbf{w}^*\|_2 \\ &= \|\mathbf{w}^{(t)} - \mathbf{w}^* - \eta\nabla\mathcal{R}(\mathbf{w}^{(t)}; \mathrm{TP}) - \eta\nabla\mathcal{R}(\mathbf{w}^{(t)}; \mathrm{FP})\|_2 \\ &\le \|\mathbf{w}^{(t)} - \mathbf{w}^* - \eta\nabla\mathcal{R}(\mathbf{w}^{(t)}; \mathrm{TP})\|_2 + \|\eta\nabla\mathcal{R}(\mathbf{w}^{(t)}; \mathrm{FP})\|_2 \end{aligned} \tag{6}$$

We upper bound the first term of Equation (6) through its square.

$$\|\mathbf{w}^{(t)} - \mathbf{w}^* - \eta\nabla\mathcal{R}(\mathbf{w}^{(t)}; \mathrm{TP})\|_2^2 = \|\mathbf{w}^{(t)} - \mathbf{w}^*\|_2^2 - 2\eta\cdot\langle\mathbf{w}^{(t)} - \mathbf{w}^*, \nabla\mathcal{R}(\mathbf{w}^{(t)}; \mathrm{TP})\rangle + \eta^2\cdot\|\nabla\mathcal{R}(\mathbf{w}^{(t)}; \mathrm{TP})\|_2^2 \tag{7}$$

We now lower bound the second term of Equation (7). Note from the randomized initialization, we have $\|\mathbf{w}^{(0)} - \mathbf{w}^*\| \le \mathbf{w}^*$. Then, noting that $\|\mathbf{w}^*\| \le R$, we have by the Cauchy-Schwarz inequality, for any

$\mathbf{x} \sim \mathcal{P}$, we have $|\mathbf{w}^{(t)} \cdot \mathbf{x}| \leq 2RB$ almost surely. Here we leverage Property 13 and note there exists a γ s.t. $\sigma'(x) \geq \gamma > 0$ for all $x \in \mathbb{R}$ s.t. $x \leq 2RB$.

$$\begin{aligned}
&2\eta \cdot \langle \mathbf{w}^{(t)} - \mathbf{w}^*, \nabla \mathcal{R}(\mathbf{w}^{(t)}; \mathrm{TP}) \rangle = \frac{4\eta}{(1-\epsilon)N} \cdot \langle \mathbf{w}^{(t)} - \mathbf{w}^*, \sum_{i \in \mathrm{TP}} (\sigma(\mathbf{w}^{(t)} \cdot \mathbf{x}_i) - y_i) \cdot \sigma'(\mathbf{w}^{(t)} \cdot \mathbf{x}_i) \cdot \mathbf{x}_i \rangle \\
&= \frac{4\eta}{(1-\epsilon)N} \cdot \langle \mathbf{w}^{(t)} - \mathbf{w}^*, \sum_{i \in \mathrm{TP}} (\sigma(\mathbf{w}^{(t)} \cdot \mathbf{x}_i) - \sigma(\mathbf{w}^* \cdot \mathbf{x}_i) + \xi_i) \cdot \sigma'(\mathbf{w}^{(t)} \cdot \mathbf{x}_i) \cdot \mathbf{x}_i \rangle \\
&\geq \frac{4\eta}{(1-\epsilon)N} \cdot \gamma^2 \lambda_{\min}(X_{\mathrm{TP}} X_{\mathrm{TP}}^\top) \|\mathbf{w}^{(t)} - \mathbf{w}^*\|_2^2 - \frac{4\eta}{(1-\epsilon)N} \cdot \|\mathbf{w}^{(t)} - \mathbf{w}^*\|_2 \Big\| \sum_{i \in \mathrm{TP}} \xi_i \sigma'(\mathbf{w}^{(t)} \cdot \mathbf{x}_i) \cdot \mathbf{x}_i \Big\|_2 \qquad (8)
\end{aligned}$$

In the above, in the final relation, we apply Cauchy-Schwarz inequality. Then from an application of Young's Inequality, we obtain

$$\begin{aligned}
&\frac{4\eta}{(1-\epsilon)N} \cdot \|\mathbf{w}^{(t)} - \mathbf{w}^*\|_2 \Big\| \sum_{i \in \mathrm{TP}} \xi_i \sigma'(\mathbf{w}^{(t)} \cdot \mathbf{x}_i) \cdot \mathbf{x}_i \Big\|_2 \\
&\quad \leq \frac{\eta}{(1-\epsilon)N} \cdot \gamma^2 \lambda_{\min}(X_{\mathrm{TP}} X_{\mathrm{TP}}^\top) \|\mathbf{w}^{(t)} - \mathbf{w}^*\|_2^2 + \frac{4\eta}{(1-\epsilon)N} \cdot \lambda_{\min}^{-1}(X_{\mathrm{TP}} X_{\mathrm{TP}}^\top) \Big\| \sum_{i \in \mathrm{TP}} \xi_i \sigma'(\mathbf{w}^{(t)} \cdot \mathbf{x}_i) \cdot \mathbf{x}_i \Big\|_2^2
\end{aligned}$$

We next upper bound the third term in Equation (7). We note that σ is $\|\sigma\|_{\mathrm{lip}}$-Lipschitz, then from an application of the triangle inequality, we obtain

$$\begin{aligned}
\eta \cdot \|\nabla \mathcal{R}(\mathbf{w}^{(t)}; \mathrm{TP})\|_2^2 &= \frac{4\eta^2}{[(1-\epsilon)N]^2} \cdot \Big\| \sum_{i \in \mathrm{TP}} (\sigma(\mathbf{w}^{(t)} \cdot \mathbf{x}_i) - \sigma(\mathbf{w}^* \cdot \mathbf{x}_i) + \xi_i) \cdot \sigma'(\mathbf{w}^{(t)} \cdot \mathbf{x}_i) \cdot \mathbf{x}_i \Big\|_2^2 \\
&\leq \frac{8\eta^2}{[(1-\epsilon)N]^2} \cdot \left(\|\sigma\|_{\mathrm{lip}}^4 \lambda_{\max}^2(X_{\mathrm{TP}} X_{\mathrm{TP}}^\top) \|\mathbf{w}^{(t)} - \mathbf{w}^*\|_2^2 + \Big\| \sum_{i \in \mathrm{TP}} \xi_i \sigma'(\mathbf{w}^{(t)} \cdot \mathbf{x}_i) \cdot \mathbf{x}_i \Big\|_2^2 \right) \quad (9)
\end{aligned}$$

Then, from choosing $\eta \leq \frac{\gamma^2 (1-\epsilon) N \lambda_{\min}(X_{\mathrm{TP}} X_{\mathrm{TP}}^\top)}{4\|\sigma\|_{\mathrm{lip}}^4 \lambda_{\max}^2(X_{\mathrm{TP}} X_{\mathrm{TP}}^\top)}$. We see the first term of Equation (9) is less than half the first term in Equation (8). We now bound the second term in Equation (6).

$$\begin{aligned}
\eta^2 \cdot \|\nabla \mathcal{R}(\mathbf{w}^{(t)}; \mathrm{FP})\|_2^2 &= \frac{4\eta^2}{[(1-\epsilon)N]^2} \cdot \Big\| \sum_{i \in \mathrm{FP}} (\sigma(\mathbf{w}^{(t)} \cdot \mathbf{x}_i) - y_i) \cdot \sigma'(\mathbf{w} \cdot \mathbf{x}_i) \cdot \mathbf{x}_i \Big\|_2^2 \\
&\overset{(i)}{\leq} \frac{4\eta^2}{[(1-\epsilon)N]^2} \cdot \|\sigma\|_{\mathrm{lip}}^2 \|X_{\mathrm{FP}} X_{\mathrm{FP}}^\top\|_2 \sum_{i \in \mathrm{FP}} (\sigma(\mathbf{w}^{(t)} \cdot \mathbf{x}_i) - y_i)^2 \\
&\overset{(ii)}{\leq} \frac{4\eta^2}{[(1-\epsilon)N]^2} \cdot \|\sigma\|_{\mathrm{lip}}^2 \|X_{\mathrm{FP}} X_{\mathrm{FP}}^\top\|_2 \sum_{i \in \mathrm{FN}} (\sigma(\mathbf{w}^{(t)} \cdot \mathbf{x}_i) - y_i)^2 \\
&= \frac{4\eta^2}{[(1-\epsilon)N]^2} \cdot \|\sigma\|_{\mathrm{lip}}^2 \|X_{\mathrm{FP}} X_{\mathrm{FP}}^\top\|_2 \sum_{i \in \mathrm{FN}} (\sigma(\mathbf{w}^{(t)} \cdot \mathbf{x}_i) - \sigma(\mathbf{w}^* \cdot \mathbf{x}_i) + \xi_i)^2 \\
&\leq \frac{8\eta^2}{[(1-\epsilon)N]^2} \cdot \|\sigma\|_{\mathrm{lip}}^2 \|X_{\mathrm{FP}} X_{\mathrm{FP}}^\top\|_2 \Big(\|\sigma\|_{\mathrm{lip}}^2 \cdot \|X_{\mathrm{FN}} X_{\mathrm{FN}}^\top\|_2 \|\mathbf{w}^{(t)} - \mathbf{w}^*\|_2^2 + \|\boldsymbol{\xi}_{\mathrm{FN}}\|_2^2 \Big) \quad (10)
\end{aligned}$$

In the above, (i) follows from Lemma 30 and (ii) follows from the optimality of the Hard-Thresholding operator. Concluding the step, we have from Equations (7) and (10),

$$\begin{aligned}
\|\mathbf{w}^{(t+1)} - \mathbf{w}^*\|_2 &\leq \|\mathbf{w}^{(t)} - \mathbf{w}^*\|_2 \left(1 - \frac{\eta}{2(1-\epsilon)N} \cdot \gamma^2 \lambda_{\min}(X_{\mathrm{TP}} X_{\mathrm{TP}}^\top) + \frac{\sqrt{8}\eta}{(1-\epsilon)N} \cdot \|\sigma\|_{\mathrm{lip}}^2 \|X_{\mathrm{FP}}\|_2 \|X_{\mathrm{FN}}\|_2 \right) \\
&+ \left(\frac{\sqrt{8}\eta}{(1-\epsilon)N} + \frac{2\sqrt{\eta}}{\sqrt{(1-\epsilon)N}} \cdot \sigma_{\min}^{-1}(X_{\mathrm{TP}}) \right) \Big\| \sum_{i \in \mathrm{TP}} \xi_i \sigma'(\mathbf{w}^{(t)} \cdot \mathbf{x}_i) \cdot \mathbf{x}_i \Big\|_2 + \frac{\sqrt{8}\eta}{(1-\epsilon)N} \cdot \|X_{\mathrm{FP}}\|_2 \|\boldsymbol{\xi}_{\mathrm{FN}}\|_2
\end{aligned}$$

