Learning and Generalization in Overparameterized Neural Networks, Going Beyond Two Layers

Zeyuan Allen-Zhu
zeyuan@csail.mit.edu
Microsoft Research AI

Yuanzhi Li
yuanzhil@stanford.edu
Stanford University

Yingyu Liang
yliang@cs.wisc.edu
University of Wisconsin-Madison

November 13, 2018*

Abstract

Neural networks have great success in many machine learning applications, but the fundamental learning theory behind them remains largely unsolved. Learning neural networks is NP-hard, but in practice, simple algorithms like stochastic gradient descent (SGD) often produce good solutions. Moreover, it is observed that overparameterization — designing networks whose number of parameters is larger than statistically needed to perfectly fit the data — improves both optimization and generalization, appearing to contradict traditional learning theory.

In this work, we extend the theoretical understanding of two and three-layer neural networks in the overparameterized regime. We prove that, using overparameterized neural networks, one can (improperly) learn some notable hypothesis classes, including two and three-layer neural networks with fewer parameters. Moreover, the learning process can be simply done by SGD or its variants in polynomial time using polynomially many samples. We also show that for a fixed sample size, the generalization error of the solution found by some SGD variant can be made almost independent of the number of parameters in the overparameterized network.

*Authors sorted in alphabetical order.

1 Introduction

Neural network learning has become a key machine learning approach and has achieved remarkable success in a wide range of real-world domains, such as computer vision, speech recognition, and game playing [KSH12, HZRS16, GMH13, SHM+16]. In contrast to the widely accepted empirical success, much less theory is known. Despite a recent increase of theoretical studies, many questions remain largely open, including fundamental ones about the optimization and generalization in learning neural networks.

A neural network of L layers is a function defined via layers of neurons: neurons in layer 0 are the coordinates of the input x; neurons in the subsequent layers 1 to L take a linear combination of the output of the previous layer, and then apply an activation function; the output y of the neural network is those of the neurons in the last layer. The weights in the linear combinations in all layers are called parameters of the network, and layers 1 to $L-1$ are called hidden layers. In the problem of learning neural networks, given training data $\{(x_i, y_i) : 1 \le i \le N\}$ where x_i is i.i.d. from some unknown distribution and y_i is the label, the goal is to find a network with a small generalization error with respect to some prescribed loss function.

One key challenge in analyzing the learning of neural networks is that the corresponding optimization is non-convex and is theoretically hard in the general case [ZLWJ17, Sha18]. This is in sharp contrast to the fact that simple optimization algorithms like stochastic gradient descent (SGD) and its variants usually produce good solutions in practice. One of the empirical tricks to overcome the learning difficulty is to use neural networks that are heavily overparameterized [ZBH+17]: the number of parameters is usually larger than the number of the training samples. Unlike traditional convex models, overparameterization for neural networks actually improves both training speed and generalization. For example, it is observed in [LSS14] that on synthetic data generated from a ground-truth network, SGD converges faster when the learnt network has more parameters than the ground truth. Perhaps more interestingly, in [AGNZ18] it was found that overparameterized networks learnt in practice can often be compressed to simpler ones with much fewer parameters, without hurting their ability to generalize; however, directly learning such simpler networks runs into worse results due to the optimization difficulty.

The practical findings suggest that, albeit optimizing a neural network along can be computationally expensive, using an overparameterized neural network as an *improper* learner for some simpler hypothesis class, especially smaller neural networks, might not actually be difficult. Towards this end, the following question is of great interest both in theory and in practice:

Question 1: *Can overparameterized networks be used as* efficient *improper learners for neural networks of fewer parameters or simpler structures?*

Improper learners are common in theory literatures. Polynomials are often used as improper learners for various purposes such as learning DNF and density estimation (e.g., [KS04, ADLS17]). Several recent works also study using kernels as improper learners for neural networks [LSS14, ZLJ16, Dan17].

However, in practice, multi-layer neural networks with the rectified linear unit (ReLU) activation function have been the dominant learners across vastly different domains. It is known that some other activation functions, especially smooth ones, can lead to provable learning guarantees. For example, [APVZ14] uses a two-layer neural network with exponential activation to learn polynomials. To the best of our knowledge, the practical universality of the non-smooth ReLU activation is still not well-understood. This motivates us to study ReLU networks.

Recently, some progress has been made towards understanding, how in two-layer networks with ReLU activations, overparameterization can make the learning process easier. In particular, [BGMS17] shows that such networks can learn linearly-separable data using just SGD. [LL18] shows that SGD learns a network with good generalization when the data come from mixtures of well-separated distributions. [DZPS18] shows that gradient descent can perfectly fit the training samples when the data is not degenerated (i.e., a matrix related to the input x and the ReLU activation has eigenvalues greater than 0). These results are only for two layers and only applicable to structured data. This leads to the following natural question:

Question 2: *Can overparameterization* simplify *the learning process* without *any structural assumptions about the input distribution?*

Most existing works analyzing the learning process of neural networks [Kaw16, SC16, XLS16, GLM17, SJL17, Tia17, BG17, ZSJ+17, LY17, BL17] need to make unrealistic assumptions about the data (such as being random Gaussian) and/or have strong assumptions about the network (such as using linear activations). A theorem *without* distributional assumptions about the data is often more desirable. Indeed, how to obtain a result that does not depend on the data distribution, but only on the hypothesis class itself, lies in the center of *PAC-learning* which is one of the foundations of machine learning theory [Val84].

Following these questions, we also note that determining the exact amount of overparameterization can be challenging without a clear knowledge of the ground truth. In practice, researchers usually create networks with a *significantly large* number of parameters, and surprisingly, the generalization error often does not increase. Thus, we would also like to understand the following question:

Question 3: *Can overparameterized networks be learnt to a small* generalization *error, using a number of samples that is (almost) independent of the number of parameters?*

This question cannot be studied under the traditional VC-dimension learning theory, since in principle, the VC dimension of the network grows with the number of parameters. Recently, several works [BFT17, NBMS17, AGNZ18, GRS18] explain generalization in the overparameterized setting by studying some other "complexity" of the learnt neural networks. Most related to the discussion here is [BFT17], where the authors prove a generalization bound in terms of the norms (of weight matrices) of each layer, as opposed to the number of parameters. However, their norms are "sparsity induced norms": in order for the norm not to scale with the number of hidden neurons m, essentially, it requires the number of neurons with non-zero weights *not* to scale with m. This more or less reduces the problem to the non-overparameterized case. More importantly, it is not clear from these results how a network with such low "complexity" and a good training loss can be produced by the training method.

1.1 Our Results

In this work, we extend the theoretical understanding of neural networks both in the algorithmic and the generalization perspectives. We give positive answers to the above questions for networks of two and *three* layers. We prove that when the network is sufficiently overparameterized, simple optimization algorithms (SGD or its variants) can learn ground-truth networks with a small generalization error in polynomial time using polynomially many samples.

To state our result in a simple form, we assume that there exists a (two or three-layer) *ground-truth network* with generalization error ε_0, and show that one can learn this hypothesis class, up to generalization error $\varepsilon_0 + \varepsilon$, using larger (two or three-layer) networks of size greater than a fixed polynomial in the size of the ground truth, in $1/\varepsilon$, and in the "complexity" of the activation function used in the ground truth. Furthermore, the sample complexity is also polynomial in these parameters, and only poly-logarithmic in the size of the overparameterized network. Our result is proved for networks with the ReLU activation where instead the ground truth can use arbitrary smooth activation functions.

