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

Arvind Rathnashyam
RPI Math, rathna@rpi.edu

Alex Gittens
RPI CS, 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. 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 $O(C_\sigma \nu \epsilon \log(1/\epsilon))$ approximation bound in ℓ_2-norm, with sample complexity $O(d/\epsilon)$ and failure probability $e^{-\Omega(d)}$, where C_σ is a constant dependent on the activation function.

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 ϵ-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 $0 \leq \epsilon < 1/2$, an adversary is allowed to inspect all samples and modify ϵn samples arbitrarily. This corrupted set of n points is then given as input to the learning algorithm.*

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., 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 $\mathrm{X} = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$ 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*

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

with probability at least $1 - \delta$.

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 (Algorithms 1 and 2). Our proof techniques

Table 1: Iterative thresholding for linear regression under strong ϵ-contamination. All results are with high probability; big-O hides constants and polylog factors. Assume centered sub-Gaussian covariates with second moment I, proxy $\Gamma \precsim I_d$, dimension d, corruption rate $\epsilon < 1/2$, and noise variance ν.

Reference	Approximation error	Overall runtime	Method
Bhatia et al. [2015]	$O(\nu)$	$O\left(nd^2 \log \frac{\|\mathbf{b}\|}{\epsilon\sqrt{n}}\right)$	Full solve
Shen and Sanghavi [2019]	$O(\nu)$	$O\left(nd^2 \log \frac{\|\mathbf{w}^*\|}{\nu}\right)$	Gradient descent
Awasthi et al. [2022]	$O\big(\nu\,\epsilon \log(1/\epsilon)\big)$	$O\left(\frac{nd^2}{\nu\,\epsilon^2}\right)$	Full solve
This work (Theorem 8)	$O\big(\nu\,\epsilon \log(1/\epsilon)\big)$	$O\left(nd^2 \log \frac{\|\mathbf{w}^*\|}{\nu\,\epsilon}\right)$	Gradient descent

Notes. (i) "Approximation error" is the statistical accuracy after convergence (optimization error contracts below this). (ii) Runtimes count arithmetic on $X \in \mathbb{R}^{d\times n}$; constants and $\log(1/\delta)$ factors suppressed. (iii) If $\Sigma \neq I$, whiten; otherwise replace I by Σ and multiply constants by $\sqrt{\lambda_{\max}(\Sigma)}/\lambda_{\min}(\Sigma)$ as appropriate. (iv) The "signal" placeholder in the first row should match that paper's stated scale (e.g., a condition-number–like ratio); keep logs dimensionless.

Table 2: Error upper bounds under covariate assumptions (high probability).

Reference	Approximation	Neuron	Covariate assumption
Theorem 8	$O\big(C_\sigma \nu\,\epsilon \log(1/\epsilon)\big)$	General	Sub-Gaussian (covariate corruption)
Theorem 13	$O\big(\nu\,\epsilon \log(1/\epsilon)\big)$	ReLU	Rotationally invariant sub-Gaussian (no covariate corruption)

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 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 ℓ_2 norm of the label 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 summarizes our results on learning single neuron models. In Theorem 13, 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 $O(C_\sigma \nu\epsilon \log(1/\epsilon))$, where C_σ is a constant dependent on the activation function.

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 near-linear filtering algorithm of Dong et al. [2019].

Paper Outline: In Section 2, we give the mathematical notation used in the main body of the paper before discussing the literature on iterative thresholding for robust learning. In Section 3, we present our formal results and a proof sketch for the learning of linear and non-linear neurons. We defer all proofs of our main results to Appendix A.

2 Preliminaries

2.1 Mathematical notation and background

Notation. We use $[T]$ to denote the set $\{1, 2, \ldots, T\}$. We say $y \lesssim x$ or $y = O(x)$ if there exists a constant

C such that $y \leq Cx$. We say $y \gtrsim x$ or $y = \Omega(x)$ if there exists a constant C s.t. $y \geq Cx$. We define $\mathbb{S}^{d-1}$ as the $d-1$-dimensional sphere $\{\mathbf{x} \in \mathbb{R}^d : \|\mathbf{x}\|_2 = 1\}$.

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 spectral norm of a matrix can be characterized as $\|A\| = \max_{\mathbf{v} \in \mathbb{S}^{n-1}} \|A\mathbf{v}\|$ for $A \in \mathbb{R}^{m \times n}$.

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, Pensia 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 $\mathbf{v} \in \mathbb{S}^{d-1}$,*

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

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

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

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*

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

We defer more technical preliminaries to the relevant proofs in the Appendix.

2.2 Related work

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 $O(\nu)$.

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 Covariate filtering

Recent works such as [Awasthi et al., 2022, Pensia et al., 2020], run a covariate filtering algorithm as a preprocessing step on then inputs. The goal of the covariate filtering algorithm is to return a weight vector $\boldsymbol{\omega} \in \mathbb{R}^n$ so the resulting dataset has the resilence property.

Definition 6 (Resilience)**.** *A set of points $\{\mathbf{x}_i\}_{i\in G}$ lying in $\mathbb{R}^d$ is (ϵ,τ)-resilient in a norm $\|\cdot\|$ if for any $S \subset T$, $|S| \geq (1-\epsilon)n$, it holds that $\|\frac{1}{|S|}\sum_{i\in S}\mathbf{x}_i - \mu\| \leq \tau$.*

From Zhu et al. [2022], sub-Gaussian distributions are $(\epsilon, \epsilon\sqrt{\log(1/\epsilon)})$-resilient. We first run the filtering algorithm for robust mean estimation Dong et al. [2019]. We are then returned an ϵ-corrupted dataset, $\{\widetilde{\mathbf{x}}_i\}_{i=1}^n \coloneqq \{\omega_i\mathbf{x}_i\}_{i=1}^n$, where $0 \leq \omega_i \leq 1$ that satisfies for any $S \subset [n]$, $|S| \geq (1-\epsilon)n$,

$$\left\|\sum\nolimits_S \omega_i\mathbf{x}_i\mathbf{x}_i^\top - I\right\| = O(n\epsilon\log(1/\epsilon)).$$

Essentially, the mean and covariance of any sufficiently large subset is close to the expected value.

3.2 Linear regression

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 general neuron problem for $\mathbf{w} \in \mathbb{R}^d$, $X \in \mathbb{R}^{d\times n}$, and $\mathbf{y} \in \mathbb{R}^n$ is

$$\mathcal{L}(\mathbf{w};S) = \sum\nolimits_{i\in S}(\sigma(\mathbf{w}\cdot\mathbf{x}_i) - y_i)^2, \quad \mathcal{R}(\mathbf{w};S) = \frac{1}{(1-\epsilon)n}\cdot\sum\nolimits_{i\in S}(\sigma(\mathbf{w}\cdot\mathbf{x}_i) - y_i)^2.$$

For the linear regression case, we consider the identity neuron $\sigma(x) = x$. We will first give some notation prior to presenting the algorithm.

Definition 7 (Hard Thresholding Operator)**.** *For any vector $\mathbf{x} \in \mathbb{R}^n$, define the order statistics as $(\mathbf{x})_{(1)} \leq (\mathbf{x})_{(2)} \leq \ldots \leq (\mathbf{x})_{(n)}$, the hard thresholding operator is given as*

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

Algorithm 1 Gradient descent iterative thresholding with covariate filtering

input: Samples $S = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$, ϵ, η, T, $\mathbf{w}^{(0)}$, σ
output: $O(C_\sigma \nu\epsilon \log(1/\epsilon))$-approximate solution $\mathbf{w} \in \mathbb{R}^d$

$\boldsymbol{\omega} = \textsc{CovariateFiltering}(S, \epsilon)$ ▷ Algorithm 4 in Dong et al. [2019]
for $t = 0, \ldots, T-1$ **do**
  $r_i^{(t)} = (\sigma(\mathbf{w}^{(t)} \cdot \mathbf{x}_i) - y_i)^2$ for all $i \le n$ ▷ Calculate $\mathcal{L}$ for each sample
  $S^{(t)} \leftarrow \mathrm{HT}(\mathbf{r}^{(t)}, (1-\epsilon)n)$ ▷ Hard thresholding (see Definition 7)
  $\mathbf{w}^{(t+1)} \leftarrow \mathrm{UPDATE}(\mathbf{w}^{(t)}, S^{(t)}, \eta, \mathbf{r}^{(t)}, \boldsymbol{\omega})$ ▷ Gradient descent update
end for
return: $\mathbf{w}^{(T)}$

With the hard thresholding operator from Definition 7 in hand, we are now ready to present our algorithm for learning a general neuron.

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 $(1-\epsilon)n$-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 $X_{S^{(t)}} X_{S^{(t)}}^\top$ 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 a $O(C_\sigma \nu\epsilon \log(1/\epsilon))$ ℓ_2-approximation error.

Theorem 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, where $\mathbf{x}_i$ are sampled from a sub-Gaussian distribution with second-moment matrix Σ and sub-Gaussian proxy Γ. Suppose $y_i = \mathbf{w}^* \cdot \mathbf{x}_i + \xi_i$ where $\xi_i \sim \mathcal{N}(0, \nu^2)$ for all $i \in G$. Then after $O(\kappa(\Sigma)\log(\|\mathbf{w}^*\|/\epsilon))$ iterations, $n = \Omega(d/\epsilon^2)$, and learning rate $\eta = 0.1 \cdot (\gamma/L)^2$, Algorithm 1 returns $\hat{\mathbf{w}}$ such that*

$$\|\hat{\mathbf{w}} - \mathbf{w}^*\| = O\left(\nu\epsilon \log(1/\epsilon)\right),$$

with probability exceeding $1 - 3Te^{-\Omega(d)}$.