Step 2: Concentration Bounds. From Lemma 26, we have with probability at least $1-\delta$,

$$\|\sum_{i\in\mathrm{TP}} \xi_i \sigma'(\mathbf{w}^{(t)}\cdot\mathbf{x}_i)\cdot\mathbf{x}_i\|_2 \le NC_\Sigma\|\sigma\|_{\mathrm{lip}}^2 C_\nu\left(\frac{2d}{N}\log(5)+\frac{2}{N}\log(1/\delta)+6\epsilon\log\epsilon^{-1}\right)^{1/2}$$

Recall that $\|X_{\mathrm{FP}}\|_2 \le \sqrt{N\epsilon C_Q}$. Then, with probability at least $1-\delta$, we have $\|X_{\mathrm{FN}}\|_2\|X_{\mathrm{FP}}\|_2 \le N\epsilon\cdot\sqrt{10C_3\log\epsilon^{-1}}$. We also have $\|\boldsymbol{\xi}_{\mathrm{FN}}\|_2 \le \sqrt{30N\epsilon\log\epsilon^{-1}}$. We then find for $\epsilon \le \frac{C^2_{\lfloor\sigma\rfloor}}{\|\sigma\|^2_{\mathrm{lip}}}\frac{\lambda_{\min}(\Sigma)}{32\sqrt{80\log(1/2)}}$ and probability exceeding $1-\delta$,

$$\|\sigma\|_{\mathrm{lip}}^2\|X_{\mathrm{FP}}\|_2\|X_{\mathrm{FN}}\|_2 \le \frac{1}{\sqrt{144}}\cdot\gamma^2\lambda_{\min}(\Sigma)$$

Then combining the results with our choice of η, we obtain,

$$\begin{aligned}\|\mathbf{w}^{(t+1)}-\mathbf{w}^*\|_2 &\le \varepsilon+\frac{1}{\lambda_{\min}(\Sigma)}\cdot\gamma^{-2}\left(8+16\sqrt{8}\kappa^2(\Sigma)\right)\left(C_\Sigma\|\sigma\|_{\mathrm{lip}}^2C_\nu\left(\frac{2d}{N}\log(5)+\frac{2}{N}\log(1/\delta)+6\epsilon\log\epsilon^{-1}\right)^{1/2}\right)\\ &\le \varepsilon+\frac{1}{\lambda_{\min}(\Sigma)}\cdot\gamma^{-2}\|\sigma\|_{\mathrm{lip}}^2\nu\left(24\sqrt{96}\kappa^2(\Sigma)\right)\sqrt{\epsilon\log\epsilon^{-1}}\end{aligned}$$

In the above, the final inequality holds when $N \ge \frac{2d\log 5+\log(1/\delta)}{6\epsilon\log\epsilon^{-1}} = \Omega\left(\frac{d+\log(1/\delta)}{\epsilon}\right)$. Our proof is copmllete. ■

B.2 Leaky-ReLU Neuron

B.2.1 Proof of Theorem 18

Proof. We will decompose the gradient into the good component and corrupted component. The first part of our proof will show that $\mathbf{w}$ moves in the direction of $\mathbf{w}^*$, then in the second part of the proof we will show the affect of the corrupted gradient. Finally, we combine step 1 and step 2 to show that there exists sufficiently small ϵ such that we can get linear convergence with a small additive error term.

Step 1: Upper bounding the ℓ_2 norm distance between $\mathbf{w}^{(t+1)}$ and $\mathbf{w}^*$. We have from Algorithm 1,

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

We will first upper bound the first term of Equation (11) through its square.

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

In the above, the a.s. relation follows from Property **??**. We will first lower bound the second term of Equation (12). We will first bound the spectrum of $\nabla^2\mathcal{L}(\mathbf{w};\mathrm{TP})$ where $\mathbf{w}\in\mathbb{R}^d$ is an arbitrary vector.

$$\nabla^2\mathcal{L}(\mathbf{w};\mathrm{TP}) = 2\cdot\sum_{i\in\mathrm{TP}}(\sigma(\mathbf{w}\cdot\mathbf{x}_i)-y_i)\cdot\sigma''(\mathbf{w}\cdot\mathbf{x}_i)\cdot\mathbf{x}_i\mathbf{x}_i^\top+2\cdot\sum_{i\in\mathrm{TP}}\left[\sigma'(\mathbf{w}\cdot\mathbf{x}_i)\right]^2\cdot\mathbf{x}_i\mathbf{x}_i^\top$$

Then, from noting that the second derivative of Leaky-ReLU is non-zero at one point, we have

$$\nabla^2\mathcal{L}(\mathbf{w};\mathrm{TP}) \overset{\text{a.s.}}{=} 2\cdot\sum_{i\in\mathrm{TP}}\left[\sigma'(\mathbf{w}\cdot\mathbf{x}_i)\right]^2\cdot\mathbf{x}_i\mathbf{x}_i^\top$$

We then obtain for any $\mathbf{w}\in\mathbb{R}^d$, almost surely,

$$2\cdot\gamma^2\lambda_{\min}(X_{\mathrm{TP}}X_{\mathrm{TP}}^\top)\cdot I \preceq \nabla^2\mathcal{L}(\mathbf{w};\mathrm{TP}) \preceq 2\cdot\|\sigma\|_{\mathrm{lip}}^2\lambda_{\max}(X_{\mathrm{TP}}X_{\mathrm{TP}}^\top)\cdot I \tag{13}$$

We can now lower bound the second term of Equation (12). We have from the convexity of the Leaky-ReLU,

$$\begin{aligned}
2\eta\cdot&\langle \mathbf{w}^{(t)} - \mathbf{w}^*, \nabla\mathcal{R}(\mathbf{w}^{(t)};\mathrm{TP}) - \nabla\mathcal{R}(\mathbf{w}^{(*)};\mathrm{TP})\rangle \\
&= \frac{4\eta}{(1-\epsilon)N}\cdot\langle \mathbf{w}^{(t)} - \mathbf{w}^*, \int_0^1 \nabla^2\mathcal{R}(\mathbf{w}^* + \theta(\mathbf{w}^* - \mathbf{w}^{(t)});\mathrm{TP})d\theta\cdot(\mathbf{w}^{(t)} - \mathbf{w}^*)\rangle \\
&\overset{(13)}{\geq} \frac{4\eta}{(1-\epsilon)N}\cdot C^2_{\lfloor\sigma\rfloor}\lambda_{\min}(X_{\mathrm{TP}}X_{\mathrm{TP}}^\top)\|\mathbf{w}^{(t)} - \mathbf{w}^*\|_2^2
\end{aligned} \tag{14}$$

We now will upper bound the third term of Equation (12) with a similar argument to our bound in Equation (14).

$$\begin{aligned}
\eta^2\cdot&\|\nabla\mathcal{R}(\mathbf{w}^{(t)};\mathrm{TP}) - \nabla\mathcal{R}(\mathbf{w}^*;\mathrm{TP})\|_2^2 \\
&= \frac{4\eta^2}{[(1-\epsilon)N]^2}\cdot\Big\|\int_0^1 \nabla^2\mathcal{R}(\mathbf{w}^* + \theta(\mathbf{w}^* - \mathbf{w}^{(t)});\mathrm{TP})d\theta\cdot(\mathbf{w}^{(t)} - \mathbf{w}^*)\Big\|_2^2 \\
&\overset{(13)}{\leq} \frac{4\eta^2}{[(1-\epsilon)N]^2}\cdot C^2_{\lceil\sigma\rceil}\lambda^2_{\max}(X_{\mathrm{TP}}X_{\mathrm{TP}}^\top)\|\mathbf{w}^{(t)} - \mathbf{w}^*\|_2^2
\end{aligned} \tag{15}$$

where (ii) follows from the $\|\sigma\|_{\mathrm{lip}}$-Lipschitzness of σ given in Property **??**. We then observe that the first term of Equation (14) is greater than half the first term in Equation (15) when $\eta \leq \frac{C_\sigma^2(1-\epsilon)N\lambda_{\min}(X_{\mathrm{TP}}X_{\mathrm{TP}}^\top)}{2\lambda^2_{\max}(X_{\mathrm{TP}}X_{\mathrm{TP}}^\top)}$.
Step 2: Upper bounding the corrupted gradient. We now upper bound the third term in Equation (12).

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

In the above, the first inequality follows from Lemma 30, the second inequality follows from the optimality of the Subquantile set, the final inequality follows from the C_2-Lipschitzness of σ. Then from Lemma 26, we have with probability at least $1-\delta$,

$$\Big\|\sum_{i\in\mathrm{TP}}\xi_i\sigma'(\mathbf{w}^{(t)}\cdot\mathbf{x}_i)\cdot\mathbf{x}_i\Big\|_2 \leq 9NC_\Sigma C_\nu\left(\frac{2d}{N}\log 12 + \frac{2Rd}{N}\log 12 + \frac{2}{N}\log(1/\delta) + 6\epsilon\log\epsilon^{-1}\right)^{1/2}$$

We now combine Steps 1 and 2 to give the linear convergence result. Noting that $\sqrt{1-2x} \leq 1-x$ when $x \leq 1/2$, we have

$$\begin{aligned}
\|\mathbf{w}^{(t+1)} - \mathbf{w}^*\|_2 \leq \|\mathbf{w}^{(t)} - \mathbf{w}^*\|_2&\left(1 - \frac{2C_\sigma^2\eta}{(1-\epsilon)N}\cdot\lambda_{\min}(X_{\mathrm{TP}}X_{\mathrm{TP}}^\top) + \frac{2\eta}{(1-\epsilon)N}\cdot\|X_{\mathrm{FP}}\|_2\|X_{\mathrm{FN}}\|_2\right) \\
&+ \frac{15C_\Sigma C_\nu d\log 5}{\sqrt{(1-\epsilon)N}} + \frac{15C_\Sigma C_\nu\log(1/\delta)}{\sqrt{(1-\epsilon)N}} + 16C_\Sigma C_\nu\sqrt{\epsilon\log\epsilon^{-1}}
\end{aligned}$$

Step 3: Concentration Bounds. We will give the relevant probabilistic bounds for the random variables in Steps 1 and 2. From Lemma 25, we have $\|X_{\mathrm{FN}}\|_2\|X_{\mathrm{FP}}\|_2 \leq \epsilon\sqrt{\lambda_{\max}(\Sigma)\cdot 10C_3N\log\epsilon^{-1}}$ with probability at least $1-\delta$ when $N \geq \frac{2}{\epsilon}\cdot\left(dC_\Sigma^2 + \frac{\log(2/\delta)}{c_K}\right)$ when $\epsilon \leq \frac{1}{60}\cdot\kappa^{-1}(\Sigma)$. From the same Lemma and under the same data conditions we have $\lambda_{\min}(X_{\mathrm{TP}}X_{\mathrm{TP}}^\top) \geq \frac{1}{4}\cdot\lambda_{\min}(\Sigma)$. Then for $\epsilon \leq \frac{C_\sigma^2\lambda_{\min}(\Sigma)}{\sqrt{32C_3\lambda_{\max}(\Sigma)}}$, we have $\|X_{\mathrm{FP}}\|_2\|X_{\mathrm{FN}}\|_2 \leq \frac{1}{2} \cdot \lambda_{\min}(\Sigma)$. We then have, after $O\Big(\kappa^2(\Sigma)\log\Big(\frac{\|\mathbf{w}^*\|_2+10d}{\varepsilon}\Big)\Big)$ iterations with high probability,

$$\begin{aligned}\|\mathbf{w}^{(T)} - \mathbf{w}^*\|_2 &\leq \varepsilon + \frac{1}{N} \cdot 144 C_\sigma^{-2}\kappa^2(\Sigma) K C_\nu d \log 5 + 432\kappa^2(\Sigma) C_\sigma^{-1} K C_\nu d\epsilon \log \epsilon^{-1} \\ &= O\big(\kappa^2(\Sigma) C_\sigma^{-2} K C_\nu d\epsilon \log \epsilon^{-1}\big)\end{aligned}$$

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

B.3 ReLU Neuron

In this section, we consider functions such as the ReLU. Our high-level analysis will be similar to the previous sub-sections however the details are considerably different and require stronger conditions we can guarantee by randomness.