Furthermore, unlike the two-layer case (so there is only one hidden layer) the optimization landscape in the overparameterized regime is almost convex [LL18, DZPS18], our result on *three-layer networks* gives the first theoretical proof that learning neural networks, even when there are sophisticated *non-convex* interactions between hidden layers, might still be non-difficult, as long as sufficient overparameterization is provided. This gives answers to the fundamental questions about the generalization and algorithmic aspects of neural network learning. Since practical neural networks are heavily overparameterized, our results may also provide theoretical insights to networks used in various applications.

We highlight a few interesting conceptual findings we used to derive our main result:

- In the overparameterized regime, *every neuron matters*. During training, information is *spread out* among all the neurons instead of collapsed into a few neurons. With this structure, we can prove a new generalization bound that is (almost) independent of the number of neurons, even when all neurons have non-negligible contributions to the output.
- In the overparameterized regime, good networks with small generalization errors are *almost everywhere*. With high probability over the random initialization, there exists a good network in the "close" neighborhood of the initialization.
- In the overparameterized regime, if one stays close enough to the random initialization, the learning process is tightly coupled with that of a "pseudo network" which has a more benign optimization landscape.

Combining the second and the third items leads to the convergence of the optimization process, and combining that with the first item gives our generalization result.

Roadmap. We formally define our (improper) learning problem in Section 2, and introduce notations in Section 3. In Section 4 we present our main theorems and give some examples. Our Section 5 in the main body summarizes our main proof ideas for our three-layer network results, and details are in Appendix B, C, D, E and F. Our two-layer proof is much easier and is included in Appendix G.

2 Problem and Assumptions

We consider learning some unknown distribution $\mathcal{D}$ of data points $z = (x, y) \in \mathbb{R}^d \times \mathbb{R}$, where $x \in \mathbb{R}^d$ is the input data point and y is the label associated with this data point. Without loss of generality, we restrict our attention to the distributions $\mathcal{D}$ where each data point $x \in \mathbb{R}^d$ in $\mathcal{D}$ is of unit Euclidean norm and satisfies $x_d = \frac{1}{2}$.[1]

We consider a loss function $L\colon \mathbb{R}^k \times \mathbb{R} \to [0, 1]$ such that for every label $y \in \mathbb{R}$, the loss function $L(\cdot, y)$ is convex, 1-Lipschitz continuous and 1-Lipschitz smooth. Assume that there exists a ground-truth function $F^* = (f_1^*, \cdots, f_k^*)\colon \mathbb{R}^d \to \mathbb{R}^k$ and some $\varepsilon_0 \geq 0$ so that

$$\mathbb{E}_{(x,y)\sim\mathcal{D}}\big[L(F^*(x), y)\big] \leq \varepsilon_0 \ . \tag{1}$$

[1] This is without loss of generality, since $\frac{1}{2}$ can always be padded to the last coordinate, and $\|x\|_2 = 1$ can always be ensured from $\|x\|_2 \leq 1$ by padding $\sqrt{1 - \|x\|_2^2}$ to the second-last coordinate. We make this assumption to simplify our notations: for instance, $x_d = \frac{1}{2}$ allows us to focus only on ground-truth networks without bias.

Our goal is to learn a neural network $F = (f_1, \cdots, f_k) \colon \mathbb{R}^d \to \mathbb{R}^k$ for a given $\varepsilon \geq 0$, satisfying

$$\mathbb{E}_{(x,y)\sim\mathcal{D}} L(F(x), y) \leq \varepsilon_0 + \varepsilon \ , \tag{2}$$

using a data set $\mathcal{Z} = \{z_1, \ldots, z_N\}$ consisting of N i.i.d. samples from the distribution $\mathcal{D}$.

In this paper, we consider F being equipped with the ReLU activation function: $\sigma(x) \stackrel{\text{def}}{=} \max\{0, x\}$. It is arguably the most widely-used activation function in practice. We assume F^* uses arbitrary smooth activation functions.

Below, we describe the details of the ground truth, our network, and the learning process for the cases of two and three layers, respectively.

2.1 Two Layer Networks

Ground truth $F^*(x)$. The ground truth $F^* = (f_1^*, \cdots, f_k^*)$ for our two-layer case is

$$f_r^*(x) = \sum_{i=1}^{p} a_{r,i}^* \phi_i(\langle w_{1,i}^*, x\rangle)\langle w_{2,i}^*, x\rangle \tag{3}$$

where each $\phi_i \colon \mathbb{R} \to \mathbb{R}$ is infinite-order smooth, $w_{1,i}^*, w_{2,i}^* \in \mathbb{R}^d$ are *ground-truth weight vectors*, and $a_{r,i}^* \in \mathbb{R}$ are weights. Without loss of generality, we assume $\|w_{1,i}^*\|_2 = \|w_{2,i}^*\|_2 = 1$ and $|a_{r,i}^*| \leq 1$.

Remark 2.1. Standard two-layer networks are only special cases of our formulation (3). Indeed, since $x_d = 1/2$, if we set $w_{2,i}^* = (0, \ldots, 0, 1)$ then

$$2f_r^*(x) = \sum_{i=1}^{p} a_{r,i}^* \phi_i(\langle w_{1,i}^*, x\rangle) \tag{4}$$

is a two-layer network with activation functions $\{\phi_i\}_{i\in[p]}$. Our formulation (3) allows for more functions, and in particular, and capture combinations of correlations between *non-linear* and linear measurements of different directions of x.

Our network $F(x; W)$. Our (improper) learners are two-layer networks $F = (f_1, f_2, \ldots, f_k)$ with

$$f_r(x) = \sum_{i=1}^{m} a_{r,i}\sigma(\langle w_i, x\rangle + b_i) \ . \tag{5}$$

Here, $\{w_i \in \mathbb{R}^d\}_{i\in[m]}$ are the *hidden weights* and each b_i is the bias. We let $W \in \mathbb{R}^{m\times d}$ denote the weight matrix with $w_1, \ldots, w_m$ as rows. We also write the function as $F(x; W)$. To simplify our analysis, only weight matrix W shall be updated and we keep b_i and $a_{r,i}$ at the random initialization. Let $w_i^{(0)}$ denote the initial value for w_i and we also use $a_r^{(0)} = a_r$ and $b_i^{(0)} = b_i$ to emphasize that they are at random initialization. Below we specify our random initialization:[2]

1. For all $i \in [m], j \in [d]$: $w_{i,j}^{(0)} \sim \mathcal{N}(0, 1/m)$ are i.i.d. Gaussian random variables.
2. For all $i \in [m]$, $b_i^{(0)} = 0$ with probability 1/2 and $b_i^{(0)} \sim \mathcal{N}(0, 2/m)$ otherwise.
3. For all $r \in [k], i \in [m]$, $a_{r,i}^{(0)} \sim \mathcal{N}(0, \varepsilon_a^2)$ for some fixed $\varepsilon_a \in (0, 1]$ to be specified later.

