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

Arvind Rathnashyam
Department of Mathematics
Rensselaer Polytechnic Institute
Troy, NY 12180, USA
rathna@rpi.edu

Alex Gittens
Department of Computer Science
Rensselaer Polytechnic Institute
Troy, NY 12180, USA
gittea@rpi.edu

Abstract

We derive approximation bounds for learning single neuron models using thresholded gradient descent when both the labels and the covariates are possibly corrupted adversarially. We assume $(\mathbf{x}, y) \sim \mathcal{D}$ satisfy

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

where σ is a nonlinear activation function, the noise ξ is sampled from $\mathcal{N}(0, \nu^2)$, and the covariate vector $\mathbf{x}$ is sampled from a sub-Gaussian distribution. We study sigmoidal, leaky-ReLU, and ReLU activation functions and derive a $O(\nu\sqrt{\epsilon \log(1/\epsilon)})$ approximation bound in ℓ_2-norm, with sample complexity $O(d/\epsilon)$ and failure probability $e^{-\Omega(d)}$.

We also study the linear regression problem, where $\sigma(x) = x$. We derive a $O(\nu\epsilon \log(1/\epsilon))$ approximation bound, improving upon the previous $O(\nu)$ approximation bounds for the gradient-descent based iterative thresholding algorithms of Bhatia et al. (NeurIPS 2015) and Shen and Sanghavi (ICML 2019). Our algorithm has a $O(\mathsf{polylog}(N, d) \log(R/\epsilon))$ runtime complexity when $\|\mathbf{w}^*\|_2 \leq R$, improving upon the $O(\mathsf{polylog}(N, d)/\epsilon^2)$ runtime complexity of Awasthi et al. (NeurIPS 2022).

1 Introduction

The learning of the parameters of Generalized Linear Models (GLMs) [Awasthi et al., 2022, Diakonikolas et al., 2019, Fischler and Bolles, 1981, Li et al., 2021, Osama et al., 2020] and linear regression models [Bhatia et al., 2017, Mukhoty et al., 2019] under the Huber contamination model is well-studied. Efficient algorithms for robust statistical estimation have been studied extensively for problems such as high-dimensional mean estimation [Cheng et al., 2020, Prasad et al., 2019] and Robust Covariance Estimation [Cheng et al., 2019, Fan et al., 2018]; see Diakonikolas and Kane [2023] for an overview. Along these lines, interest has developed in the development of robust gradient-descent based approaches to machine learning problems [Diakonikolas et al., 2019, Prasad et al., 2018]. This work advances this line of research by providing gradient-descent based approaches for learning single neuron models under the strong contamination model.

Definition 1 (Strong ϵ-Contamination Model)**.** *Given a corruption parameter $0 \leq \epsilon < 0.5$, and data samples $(\mathbf{x}_i, y_i)$ for $i \in [N]$, 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. We define the corrupted dataset $\mathrm{X} = \mathrm{P} \cup \mathrm{Q}$ where P contains the remaining target points and Q contains the points modified by the adversary.*

Current approaches for robust learning across various machine learning tasks often use gradient descent over a robust objective (see e.g. Tilted Empirical Risk Minimization (TERM) [Li et al.,

Preprint. Under review.

2021]). These robust objectives tend to not be convex and therefore are difficult to obtain strong approximation bounds for general classes of models. Another popular approach is filtering, where at each iteration of training, points deemed to be as outliers are removed from training (see e.g. SEVER [Diakonikolas et al., 2019]). However, filtering algorithms such as SEVER require careful parameter selection to be useful in practice.

Iterative Thresholding is a popular framework for robust statistical estimation that was introduced in the 19th century by Legendre [Legendre, 1806]. Part of its appeal lies in its simplicity, as at each iteration of training we simply ignore points with error above a certain threshold. Despite the venerablity of this approach, it was not until recently that some iterative thresholding algorithms were proven to deliver robust parameter estimates in polynomial time. One of the first theoretical results in the literature proving the effectiveness of iterative thresholding considers its use in estimating the parameters of regression models under additive adversarial corruptions.

Theorem 2 (Theorem 5 in Bhatia et al. [2015])**.** *Let X be a sub-Gaussian data matrix, and $\mathbf{y} = X^\top \mathbf{w}^* + \mathbf{b}$ where $\mathbf{b}$ is the additive and possibly adversarial corruption. Then there exists a gradient-descent algorithm such that $\|\mathbf{w}^{(t)} - \mathbf{w}^*\|_2 \leq \varepsilon$ after $t = O\left(\log\left(\frac{1}{\sqrt{n}}\frac{\|\mathbf{b}\|_2}{\varepsilon}\right)\right)$ iterations.*

The algorithm referred to in Theorem 2 uses gradient-descent based iterative thresholding, exhibits a logarithmic dependence on $\|\mathbf{b}\|_2$, and is applicable to the *realizable* setting, i.e. there must be no stochastic noise in $\mathbf{y}$. More recently, Awasthi et al. [2022] studied the iterative trimmed maximum likelihood estimator. In their algorithm, at each step they find $\mathbf{w}^*$ which maximizes the likelihood of the samples in the trimmed set.

Theorem 3 (Theorem 4.2 in Awasthi et al. [2022])**.** *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 $\eta_i \sim \mathcal{N}(0, \nu^2)$ 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 $1-\delta$,*

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

Our first result recovers Theorem 3 as a special case and also allows for vector targets. Furthermore, we obtain the approximation in time $O(\mathsf{poly}(N,d)\log(R/\epsilon))$, improving upon the $O(\mathsf{polylog}(N,d)/\epsilon^2)$ runtime complexity of [Awasthi et al., 2022].

1.1 Contributions

Our main contribution consists of algorithms and corresponding approximation bounds for gradient-based iterative thresholding for several non-linear learning problems (Algorithm 2) and for multitarget linear regression (Algorithm 1). Our proof techniques extend [Bhatia et al., 2015, Shen and Sanghavi, 2019, Awasthi et al., 2022], as we suppose the adversary also corrupts the covariates. To our knowledge, we are the first to provide guarantees on the performance of iterative thresholding for learning non-linear models, outside of generalized linear models [Awasthi et al., 2022], under the strong contamination model.

Table 1: Summary of related work on iterative thresholding algorithms for linear regression under the strong ϵ-contamination model, and our contributions. We assume the target data is sampled from a centered sub-Gaussian distribution with second-moment matrix Σ, sub-Gaussian norm C_Σ, and dimension d. We assume the variance of the optimal estimator is ν.

Reference	Approximation	Runtime	Algorithm
Bhatia et al. [2015]	$O(\nu)$	$O\left(Nd^2\log\left(\frac{1}{\sqrt{n}}\frac{\|\mathbf{b}\|_2}{\epsilon}\right)\right)$	Full Solve
Shen and Sanghavi [2019]	$O(\nu)$	$O\left(Nd^2\log\left(\frac{\|\mathbf{w}^*\|_2}{\nu}\right)\right)$	Gradient Descent
Awasthi et al. [2022]	$O(\nu\epsilon\log(1/\epsilon))$	$O\left(\left(Nd^2+d^3\right)\left(\frac{1}{\nu\epsilon^2}\right)\right)$	Full Solve
Corollary 8	$O(\nu\epsilon\log(1/\epsilon))$	$O\left(Nd^2\log\left(\frac{\|\mathbf{w}^*\|}{\nu\epsilon}\right)\right)$	Gradient Descent

Table 1 summarizes the approximation and runtime guarantees for linear regression provided in this work and in the literature. Comparing to Bhatia et al. [2015], our runtime does not depend on

the norm of the adversarial corruption, which can be made arbitrarily large by the adversary. We extend upon Shen and Sanghavi [2019] by putting dependence on ϵ with a small logarithmic factor, improving the error bound significantly. We also offer a significant run-time improvement over the study in Awasthi et al. [2022], from $O(1/\varepsilon^2)$ to $O(\log(R/\varepsilon))$, where $\|\mathbf{w}^*\|_2 \leq R$.

Table 2: Our results and Distributional assumptions for learning different neuron activation functions. In the approximation bounds, $\kappa(\Sigma) = \lambda_{\max}(\Sigma)/\lambda_{\min}(\Sigma)$ where $\mathbf{E}_{\mathbf{x}\sim\mathcal{D}}[\mathbf{x}\mathbf{x}^\top] = \Sigma$, for all $\mathbf{x} \sim \mathcal{D}$ and $t \in [T]$ we have $\sigma'(\mathbf{w}^{(t)} \cdot \mathbf{x}) \geq \gamma$ and $\sigma'(\mathbf{w}^* \cdot \mathbf{x}) \geq \gamma$, for any $x, y \in \mathbb{R}$ it holds that $\sigma(x-y) \leq \|\sigma\|_{\text{lip}}|x-y|$, and ν is the variance of the optimal estimator.

Reference	Approximation	Neuron	Covariate Distribution
Theorem 7	$O(\nu\epsilon\log(1/\epsilon))$	Linear	Sub-Gaussian
Theorem 12	$O\left(\gamma^{-2}\|\sigma\|_{\text{lip}}^2\nu\kappa(\Sigma)\sqrt{\epsilon\log(1/\epsilon)}\right)$	Sigmoid	Bounded Sub-Gaussian
Theorem 13	$O\left(\gamma^{-2}\nu\kappa(\Sigma)\sqrt{\epsilon\log(1/\epsilon)}\right)$	Leaky-ReLU	Sub-Gaussian
Theorem 15	$O\left(\nu\kappa(\Sigma)\sqrt{\epsilon\log(1/\epsilon)}\right)$	ReLU	$L_4 - L_2$ Hypercontractive

Table 2 summarizes our results on learning single neuron models. The bounded assumption in Theorem 12 is necessary to give a lower bound on $\sigma'(\mathbf{w}\cdot\mathbf{x})$ for $\mathbf{x}\sim\mathcal{D}$ and $\mathbf{w}$ is equal to $\mathbf{w}^{(t)}$ for any $t \in [T]$ or $\mathbf{w}^*$. In Theorem 15, the $L_4 \to L_2$ hypercontractive assumption allows us to bound the minimum eigenvalue of the sample covariance matrix in the intersection of halfspacse defined by $\mathbf{x}\mathbf{x}^\top \cdot \mathbf{1}\{\mathbf{w}^*\cdot\mathbf{x} \geq 0\} \cdot \mathbf{1}\{\mathbf{w}^{(t)}\cdot\mathbf{x} \geq 0\}$ for any $t \in [T]$. In general, our nonlinear learning algorithms have approximation error on the order of $O(\sqrt{\epsilon\log(1/\epsilon)})$.

All our main results allow corruption to be present in both the covariates and the labels. We are able to show with only knowledge of the minimum and maximum eigenvalues of the sample covariance matrix, iterative thresholding is capable of directly handling corrupted covariates. In comparison, the algorithm of Awasthi et al. [2022] considers corruption only in the labels; they extend it to handle corruption in the covariates by preprocessing the covariates using the filtering algorithm of Dong et al. [2019].

2 Preliminaries

2.1 Mathematical Notation and Background

Notation. We use $[T]$ to denote the set $\{1, 2, \ldots, T\}$. We say $y = O(x)$ if there exists x_0 such that for all $x \geq x_0$ there exists C such that $y \leq Cx$. We say $y = \Omega(x)$ if there exists x_0 such that for all $x \geq x_0$ there exists C s.t. $y \geq Cx$. We use the notation $a \vee b \triangleq \max(a, b)$ and $a \wedge b \triangleq \min(a, b)$. We define $\mathbb{S}^{d-1}$ as the $d-1$-dimensional sphere $\{\mathbf{x} \in \mathbb{R}^d : \|\mathbf{x}\|_2 = 1\}$. We denote the Hadamard product between two vectors of the same size as $\mathbf{x} \circ \mathbf{y}$; that is, $(\mathbf{x}\circ\mathbf{y})_i = x_i y_i$.

Matrices. For a matrix A, let $\lambda_{\max}(A)$ and $\lambda_{\min}(A)$ represent the maximum and minimimum eigenvalues of A, respectively. Let $\sigma_i(A)$ denote the ith largest singular values of A; as such, $\sigma_1(A) \geq \sigma_2(A) \geq \cdots \geq \sigma_{m\wedge n}(A)$. The trace of a square $n \times n$ matrix is given by $\text{tr}(A) = \sum_{i\in[n]}\sigma_i(A)$. We use the following matrix norms for a matrix $A \in \mathbb{R}^{m\times n}$:

$$\text{Spectral Norm: } \|A\|_2 = \max_{\mathbf{x}\in\mathbb{S}^{n-1}}\|A\mathbf{x}\|_2 = \sigma_1(A) \tag{1}$$

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

Probability. We now discuss the probabilitistic concepts used in this work. We consider the general sub-Gaussian design, which is prevalent in the study of robust statistics (see e.g. [Awasthi et al., 2022, Bhatia et al., 2015, Jambulapati et al., 2020]).

Definition 4 (Sub-Gaussian Distribution)**.** *We say a vector $\mathbf{x}$ is sampled from a sub-Gaussian distribution with second-moment matrix Σ and sub-Gaussian proxy Γ if, for any $\mathbf{v} \in \mathbb{S}^{d-1}$,*

$$\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}}[\exp(t\mathbf{x}\cdot\mathbf{v})] \leq \exp\left(-\frac{t^2\|\Gamma\|_2}{2}\right).$$

A scalar random variable X is sub-Gaussian if there exists $K > 0$ such that for all $p \in \mathbb{N}$,

$$\|X\|_{L_p} \triangleq (\mathbf{E}|X|^p)^{1/p} \leq K\sqrt{p}.$$

Sub-Gaussian distributions are convenient to work with in robust statistics as the empirical mean of any subset of a set of i.i.d realizations of sub-Gaussian random variables is close to the mean of the distribution (see e.g. Steinhardt et al. [2018]). Sub-Gaussian distributions have tails that decay exponentially, i.e. at least as fast as Gaussian random variables. In Table 2 we introduce $L_4 \to L_2$ hypercontractivity, which we will formally define here.

Definition 5. *A distribution $\mathcal{D}$ is $L_4 \to L_2$ hypercontractive if there exists a constant L such that*

$$\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}} \|\mathbf{x}\|_2^4 \leq L \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}} \|\mathbf{x}\|_2^2 = L \operatorname{tr}(\Sigma).$$

2.2 Related Work

The use iterative thresholding for robust statistical estimation dates back to 1806 by Legendre Legendre [1806]. Iterative thresholding has been studied theoretically and applied empirically to various machine learning problems including linear regression, GLMs, and generative adversarial networks (GANs) [Bhatia et al., 2015, Hu et al., 2023, Mukhoty et al., 2019]. However, theoretical guarantees for its efficacy in learning nonlinear models are sparse and not sufficiently strong to justify practical usage of iterative thresholding.

We first introduce the statistical idea of a breakdown point. The breakdown point is defined as the smallest fraction of observations that can be replaced with arbitrary values to cause an estimator to take on arbitrarily large incorrect values. The algorithms presented in this paper have breakdown point $\Omega(1)$. This is an improvement over the robust algorithm given in Chen et al. [2013] which has breakdown point $\Omega(1/\sqrt{d})$. Recent papers on iterative thresholding (see e.g. [Bhatia et al., 2015, Shen and Sanghavi, 2019, Awasthi et al., 2022] also have breakdown point $\Omega(1)$.

Bhatia et al. [2015] study iterative thresholding for least squares regression / sparse recovery. In particular, one of their contributions is a gradient descent algorithm, TORRENT-GD, applicable when the covariates are sampled from a sub-Gaussian distribution. Their approximation bound (Theorem 2) relies on the fact that $\lambda_{\min}(\Sigma) = \lambda_{\max}(\Sigma)$, so that with sufficiently large sample size and sufficiently small corruption parameter ϵ, the condition number $\kappa(X)$ approaches 1. Bhatia et al. [2015] also provide guarantees on the performance of a full solve algorithm, TORRENT-FC, which after each thresholding step to obtain $(1-\epsilon)N$ samples sets $\mathbf{w}^{(t)}$ to be the minimizer of the squared loss over the selected $(1-\epsilon)N$ points. They study this algorithm in the presence of both adversarial and intrinsic noise. Their analysis guarantees $O(\nu)$ error when the intrinsic noise is sub-Gaussian with sub-Gaussian norm $O(\nu^2)$.

Shen and Sanghavi [2019] study iterative thresholding for learning generalized linear models (GLMs). In both the linear and non-linear case, their algorithms exhibit linear convergence. Their results imply a bound of $O(\nu)$ in the linear case. They further provide experimental evidence of the success of iterative thresholding when applied to neural networks.

More recently, Awasthi et al. [2022] 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, $O(\nu\epsilon\log(1/\epsilon))$. The algorithm studied by Awasthi et al. [2022] natively handles corruptions in labels, and to handle the case of corrupted variates, they first run a near-linear filtering algorithm from Dong et al. [2019] to obtain covariates that are sub-Gaussian with close to identity covariance.

3 Iterative Thresholding for gradient-based learning

In this section we introduce our algorithms for iterative thresholding gradient-based robust learning of linear and non-linear models. We start with the simple case of regression with multiple targets, as this provides a simple introduction to and instantiation of our general proof technique. As a corollary, we find a result regarding linear regression in the sub-Gaussian setting without covariate corruption; this result is compared with the existing literature [Awasthi et al., 2022, Bhatia et al., 2017, Shen

and Sanghavi, 2019]. Next, we consider the learning of non-linear neurons. This results in a suite of novel results regarding the use of iterative thresholding gradient-based learning that are incomparable with the existing literature.

3.1 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 Bhatia et al. [2017, Theorem 5] and Awasthi et al. [2022, Lemma A.1] by including covariate corruption without requiring a filtering algorithm, allowing variance in the optimal estimator, and accommodating a second-moment matrix for the uncorrupted data that is not the identity. 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 $Y \in \mathbb{R}^{K\times n}$ is

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

We furthermore define $\mathcal{R}(W; X, Y) = \frac{1}{(1-\epsilon)N}\mathcal{L}(W; X, Y)$ as the empirical risk function. We will first give some notation prior to presenting the algorithm.