B.3.1 Proof of Theorem 20

Proof. We will now begin our standard analysis.

$$\begin{aligned}\|\mathbf{w}^{(t+1)} - \mathbf{w}^*\|_2 &= \|\mathbf{w}^{(t)} - \mathbf{w}^* - \eta\nabla\mathcal{R}(\mathbf{w}; \mathrm{S}^{(t)})\|_2 \\ &= \|\mathbf{w}^{(t)} - \mathbf{w}^* - \eta\nabla\mathcal{R}(\mathbf{w}^{(t)}; \mathrm{TP}) - \eta\nabla\mathcal{R}(\mathbf{w}^{(t)}; \mathrm{FP})\|_2 \\ &\leq \|\mathbf{w}^{(t)} - \mathbf{w}^* - \eta\nabla\mathcal{R}(\mathbf{w}^{(t)}; \mathrm{TP})\|_2 + \|\eta\nabla\mathcal{R}(\mathbf{w}^{(t)}; \mathrm{FP})\|_2 \end{aligned} \tag{16}$$

We will now upper bound the first term of Equation (16) through its square,

$$\begin{aligned}&\|\mathbf{w}^{(t)} - \mathbf{w}^* - \eta\nabla\mathcal{R}(\mathbf{w}^{(t)}; \mathrm{TP})\|_2^2 = \\ &\qquad \|\mathbf{w}^{(t)} - \mathbf{w}^*\|_2^2 - 2\eta \cdot \langle \mathbf{w}^{(t)} - \mathbf{w}^*, \nabla\mathcal{R}(\mathbf{w}^{(t)}; \mathrm{TP})\rangle + \eta^2 \cdot \|\nabla\mathcal{R}(\mathbf{w}^{(t)}; \mathrm{TP})\|_2^2\end{aligned} \tag{17}$$

We will lower bound the second term of Equation (17). We will first adopt the notation from [ZYWG19], let $\Sigma_{\mathrm{TP}}(\mathbf{w}, \hat{\mathbf{w}}) = X_{\mathrm{TP}}^\top X_{\mathrm{TP}}^\top \cdot \mathbf{1}\{X_{\mathrm{TP}}^\top \mathbf{w} \geq \mathbf{0}\} \cdot \mathbf{1}\{X_{\mathrm{TP}}^\top \hat{\mathbf{w}} \geq \mathbf{0}\}$, it then follows

$$\begin{aligned}&2\eta \cdot \langle \mathbf{w}^{(t)} - \mathbf{w}^*, \nabla\mathcal{R}(\mathbf{w}^{(t)}; \mathrm{TP})\rangle \\ &\stackrel{\text{def}}{=} \frac{4\eta}{(1-\epsilon)N} \cdot \langle \mathbf{w}^{(t)} - \mathbf{w}^*, \sum_{i\in\mathrm{TP}} (\sigma(\mathbf{w}^{(t)} \cdot \mathbf{x}_i) - y_i) \cdot \mathbf{x}_i \cdot \mathbf{1}\{\mathbf{w}^{(t)} \cdot \mathbf{x}_i \geq 0\}\rangle \\ &= \frac{4\eta}{(1-\epsilon)N} \cdot \langle \mathbf{w}^{(t)} - \mathbf{w}^*, \Sigma_{\mathrm{TP}}(\mathbf{w}^{(t)}, \mathbf{w}^{(t)})\mathbf{w}^{(t)} - \Sigma_{\mathrm{TP}}(\mathbf{w}^{(t)}, \mathbf{w}^*)\mathbf{w}^*\rangle \\ &= \frac{4\eta}{(1-\epsilon)N} \cdot \langle \mathbf{w}^{(t)} - \mathbf{w}^*, \Sigma_{\mathrm{TP}}(\mathbf{w}^{(t)}, \mathbf{w}^*)(\mathbf{w}^{(t)} - \mathbf{w}^*) + \Sigma_{\mathrm{TP}}(\mathbf{w}^{(t)}, -\mathbf{w}^*)\mathbf{w}^{(t)}\rangle \\ &\geq \frac{4\eta}{(1-\epsilon)N} \cdot \lambda_{\min}(\Sigma_{\mathrm{TP}}(\mathbf{w}^{(t)}, \mathbf{w}^*))\|\mathbf{w}^{(t)} - \mathbf{w}^*\|_2^2\end{aligned}$$

In the above, the final inequality holds from the following relation,

$$\begin{aligned}\langle \mathbf{w}^{(t)} - \mathbf{w}^*, \Sigma_{\mathrm{TP}}(\mathbf{w}^{(t)}, -\mathbf{w}^*)\mathbf{w}^{(t)}\rangle &= \langle \mathbf{w}^{(t)} - \mathbf{w}^*, \sum_{i\in\mathrm{TP}} \mathbf{x}_i\mathbf{w}^{(t)} \cdot \mathbf{x}_i \cdot \mathbf{1}\{\mathbf{w}^{(t)} \cdot \mathbf{x}_i \geq 0\} \cdot \mathbf{1}\{\mathbf{w}^* \cdot \mathbf{x}_i \leq 0\}\rangle \\ &= \sum_{i\in\mathrm{TP}} (\mathbf{w}^{(t)} \cdot \mathbf{x}_i - \mathbf{w}^* \cdot \mathbf{x}_i)(\mathbf{w}^{(t)} \cdot \mathbf{x}_i) \cdot \mathbf{1}\{\mathbf{w}^{(t)} \cdot \mathbf{x}_i \geq 0\} \cdot \mathbf{1}\{\mathbf{w}^* \cdot \mathbf{x}_i \leq 0\} \geq 0\end{aligned}$$

In the above, in the final relation we can note that when the indicators are positive, it must follow that both $\mathbf{w}^{(t)} \cdot \mathbf{x}_i$ is positive and $\mathbf{w}^{(t)} \cdot \mathbf{x}_i \geq \mathbf{w}^* \cdot \mathbf{x}_i$ as $\mathbf{w}^* \cdot \mathbf{x}_i \leq 0$. We have from Weyl's Inequality,

$$\lambda_{\min}\Big(\Sigma_{\mathrm{TP}}(\mathbf{w}^{(t)}, \mathbf{w}^*)\Big) \geq \lambda_{\min}\Big(\mathbf{E}\Big[\Sigma_{\mathrm{TP}}(\mathbf{w}^{(t)}, \mathbf{w}^*)\Big]\Big) - \|\Sigma_{\mathrm{TP}}(\mathbf{w}^{(t)}, \mathbf{w}^*) - \mathbf{E}\Big[\Sigma_{\mathrm{TP}}(\mathbf{w}^{(t)}, \mathbf{w}^*)\Big]\|_2$$

Let $\Omega = \left\{\mathbf{x} \in \mathbb{R}^d : \mathbf{x}^\top \mathbf{w}^{(t)} \geq 0, \mathbf{x}^\top \mathbf{w}^* \geq 0\right\}$, then

$$\begin{aligned}
\mathbf{E}\Big[\Sigma_{\mathrm{TP}}(\mathbf{w}^{(t)}, \mathbf{w}^*)\Big] &= \sum_{i \in \mathrm{TP}} \mathbf{E}\Big[\mathbf{x}_i \mathbf{x}_i^\top \cdot \mathbf{1}\{\mathbf{w}^{(t)} \cdot \mathbf{x}_i \geq 0\} \cdot \mathbf{1}\{\mathbf{w}^* \cdot \mathbf{x}_i \geq 0\}\Big] \\
&\overset{(i)}{\succeq} N(1-2\epsilon) \cdot \left(\pi - \Theta^{(t)} - \sin \Theta^{(t)}\right) \cdot I \\
&\succeq N(1-2\epsilon) \cdot \left(\pi - 2 \arcsin\left(\frac{\|\mathbf{w}^{(t)} - \mathbf{w}^*\|}{\|\mathbf{w}^*\|}\right)\right) \cdot I \\
&\succeq N(1-2\epsilon) \cdot \pi \left(1 - \frac{\|\mathbf{w}^{(t)} - \mathbf{w}^*\|}{\|\mathbf{w}^*\|}\right) \cdot I
\end{aligned}$$

In the above, (i) follows from Lemma 28, (ii) follows from the guarantee in the randomized initialization. We can then note We now bound the second-moment matrix approximation. Let $dP(\mathbf{x})$ be a Dirac-measure for $\mathbf{x} \in \mathrm{TP}$. We will now upper bound the third term in Equation (17).

$$\begin{aligned}
\eta^2 \cdot \|\nabla \mathcal{R}(\mathbf{w}^{(t)}; \mathrm{TP})\|_2^2 &= \frac{4\eta^2}{[(1-\epsilon)N]^2} \cdot \Big\| \sum_{i \in \mathrm{TP}} (\sigma(\mathbf{w}^{(t)} \cdot \mathbf{x}_i) - \sigma(\mathbf{w}^* \cdot \mathbf{x}_i) - \xi_i) \cdot \mathbf{x}_i \cdot \mathbf{1}\{\mathbf{w}^{(t)} \cdot \mathbf{x}_i \geq 0\} \Big\|_2^2 \\
&\leq \frac{8\eta^2}{[(1-\epsilon)N]^2} \cdot \Big\| \sum_{i \in \mathrm{TP}} (\sigma(\mathbf{w}^{(t)} \cdot \mathbf{x}_i) - \sigma(\mathbf{w}^* \cdot \mathbf{x}_i)) \cdot \mathbf{x}_i \cdot \mathbf{1}\{\mathbf{w}^{(t)} \cdot \mathbf{x}_i \geq 0\} \Big\|_2^2 \\
&\quad + \frac{8\eta^2}{[(1-\epsilon)N]^2} \cdot \Big\| \sum_{i \in \mathrm{TP}} \xi_i \mathbf{x}_i \cdot \mathbf{1}\{\mathbf{w}^{(t)} \cdot \mathbf{x}_i \geq 0\} \Big\|_2^2
\end{aligned} \tag{18}$$