Learning processwe. Given data set $\mathcal{Z} = \{z_1, \ldots, z_N\}$ where each $z_i = (x_i, y_i)$, the network F is first randomly initialized and then updated by SGD. Let $w_i^{(t)} = \Delta w_i^{(t)} + w_i^{(0)}$ denote the value of

[2]Our random choice of $b_i^{(0)}$ seems a bit unconventional. We assume this for the purpose of simplifying proofs. It can be replaced with simply $\mathcal{N}(0, \frac{2}{m})$ at the expense of complicating the proofs of Lemma 5.1.

Algorithm 1 SGD for two-layer networks

Input: Data set $\mathcal{Z}$, initialization $W^{(0)}$, step size η

1: $W_0 = 0$
2: **for** $t = 1, 2, \ldots$ **do**
3: Randomly sample $z^{(t)} = (x^{(t)}, y^{(t)})$ from the data set $\mathcal{Z}$
4: Update: $W_t = W_{t-1} - \eta \nabla L_F(z^{(t)}, W_{t-1} + W^{(0)})$
5: **end for**

w_i at the t-th iteration of SGD. Let $W^{(0)}$ denote the initial weights, and let W_t denote the matrix with $\Delta w_i^{(t)}$ as the i-th row (note that W_t is the matrix of *increments*). For $z = (x, y)$, define

$$L_F(z, W_t) \stackrel{\text{def}}{=} L(F(x; W^{(0)} + W_t), y).$$

Given step size η and $W^{(0)}$, the SGD algorithm is presented in Algorithm 1. We remark that the (sub-)gradient $\nabla L_F(z, W^{(0)} + W_{t-1})$ is taken with respect to W_{t-1}.[3]

2.2 Three Layer Networks

Ground truth $F^*(x)$. The ground truth $F^* = (f_1^*, \cdots, f_k^*)$ for our three-layer case is

$$f_r^*(x) = \sum_{i\in[p_1]} a_{r,i}^* \Phi_i \left(\sum_{j\in[p_2]} v_{1,i,j}^* \phi_{1,j}(\langle w_{1,j}^*, x\rangle) \right) \left(\sum_{j\in[p_2]} v_{2,i,j}^* \phi_{2,j}(\langle w_{2,j}^*, x\rangle) \right) \tag{6}$$

where each function $\phi_{1,j}, \phi_{2,j}, \Phi_i \colon \mathbb{R} \to \mathbb{R}$ is infinite-order smooth, vectors $w_{1,i}^*, w_{2,i}^* \in \mathbb{R}^d$ are the *ground-truth weights of the first layer*, vectors $v_{1,i}^*, v_{2,i}^* \in \mathbb{R}^{p_2}$ are the *ground-truth weights of the second layer*, and reals $a_{r,i}^* \in \mathbb{R}$ are weights. Without loss of generality, we assume $\|w_{1,j}^*\|_2 = \|w_{2,j}^*\|_2 = \|v_{1,i}\|_2 = \|v_{2,i}\|_2 = 1$ and $|a_{r,i}^*| \le 1$.

Remark 2.2. Standard three-layer networks are only special case of our formulation (6). If we set $\phi_{2,j} \equiv 1/p_2$ as constant functions and $v_{2,i,j}^* = 1$, then

$$f_r^*(x) = \sum_{i\in[p_1]} a_{r,i}^* \Phi_i \left(\sum_{j\in[p_2]} v_{i,j}^* \phi_j(\langle w_j^*, x\rangle) \right) \tag{7}$$

is a three-layer network with activation functions $\{\phi_j\}, \{\Phi_i\}$. Our formulation (6) is much more general. As an interesting example, even when in the special case of $\Phi_i(z) = z$, the ground truth

$$\sum_{i\in[p_1]} a_{r,i}^* \left(\sum_{j\in[p_2]} v_{1,i,j}^* \phi_1(\langle w_{1,j}^*, x\rangle) \right) \left(\sum_{j\in[p_2]} v_{2,i,j}^* \phi_2(\langle w_{2,j}^*, x\rangle) \right) \tag{8}$$

captures combinations of correlations of combinations of *non-linear* measurements in different directions of x. This we do not know how to compute using two-layer networks.

Remark 2.3. In fact, our results of this paper *even apply* to the following general form:

$$f_r^*(x) = \sum_{i\in[p_1]} a_{r,i}^* \Phi_i \left(\sum_{j\in[p_2]} v_{1,i,j}^* \phi_{1,j}(\langle w_{1,j}^*, x\rangle)\langle w_{3,j}^*, x\rangle \right) \left(\sum_{j\in[p_2]} v_{2,i,j}^* \phi_{2,j}(\langle w_{2,j}^*, x\rangle)\langle w_{4,j}^*, x\rangle \right). \tag{9}$$

[3]Strictly speaking, $L(F(x; W), y)$ does not have gradient everywhere due to the non-differentiability of ReLU. Throughout the paper, ∇ is used to denote the value computed by setting $\nabla\sigma(x) = \mathbb{I}[x \ge 0]$, which is also what is used in practical auto-differentiation softwares.

We choose to present the slightly weaker formulation (6) for the presenting the cleaning proofs.

Our network $F(x;W,V)$. Our (improper) learners are three-layer networks $F = (f_1, f_2, \ldots, f_k)$ with

$$f_r(x) = \sum_{i=1}^{m_2} a_{r,i}\sigma\left(\langle v_i, \mathfrak{N}(x)\rangle + b_{2,i}\right), \tag{10}$$

$$\mathfrak{N}(x) = (\mathfrak{n}_1(x), \cdots, \mathfrak{n}_{m_1}(x)), \tag{11}$$

$$\text{where each } \mathfrak{n}_i(x) = \sigma\left(\langle w_i, x\rangle + b_{1,i}\right) \tag{12}$$

There are m_1 and m_2 hidden neurons in the first and second layers respectively.[4] The vectors $\{w_i \in \mathbb{R}^d\}_{i\in[m_1]}$ represent the *weights of the first layer* and vectors $\{v_i \in \mathbb{R}^{m_1}\}_{i\in[m_2]}$ represent the *weights of the second layer*. The reals $b_{1,i} \in \mathbb{R}$ represent the bias for the first layer, and the reals $b_{2,i} \in \mathbb{R}$ represent the bias for the second layer. Let $W \in \mathbb{R}^{m_1\times d}$ and $V \in \mathbb{R}^{m_2\times m_1}$ denote the weight matrices with $w_1, \ldots, w_{m_1}$ and $v_1, \ldots, v_{m_2}$ being the rows respectively.

We denote the function also as $F(x;W,V)$. To simplify our analysis, the weight matrices W and V are updated but b_1, b_2 and $a_{r,i}$ are unchanged. We denote by $W^{(0)}$ and $V^{(0)}$ the initial value of W and V, and use $a^{(0)} = a$, $b_1^{(0)} = b_1$ and $b_2^{(0)} = b_2$ to emphasize that their are at random initialization. Below we specify our random initialization:

1. For all $i \in [m_1], j \in [d]$: $w_{i,j}^{(0)} \sim \mathcal{N}(0, 1/m_1)$ are i.i.d. Gaussian random variables.
2. For all $i \in [m_1]$, $b_{1,i}^{(0)} = 0$ with probability 1/2 and $b_{1,i}^{(0)} \sim \mathcal{N}(0, 2/m_1)$ otherwise.
3. For all $i \in [m_2], j \in [m_1]$: $v_{i,j}^{(0)} \sim \mathcal{N}(0, 1/m_2)$ are i.i.d. Gaussian random variables.
4. For all $i \in [m_2]$, $b_{2,i}^{(0)} = 0$ with probability 1/2 and $b_{2,i}^{(0)} \sim \mathcal{N}(0, 2/m_2)$ otherwise.
5. For all $r \in [k], i \in [m_2]$, $a_{r,i}^{(0)} \sim \mathcal{N}(0, \varepsilon_a^2)$ for some fixed $\varepsilon_a \in (0, 1]$ to be specified later.