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

We are able to recover the result of Lemma 4.2 in Awasthi et al. [2022] when the covariates (corrupted and un-corrupted) are sampled from a sub-Gaussian distribution with second-moment matrix I and $\Gamma \precsim I$. The full solve algorithm studied in Awasthi et al. [2022] returns a $O(\nu\epsilon\log(1/\epsilon))$ in time $O(nd^2/\epsilon^2)$ and the same algorithm studied in Bhatia et al. [2015], TORRENT-FC otains $O(\nu)$ approximation error in run-time $O\left(nd^2 \log\left(\frac{1}{\sqrt{n}} \frac{\|\mathbf{b}\|}{\nu\epsilon\log(1/\epsilon)}\right)\right)$, with the gradient descent based approach, we are able to improve the runtime to $O\left(nd^2 \log\left(\frac{\|\mathbf{w}^*\|}{\nu\epsilon}\right)\right)$ for the same approximation bound. 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.

3.3 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, monotonically increasing, and differentiable almost everywhere.*

Property 10. *σ is Lipschitz, i.e. $|\sigma(x) - \sigma(y)| \le L|x-y|$.*

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

3.4 Learning a general non-linear neuron

We now study the problem of minimizing the ℓ_2 loss for a general non-linear activation function that satisfies Properties 9 to 11. For a single training sample $(\mathbf{x}_i, y_i) \sim \mathcal{D}$, the loss is given as follows,

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

Algorithm 2 Gradient Descent Iterative Thresholding with randomized initialization

input: Samples $S = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$, ϵ, η, T
output: $O(\nu\epsilon \log(1/\epsilon))-$Approximate solution $\mathbf{w} \in \mathbb{R}^d$ to minimize $\|\mathbf{w} - \mathbf{w}^*\|_2$.

1: $\mathbf{w}^{(0)} \sim \mathcal{B}_d(\alpha\|\mathbf{w}^*\|)$ ▷ cf. Lemma 12
2: **for** $t = 0, \ldots, T-1$ **do**
3: $\quad r_i^{(t)} = (\sigma(\mathbf{x}_i^\top \mathbf{w}^{(t)}) - y_i)^2$ for all $i \le n$ ▷ Calculate $\mathcal{L}$ for each point
4: $\quad S^{(t)} \leftarrow \mathrm{HT}(\mathbf{r}^{(t)}, (1-\epsilon)n)$ ▷ cf. Definition 7
5: $\quad \mathbf{w}^{(t+1)} \leftarrow \mathbf{w}^{(t)} - \eta \nabla \mathcal{R}(\mathbf{w}^{(t)}; S^{(t)})$ ▷ Gradient Descent Update
6: **end for**
return: $\mathbf{w}^{(T)}$

Proof of Sketch. We will give a general sketch of the proof for our general neuron result, deferring the more technical ideas to the proof in the appendix. For any $1 \le t \le T$, we have

$$\begin{aligned}\|\mathbf{w}^{(t+1)} - \mathbf{w}^*\| &= \|\mathbf{w}^{(t)} - \mathbf{w}^* - \eta \nabla \mathcal{R}(\mathbf{w}^{(t)}; S^{(t)})\| \\ &\le \|\mathbf{w}^{(t)} - \mathbf{w}^* - \eta \nabla \mathcal{R}(\mathbf{w}^{(t)}; S^{(t)} \cap G)\| + \underbrace{\|\eta \nabla \mathcal{R}(\mathbf{w}^{(t)}; S^{(t)} \cap E)\|}_{\text{corrupted gradient}} := \Xi_1 + \Xi_2. \end{aligned} \tag{1}$$

In the above, the second relation follows from the linearity of the loss function, which gives

$$\nabla \mathcal{R}(\mathbf{w}^{(t)}; S^{(t)}) = \nabla \mathcal{R}(\mathbf{w}^{(t)}; S^{(t)} \cap G) + \nabla \mathcal{R}(\mathbf{w}^{(t)}; S^{(t)} \cap E),$$

and then applying the triangle inequality. We upper bound Ξ_1 through its square,

$$\Xi_1^2 = \|\mathbf{w}^{(t)} - \mathbf{w}^*\|_2^2 - \underbrace{2\eta \cdot \langle \mathbf{w}^{(t)} - \mathbf{w}^*, \nabla \mathcal{R}(\mathbf{w}^{(t)}; S^{(t)} \cap G) \rangle}_{\Xi_{11}} + \underbrace{\eta^2 \cdot \|\nabla \mathcal{R}(\mathbf{w}^{(t)}; S^{(t)} \cap G)\|^2}_{\Xi_{12}}.$$

In this step the finer details of the proof for the general neuron and the ReLU neuron will differ, however the structure remains the same. In the appendix, we prove the existence of a constant $C_1 = \Omega(1)$ based on the sub-Gaussian design

$$\Xi_{11} \ge \eta \left(C_1 \gamma^2 \cdot \|\mathbf{w}^{(t)} - \mathbf{w}^*\|^2 - \nu\epsilon \log(1/\epsilon) \cdot \|\mathbf{w}^{(t)} - \mathbf{w}^*\| \right),$$

Next, an application of Peter-Paul's inequality[1] gives us,

$$\Xi_{11} \gtrsim \eta \left(C_1 \gamma^2 \cdot \|\mathbf{w}^{(t)} - \mathbf{w}^*\|^2 - C_2 \gamma^{-2} \nu^2 \epsilon^2 \log^2(1/\epsilon) \right),$$

where γ is the gradient lower bound in Property 11. We next show there exists a positive constant $C_2 = O(1)$, such that

$$\Xi_{12} \le \eta^2 \left(C_3 L^4 \cdot \|\mathbf{w}^{(t)} - \mathbf{w}^*\|^2 + L^2 \nu^2 \epsilon^2 \log^2(1/\epsilon) \right),$$

where L is the Lipschitz constant in Property 10. We must now control the corrupted gradient term, Ξ_2. The key idea is to note that from the optimality of the hard thresholding step,

$$\sum\nolimits_{i \in S^{(t)} \cap E} (\sigma(\mathbf{w}^{(t)} \cdot \mathbf{x}_i) - y_i)^2 \le \sum\nolimits_{i \in G \setminus S^{(t)}} (\sigma(\mathbf{w}^{(t)} \cdot \mathbf{x}_i) - y_i)^2 \tag{2}$$

[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 $ab \le \frac{t^p a^p}{p} + \frac{t^{-q} b^q}{q}$. Consider young's Inequality and replace a with at^p and b with bt^{-p}.

because $|S^{(t)} \cap E| = |G \setminus S^{(t)}|$. We then prove the existence of a constant C_4 such that,

$$\Xi_2 \leq C_4 \eta \cdot L^2 \epsilon \log(1/\epsilon) \|\mathbf{w}^{(t)} - \mathbf{w}^*\|^2.$$

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

$$\|\mathbf{w}^{(t+1)} - \mathbf{w}^*\| \leq \left(1 - C_5 \eta \gamma^2\right) \cdot \|\mathbf{w}^{(t)} - \mathbf{w}^*\|_2 + L\nu\epsilon \log(1/\epsilon), \tag{3}$$

when we choose $\eta = O(\gamma^2/L^2)$. In Equation (3), we obtain a bound that is of the form,

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

We unroll the recurrence relation

$$\|\mathbf{w}^{(t)} - \mathbf{w}^*\|_2 \leq \|\mathbf{w}^{(0)} - \mathbf{w}^*\|_2 (1 - \lambda)^t + \sum_k^t (1 - \lambda)^k \Lambda.$$

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

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

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

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

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 12 that $\|\mathbf{w}^* - \mathbf{w}^{(0)}\| \leq \|\mathbf{w}^*\|$ with high probability, giving us the desired bound. ■

3.4.1 Algorithmic thresholding parameter

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 $|S_\epsilon \cap E| \geq |S_{\epsilon^*} \cap E|$ as $(1-\epsilon)n \geq (1-\epsilon^*)n$. We similarly have $|G \setminus S_{\epsilon^*}^{(t)}| \leq |G \setminus S_\epsilon^{(t)}|$ as the thresholded set is smaller in cardinality. We thus have our key step, Equation (2), will still hold when $\epsilon \geq \epsilon^*$ and we can thus obtain the same approximation bounds with only a lower bound on ϵ^*.

3.5 Learning ReLU neurons

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

Lemma 12 (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 probabilistic guarantees on the convergence of learning a ReLU neuron.

Theorem 13. *Let $X = [\mathbf{x}_1, \ldots, \mathbf{x}_n] \in \mathbb{R}^{d \times n}$ 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 norm K 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 $i \in G$. Then after $O\left(\kappa(\Sigma) \log\left(\frac{\|\mathbf{w}^*\|}{\varepsilon}\right)\right)$ gradient descent iterations and $n = \Omega\left(\frac{d + \log(1/\delta)}{\epsilon^2}\right)$, then with probability exceeding $1 - 3T\delta$, Algorithm 2 with learning rate $\eta = O\left(\frac{\lambda_{\min}(\Sigma)}{\lambda^2_{\max}(\Sigma)}\right)$ returns $\mathbf{w}^{(T)}$ such that*

$$\|\mathbf{w}^{(t)} - \mathbf{w}^*\| \leq O(\nu\epsilon \log(1/\epsilon)),$$

with probability exceeding $1 - 3T\delta$.