Recall we have from Lemma 26, we have an upper bound on the second term of Equation (18). We give a bound on the first term of Equation (18).

$$\begin{aligned}
&\frac{8\eta^2}{[(1-\epsilon)N]^2} \cdot \Big\| \sum_{i \in \mathrm{TP}} (\sigma(\mathbf{w}^{(t)} \cdot \mathbf{x}_i) - \sigma(\mathbf{w}^* \cdot \mathbf{x}_i)) \cdot \mathbf{x}_i \cdot \mathbf{1}\{\mathbf{w}^{(t)} \cdot \mathbf{x}_i \geq 0\} \Big\|_2^2 \\
&\qquad \leq \frac{8\eta^2}{[(1-\epsilon)N]^2} \cdot \big\| X_{\mathrm{TP}} X_{\mathrm{TP}}^\top \big\|_2 \sum_{i \in \mathrm{TP}} (\sigma(\mathbf{w}^{(t)} \cdot \mathbf{x}_i) - \sigma(\mathbf{w}^* \cdot \mathbf{x}_i))^2 \\
&\qquad \leq \frac{8\eta^2}{[(1-\epsilon)N]^2} \cdot \big\| X_{\mathrm{TP}} X_{\mathrm{TP}}^\top \big\|_2^2 \|\mathbf{w}^{(t)} - \mathbf{w}^*\|_2^2
\end{aligned} \tag{19}$$

Then, by choosing $\eta \leq \frac{\lambda_{\min}(\Sigma)}{80\lambda_{\max}^2(\Sigma)}$, we have that the RHS in Equation (19) will be less than $\frac{\lambda_{\min}(\Sigma)}{8}$.

Step 3: Upper bounding the corrupted gradient. We now upper bound the third term in Equation (17).

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

In the above, the first inequality follows from the same argument as Lemma 30, the second inequality follows from the optimality of the Subquantile set, and the final inequality follows from noting that σ is 1-Lipschitz. We now conclude Steps 1-3 with our linear convergence result.

$$\|\mathbf{w}^{(t+1)} - \mathbf{w}^*\|_2 \le \|\mathbf{w}^{(t)} - \mathbf{w}^*\| \left(1 - \frac{\eta}{32} \cdot \lambda_{\min}(\Sigma) + \frac{2\eta}{(1-\epsilon)N} \cdot \|X_{\mathrm{FP}}\|_2 \|X_{\mathrm{FN}}\|\right) + \frac{16}{3} K C_\nu \cdot \left(\frac{2d}{N}\log(5) + \frac{2}{N}\log(1/\delta) + 6\epsilon \log \epsilon^{-1}\right)^{1/2}$$

Step 4: Concentration Inequalities. From our previous theorems, we have $\|X_{\mathrm{FP}}\|_2 \|X_{\mathrm{FN}}\|_2 \le N\epsilon\sqrt{10C_3 \log \epsilon^{-1}}$ with probability exceeding $1-\delta$ and $N = \Omega\Big(\frac{d+\log(2/\delta)}{\epsilon}\Big)$. If $N = \Omega\Big(\frac{2d\log 5 + \log(1/\delta)}{3\epsilon \log \epsilon^{-1}}\Big)$ and $\epsilon \le \frac{1}{64\sqrt{80\log(2)}}$, we obtain after $T = O\Big(\kappa^2(\Sigma)\log\Big(\frac{\|\mathbf{w}^*\|_2}{\varepsilon}\Big)\Big)$ gradient descent iterations,

$$\|\mathbf{w}^{(t+1)} - \mathbf{w}^*\|_2 \le \varepsilon + 1024\sqrt{\epsilon \log \epsilon^{-1}} = O\Big(\sqrt{\epsilon \log \epsilon^{-1}}\Big)$$

when we choose $\varepsilon = O\Big(\sqrt{\epsilon \log \epsilon^{-1}}\Big)$. Our proof is complete. ■

C Probability Theory

In this section we will give various concentration inequalities on functions of the good data.

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*

$$\mathbf{Pr}\big\{\|\boldsymbol{\xi}\|_2^2 \ge \sigma\big(n + 2\sqrt{nx} + 2x\big)\big\} \le e^{-x}$$

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 $\epsilon \in (0, 0.5)$ and let $\mathcal{N}_2$ represent all such subsets. Given a failure probability $\delta \in (0,1)$, when $n \ge \log(1/\delta)$, with probability exceeding $1-\delta$,*

$$\max_{S \in \mathcal{N}_2} \|\boldsymbol{\xi}_{\mathrm{S}}\|_2^2 \le \sigma\big(30 n\epsilon \log \epsilon^{-1}\big)$$

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

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

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

$$\mathbf{Pr}\Big\{\max_{S \in \mathcal{N}_2} \|\boldsymbol{\xi}\|_2^2 \ge \sigma\big(\epsilon n + 2\sqrt{\epsilon n x} + 2x\big)\Big\} \le \Big(\frac{e}{\epsilon}\Big)^{\epsilon n} \mathbf{Pr}\big\{\|\boldsymbol{\xi}\|_2^2 \ge \sigma\big(\epsilon n + 2\sqrt{\epsilon n x} + 2x\big)\big\} \le \Big(\frac{e}{\epsilon}\Big)^{\epsilon n} e^{-x}$$

In the first inequality we apply a union bound over $\mathcal{N}_2$ with Lemma 32, and in the second inequality we use Lemma 21. We then obtain with probability exceeding $1-\delta$,

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

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

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

$$\lambda_{\max}(XX^\top) \leq n \cdot \lambda_{\max}(\Sigma) + \lambda_{\max}(\Sigma) \cdot \left(C_\Sigma \cdot \sqrt{dn} + t \cdot \sqrt{n}\right)$$

Lemma 24 (Sub-Gaussian Matrix Frobenius Norm Estimation)**.** *Let $X \in \mathbb{R}^{d\times n}$ have columns sampled from a sub-Gaussian distribution with second-moment matrix Σ and sub-Gaussian norm C_Σ, then with probability at least $1 - \delta$,*

$$\|XX^\top\|_{\mathrm{F}} \leq O(\mathrm{Tr}(\Sigma))$$

Proof.

$$\begin{aligned}\|XX^\top\|_{\mathrm{F}} &\leq n\,\mathrm{Tr}(\Sigma) + d\lambda_{\max}(\Sigma) \cdot \|XX^\top - n\Sigma\|_2 \\ &\leq n\,\mathrm{Tr}(\Sigma) + d\lambda_{\max}(\Sigma) \cdot \left(C_\Sigma\sqrt{dn} + \sqrt{\frac{n}{c_K}\log(1/\delta)}\right)\end{aligned}$$

■

Lemma 25. *Let $X \in \mathbb{R}^{d\times n}$ have columns sampled from a sub-Gaussian distribution with sub-Gaussian norm K and second-moment matrix Σ. Let $S \subset [n]$ such that $|S| = \epsilon n$ for $\epsilon \in (0, 0.5)$ and let $\mathcal{N}_2$ represent all such subsets. Then with probability at least $1 - \delta$,*

$$\begin{aligned}&\max_{S\in\mathcal{N}_2} \lambda_{\max}(X_{\mathrm{S}}X_{\mathrm{S}}^\top) \leq \lambda_{\max}(\Sigma) \cdot \left(10n\epsilon\log\epsilon^{-1}\right) \\ &\min_{S\in\mathcal{N}_2} \lambda_{\min}(X_{[n]\setminus \mathrm{S}}X_{[n]\setminus\mathrm{S}}^\top) \geq \frac{n}{4} \cdot \lambda_{\min}(\Sigma)\end{aligned}$$

when

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

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

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

$$\begin{aligned}&\mathbf{Pr}\left\{\max_{S\in\mathcal{N}_2} \lambda_{\max}(X_{\mathrm{S}}X_{\mathrm{S}}^\top) \geq n\epsilon \cdot \lambda_{\max}(\Sigma) + \lambda_{\max}(\Sigma) \cdot \left(C_\Sigma \cdot \sqrt{dn\epsilon} + t\sqrt{n\epsilon}\right)\right\} \\ &\quad\leq \left(\frac{e}{\epsilon}\right)^{\epsilon n} \mathbf{Pr}\left\{\lambda_{\max}(X_{\mathrm{S}}X_{\mathrm{S}}^\top) \geq n\epsilon \cdot \lambda_{\max}(\Sigma) + \lambda_{\max}(\Sigma) \cdot \left(C_\Sigma \cdot \sqrt{dn\epsilon} + t\sqrt{n\epsilon}\right)\right\} \leq 2 \cdot \left(\frac{e}{\epsilon}\right)^{\epsilon n} e^{-c_K t^2}\end{aligned}$$

In the above, the first inequality follows from a union bound over $\mathcal{N}_2$ and Lemma 32, the second inequality follows from Lemma 23. Then from elementary inequalities, we obtain with probability $1 - \delta$,

$$\begin{aligned}\max_{S\in\mathcal{N}_2} \lambda_{\max}(X_{\mathrm{S}}X_{\mathrm{S}}^\top) &\leq n\epsilon \cdot \lambda_{\max}(\Sigma) + \lambda_{\max}(\Sigma) \cdot \left(C_\Sigma \cdot \sqrt{dn\epsilon} + \sqrt{\frac{1}{c_K}(n\epsilon \cdot \log(2/\delta) + 3n^2\epsilon^2\log\epsilon^{-1})}\right) \\ &\leq n \cdot \lambda_{\max}(\Sigma) \cdot (\epsilon + 3^{3/4}\epsilon\log\epsilon^{-1}) + \lambda_{\max}(\Sigma) \cdot \left(C_\Sigma \cdot \sqrt{dn\epsilon} + \sqrt{\frac{1}{c_K}n\epsilon \cdot \log(2/\delta)}\right) \\ &\leq \lambda_{\max}(\Sigma) \cdot \left(6n\epsilon\log\epsilon^{-1}\right) + \lambda_{\max}(\Sigma) \cdot \left(C_\Sigma \cdot \sqrt{dn\epsilon} + \sqrt{\frac{1}{c_K}n\epsilon \cdot \log(2/\delta)}\right) \\ &\leq \lambda_{\max}(\Sigma) \cdot \left(10n\epsilon\log\epsilon^{-1}\right)\end{aligned}$$

In the above, the last inequality holds when

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

and our proof of the upper bound for the maximal eigenvalue is complete. We have from Weyl's Inequality for any $S \in \mathcal{N}_2$,

$$\lambda_{\min}(X_{\mathrm{X}\setminus\mathrm{S}} X_{\mathrm{X}\setminus\mathrm{S}}^\top) = \lambda_{\min}(XX^\top - X_{\mathrm{S}} X_{\mathrm{S}}^\top) \geq \lambda_{\min}(XX^\top) - \lambda_{\max}(X_{\mathrm{S}} X_{\mathrm{S}}^\top)$$