Definition 6 (Hard Thresholding Operator)**.** *For any vector $\mathbf{x} \in \mathbb{R}^n$, define the permutation $\mathbf{x}_{(1)} \leq \mathbf{x}_{(2)} \ldots \leq \mathbf{x}_{(n)}$, the hard thresholding operator is given as*

$$\mathrm{HT}(\mathbf{x}; k) = \left\{i \in [n] : \mathbf{x}_i \in \left\{\mathbf{x}_{(1)}, \ldots, \mathbf{x}_{(k)}\right\}\right\}.$$

With the hard thresholding operator from Definition 6 in hand, we are now ready to present our algorithm for multi-linear regression.

Algorithm 1 Deterministic Gradient Descent Iterative Thresholding for Multi-Linear Regression

input: Possibly corrupted $X \in \mathbb{R}^{d\times N}$ with outputs $Y \in \mathbb{R}^{K\times N}$, corruption parameter $\epsilon = O(1)$, and $\|\mathbf{x}_i\| \leq B$ for all $i \in \mathrm{Q}$

output: $\nu\sqrt{BK\epsilon\log(1/\epsilon)}-$Approximate solution $W \in \mathbb{R}^{K\times d}$ to minimize $\|W - W^*\|_{\mathrm{F}}$

$W^{(0)} \leftarrow 0$

$\eta \leftarrow 0.1\kappa(\Sigma)$

$T \leftarrow O\left(\kappa^2(\Sigma)\log\left(\frac{\|W^*\|_{\mathrm{F}}}{\varepsilon}\right)\right)$

for $t \in [T]$ **do**

  $ri^{(t)} = \|W\mathbf{x}_i - \mathbf{y}_i\|^2 \quad \forall i \in [N]$ ▷ Calculate $\mathcal{L}$ for each sample

  $\mathrm{S}^{(t)} \leftarrow \mathrm{HT}(\boldsymbol{\nu}^{(t)}, (1-\epsilon)N)$ ▷ See Definition 6

  $W^{(t+1)} \leftarrow W^{(t)} - \eta\nabla\mathcal{R}(W^{(t)}; \mathrm{S}^{(t)})$ ▷ Gradient Descent Update

return: $W^{(T)}$

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 $O\left(Nd^2\log\left(\frac{\|\mathbf{w}^*\|_2}{\nu\epsilon}\right)\right)$ to obtain $O(\nu\epsilon\log(1/\epsilon))$ ℓ_2-approximation error.

Theorem 7. *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 $j \in [K]$, $[\mathbf{e}_i]_j \sim \mathcal{N}(0, \nu^2)$. Then after $O\left(\kappa(\Sigma)\log\left(\frac{\|W^*\|_{\mathrm{F}}}{\varepsilon}\right)\right)$ gradient descent iterations, $N = \Omega\left(\frac{d+\log(1/\delta)}{\epsilon^2}\right)$, and learning rate $\eta = 0.1\lambda_{\max}^{-2}(\Sigma)$, with probability exceeding $1 - 3T\delta$, Algorithm 1 returns $W^{(T)}$ such that*

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

We are able to recover the result of Lemma 4.2 in Awasthi et al. [2022] 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 Awasthi et al. [2022] returns a $O(\nu\epsilon\log(1/\epsilon))$ in time $O\big(\frac{1}{\epsilon^2}\big(Nd^2 + d^3\big)\big)$ and the same algorithm studied in Bhatia et al. [2015], TORRENT-FC otains $O(\nu)$ approximation error in run-time $O\Big(\log\Big(\frac{1}{\sqrt{n}}\frac{\|\mathbf{b}\|}{\nu\epsilon\log(1/\epsilon)}\Big)(Nd^2 + d^3)\Big)$, with the gradient descent based approach, we are able to improve the runtime to $O\Big(\log\Big(\frac{\|\mathbf{w}^*\|}{\nu\epsilon\log(1/\epsilon)}\Big)Nd^2\Big)$ 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 Bhatia et al. [2015], 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 Shen and Sanghavi [2019] 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 8. *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 $\mathbf{x}_i$ are sampled from a sub-Gaussian distribution with second-moment matrix I and sub-Gaussian proxy I and $\mathbf{y}_i = \mathbf{w}^* \cdot \mathbf{x}_i + \xi_i$ where $\xi_i \sim \mathcal{N}(0, \nu^2)$ for $i \in \mathrm{P}$. Then after $O\left(\log\left(\frac{\|\mathbf{w}^*\|_2}{\varepsilon}\right)\right)$ gradient descent iterations, sample size $N = \Omega\Big(\frac{d+\log(1/\delta)}{\epsilon}\Big)$, and learning rate $\eta = 0.1$, with probability exceeding $1-\delta$, Algorithm 1 returns $\mathbf{w}^{(T)}$ such that*

$$\|\mathbf{w}^{(T)} - \mathbf{w}^*\|_2 \le \varepsilon + O\Big(\nu\epsilon\sqrt{KB\log(1/\epsilon)}\Big)$$

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,

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

The second relation given in Corollary 8 matches the best known bound for robust linear regression with iterative thresholding. The first relation given in Corollary 8 is the extension to handle second-moment matrices which do not have unitary condition number as well as corrupted covariates.

3.2 Activation Functions

We first give properties of the non-linear functions we will be learning. All functions we henceforth study will have some subset of the below listed properties.

Property 9. *σ is a continuous, monotically increasing, and differentiable almost everywhere.*

Property 10. *σ is Lipschitz, i.e. $|\sigma(x) - \sigma(y)| \le \|\sigma\|_{\mathrm{lip}}|x - y|$.*

Property 11. *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 9, 10, and 11. Property 11 does not hold for the ReLU function, and therefore we require stronger conditions for our approximation bounds to hold.

3.3 Learning Sigmoidal Neurons

We now study the problem of minimizing the ℓ_2 loss for sigmoidal type neurons. For a single training sample $(\mathbf{x}_i, y_i) \sim \mathcal{D}$, the loss is given as follows,

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

Theorem 12. *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 proxy Γ, 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 9, 10,*

Algorithm 2 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 ν, corruption parameter $\epsilon = O(1)$, and small constant α.

output: $\nu\sqrt{\epsilon\log(1/\epsilon)}-$Approximate solution $\mathbf{w} \in \mathbb{R}^d$ to minimize $\|\mathbf{w} - \mathbf{w}^*\|_2$.

$\mathbf{w}^{(0)} \sim \mathcal{B}_d(\alpha\|\mathbf{w}^*\|)$
$\eta \leftarrow 0.1\kappa^{-2}(\Sigma)$
$T \leftarrow O\Big(\kappa^2(\Sigma)\log\Big(\frac{\|\mathbf{w}^*\|_2}{\varepsilon}\Big)\Big)$
for $t \in [T]$ **do**
  $\nu_i^{(t)} = (\sigma(\mathbf{x}_i^\top \mathbf{w}^{(t)}) - y_i)^2 \quad \forall i \in [N]$ ▷ Calculate $\mathcal{L}$ for each point
  $\mathrm{S}^{(t)} \leftarrow \mathrm{HT}(\boldsymbol{\nu}^{(t)}, (1-\epsilon)N)$ ▷ See Definition 6
  $\mathbf{w}^{(t+1)} \leftarrow \mathbf{w}^{(t)} - \eta\nabla\mathcal{R}(\mathbf{w}^{(t)}; \mathrm{S}^{(t)})$ ▷ Gradient Descent Update

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

and 11. Then after $O\left(\kappa(\Sigma)\log\left(\frac{\|\mathbf{w}^*\|}{\varepsilon}\right)\right)$ *gradient descent iterations, then with probability exceeding* $1-\delta$,

$$\|\mathbf{w}^{(T)} - \mathbf{w}^*\|_2 \leq O\Big(\gamma^{-2}\lambda_{\min}^{-1}(\Sigma)\|\sigma\|_{\mathrm{lip}}^2\epsilon\sqrt{B\log(1/\epsilon)}\Big) + O\Big(\gamma^{-2}\|\sigma\|_{\mathrm{lip}}^2\nu\kappa(\Sigma)\sqrt{\epsilon\log(1/\epsilon)}\Big)$$

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

The result is slightly weaker than the approximation bound achieved in Theorem 7 by a square root factor. In the proof for our linear regression result, we are able to leverage the consistency of linear regression estimators, i.e. as $N \to \infty$, we have $\arg\min_{\mathbf{w}} \mathcal{L}(\mathbf{w}, \mathcal{D}) \to \arg\min_{\mathbf{w}} \mathbf{E}_{\mathcal{D}}[\mathcal{L}(\mathbf{w}, \mathcal{D})]$.

Proof.[Sketch] We will give a general sketch of the proof for our sigmoidal neuron result. Let $t \in [T]$, we then have,

$$\begin{aligned}\|\mathbf{w}^{(t+1)} - \mathbf{w}^*\| &\leq \|\mathbf{w}^{(t)} - \eta\nabla\mathcal{R}(\mathbf{w}^{(t)}; \mathrm{S}^{(t)}) - \mathbf{w}^*\|_2 \\ &\leq \underbrace{\|\mathbf{w}^{(t)} - \mathbf{w}^* - \eta\nabla\mathcal{R}(\mathbf{w}^{(t)}; \mathrm{S}^{(t)} \cap \mathrm{P})\|}_{I} + \underbrace{\|\eta\nabla\mathcal{R}(\mathbf{w}^{(t)}; \mathrm{S}^{(t)} \cap \mathrm{Q})\|_2}_{II}. \end{aligned} \tag{3}$$

In the above, the second relation follows from noting

$$\nabla\mathcal{R}(\mathbf{w}^{(t)}; \mathrm{S}^{(t)}) = \nabla\mathcal{R}(\mathbf{w}^{(t)}; \mathrm{S}^{(t)} \cap \mathrm{P}) + \nabla\mathcal{R}(\mathbf{w}^{(t)}; \mathrm{S}^{(t)} \cap \mathrm{Q}).$$

and then applying the triangle inequality. We upper bound I through its square,

$$I^2 = \|\mathbf{w}^{(t)} - \mathbf{w}^*\|_2^2 - \underbrace{2\eta \cdot \langle \mathbf{w}^{(t)} - \mathbf{w}^*, \nabla\mathcal{R}(\mathbf{w}^{(t)}; \mathrm{S}^{(t)} \cap \mathrm{P})\rangle}_{I_1} + \underbrace{\eta^2 \cdot \|\nabla\mathcal{R}(\mathbf{w}^{(t)}; \mathrm{S}^{(t)} \cap \mathrm{P})\|^2}_{I_2}.$$

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,

$$I_1 \geq (1-2\epsilon)c_1\|\mathbf{w}^{(t)} - \mathbf{w}^*\|_2^2 - c_3\|\mathbf{w}^{(t)} - \mathbf{w}^*\|_2$$

where c_3 is a term that is dependent on the variance of the Gaussian noise, one of our contributions is that $c_3 = O(\nu\sqrt{\epsilon\log(1/\epsilon)})$ when N sufficiently large for the activation functions studied in the text. Next, an application of Peter-Paul's inequality[1] gives us,

$$I_1 \geq ((1-2\epsilon)c_1 - c_3c_1)\|\mathbf{w}^{(t)} - \mathbf{w}^*\|_2^2 - c_3c_1^{-1}$$

We next show there exists a positive constant c_2, such that

$$I_2^2 \leq (1-\epsilon)c_2\|\mathbf{w}^{(t)} - \mathbf{w}^*\|_2^2$$

[1]Peter-Paul's inequality states that for any $p, q > 1$ such that $\frac{1}{p} + \frac{1}{q} = 1$, then for every t, we have $ab \leq \frac{t^p a^p}{p} + \frac{t^{-q}b^q}{q}$. Consider Young's Inequality and replace a with at^p and b with bt^{-p}.

Then, solving a simple quadratic equation, we have that $\eta^2 C_2 \leq \eta c_1 c_3$ for a $c_3 \in (0,1)$ when we choose $\eta \leq \frac{c_1 c_3}{c_2}$ 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,

$$\sum_{i \in \mathrm{S}^{(t)} \cap \mathrm{Q}} \mathcal{L}(\mathbf{w}^{(t)}; \mathbf{x}_i, y_i) \leq \sum_{i \in \mathrm{P} \setminus \mathrm{S}^{(t)}} \mathcal{L}(\mathbf{w}^{(t)}; \mathbf{x}_i, y_i) \tag{4}$$

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,

$$II \leq \epsilon c_4 \|\mathbf{w}^{(t)} - \mathbf{w}^*\|_2^2$$

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

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

In our deterministic algorithm, we choose $\mathbf{w}^{(0)} = \mathbf{0}$ and thus we have $\|\mathbf{w}^* - \mathbf{w}^{(0)}\|_2 = \|\mathbf{w}^*\|_2$. In our randomized algorithm, we have from Lemma 14, with high probability $\|\mathbf{w}^* - \mathbf{w}^{(0)}\| \leq \|\mathbf{w}^*\|$, giving us the desired bound. ■

3.3.1 Algorithmic ϵ

In practice, one does not have access to the true corruption rate in the dataset. We therefore differentiate between the algorithmic corruption parameter ϵ and the true corruption parameter of the dataset ϵ^*. Consider the case when $\epsilon \geq \epsilon^*$ i.e., we overestimate the corruption rate of the dataset. We have at any iteration $t \in [T]$, that $|\mathrm{S}_\epsilon \cap \mathrm{Q}| \geq |\mathrm{S}_{\epsilon^*} \cap \mathrm{Q}|$ as $(1-\epsilon)N \geq (1-\epsilon^*)N$. We similarly have $|\mathrm{P} \setminus \mathrm{S}_{\epsilon^*}^{(t)}| \leq |\mathrm{P} \setminus \mathrm{S}_\epsilon^{(t)}|$ as the thresholded set is smaller in cardinality. We thus have our key step, Equation (4), will still hold when $\epsilon \geq \epsilon^*$.

3.4 Learning Leaky-ReLU Neurons

We will now consider learning a neuron with the Leaky-ReLU function. We can note that Property 9, 10, and 11 all hold for the Leaky-ReLU. More conveniently, we have γ in Property 11 is constant over $\mathbb{R}$. In our proof we are able to leverage the fact that the second derivative is zero almost surely.

Theorem 13. *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 $y_i = \sigma(\mathbf{x}^\top \mathbf{w}^*) + \xi_i$ for $\xi_i \sim \mathcal{N}(0, \nu^2)$ where $\sigma(x) = \max\{\gamma x, x\}$. Then after $O\left(\gamma^{-2}\kappa(\Sigma)\log\left(\frac{\|W^*\|_\mathrm{F}}{\varepsilon}\right)\right)$ gradient descent iterations and $\epsilon \leq \frac{\gamma^2 \lambda_{\min}(\Sigma)}{\sqrt{32 B \lambda_{\max}(\Sigma)}}$, with probability exceeding $1 - 4\delta$, Algorithm 2 with learning rate $\eta = O(\kappa^{-2}(\Sigma))$ returns $\mathbf{w}^{(T)}$ such that*

$$\|\mathbf{w}^{(T)} - \mathbf{w}^*\|_2 \leq O\left(\nu \|\Gamma\|_2 \lambda_{\min}^{-1}(\Sigma) \sqrt{\epsilon \log(1/\epsilon)}\right) + O\left(\gamma^{-2}\kappa(\Sigma)\nu\epsilon\sqrt{B\log(1/\epsilon)}\right)$$

Proof. The proof is deferred to Appendix B.2.1. ■

Our proof and result also hold for a smoothened version of the Leaky-ReLU. To avoid the kink at $x = 0$, one can consider the Smooth-Leaky-ReLU, defined as

$$\text{Smooth-Leaky-ReLU}(x) = \alpha x + (1 - \alpha)\log(1 + e^x)$$

for an $\alpha \in (0, 1)$. The Smooth-Leaky-ReLU satisfies Properties 9, 10, and 11, and is convex, indicating that Theorem 13 can be applied.

3.5 Learning ReLU Neurons

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

Lemma 14 (Theorem 3.4 in Du et al. [2018]). *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 probabistic guarantees on the convergence of learning a ReLU neuron.

Theorem 15. *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 $\mathbf{x}_i$ are sampled from a sub-Gaussian distribution with second-moment matrix Σ and sub-Gaussian proxy Γ and the output is given as $y_i = \sigma(\mathbf{w}^* \cdot \mathbf{x}_i) + \xi_i$ for $\xi_i \sim \mathcal{N}(0, \nu^2)$ for all $i \in$ P. Then after $O\Big(\kappa(\Sigma)\log\Big(\frac{\|\mathbf{w}^*\|}{\epsilon}\Big)\Big)$ gradient descent iterations and $N = \Omega\Big(\frac{d + \log(1/\delta)}{\epsilon}\Big)$, then with probability exceeding $1 - 3\delta$, Algorithm 2 with learning rate $\eta = O\Big(\frac{\lambda_{\min}(\Sigma)}{\lambda^2_{\max}(\Sigma)}\Big)$ returns $\mathbf{w}^{(T)}$ such that*

$$\|\mathbf{w}^{(T)} - \mathbf{w}^*\|_2 \leq O\Big(\nu\|\Gamma\|_2\lambda^{-1}_{\min}(\Sigma)\sqrt{\epsilon\log(1/\epsilon)}\Big) + O\Big(\kappa(\Sigma)\nu\epsilon\sqrt{B\log(1/\epsilon)}\Big)$$

Proof. The proof is deferred to Appendix B.3 ■

Our proof for learning ReLU neurons follows the same high level structure as learning sigmoidal or leaky-ReLU neurons, however it is significantly more technical. We also note that Lemma 14 implies that randomized restarts with high probability will return a vector with $O(\sqrt{\epsilon\log(1/\epsilon)})$ ℓ_2 approximation error.