Proof. The proof is deferred to Appendix A.2. ■

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, as we require bounds on the spectra of empirical covariance matrices over the intersection of half-spaces. We also note that Lemma 12 implies that randomized restarts with high probability will return a vector with $O(\nu\epsilon\log(1/\epsilon))$ ℓ_2 approximation error. With access to an uncorrupted test set, using concentration of measure inequalities for sub-Gaussian distributions, it is possible to use adaptive algorithms to determine how many randomized restarts is enough to obtain $O(\nu\epsilon\log(1/\epsilon))$ 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 a linear combination 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.

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.

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.

Ankit Pensia, Varun Jog, and Po-Ling Loh. Robust regression with covariate filtering: Heavy tails and adversarial contamination. *arXiv preprint arXiv:2009.12976*, 2020.

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.

Banghua Zhu, Jiantao Jiao, and Jacob Steinhardt. Generalized resilience and robust statistics. *The Annals of Statistics*, 50(4):2256–2283, 2022.

A Proofs for learning nonlinear neurons

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

Notation. We first state some notational preliminaries. Partition the indices of the training data set S into the indices of the inliers, T, and the indices of the corrupted points, E. That is, let $[n] = G \cup E$ where $|G| = (1-\epsilon)n$ and $|E| = \epsilon n$, where $i \in G$ iff it was not modified by the adversary, and $i \in E$ iff $(\mathbf{x}_i, y_j)$ has been seen and possibly modified by the adversary. For $t \in [T]$, we denote $S^{(t)}$ as the set of indices selected from the hard thresholding at iteration t. We decompose $S^{(t)}$ into the indices of *True Positives* (inliers) and the indices of *False Positives* (corrupted samples), i.e. $S^{(t)} = (S^{(t)} \cap T) \cup (S^{(t)} \cap E) :== \mathrm{TP} \cup \mathrm{FP}$. We similarly decompose the points discarded at each iteration into *False Negatives* and *True Negatives*, i.e. $G \setminus S^{(t)} = ((G \setminus S^{(t)}) \cap T) \cup ((G \setminus S^{(t)}) \cap E) \coloneqq \mathrm{FN} \cup \mathrm{TN}$.

A.1 General neuron

In this section, we consider the general approximation bound for a nonlinear neuron that satisfies Properties 9, 10, and 11.

Proof of Theorem 8. Recall that $X \in \mathbb{R}^{d \times n}$ and $\mathbf{y} \in \mathbb{R}^n$. Noting that $\nabla \mathcal{R}(\mathbf{w}^{(t)}; \mathrm{S}^{(t)}) = \nabla \mathcal{R}(\mathbf{w}^{(t)}; \mathrm{TP}) + \nabla \mathcal{R}(\mathbf{w}^{(t)}; \mathrm{FP})$, we have

$$\|\mathbf{w}^{(t+1)} - \mathbf{w}^*\| \leq \|\mathbf{w}^{(t)} - \mathbf{w}^* - \eta \nabla \mathcal{R}(\mathbf{w}^{(t)}; \mathrm{TP})\| + \eta \cdot \|\nabla \mathcal{R}(\mathbf{w}^{(t)}; \mathrm{FP})\| \coloneqq T_1 + T_2.$$

We expand I through it's square,

$$T_1^2 = \|\mathbf{w}^{(t)} - \mathbf{w}^*\|^2 - 2\eta \cdot \langle \nabla \mathcal{R}(\mathbf{w}^{(t)}; \mathrm{TP}), \mathbf{w}^{(t)} - \mathbf{w}^* \rangle + \eta^2 \cdot \|\nabla \mathcal{R}(\mathbf{w}^{(t)}; \mathrm{TP})\|^2 \coloneqq T_{11}^2 - T_{12} + T_{13}^2.$$

We first consider a lower bound for I_2,

$$\begin{aligned} T_{12} &= \frac{4\eta}{(1-\epsilon)n} \cdot \left\langle \mathbf{w}^{(t)} - \mathbf{w}^*, \sum\nolimits_{\mathrm{TP}} (\sigma(\mathbf{w}^{(t)} \cdot \mathbf{x}_i) - y_i) \cdot \sigma'(\mathbf{w}^{(t)} \cdot \mathbf{x}_i) \cdot \omega_i \mathbf{x}_i \right\rangle \\ &= \frac{4\eta}{(1-\epsilon)n} \cdot \left\langle \mathbf{w}^{(t)} - \mathbf{w}^*, \sum\nolimits_{\mathrm{TP}} (\sigma(\mathbf{w}^{(t)} \cdot \mathbf{x}_i) - \sigma(\mathbf{w}^* \cdot \mathbf{x}_i) - \xi_i) \cdot \sigma'(\mathbf{w}^{(t)} \cdot \mathbf{x}_i) \cdot \omega_i \mathbf{x}_i \right\rangle \end{aligned}$$

From the mean value theorem (MVT), there exists a constant c_i, such that

$$\sigma(\mathbf{w}^{(t)} \cdot \mathbf{x}_i) - \sigma(\mathbf{w}^* \cdot \mathbf{x}_i) = \sigma'(c_i)(\mathbf{w}^{(t)} - \mathbf{w}^*) \cdot \mathbf{x}_i. \tag{4}$$

Let $\widetilde{X}_{\mathrm{TP}} \coloneqq X_{\mathrm{TP}} \mathrm{diag}(\boldsymbol{\omega})$. Then from Property 11 and Cauchy-Schwarz inequality,

$$T_{12} \geq \frac{4\eta}{(1-\epsilon)n} \left(\gamma^2 \lambda_{\min}(\widetilde{X}_{\mathrm{TP}} \widetilde{X}_{\mathrm{TP}}^\top) \cdot \|\mathbf{w}^{(t)} - \mathbf{w}^*\|^2 - \left\| \sum\nolimits_{\mathrm{TP}} \xi_i \sigma'(\mathbf{w}^{(t)} \cdot \mathbf{x}_i) \cdot \omega_i \mathbf{x}_i \right\| \cdot \|\mathbf{w}^{(t)} - \mathbf{w}^*\| \right).$$

From resilience and Lemma 26 (with Property 10),

$$T_{12} \gtrsim \eta \left(\gamma^2 \cdot \|\mathbf{w}^{(t)} - \mathbf{w}^*\|^2 - L\nu\epsilon \log(1/\epsilon) \cdot \|\mathbf{w}^{(t)} - \mathbf{w}^*\| \right),$$

then from an application of Peter-Paul's inequality

$$T_{12} \gtrsim \|\mathbf{w}^{(t)} - \mathbf{w}^*\|^2 \cdot \eta^2 \gamma^2 - L^2 \nu^2 \cdot \epsilon^2 \log^2(1/\epsilon),$$

with failure probability at most $2\mathrm{e}^{-d}$.

$$\begin{aligned} \nabla \mathcal{L}(\mathbf{w}^{(t)}; \mathrm{TP}) &= \sum\nolimits_{\mathrm{TP}} (\sigma(\mathbf{w}^{(t)} \cdot \mathbf{x}_i) - y_i) \cdot \sigma'(\mathbf{w}^{(t)} \cdot \mathbf{x}_i) \cdot \omega_i \mathbf{x}_i \\ &= \sum\nolimits_{\mathrm{TP}} (\sigma(\mathbf{w}^{(t)} \cdot \mathbf{x}_i) - \sigma(\mathbf{w}^* \cdot \mathbf{x}_i) - \xi_i) \cdot \sigma'(\mathbf{w}^{(t)} \cdot \mathbf{x}_i) \cdot \omega_i \mathbf{x}_i \end{aligned}$$

Then from an application of the triangle inequality,

$$\|\nabla\mathcal{L}(\mathbf{w}^{(t)};\mathrm{TP})\| \leq \left\|\sum\nolimits_{\mathrm{TP}}(\sigma(\mathbf{w}^{(t)}\cdot\mathbf{x}_i)-\sigma(\mathbf{w}^{*}\cdot\mathbf{x}_i))\cdot\omega_i\mathbf{x}_i\right\| + \left\|\sum\nolimits_{\mathrm{TP}}\xi_i\sigma'(\mathbf{w}^{(t)}\cdot\mathbf{x}_i)\cdot\omega_i\mathbf{x}_i\right\| \coloneqq T_{131}+T_{132}.$$

We find from an application of the MVT (see Equation (4)) that $T_{131}\leq L\|\widetilde{X}_{\mathrm{TP}}\widetilde{X}_{\mathrm{TP}}^{\top}\|\|\mathbf{w}^{(t)}-\mathbf{w}^{*}\|$ and from resilence $T_{131}\leq 2Ln\|\mathbf{w}^{(t)}-\mathbf{w}^{*}\|$. From Lemma 26, $T_{132}\leq Ln\nu\epsilon\log(1/\epsilon)$ with high probability.

$$T_{13}\lesssim \eta L^2\cdot\|\mathbf{w}^{(t)}-\mathbf{w}^{*}\|+\eta L\nu\epsilon\log(1/\epsilon),$$

with probability at least $1-2\mathrm{e}^{-d}$ when $n=\Omega(d/\epsilon)$. We now upper bound II using similar argument to our upper bound of the norm of $\nabla\mathcal{R}(\mathbf{w}^{(t)};\mathrm{TP})$,

$$\|\nabla\mathcal{R}(\mathbf{w}^{(t)};\mathrm{FP})\|\leq L\|X_{\mathrm{FP}}\|\left(\sum\nolimits_{\mathrm{FP}}(\sigma(\mathbf{w}^{(t)}\cdot\mathbf{x}_i)-y_i)^2\right)^{1/2}.$$