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

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

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

Lemma 26. *Let $X = [\mathbf{x}_1, \dots, \mathbf{x}_N]$ be the data matrix such that for $i \in [N]$, $\mathbf{x}_i$ are sampled from a sub-Gaussian distribution with second-moment matrix Σ with sub-Gaussian norm C_Σ and ξ_i are sampled from sub-Gaussian distribution with sub-Gaussian norm C_σ. Assume $f : \mathbb{R}^d \to \mathbb{R}$ is bounded over $\mathbb{R}$ and Lipschitz. Let $\mathcal{S}$ represent all subsets of $[N]$ of size empty to $(1-\epsilon)N$. Set a failure probability $\delta \in (0,1)$, then with probability at least $1 - \delta$,*
Lipschitz and Bounded Function:

$$\max_{\mathrm{S}\in\mathcal{S}} \sup_{\mathbf{w}\in\mathcal{B}(\mathbf{0},R)} \|(\boldsymbol{\xi}_{\mathrm{S}} \circ f(X_{\mathrm{S}}^\top \mathbf{w}))^\top X_{\mathrm{S}}\|_2 \leq \frac{1}{c} 2\sqrt{2} N C_\Sigma C_\sigma \|f\|_\infty \left(\frac{2d}{N}\log(6) + \frac{2Rd}{N}\log(6) + \frac{2}{N}\log(2/\delta) + 6\epsilon \log \epsilon^{-1}\right)^{1/2}$$

Indicator Function:

$$\max_{\mathrm{S}\in\mathcal{S}} \sup_{\mathbf{w}\in\mathcal{B}(\mathbf{0},R)} \|(\boldsymbol{\xi}_{\mathrm{S}} \circ f(X_{\mathrm{S}}^\top \mathbf{w}))^\top X_{\mathrm{S}}\|_2 \leq \frac{1}{c} 2\sqrt{2} N C_\Sigma C_\sigma \|f\|_\infty \left(\frac{2d}{N}\log(6) + \frac{2Rd}{N}\log(6) + \frac{2}{N}\log(2/\delta) + 6\epsilon \log \epsilon^{-1}\right)^{1/2}$$

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

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

We will first show that $f(\mathbf{w}\cdot\mathbf{x}_i)\mathbf{x}_i$ is sub-Gaussian. We first note for any $\mathbf{v} \in \mathbb{S}^{d-1}$, the random variable, $\mathbf{x}_i^\top \mathbf{v}$ is sub-Gaussian by definition. We then have,

$$\Big(\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}} |f(\mathbf{w}\cdot\mathbf{x}_i)\mathbf{x}_i^\top \mathbf{v}|^p\Big)^{1/p} \overset{(i)}{\leq} \Big(\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}} |f(\mathbf{w}\cdot\mathbf{x}_i)|^{2p} \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}} |\mathbf{x}^\top\mathbf{v}|^{2p}\Big)^{1/2p} \overset{(ii)}{\leq} \Big(\|f\|_\infty C_\Sigma \sqrt{2}\Big)\sqrt{p}$$

In the above, (i) follows from Hölder's Inequality, (ii) follows from noting from letting $q = 2p$ and noting from Definition 7 that $\|\mathbf{x}_i^\top \mathbf{v}\|_{L_q}$ is upper bounded by $C_\Sigma \sqrt{q}$. We thus have $f(\mathbf{w}\cdot\mathbf{x}_i)\mathbf{x}_i^\top \mathbf{v}$ is sub-Gaussian for any $\mathbf{w} \in \mathbb{R}^d$ and $\|f(\mathbf{w}\cdot\mathbf{x}_i)\mathbf{x}_i^\top \mathbf{v}\|_{\psi_2} \leq \sqrt{2} C_\Sigma \|f\|_\infty$. Recall $C_\sigma \triangleq \|\xi_i\|_{\psi_2}$, then from Lemma 8, the random variable $\xi_i f(\mathbf{w}\cdot\mathbf{x}_i)\mathbf{x}_i^\top \mathbf{v}$ is sub-exponential s.t. $\|\xi_i f(\mathbf{w}\cdot\mathbf{x}_i)\mathbf{x}_i^\top \mathbf{v}\|_{\psi_1} \leq \sqrt{2} C_\Sigma C_\sigma \|f\|_\infty$. Let $\widetilde{\mathbf{w}} \in \mathcal{N}_1$ such that $\widetilde{\mathbf{w}} = \arg\min_{\mathbf{u}\in\mathcal{N}_1} \|\mathbf{w} - \mathbf{u}\|_2$, where $\mathcal{N}_1$ is a ε-cover of $\mathcal{B}(\mathbf{0}, R)$.
Step 1: Probabilistic Decomposition. We use the decomposition given in Lemma A.4 of [ZYWG19].

$$\begin{aligned}&\mathbf{Pr}\left\{\max_{\mathrm{S}\in\mathcal{S}} \sup_{\mathbf{w}\in\mathcal{B}(\mathbf{0},R)} \frac{1}{(1-\epsilon)N} \cdot \Big\|\sum_{i\in\mathrm{S}} \xi_i f(\mathbf{w}\cdot\mathbf{x}_i)\mathbf{x}_i\Big\| \geq t\right\} \\ &\leq \mathbf{Pr}\left\{\max_{\mathrm{S}\in\mathcal{S}} \max_{\mathbf{u}\in\mathcal{N}_1} \frac{1}{(1-\epsilon)N} \cdot \Big\|\sum_{i\in\mathrm{S}} \xi_i f(\mathbf{x}_i\cdot\mathbf{u})\mathbf{x}_i\Big\|_2 \geq \frac{t}{2}\right\} \\ &+ \mathbf{Pr}\left\{\max_{\mathrm{S}\in\mathcal{S}} \sup_{\mathbf{w}\in\mathcal{B}(\mathbf{0},R)} \frac{1}{(1-\epsilon)N} \cdot \Big\|\sum_{i\in\mathrm{S}} \xi_i f(\mathbf{w}\cdot\mathbf{x}_i)\mathbf{x}_i - \sum_{i\in\mathrm{S}} \xi_i f(\mathbf{x}_i^\top \widetilde{\mathbf{w}})\mathbf{x}_i\Big\| \geq \frac{t}{2}\right\}\end{aligned} \tag{20}$$

We will now use a ε-covering argument to bound the first term of Equation (20). Let $\mathcal{N}_2$ be a ε-net of $\mathbb{S}^{d-1}$ such that for any $\mathbf{v} \in \mathbb{S}^{d-1}$, there exists $\mathbf{u} \in \mathcal{N}_2$ such that $\|\mathbf{u} - \mathbf{v}\|_2 \leq \varepsilon$. Let $\mathbf{u}^* = \arg\max_{\mathbf{u} \in \mathcal{N}_2} |(\boldsymbol{\xi} \circ f(X^\top \widetilde{\mathbf{w}}))^\top X\mathbf{u}|$ and $\mathbf{v}^* = \arg\max_{\mathbf{v} \in \mathbb{S}^{d-1}} |(\boldsymbol{\xi} \circ f(X^\top \widetilde{\mathbf{w}}))^\top X\mathbf{v}|$. We then have from the triangle inequality,

$$|(\boldsymbol{\xi} \circ f(X^\top \widetilde{\mathbf{w}}))^\top X\mathbf{v}^* - (\boldsymbol{\xi} \circ f(X^\top \widetilde{\mathbf{w}}))^\top X\mathbf{u}^*| \leq \|(\boldsymbol{\xi} \circ f(X^\top \widetilde{\mathbf{w}}))^\top X\|_2 \|\mathbf{u}^* - \mathbf{v}^*\|_2 \leq \varepsilon \cdot \|(\boldsymbol{\xi} \circ f(X^\top \widetilde{\mathbf{w}}))^\top X\|_2$$

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

$$\begin{aligned}|(\boldsymbol{\xi} \circ f(X^\top \widetilde{\mathbf{w}}))^\top X\mathbf{u}^*| &\geq |(\boldsymbol{\xi} \circ f(X^\top \widetilde{\mathbf{w}}))^\top X\mathbf{v}^*| - |(\boldsymbol{\xi} \circ f(X^\top \widetilde{\mathbf{w}}))^\top X\mathbf{u}^* - (\boldsymbol{\xi} \circ f(X^\top \widetilde{\mathbf{w}}))^\top X\mathbf{v}^*| \\ &\geq (1-\varepsilon)|(\boldsymbol{\xi} \circ f(X^\top \widetilde{\mathbf{w}}))^\top X\mathbf{v}^*|\end{aligned}$$

From rearranging, we have

$$|(\boldsymbol{\xi} \circ f(X^\top \widetilde{\mathbf{w}}))^\top X\mathbf{v}^*| \leq \frac{1}{1-\varepsilon} \cdot |(\boldsymbol{\xi} \circ f(X^\top \widetilde{\mathbf{w}}))^\top X\mathbf{u}^*|$$

With this result we are ready to make the probabilistic bounds. Suppose $\mathcal{S}$ represents all subsets of $[N]$ of size empty to $(1-\epsilon)N$. Suppose $\mathcal{N}_2$ is a ε_2 net of $\mathbb{S}^{d-1}$ and $\mathcal{N}_1$ is a ε_1 net of $\mathcal{B}(\mathbf{w}^*, R)$, we can then note that $\|\mathbf{w}^*\| \leq R$. Then,

$$\begin{aligned}&\mathbf{Pr}\left\{\frac{1}{(1-\epsilon)N} \cdot \max_{\mathrm{S} \in \mathcal{S}} \max_{\mathbf{w} \in \mathcal{N}_1} \|(\boldsymbol{\xi}_{\mathrm{S}} \circ f(X_{\mathrm{S}}^\top \mathbf{w}))^\top X_{\mathrm{S}}\| \geq \frac{t}{2}\right\} \\ &\quad \leq \mathbf{Pr}\left\{\frac{1}{1-\varepsilon_2} \cdot \frac{1}{(1-\epsilon)N} \cdot \max_{\mathrm{S} \in \mathcal{S}} \max_{\mathbf{v} \in \mathcal{N}_2} \max_{\mathbf{w} \in \mathcal{N}_1} |(\boldsymbol{\xi}_{\mathrm{S}} \circ f(X_{\mathrm{S}}^\top \mathbf{w}))^\top X_{\mathrm{S}}\mathbf{v}| \geq \frac{t}{2}\right\} \\ &\quad \leq \sum_{j \in [|\mathcal{S}|]} \sum_{i \in [|\mathcal{N}_1|]} \sum_{k \in [|\mathcal{N}_2|]} \mathbf{Pr}\left\{\frac{1}{1-\varepsilon_2} \cdot \frac{1}{(1-\epsilon)N} \cdot |(\boldsymbol{\xi}_{\mathrm{S}_j} \circ f(X_{\mathrm{S}_j}^\top \mathbf{w}_i))^\top X_{\mathrm{S}}\mathbf{v}_k| \geq \frac{t}{2}\right\} \\ &\quad \overset{(iii)}{\leq} 2\left(\frac{3}{\varepsilon_1}\right)^{Rd}\left(\frac{3}{\varepsilon_2}\right)^{d} \cdot \left(\frac{e}{\epsilon}\right)^{N\epsilon} \exp\left(-c \cdot \left(\frac{t^2(1-\epsilon)^2N^2(1-\varepsilon_2)^2}{8C_\Sigma^2 C_\sigma^2 \|f\|_\infty^2 |\mathrm{S}|} \wedge \frac{t(1-\epsilon)N(1-\varepsilon_2)}{2\sqrt{2}C_\Sigma C_\sigma \|f\|_\infty}\right)\right) \leq \frac{\delta}{2}\end{aligned}$$