Learning process.. As in the two-layer case, we use $W_t + W^{(0)}$ and $V_t + V^{(0)}$ to denote the weight matrices at the t-th iteration of the optimization algorithm (so that W_t and V_t are the *increments*). For three-layer networks, we consider two variants of SGD to be discussed below.

Given sample $z = (x, y)$ and $\lambda > 0$, define function

$$L_F(z, \lambda, W, V) \stackrel{\text{def}}{=} L(\lambda F(x;W,V), y) \ .$$

where the role of λ is to scale down the entire function F (because a ReLU network is positive homogenous). Both variants of SGD optimize with respect to matrices W, V as well as this parameter $\lambda > 0$. The algorithms shall ensure that λ is *non-increasing* throughout the updates — this is similar to *weight decay* used in practice.

Remark 2.4. The scaling $\lambda > 0$ can be view as a simplified version of weight decay. Intuitively, in the training process, it is *easy* to add new information (from the ground truth) to our current network, but *hard* to forget "false" information that is already in network. Such false information can be accumulated from randomness of SGD, non-convex landscapes, and so on. Thus, by scaling down the weights of the current network, we can effectively forget false information.

Algorithm 2 presents the details. Choosing $L_R = L_1$ gives our first variant of SGD using objective function (13), and choosing $L_R = L_2$ gives our second variant using objective (14).

[4]Our results apply to the case of $m_1 = m_2$, but we state the more general case of $m_1 \neq m_2$ for some lemmas because they may be of independent interests.

Algorithm 2 SGD for three-layer networks

Input: Data set $\mathcal{Z}$, initialization $W^{(0)}, V^{(0)}$, step size η, time T, $\sigma_w, \sigma_v, \sigma_n, \lambda_w, \lambda_v$. Loss $L_R \in \{L_1, L_2\}$ as defined in (13) and (14).

$W_0 = 0, V_0 = 0, \lambda_1 = 1, t_0 = 1$, $T_w = \Theta\left(\eta T / \log(1/\varepsilon)\right)$.
for $t = 1, 2, \ldots, T$ **do**
  Sample $z^{(t)} = (x^{(t)}, y^{(t)})$ from the data set $\mathcal{Z}$
  Sample $W^\rho \in \mathbb{R}^{m_1 \times d}, V^\rho \in \mathbb{R}^{m_2 \times m_1}$, with $W^\rho_{ij} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma_w^2), V^\rho_{ij} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma_v^2)$
  Sample noise $\xi_w \in \mathbb{R}^{m_1 \times d}$, $\xi_v \in \mathbb{R}^{m_2 \times m_1}$ with $(\xi_w)_{ij} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma_n^2), (\xi_v)_{ij} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma_n^2)$
  If $L_R = L_2$, randomly sample diagonal matrix $\mathbf{\Sigma} \in \mathbb{R}^{m_1 \times m_1}$ with diagonal entries i.i.d. uniform on $\{1, -1\}$.
  Update $W_t \leftarrow W_{t-1} - \eta \frac{\partial L_R^{(t)}}{\partial W_{t-1}} + \eta \xi_w, \quad V_t \leftarrow V_{t-1} - \eta \frac{\partial L_R^{(t)}}{\partial V_{t-1}} + \eta \xi_v$
  if $t = t_0$ **then**
    $\lambda_{t+1} = (1-\eta)\lambda_t$. Randomly sample $t' = [t + T_w, t + 2T_w]$ and let $t_0 = t'$.
  else
    $\lambda_{t+1} = \lambda_t$.
  end if
end for
If $L_R = L_2$, randomly sample diagonal matrix $\mathbf{\Sigma} \in \mathbb{R}^{m_1 \times m_1}$ with diagonal entries i.i.d. uniform on $\{1, -1\}$ and let $W'_T = \mathbf{\Sigma} W_T$, $V'_T = V_T \mathbf{\Sigma}$. Otherwise $W'_T = W_T, V'_T = V_T$
Randomly sample $\widetilde{\Theta}(1/\varepsilon^2)$ many noise matrices $\left\{W^{\rho,j}, V^{\rho,j}\right\}$. Let

$$j^* = \arg\min_j \left\{ \mathbb{E}_{z \in \mathcal{Z}} L_F(z, W^{(0)} + W^{\rho,j} + W'_T, V^{(0)} + V^{\rho,j} + V'_T) \right\}$$

Output $W_T^{(out)} = W^{(0)} + W^{\rho,j^*} + W'_T$, $V_T^{(out)} = V^{(0)} + V^{\rho,j^*} + V'_T$.

First variant of SGD. In each iteration t, let us use the following objective function:

$$L_1^{(t)} \overset{\text{def}}{=} \left[L_F(z^{(t)}, \lambda_t, W_{t-1} + W^\rho + W^{(0)}, V_{t-1} + V^\rho + V^{(0)}) \right] + \lambda_w \|W_{t-1}\|_{2,4}^4 + \lambda_v \|V_{t-1}\|_{2,2}^2 \ . \tag{13}$$

The SGD algorithm updates $W_t = W_{t-1} - \eta \frac{\partial L_R^{(t)}}{\partial W_{t-1}} + \eta \xi_w$ and $V_t = V_{t-1} - \eta \frac{\partial L_R^{(t)}}{\partial V_{t-1}} + \eta \xi_v$. Here, the additional terms ξ_w, ξ_v are Gaussian noise for the theoretical purpose to helping SGD escape from saddle points. Their magnitudes can be inversely polynomially in m_1 and m_2, and may not be necessary in practice. The regularizer $\|W_{t-1}\|_{2,4}^4$ (see definition in Section 3) can be replaced with $\|W_{t-1}\|_{2,2+\alpha}^{2+\alpha}$ for any constant $\alpha > 0$. We choose $\alpha = 2$ to present the cleanest proof.

Second variant of SGD. This is a slight modification from the first variant to make it more sample efficient. Consider the following objective:

$$L_2^{(t)} \overset{\text{def}}{=} \left[L_F(z^{(t)}, \lambda_t, \mathbf{\Sigma} W_{t-1} + W^\rho + W^{(0)}, V_{t-1}\mathbf{\Sigma} + V^\rho + V^{(0)}) \right] + \lambda_w \|W_{t-1}\|_{2,4}^4 + \lambda_v \|V_{t-1}\|_{2,2}^2 \tag{14}$$

where $\mathbf{\Sigma} \in \mathbb{R}^{m_1 \times m_1}$ is a random diagonal matrix with diagonal entries i.i.d. uniformly drawn from $\{1, -1\}$. Here, the random diagonal matrix Σ is similar to the Dropout technique [SHK+14] used in practice which randomly masks out neurons. We make a slight modification that we only mask out the weight increments W_t, V_t, but not the initialization $W^{(0)}, V^{(0)}$.