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 [Hu et al., 2023, Shen and Sanghavi, 2019] 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. There are many directions for future work. Regarding iterative thresholding, our paper has established upper bounds on the approximation error of activation functions, an interesting next step is on upper bounds for the sum of activation functions, i.e. one hidden-layer neural networks. In the linear regression case, Gao [2020] derived the minimax optimal error of $O(\sigma\epsilon)$. Establishing this result for sigmoidal, leaky-ReLU, and ReLU functions would be helpful in the discussing the strength of our bounds. Deriving upper and lower bounds for iterative thresholding for binary classification is a good direction for future research. In ± 1 classification, considering $y = \text{sign}(\mathbf{w}^* \cdot \mathbf{x} + \xi)$, the sign function adds adds an interesting complication. A study on if our current techniques can also handle the sign function would be interesting.

Acknowledgments and Disclosure of Funding

The authors thank Weihao Kong for helpful discussions.

References

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.

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.

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.

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

Yudong Chen, Constantine Caramanis, and Shie Mannor. Robust sparse regression under adversarial corruption. In *International conference on machine learning*, pages 774–782. PMLR, 2013.

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.

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.

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

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

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

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.

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.

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

Martin A. Fischler and Robert C. Bolles. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Commun. ACM*, 24(6):381–395, jun 1981. ISSN 0001-0782. doi: 10.1145/358669.358692.

Chao Gao. Robust regression via mutivariate regression depth. *Bernoulli*, 26(2):1139 – 1170, 2020. doi: 10.3150/19-BEJ1144. URL `https://doi.org/10.3150/19-BEJ1144`.

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

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.

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

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

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

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

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.

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.

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

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.

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

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

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.

Jacob Steinhardt, Moses Charikar, and Gregory Valiant. Resilience: A criterion for learning in the presence of arbitrary outliers. In Anna R. Karlin, editor, *9th Innovations in Theoretical Computer Science Conference, ITCS 2018, January 11-14, 2018, Cambridge, MA, USA*, volume 94 of *LIPIcs*, pages 45:1–45:21. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2018. doi: 10.4230/LIPIcs.ITCS.2018.45. URL `https://doi.org/10.4230/LIPIcs.ITCS.2018.45`.

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

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

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 $(\mathbf{x}_j, y_j)$ for $j \in \mathrm{Q}$ has been arbitrarily corrupted 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 7

We will first give some preliminaries for our proof. 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 16. *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 17. *Suppose A, B, C are conformal matrices, then*

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

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 Petersen et al. [2008] Equations (102) and (119) (where we set $B = I$ and $C = 0$). 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 Appendix 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

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

In contrast with our proof sketch in Section 3.3, we introduce a consistency estimate in IV. We first will upper bound I through the expansion of its square.

$$\begin{aligned}&\|W^{(t)} - W^* - \eta \nabla \mathcal{R}(W^{(t)}; \mathrm{TP}) + \eta \nabla \mathcal{R}(W^*; \mathrm{TP})\|_{\mathrm{F}}^2 = \|W^{(t)} - W^*\|_{\mathrm{F}}^2 \\ &\underbrace{-2\eta \cdot \mathrm{tr}\big((W^{(t)} - W^*)^\top (\nabla \mathcal{R}(W^{(t)}; \mathrm{TP}) - \nabla \mathcal{R}(W^*; \mathrm{TP}))\big)}_{I_1} + \underbrace{\|\eta \nabla \mathcal{R}(W^{(t)}; \mathrm{TP}) - \eta \nabla \mathcal{R}(W^{(t)}; \mathrm{TP})\|_{\mathrm{F}}^2}_{I_2}\end{aligned}$$

We first lower bound I_1, we will obtain a lower bound that is less than I_2, allowing us to cancel the I_2 term.

$$\begin{aligned}
I_1 &= 2\eta \cdot \mathrm{tr}\big((W^{(t)} - W^*)^\top (\nabla \mathcal{R}(W^{(t)}; \mathrm{TP}) - \nabla \mathcal{R}(W^*; \mathrm{TP}))\big) \\
&\overset{\text{def}}{=} \frac{4\eta}{(1-\epsilon)N} \cdot \mathrm{tr}\big((W^{(t)} - W^*)^\top \big((W^{(t)} X_{\mathrm{TP}} - Y_{\mathrm{TP}}) X_{\mathrm{TP}}^\top - E_{\mathrm{TP}} X_{\mathrm{TP}}^\top\big)\big) \\
&\overset{(i)}{=} \frac{4\eta}{(1-\epsilon)N} \cdot \mathrm{tr}\Big((W^{(t)} - W^*)^\top (W^{(t)} - W^*) X_{\mathrm{TP}} X_{\mathrm{TP}}^\top\Big)
\end{aligned}$$

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

$$\begin{aligned}
I_1 &\overset{(ii)}{=} \frac{4\eta}{(1-\epsilon)N} \cdot \mathrm{tr}\Big((W^{(t)} - W^*) X_{\mathrm{TP}} X_{\mathrm{TP}}^\top (W^{(t)} - W^*)^\top\Big) \\
&\overset{(iii)}{=} \frac{4\eta}{(1-\epsilon)N} \cdot \langle \mathrm{vec}((W^{(t)} - W^*)^\top), \mathrm{vec}(X_{\mathrm{TP}} X_{\mathrm{TP}}^\top (W^{(t)} - W^*)^\top) \rangle \\
&\overset{(iv)}{=} \frac{4\eta}{(1-\epsilon)N} \cdot \langle \mathrm{vec}((W^{(t)} - W^*)^\top), (I \otimes X_{\mathrm{TP}} X_{\mathrm{TP}}^\top) \mathrm{vec}((W^{(t)} - W^*)^\top) \rangle \\
&\overset{(v)}{=} \frac{4\eta}{(1-\epsilon)N} \cdot \sum_{k \in [K]} \big\langle \mathbf{w}_k^{(t)} - \mathbf{w}_k^*, X_{\mathrm{TP}} X_{\mathrm{TP}}^\top (\mathbf{w}_k^{(t)} - \mathbf{w}_k^*) \big\rangle \\
&\overset{(vi)}{\geq} \frac{2\eta}{(1-\epsilon)N} \cdot \sum_{k \in [K]} \big(\lambda_{\min}(X_{\mathrm{TP}} X_{\mathrm{TP}}^\top) \|\mathbf{w}_k^{(t)} - \mathbf{w}_k^*\|_2^2 + \|X_{\mathrm{TP}} X_{\mathrm{TP}}^\top\|_2^{-1} \|X_{\mathrm{TP}} X_{\mathrm{TP}}^\top (\mathbf{w}_k^{(t)} - \mathbf{w}_k^*)\|_2^2\big) \\
&= \frac{2\eta}{(1-\epsilon)N} \cdot \lambda_{\min}(X_{\mathrm{TP}} X_{\mathrm{TP}}^\top) \|W^{(t)} - W^*\|_{\mathrm{F}}^2 + \underbrace{\frac{2\eta}{(1-\epsilon)N} \cdot \|X_{\mathrm{TP}} X_{\mathrm{TP}}^\top\|_2^{-1} \|(W^{(t)} - W^*) X_{\mathrm{TP}} X_{\mathrm{TP}}^\top\|_{\mathrm{F}}^2}_{I_{12}}
\end{aligned}$$

In the above, (ii) follows from the cyclic property of the trace, (iii) follows from the relation given in Lemma 16, (iv) holds from the relation given in Lemma 17, the inequality in (vi) follows from Lemma 37, and in (v) we apply Lemma 17, 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 corrupted gradient term.

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

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 Hard Thresholding Operator (see Definition 6), and (ix) follows from the sub-additivity of the Frobenius norm. We will now upper bound III.

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

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, IV.

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

We can then have from Lemma 33,

$$\|E_{\mathrm{P}}X_{\mathrm{P}}^\top\|_{\mathrm{F}}^2 \leq 2\sigma^2\big(K\|X_{\mathrm{P}}\|_{\mathrm{F}}^2\log\big(2N^2/\delta\big)\big) \leq 2\sigma^2\big((1-\epsilon)Nd\lambda_{\max}(\Sigma)\log\big(2N^2/\delta\big)\big)$$

We then have with failure probability at most δ,

$$\frac{2\eta}{(1-\epsilon)N}\cdot\|E_{\mathrm{P}}X_{\mathrm{P}}^\top\|_{\mathrm{F}} \leq \eta\cdot\sqrt{\frac{8\sigma^2d\lambda_{\max}(\Sigma)\log(2N^2/\delta)}{(1-\epsilon)N}}$$

We then have from our choice of $\eta = 0.1\lambda_{\max}^{-1}(\Sigma)$, we have $I_2 \leq I_{12}$. We thus obtain from noting that $\sqrt{1-2x} \leq 1-x$ for any $x \leq 1/2$,

$$\|W^{(t+1)}-W^*\|_{\mathrm{F}} \leq \|W^{(t)}-W^*\|_{\mathrm{F}}\left(1-\frac{2\eta}{(1-\epsilon)N}\cdot\lambda_{\min}(X_{\mathrm{TP}}X_{\mathrm{TP}}^\top)+\frac{2\eta}{(1-\epsilon)N}\cdot\|X_{\mathrm{FP}}\|_2\|X_{\mathrm{FN}}\|_2\right)$$

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

Step 3: Concentration Bounds. From Proposition 19 and our ℓ_2 bounded corrupted covariate assumption, we obtain with failure probability at most δ,

$$\|E_{\mathrm{FN}}\|_{\mathrm{F}}\|X_{\mathrm{FP}}\|_{\mathrm{F}} \leq N\epsilon\sigma\sqrt{30KB\log(1/\epsilon)}$$

From Lemma 33, we have when $N \geq \log(1/\delta)$, with failure probability at most δ,

$$\|E_{\mathrm{FN}}X_{\mathrm{FN}}^\top\|_{\mathrm{F}} \leq \sqrt{6K\log N}\|X_{\mathrm{FN}}\|_2$$

Then utilizing the result from Lemma 21, we have with failure probability at most δ,

$$\sqrt{6K\log N}\|X_{\mathrm{FN}}\|_2 \leq \sqrt{60K\lambda_{\max}(\Sigma)N\log N\cdot\epsilon\log(1/\epsilon)}$$

Noting that $|\mathrm{S}^{(t)}\cap P| \geq (1-2\epsilon)N$, we have from Lemma 21 for $\epsilon \leq \frac{1}{60}\cdot\kappa^{-1}(\Sigma)$, the minimum eigenvalue satisfies with failure probability at most δ,

$$\lambda_{\min}(X_{\mathrm{TP}}X_{\mathrm{TP}}^\top) \geq \frac{N}{4}\cdot\lambda_{\min}(\Sigma)$$

Then when $\epsilon \leq \sqrt{\frac{1}{960B}\cdot\kappa^{-1}(\Sigma)\lambda_{\min}(\Sigma)}$, we have with high probability,

$$\|X_{\mathrm{FP}}\|_2\|X_{\mathrm{FN}}\|_2 - \lambda_{\min}(X_{\mathrm{TP}}X_{\mathrm{TP}}^\top) \leq \frac{\eta}{4(1-\epsilon)}\cdot\lambda_{\min}(\Sigma)$$

Combining our estimates, we have

$$\|W^{(t+1)}-W^*\|_{\mathrm{F}} \leq \|W^{(t)}-W^*\|_{\mathrm{F}}\left(1-\frac{\eta}{4(1-\epsilon)}\cdot\lambda_{\min}(\Sigma)\right)+\eta\cdot\sigma\epsilon\sqrt{480KB\log(1/\epsilon)}$$

$$+\frac{\eta}{(1-\epsilon)\sqrt{N}}\cdot\sqrt{240K\log(N)\lambda_{\max}(\Sigma)\epsilon\log(1/\epsilon)}+\frac{\eta}{(1-\epsilon)\sqrt{N}}\cdot\sigma\sqrt{11d\lambda_{\max}(\Sigma)\log(2N^2/\delta)}$$

with probability exceeding $1-3\delta$. Then, when $N = \widetilde{\Omega}\left(\frac{d\lambda_{\max}(\Sigma)+\log(1/\delta)}{\epsilon^2KB}\right)$ we have

$$\|W^{(t+1)}-W^*\|_{\mathrm{F}} \leq \|W^{(t)}-W^*\|_{\mathrm{F}}\left(1-\frac{\eta}{4(1-\epsilon)}\cdot\lambda_{\min}(\Sigma)\right)+\eta\cdot\sigma\epsilon\sqrt{4320KB\log(1/\epsilon)}$$

Then, solving for the induction with an infinite sum, referring to the proof sketch in Section 3.3 and our choice of η, we have after $O\left(\kappa(\Sigma)\cdot\log\left(\frac{\|W^*\|_{\mathrm{F}}}{\varepsilon}\right)\right)$ iterations,

$$\|W^{(T)}-W^*\|_{\mathrm{F}} \leq \varepsilon + \frac{\sigma\epsilon\sqrt{34560KB\log(1/\epsilon)}}{\lambda_{\min}(\Sigma)}$$

Our proof is complete. ■

B Proofs for Learning Nonlinear Neurons

In this section we present our omitted proofs for the approximation bounds for learning nonlinear neurons with Algorithm 2.

B.1 Sigmoidal Neurons

In this section we will give the ommited proofs for the ℓ_2 approximation bounds for learning sigmoidal neurons with Algorithm 2 in the Strong ϵ-Contamination Model.

B.1.1 Proof of Theorem 12

Proof. From Algorithm 2, 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 $O(\epsilon \log(1/\epsilon))$.
Step 1: Linear Convergence.

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

We will expand I through its square and give an upper bound.

$$I^2 = \|\mathbf{w}^{(t)} - \mathbf{w}^*\|_2^2 - \underbrace{2\eta \cdot \langle \mathbf{w}^{(t)} - \mathbf{w}^*, \nabla \mathcal{R}(\mathbf{w}^{(t)}; \mathrm{TP}) \rangle}_{I_1} + \underbrace{\eta^2 \cdot \|\nabla \mathcal{R}(\mathbf{w}^{(t)}; \mathrm{TP})\|_2^2}_{I_2}$$

We now lower bound I_1. Note from the randomized initialization, we have $\|\mathbf{w}^{(0)} - \mathbf{w}^*\| \leq \mathbf{w}^*$. Then, noting that $\|\mathbf{w}^*\| \leq R$, 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 11 and note there exists a γ s.t. $\sigma'(x) \geq \gamma > 0$ for all $x \in \mathbb{R}$ s.t. $x \leq 2RB$.

$$\begin{aligned}
I_1 &= \frac{4\eta}{(1-\epsilon)N} \cdot \langle \mathbf{w}^{(t)} - \mathbf{w}^*, \sum_{i \in \mathrm{TP}} (\sigma(\mathbf{w}^{(t)} \cdot \mathbf{x}_i) - y_i) \cdot \sigma'(\mathbf{w}^{(t)} \cdot \mathbf{x}_i) \cdot \mathbf{x}_i \rangle \\
&= \frac{4\eta}{(1-\epsilon)N} \cdot \langle \mathbf{w}^{(t)} - \mathbf{w}^*, \sum_{i \in \mathrm{TP}} (\sigma(\mathbf{w}^{(t)} \cdot \mathbf{x}_i) - \sigma(\mathbf{w}^* \cdot \mathbf{x}_i) + \xi_i) \cdot \sigma'(\mathbf{w}^{(t)} \cdot \mathbf{x}_i) \cdot \mathbf{x}_i \rangle \\
&\overset{(i)}{\geq} \underbrace{\frac{4\eta}{(1-\epsilon)N} \cdot \gamma^2 \lambda_{\min}(X_{\mathrm{TP}} X_{\mathrm{TP}}^\top) \|\mathbf{w}^{(t)} - \mathbf{w}^*\|_2^2}_{I_{11}} - \frac{4\eta}{(1-\epsilon)N} \cdot \|\mathbf{w}^{(t)} - \mathbf{w}^*\|_2 \Big\| \sum_{i \in \mathrm{TP}} \xi_i \sigma'(\mathbf{w}^{(t)} \cdot \mathbf{x}_i) \cdot \mathbf{x}_i \Big\|_2
\end{aligned}$$

In the above, (i) follows from defining the function $h : \mathbb{R}^d \mapsto \mathbb{R}$,

$$\mathbf{w} \mapsto \int_{\mathbb{R}^d} \left(\int \sigma \right) (\mathbf{w} \cdot \mathbf{x}) dP(\mathbf{x})$$

where $dP(\mathbf{x}) = \mathbf{1}\{\mathbf{x} \in \mathrm{TP}\}$. Then from Property 11, we have that σ' is strictly positive over a compact domain, this implies that $\int \sigma$ is strongly convex over a compact domain. Then we can calculate the Hessian,

$$\nabla^2 h(\mathbf{w}) = \int_{\mathbb{R}^d} \sigma'(\mathbf{w} \cdot \mathbf{x}) \cdot \mathbf{x}\mathbf{x}^\top dP(\mathbf{x}) \succeq \gamma \lambda_{\min}(X_{\mathrm{TP}} X_{\mathrm{TP}}^\top) \cdot I$$

With the strong convexity in hand, we have,

$$\langle \mathbf{w}^{(t)} - \mathbf{w}^*, \sum_{i \in \mathrm{TP}} (\sigma(\mathbf{w}^{(t)} \cdot \mathbf{x}_i) - \sigma(\mathbf{w}^* \cdot \mathbf{x}_i) \cdot \mathbf{x}_i \rangle$$
$$= \langle \mathbf{w}^{(t)} - \mathbf{w}^*, \int_0^1 \nabla^2 h(\mathbf{w}^* + \theta(\mathbf{w}^{(t)} - \mathbf{w}^*)) d\theta \cdot (\mathbf{w}^{(t)} - \mathbf{w}^*) \rangle \overset{(ii)}{\geq} \gamma \lambda_{\min}(X_{\mathrm{TP}} X_{\mathrm{TP}}^\top) \cdot \|\mathbf{w}^{(t)} - \mathbf{w}^*\|_2^2$$

In the above, (ii) follows from noting that for any $\theta \in [0, 1]$ we have,

$$\|(1-\theta)\mathbf{w}^* + \theta \mathbf{w}^{(t)}\|_2 \leq \theta \|\mathbf{w}^* - \mathbf{w}^{(t)}\|_2 + \|\mathbf{w}^*\|_2 \leq 2\|\mathbf{w}^*\|_2 = 2R$$