In the above, (iii) follows from Bernstein's Inequality (see Lemma 9). We can now note that $\log\binom{N}{(1-\epsilon)N} = \log\binom{N}{\epsilon N}$. Then to satisfy the above probabilistic condition, it must hold that

$$t \geq \frac{1}{c} 2\sqrt{2} C_\Sigma C_\sigma \|f\|_\infty \left(\frac{2d}{N}\log\left(\frac{3}{\varepsilon_2}\right) + \frac{2Rd}{N}\log\left(\frac{3}{\varepsilon_1}\right) + \frac{2}{N}\log(2/\delta) + 6\epsilon\log\epsilon^{-1}\right)^{1/2}$$

We now bound the second term of Equation (20). For any $\mathbf{w}$, recall $\widetilde{\mathbf{w}} = \arg\min_{\mathbf{u} \in \mathcal{N}_1} \|\mathbf{w} - \mathbf{u}\|_2$ and thus for any $\mathbf{w}$, we have $\|\mathbf{w} - \widetilde{\mathbf{w}}\|_2 \leq \varepsilon_1$.

Step 2a: Lipschitz Functions. We have

$$\begin{aligned}&\sup_{\mathbf{w} \in \mathcal{B}(\mathbf{0}, R)} \frac{1}{(1-\epsilon)N} \cdot \Big\|\sum_{i \in \mathrm{S}} \xi_i f(\mathbf{w} \cdot \mathbf{x}_i)\mathbf{x}_i - \sum_{i \in \mathrm{S}} \xi_i f(\widetilde{\mathbf{w}} \cdot \mathbf{x}_i)\mathbf{x}_i\Big\|_2 \\ &\qquad \leq \sup_{\mathbf{w} \in \mathcal{B}(\mathbf{0}, R)} \max_{i \in [N]} \frac{1}{(1-\epsilon)N} \cdot \|\xi_i \mathbf{x}_i\|_2 \|f(\mathbf{w} \cdot \mathbf{x}_i) - f(\widetilde{\mathbf{w}} \cdot \mathbf{x}_i)\|_2 \\ &\qquad \leq \sup_{\mathbf{w} \in \mathcal{B}(\mathbf{0}, R)} \max_{i \in [N]} \frac{\|f\|_{\mathrm{lip}}}{(1-\epsilon)N} \cdot \|\xi_i \mathbf{x}_i\|_2 \|(\mathbf{w} - \widetilde{\mathbf{w}}) \cdot \mathbf{x}_i\|_2 \\ &\qquad \leq \frac{\varepsilon_1 \|f\|_{\mathrm{lip}}}{(1-\epsilon)N} \cdot \max_{i \in [N]} \|\xi_i \mathbf{x}_i\|_2 \max_{i \in [N]} \|\mathbf{x}_i\|_2 \\ &\qquad \overset{(iv)}{\leq} \frac{\varepsilon_1 \|f\|_{\mathrm{lip}}}{2(1-\epsilon)N} \cdot \left(\max_{i \in [N]} \|\xi_i \mathbf{x}_i\|_2^2 + \max_{i \in [N]} \|\mathbf{x}_i\|_2^2\right)\end{aligned}$$

In the above, (iv) follows from Young's Inequality. Note that $\mathbf{x}_i^\top \mathbf{v}$ for any $\mathbf{v} \in \mathbb{S}^{d-1}$ and $i \in [N]$ is sub-Gaussian, we thus have,

$$\mathbf{Pr}\left\{\frac{\varepsilon_1 \|f\|_{\mathrm{lip}}}{(1-\epsilon)N} \cdot \max_{j \in [N]} \|\mathbf{x}_j\|_2^2 \geq \frac{t}{2}\right\} \leq \sum_{i \in \mathcal{N}_2} \sum_{j \in [N]} \mathbf{Pr}\left\{\frac{\varepsilon_1 \|f\|_{\mathrm{lip}}}{(1-\epsilon)N} \cdot \frac{1}{1-\varepsilon_2} \cdot |\mathbf{x}_j \cdot \mathbf{v}_i|^2 \geq \frac{t}{2}\right\}$$

$$\leq N \cdot \left(\frac{3}{\varepsilon_2}\right)^d \cdot \exp\left[-c \cdot \left(\frac{t(1-\varepsilon_2)(1-\epsilon)N}{2C_\Sigma\varepsilon_1}\right)^2 \wedge \left(\frac{t(1-\varepsilon_2)(1-\epsilon)N}{2C_\Sigma\varepsilon_1}\right)\right] \leq \frac{\delta}{4}$$

where in the above, the final inequality follows when

$$t \geq \left(\frac{C_\Sigma\varepsilon_1}{c(1-\varepsilon_2)} \cdot \left(\frac{2\log N}{N} + \frac{2d}{N}\log\left(\frac{3}{\varepsilon_2}\right) + \frac{2}{N}\log(4/\delta)\right)\right)^{1/2}$$

To bound $\max_{i\in S}\|\xi_i \mathbf{x}_i\|_2$, we note that for any $\mathbf{v} \in \mathbb{R}^d$ that $\xi_i \mathbf{x}_i^\top \mathbf{v}$ is sub-exponential with norm $\|\xi_i \mathbf{x}_i^\top \mathbf{v}\|_{\psi_1} \leq C_\sigma C_\Sigma$. Similarly,

$$\mathbf{Pr}\left\{\frac{\varepsilon_1\|f\|_{\text{lip}}}{2(1-\epsilon)N} \cdot \max_{i\in S}\|\xi_i\mathbf{x}_i\|_2^2 \geq \frac{t}{4}\right\} \leq \frac{\delta}{4}$$

The final inequality holds when,

$$t \geq \left(\frac{C_\Sigma C_\sigma\varepsilon_1}{c(1-\varepsilon_2)} \cdot \left(\frac{2\log N}{N} + \frac{2d}{N}\log\left(\frac{3}{\varepsilon_2}\right) + \frac{2}{N}\log(4/\delta)\right)\right)^{1/2}$$

Step 2b: Indicator-Type Functions. We have,

$$\begin{aligned}
&\sup_{\mathbf{w}\in\mathcal{B}(\mathbf{0},R)} \frac{1}{(1-\epsilon)N} \cdot \Big\|\sum_{i\in S}\xi_i f(\mathbf{w}\cdot\mathbf{x}_i)\mathbf{x}_i - \sum_{i\in S}\xi_i f(\widetilde{\mathbf{w}}\cdot\mathbf{x}_i)\mathbf{x}_i\Big\|_2 \\
&\quad\leq \sup_{\mathbf{w}\in\mathcal{B}(\mathbf{0},R)} \max_{i\in[N]} \frac{1}{(1-\epsilon)N} \cdot \|\xi_i\mathbf{x}_i\|_2\|f(\mathbf{w}\cdot\mathbf{x}_i) - f(\widetilde{\mathbf{w}}\cdot\mathbf{x}_i)\|_2 \\
&\quad\leq \max_{i\in[N]} \frac{1}{(1-\epsilon)N} \cdot \|\xi_i\mathbf{x}_i\|_2 \sup_{\mathbf{w}\in\mathcal{B}(\mathbf{w}^*,R)} (\mathbf{1}\{\mathbf{w}\cdot\mathbf{x}_i \geq 0, \widetilde{\mathbf{w}}\cdot\mathbf{x}_i \leq 0\} + \mathbf{1}\{\mathbf{w}\cdot\mathbf{x}_i \leq 0, \widetilde{\mathbf{w}}\cdot\mathbf{x}_i \geq 0\}) \\
&\quad\leq \max_{i\in[N]} \frac{2\operatorname{Tr}(\Sigma)}{(1-\epsilon)N} \cdot \|\xi_i\mathbf{x}_i\|_2 \cdot \Theta(\mathbf{w},\widetilde{\mathbf{w}}) \leq \max_{i\in[N]} \frac{4\operatorname{Tr}(\Sigma)}{(1-\epsilon)N} \cdot \|\xi_i\mathbf{x}_i\|_2 \cdot \frac{\varepsilon_1}{\|\mathbf{w}^*\| - R}
\end{aligned}$$

This upper bound holds for both when f is indicator, and when f is the derivative of leaky-relu. We then have,

$$\begin{aligned}
\mathbf{Pr}\left\{\frac{4\operatorname{Tr}(\Sigma)\varepsilon_1 c_1}{(1-\epsilon)N} \cdot \max_{i\in[N]}\|\xi_i\mathbf{x}_i\|_2 \geq \frac{t}{2}\right\} &\leq N \cdot \mathbf{Pr}\left\{\frac{4\operatorname{Tr}(\Sigma)\varepsilon_1 c_1}{(1-\epsilon)N} \cdot \frac{1}{1-\varepsilon_2} \cdot \max_{\mathbf{v}\in\mathcal{N}_2}|\xi_i\mathbf{x}_i\cdot\mathbf{v}| \geq \frac{t}{2}\right\} \\
&\leq N\left(\frac{3}{\varepsilon_1}\right)^d \exp\left[-c \cdot \left(\left(\frac{t(1-\varepsilon_2)(1-\epsilon)N}{4\operatorname{Tr}(\Sigma)\varepsilon_1 c_1}\right)^2 \wedge \left(\frac{t(1-\varepsilon_2)(1-\epsilon)N}{4\operatorname{Tr}(\Sigma)\varepsilon_1 c_1}\right)\right)\right] \leq \frac{\delta}{4}
\end{aligned}$$

The final inequality holds when,

$$t \geq \left(\frac{4\operatorname{Tr}(\Sigma)\varepsilon_1 c_1}{(1-\varepsilon_2)N}\right)()^{1/2}$$