We note that $|\mathrm{FP}|=|\mathrm{FN}|$, and therefore

$$\sum\nolimits_{\mathrm{FP}}(\sigma(\mathbf{w}^{(t)}\cdot\mathbf{x}_i)-y_i)^2\leq\sum\nolimits_{\mathrm{FN}}(\sigma(\mathbf{w}^{(t)}\cdot\mathbf{x}_i)-y_i)^2.$$

We then obtain

$$\|\nabla\mathcal{R}(\mathbf{w}^{(t)};\mathrm{FP})\|\leq 2L^2\|X_{\mathrm{FP}}\|\|X_{\mathrm{FN}}\|\cdot\|\mathbf{w}^{(t)}-\mathbf{w}^{*}\|+2L\|X_{\mathrm{FP}}\|\|\boldsymbol{\xi}_{\mathrm{FN}}\|.$$

We then find

$$T_2\lesssim\eta\left(L^2\epsilon\log(1/\epsilon)\cdot\|\mathbf{w}^{(t)}-\mathbf{w}^{*}\|+L\cdot\nu\epsilon\log(1/\epsilon)\right),$$

with probability at least $1-2\mathrm{e}^{-d}$ when $n=\Omega(d/\epsilon)$.

$$\|\mathbf{w}^{(t+1)}-\mathbf{w}^{*}\|\leq\left(1-C\eta\gamma^2\right)\cdot\|\mathbf{w}^{(t)}-\mathbf{w}^{*}\|+\eta L\cdot\nu\epsilon\log(1/\epsilon).$$

We then obtain an error of $O((L/\gamma^2)\nu\epsilon\log(1/\epsilon))$. Our proof is complete. ■

A.2 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.

Proof of Theorem 13. We will now begin our standard analysis.

$$\|\mathbf{w}^{(t+1)}-\mathbf{w}^{*}\|_2\leq\|\mathbf{w}^{(t)}-\mathbf{w}^{*}-\eta\nabla\mathcal{R}(\mathbf{w}^{(t)};\mathrm{TP})\|_2+\|\eta\nabla\mathcal{R}(\mathbf{w}^{(t)};\mathrm{FP})\|_2\coloneqq\Xi_1+\Xi_2.$$

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

$$\Xi_1^2=\|\mathbf{w}^{(t)}-\mathbf{w}^{*}\|^2-2\eta\cdot\langle\mathbf{w}^{(t)}-\mathbf{w}^{*},\nabla\mathcal{R}(\mathbf{w}^{(t)};\mathrm{TP})\rangle+\eta^2\cdot\|\nabla\mathcal{R}(\mathbf{w}^{(t)};\mathrm{TP})\|^2\coloneqq\Xi_{11}^2-\Xi_{12}+\Xi_{13}^2.$$

We will first lower bound I_2.

$$\begin{aligned}\langle\mathbf{w}^{(t)}-\mathbf{w}^{*},\nabla\mathcal{L}(\mathbf{w}^{(t)};\mathrm{TP})\rangle &\stackrel{\mathrm{def}}{=}\left\langle\mathbf{w}^{(t)}-\mathbf{w}^{*},\sum\nolimits_{\mathrm{TP}}(\sigma(\mathbf{w}^{(t)}\cdot\mathbf{x}_i)-y_i)\cdot\mathbf{x}_i\cdot\mathbf{1}\{\mathbf{w}^{(t)}\cdot\mathbf{x}_i\geq 0\}\right\rangle\\ &=\left\langle\mathbf{w}^{(t)}-\mathbf{w}^{*},\sum\nolimits_{\mathrm{TP}}(\sigma(\mathbf{w}^{(t)}\cdot\mathbf{x}_i)-\sigma(\mathbf{w}^{*}\cdot\mathbf{x}_i)-\xi_i)\cdot\mathbf{x}_i\cdot\mathbf{1}\{\mathbf{w}^{(t)}\cdot\mathbf{x}_i\geq 0\}\right\rangle.\end{aligned}$$

We will first adopt the notation from Zhang et al. [2019], we define

$$\Sigma_{\mathrm{TP}}(\mathbf{w},\hat{\mathbf{w}})\coloneqq\sum\nolimits_{\mathrm{TP}}\mathbf{x}_i\mathbf{x}_i^{\top}\cdot\mathbf{1}\{\mathbf{x}_i\cdot\mathbf{w}\geq 0\}\cdot\mathbf{1}\{\mathbf{x}_i\cdot\hat{\mathbf{w}}\geq 0\}. \tag{5}$$

Using the notation from Equation (5), we obtain

$$\sum\nolimits_{\mathrm{TP}}(\sigma(\mathbf{w}^{(t)}\cdot\mathbf{x}_i)-\sigma(\mathbf{w}^{*}\cdot\mathbf{x}_i))\cdot\mathbf{x}_i\cdot\mathbf{1}\{\mathbf{w}^{(t)}\cdot\mathbf{x}_i\geq 0\}=\Sigma_{\mathrm{TP}}(\mathbf{w}^{(t)},\mathbf{w}^{(t)})\mathbf{w}^{(t)}-\Sigma_{\mathrm{TP}}(\mathbf{w}^{*},\mathbf{w}^{(t)})\mathbf{w}^{*}.$$

We have the following equivalence,

$$\Sigma_{\text{TP}}(\mathbf{w}^{(t)}, \mathbf{w}^{(t)})\mathbf{w}^{(t)} - \Sigma_{\text{TP}}(\mathbf{w}^*, \mathbf{w}^{(t)})\mathbf{w}^* = \Sigma_{\text{TP}}(\mathbf{w}^{(t)}, \mathbf{w}^*)(\mathbf{w}^{(t)} - \mathbf{w}^*) + \Sigma_{\text{TP}}(\mathbf{w}^{(t)}, -\mathbf{w}^*)\mathbf{w}^{(t)}.$$

We then observe that

$$\left\langle \mathbf{w}^{(t)} - \mathbf{w}^*, \Sigma_{\text{TP}}(\mathbf{w}^{(t)}, -\mathbf{w}^*)\mathbf{w}^{(t)} \right\rangle \geq 0,$$

as $\mathbf{w}^{(t)} \cdot \mathbf{x}_i \geq \mathbf{w}^* \cdot \mathbf{x}_i$ and $\mathbf{w}^{(t)} \cdot \mathbf{x}_i \geq 0$ when $\mathbf{1}\{\mathbf{w}^{(t)} \cdot \mathbf{x}_i \geq 0\} \cdot \mathbf{1}\{\mathbf{w}^* \cdot \mathbf{x}_i < 0\} = 1$. Then from Cauchy-Schwarz,

$$\Xi_{12} \geq \frac{4\eta}{(1-\epsilon)n} \left(\lambda_{\min}(\Sigma_{\text{TP}}(\mathbf{w}^{(t)}, \mathbf{w}^*)) \cdot \|\mathbf{w}^{(t)} - \mathbf{w}^*\|^2 - \left\| \sum\nolimits_{\text{TP}} \xi_i \mathbf{x}_i \cdot \mathbf{1}\{\mathbf{w}^{(t)} \cdot \mathbf{x}_i \geq 0\} \right\| \cdot \|\mathbf{w}^{(t)} - \mathbf{w}^*\| \right).$$

We will now consider a lower bound on the minimum eigenvalue for the covariance matrix in the intersection of two half-spaces. We have from Weyl's Inequality,

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

We bound the minimum value of the expected covariance in the intersection of half-spaces from Lemma 28,

$$\mathbf{E}\Big[\Sigma_{\text{TP}}(\mathbf{w}^{(t)}, \mathbf{w}^*)\Big] \succeq n(1-2\epsilon) \cdot \left(\frac{\pi - \Theta - \sin\Theta}{2\pi} \right) \cdot I.$$

Since $0 \leq \Theta \leq \pi/2$, the minimum eigenvalue is $\Omega(n)$. We then have from Lemma 27,

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

We then find there exists sufficiently small ϵ such that the minimimum eigenvalue of $\Sigma_{\text{TP}}(\mathbf{w}^{(t)}, \mathbf{w}^*)$ satisfies $\lambda_{\min}(\Sigma_{\text{TP}}(\mathbf{w}^{(t)}, \mathbf{w}^*)) \geq 1/4$. We now bound the second-moment matrix approximation.

$$\left\| \sum\nolimits_{\text{TP}} \xi_i \mathbf{x}_i \cdot \mathbf{1}\{\mathbf{w}^{(t)} \cdot \mathbf{x}_i \geq 0\} \right\| \leq \sqrt{4800\lambda_{\max}(\Sigma)} n\nu\epsilon \log(1/\epsilon),$$

with high probability from Proposition 25. We now upper bound Ξ_2.

$$\|\nabla \mathcal{L}(\mathbf{w}^{(t)}; \text{FP})\| \leq 2\|X_{\text{FP}}\|\|X_{\text{FN}}\| \cdot \|\mathbf{w}^{(t)} - \mathbf{w}^*\| + 2\|X_{\text{FP}}\|\|\boldsymbol{\xi}_{\text{FN}}\|.$$