We can then use the same γ as previously defined. Then from an application of Peter-Paul's Inequality, we obtain

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

We next upper bound I_2. We note that σ is $\|\sigma\|_{\mathrm{lip}}$-Lipschitz, then from an application of the triangle inequality, we obtain

$$I_2 \overset{\mathrm{def}}{=} \frac{4\eta^2}{[(1-\epsilon)N]^2} \cdot \Big\| \sum_{i \in \mathrm{TP}} (\sigma(\mathbf{w}^{(t)} \cdot \mathbf{x}_i) - \sigma(\mathbf{w}^* \cdot \mathbf{x}_i) + \xi_i) \cdot \sigma'(\mathbf{w}^{(t)} \cdot \mathbf{x}_i) \cdot \mathbf{x}_i \Big\|_2^2$$
$$\overset{(i)}{=} \frac{4\eta^2}{[(1-\epsilon)N]^2} \cdot \Big\| \sum_{i \in \mathrm{TP}} \mathbf{x}_i \mathbf{x}_i^\top (\mathbf{w} - \mathbf{w}^*) \cdot \sigma'(c_i) \sigma'(\mathbf{w}^{(t)} \cdot \mathbf{x}_i) + \xi_i \cdot \sigma'(\mathbf{w}^{(t)} \cdot \mathbf{x}_i) \cdot \mathbf{x}_i \Big\|_2^2$$
$$\leq \underbrace{\frac{8\eta^2}{[(1-\epsilon)N]^2} \cdot \|\sigma\|_{\mathrm{lip}}^4 \lambda_{\max}^2(X_{\mathrm{TP}} X_{\mathrm{TP}}^\top) \|\mathbf{w}^{(t)} - \mathbf{w}^*\|_2^2}_{I_{21}} + \frac{8\eta^2}{[(1-\epsilon)N]^2} \cdot \Big\| \sum_{i \in \mathrm{TP}} \xi_i \sigma'(\mathbf{w}^{(t)} \cdot \mathbf{x}_i) \cdot \mathbf{x}_i \Big\|_2^2$$

In the above, (i) follows from noting there exists a constant $c_i \in [\mathbf{w}^{(t)} \cdot \mathbf{x}_i, \mathbf{w}^* \cdot \mathbf{x}_i]$ such that

$$\sigma'(c_i)(\mathbf{w}^{(t)} - \mathbf{w}^*) \cdot \mathbf{x}_i = \sigma(\mathbf{w}^{(t)} \cdot \mathbf{x}_i) - \sigma(\mathbf{w}^* \cdot \mathbf{x}_i)$$

from the Mean-Value Theorem. 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 have $I_{21} \leq 0.5 I_{11}$. We now bound the corrupted gradient term.

$$II^2 = \frac{4\eta^2}{[(1-\epsilon)N]^2} \cdot \Big\| \sum_{i \in \mathrm{FP}} (\sigma(\mathbf{w}^{(t)} \cdot \mathbf{x}_i) - y_i) \cdot \sigma'(\mathbf{w} \cdot \mathbf{x}_i) \cdot \mathbf{x}_i \Big\|_2^2$$
$$\overset{(ii)}{\leq} \frac{4\eta^2}{[(1-\epsilon)N]^2} \cdot \|\sigma\|_{\mathrm{lip}}^2 \|X_{\mathrm{FP}} X_{\mathrm{FP}}^\top\|_2 \sum_{i \in \mathrm{FP}} (\sigma(\mathbf{w}^{(t)} \cdot \mathbf{x}_i) - y_i)^2$$
$$\overset{(iii)}{\leq} \frac{4\eta^2}{[(1-\epsilon)N]^2} \cdot \|\sigma\|_{\mathrm{lip}}^2 \|X_{\mathrm{FP}} X_{\mathrm{FP}}^\top\|_2 \sum_{i \in \mathrm{FN}} (\sigma(\mathbf{w}^{(t)} \cdot \mathbf{x}_i) - y_i)^2$$
$$= \frac{4\eta^2}{[(1-\epsilon)N]^2} \cdot \|\sigma\|_{\mathrm{lip}}^2 \|X_{\mathrm{FP}} X_{\mathrm{FP}}^\top\|_2 \sum_{i \in \mathrm{FN}} (\sigma(\mathbf{w}^{(t)} \cdot \mathbf{x}_i) - \sigma(\mathbf{w}^* \cdot \mathbf{x}_i) + \xi_i)^2$$
$$\overset{(iv)}{\leq} \frac{8\eta^2}{[(1-\epsilon)N]^2} \cdot \|\sigma\|_{\mathrm{lip}}^2 \|X_{\mathrm{FP}} X_{\mathrm{FP}}^\top\|_2 \Big(\|\sigma\|_{\mathrm{lip}}^2 \cdot \|X_{\mathrm{FN}} X_{\mathrm{FN}}^\top\|_2 \|\mathbf{w}^{(t)} - \mathbf{w}^*\|_2^2 + \|\boldsymbol{\xi}_{\mathrm{FN}}\|_2^2 \Big)$$

In the above, (ii) follows from Lemma 36, (iii) follows from the optimality of the Hard-Thresholding operator, (iv) follows from noting σ is Lipschitz and the elementary inequality $(a+b)^2 \leq 2a^2 + 2b^2$ for any $a, b \in \mathbb{R}$. Concluding the step, we have,

$$\|\mathbf{w}^{(t+1)} - \mathbf{w}^*\|_2 \leq \|\mathbf{w}^{(t)} - \mathbf{w}^*\|_2 \left(1 - \frac{\eta}{2(1-\epsilon)N} \cdot \gamma^2 \lambda_{\min}(X_{\mathrm{TP}} X_{\mathrm{TP}}^\top) + \frac{\sqrt{8}\eta}{(1-\epsilon)N} \cdot \|\sigma\|_{\mathrm{lip}}^2 \|X_{\mathrm{FP}}\|_2 \|X_{\mathrm{FN}}\|_2 \right)$$

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

Step 2: Concentration Bounds. From Lemma 29, we have with probability at least $1-\delta$,

$$\|\sum_{i\in\mathrm{TP}}\xi_i\sigma'(\mathbf{w}^{(t)}\cdot\mathbf{x}_i)\cdot\mathbf{x}_i\|_2\lesssim\nu\|\Gamma\|_2\|\sigma\|_{\mathrm{lip}}N\sqrt{\epsilon\log(1/\epsilon)}$$

Recall that $\|X_{\mathrm{FP}}\|_2\leq\sqrt{\epsilon NB}$. From Lemma 21, with probability at least $1-\delta$, we have

$$\|X_{\mathrm{FN}}\|_2\|X_{\mathrm{FP}}\|_2\leq N\epsilon\cdot\sqrt{10B\log(1/\epsilon)}$$

From the second relation in Lemma 21, we have when $N=\Omega\Big(\frac{d+\log(1/\delta)}{\epsilon}\Big)$, with probability exceeding $1-\delta$, the minimimum eigenvalue satisfies,

$$\lambda_{\min}(X_{\mathrm{TP}}X_{\mathrm{TP}}^{\top})\geq\frac{N}{4}\lambda_{\min}(\Sigma)$$

We then find for

$$\epsilon^2\leq\frac{\gamma^2}{\|\sigma\|_{\mathrm{lip}}^2}\frac{\lambda_{\min}(\Sigma)}{\sqrt{B\lambda_{\max}(\Sigma)}}$$

and probability exceeding $1-\delta$,

$$(\gamma^2/2)\lambda_{\min}(X_{\mathrm{TP}}X_{\mathrm{TP}}^{\top})-\sqrt{8}\|\sigma\|_{\mathrm{lip}}^2\|X_{\mathrm{FP}}\|_2\|X_{\mathrm{FN}}\|_2\geq\frac{N\gamma^2}{16}\lambda_{\min}(\Sigma)$$

Combining the ℓ_2 boundedness of the corrupted covariates and Proposition 19, we have with probability exceeding $1-\delta$,

$$\|X_{\mathrm{FP}}\|_2\|\boldsymbol{\xi}_{\mathrm{FN}}\|_2\leq N\epsilon\nu\cdot\sqrt{30B\log(1/\epsilon)}$$

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

$$\|\mathbf{w}^{(t+1)}-\mathbf{w}^*\|_2\leq\|\mathbf{w}^{(t)}-\mathbf{w}^*\|_2\left(1-\frac{\eta}{16(1-\epsilon)N}\cdot\gamma^2\lambda_{\min}(\Sigma)\right)+\eta\epsilon\|\sigma\|_{\mathrm{lip}}^2\sqrt{80B\log(1/\epsilon)}$$

$$+\left(\sqrt{8}\eta+2\sqrt{\eta\lambda_{\min}(\Sigma)}\right)C_{\Sigma}\|\sigma\|_{\mathrm{lip}}^2C_{\nu}\left(\frac{2d}{N}\log(12)+\frac{2Rd}{N}\log(12)+\frac{2}{N}\log(1/\delta)+6\epsilon\log(1/\epsilon)\right)^{1/2}$$

In the above, the final inequality holds when $N\geq\frac{Rd\log 12+2d\log 12+\log(1/\delta)}{6\epsilon\log(1/\epsilon)}=\Omega\Big(\frac{d+\log(1/\delta)}{\epsilon}\Big)$. Then, following from our sketch in Section 3.3, we have

$$\|\mathbf{w}^{(t+1)}-\mathbf{w}^*\|_2\leq\varepsilon+O\Big(\gamma^{-2}\lambda_{\min}^{-1}(\Sigma)\|\sigma\|_{\mathrm{lip}}^2\epsilon\sqrt{B\log(1/\epsilon)}\Big)+O\Big(\gamma^{-2}\|\sigma\|_{\mathrm{lip}}^2\kappa(\Sigma)\sqrt{\epsilon\log(1/\epsilon)}\Big)$$

Our proof is complete. ■

B.2 Leaky-ReLU Neuron

In this section we will derive our ℓ_2-approximation bound for the Leaky-ReLU neuron.

B.2.1 Proof of Theorem 13

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 2,

$$\|\mathbf{w}^{(t+1)}-\mathbf{w}^*\|_2=\|\mathbf{w}^{(t)}-\mathbf{w}^*-\eta\nabla\mathcal{R}(\mathbf{w}^{(t)};\mathrm{S}^{(t)})\|_2$$

$$=\|\mathbf{w}^{(t)}-\mathbf{w}^*-\eta\nabla\mathcal{R}(\mathbf{w}^{(t)};\mathrm{TP})+\eta\nabla\mathcal{R}(\mathbf{w}^*;\mathrm{TP})-\eta\nabla\mathcal{R}(\mathbf{w}^*;\mathrm{TP})-\nabla\mathcal{R}(\mathbf{w}^{(t)};\mathrm{FP})\|_2$$

$$\leq\underbrace{\|\mathbf{w}^{(t)}-\mathbf{w}^*-\eta\nabla\mathcal{R}(\mathbf{w}^{(t)};\mathrm{TP})+\eta\nabla\mathcal{R}(\mathbf{w}^*;\mathrm{TP})\|_2}_{I}+\underbrace{\|\eta\nabla\mathcal{R}(\mathbf{w}^*;\mathrm{TP})\|_2}_{II}+\underbrace{\|\eta\nabla\mathcal{R}(\mathbf{w}^{(t)};\mathrm{FP})\|_2}_{III}$$

We will first upper bound the I_1 by an expansion of its square.

$$I^2 = \|\mathbf{w}^{(t)} - \mathbf{w}^*\|_2^2 - \underbrace{2\eta \cdot \langle \mathbf{w}^{(t)} - \mathbf{w}^*, \nabla\mathcal{R}(\mathbf{w}^{(t)}; \mathrm{TP})\rangle - \nabla\mathcal{R}(\mathbf{w}^{(t)}; \mathrm{TP})}_{I_1}$$
$$+ \underbrace{\eta^2 \cdot \|\nabla\mathcal{R}(\mathbf{w}^{(t)}; \mathrm{TP}) - \nabla\mathcal{R}(\mathbf{w}^*; \mathrm{TP})\|^2}_{I_2}$$

In the above, the a.s. relation follows from Property 9. We will first lower bound I_1. We will first bound the spectrum of $\nabla^2\mathcal{L}(\mathbf{w}; \mathrm{TP})$ for any $\mathbf{w} \in \mathbb{R}^d$.

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

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

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

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

$$2 \cdot \gamma^2 \lambda_{\min}(X_{\mathrm{TP}}X_{\mathrm{TP}}^\top) \cdot I \preceq \nabla^2\mathcal{L}(\mathbf{w}; \mathrm{TP}) \preceq 2 \cdot \|\sigma\|_{\mathrm{lip}}^2 \lambda_{\max}(X_{\mathrm{TP}}X_{\mathrm{TP}}^\top) \cdot I \tag{5}$$

We can now lower bound I_1 from the convexity of the Leaky-ReLU,

$$I_1 = \frac{4\eta}{(1-\epsilon)N} \cdot \langle \mathbf{w}^{(t)} - \mathbf{w}^*, \int_0^1 \nabla^2\mathcal{R}(\mathbf{w}^* + \theta(\mathbf{w}^* - \mathbf{w}^{(t)}); \mathrm{TP}) d\theta \cdot (\mathbf{w}^{(t)} - \mathbf{w}^*)\rangle$$
$$\overset{(5)}{\geq} \frac{4\eta}{(1-\epsilon)N} \cdot \gamma^2 \lambda_{\min}(X_{\mathrm{TP}}X_{\mathrm{TP}}^\top) \|\mathbf{w}^{(t)} - \mathbf{w}^*\|_2^2$$

We now will upper bound I_2 with a similar argument.

$$I_2 = \frac{4\eta^2}{[(1-\epsilon)N]^2} \cdot \Big\| \int_0^1 \nabla^2\mathcal{R}(\mathbf{w}^* + \theta(\mathbf{w}^* - \mathbf{w}^{(t)}); \mathrm{TP}) d\theta \cdot (\mathbf{w}^{(t)} - \mathbf{w}^*) \Big\|_2^2$$
$$\overset{(5)}{\leq} \frac{4\eta^2}{[(1-\epsilon)N]^2} \cdot \|\sigma\|_{\mathrm{lip}}^2 \lambda_{\max}^2(X_{\mathrm{TP}}X_{\mathrm{TP}}^\top) \|\mathbf{w}^{(t)} - \mathbf{w}^*\|_2^2$$

where (ii) follows from the $\|\sigma\|_{\mathrm{lip}}$-Lipschitzness of σ given in Property 10. We then observe that $I_2 \leq 0.5 I_1$ when we choose

$$\eta \leq \frac{\gamma^2}{\|\sigma\|_{\mathrm{lip}}^2} \frac{N\lambda_{\min}(X_{\mathrm{TP}}X_{\mathrm{TP}}^\top)}{2\lambda_{\max}^2(X_{\mathrm{TP}}X_{\mathrm{TP}}^\top)}$$

Step 2: Upper bounding the corrupted gradient. We now upper bound the corrupted gradient term.

$$III^2 \overset{\text{def}}{=} \frac{4\eta^2}{[(1-\epsilon)N]^2} \cdot \Big\| \sum_{i \in \mathrm{FP}} (\sigma(\mathbf{w} \cdot \mathbf{x}_i) - y_i) \cdot \sigma'(\mathbf{w}^{(t)} \cdot \mathbf{x}_i) \cdot \mathbf{x}_i \Big\|_2^2$$
$$\leq \frac{4\eta^2}{[(1-\epsilon)N]^2} \cdot \|\sigma\|_{\mathrm{lip}}^2 \|X_{\mathrm{FP}}X_{\mathrm{FP}}^\top\|_2 \sum_{i \in \mathrm{FP}} \big(\sigma(\mathbf{w} \cdot \mathbf{x}_i) - \sigma(\mathbf{w}^* \cdot \mathbf{x}_i)\big)^2$$
$$\leq \frac{4\eta^2}{[(1-\epsilon)N]^2} \cdot \|\sigma\|_{\mathrm{lip}}^2 \|X_{\mathrm{FP}}X_{\mathrm{FP}}^\top\|_2 \sum_{i \in \mathrm{FN}} \big(\sigma(\mathbf{w} \cdot \mathbf{x}_i) - \sigma(\mathbf{w}^* \cdot \mathbf{x}_i) + \xi_i\big)^2$$
$$\leq \frac{8\eta^2}{[(1-\epsilon)N]^2} \cdot \|\sigma\|_{\mathrm{lip}}^2 \|X_{\mathrm{FP}}X_{\mathrm{FP}}^\top\|_2 \Big(\|\sigma\|_{\mathrm{lip}}^2 \|X_{\mathrm{FN}}X_{\mathrm{FN}}^\top\|_2 \|\mathbf{w}^{(t)} - \mathbf{w}^*\|_2^2 + \|\boldsymbol{\xi}_{\mathrm{FN}}\|_2^2 \Big)$$

In the above, the first inequality follows from Lemma 36, the second inequality follows from the optimality of the Subquantile set, the final inequality follows from the $\|\sigma\|_{\mathrm{lip}}$-Lipschitzness of σ. Then from Lemma 30, we have with probability at least $1 - \delta$,

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

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 for $N = \Omega\Big(\frac{Rd+\log(1/\delta)}{\epsilon}\Big)$,

$$\|\mathbf{w}^{(t+1)}-\mathbf{w}^*\|_2 \leq \|\mathbf{w}^{(t)}-\mathbf{w}^*\|_2\Bigg(1-\frac{2\gamma^2\eta}{(1-\epsilon)N}\cdot\lambda_{\min}(X_{\mathrm{TP}}X_{\mathrm{TP}}^\top)+\frac{2\eta}{(1-\epsilon)N}\cdot\|X_{\mathrm{FP}}\|_2\|X_{\mathrm{FN}}\|_2\Bigg)$$
$$+\frac{15C_\Sigma C_\nu d\log 5}{\sqrt{(1-\epsilon)N}}+\frac{15C_\Sigma C_\nu\log(1/\delta)}{\sqrt{(1-\epsilon)N}}+16C_\Sigma C_\nu\sqrt{\epsilon\log(1/\epsilon)}+\frac{\sqrt{8}\eta}{(1-\epsilon)N}\cdot\|X_{\mathrm{FP}}\|_2\|\boldsymbol{\xi}_{\mathrm{FN}}\|_2$$

Step 3: Concentration Bounds. We will give the relevant probabilistic bounds for the random variables in Steps 1 and 2. From Lemma 21, we have

$$\|X_{\mathrm{FN}}\|_2\|X_{\mathrm{FP}}\|_2 \leq \epsilon\sqrt{\lambda_{\max}(\Sigma)\cdot 10BN\log(1/\epsilon)}$$

with probability at least $1-\delta$ when

$$N \geq \frac{2}{\epsilon}\cdot\left(dC_K^2+\frac{\log(2/\delta)}{c_K}\right) \text{ and } \epsilon \leq \frac{1}{60}\cdot\kappa^{-1}(\Sigma)$$

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 when the corruption rate satisfies

$$\epsilon \leq \frac{\gamma^2\lambda_{\min}(\Sigma)}{\sqrt{32B\lambda_{\max}(\Sigma)}}$$

We have

$$\|X_{\mathrm{FP}}\|_2\|X_{\mathrm{FN}}\|_2-\gamma^2\lambda_{\min}(X_{\mathrm{TP}}X_{\mathrm{TP}}^\top) \geq \frac{1}{8}\cdot\lambda_{\min}(\Sigma)$$

We then have, after $O\Big(\kappa^2(\Sigma)\log\Big(\frac{\|\mathbf{w}^*\|_2}{\varepsilon}\Big)\Big)$ iterations with high probability,

$$\|\mathbf{w}^{(T)}-\mathbf{w}^*\|_2 \leq \varepsilon+O\Big(C_\Sigma C_\nu\sqrt{\epsilon\log(1/\epsilon)}\Big)+O\Big(\gamma^{-2}\kappa(\Sigma)\nu\epsilon\sqrt{B\log(1/\epsilon)}\Big)$$

In the final inequality above, we set $\varepsilon = O\Big(C_\Sigma\nu\sqrt{\epsilon\log(1/\epsilon)}\Big)$ for

$$N = \Omega\left(\frac{d+\log(1/\delta)}{\epsilon}\right)$$

Our proof is complete. ■

B.3 ReLU Neuron

In this section, we consider ReLU type functions. 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 15

Proof. We will now begin our standard analysis.