3 Notations

We denote by $\|w\|_2$ and $\|w\|_\infty$ the Euclidean and infinity norms of vectors w, and $\|w\|_0$ the number of non-zeros of w. We denote the row l_p (for $p \geq 1$) norm for $W \in \mathbb{R}^{m \times d}$ as

$$\|W\|_{2,p} \stackrel{\text{def}}{=} \left(\sum_{i \in [m]} \|w_i\|_2^p \right)^{1/p} . \tag{15}$$

By definition, $\|W\|_{2,2} = \|W\|_F$ is the Frobenius norm of W.

For a matrix $W \in \mathbb{R}^{m \times d}$, we use W_i or sometimes w_i to denote the i-th row of W.

The ℓ_2 Wasserstein distance between random variables A, B is

$$\mathcal{W}_2(A, B) \stackrel{\text{def}}{=} \sqrt{\inf_{(X,Y) \text{ s.t. } X \sim A, Y \sim B} \mathbb{E}\big[|X - Y|^2\big]} \tag{16}$$

where the infimum is taken over all possible joint distributions over (X, Y) where the marginal on X (resp. Y) is distributed in the same way as A (resp. B).

For notation simplicity, with high probability (or w.h.p.) means with probability $1 - e^{-c \log^2 m}$ for a sufficiently large constant c for two-layer network, and $1 - e^{-c \log^2 (m_1 m_2)}$ for three-layer network. In this paper, $\widetilde{O}$ hides factors of $\mathbf{poly}(\log m)$ for two-layer networks, or $\mathbf{poly}(\log m_1, \log m_2)$ for three-layer networks. We use $x = y \pm z$ to denote that $x \in [y - z, y + z]$. $\mathbb{I}_E$ or $\mathbb{I}[E]$ denote the indicator function of the event E.

The following notion measures the complexity of any activation function $\phi(z)$ used in the ground-truth network. Suppose $\phi(z) = \sum_{i=0}^{\infty} c_i z^i$. Given non-negative R, the complexity $\mathfrak{C}(\phi, R)$ is

$$\mathfrak{C}(\phi, R) \stackrel{\text{def}}{=} \sum_{i=0}^{\infty} (CR)^i |c_i| \tag{17}$$

where C is a sufficiently large constant (e.g., 10^4). We also use

$$\mathfrak{C}(\phi, R) = \max_{j \in [p_1], s \in [1,2]} \{\mathfrak{C}(\phi_{s,j}, R)\}, \quad \mathfrak{C}(\Phi, R) = \max_{j \in [p_1]} \{\mathfrak{C}(\Phi_j, R)\} \tag{18}$$

as the complexity of the first and second hidden layers, respectively. We assume throughout the paper that $\phi_{s,j}$ and Φ_j are sufficiently smooth such that these terms are bounded.

4 Main Results

4.1 Two Layer Networks

We have the following main theorem for two-layer networks. (Recall that p is the number of hidden neurons in the ground-truth network.)

Theorem 4.1. *For every $\varepsilon \in (0, \frac{1}{pk})$, by setting $\varepsilon_a = \varepsilon / \widetilde{\Theta}(1)$ in the initialization, there exists $M = \mathbf{poly}(\mathfrak{C}(\phi, \sqrt{\log(p/\varepsilon)}), 1/\varepsilon)$ such that for every $m \geq M$, we can choose a learning rate $\eta = 1/(m \cdot \mathbf{poly}(M, \log m))$ such that as long as*

$$N \geq \widetilde{\Omega}(M)$$

there is $T = \mathbf{poly}(M, \log m)$ such that with high probability, SGD in iteration T satisfies

$$\mathbb{E}_{(x,y) \sim \mathcal{D}} L(F(x; W^{(0)} + W_T), y) \leq \varepsilon_0 + \varepsilon.$$

One can verify that $\mathfrak{C}(\phi, \sqrt{\log p/\varepsilon})$ is $o(p/\varepsilon)$ for functions such as $\phi(z) = e^z, \sin z, \sinh z$ or low degree polynomials. Our theorem indicates that for ground-truth networks with such activation functions, we can efficiently learn them using a network of size $\mathbf{poly}(1/\varepsilon, k, p)$ with ReLU activation. The number of samples needed is at most $\mathbf{poly}(1/\varepsilon, k, p, \log m)$, (almost) independent of m, the amount of overparametrization in our network.

If $\phi(z)$ is sigmoid, to get ε approximation we can truncate its Taylor series at degree $O(\log \frac{1}{\varepsilon})$. One can verify that $\mathfrak{C}(\phi, \sqrt{\log p/\varepsilon})$ is at most $(\log(p/\varepsilon))^{O(1/\varepsilon)} \leq \left(\frac{1}{\varepsilon}\right)^{O(\log \log \frac{p}{\varepsilon})}$, which is slightly super-polynomial.

One might want to compare our result to [APVZ14]. First of all, our result allows activation functions in the ground truth to be *infinite-degree* polynomials, which is not captured by [APVZ14]. More importantly, to learn a polynomial with degree r, their sample complexity is d^r but our result is *dimension free.*

One can also view Theorem 4.1 as a nonlinear analogue to the margin theory for linear classifiers. The ground-truth network with a small generalization error (and of bounded norm) can be viewed as a "large margin non-linear classifier." In this view, Theorem 4.1 shows that assuming the existence of such large-margin classifier, SGD finds a good solution with sample complexity mostly determined by the margin, instead of the *dimension* of the data.

4.2 Three Layer Networks

We give the main theorem for using SGD (Algorithm 2 with $L_R = L_1$) to train three-layer networks.

Theorem 4.2. *If $L_R = L_1$, then for every* constant *$\gamma > 0$, every $\varepsilon \in (0, 1/(p_1 p_2 k))$, there exists*[5]

$$M = \mathbf{poly}\left(\mathfrak{C}(\Phi, \sqrt{p_2}\mathfrak{C}(\phi, \sqrt{\log(1/\varepsilon)})\sqrt{\log(1/\varepsilon)}), \frac{1}{\varepsilon}\right)$$

such that for every $m_2 = m_1 \geq M$, and properly set $\varepsilon_a, \lambda_w, \lambda_v, \sigma_w, \sigma_v$,[6] *as long as*

$$N \geq \widetilde{\Omega}(M(m_2)^2)$$

there is a choice $\eta = 1/\mathbf{poly}(m_1, m_2)$ and $T = \mathbf{poly}(m_1, m_2)$ such that with probability at least 99/100,[7]

$$\mathbb{E}_{(x,y)\sim\mathcal{D}} L(\lambda_T F(x; W_T^{(out)}, V_T^{(out)}), y) \leq (1+\gamma)\varepsilon_0 + \varepsilon.$$

Note that the expected loss is over $\mathcal{D}$, the real data distribution, instead of $\mathcal{Z}$, the empirical distribution. Theorem 4.2 shows that using N samples we can find a network with a small *generalization* error.