From Proposition 17, we have that the maximum eigenvalues of X_{FP} and X_{FN} are $O(\sqrt{n\epsilon \log(1/\epsilon)})$ with probability with at least $1 - 2\mathrm{e}^{-d}$ when $n = \Omega(d/\epsilon)$. From Proposition 15, $\|\boldsymbol{\xi}_{\text{FN}}\| = O(\sqrt{\nu n\epsilon \log(1/\epsilon)})$ with probability at least e^{-d} when $n = \Omega(d)$. We then find that

$$\Xi_2 \lesssim \epsilon \log(1/\epsilon) \cdot \|\mathbf{w}^{(t)} - \mathbf{w}^*\| + \nu\epsilon \log(1/\epsilon).$$

We now conclude with our linear convergence result.

$$\|\mathbf{w}^{(t+1)} - \mathbf{w}^*\| \leq \|\mathbf{w}^{(t)} - \mathbf{w}^*\| \cdot (1 - C) + \nu\epsilon \log(1/\epsilon).$$

Solving for the recurrence, we find that $\|\mathbf{w}^{(T)} - \mathbf{w}^*\| = O(\nu\epsilon \log(1/\epsilon))$. Our proof is complete. ■

B Probability Theory

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

B.1 Chi-squared random variables

Lemma 14 (Upper bound on sum of Chi-squared variables [Laurent and Massart, 2000]). *Suppose $\xi_i \sim \mathcal{N}(0, \nu^2)$ for $1 \le i \le n$, then*

$$\mathbf{Pr}\{\|\boldsymbol{\xi}\|^2 \ge \nu^2(n + 2\sqrt{nx} + 2x)\} \le \mathrm{e}^{-x}.$$

Proposition 15 (Probabilistic Upper Bound on Sum of Chi-Squared Variables). *Suppose $\xi_i \sim \mathcal{N}(0, \nu^2)$ 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. Choose $\delta < 1$, then*

$$\max_{S \in \mathcal{W}} \|\boldsymbol{\xi}_S\|_2^2 \le 30\nu^2 n\epsilon \log(1/\epsilon),$$

with probability exceeding $1 - \delta$ when $n = \Omega(\log(1/\delta))$.

Proof of Proposition 15. Directly from Lemma 14, we have with probability exceeding $1 - \delta$.

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

We now can prove the claimed bound using the union bound,

$$\mathbf{Pr}\Big\{\max_{S \in \mathcal{W}} \|\boldsymbol{\xi}\|_2^2 \ge \nu^2(\epsilon n + 2\sqrt{\epsilon n x} + 2x)\Big\} \le \binom{\mathrm{e}}{\epsilon}^{\epsilon n} \mathbf{Pr}\{\|\boldsymbol{\xi}\|_2^2 \ge \nu^2(\epsilon n + 2\sqrt{\epsilon n x} + 2x)\} \le \binom{\mathrm{e}}{\epsilon}^{\epsilon n} \mathrm{e}^{-x}.$$

In the first inequality we apply a union bound over $\mathcal{W}$ with Lemma 29, and in the second inequality we use Lemma 14. We then obtain with probability exceeding $1 - \delta$,

$$\max_{S \in \mathcal{W}} \|\boldsymbol{\xi}_S\|_2^2 \le \nu\Big(\epsilon n + 2\sqrt{n\epsilon\log(1/\delta) + 3n^2\epsilon^2\log(1/\epsilon)} + 2\log(1/\delta) + 6n\epsilon\log(1/\epsilon)\Big).$$

We note that $\log\binom{n}{\epsilon n} \le 3n\epsilon\log(1/\epsilon)$ as $\epsilon < 0.5$ and $\sqrt{\log(1/\epsilon)} \le (\log(2))^{-1/2}\log(1/\epsilon) \le \sqrt{3}\log(1/\epsilon)$ when $\epsilon < 0.5$. When $n \ge \log(1/\delta)$, we find that

$$\max_{S \in \mathcal{W}} \|\boldsymbol{\xi}_S\|_2^2 \le 30\nu^2 n\epsilon\log(1/\epsilon),$$

with probability at least $1 - \delta$. The proof is complete. ■

B.2 Eigenvalue Concentration Inequalities

In this section, we derive concentration for the minimal and maximal eigenvalues of the sample covariance matrix with $n\epsilon$ samples removed by the adversary.

Lemma 16 (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, C_K, dependent on the sub-Gaussian norm such that,*

$$\lambda_{\max}(XX^\top) \le n \cdot \lambda_{\max}(\Sigma) + \lambda_{\max}(\Sigma) \cdot \Big(C_K\sqrt{dn} + t\sqrt{n}\Big),$$

with probablity at least $1 - 2\mathrm{e}^{-c_K t^2}$.

Proposition 17. *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 $\mathcal{W} \coloneqq \{S \subset [n] : |S| = \epsilon n\}$. Then there exist constants c_K, C_K, such that*

$$\max_{S \in \mathcal{W}} \lambda_{\max}(X_S X_S^\top) \le 10n\epsilon\log(1/\epsilon) \cdot \lambda_{\max}(\Sigma)$$
$$\min_{S \in \mathcal{W}} \lambda_{\min}(X_{[n]\setminus S} X_{[n]\setminus S}^\top) \ge 0.1n \cdot \lambda_{\min}(\Sigma)$$

with probability at least $1 - \delta$ when

$$n = \Omega\left(\frac{d + \log(1/\delta)}{\epsilon \cdot \lambda_{\min}^2(\Sigma)}\right) \text{ and } \epsilon \le 0.014 \cdot \kappa^{-5/3}(\Sigma).$$

Proof of Proposition 17. We will use a union bound to obtain our claimed error bound.

$$\mathbf{Pr}\left\{\max_{S\in\mathcal{W}}\lambda_{\max}(X_SX_S^\top)\geq n\epsilon\cdot\lambda_{\max}(\Sigma)+\lambda_{\max}(\Sigma)\cdot\left(C_K\cdot\sqrt{dn\epsilon}+t\sqrt{n\epsilon}\right)\right\}\leq 2\cdot\left(\frac{\mathrm{e}}{\epsilon}\right)^{\epsilon n}\mathrm{e}^{-c_Kt^2}.$$

In the above, we upper bound through a union bound over $\mathcal{W}$ with cardinality upper bounded in Lemma 29, and then apply Lemma 16. We then obtain with probability $1-\delta$,

$$\lambda_{\max}(X_SX_S^\top)\leq n\epsilon\cdot\lambda_{\max}(\Sigma)+\lambda_{\max}(\Sigma)\cdot\left(C_K\cdot\sqrt{dn\epsilon}+\sqrt{c_K^{-1}(n\epsilon\cdot\log(2/\delta)+3n^2\epsilon^2\log(1/\epsilon))}\right).$$

Then, when the sample size satisfies

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

the maximum eigenvalue satisfies, $\lambda_{\max}(X_SX_S^\top)\leq 10n\epsilon\log(1/\epsilon)$ and our proof of the upper bound for the maximal eigenvalue is complete. We now lower bound the minimum eigenvalue. We have from Weyl's inequality for any $S\in\mathcal{W}$,

$$\lambda_{\min}(X_{[n]\setminus S}X_{[n]\setminus S}^\top)=\lambda_{\min}(XX^\top-X_SX_S^\top)\geq\lambda_{\min}(XX^\top)-\lambda_{\max}(X_SX_S^\top)$$

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

$$\begin{aligned}\lambda_{\min}(X_{[n]\setminus S}X_{[n]\setminus S}^\top)&\geq n\cdot\lambda_{\min}(\Sigma)-C_K\cdot\sqrt{dn}-\sqrt{\frac{1}{c_K}\cdot n\cdot\log(2/\delta)}-\lambda_{\max}(\Sigma)\cdot(10n\epsilon\log(1/\epsilon))\\&\geq 0.9n\cdot\lambda_{\min}(\Sigma)-\lambda_{\max}(\Sigma)\cdot 10n\epsilon\log(1/\epsilon)\geq 0.1n\cdot\lambda_{\min}(\Sigma).\end{aligned}$$

In the above, the first inequality follows when $n\geq\frac{200}{\lambda_{\min}^2(\Sigma)}\left(C_K^2\cdot d+\frac{1}{c_K}\cdot\log(2/\delta)\right)$. From some algebra we find that $\epsilon\log(1/\epsilon)\leq\epsilon^{1-\mathrm{e}^{-1}}<\epsilon^{0.6}$. Therefore, the last inequality holds when $\epsilon\leq 0.014\kappa^{-5/3}(\Sigma)$ The proof is complete. ■

B.3 Sum of product of nonlinear random variables

We will first define the Orlicz norm for sub-Gaussian random variables.

Definition 18. *The sub-Gaussian norm of a random variable X is denoted as $\|X\|_{\psi_2}$, and is defined as*

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

We first note that Gaussian scalars are sub-Gaussian and $\|X\|_{\psi_2}=O(\nu)$ for $X\sim\mathcal{N}(0,\nu^2)$.

Lemma 19. *Let $X\sim\mathcal{N}(0,\nu^2)$, then $\|X\|_{\psi_2}=\sqrt{8/3}\nu$.*

Proof of Lemma 19. We have from Definition 18 that

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

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

$$\mathbf{E}\left[\exp(X^2/c^2)\right]=\frac{1}{\nu\sqrt{2\pi}}\int_{-\infty}^{\infty}\exp\left(X^2\left(\frac{1}{c^2}-\frac{1}{2\nu^2}\right)\right)\mathrm{d}X=\frac{1}{\nu\sqrt{2}}\left(\frac{1}{c^2}-\frac{1}{2\nu^2}\right)^{-1/2}=2.$$

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

We now will define the Orlicz norm for sub-exponential random variables.