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

We will now upper bound I through its square in accordance with our proof sketch,

$$I^2 = \|\mathbf{w}^{(t)}-\mathbf{w}^*\|_2^2-\underbrace{2\eta\cdot\langle\mathbf{w}^{(t)}-\mathbf{w}^*,\nabla\mathcal{R}(\mathbf{w}^{(t)};\mathrm{TP})\rangle}_{I_1}+\underbrace{\eta^2\cdot\|\nabla\mathcal{R}(\mathbf{w}^{(t)};\mathrm{TP})\|_2^2}_{I_2}$$

We will first lower bound I_1. We will first adopt the notation from Zhang et al. [2019], let $\Sigma_{\mathrm{TP}}(\mathbf{w}, \hat{\mathbf{w}}) = X_{\mathrm{TP}} X_{\mathrm{TP}}^\top \cdot \mathbf{1}\{X_{\mathrm{TP}}^\top \mathbf{w} \geq \mathbf{0}\} \cdot \mathbf{1}\{X_{\mathrm{TP}}^\top \hat{\mathbf{w}} \geq \mathbf{0}\}$, it then follows

$$
\begin{aligned}
I_1 &\stackrel{\text{def}}{=} \frac{4\eta}{(1-\epsilon)N} \cdot \langle \mathbf{w}^{(t)} - \mathbf{w}^*, \sum_{i \in \mathrm{TP}} (\sigma(\mathbf{w}^{(t)} \cdot \mathbf{x}_i) - y_i) \cdot \mathbf{x}_i \cdot \mathbf{1}\{\mathbf{w}^{(t)} \cdot \mathbf{x}_i \geq 0\} \rangle \\
&= \frac{4\eta}{(1-\epsilon)N} \cdot \langle \mathbf{w}^{(t)} - \mathbf{w}^*, \Sigma_{\mathrm{TP}}(\mathbf{w}^{(t)}, \mathbf{w}^{(t)}) \mathbf{w}^{(t)} - \Sigma_{\mathrm{TP}}(\mathbf{w}^{(t)}, \mathbf{w}^*) \mathbf{w}^* \rangle \\
&\quad - \frac{4\eta}{(1-\epsilon)N} \cdot \langle \mathbf{w}^{(t)} - \mathbf{w}^*, \sum_{i \in \mathrm{TP}} \xi_i \mathbf{x}_i \cdot \mathbf{1}\{\mathbf{w}^{(t)} \cdot \mathbf{x}_i \geq 0\} \rangle \\
&\geq \frac{4\eta}{(1-\epsilon)N} \cdot \langle \mathbf{w}^{(t)} - \mathbf{w}^*, \Sigma_{\mathrm{TP}}(\mathbf{w}^{(t)}, \mathbf{w}^*)(\mathbf{w}^{(t)} - \mathbf{w}^*) + \Sigma_{\mathrm{TP}}(\mathbf{w}^{(t)}, -\mathbf{w}^*) \mathbf{w}^{(t)} \rangle \\
&\quad - \frac{4\eta}{(1-\epsilon)N} \cdot \|\mathbf{w}^{(t)} - \mathbf{w}^*\|_2 \Big\| \sum_{i \in \mathrm{TP}} \xi_i \mathbf{x}_i \cdot \mathbf{1}\{\mathbf{w}^{(t)} \cdot \mathbf{x}_i \geq 0\} \Big\|_2 \\
&\stackrel{(i)}{\geq} \frac{4\eta}{(1-\epsilon)N} \cdot \lambda_{\min}(\Sigma_{\mathrm{TP}}(\mathbf{w}^{(t)}, \mathbf{w}^*)) \|\mathbf{w}^{(t)} - \mathbf{w}^*\|_2^2 \\
&\quad - \frac{4\eta}{(1-\epsilon)N} \cdot \|\mathbf{w}^{(t)} - \mathbf{w}^*\|_2 \Big\| \sum_{i \in \mathrm{TP}} \xi_i \mathbf{x}_i \cdot \mathbf{1}\{\mathbf{w}^{(t)} \cdot \mathbf{x}_i \geq 0\} \Big\|_2 \\
&\stackrel{(ii)}{\geq} \frac{2\eta}{(1-\epsilon)N} \cdot \lambda_{\min}(\Sigma_{\mathrm{TP}}(\mathbf{w}^{(t)}, \mathbf{w}^*)) \|\mathbf{w}^{(t)} - \mathbf{w}^*\|_2^2 \\
&\quad - \frac{2\eta}{(1-\epsilon)N} \cdot \lambda_{\min}^{-1}(\Sigma_{\mathrm{TP}}(\mathbf{w}^{(t)}, \mathbf{w}^*)) \cdot \Big\| \sum_{i \in \mathrm{TP}} \xi_i \mathbf{x}_i \cdot \mathbf{1}\{\mathbf{w}^{(t)} \cdot \mathbf{x}_i \geq 0\} \Big\|_2^2
\end{aligned}
$$

In the above, (ii) follows from Young's Inequality and (i) holds from the following relation,

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

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

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

In the above, (i) follows from Lemma 32, (ii) follows from the guarantee in the randomized initialization. We then have from Lemma 31,

$$
\|\Sigma_{\mathrm{TP}}(\mathbf{w}^{(t)}, \mathbf{w}^*) - \mathbf{E}\Big[\Sigma_{\mathrm{TP}}(\mathbf{w}^{(t)}, \mathbf{w}^*)\Big]\|_2 \lesssim N\|\Gamma\|_2^2 \sqrt{\epsilon \log(1/\epsilon)}.
$$

We then find there exists sufficiently small ϵ such that the minimimum eigenvalue of $\Sigma_{\mathrm{TP}}(\mathbf{w}^{(t)}, \mathbf{w}^*)$ satisfies

$$\lambda_{\min}(\Sigma_{\mathrm{TP}}(\mathbf{w}^{(t)}, \mathbf{w}^*)) \geq \frac{\lambda_{\min}(\Sigma)}{4}.$$

We now bound the second-moment matrix approximation. Let $dP(\mathbf{x})$ be a Dirac-measure for $\mathbf{x} \in \mathrm{TP}$. We now upper bound I_2 by splitting it into two sperate terms,

$$\begin{aligned}
I_2 &= \frac{4\eta^2}{\left[(1-\epsilon)N\right]^2} \cdot \Big\| \sum_{i \in \mathrm{TP}} (\sigma(\mathbf{w}^{(t)} \cdot \mathbf{x}_i) - \sigma(\mathbf{w}^* \cdot \mathbf{x}_i) - \xi_i) \cdot \mathbf{x}_i \cdot \mathbf{1}\{\mathbf{w}^{(t)} \cdot \mathbf{x}_i \geq 0\} \Big\|_2^2 \\
&\leq \underbrace{\frac{8\eta^2}{\left[(1-\epsilon)N\right]^2} \cdot \Big\| \sum_{i \in \mathrm{TP}} (\sigma(\mathbf{w}^{(t)} \cdot \mathbf{x}_i) - \sigma(\mathbf{w}^* \cdot \mathbf{x}_i)) \cdot \mathbf{x}_i \cdot \mathbf{1}\{\mathbf{w}^{(t)} \cdot \mathbf{x}_i \geq 0\} \Big\|_2^2}_{I_{21}} \\
&\quad + \underbrace{\frac{8\eta^2}{\left[(1-\epsilon)N\right]^2} \cdot \Big\| \sum_{i \in \mathrm{TP}} \xi_i \mathbf{x}_i \cdot \mathbf{1}\{\mathbf{w}^{(t)} \cdot \mathbf{x}_i \geq 0\} \Big\|_2^2}_{I_{22}}
\end{aligned}$$

Recall from Lemma 30, we have an upper bound on I_{22}. We next upper bound I_{21}.

$$\begin{aligned}
I_{22} &= \frac{8\eta^2}{\left[(1-\epsilon)N\right]^2} \cdot \Big\| \sum_{i \in \mathrm{TP}} (\sigma(\mathbf{w}^{(t)} \cdot \mathbf{x}_i) - \sigma(\mathbf{w}^* \cdot \mathbf{x}_i)) \cdot \mathbf{x}_i \cdot \mathbf{1}\{\mathbf{w}^{(t)} \cdot \mathbf{x}_i \geq 0\} \Big\|_2^2 \\
&\leq \frac{8\eta^2}{\left[(1-\epsilon)N\right]^2} \cdot \big\| X_{\mathrm{TP}} X_{\mathrm{TP}}^\top \big\|_2 \sum_{i \in \mathrm{TP}} (\sigma(\mathbf{w}^{(t)} \cdot \mathbf{x}_i) - \sigma(\mathbf{w}^* \cdot \mathbf{x}_i))^2 \\
&\leq \frac{8\eta^2}{\left[(1-\epsilon)N\right]^2} \cdot \big\| X_{\mathrm{TP}} X_{\mathrm{TP}}^\top \big\|_2^2 \|\mathbf{w}^{(t)} - \mathbf{w}^*\|_2^2
\end{aligned}$$

Then, by choosing $\eta \leq \frac{\lambda_{\min}(\Sigma)}{80\lambda_{\max}^2(\Sigma)}$, we have that $I_{22} \leq \frac{\lambda_{\min}(\Sigma)}{8}$.

Step 3: Upper bounding the corrupted gradient. We now upper bound II.

$$\begin{aligned}
II &= \frac{4\eta^2}{\left[(1-\epsilon)N\right]^2} \cdot \Big\| \sum_{i \in \mathrm{FP}} (\sigma(\mathbf{w}^{(t)} \cdot \mathbf{x}_i) - y_i) \cdot \mathbf{x}_i \cdot \mathbf{1}\{\mathbf{w}^{(t)} \cdot \mathbf{x}_i \geq 0\} \Big\|_2^2 \\
&\leq \frac{4\eta^2}{\left[(1-\epsilon)N\right]^2} \cdot \|\Sigma_{\mathrm{FP}}(\mathbf{w}^{(t)}, \mathbf{w}^{(t)})\|_2 \sum_{i \in \mathrm{FP}} (\sigma(\mathbf{w}^{(t)} \cdot \mathbf{x}_i) - \sigma(+)\xi_i)^2 \\
&\leq \frac{4\eta^2}{\left[(1-\epsilon)N\right]^2} \cdot \|X_{\mathrm{FP}} X_{\mathrm{FP}}^\top\|_2 \sum_{i \in \mathrm{FN}} (\sigma(\mathbf{w}^{(t)} \cdot \mathbf{x}_i) - \sigma(\mathbf{w}^* \cdot \mathbf{x}_i) + \xi_i)^2 \\
&\leq \frac{8\eta^2}{\left[(1-\epsilon)N\right]^2} \cdot \|X_{\mathrm{FP}} X_{\mathrm{FP}}^\top\|_2 \Big(\|X_{\mathrm{FN}} X_{\mathrm{FN}}^\top\|_2 \|\mathbf{w}^{(t)} - \mathbf{w}^*\|_2^2 + \|\boldsymbol{\xi}_{\mathrm{FN}}\|_2^2 \Big)
\end{aligned}$$

In the above, the first inequality follows from the same argument as Lemma 36, 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.

$$\begin{aligned}
\|\mathbf{w}^{(t+1)} - \mathbf{w}^*\|_2 &\lesssim \|\mathbf{w}^{(t)} - \mathbf{w}^*\| \left(1 - \frac{\eta}{32} \cdot \lambda_{\min}(\Sigma) + \frac{4\eta}{(1-\epsilon)N} \cdot \|X_{\mathrm{FP}}\|_2 \|X_{\mathrm{FN}}\|_2 \right) \\
&\quad + \left(\frac{8\eta}{N} + \frac{4\eta}{N} \cdot \sqrt{\lambda_{\min}^{-1}(\Sigma)} \right) \cdot \Big\| \sum_{i \in \mathrm{TP}} \xi_i \mathbf{x}_i \cdot \mathbf{1}\{\mathbf{w}^{(t)} \cdot \mathbf{x}_i \geq 0\} \Big\|_2 + \frac{\eta}{(1-\epsilon)N} \|X_{\mathrm{FP}}\|_2 \|\boldsymbol{\xi}_{\mathrm{FN}}\|_2
\end{aligned}$$

Step 4: Concentration Inequalities. From our previous theorems, we have

$$\|X_{\mathrm{FP}}\|_2 \|X_{\mathrm{FN}}\|_2 \leq N\epsilon\sqrt{10B\log(1/\epsilon)}$$

with probability exceeding $1-\delta$ and $N = \Omega\Big(\frac{d+\log(2/\delta)}{\epsilon}\Big)$. From Lemma 30, we have with probability exceeding $1-\delta$,

$$\Big\| \sum_{i \in \mathrm{TP}} \xi_i \mathbf{x}_i \cdot \mathbf{1}\{\mathbf{w}^{(t)} \cdot \mathbf{x}_i \geq 0\} \Big\|_2 \lesssim \nu \|\Gamma\|_2 \sqrt{\epsilon \log(1/\epsilon)}.$$

We furthermore have with failure probability less than δ from Lemma 18,

$$\|\boldsymbol{\xi}_{\mathrm{FN}}\|_2 \leq \nu\sqrt{30\epsilon\log(1/\epsilon)}$$

If $\epsilon \leq \frac{1}{64\sqrt{80\log(2)}}$, we obtain after $T = O\Big(\kappa^2(\Sigma)\log\Big(\frac{\|\mathbf{w}^*\|_2}{\varepsilon}\Big)\Big)$ gradient descent iterations,

$$\begin{aligned}\|\mathbf{w}^{(t+1)} - \mathbf{w}^*\|_2 &\lesssim \varepsilon + \nu\|\Gamma\|_2\lambda_{\min}^{-1}(\Sigma)\sqrt{\epsilon\log(1/\epsilon)} + \nu\epsilon\|\Gamma\|_2\sqrt{B\log(1/\epsilon)} \\ &= O\Big(\nu\|\Gamma\|_2\lambda_{\min}^{-1}(\Sigma)\sqrt{\epsilon\log(1/\epsilon)}\Big) + O\Big(\nu\epsilon\|\Gamma\|_2\sqrt{B\log(1/\epsilon)}\Big)\end{aligned}$$

when we choose $\varepsilon = O\Big(\nu\|\Gamma\|_2\lambda_{\min}^{-1}(\Sigma)\sqrt{\epsilon\log(1/\epsilon)}\Big) + O\Big(\nu\epsilon\|\Gamma\|_2\sqrt{B\log(1/\epsilon)}\Big)$. Our proof is complete. ■

C Probability Theory

In this section, we will present and prove various concentration inequalities and upper bounds for random variables.

C.1 χ-Squared Random Variables

Lemma 18 (Upper Bound on Sum of Chi-Squared Variables [Laurent and Massart, 2000])**.** *Suppose $\xi_i \sim \mathcal{N}(0,\sigma^2)$ for $i \in [n]$, then*

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

Proposition 19 (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 $\mathcal{W}$ represent all such subsets. Given a failure probability $\delta \in (0,1)$, when $n \geq \log(1/\delta)$, with probability exceeding $1-\delta$,*

$$\max_{\mathrm{S}\in\mathcal{W}}\|\boldsymbol{\xi}_{\mathrm{S}}\|_2^2 \leq \sigma(30n\epsilon\log(1/\epsilon))$$

Proof. Directly from Lemma 18, we have with probability exceeding $1-\delta$.