Recall from Section 2 that a major advantage of using a three-layer network, compared to a two-layer one, is the ability to learn (combinations of) *correlations* between non-linear measurements of the data. This corresponds to the special case $\Phi(z) = z$ and our three-layer ground truth (for the r-th output) can be

$$\sum_{i\in[p_1]} a^*_{r,i} \left(\sum_{j\in[p_2]} v^*_{1,i,j}\phi_{1,j}(\langle w^*_{1,j}, x\rangle)\right)\left(\sum_{j\in[p_2]} v^*_{2,i,j}\phi_{2,j}(\langle w^*_{2,j}, x\rangle)\right) \ . \tag{19}$$

Since $\Phi(z) = z$, we have $\mathfrak{C}(\Phi, \sqrt{p_2}\mathfrak{C}(\phi, \sqrt{\log(1/\varepsilon)})\sqrt{\log(1/\varepsilon)}) \approx O(\sqrt{p_2}\mathfrak{C}(\phi, \sqrt{\log(1/\varepsilon)}))$, learning this three-layer network is essentially in the *same complexity* as learning each $\phi_{s,j}$ in two-layer

[5] The coefficients and degree of **poly** can depend on γ.

[6] Their choices can be found in Lemma 5.3.

[7] The probability is taken over the randomness of the initialization weights $W^{(0)}, V^{(0)}$, and all the randomness used in the algorithm.

networks. As a concrete example, a three-layer network can learn $\cos(100\langle w_1^*, x\rangle) \cdot \cos(100\langle w_2^*, x\rangle)$ up to accuracy ε in complexity $\mathbf{poly}(1/\varepsilon)$, while it is unclear how to do so using two-layer networks.

For general Φ, ignoring small factors $\sqrt{\log(1/\varepsilon)}$ and p_2, the complexity of three-layer networks is essentially $\mathfrak{C}(\Phi, \mathfrak{C}(\phi))$. This is necessary in some sense: consider the case when $\phi(x) = Cx$ for a very large parameter C, then $\Phi(\phi(z))$ is just a function $\Phi'(z) = \Phi(Cz)$, and we have $\mathfrak{C}(\Phi', 1) = \mathfrak{C}(\Phi, C) = \mathfrak{C}(\Phi, \Theta(\mathfrak{C}(\phi, 1)))$.

We next give the main theorem for SGD with Dropout-type noise (Algorithm 2 with $L_R = L_2$) to train three-layer networks. It has better sample efficiency comparing to Theorem 4.2.

Theorem 4.3. *If $L_R = L_2$, in the same setting as Theorem 4.2, as long as*

$$N \geq \widetilde{\Omega}(M)$$

there is a choice $\eta = 1/\mathbf{poly}(m_1, m_2)$, $T = \mathbf{poly}(m_1, m_2)$ such that with probability at least $99/100$*,*

$$\mathbb{E}_{(x,y)\sim\mathcal{D}} L(\lambda_T F(x; W_T^{(out)}, V_T^{(out)}), y) \leq (1+\gamma)\varepsilon_0 + \varepsilon.$$

As mentioned, the training algorithm for this version of SGD uses a random diagonal scaling $\mathbf{\Sigma}$, which is similar to the Dropout trick used in practice to reduce sample complexity by turning on and off each hidden neuron. Theorem 4.3 shows that in this case, the sample complexity needed to achieve a small generalization error scales only polynomially with the complexity of the ground-truth network, and is *(almost) independent* of m, the amount of overparameterization in our network.

5 Main Lemmas for Three Layer Networks

We present the key technical lemmas we used for proving the three-layer network Theorem 4.2 and 4.3. The two-layer result is based on similar ideas but simpler. We defer that to Appendix G.

We show the existence of good networks near initialization in Section 5.1, and present key lemma about the optimization procedure in Section 5.2. We present our coupling technique in Section 5.3, and then conclude with the lemma for generalization in Section 5.4.

5.1 Ground Truth Out of Randomness

To construct a network to approximate ground truth, our high-level idea is to represent it as an expectation and then apply sampling. We begin with representation of the activation function.

Lemma 5.1 (from indicator to functions)**.** *For every smooth function ϕ, every $\varepsilon \in (0, 1/\mathfrak{C}(\phi, 1))$, we have that there exists a function $h : \mathbb{R}^2 \to \left[-\mathfrak{C}(\phi, \sqrt{\log(1/\varepsilon)}), \mathfrak{C}(\phi, \sqrt{\log(1/\varepsilon)})\right]$ such that for every $x_1 \in [-1, 1]$:*

$$\left|\mathbb{E}\left[\mathbb{I}_{\alpha_1 x_1 + \beta_1\sqrt{1-x_1^2} + b_0 \geq 0} h(\alpha_1, b_0)\right] - \phi(x_1)\right| \leq 2\varepsilon \tag{20}$$

where $\alpha_1, \beta_1 \sim \mathcal{N}(0, 1)$ and b_0 is 0 with half probability and is drawn from $\mathcal{N}(0, 2)$ with the other half probability. α_1, β_1 and b_0 are independent random variables.

Given Lemma 5.1, we can directly apply it to the two-layer case. As for the three-layer case, we need to first apply it to the second hidden layer. Let us consider the input (without bias) to a single neuron of the second layer at random initialization, given as:

$$n_1(x) = \sum_{i\in[m_1]} v_{1,i}^{(0)} \sigma\left(\langle w_i^{(0)}, x\rangle + b_{1,i}^{(0)}\right). \tag{21}$$

We show the following lemma:

Lemma 5.2 (information out of randomness)**.** *For every $w^* \in \mathbb{R}^d$ with $\|w^*\|_2 = 1$, for every $\varepsilon > 0$, there exists real-valued functions $\rho(v_1^{(0)}, W^{(0)}, b_1^{(0)})$ and $B(x, v_1^{(0)}, W^{(0)}, b_1^{(0)}), R(x, v_1^{(0)}, W^{(0)}, b_1^{(0)})$, and function $\phi_\varepsilon(x)$ such that for every x:*

$$n_1(x) = \rho\left(v_1^{(0)}, W^{(0)}, b_1^{(0)}\right)\phi_\varepsilon(x) \tag{22}$$

$$+ B\left(x, v_1^{(0)}, W^{(0)}, b_1^{(0)}\right) + R\left(x, v_1^{(0)}, W^{(0)}, b_1^{(0)}\right). \tag{23}$$

Moreover, letting $C = \mathfrak{C}(\phi, \sqrt{\log(1/\varepsilon)})$ be the complexity of ϕ, we have

1. *For every fixed x, $\rho\left(v_1^{(0)}, W^{(0)}, b_1^{(0)}\right)$ is independent of $B\left(x, v_1^{(0)}, W^{(0)}, b_1^{(0)}\right)$.*
2. *$\rho\left(v_1^{(0)}, W^{(0)}, b_1^{(0)}\right) \sim \mathcal{N}\left(0, \frac{1}{100C^2 m_2}\right)$.*
3. *For every x with $\|x\|_2 = 1$, $|\phi_\varepsilon(x) - \phi(\langle w^*, x\rangle)| \le 2\varepsilon$.*
4. *For every fixed x with $\|x\|_2 = 1$, with high probability $\left|R\left(x, v_1^{(0)}, W^{(0)}, b_1^{(0)}\right)\right| \le \widetilde{O}\left(\frac{1}{\sqrt{m_1 m_2}}\right)$.*

Lemma 5.2 shows that we can "view" the input to any neuron of the second layer essentially as "a Gaussian variable ρ" times "the target function $\phi(\langle w^*, x\rangle)$" in the first layer of the ground-truth network. This may sound weird at first look because random initialization cannot carry any information about the ground truth.