Definition 20. *The sub-exponential norm of a random variable X is denoted as $\|X\|_{\psi_1}$, and is defined as*

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

Sub-exponential random variables appear frequently in our analysis from the consequence of the following lemma.

Lemma 21 (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 22 (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 $t \geq 0$,*

$$\mathbf{Pr}\left\{\left|\sum\nolimits_i a_i X_i\right| \geq t\right\} \leq 2\exp\left[-c\min\left(\frac{t^2}{K^2\|\mathbf{a}\|_2^2}, \frac{t}{K\|\mathbf{a}\|_\infty}\right)\right].$$

We are now ready to prove our main results of the section.

Lemma 23. *Suppose $\mathbf{x}_1, \ldots, \mathbf{x}_n$ are sampled i.i.d from a sub-Gaussian distribution with second-moment matrix Σ and sub-Gaussian proxy Γ. Suppose $\xi_1, \ldots, \xi_n$ are i.i.d sampled from $\mathcal{N}(0, \nu^2)$, then*

$$\left\|\sum\nolimits_i \xi_i f(\mathbf{w}\cdot\mathbf{x}_i)\mathbf{x}_i\right\| \leq C\|\Gamma\|\|f\|_\infty \nu\sqrt{nd\log(12R/\delta)},$$

with probability at least $1-\delta$.

Proof of Lemma 23. We use the following characterization of the spectral norm,

$$\left\|\sum\nolimits_S \xi_i f(\mathbf{w}\cdot\mathbf{x}_i)\mathbf{x}_i\right\| = \sup_{\|\mathbf{v}\|=1}\left|\sum\nolimits_S \xi_i f(\mathbf{w}\cdot\mathbf{x}_i)\mathbf{x}_i\cdot\mathbf{v}\right|.$$

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,

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

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 $\|f(\mathbf{w}\cdot\mathbf{x}_i)\mathbf{x}_i\cdot\mathbf{v}\|_{\psi_2} \lesssim \|\Gamma\|_2\|f\|_\infty$. We have $\|\xi_i\|_{\psi_2} = \sqrt{8/3}\nu$ from Lemma 19, then from Lemma 21, the random variable $\xi_i f(\mathbf{w}\cdot\mathbf{x}_i)\mathbf{x}_i\cdot\mathbf{v}$ is sub-exponential s.t. $\|\xi_i f(\mathbf{w}\cdot\mathbf{x}_i)\mathbf{x}_i\cdot\mathbf{v}\|_{\psi_1} \lesssim \|\Gamma\|_2\|f\|_\infty\nu$. 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

$$\mathbf{u}^* = \arg\max_{\mathbf{u}\in\mathcal{N}_1}\left|\sum\nolimits_S \xi_i f(\widetilde{\mathbf{w}}\cdot\mathbf{x}_i)\,\mathbf{x}_i\cdot\mathbf{u}\right|, \quad \mathbf{v}^* = \arg\max_{\mathbf{v}\in\mathbb{S}^{d-1}}\left|\sum\nolimits_S \xi_i f(\widetilde{\mathbf{w}}\cdot\mathbf{x}_i)\,\mathbf{x}_i\cdot\mathbf{v}\right|.$$

We then have from the Cauchy-Schwarz inequality,

$$\left|\sum\nolimits_S \xi_i f(\widetilde{\mathbf{w}}\cdot\mathbf{x}_i)\,\mathbf{x}_i\cdot(\mathbf{u}^*-\mathbf{v}^*)\right| \leq \varepsilon\cdot\left\|\sum\nolimits_S \xi_i f(\mathbf{w}\cdot\mathbf{x}_i)\,\mathbf{x}_i\right\|, \tag{6}$$

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

$$\left|\sum\nolimits_S \xi_i f(\widetilde{\mathbf{w}}\cdot\mathbf{x}_i)\mathbf{x}_i\cdot\mathbf{u}^*\right| \geq \left|\sum\nolimits_S \xi_i f(\widetilde{\mathbf{w}}\cdot\mathbf{x}_i)\mathbf{x}_i\cdot\mathbf{v}^*\right| - \left|\sum\nolimits_S \xi_i f(\widetilde{\mathbf{w}}\cdot\mathbf{x}_i)\mathbf{x}_i\cdot(\mathbf{u}^*-\mathbf{v}^*)\right|$$

Then from Equation (6),

$$\left|\sum\nolimits_S \xi_i f(\widetilde{\mathbf{w}} \cdot \mathbf{x}_i)\mathbf{x}_i \cdot \mathbf{u}^*\right| \geq (1-\varepsilon) \cdot \left|\sum\nolimits_S \xi_i f(\widetilde{\mathbf{w}} \cdot \mathbf{x}_i)\mathbf{x}_i \cdot \mathbf{v}^*\right|,$$

so from rearranging we obtain

$$\left\|\sum\nolimits_S \xi_i f(\widetilde{\mathbf{w}} \cdot \mathbf{x}_i)\mathbf{x}_i\right\| \leq \frac{1}{1-\varepsilon} \cdot \left|\sum\nolimits_S \xi_i f(\widetilde{\mathbf{w}} \cdot \mathbf{x}_i)\mathbf{x}_i \cdot \mathbf{u}^*\right| \tag{7}$$

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 $\mathcal{B}(\mathbf{w}^*, r)$, then from Bernstein's inequality (see Lemma 22)

$$\left|\sum\nolimits_S \xi_i f(\widetilde{\mathbf{w}} \cdot \mathbf{x}_i)\mathbf{x}_i \cdot \mathbf{u}^*\right| \leq C\|\Gamma\|\|f\|_\infty \nu \sqrt{nd\log(12r/\delta)},$$

with probability at least $1-\delta$. ■

Lemma 24. *Suppose $\mathbf{x}_1, \ldots, \mathbf{x}_n$ are sampled i.i.d from a sub-Gaussian distribution with second-moment matrix Σ and sub-Gaussian proxy Γ. Suppose $\xi_1, \ldots, \xi_n$ are i.i.d sampled from $\mathcal{N}(0, \nu^2)$. For any set $S \subset [n]$ such that $|S| \leq 2\epsilon n$ for $\epsilon < 1/2$*

$$\left\|\sum\nolimits_S \xi_i f(\mathbf{w} \cdot \mathbf{x}_i)\mathbf{x}_i\right\| \leq \sqrt{1200\nu\lambda_{\max}(\Sigma)}\|f\|_\infty n\epsilon\log(1/\epsilon),$$

with probability at least $1-2\mathrm{e}^{-d}$ when $n = \Omega(d/\epsilon)$.

Proof of Lemma 24. We use the following characterization of the two norm,

$$\left\|\sum\nolimits_S \xi_i f(\mathbf{w} \cdot \mathbf{x}_i)\mathbf{x}_i\right\| = \sup_{\mathbf{v} \in \mathbb{S}^{d-1}} \left|\sum\nolimits_S \xi_i f(\mathbf{w} \cdot \mathbf{x}_i)\mathbf{x}_i \cdot \mathbf{v}\right|.$$

Then, from Cauchy-Schwarz inequality,

$$\sup_{\mathbf{v} \in \mathbb{S}^{d-1}} \left|\sum\nolimits_S \xi_i f(\mathbf{w} \cdot \mathbf{x}_i)\mathbf{x}_i \cdot \mathbf{v}\right| \leq \|f\|_\infty \sup_{\mathbf{v} \in \mathbb{S}^{d-1}} \left(\sum\nolimits_S \xi_i^2\right)^{1/2} \left(\sum\nolimits_S \langle \mathbf{x}_i, \mathbf{v}\rangle^2\right)^{1/2}.$$

Then, when the sample size satistfies $n = \Omega(d/\epsilon)$,

$$\left\|\sum\nolimits_S \xi_i f(\mathbf{w} \cdot \mathbf{x}_i)\mathbf{x}_i\right\| \leq \sqrt{1200\lambda_{\max}(\Sigma)}\|f\|_\infty n\nu\epsilon\log(1/\epsilon),$$

with probability at least $1-2\mathrm{e}^{-d}$ (see Propositions 15 and 17). ■

Proposition 25. *Suppose $\mathbf{x}_1, \ldots, \mathbf{x}_n$ are sampled i.i.d from a sub-Gaussian distribution with second-moment matrix Σ and sub-Gaussian proxy Γ. Suppose $\xi_1, \ldots, \xi_n$ are i.i.d sampled from $\mathcal{N}(0, \nu^2)$. For any set $S \subset [n]$ such that $(1-2\epsilon)n \leq |S| \leq (1-\epsilon)n$ for $\epsilon < 1/2$*

$$\left\|\sum\nolimits_S \xi_i f(\mathbf{w} \cdot \mathbf{x}_i)\mathbf{x}_i\right\| \leq \sqrt{4800\lambda_{\max}(\Sigma)}\|f\|_\infty n\nu\epsilon\log(1/\epsilon),$$

with probability at least $1-2\mathrm{e}^{-d}-12rd^{-10}$.