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

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

$$\mathbf{Pr}\Big\{\max_{\mathrm{S}\in\mathcal{W}}\|\boldsymbol{\xi}\|_2^2 \geq \sigma\big(\epsilon n + 2\sqrt{\epsilon nx} + 2x\big)\Big\} \leq \binom{\mathrm{e}}{\epsilon}^{\epsilon n}\mathbf{Pr}\big\{\|\boldsymbol{\xi}\|_2^2 \geq \sigma\big(\epsilon n + 2\sqrt{\epsilon nx} + 2x\big)\big\} \leq \binom{\mathrm{e}}{\epsilon}^{\epsilon n}\mathrm{e}^{-x}$$

In the first inequality we apply a union bound over $\mathcal{W}$ with Lemma 38, and in the second inequality we use Lemma 18. We then obtain with probability exceeding $1-\delta$,

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

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

C.2 Eigenvalue Concentration Inequalities

Lemma 20 (Sub-Gaussian Covariance Matrix Estimation Vershynin [2010]). *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 $c_k, \|\Gamma\|_2$, dependent on the sub-Gaussian norm such that with probability at least $1 - 2\mathrm{e}^{-c_K t^2}$,*

$$\lambda_{\max}(XX^\top) \leq n \cdot \lambda_{\max}(\Sigma) + \lambda_{\max}(\Sigma) \cdot \left(\|\Gamma\|_2 \cdot \sqrt{dn} + t \cdot \sqrt{n}\right)$$

Lemma 21. *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 $\mathcal{W}$ represent all such subsets. Then with probability at least $1 - \delta$,*

$$\max_{\mathrm{S}\in\mathcal{W}} \lambda_{\max}(X_{\mathrm{S}} X_{\mathrm{S}}^\top) \leq \lambda_{\max}(\Sigma) \cdot (10n\epsilon \log(1/\epsilon))$$
$$\min_{S\in\mathcal{W}} \lambda_{\min}(X_{[n]\setminus\mathrm{S}} X_{[n]\setminus\mathrm{S}}^\top) \geq \frac{n}{4} \cdot \lambda_{\min}(\Sigma)$$

when

$$n \geq \frac{2}{\epsilon} \cdot \left(\|\Gamma\|_2^2 \cdot d + \frac{\log(2/\delta)}{c_K}\right) \text{ and } \epsilon \leq \frac{1}{30} \cdot \kappa^{-1}(\Sigma)$$

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

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

In the above, the first inequality follows from a union bound over $\mathcal{W}$ and Lemma 38, the second inequality follows from Lemma 20. Then from elementary inequalities, we obtain with probability $1 - \delta$,

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

In the above, the last inequality holds when

$$n \geq \frac{2}{\epsilon} \cdot \left(\|\Gamma\|_2^2 \cdot d + \frac{\log(2/\delta)}{c_K}\right)$$

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

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

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

$$\begin{aligned}
\lambda_{\min}(X_{\mathrm{X}\setminus\mathrm{S}} X_{\mathrm{X}\setminus\mathrm{S}}^\top) &\geq n \cdot \lambda_{\min}(\Sigma) - \|\Gamma\|_2 \cdot \sqrt{dn} - \sqrt{\frac{1}{c_K} \cdot n \cdot \log(2/\delta)} - \lambda_{\max}(\Sigma) \cdot (10n\epsilon \log(1/\epsilon)) \\
&\geq \frac{3n}{4} \cdot \lambda_{\min}(\Sigma) - \lambda_{\max}(\Sigma) \cdot (10n\epsilon \log(1/\epsilon)) \geq \frac{n}{4} \cdot \lambda_{\min}(\Sigma)
\end{aligned}$$

In the above, the first inequality follows when $n \geq \frac{32}{\lambda_{\min}^2(\Sigma)}\left(\|\Gamma\|_2 \cdot d + \frac{1}{c_K} \cdot \log(2/\delta)\right)$, and from some algebra, we find the last inequality holds when $\epsilon \leq \frac{1}{30} \cdot \kappa^{-1}(\Sigma)$ by noting that $\epsilon < 0.5$. The proof is complete. ■

C.3 Sum of product of nonlinear random variables

We will first define the Orlicz Norm for sub-Gaussian and sub-exponential random variables.

Definition 22. *The sub-Gaussian norm of a random variable X is denoted as $\|X\|_{\psi_2}$, and is defined as*

$$\|X\|_{\psi_2} = \inf\{t > 0 : \mathbf{E}[\exp(X^2/t^2)] \leq 2\}.$$

Definition 23. *The sub-exponential norm of a random variable X is denoted as $\|X\|_{\psi_1}$, and is defined as*

$$\|X\|_{\psi_1} = \inf\{t > 0 : \mathbf{E}[\exp(|X|/t)] \leq 2\}.$$

In this section, we will work with the products of sub-Gaussian random variables.

Lemma 24 (Lemma 2.7.7 in Vershynin [2020])**.** *Let X, Y be sub-Gaussian random variables, then XY is sub-exponential, furthermore,*

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

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

Lemma 25 (Proposition 5.16 in Vershynin [2010])**.** *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 $t \geq 0$,*

$$\mathbf{Pr}\Big\{\big|\sum_{i\in[N]} a_i X_i\big| \geq t\Big\} \leq 2\exp\Big[-c\min\Big(\frac{t^2}{K^2\|\mathbf{a}\|_2^2}, \frac{t}{K\|\mathbf{a}\|_\infty}\Big)\Big]$$

Lemma 26. *Let $X = [\mathbf{x}_1, \ldots, \mathbf{x}_N]$ 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 proxy Γ and ξ_i are sampled from $\mathcal{N}(0, \nu^2)$. Assume $f : \mathbb{R}^d \mapsto \mathbb{R}$ is bounded over $\mathbb{R}$. Let $\mathcal{W}$ represent all subsets of $[N]$ of size empty to $(1-\epsilon)N$ and $\mathcal{N}_1$ be a ε-cover of $\mathcal{S}^{d-1}$, then for any $\mathbf{u} \in \mathcal{S}^{d-1}$ and $\mathrm{S} \in \mathcal{W}$. With probability at least $1-\delta$,*

$$\big\|\sum_{i\in\mathrm{S}} \xi_i f(\mathbf{x}_i \cdot \mathbf{u})\mathbf{x}_i\big\| \lesssim \nu\|\Gamma\|_2\|f\|_\infty N\sqrt{\epsilon\log(1/\epsilon)},$$

where c is an absolute constant and the sample size satisfies

$$N = \Omega\Big(\frac{Rd + \log(1/\delta)}{\epsilon}\Big).$$

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

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

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, $\mathbf{x}_i \cdot \mathbf{v}$ is sub-Gaussian by definition. We then have,

$$\Big(\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}}|f(\mathbf{w}\cdot\mathbf{x})\mathbf{x}\cdot\mathbf{v}|^p\Big)^{1/p} \overset{(i)}{\leq} \Big(\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}}|f(\mathbf{w}\cdot\mathbf{x})|^{2p}\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}}|\mathbf{x}\cdot\mathbf{v}|^{2p}\Big)^{1/2p} \overset{(ii)}{\leq} \Big(\|f\|_\infty\|\Gamma\|_2\sqrt{2}\Big)\sqrt{p}.$$

In the above, (i) follows from Hölder's Inequality, (ii) follows from noting from letting $q = 2p$ and noting from Definition 4 that $\|\mathbf{x}_i\cdot\mathbf{v}\|_{L_q}$ is upper bounded by $\|\Gamma\|_2\sqrt{q}$. We thus have $f(\mathbf{w}\cdot\mathbf{x}_i)\mathbf{x}_i\cdot\mathbf{v}$ is sub-Gaussian for any $\mathbf{w} \in \mathbb{R}^d$ and $\|f(\mathbf{w}\cdot\mathbf{x}_i)\mathbf{x}_i\cdot\mathbf{v}\|_{\psi_2} \lesssim \sqrt{2}\|\Gamma\|_2\|f\|_\infty$. We have $\|\xi_i\|_{\psi_2} = \sqrt{8/3}\nu$ from Lemma 35, then from **??**, the random variable $\xi_i f(\mathbf{w}\cdot\mathbf{x}_i)\mathbf{x}_i\cdot\mathbf{v}$ is sub-exponential s.t. $\|\xi_i f(\mathbf{w}\cdot\mathbf{x}_i)\mathbf{x}_i\cdot\mathbf{v}\|_{\psi_1} \lesssim \sqrt{16/3}\|\Gamma\|_2\|f\|_\infty\nu$. 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 $\mathcal{B}(\mathbf{0}, R)$. 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,

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

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 obtain

$$|(\boldsymbol{\xi} \circ f(X^\top \widetilde{\mathbf{w}}))^\top X\mathbf{v}^*| \leq \frac{1}{1-\varepsilon} \cdot |(\boldsymbol{\xi} \circ f(X^\top \widetilde{\mathbf{w}}))^\top X\mathbf{u}^*|. \tag{6}$$

With this result we are ready to make the probabilistic bounds. Suppose $\mathcal{W}$ represents all subsets of $[N]$ of size empty to $(1-\epsilon)N$. Suppose $\mathcal{N}_2$ is a $1/2$-net of $\mathbb{S}^{d-1}$ and $\mathcal{N}_1$ is a $1/2$-net of $\mathcal{B}(\mathbf{w}^*, R)$, we can then note that $\|\mathbf{w}^*\| \leq R$. Then,

$$\begin{aligned}&\mathbf{Pr}\left\{\max_{\mathrm{S}\in\mathcal{W}} \max_{\mathbf{w}\in\mathcal{N}_1} \|(\boldsymbol{\xi}_\mathrm{S} \circ f(X_\mathrm{S}^\top \mathbf{w}))^\top X_\mathrm{S}\| \geq t\right\} \\ &\overset{(6)}{\leq} \mathbf{Pr}\left\{\max_{\mathrm{S}\in\mathcal{W}} \max_{\mathbf{v}\in\mathcal{N}_2} \max_{\mathbf{w}\in\mathcal{N}_1} |(\boldsymbol{\xi}_\mathrm{S} \circ f(X_\mathrm{S}^\top \mathbf{w}))^\top X_\mathrm{S}\mathbf{v}| \geq 2t\right\} \\ &\overset{(iii)}{\leq} 2 \cdot 6^{Rd+d}\left(\frac{\mathrm{e}}{\epsilon}\right)^{N\epsilon} \exp\left[-c\min\left(\frac{t^2}{(4/3)c_1^2\nu^2\|\Gamma\|_2^2\|f\|_\infty^2|\mathrm{S}|}, \frac{t}{\sqrt{4/3}c_1\nu\|\Gamma\|_2\|f\|_\infty}\right)\right] \leq \delta\end{aligned}$$

In the above, (iii) follows from a union bound over $\mathcal{W}$, $\mathcal{N}_1$, and $\mathcal{N}_2$, and then applying Bernstein's Inequality (see Lemma 25). We also note that $\log\binom{N}{(1-\epsilon)N} = \log\binom{N}{\epsilon N}$, and then applying Lemma 38 gives the inequality. Then to satisfy the above probabilistic condition, it must hold that

$$t \geq \sqrt{4/3}c^{-1}c_1\nu\|\Gamma\|_2\|f\|_\infty\big(2dN\log(6) + 2NRd\log(6) + 2N\log(2/\delta) + 6N^2\epsilon\log(1/\epsilon)\big)^{1/2}$$

Then, when the sample size satisfies

$$N \geq \frac{(Rd+d)\log(6) + \log(2/\delta)}{3\epsilon} = \Omega\left(\frac{Rd + \log(1/\delta)}{\epsilon}\right).$$

We obtain for any $\mathrm{S} \in \mathcal{W}$ and $\mathbf{u} \in \mathcal{N}_1$,

$$\Big\|\sum_{i\in\mathrm{TP}} f(\mathbf{u}\cdot\mathbf{x}_i)\xi_i\mathbf{x}_i\Big\|_2 \lesssim \nu\|\Gamma\|_2\|f\|_\infty N\sqrt{\epsilon\log(1/\epsilon)}$$

Our proof is complete. ■

Lemma 27. *Let $X = [\mathbf{x}_1, \ldots, \mathbf{x}_N]$ 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 proxy Γ and ξ_i are sampled from $\mathcal{N}(0, \nu^2)$. Assume $f : \mathbb{R}^d \to \mathbb{R}$ is bounded over $\mathbb{R}$ and Lipschitz. Let $\mathcal{W}$ represent all subsets of $[N]$ of size empty to $(1-\epsilon)N$. Suppose $\mathbf{w} \in \mathcal{B}(\mathbf{w}^*, R)$ and $\mathrm{S} \in \mathcal{W}$. Let $\mathcal{N}_1$ be a ε-net of $\mathbb{S}^{d-1}$ and define $\widetilde{\mathbf{w}} = \arg\min_{\mathbf{u}\in\mathcal{N}_1}\|\mathbf{w}-\mathbf{u}\|_2$. Set a failure probability $\delta \in (0,1)$, then with probability at least $1-\delta$,*

$$\begin{aligned}\max_{\mathrm{S}\in\mathcal{W}} \sup_{\mathbf{w}\in\mathcal{B}(\mathbf{0},R)} \Big\|\sum_{i\in\mathrm{S}} \xi_i f(\mathbf{w}\cdot\mathbf{x}_i)\mathbf{x}_i - \sum_{i\in\mathrm{S}} \xi_i f(\widetilde{\mathbf{w}}\cdot\mathbf{x}_i)\mathbf{x}_i\Big\|_2 \\ \leq \sqrt{N}\|f\|_{\mathrm{lip}}\|\Gamma\|_2\nu\big(\log^2(N/\delta) + d\log(N/\delta)\big)\end{aligned}$$

Proof. We have from the Cauchy-Schwarz inequality,

$$\begin{aligned}&\sup_{\mathbf{w}\in\mathcal{B}(\mathbf{0},R)} \Big\|\sum_{i\in\mathrm{S}} \xi_i f(\mathbf{w}\cdot\mathbf{x}_i)\mathbf{x}_i - \sum_{i\in\mathrm{S}} \xi_i f(\widetilde{\mathbf{w}}\cdot\mathbf{x}_i)\mathbf{x}_i\Big\|_2 \\ &\leq N \sup_{\mathbf{w}\in\mathcal{B}(\mathbf{0},R)} \max_{i\in[N]} \|\xi_i\mathbf{x}_i\|_2\|f(\mathbf{w}\cdot\mathbf{x}_i) - f(\widetilde{\mathbf{w}}\cdot\mathbf{x}_i)\|_2 \\ &\leq N\|f\|_{\mathrm{lip}} \sup_{\mathbf{w}\in\mathcal{B}(\mathbf{0},R)} \max_{i\in[N]} \|\xi_i\mathbf{x}_i\|_2\|(\mathbf{w}-\widetilde{\mathbf{w}})\cdot\mathbf{x}_i\|_2 \\ &\leq N\|f\|_{\mathrm{lip}}\varepsilon_1 \max_{i\in[N]} \|\xi_i\mathbf{x}_i\|_2 \max_{i\in[N]} \|\mathbf{x}_i\|_2.\end{aligned}$$

We will bound the two maximal terms seperately. We first consider the $\|\mathbf{x}_i\|$ term. We have

$$\mathbf{Pr}\left\{\max_{j\in[N]}\|\mathbf{x}_j\|_2 \geq t_1\right\} \leq N\,\mathbf{Pr}\left\{2\max_{\mathbf{v}\in\mathcal{N}_2}|\mathbf{x}_j\cdot\mathbf{v}| \geq t_1\right\} \leq 2N6^d\exp\left(-\frac{t_1^2}{2\|\Gamma\|_2}\right) \leq \delta. \quad (7)$$

The final inequality above follows when

$$t_1 \geq \sqrt{2\|\Gamma\|_2}(\log(N) + d\log(6) + \log(2/\delta))^{1/2}.$$

We now consider the term $\xi_i\mathbf{x}_i$. We note that for any $\mathbf{v}\in\mathcal{N}_2$ that $\mathbf{x}_i\cdot\mathbf{v}$ is sub-Gaussian with norm $\|\mathbf{x}_i\cdot\mathbf{v}\|_{\psi_2} \lesssim \|\Gamma\|_2$. We have from the PDF of a Gaussian random variable,

$$\mathbf{Pr}\left\{\max_{i\in\mathrm{S}}|\xi_i| \geq t_2\right\} \leq \frac{\sqrt{2}N}{\sqrt{\pi}}\int_t^\infty e^{-\frac{x^2}{2\nu^2}}dx \leq \frac{\sqrt{2}N}{\sqrt{\pi}}\int_t^\infty \frac{xe^{-\frac{x^2}{2\nu^2}}}{\sqrt{t}}dx = \frac{\sqrt{2}N}{\sqrt{\pi}}e^{-\frac{t}{2\nu}} \leq \delta. \quad (8)$$

The final inequality holds when,

$$t_2 \geq 2\nu(\log(N) + \log(1/\delta)).$$

Combining our estimates, we have

$$\mathbf{Pr}_{\mathbf{x}\sim\mathcal{D}}\left\{\max_{i\in[n]}\|\xi_i\mathbf{x}_i\|_2\max_{i\in[n]}\|\mathbf{x}_i\|_2 \geq t_1t_2\right\} \leq \delta$$

We then choose $\varepsilon = 1/\sqrt{N}$ and we have with probability at least $1-\delta$,

$$\begin{aligned}\sup_{\mathbf{w}\in\mathcal{B}(\mathbf{0},R)}&\Big\|\sum_{i\in\mathrm{S}}\xi_i f(\mathbf{w}\cdot\mathbf{x}_i)\mathbf{x}_i - \sum_{i\in\mathrm{S}}\xi_i f(\widetilde{\mathbf{w}}\cdot\mathbf{x}_i)\mathbf{x}_i\Big\|_2\\ &\lesssim \sqrt{N}\|f\|_{\mathrm{lip}}\|\Gamma\|_2\nu\Big(\log^2(N/\delta) + d^{1/2}\log(N/\delta)\Big)\end{aligned}$$