There is no contradiction here, because we will show, B is essentially another Gaussian (with the same distribution as ρ) times $\sqrt{\text{constant} - \phi^2(\langle w^*, x\rangle)}$, thus $n_1(x)$ can still be *independent* of the value of $\phi(\langle w^*, x\rangle)$. Nevertheless, the decomposition in Lemma 5.2 shall enable us to show that, when we start to modify the hidden weights W, V, the learning process will start to *discover* this structure and make the weight of the term relating to $\phi(\langle w^*, x\rangle)$ *stand out* from other terms.

Using this lemma, we will show one of our main structural properties of overparameterized networks: solutions with good generalization errors are *dense* in the parameter space, in the sense that with high probability over the random initialization, there exists a good solution in the close neighborhood of the initialized weights. This is formally written as Lemma 5.6 in Section 5.3.

5.2 Optimization

Naïve approach. The tentative approach is using the property that solutions with good generalization errors are dense in the parameter space, we'd like to show that the optimization landscape of the overparameterized three-layer neural network is benign: it has no spurious local minimal or any ≥ 3 order critical points. Thus, we can use the existing theorem on escaping saddle points with SGD to prove convergence. However, even before digging into the optimization landscape, there is already a big hole in this idea.

Key issue. ReLU network is not second-order-differentiable: a ReLU activation does not have a well-defined Hessian/sub-Hessian at zero. One may naïvely think that since a ReLU network is infinite-order differentiable everywhere except a measure zero set, so we can safely ignore the Hessian issue and proceed by pretending that the Hessian of ReLU is always zero. This intuition is *very wrong*. Following it, we could have run into the absurd conclusion that any piece-wise linear function is convex, since the Hessian of it is *zero* almost everywhere. The issue is that the only non-smooth point of ReLU has a Hessian value equal to the δ-function, thus, these non-smooth points, albeit being measure zero, are actually turning points in the landscape. If we want a meaningful second-order statement of the ReLU network, *we must not naïvely ignore the "Hessian" of a ReLU network at zeros.*

Smoothing. To fix the naïve approach, we use Gaussian smoothing. Given any bounded function $f : \mathbb{R}^{m_3} \to \mathbb{R}$, we have that $\mathbb{E}_{\rho \sim \mathcal{N}(0,\sigma^2 \mathbf{I})}[f(x+\rho)]$ is a infinite-order differentiable function as long as $\sigma > 0$. Thus, we can consider the smoothed version of the neural network: $F(x; W + W^\rho, V + V^\rho)$ where W^ρ, V^ρ are random Gaussian matrices. As for the main result, we show that essentially, $\mathbb{E}[L_F(z, W + W^\rho, V + V^\rho)]$ also has the desire property that all ≥ 3 order critical points are global optimal. Perhaps worth pointing out, the Hessian of this smoothed function can be *significantly different* from the original one, for example, $\mathbb{E}_{\rho \sim \mathcal{N}(0,1)}[\sigma(x+\rho)]$ has a Hessian value ≈ 1 at all $x = [-1, 1]$, while in the original ReLU function σ, the Hessian is 0 *almost everywhere*.

In practice, since the solution W_t, V_t are found by *stochastic* gradient descent starting from *random initialization*, they will have a non-negligible amount of *intrinsic* noise. Thus, the additional smoothing in the algorithm might not be needed by an observation in [KLY18]. Smooth analysis [ST01] might also be used for analyzing the effect of such noise, but this is beyond the scope of this paper.

Actual algorithm. Let us consider the following smoothed, and regularized objective:

$$L'(\lambda_t, W_t, V_t) = \mathbb{E}_{W^\rho, V^\rho, x, y \sim \mathcal{Z}} \left[L\left(\lambda_t F\left(x; W^{(0)} + W^\rho + W_t, V^{(0)} + V^\rho + V_t \right), y \right) \right] + R(W_t, V_t) \tag{24}$$

where W^ρ, V^ρ are Gaussian random matrices with each entry i.i.d. from $\mathcal{N}(0, \sigma_w^2)$ and $\mathcal{N}(0, \sigma_v^2)$, respectively. $R(W_t, V_t) = \lambda_v \|V_t\|_F^2 + \lambda_w \|W_t\|_{2,4}^4$ and λ_v, λ_w are set such that $\lambda_v \|V^*\|_F^2 = \lambda_w \|W^*\|_{2,4}^4 = \varepsilon^2$. See Algorithm 2 (with $L_R = L_1$) for the details.

We shall prove the following lemma that shows the smoothed objective has the desire property that all ≥ 3 order critical points are globally optimal.

Lemma 5.3 (Optimization)**.** *For every $\varepsilon \in [0, 1]$, when $m_1 = m_2 \geq M$ as defined in Theorem 4.2 and $C_1 = \mathbf{poly}\left(\mathfrak{C}(\Phi, \sqrt{p_2}\mathfrak{C}(\phi, \sqrt{\log(1/\varepsilon)})\sqrt{\log(1/\varepsilon)}), k, p_1, p_2, \frac{1}{\varepsilon} \right)$, for every constant $\gamma \in (0, 1/4]$, let $\mathbf{\Sigma} \in \mathcal{R}^{m_1 \times m_1}$ be a diagonal matrix with each diagonal entry i.i.d. drawn from $\{-1, 1\}$. Consider the following parameter choices*[8]*:*

$$\tau_v = (\varepsilon')^{\Theta(1)} \sqrt{\frac{m_1}{m_2}}, \quad \tau'_w = \frac{1}{(\varepsilon')^{\Theta(1)} m_1^{3/4}}, \quad \lambda_w = \frac{2}{(\tau'_w)^4}, \quad \lambda_v = \frac{2}{\tau_v^2}, \tag{25}$$

$$\sigma_v = \frac{(\varepsilon')^{\Theta(1)}}{\sqrt{m_2}}, \quad \sigma_w = \frac{1}{m_1^{99/100}}, \quad \tau_w = \sigma_w m_1^{1/4}, \tag{26}$$

$$m_2 = m_1, \varepsilon' = \varepsilon / C_1, \varepsilon_a = \varepsilon, \tag{27}$$

then as long as

$$L'(\lambda_t, W_t, V_t) \in [(1+\gamma)\varepsilon_0 + \widetilde{\Omega}(\varepsilon), 2] \tag{28}$$

$$\lambda_t \geq (\varepsilon')^{-O(1)} \tag{29}$$

there exists W^*, V^* with $\|W^*\|_F, \|V^*\|_F \leq 1$ such that for every $\eta \in \left[0, \frac{1}{\mathbf{poly}(m_1, m_2)}\right]$:

$$\min\{\mathbb{E}_{\mathbf{\Sigma}}\left[L'\left(\lambda_t, W_t + \sqrt{\eta}\mathbf{\Sigma} W^*, V_t + \sqrt{\eta} V^* \mathbf{\Sigma}\right) \right], L'((1-\eta)\lambda_t, W_t, V_t)\} \tag{30}$$

$$\leq (1 - \eta\gamma/4)(L'(\lambda_t, W_t, V_t)). \tag{31}$$

This lemma then leads to the final convergence for Algorithm 2.