Proof of Proposition 25. Noting that $S = T \setminus (T \setminus S) \coloneqq T \setminus R$,

$$\left\|\sum\nolimits_S \xi_i f(\mathbf{w} \cdot \mathbf{x}_i)\mathbf{x}_i\right\| \leq \left\|\sum\nolimits_T \xi_i f(\mathbf{w} \cdot \mathbf{x}_i)\mathbf{x}_i\right\| + \left\|\sum\nolimits_R \xi_i f(\mathbf{w} \cdot \mathbf{x}_i)\mathbf{x}_i\right\|,$$

where $|R| \leq 2\epsilon$. Then from Lemmas 23 and 24, we obtain

$$\left\|\sum\nolimits_S \xi_i f(\mathbf{w} \cdot \mathbf{x}_i)\mathbf{x}_i\right\| \leq \sqrt{4800\nu\lambda_{\max}(\Sigma)}\|f\|_\infty n\epsilon\log(1/\epsilon),$$

with high probability when $n = \Omega(d/\epsilon^2)$ with probability at least $1-2\mathrm{e}^{-d}-12rd^{-10}$. ■

B.4 Resilience

In this section, we consider results using the resilience property (see Definition 6) of the sub-Gaussian distribution.

Lemma 26. *Suppose $\mathbf{x}_1, \dots, \mathbf{x}_n$ satisfy resilience properties. Suppose $\xi_1, \dots, \xi_n$ are sampled i.i.d. from $\mathcal{N}(0, \nu^2)$. For any set $S \subset [n]$ such that $(1-2\epsilon)n \le |S| \le (1-\epsilon)n$ for $\epsilon < 1/2$*

$$\left\| \sum\nolimits_S \xi_i f(\mathbf{w} \cdot \mathbf{x}_i)\, \mathbf{x}_i \right\| \le C\nu n\epsilon \log(1/\epsilon),$$

with probability at least $1-2\delta$ when $n = \Omega\left(\frac{d+\log(1/\delta)}{\epsilon^2}\right)$.

Proof of Lemma 26. Noting that $S = T \setminus (T \setminus S) \coloneqq T \setminus R$,

$$\left\| \sum\nolimits_S \xi_i f(\mathbf{w} \cdot \mathbf{x}_i)\mathbf{x}_i \right\| \le \left\| \sum\nolimits_T \xi_i f(\mathbf{w} \cdot \mathbf{x}_i)\mathbf{x}_i \right\| + \left\| \sum\nolimits_R \xi_i f(\mathbf{w} \cdot \mathbf{x}_i)\mathbf{x}_i \right\| \coloneqq T_1 + T_2,$$

where $|R| \le 2\epsilon n$. From Lemma 23, $T_1 = O(\nu\sqrt{nd})$ with high probability.

$$T_2 \le \max_{\mathbf{v} \in \mathbb{S}^{d-1}} \|f\|_\infty \left(\sum\nolimits_R \xi_i^2 \right)^{1/2} \left(\sum\nolimits_R \langle \mathbf{x}_i, \mathbf{v} \rangle^2 \right)^{1/2}.$$

From Proposition 15, we have $\|\boldsymbol{\xi}_R\| = O(\nu\sqrt{\epsilon \log(1/\epsilon)})$ and from resilence $\|X_R X_R^\top\|^{1/2} = O(\sqrt{\epsilon \log(1/\epsilon)})$. We thus find that $T_2 = O(\nu\epsilon \log(1/\epsilon))$ with probability at least $1-2\delta$ when $n = \Omega\left(\frac{d+\log(1/\delta)}{\epsilon^2}\right)$. ■

B.5 Covariance matrix estimation in an intersection of half-spaces

Lemma 27. *Fix $\mathbf{w}^* \in \mathbb{R}^{d-1}$ and suppose $\mathbf{w} \in \mathcal{B}(\mathbf{w}^*, r)$ for a constant $r < \|\mathbf{w}^*\|$. Suppose $\mathbf{x}_1, \dots, \mathbf{x}_n$ are sampled i.i.d from a rotationally invariant sub-Gaussian distribution, $\mathcal{D}$, with second-moment matrix Σ and sub-Gaussian proxy Γ. For any $\mathrm{S} \subset [n]$ s.t. $|\mathrm{S}| \le (1-\epsilon)n$,*

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

with probability at least $1-\delta$ when $n = \Omega(d \log(r/\delta)/\epsilon)$.

Proof of Lemma 27. Recall the notation $\Sigma_{\mathrm{S}}(\mathbf{w}^{(1)}, \mathbf{w}^{(2)}) \coloneqq \sum_{\mathrm{S}} \mathbf{x}_i \mathbf{x}_i^\top \cdot \mathbf{1}\{\mathbf{w}^{(1)} \cdot \mathbf{x}_i \ge 0\} \cdot \mathbf{1}\{\mathbf{w}^{(2)} \cdot \mathbf{x}_i \ge 0\}$. Let $\mathcal{N}_1$ be an ε-cover of $\mathcal{B}(\mathbf{w}^*, r)$ and $\mathcal{N}_2$ be an ε-cover of $\mathbb{S}^{d-1}$. Let $\widetilde{\mathbf{w}} \coloneqq \arg\min_{\mathbf{v} \in \mathcal{N}_1} \|\mathbf{w} - \mathbf{v}\|_2$ for each $\mathbf{w} \in \mathcal{B}(\mathbf{w}^*, r)$. We will use the decomposition given in Theorem 1 of Mei et al. [2016] to obtain

$$\begin{aligned}
&\mathbf{Pr}\left\{ \max_{\mathrm{S} \in \mathcal{W}} \sup_{\mathbf{w} \in \mathcal{B}(\mathbf{w}^*; r)} \|\Sigma_{\mathrm{S}}(\mathbf{w}, \mathbf{w}^*) - \mathbf{E}_{\mathbf{x} \sim \mathcal{D}}[\Sigma_{\mathrm{S}}(\mathbf{w}, \mathbf{w}^*)]\|_2 \ge t \right\} \\
&\le \mathbf{Pr}\left\{ \max_{\mathrm{S} \in \mathcal{W}} \sup_{\mathbf{w} \in \mathcal{B}(\mathbf{w}^*; r)} \|\Sigma_{\mathrm{S}}(\mathbf{w}, \mathbf{w}^*) - \Sigma_{\mathrm{S}}(\widetilde{\mathbf{w}}, \mathbf{w}^*)\|_2 \ge \frac{t}{3} \right\} \\
&\quad + \mathbf{Pr}\left\{ \max_{\mathrm{S} \in \mathcal{W}} \max_{\widetilde{\mathbf{w}} \in \mathcal{N}_1} \|\Sigma_{\mathrm{S}}(\widetilde{\mathbf{w}}, \mathbf{w}^*) - \mathbf{E}_{\mathbf{x} \sim \mathcal{D}}[\Sigma_{\mathrm{S}}(\widetilde{\mathbf{w}}, \mathbf{w}^*)]\|_2 \ge \frac{t}{3} \right\} \\
&\quad + \mathbf{Pr}\left\{ \sup_{\mathbf{w} \in \mathcal{B}(\mathbf{w}^*; r)} \|\mathbf{E}_{\mathbf{x} \sim \mathcal{D}}[\Sigma_{\mathrm{S}}(\widetilde{\mathbf{w}}, \mathbf{w}^*)] - \mathbf{E}_{\mathbf{x} \sim \mathcal{D}}[\Sigma_{\mathrm{S}}(\mathbf{w}, \mathbf{w}^*)]\|_2 \ge \frac{t}{3} \right\} \coloneqq T_1 + T_2 + T_3.
\end{aligned}$$

We first consider an upper bound bound for T_1.

$$\Sigma_{\mathrm{S}}(\mathbf{w}, \mathbf{w}^*) - \Sigma_{\mathrm{S}}(\widetilde{\mathbf{w}}, \mathbf{w}^*) \preceq \sum\nolimits_{\mathrm{S}} \mathbf{x}_i \mathbf{x}_i^\top \left(\mathbf{1}\{\mathbf{w} \cdot \mathbf{x}_i \ge 0\} - \mathbf{1}\{\widetilde{\mathbf{w}} \cdot \mathbf{x}_i \ge 0\} \right) \cdot \mathbf{1}\{\mathbf{w}^* \cdot \mathbf{x}_i \ge 0\}.$$

Then,

$$\|\Sigma_{\mathrm{S}}(\mathbf{w}, \mathbf{w}^*) - \Sigma_{\mathrm{S}}(\widetilde{\mathbf{w}}, \mathbf{w}^*)\| \le \left(\max\nolimits_{i \le n} \|\mathbf{x}_i\|^2 \right) \cdot \sum\nolimits_i |\mathbf{1}\{\mathbf{w} \cdot \mathbf{x}_i \ge 0\} - \mathbf{1}\{\widetilde{\mathbf{w}} \cdot \mathbf{x}_i \ge 0\}|. \tag{8}$$