Our proof is complete. ■

Lemma 28. *Let $X = [\mathbf{x}_1,\ldots,\mathbf{x}_N]$ 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 proxy Γ and ξ_i are sampled from $\mathcal{N}(0,\nu^2)$. Assume $f:\mathbb{R}^d\to\mathbb{R}$ is of the form,*

$$f(x) = \begin{cases}1 & x > 0\\ c & x \leq 0\end{cases}.$$

Let $\mathcal{W}$ represent all subsets of $[N]$ of size empty to $(1-\epsilon)N$. Suppose $\mathbf{w}\in\mathcal{B}(\mathbf{w}^*,R)$ and $\mathrm{S}\in\mathcal{W}$. Let $\mathcal{N}_1$ be a ε-net of $\mathbb{S}^{d-1}$ and define $\widetilde{\mathbf{w}} = \arg\min_{\mathbf{u}\in\mathcal{N}_1}\|\mathbf{w}-\mathbf{u}\|_2$. Set a failure probability $\delta\in(0,1)$, then with probability at least $1-\delta$,

$$\begin{aligned}\max_{\mathrm{S}\in\mathcal{W}}\sup_{\mathbf{w}\in\mathcal{B}(\mathbf{0},R)}&\Big\|\sum_{i\in\mathrm{S}}\xi_i f(\mathbf{w}\cdot\mathbf{x}_i)\mathbf{x}_i - \sum_{i\in\mathrm{S}}\xi_i f(\widetilde{\mathbf{w}}\cdot\mathbf{x}_i)\mathbf{x}_i\Big\|_2\\ &\lesssim \sqrt{N}\|f\|_{\mathrm{lip}}\|\Gamma\|_2\nu\big(\log^2(N/\delta) + d\log(N/\delta)\big)\end{aligned}$$

Proof. 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 $\|\mathbf{w}-\widetilde{\mathbf{w}}\|_2 \leq \varepsilon_1$. We will now consider the two seperate function classes. We have,

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

This upper bound holds for both when f is of the form,

$$f(x) = \begin{cases}1 & x > 0\\ c & x \leq 0\end{cases}$$

for any $c \in [0,1)$. We have with probability at least $1-\delta$,

$$\max_{i\in[N]} \|\xi_i \mathbf{x}_i\|_2 \lesssim \nu\sqrt{\|\Gamma\|_2}\Big(\log^{3/2}(N/\delta) + d^{1/2}\log(N/\delta)\Big)$$

We next consider the second term,

$$\begin{aligned}
&\Pr_{\mathbf{x}\sim\mathcal{D}}\left\{\sup_{\mathbf{w}\in\mathcal{B}(\mathbf{w}^*,R)} \mathbf{1}\{\mathbf{w}\cdot\mathbf{x}_i \le 0, \widetilde{\mathbf{w}}\cdot\mathbf{x}_i \ge 0\} \ge t_2\right\} \\
&\quad\le t_2^{-1} \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}}\left[\sup_{\mathbf{w}\in\mathcal{B}(\mathbf{w}^*,R)} \mathbf{1}\{\mathbf{w}\cdot\mathbf{x}_i \le 0, \widetilde{\mathbf{w}}\cdot\mathbf{x}_i \ge 0\}\right] \\
&\quad\le \frac{\Theta(\mathbf{w},\widetilde{\mathbf{w}})}{2\pi t_2} \le \frac{1}{2\pi t_2}\arcsin\left(\frac{\|\mathbf{w}-\widetilde{\mathbf{w}}\|}{\|\mathbf{w}\|}\right) \le \frac{1}{2\pi t_2}\frac{\varepsilon_1}{\|\mathbf{w}^*\|-r} \le t_2^{-1}\varepsilon_1 c_1,
\end{aligned}$$

where c_1 is a constant and $r < R$. We obtain similarly that

$$\Pr_{\mathbf{x}\sim\mathcal{D}}\left\{\sup_{\mathbf{w}\in\mathcal{B}(\mathbf{w}^*,R)} \mathbf{1}\{\mathbf{w}\cdot\mathbf{x}_i \ge 0, \widetilde{\mathbf{w}}\cdot\mathbf{x}_i \le 0\} \ge t_2\right\} \le t_2^{-1}\varepsilon_1 c_1.$$

Now we can combine our estimates and obtain with probability exceeding $1-\delta$,

$$\begin{aligned}
&\max_{\mathrm{S}\in\mathcal{W}} \sup_{\mathbf{w}\in\mathcal{B}(\mathbf{0},R)} \Big\|\sum_{i\in\mathrm{S}} \xi_i f(\mathbf{w}\cdot\mathbf{x}_i)\mathbf{x}_i - \sum_{i\in\mathrm{S}} \xi_i f(\widetilde{\mathbf{w}}\cdot\mathbf{x}_i)\mathbf{x}_i\Big\|_2 \\
&\qquad\lesssim \varepsilon_1\nu\sqrt{\|\Gamma\|_2}\delta^{-1}\Big(\log^{3/2}(N/\delta) + d^{1/2}\log(N/\delta)\Big),
\end{aligned}$$

Our proof is complete. ■

Lemma 29. *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 proxy Γ and ξ_i are sampled from $\mathcal{N}(0,\nu^2)$. Assume $f:\mathbb{R}^d\to\mathbb{R}$ is bounded over $\mathbb{R}$ and Lipschitz. Let $\mathcal{W}$ represent all subsets of $[N]$ of size empty to $(1-\epsilon)N$. Suppose $\mathbf{w}\in\mathcal{B}(\mathbf{w}^*,R)$ and $\mathrm{S}\in\mathcal{W}$. Set a failure probability $\delta\in(0,1)$, then with probability at least $1-\delta$,*

$$\|(\boldsymbol{\xi}_\mathrm{S}\circ f(X_\mathrm{S}^\top\mathbf{w}))^\top X_\mathrm{S}\|_2 \lesssim \nu\|\Gamma\|_2\|f\|_\infty N\sqrt{\epsilon\log(1/\epsilon)}$$

Proof. We use the decomposition given in Zhang et al. [2019, Lemma A.4]. We obtain,

$$\begin{aligned}
\Pr\left\{\max_{\mathrm{S}\in\mathcal{W}}\sup_{\mathbf{w}\in\mathcal{B}(\mathbf{0},R)}\Big\|\sum_{i\in\mathrm{S}}\xi_i f(\mathbf{w}\cdot\mathbf{x}_i)\mathbf{x}_i\Big\| \ge t\right\} &\le \Pr\left\{\max_{\mathrm{S}\in\mathcal{W}}\max_{\mathbf{u}\in\mathcal{N}_1}\Big\|\sum_{i\in\mathrm{S}}\xi_i f(\mathbf{x}_i\cdot\mathbf{u})\mathbf{x}_i\Big\|_2 \ge \frac{t}{2}\right\} \\
&+ \Pr\left\{\max_{\mathrm{S}\in\mathcal{W}}\sup_{\mathbf{w}\in\mathcal{B}(\mathbf{0},R)}\Big\|\sum_{i\in\mathrm{S}}\xi_i f(\mathbf{w}\cdot\mathbf{x}_i)\mathbf{x}_i - \sum_{i\in\mathrm{S}}\xi_i f(\widetilde{\mathbf{w}}\cdot\mathbf{x}_i)\mathbf{x}_i\Big\|_2 \ge \frac{t}{2}\right\}. \qquad (9)
\end{aligned}$$

From Lemma 26, we have the first term of Equation (10) is less than $\delta/2$ when

$$t \gtrsim \nu\|\Gamma\|_2\|f\|_\infty N\sqrt{\epsilon\log(1/\epsilon)}.$$

From Lemma 27, we have the second term of Equation (10) is less than $\delta/2$ when

$$t \gtrsim \sqrt{N}\|f\|_{\mathrm{lip}}\|\Gamma\|_2\nu\big(\log^2(N/\delta) + d\log(N/\delta)\big)$$

Then, when the sample complexity satisfies

$$N = \Omega\left(\frac{\log^4(N/\delta) + d\log^2(N/\delta)}{\epsilon}\right)$$

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

$$\|(\boldsymbol{\xi}_\mathrm{S}\circ f(X_\mathrm{S}^\top\mathbf{w}))^\top X_\mathrm{S}\|_2 \lesssim \nu\|\Gamma\|_2\|f\|_\infty N\sqrt{\epsilon\log(1/\epsilon)}$$

Our proof is complete. ■

Lemma 30. *Let $X = [\mathbf{x}_1, \ldots, \mathbf{x}_N]$ 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 proxy Γ and ξ_i are sampled from $\mathcal{N}(0, \nu^2)$. Assume $f : \mathbb{R}^d \to \mathbb{R}$ is of the form,*

$$f(x) = \begin{cases} 1 & x \geq 0 \\ c & x \leq 0 \end{cases}$$

Let $\mathcal{W}$ represent all subsets of $[N]$ of size empty to $(1-\epsilon)N$. Suppose $\mathbf{w} \in \mathcal{B}(\mathbf{w}^*, R)$ and $\mathrm{S} \in \mathcal{W}$. Set a failure probability $\delta \in (0,1)$, then with probability at least $1-\delta$,

$$\|(\boldsymbol{\xi}_\mathrm{S} \circ f(X_\mathrm{S}^\top \mathbf{w}))^\top X_\mathrm{S}\|_2 \lesssim \nu \|\Gamma\|_2 \|f\|_\infty N \sqrt{\epsilon \log(1/\epsilon)}$$

Proof. We use the decomposition given in Zhang et al. [2019, Lemma A.4]. We obtain,

$$\begin{aligned} \mathbf{Pr}\Big\{ \max_{\mathrm{S}\in\mathcal{W}} \sup_{\mathbf{w}\in\mathcal{B}(\mathbf{0},R)} \big\| \sum_{i\in\mathrm{S}} \xi_i f(\mathbf{w}\cdot\mathbf{x}_i)\mathbf{x}_i \big\| \geq t \Big\} &\leq \mathbf{Pr}\Big\{ \max_{\mathrm{S}\in\mathcal{W}} \max_{\mathbf{u}\in\mathcal{N}_1} \big\| \sum_{i\in\mathrm{S}} \xi_i f(\mathbf{x}_i\cdot\mathbf{u})\mathbf{x}_i \big\|_2 \geq \frac{t}{2} \Big\} \\ &+ \mathbf{Pr}\Big\{ \max_{\mathrm{S}\in\mathcal{W}} \sup_{\mathbf{w}\in\mathcal{B}(\mathbf{0},R)} \big\| \sum_{i\in\mathrm{S}} \xi_i f(\mathbf{w}\cdot\mathbf{x}_i)\mathbf{x}_i - \sum_{i\in\mathrm{S}} \xi_i f(\widetilde{\mathbf{w}}\cdot\mathbf{x}_i)\mathbf{x}_i \big\|_2 \geq \frac{t}{2} \Big\}. \end{aligned} \tag{10}$$

From Lemma 26, we have the first term of Equation (10) is less than $\delta/2$ when

$$t \gtrsim \nu \|\Gamma\|_2 N \sqrt{\epsilon \log(1/\epsilon)}.$$

From Lemma 28, we have the second term of Equation (10) is less than $\delta/2$ when

$$t \gtrsim \nu \sqrt{\|\Gamma\|_2} \delta^{-1} \Big(\log^{3/2}(N/\delta) + d^{1/2} \log(N/\delta) \Big)$$

Then, when the sample complexity satisfies

$$N = \Omega\bigg(\frac{\log^3(N/\delta) + d\log^2(N/\delta)}{\delta\epsilon} \bigg).$$

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

$$\|(\boldsymbol{\xi}_\mathrm{S} \circ f(X_\mathrm{S}^\top \mathbf{w}))^\top X_\mathrm{S}\|_2 \lesssim \nu \|\Gamma\|_2 N \sqrt{\epsilon \log(1/\epsilon)}$$

Our proof is complete. ■

Lemma 31. *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 $\|\Gamma\|_2$. Suppose $\mathrm{S} \subset [N]$ s.t. $|\mathrm{S}| \leq (1-\epsilon)N$. Then with probability at least $1-\delta$,*

$$\big\| \Sigma_\mathrm{S}(\mathbf{w}^{(t)}, \mathbf{w}^*) - \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}}[\Sigma_\mathrm{S}(\mathbf{w}^{(t)}, \mathbf{w}^*)] \big\|_2 \lesssim N\|\Gamma\| \sqrt{\epsilon \log(1/\epsilon)}$$

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 $\mathbb{S}^{d-1}$. Let $\widetilde{\mathbf{w}} = \arg\min_{\mathbf{v}\in\mathcal{N}_1} \|\mathbf{w} - \mathbf{v}\|_2$ throughout the relations. We will use the decomposition given in Theorem 1 of Mei et al. [2016] to obtain

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

We bound all terms separately. For the first term,

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

Then from Equation (7), we have with probability exceeding $1-\delta$,

$$\max_{i\in[N]} \|\mathbf{x}_i\mathbf{x}_i^\top\|_2 \le 2\|\Gamma\|_2(\log(N) + d\log(6) + \log(2/\delta)).$$

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

$$\begin{aligned}
&\mathbf{Pr}\left\{\max_{\mathrm{S}\in\mathcal{W}}\max_{\widetilde{\mathbf{w}}\in\mathcal{N}_1} \|\Sigma_{\mathrm{S}}(\widetilde{\mathbf{w}},\mathbf{w}^*) - \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}}[\Sigma_{\mathrm{S}}(\widetilde{\mathbf{w}},\mathbf{w}^*)]\|_2 \ge \frac{t}{3}\right\} \\
&\le \mathbf{Pr}\left\{\max_{\mathrm{S}\in\mathcal{W}}\max_{\widetilde{\mathbf{w}}\in\mathcal{N}_1}\max_{\mathbf{v}\in\mathcal{N}_2} |\|\widetilde{X}_{\mathrm{S}}^\top\mathbf{v}\|_2^2 - \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}}\|\widetilde{X}_{\mathrm{S}}^\top\mathbf{v}\|_2^2| \ge \frac{2t}{3}\right\} \\
&\overset{(i)}{\le} 2\left(\frac{\mathrm{e}}{\epsilon}\right)^{N\epsilon}(3/\varepsilon_1)^{Rd}12^d \exp\left[-c\min\left(\frac{t^2}{144\cdot 256\|\Gamma\|_2^2|\mathrm{S}|}, \frac{t}{4\cdot 48\|\Gamma\|_2}\right)\right] \le \frac{\delta}{2}
\end{aligned}$$

In (i) we note from Lemma 1.12 in Rigollet and Hütter [2023], 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 and $\||\widetilde{\mathbf{x}}\cdot\mathbf{v}|^2 - \mathbf{E}_{\mathbf{x}\sim\mathcal{D}}|\widetilde{\mathbf{x}}\cdot\mathbf{v}|^2\|_{\psi_1} \le 16\|\Gamma\|_2$, we can then apply Bernstein's Inequality (see Lemma 25). The probabilistic condition above is then satisfied when,

$$t \gtrsim \left(N\|\Gamma\|_2^2(Rd\log(3/\varepsilon_1) + d\log(3/\varepsilon_2) + 3N\epsilon\log(1/\epsilon) + \log(4/\delta))\right)^{1/2}$$

We now consider the third term.

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

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 $\mathbf{E}_{\mathbf{x}\sim\mathcal{D}}\|\mathbf{x}\|^4 \le L\,\mathbf{E}_{\mathbf{x}\sim\mathcal{D}}\|\mathbf{x}\|^2 = L\,\mathrm{tr}(\Sigma)$. We now consider the second term,

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

Then we obtain zero probability as indicated in the statement when

$$t \gtrsim N\sqrt{L\,\mathrm{tr}(\Sigma)\varepsilon_1}$$

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

$$\max_{\mathrm{S}\in\mathcal{W}} \sup_{\mathbf{w}\in\mathcal{B}(\mathbf{w}^*,R)} \|\Sigma_{\mathrm{S}}(\mathbf{w},\mathbf{w}^*) - \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}}[\Sigma_{\mathrm{S}}(\mathbf{w},\mathbf{w}^*)]\|_2 \lesssim N\|\Gamma\|_2\sqrt{\epsilon\log(1/\epsilon)}$$

Our proof is complete. ■

Lemma 32. *Suppose* $\mathbf{E}_{\mathbf{x}\sim\mathcal{D}}[\mathbf{x}] = \mathbf{0}$ *and* $\mathbf{E}_{\mathbf{x}\sim\mathcal{D}}\left[\mathbf{x}\mathbf{x}^\top\right] = I$ *for* $\mathbf{x} \sim \mathcal{D}$ *where* $\mathcal{D}$ *is a rotationally invariant distribution. Fix* $\mathbf{w}_1, \mathbf{w}_2 \in \mathbb{R}^d$ *and define* $\Theta = \arccos\left(\frac{\mathbf{w}_1\cdot\mathbf{w}_2}{\|\mathbf{w}_1\|\|\mathbf{w}_2\|}\right) \leq \frac{\pi}{2}$. *Then,*

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

Proof. Since we have $\mathcal{D}$ is an isotropic distribution, we then have the distribution of $U\mathbf{x}$ is isotropic for any unitary U. Then consider the unitary matrix,

$$U = \begin{bmatrix} \frac{\mathbf{w}_1}{\|\mathbf{w}_1\|} & \frac{\mathbf{w}_2 - \mathrm{Proj}_{\mathbf{w}_2}\mathbf{w}_1}{\|\mathbf{w}_2 - \mathrm{Proj}_{\mathbf{w}_2}\mathbf{w}_1\|} & \mathbf{u}_3 & \dots & \mathbf{u}_d \end{bmatrix}$$

where $\mathbf{u}_3, \dots, \mathbf{u}_d$ represents some orthonormal basis of the complementary subspace to the subspace spanned by $\mathbf{w}_1$ and $\mathbf{w}_2$. 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 $\Theta \leq \frac{\pi}{2}$.

$$\begin{aligned}
&\mathop{\mathbf{E}}_{X\sim\mathcal{D}}\left[XX^\top \cdot \mathbf{1}\left\{X^\top\mathbf{w}_1 \geq \mathbf{0}\right\} \cdot \mathbf{1}\left\{X^\top\mathbf{w}_2 \geq \mathbf{0}\right\}\right] \\
&\quad= \mathop{\mathbf{E}}_{X\sim\mathcal{D}}\left[U^\top UXX^\top U^\top U \cdot \mathbf{1}\left\{X^\top\mathbf{w}_1 \geq \mathbf{0}\right\} \cdot \mathbf{1}\left\{X^\top\mathbf{w}_2 \geq \mathbf{0}\right\}\right] \\
&\quad= U^\top \mathop{\mathbf{E}}_{X\sim\mathcal{D}}\left[UXX^\top U^\top \cdot \mathbf{1}\left\{X^\top\mathbf{w}_1 \geq \mathbf{0}\right\} \cdot \mathbf{1}\left\{X^\top\mathbf{w}_2 \geq \mathbf{0}\right\}\right] U
\end{aligned}$$