Step 3: Combining Estimates. We now choose $\varepsilon_1 = \frac{1}{2}$ and $\varepsilon_2 = \frac{1}{2}$.Combining our estimates, we obtain,

$$\mathbf{Pr}\left\{\frac{1}{(1-\epsilon)N} \cdot \max_{S\in\mathcal{S}}\|(\boldsymbol{\xi}_S \circ f(X_S^\top\mathbf{w}))^\top X_S\| \geq t\right\} \leq \delta$$

when

$$t \geq \frac{1}{c}2\sqrt{2}C_\Sigma C_\sigma\|f\|_\infty\left(\frac{2d}{N}\log(6) + \frac{2Rd}{N}\log(6) + \frac{2}{N}\log(2/\delta) + 6\epsilon\log\epsilon^{-1}\right)^{1/2}$$

when $N \geq \frac{\log(6)(Rd+d)+2\log(4/\delta)}{\epsilon\log\epsilon^{-1}} = \Omega\left(\frac{Rd+\log(1/\delta)}{\epsilon}\right)$. Our proof is complete. ∎

Lemma 27. *Fix $\mathbf{w}^* \in \mathbb{R}^{d-1}$ and suppose $\mathbf{w} \in \mathcal{B}(\mathbf{w}^*, R)$ for a constant $R < \|\mathbf{w}^*\|$. Sample $\mathbf{x}_1, \ldots, \mathbf{x}_N$ i.i.d from a sub-Gaussian distribution with second-moment matrix Σ and sub-Gaussian norm C_Σ. Suppose $S \subset [N]$ s.t. $|S| \leq (1-\epsilon)N$. Then with probability at least $1-\delta$,*

$$\Big\|\sum_{i\in S}\mathbf{x}_i\mathbf{x}_i^\top \cdot \mathbf{1}\{\mathbf{w}^*\cdot\mathbf{x}_i \geq 0\} \cdot \mathbf{1}\{\mathbf{w}^{(t)}\cdot\mathbf{x}_i \geq 0\} - \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}}[\mathbf{x}\mathbf{x}^\top \cdot \mathbf{1}\{\mathbf{w}^*\cdot\mathbf{x}_i \geq 0\} \cdot \mathbf{1}\{\mathbf{w}^{(t)}\cdot\mathbf{x}_i \geq 0\}]\Big\|_2 \leq \Xi$$

Proof. Let $\mathcal{N}_1$ be an ε_1-cover of $\mathcal{B}(\mathbf{w}^*, R)$ and $\mathcal{N}_2$ be an ε_2-cover of $\mathbb{S}^{d-1}$. We will use the decomposition given in Theorem 1 of [MBM16]. Let $\widetilde{\mathbf{w}} = \arg\min_{\mathbf{v}\in\mathcal{N}_1}\|\mathbf{w}-\mathbf{v}\|_2$ throughout the relations.

$$\begin{aligned}
&\mathbf{Pr}\left\{\max_{\mathrm{S}\in\mathcal{S}}\sup_{\mathbf{w}\in\mathcal{B}(\mathbf{w}^*;R)}\frac{1}{(1-\epsilon)N}\cdot\|\Sigma_{\mathrm{S}}(\mathbf{w},\mathbf{w}^*)-\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}}[\Sigma_{\mathrm{S}}(\mathbf{w},\mathbf{w}^*)]\|_2\geq t\right\}\\
&\qquad\leq\mathbf{Pr}\left\{\max_{\mathrm{S}\in\mathcal{S}}\sup_{\mathbf{w}\in\mathcal{B}(\mathbf{w}^*;R)}\frac{1}{(1-\epsilon)N}\cdot\|\Sigma_{\mathrm{S}}(\mathbf{w},\mathbf{w}^*)-\Sigma_{\mathrm{S}}(\widetilde{\mathbf{w}},\mathbf{w}^*)\|_2\geq\frac{t}{3}\right\}\\
&\qquad+\mathbf{Pr}\left\{\max_{\mathrm{S}\in\mathcal{S}}\max_{\widetilde{\mathbf{w}}\in\mathcal{N}_1}\frac{1}{(1-\epsilon)N}\cdot\|\Sigma_{\mathrm{S}}(\widetilde{\mathbf{w}},\mathbf{w}^*)-\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}}[\Sigma_{\mathrm{S}}(\widetilde{\mathbf{w}},\mathbf{w}^*)]\|_2\geq\frac{t}{3}\right\}\\
&\qquad+\mathbf{Pr}\left\{\sup_{\mathbf{w}\in\mathcal{B}(\mathbf{w}^*;R)}\frac{1}{(1-\epsilon)N}\cdot\|\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}}[\Sigma_{\mathrm{S}}(\widetilde{\mathbf{w}},\mathbf{w}^*)]-\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}}[\Sigma_{\mathrm{S}}(\mathbf{w},\mathbf{w}^*)]\|_2\geq\frac{t}{3}\right\}
\end{aligned}$$

We bound all terms separately. For the first term,

$$\begin{aligned}
\max_{\mathrm{S}\in\mathcal{S}}\sup_{\mathbf{w}\in\mathcal{B}(\mathbf{w}^*;R)}\|\Sigma_{\mathrm{S}}(\mathbf{w},\mathbf{w}^*)-\Sigma_{\mathrm{S}}(\widetilde{\mathbf{w}},\mathbf{w}^*)\|_2&\leq\max_{\mathrm{S}\in\mathcal{S}}\sup_{\mathbf{w}\in\mathcal{B}(\mathbf{w}^*;R)}\|\Sigma_{\mathrm{S}}(\mathbf{w},-\widetilde{\mathbf{w}})\|_2+\max_{\mathrm{S}\in\mathcal{S}}\sup_{\mathbf{w}\in\mathcal{B}(\mathbf{w}^*;R)}\|\Sigma_{\mathrm{S}}(-\mathbf{w},\widetilde{\mathbf{w}})\|_2\\
&\leq 2\max_{i\in[N]}\|\mathbf{x}_i\mathbf{x}_i^\top\|_2\arcsin\left(\frac{\|\widetilde{\mathbf{w}}-\mathbf{w}\|_2}{\|\mathbf{w}\|}\right)\leq\varepsilon_1\cdot\frac{\pi}{\|\mathbf{w}^*\|-R}\cdot\max_{i\in[N]}\|\mathbf{x}_i\mathbf{x}_i^\top\|
\end{aligned}$$

We then have for a constant c_1,

$$\begin{aligned}
&\mathbf{Pr}\left\{\frac{\varepsilon_1c_1}{(1-\epsilon)N}\cdot\max_{i\in[N]}\|\mathbf{x}_i\mathbf{x}_i^\top\|\geq\frac{t}{3}\right\}\\
&\leq N\cdot\mathbf{Pr}\left\{\|\mathbf{x}\|_2^2\geq\frac{t}{3\varepsilon_1c_1}\cdot(1-\epsilon)N\right\}\\
&\leq N\cdot\mathbf{Pr}\left\{\max_{\mathbf{v}\in\mathcal{N}_2}|\mathbf{x}\cdot\mathbf{v}|^2\geq\frac{t}{3\varepsilon_1}\cdot c_1(1-\epsilon)N(1-\varepsilon_2)\right\}\\
&\leq N\cdot\sum_{i\in[|\mathcal{N}_2|]}\mathbf{Pr}\left\{|\mathbf{x}\cdot\mathbf{v}_i|\geq\left(\frac{t}{3\varepsilon_1c_1}\cdot(1-\epsilon)N(1-\varepsilon_2)\right)^{1/2}\right\}\\
&\overset{(i)}{\leq}2N\left(\frac{3}{\varepsilon_2}\right)^d\exp\left(-\frac{1}{2C_\sigma^2}\cdot\left(\frac{t}{3\varepsilon_1c_1}\cdot(1-\epsilon)N(1-\varepsilon_2)\right)\right)\leq\frac{\delta}{2}
\end{aligned}$$

In the above, (i) follows from the tails on a sub-Gaussian. The above probabilistic condition is satisfied when,

$$t\geq 6C_\sigma^2\varepsilon_1c_1\left(\frac{\log N}{N}+\frac{d}{N}\log\left(\frac{3}{\varepsilon_2}\right)+\frac{\log(4/\delta)}{N}\right)$$

For the second term, let $\widetilde{\mathbf{x}}=\mathbf{x}\cdot\mathbf{1}\{\widetilde{\mathbf{w}}\cdot\mathbf{x}\geq 0\}\cdot\mathbf{1}\{\mathbf{w}^*\cdot\mathbf{x}\geq 0\}$. We then have,

$$\begin{aligned}
&\mathbf{Pr}\left\{\max_{\mathrm{S}\in\mathcal{S}}\max_{\widetilde{\mathbf{w}}\in\mathcal{N}_1}\frac{1}{(1-\epsilon)N}\cdot\|\Sigma_{\mathrm{S}}(\widetilde{\mathbf{w}},\mathbf{w}^*)-\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}}[\Sigma_{\mathrm{S}}(\widetilde{\mathbf{w}},\mathbf{w}^*)]\|_2\geq\frac{t}{3}\right\}\\
&\leq\mathbf{Pr}\left\{\max_{\mathrm{S}\in\mathcal{S}}\max_{\widetilde{\mathbf{w}}\in\mathcal{N}_1}\max_{\mathbf{v}\in\mathcal{N}_2}\frac{1}{1-2\varepsilon_2}\cdot\frac{1}{(1-\epsilon)N}\cdot\left|\|\widetilde{X}_{\mathrm{S}}^\top\mathbf{v}\|_2^2-\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}}\|\widetilde{X}_{\mathrm{S}}^\top\mathbf{v}\|_2^2\right|\geq\frac{t}{3}\right\}\\
&\overset{(i)}{\leq}2\left(\frac{e}{\epsilon}\right)^{N\epsilon}\left(\frac{3}{\varepsilon_1}\right)^{Rd}\left(\frac{3}{\varepsilon_2}\right)^d\exp\left[-c\min\left(\frac{t^2(1-\epsilon)^2N^2(1-2\varepsilon_2)^2}{9\cdot 512C_\Sigma^2|\mathrm{S}|},\frac{t(1-\epsilon)N(1-\varepsilon_2)}{48\sqrt{2}C_\Sigma}\right)\right]\leq\frac{\delta}{2}
\end{aligned}$$

In (i) we note from Lemma 1.12 in [RH23], that the random variable $|\widetilde{\mathbf{x}}\cdot\mathbf{v}|^2-\mathbf{E}_{\mathbf{x}\sim\mathcal{D}}|\widetilde{\mathbf{x}}\cdot\mathbf{v}|^2$ is sub-exponential s.t. $\||\widetilde{\mathbf{x}}\cdot\mathbf{v}|^2-\mathbf{E}_{\mathbf{x}\sim\mathcal{D}}|\widetilde{\mathbf{x}}\cdot\mathbf{v}|^2\|_{\psi_1}\leq 16\sqrt{2}C_\Sigma$, we can then apply Bernstein's Inequality. The probabilistic condition above is then satisfied when,

$$t\geq\left(\frac{1}{c}\cdot\frac{9\cdot 512C_\Sigma^2|\mathrm{S}|}{(1-\epsilon)^2N^2(1-2\varepsilon_c^2)}\left(Rd\log\left(\frac{3}{\varepsilon_1}\right)+d\log\left(\frac{3}{\varepsilon_2}\right)+3N\epsilon\log\epsilon^{-1}+\log(4/\delta)\right)\right)^{1/2}$$

from which we obtain the simplified bound,

$$t \geq \left(\frac{18432 C_\Sigma^2}{c(1-2\varepsilon_c^2)} \left(\frac{Rd}{N} \log\left(\frac{3}{\varepsilon_1}\right) + \frac{d}{N}\log\left(\frac{3}{\varepsilon_2}\right) + 3\epsilon \log \epsilon^{-1} + \frac{1}{N}\log(4/\delta) \right)\right)^{1/2}$$