Let $D_i \coloneqq |\mathbf{1}\{\mathbf{w}\cdot\mathbf{x}_i \geq 0\} - \mathbf{1}\{\widetilde{\mathbf{w}}\cdot\mathbf{x}_i \geq 0\}|$, using the fact that $\mathcal{D}$ is rotationally invariant, we find that $\mathbf{Pr}(D_i = 1) = \Theta(\mathbf{w}, \widetilde{\mathbf{w}})/\pi$. Then we note that $\Theta(\mathbf{w}, \widetilde{\mathbf{w}}) \leq \arcsin\left(\|\mathbf{w} - \widetilde{\mathbf{w}}\|/\|\mathbf{w}\|\right)$ and $\arcsin\left(\|\mathbf{w} - \widetilde{\mathbf{w}}\|/\|\mathbf{w}\|\right) \leq \varepsilon/\|\mathbf{w}\|$ when $\|\mathbf{w} - \widetilde{\mathbf{w}}\| \leq \|\mathbf{w}\|$. Therefore,

$$\sum\nolimits_i |\mathbf{1}\{\mathbf{w}\cdot\mathbf{x}_i \geq 0\} - \mathbf{1}\{\widetilde{\mathbf{w}}\cdot\mathbf{x}_i \geq 0\}| \leq n\left(\frac{\varepsilon}{\|\mathbf{w}\|}\right) + \sqrt{n\log(1/\delta_1)}, \tag{9}$$

with probability at least δ_1. We have the following inequality for sub-Gaussian vectors, $\mathbf{Pr}(\|\mathbf{x}_i\|^2 \geq C\|\Gamma\|(d + t)) \leq \mathrm{e}^{-dt}$. Applying a union bound over the n samples,

$$\max\nolimits_{i\leq n} \|\mathbf{x}_i\|^2 \leq C\|\Gamma\|(d + \log(n/\delta_2)),$$

with probability at least $1 - \delta_2$. We use our results from Equations (8) and (9) and find that $T_2 \leq \delta/3$ when $t = O(\lambda_{\max}(\Gamma)d\sqrt{n})$ after setting $\varepsilon_1 = O(\|\mathbf{w}\|/\sqrt{n})$. For the second term, let $\widetilde{\mathbf{x}} = \mathbf{x}\cdot\mathbf{1}\{\widetilde{\mathbf{w}}\cdot\mathbf{x} \geq 0\}\cdot\mathbf{1}\{\mathbf{w}^*\cdot\mathbf{x} \geq 0\}$ and $\widetilde{X}$ represent the matrix consisting of $\mathbf{x}_i$ that are in the half-space. We then have from the same arguments in Equation (7),

$$T_2 \leq \mathbf{Pr}\left\{\max_{\mathrm{S}\in\mathcal{W}}\max_{\widetilde{\mathbf{w}}\in\mathcal{N}_1}\max_{\mathbf{v}\in\mathcal{N}_2} |\|\widetilde{X}_{\mathrm{S}}^\top\mathbf{v}\|_2^2 - \mathbf{E}_{\mathbf{x}\sim\mathcal{D}}\|\widetilde{X}_{\mathrm{S}}^\top\mathbf{v}\|_2^2| \geq \frac{2t}{3}\right\},$$

when $\mathcal{N}_2$ is a $1/2$-net of $\mathbb{S}^{d-1}$. 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 $\||\widetilde{\mathbf{x}}\cdot\mathbf{v}|^2 - \mathbf{E}_{\mathbf{x}\sim\mathcal{D}}|\widetilde{\mathbf{x}}\cdot\mathbf{v}|^2\|_{\psi_1} \leq 16\|\Gamma\|_2$, we can then apply Bernstein's Inequality (see Lemma 22). We then find that $T_2 \leq \delta/2$ when $t = \Omega(\lambda_{\max}(\Gamma)n\epsilon\log(1/\epsilon))$ and $n = \Omega(d\log(r)/\epsilon)$. We now consider the third term.

$$\begin{aligned}
T_3 &\leq \mathbf{Pr}\left\{n\sup_{\mathbf{w}\in\mathcal{B}(\mathbf{w}^*;R)} \|\mathbf{E}_{\mathbf{x}\sim\mathcal{D}}[\mathbf{x}\mathbf{x}^\top\cdot(\mathbf{1}\{\widetilde{\mathbf{w}}\cdot\mathbf{w} \geq 0\} - \mathbf{1}\{\mathbf{w}\cdot\mathbf{x} \geq 0\})\cdot\mathbf{1}\{\mathbf{w}^*\cdot\mathbf{x} \geq 0\}]\|_2 \geq \frac{t}{3}\right\} \\
&\overset{(i)}{\leq} \mathbf{Pr}\left\{n\sup_{\mathbf{w}\in\mathcal{B}(\mathbf{w}^*;R)} \mathbf{E}_{\mathbf{x}\sim\mathcal{D}}\left[\|\mathbf{x}\mathbf{x}^\top\|_2|\mathbf{1}\{\widetilde{\mathbf{w}}\cdot\mathbf{x} \geq 0\} - \mathbf{1}\{\mathbf{w}\cdot\mathbf{x} \geq 0\}|\right] \geq \frac{t}{3}\right\} \\
&\overset{(ii)}{\leq} \mathbf{Pr}\left\{n\sup_{\mathbf{w}\in\mathcal{B}(\mathbf{w}^*;r)} \left(\mathbf{E}_{\mathbf{x}\sim\mathcal{D}}\|\mathbf{x}\mathbf{x}^\top\|_2^2\,\mathbf{E}_{\mathbf{x}\sim\mathcal{D}}|\mathbf{1}\{\widetilde{\mathbf{w}}\cdot\mathbf{x} \geq 0\} - \mathbf{1}\{\mathbf{w}\cdot\mathbf{x} \geq 0\}|\right)^{1/2} \geq \frac{2t}{3}\right\}
\end{aligned}$$

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

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

Then from our choice of $\varepsilon_1 = O(\|\mathbf{w}\|/\sqrt{n})$ and $t = O(\lambda_{\max}(\Gamma)n\epsilon\log(1/\epsilon))$, we find that $T_3 = 0$. We then find when $n = \Omega(d\log(r/\delta)/\epsilon)$,

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

with probability at least $1 - 2\delta$. Our proof is complete. ■

Lemma 28. *Suppose* $\mathbf{E}_{\mathbf{x}\sim\mathcal{D}}[\mathbf{x}] = \mathbf{0}$ *and* $\mathbf{E}_{\mathbf{x}\sim\mathcal{D}}\left[\mathbf{x}\mathbf{x}^\top\right] = I$ *where* $\mathcal{D}$ *is a rotationally invariant distribution. Fix* $\mathbf{w}_1, \mathbf{w}_2 \in \mathbb{R}^d$ *and define* Θ *as the angle between* $\mathbf{w}_1, \mathbf{w}_2$ *such that* $0 < \Theta \leq \pi/2$. *Then,*

$$\mathbf{E}_{\mathbf{x}\sim\mathcal{D}}\left[\mathbf{x}\mathbf{x}^\top\cdot\mathbf{1}\{\mathbf{w}_1\cdot\mathbf{x} \geq \mathbf{0}\}\cdot\mathbf{1}\{\mathbf{w}_2\cdot\mathbf{x} \geq \mathbf{0}\}\right] \succeq \left(\frac{\pi - \Theta - \sin\Theta}{2\pi}\right)\cdot I.$$

Proof of Lemma 28. Since $\mathcal{D}$ is an isotropic distribution, we have that $U\mathbf{x} \sim \mathcal{D}$ for any unitary U. Consider U with basis spanning $\{\mathbf{w}_1, (I - \mathrm{Proj}_{\mathbf{w}_1})\mathbf{w}_2, \mathbf{u}_3, \ldots, \mathbf{u}_d\}$, where $\mathbf{w}_1, \mathbf{w}_2 \perp \mathrm{span}\{\mathbf{u}_3, \ldots, \mathbf{u}_d\}$. Then we find that $U\mathbf{w}_1 = \mathbf{e}_1$ and $U\mathbf{w}_2 = (\cos\Theta, \sin\Theta, 0, \ldots, 0)$. We thus obtain

$$\mathbf{E}_{\mathbf{x}\sim\mathcal{D}}\left[\mathbf{x}\mathbf{x}^\top\cdot\mathbf{1}\left\{\mathbf{w}_1\cdot\mathbf{x} \geq \mathbf{0}\right\}\cdot\mathbf{1}\{\mathbf{w}_2\cdot\mathbf{x} \geq \mathbf{0}\}\right] = \mathbf{E}_{\mathbf{x}\sim\mathcal{D}}\left[\mathbf{x}\mathbf{x}^\top\cdot\mathbf{1}\{x_1 \geq 0\}\cdot\mathbf{1}\{x_1\cos\Theta + x_2\sin\Theta \geq 0\}\right].$$

We now consider the principal two by two matrix of the covariance,

$$\begin{aligned} M &\coloneqq \left[\mathbf{E}_{\mathbf{x}\sim\mathcal{D}}\left[\mathbf{x}\mathbf{x}^\top \cdot \mathbf{1}\{x_1 \geq 0\} \cdot \mathbf{1}\{x_1 \cos\Theta + x_2 \sin\Theta \geq 0\}\right]\right]_{2,2} \\ &= \frac{1}{2\pi}\int_0^\infty \int_{\Theta-\pi/2}^{\pi/2} \begin{pmatrix} \cos^2\phi & \cos\phi\sin\phi \\ \cos\phi\sin\phi & \sin^2\phi \end{pmatrix} \mathrm{e}^{-r^2/2} r^3 \,\mathrm{d}\phi\,\mathrm{d}r \\ &= \frac{1}{\pi}\begin{pmatrix} \pi - \Theta + \cos\Theta\sin\Theta & \sin^2\Theta \\ \sin^2\Theta & \pi - \Theta - \cos\Theta\sin\Theta \end{pmatrix}. \end{aligned}$$

Solving for the eigenvalues of the matrix, we find that $\lambda_{\min}(M) = (\pi - \Theta - \sin\Theta)/\pi$. We then find the eigenvalues of the covariance matrix are $\{(\pi - \Theta \pm \sin\Theta)/\pi, (\pi - \Theta)/2\pi)\}$. ■

C Mathematical Tools

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

Lemma 29 (Sum of Binomial Coefficients [Cormen et al., 2022])**.** *Let $k, n \in \mathbb{N}$ such that $k \leq n$, then*

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

Lemma 30 (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,*

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