In the above, the first equality follows from noting that U is unitary, and the second equality follows from the linearity of expectation. Let $\Omega_\xi = \left\{\mathbf{x} \in \mathbb{R}^d : \mathbf{x}^\top U^\top \mathbf{w}_\xi \geq 0\right\}$ for $\xi \in \{1, 2\}$. Then,

$$\begin{aligned}
&\mathop{\mathbf{E}}_{X\sim\mathcal{D}}\left[UXX^\top U^\top \cdot \mathbf{1}\left\{X^\top\mathbf{w}_1 \geq \mathbf{0}\right\} \cdot \mathbf{1}\left\{X^\top\mathbf{w}_2 \geq \mathbf{0}\right\}\right] \\
&\quad= \mathop{\mathbf{E}}_{Y\sim\mathcal{D}}\left[YY^\top \cdot \mathbf{1}\{Y_1 \geq \mathbf{0}\} \cdot \mathbf{1}\{\alpha Y_1 + \beta Y_2 \geq \mathbf{0}\}\right] \\
&\quad= \begin{bmatrix} \mathbf{E}_{Y\sim\mathcal{D}}\left[Y_1^2 \cdot \mathbf{1}\{X \in \Omega\}\right] & \cdots & \mathbf{E}_{X\sim\mathcal{D}}[(Y_1)(Y_d) \cdot \mathbf{1}\{X \in \Omega\}] \\ \vdots & \ddots & \vdots \\ \mathbf{E}_{Y\sim\mathcal{D}}[Y_dY_1 \cdot \mathbf{1}\{X \in \Omega\}] & \cdots & \mathbf{E}_{X\sim\mathcal{D}}\left[Y_d^2 \cdot \mathbf{1}\{Y \in \Omega\}\right] \end{bmatrix}
\end{aligned}$$

In the above, the first relation follows from rotational invariance, which gives us that $UX \overset{d}{=} X$ for any unitary U. Our main tool is a change of coordinates from the euclidean space to n-spherical coordinates as all the angles will be independent. We first consider the non-diagonal elements. Suppose $(i, j) \in [d] \setminus [2] \times [d] \setminus [2]$ and w.l.o.g $i < j$, from the rotational invariance, we have

$$\mathop{\mathbf{E}}_{X\sim\mathcal{D}}[Y_jY_i \cdot \mathbf{1}\{X \in \Omega\}] = \mathop{\mathbf{E}}_{X\sim\mathcal{D}}\left[r^2 \prod_{k_1\in[i-1]} \sin(\phi_{k_1})\cos(\phi_i) \prod_{k_2\in[j-1]} \sin\phi_{k_2}\cos\phi_j\right] = \mathbf{0}$$

In the above, the final inequality follows from noting that $\int_0^\pi \cos\theta d\theta = 0$ and if $j = n$ we have $\int_0^{2\pi} \sin\theta d\theta = 0$. Then for the diagonal elements, we have $i = j$, and obtain

$$\begin{aligned}
&\mathop{\mathbf{E}}_{X\sim\mathcal{D}}\left[(X^\top U_j)^2 \cdot \mathbf{1}\{X \in \Omega\}\right] = \mathop{\mathbf{E}}_{X\sim\mathcal{D}}\left[r^2 \prod_{k\in[j-1]} \sin^2(\phi_k)\cos^2(\phi_i)\right] \\
&\quad= \frac{1}{\mathsf{S}_j} \underbrace{\int_0^\infty r^2 d\mu(r)}_{I} \underbrace{\int_{-\pi/2+\Theta}^{\pi/2} \sin^2(\phi_1)d\phi_1}_{II} \underbrace{\int_0^{2\pi} \cos^2(\phi_{j-1})d\phi_{j-1} \prod_{k\in[j-1]\setminus\{1\}} \int_0^\pi \sin^2(\phi_k)d\phi_k}_{III}
\end{aligned}$$

The term of interest here is II. Recall Θ is the angle between $\mathbf{w}_1$ and $\mathbf{w}_2$, we then have

$$II = \int_{-\pi/2+\Theta}^{\pi/2} \sin^2(\phi_1)d\phi_1 = \frac{2\pi - 2\Theta - \sin(2\Theta)}{4} \geq \frac{\pi - 2\Theta}{2}$$

Before proceeding, let us note the following,

$$\mathop{\mathbf{E}}_{X\sim\mathcal{D}}\left[(X^\top U_j)^2\right] = \frac{1}{\mathsf{S}_j} \cdot I \cdot III \cdot \int_{-\pi/2}^{\pi/2} \sin^2(\phi_1)d\phi_1 = 1$$

In the above, the final equality follows from noting $\mathcal{D}$ is isotropic and thus diagonal covariance elements are unitary. We will now consider the normalizing term, note that $\mathbf{E}_{X\sim\mathcal{D}}\big[(X^\top U_j)^2\big] = 1$ as $\mathcal{D}$ is isotropic. Then, noting that only II is modified when considering $\mathbf{1}\{X \in \Omega\}$. We then obtain $S_j = I \cdot III \cdot \int_0^\pi \sin^2(\phi_1) d\phi_1$. Then, from rearranging, we have

$$\mathop{\mathbf{E}}_{X\sim\mathcal{D}}\big[(X^\top U_j)^2 \cdot \mathbf{1}\{X \in \Omega\}\big] \geq \frac{\pi - 2\Theta}{\pi}$$

We will now consider the principal 2×2 matrix. Then, consider the vector in $\mathbb{R}^2$ in polar coordinates as $(r\cos\Theta, r\sin\Theta)$ where Θ is uniformly distributed in $[0, 2\pi)$ and r is independent. Then, to calculate the expected outer product, we integrate over the two dimensional space of the intersection of $\mathbf{w}_1$ and $\mathbf{w}_2$,

$$\begin{aligned}
&\Big[\mathop{\mathbf{E}}_{X\sim\mathcal{D}}\big[UXX^\top U^\top \cdot \mathbf{1}\big\{X^\top U^\top \mathbf{w}_1 \geq \mathbf{0}\big\} \cdot \mathbf{1}\big\{X^\top U^\top \mathbf{w}_2 \geq \mathbf{0}\big\}\big]\Big]_{2,2} \\
&= \Big[\mathop{\mathbf{E}}_{X\sim\mathcal{D}}\big[YY^\top \cdot \mathbf{1}\{Y_1 \geq \mathbf{0}\} \cdot \mathbf{1}\{\alpha Y_1 + \beta Y_2 \geq \mathbf{0}\}\big]\Big]_{2,2} \\
&= \int_0^\infty \int_{-\pi/2+\Theta}^{\pi/2} \begin{pmatrix}\cos\Theta \\ \sin\Theta\end{pmatrix} (\cos\Theta, \sin\Theta) r d\Theta dr \\
&= \int_0^\infty \frac{r}{2} \begin{pmatrix}\pi - \Theta + \sin\Theta\cos\Theta & \sin^2\Theta \\ \sin^2\Theta & \pi - \Theta - \cos\Theta\sin\Theta\end{pmatrix} dr \\
&= (1/4) \cdot \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}}\big[r^2\big] \begin{pmatrix}\pi - \Theta + \sin\Theta\cos\Theta & \sin^2\Theta \\ \sin^2\Theta & \pi - \Theta - \cos\Theta\sin\Theta\end{pmatrix} \\
&\succeq \left(\frac{\pi - \Theta - \sin\Theta}{2}\right) \cdot dI
\end{aligned}$$

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 33. *Fix $S \in \mathbb{R}^{k\times n}, T \in \mathbb{R}^{m\times \ell}$, then sample a matrix $G \in \mathbb{R}^{n\times m}$ with entries sampled i.i.d from $\mathcal{N}(0, \sigma^2)$, then with probability exceeding $1 - \delta$,*

$$\|SGT\|_{\mathrm{F}} \leq \|S\|_{\mathrm{F}} \|T\|_{\mathrm{F}} \cdot \sigma\sqrt{2\log(2nm/\delta)}$$

Proof. The proof will be a calculation.

$$\|SGT\|_{\mathrm{F}}^2 = \sum_{i\in[k]} \sum_{j\in[\ell]} \sum_{k_1,k_2\in[n]\times[m]} (S_{i,k_1} G_{k_1,k_2} T_{k_2,j})^2 \leq \|S\|_{\mathrm{F}}^2 \|T\|_{\mathrm{F}}^2 \max_{i,j\in[n]\times[m]} (G_{i,j})^2$$

It then suffices to bound the maximum value of a Gaussian squared over N^2 samples. Then, we have from a union bound and the definition of a Gaussian random variable,

$$\begin{aligned}
\mathop{\mathbf{Pr}}_{G_{i,j}\sim\mathcal{N}(0,1)}\left\{\max_{(i,j)\in[n]\times[m]} G_{i,j}^2 \geq t\right\} &= \mathop{\mathbf{Pr}}_{G_{i,j}\sim\mathcal{N}(0,1)}\left\{\max_{(i,j)\in[n]\times[m]} |G_{i,j}| \geq \sqrt{t}\right\} \\
&\leq \frac{\sqrt{2}nm}{\pi} \int_{\sqrt{t}}^\infty \mathrm{e}^{-\frac{x^2}{2\sigma^2}} dx \leq \frac{\sqrt{2}nm}{\pi} \int_{\sqrt{t}}^\infty \frac{x\mathrm{e}^{-\frac{x^2}{2\sigma^2}}}{\sqrt{t}} dx = \frac{\sqrt{2}nm}{\sqrt{\pi}} \mathrm{e}^{-\frac{t}{2\sigma^2}} \leq \delta
\end{aligned}$$

In the above, (i) follows from a union bound. We thus obtain from elementary inequalities, with failure probability at most δ,

$$\max_{(i,j)\in[n]\times[m]} G_{i,j}^2 \leq \sigma^2 \cdot (2\log(2nm/\delta))$$

Our proof is complete. ■

Lemma 34. *Let $X = [\mathbf{x}_1, \ldots, \mathbf{x}_n]$ be the data matrix such that $\mathbf{x}_i$ are sampled from a sub-Gaussian Distribution with second-moment matrix Σ and sub-Gaussian proxy Γ for $i \in [N]$. Let ξ_i be sampled from $\mathcal{N}(0, \nu^2)$ for $i \in [N]$. Let $\mathcal{W}$ represent all subsets of $[N]$ of size empty to $(1 - \epsilon)N$. For any $\delta \in (0, 1)$, with probability at least $1 - \delta$,*

$$\Big\|\sum_{i\in\mathrm{S}} \xi_i \mathbf{x}_i\Big\|_2 \lesssim \nu\|\Gamma\|_2 \sqrt{N\log(2/\delta) + Nd\log(6) + 3N^2\epsilon\log(1/\epsilon)}$$

Proof. We have by the definition of the ℓ_2 norm,

$$\big\|\sum_{i\in \mathrm{S}} \xi_i \mathbf{x}_i\big\|_2 = \max_{\mathbf{v}\in\mathbb{S}^{d-1}} \big|\sum_{i\in \mathrm{S}} \xi_i \mathbf{x}_i \cdot \mathbf{v}\big| \triangleq \big|\sum_{i\in \mathrm{S}} \xi_i \mathbf{x}_i \cdot \mathbf{v}^*\big|$$

Let $\mathcal{W}$ be a ε-net of $\mathbb{S}^{d-1}$, from the definition of a ε-net, we have the existence of $\mathbf{u} \in \mathcal{W}$ such that $\|\mathbf{u} - \mathbf{v}^*\| \leq \varepsilon$. Then, from the reverse triangle inequality, we have

$$\big\|\sum_{i\in \mathrm{S}} \xi_i \mathbf{x}_i \cdot \mathbf{u}\big\|_2 \geq \big\|\sum_{i\in \mathrm{S}} \xi_i \mathbf{x} \cdot \mathbf{v}^*\big\|_2 - \big\|\sum_{i\in \mathrm{S}} \xi_i \mathbf{x}_i \cdot (\mathbf{v}^* - \mathbf{u})\big\|_2 \geq (1-\varepsilon)\big\|\sum_{i\in \mathrm{S}} \xi_i \mathbf{x}_i\big\|_2$$

Then from rearranging, we have

$$\big\|\sum_{i\in \mathrm{S}} \xi_i \mathbf{x}_i\big\|_2 \leq \frac{1}{1-\varepsilon} \cdot \max_{\mathbf{u}\in\mathcal{W}} \big|\sum_{i\in \mathrm{S}} \xi_i \mathbf{x}_i \cdot \mathbf{u}\big|$$

Next, from Lemma 2.7.7 in Vershynin [2020], we can note that for any $\mathbf{u} \in \mathbb{S}^{d-1}$, we have $\xi_i \mathbf{x}_i \cdot \mathbf{u}$ is sub-exponential and $\|\xi_i \mathbf{x}_i \cdot \mathbf{u}\|_{\psi_1} \leq \nu \|\Gamma\|_2$. We next choose $\varepsilon = 1/2$. We can now make our probabilistic bounds.

$$\begin{aligned}
\Pr_{X\sim\mathcal{D}}\left\{\max_{\mathrm{S}\in\mathcal{W}} \big\|\sum_{i\in \mathrm{S}} \xi_i \mathbf{x}_i\big\|_2 \geq t\right\} &\leq \binom{N}{N\epsilon} \Pr_{X\sim\mathcal{D}}\left\{\max_{\mathbf{u}\in\mathcal{W}} \big|\sum_{i\in \mathrm{S}} \xi_i \mathbf{x}_i \cdot \mathbf{u}\big| \geq \frac{t}{2}\right\} \\
&\leq \left(\frac{\mathrm{e}}{\epsilon}\right)^{N\epsilon} 6^d \Pr_{X\sim\mathcal{D}}\left\{\big|\sum_{i\in \mathrm{S}} \xi_i \mathbf{x}_i \cdot \mathbf{u}\big| \geq t\right\} \\
&\leq 2\left(\frac{\mathrm{e}}{\epsilon}\right)^{N\epsilon} 6^d \exp\left[-c\min\left(\frac{t^2}{C_\sigma^2 \|\Gamma\|_2^2 |\mathrm{S}|}, \frac{t}{C_\sigma \|\Gamma\|_2}\right)\right] \leq \delta
\end{aligned}$$

In the above, the final condition holds when

$$t \geq \sqrt{c^{-1} C_\sigma^2 \|\Gamma\|_2^2 N(\log(2/\delta) + d\log(6) + 3N\epsilon\log(1/\epsilon))}$$

Our proof is complete. ■

Lemma 35. *Let $X \sim \mathcal{N}(0, \nu^2)$, then $\|X\|_{\psi_2} = \sqrt{8/3}\nu$.*

Proof. We have from Definition 4 that

$$\|X\|_{\psi_2} = \inf\big\{c \geq 0 : \mathbf{E}\big[\exp(X^2/c^2)\big] \leq 2\big\}.$$

We will now solve for the minimizing c. From the PDF of a standard Gaussian,

$$\begin{aligned}
\mathbf{E}\big[\exp(X^2/c^2)\big] &= \frac{1}{\nu\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp\left(X^2\left(\frac{1}{c^2} - \frac{1}{2\nu^2}\right)\right) dX \\
&= \frac{1}{\nu\sqrt{2}}\left(\frac{1}{c^2} - \frac{1}{2\nu^2}\right)^{-1/2} = 2.
\end{aligned}$$

From some algebra, we find the final inequality above holds when $c = \sqrt{8/3}\nu$. ■

D Mathematical Tools

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

Lemma 36. *Let $\mathbf{a}, \mathbf{b} \in \mathbb{R}^n$ and $X \in \mathbb{R}^{p\times n}$, then*

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

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

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

where the final inequality comes from noting

$$\|\mathbf{a}\circ\mathbf{b}\|^2 = \sum_{i\in[n]} a_i^2 b_i^2 \leq \max_{i\in[n]} a_i^2 \cdot \sum_{i\in[n]} b_i^2 = \|\mathbf{a}\|_\infty^2 \|\mathbf{b}\|_2^2$$

Our proof is complete. ■

Lemma 37 (Lemma 3.11 in Bubeck et al. [2015])**.** *Let f be β-smooth and α-strongly convex over $\mathbb{R}^n$, then for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$,*

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

Lemma 38 (Sum of Binomial Coefficients [Cormen et al., 2022])**.** *Let $k, n \in \mathbb{N}$ such that $k \leq n$, then*

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

Lemma 39 (Corollary 4.2.13 in Vershynin [2020])**.** *The covering number of the ℓ_2-norm ball $\mathcal{B}(\mathbf{0}; 1)$ for $\varepsilon < 0$, satisfies,*

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

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We present our main theoretical claims in the abstract and go into further detail in the introduction. In the introduction we present results which accurately reflect our main contributions in Section 3.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: All distributional assumptions are made clear in the theorem statements in Section 3.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: The theorem statements include all assumptions necessary for the claims. A sketch of the proofs of our main theorems is given in Section 3.3. All the complete proofs of the main theorems are given in Appendices A and B. The proofs in Appendices A and B utilize results from Appendix C. The results in Appendix C are given with complete proofs.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [NA]

Justification: Our code does not include experiments.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [NA]

   Justification: Our paper does not include the use of code or data.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [NA]

   Justification: Our paper's contribution is theoretical, and we thus do not run experiments.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [NA]

   Justification: Our paper does not include any experiments.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [NA]

   Justification: Our paper does not include any empirical experiments.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We preserve anonymity and abide by the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Our work is a theoretical analysis of an algorithm that has been used for over two centuries. In the introduce, we discuss the positive societal impacts of learning in the presence of a possibly adversarially corrupted dataset.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: Our paper does not pose any risks as we do not release any data or models.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [NA]

    Justification: Our paper does not use any existing assets.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.

- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: Our paper does not release any assets.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: Our research does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: Our research does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.