We now consider the third term.

$$\begin{aligned}
&\mathbf{Pr}\left\{ \sup_{\mathbf{w}\in\mathcal{B}(\mathbf{w}^*;R)} \frac{1}{(1-\epsilon)N} \cdot \| \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}}[\Sigma_{\mathrm{S}}(\widetilde{\mathbf{w}}, \mathbf{w}^*)] - \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}}[\Sigma_{\mathrm{S}}(\mathbf{w}, \mathbf{w}^*)]\|_2 \geq \frac{t}{3} \right\} \\
&= \mathbf{Pr}\left\{ \sup_{\mathbf{w}\in\mathcal{B}(\mathbf{w}^*;R)} \frac{1}{(1-\epsilon)N} \cdot \| \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}}[\mathbf{x}\mathbf{x}^\top \cdot (\mathbf{1}\{\widetilde{\mathbf{w}}\cdot\mathbf{x}\geq 0\} - \mathbf{1}\{\mathbf{w}\cdot\mathbf{x}\geq 0\}) \cdot \mathbf{1}\{\mathbf{w}^*\cdot\mathbf{x}\geq 0\}]\|_2 \geq \frac{t}{3} \right\} \\
&\overset{(ii)}{\leq} \mathbf{Pr}\left\{ \sup_{\mathbf{w}\in\mathcal{B}(\mathbf{w}^*;R)} \frac{1}{(1-\epsilon)N} \cdot \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}}\big[\|\mathbf{x}\mathbf{x}^\top\|_2 |\mathbf{1}\{\widetilde{\mathbf{w}}\cdot\mathbf{x}\geq 0\} - \mathbf{1}\{\mathbf{w}\cdot\mathbf{x}\geq 0\}|\big] \geq \frac{t}{3} \right\} \\
&\overset{(iii)}{\leq} \mathbf{Pr}\left\{ \sup_{\mathbf{w}\in\mathcal{B}(\mathbf{w}^*;R)} \frac{1}{2(1-\epsilon)N} \cdot \Big(\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}}\|\mathbf{x}\mathbf{x}^\top\|_2^2 \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}}|\mathbf{1}\{\widetilde{\mathbf{w}}\cdot\mathbf{x}\geq 0\} - \mathbf{1}\{\mathbf{w}\cdot\mathbf{x}\geq 0\}| \Big)^{1/2} \geq \frac{t}{3} \right\} = 0
\end{aligned}$$

In the above, (ii) follows from first applying Cauchy-Schwarz inequality and then applying Jensen's inequality, and (iii) follows from Hölder's Inequality. Then from $L_4 \to L_2$ hypercontractivity of $\mathcal{D}$, we have that $\mathbf{E}_{\mathbf{x}\sim\mathcal{D}}\|\mathbf{x}\|^4 \leq L\, \mathbf{E}_{\mathbf{x}\sim\mathcal{D}}\|\mathbf{x}\|^2 = L\,\mathrm{Tr}(\Sigma)$.

$$\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}}|\mathbf{1}\{\widetilde{\mathbf{w}}\cdot\mathbf{x}\geq 0\} - \mathbf{1}\{\mathbf{w}\cdot\mathbf{x}\geq 0\}| \leq \frac{\Theta(\mathbf{w}, \widetilde{\mathbf{w}})}{\pi} \leq \frac{2}{\pi}\arcsin\left(\frac{\|\mathbf{w}-\widetilde{\mathbf{w}}\|}{\|\mathbf{w}^*\| - R}\right) \leq \varepsilon_1 c_1$$

Then we obtain zero probability as indicated in the statement when

$$t \geq \frac{3}{N}\sqrt{\mathrm{Tr}(\Sigma)\varepsilon_1 c_1}$$

Combining our results, we choose $\varepsilon_1 = \frac{1}{2}$ and $\varepsilon_2 = \frac{1}{2}$ for sufficiently large $N = \Omega\Big(\frac{Rd+d+\log(4/\delta)}{\epsilon}\Big)$ ■

Lemma 28. *Suppose $\mathbf{E}_{\mathbf{x}\sim\mathcal{D}}[\mathbf{x}] = \mathbf{0}$ and $\mathbf{E}_{\mathbf{x}\sim\mathcal{D}}\big[\mathbf{x}\mathbf{x}^\top\big] = I$ for $\mathbf{x}\sim\mathcal{D}$. Fix $\mathbf{w}_1, \mathbf{w}_2 \in \mathbb{R}^d$ and define $\Theta = \arccos\Big(\frac{\mathbf{w}_1\cdot\mathbf{w}_2}{\|\mathbf{w}_1\|\|\mathbf{w}_2\|}\Big)$. For any rotationally invariant distribution, we have*

$$\mathop{\mathbf{E}}_{X\sim\mathcal{D}}\big[XX^\top \cdot \mathbf{1}\big\{X^\top \mathbf{w}_1 \geq \mathbf{0}\big\} \cdot \mathbf{1}\big\{X^\top\mathbf{w}_2 \geq \mathbf{0}\big\}\big] \succeq (\pi - \Theta - \sin\Theta)\cdot I$$

Proof. Consider the plane spanned by $\mathbf{w}_1$ and $\mathbf{w}_2$. Then, w.l.o.g let $\mathbf{w}_1 = (1,0)$ and rotate $\mathbf{w}_2$ such that it is in the first quadrant with angle Θ from $\mathbf{w}_1$ from noting that $\Theta \leq \frac{\pi}{2}$.

$$\begin{aligned}
\mathop{\mathbf{E}}_{X\sim\mathcal{D}}\big[XX^\top \cdot \mathbf{1}\big\{X^\top \mathbf{w}_1 \geq \mathbf{0}\big\} \cdot \mathbf{1}\big\{X^\top\mathbf{w}_2 \geq \mathbf{0}\big\}\big] &= \int_0^\infty \int_{-\pi/2+\Theta}^{\pi/2} \begin{pmatrix}\cos\Theta\\ \sin\Theta\end{pmatrix}(\cos\Theta, \sin\Theta) r d\Theta dr \\
&= \int_0^\infty \frac{r}{2}\begin{pmatrix}\pi - \Theta + \sin\Theta\cos\Theta & \sin^2\Theta \\ \sin^2\Theta & \pi - \Theta - \cos\Theta\sin\Theta\end{pmatrix} dr \\
&= (1/2)\cdot \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}}\big[r^2\big]\begin{pmatrix}\pi - \Theta + \sin\Theta\cos\Theta & \sin^2\Theta \\ \sin^2\Theta & \pi - \Theta - \cos\Theta\sin\Theta\end{pmatrix} \\
&\succeq (\pi - \Theta - \sin\Theta)\cdot I
\end{aligned}$$

Our proof is complete by noting that the planes perpendicular to that of the plane integrated over remain unchanged by the indicator functions and thus have unitary expectation. ■

Lemma 29. *Fix $S \in \mathbb{R}^{K\times N\epsilon}, T \in \mathbb{R}^{N\epsilon\times L}$, then sample a matrix $G \in \mathbb{R}^{N\epsilon\times N\epsilon}$ such that each column of G represents a ϵ-subset of a n-dimensional vector sampled from $\mathcal{N}(\mathbf{0}, \sigma^2 \cdot I)$, then with probability exceeding $1-\delta$,*

$$\|SGT\|_{\mathrm{F}}^2 \le \|S\|_{\mathrm{F}}^2\|T\|_{\mathrm{F}}^2 \cdot \sigma\sqrt{2\log\left(\frac{2N^2}{\delta}\right)}$$

Proof. The proof will be a calculation.

$$\|SGT\|_{\mathrm{F}}^2 = \sum_{i\in[K]}\sum_{j\in[L]}\sum_{k_1,k_2\in[N\epsilon]\times[N\epsilon]}\left(S_{i,k_1}G_{k_1,k_2}T_{k_2,j}\right)^2 \le \|S\|_{\mathrm{F}}^2\|T\|_{\mathrm{F}}^2 \max_{i,j\in[N\epsilon]\times[N\epsilon]}(G_{i,j})^2$$

It then suffices to bound the maximum value of a Gaussian squared over N^2 samples. From Lemma 33, we have

$$\mathop{\mathbf{Pr}}_{G\sim\mathcal{N}(0,1)}\left\{\max_{(i,j)\in[N]\times[N]} G_{i,j}^2 \ge t\right\} = \mathop{\mathbf{Pr}}_{G\sim\mathcal{N}(0,1)}\left\{\max_{(i,j)\in[N]\times[N]}|G_{i,j}| \ge \sqrt{t}\right\} \overset{(i)}{\le} 2N^2 e^{-\frac{t}{2\sigma^2}}$$

In the above, (i) follows from a union bound. We thus obtain from elementary inequalities, with probability at least $1-\delta$,

$$\max_{i,j} G_{i,j}^2 \le \sigma\left(2\log\left(2N^2/\delta\right)\right)$$

Our proof is complete. ■

D Mathematical Tools

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

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

$$\Big\|\sum_{i\in[n]} a_i b_i \mathbf{x}_i\Big\|^2 \le \|\mathbf{a}\|_\infty^2\|\mathbf{b}\|_2^2\|XX^\top\|_2$$

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

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

where the final inequality comes from noting

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

Our proof is complete. ■

Lemma 31 (Lemma 3.11 [B+15])**.** *Let f be β-smooth and α-strongly convex over $\mathbb{R}^n$, then for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$,*

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

Lemma 32 (Sum of Binomial Coefficients [CLRS22])**.** *Let $k, n \in \mathbb{N}$ such that $k \le n$, then*

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

Lemma 33 (Max Gaussian [RH23])**.** *Let $x_1, \ldots, x_n$ be sampled i.i.d from $\mathcal{N}(0, \sigma^2)$. Then,*

$$\mathbf{Pr}\{\|\mathbf{x}\|_\infty > t\} \le N \exp\left(-\frac{t^2}{\frac{16}{3}\sigma^2}\right)$$

Lemma 34 (Corollary 4.2.13 in [Ver20])**.** *The covering number of the ℓ_2-norm ball $\mathcal{B}(\mathbf{0}; 1)$ for $\varepsilon < 0$, satisfies,*

$$\mathcal{N}(\mathcal{B}^d_{\ell_2}(\mathbf{0}, 1), \varepsilon) \le \left(\frac{3}{\varepsilon}\right)^d$$