[8] The constant inside Θ can depend on γ

Lemma 5.4 (Convergence). *In the setting of Theorem 4.2, Algorithm 2 converges in time $T = \mathbf{poly}(m_1, m_2)$ to an $(1+\gamma)\varepsilon_0 + \varepsilon$ solution w.p. $\geq 99/100$.*

5.3 Coupling between Networks

To show that the existence of good networks, we will first use the technical lemmas in Section 5.1 to construct a pseudo network, and then show that it is coupled with our real network being learnt.

Let matrices V^ρ, W^ρ to be random Gaussian matrices such that:

$$V^\rho_{i,j} \sim \mathcal{N}(0, \sigma_v^2) \quad \text{and} \quad W^\rho_{i,j} \sim \mathcal{N}(0, \sigma_w^2) \tag{32}$$

for some $\sigma_v, \sigma_w \in [1/(m_1 m_2), 1]$.

Let us denote $D_{w,x}$ as a diagonal matrix indicating the sign of the ReLU's for the first layer at *random initialization*, i.e., $[D_{w,x}]_{i,i} = \mathbb{I}[\langle w_i^{(0)}, x\rangle + b_{1,i}^{(0)} \geq 0]$, and $D_{v,x}$ the sign of the second layer at random initialization.

Now, suppose we are currently at weights $W^{(0)} + W' + W^\rho$, $V^{(0)} + V' + V^\rho$. Denote $D_{w,x} + D'_{w,x}$ as the sign of the ReLU's for first layer at this weight, i.e. $[D_{w,x} + D'_{w,x}]_{i,i} = \mathbb{I}[\langle w_i^{(0)} + w'_i + w_i^{(\rho)}, x\rangle + b_{1,i}^{(0)} \geq 0]$ and $D_{v,x} + D'_{v,x}$ sign of the second layer at this weight.

For a fixed $r \in [k]$, let us denote $a_r = (a_{i,r})_{i\in[m_2]}$, let us define the pseudo network (and its semi-bias, bias-free versions) as

$$g_r(x; W, V) = a_r(D_{v,x} + D'_{v,x})\left(V(D_{w,x} + D'_{w,x})(Wx + b_1) + b_2\right) \tag{33}$$

$$g_r^{(b)}(x; W, V) = a_r(D_{v,x} + D'_{v,x})V(D_{w,x} + D'_{w,x})(Wx + b_1) \tag{34}$$

$$g_r^{(b,b)}(x; W, V) = a_r(D_{v,x} + D'_{v,x})V(D_{w,x} + D'_{w,x})Wx \tag{35}$$

By this definition, at exactly $W^{(0)} + W' + W^\rho$, $V^{(0)} + V' + V^\rho$ the pseudo network should equal to the true network.

$$g_r\left(x; W^{(0)} + W' + W^\rho, V^{(0)} + V' + V^\rho\right) = f_r\left(x; W^{(0)} + W' + W^\rho, V^{(0)} + V' + V^\rho\right) \tag{36}$$

We prove the following lemma:

Lemma 5.5 (Coupling). *Suppose $\tau_v \in \left[\frac{1}{\sqrt{m_2}}, \sqrt{\frac{m_1}{m_2}}\right]$, $\tau_w \in \left[\frac{1}{m_1^{3/4}}, \frac{1}{m_1^{1/2}}\right]$, $\sigma_w \leq \frac{\tau_w}{m_1^{1/4}}$, $\sigma_v \leq \frac{1}{4\sqrt{m_2}}$, $m_2 \geq m_1$, and $\eta > 0$. Given fixed unit vector x, and perturbation matrices W', V', W'', V'' (that may depend on the randomness of $W^{(0)}, b_1^{(0)}, V^{(0)}, b_2^{(0)}$) satisfying*

$$\|W'\|_{2,4} \leq \tau_w, \|V'\|_F \leq \tau_v, \|W''\|_{2,4} \leq \tau_w, \|V''\|_F \leq \tau_v \ ,$$

and random diagonal matrix $\mathbf{\Sigma}$ with each diagonal entry i.i.d. drawn from $\{\pm 1\}$, then with high probability the following holds:

1. *(Sparse sign change). $\|D'_{w,x}\|_0 \leq \widetilde{O}(\tau_w^{4/5} m_1^{6/5})$, $\|D'_{v,x}\|_0 \leq \widetilde{O}\left(\sigma_v m_2^{3/2} + \tau_v^{2/3} m_2 + \tau_w^{2/3} m_1^{1/6} m_2\right)$.*
2. *(Diagonal cross term vanish).*

$$g_r(x; W^{(0)} + W^\rho + W' + \eta\mathbf{\Sigma}W'', V^{(0)} + V^\rho + V' + \eta V''\mathbf{\Sigma}) \tag{37}$$

$$= g_r\left(x; W^{(0)} + W^\rho + W', V^{(0)} + V^\rho + V'\right) + g_r^{(b,b)}(x; \eta\mathbf{\Sigma}W'', \eta V''\mathbf{\Sigma}) + g'_r(x) \tag{38}$$

where $\mathbb{E}_{\mathbf{\Sigma}}[g'_r(x)] = 0$ and with high probability $|g'_r(x)| \leq \eta\widetilde{O}\left(\frac{\sqrt{m_2}\tau_v}{\sqrt{m_1}} + m_2^{1/2}\tau_w\right)$.

Using this lemma, we can show the main existential lemma of this paper:

Lemma 5.6 (Existential)**.** *For every $\varepsilon > 0$, let $\varepsilon_a = \varepsilon$, there exists*

$$M = \mathbf{poly}\left(\mathfrak{C}(\Phi, \sqrt{p_2}\mathfrak{C}(\phi, \sqrt{\log(1/\varepsilon)})\sqrt{\log(1/\varepsilon)}), k, p_1, p_2, \frac{1}{\varepsilon}\right) \tag{39}$$

such that if $m_1, m_2 \geq M$, and $\tau_v, \tau_w, \sigma_v, \sigma_w$ satisfies the bound in Lemma 5.3, then with high probability, there exists weights W^*, V^* (independent of W^ρ, V^ρ) with norm bounded by:

$$\|W^*\|_{2,\infty} = \max_i \|W_i^*\|_2 \leq \frac{C_0^2}{\varepsilon^{11} m_1}, \quad \|V\|_{2,\infty} = \max_i \|V_i^*\|_2 \leq \frac{\varepsilon^8 \sqrt{m_1}}{m_2} \tag{40}$$

for some fixed

$$C_0 = \mathbf{poly}\left(\mathfrak{C}(\Phi, \sqrt{p_2}\mathfrak{C}(\phi, \sqrt{\log(1/\varepsilon)})\sqrt{\log(1/\varepsilon)}), k, p_1, p_2, \log m_1, \log m_2\right) \tag{41}$$

such that

$$\mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\sum_{r=1}^{k}\left|f_r^*(x) - g_r^{(b,b)}(x, W^*, V^*)\right|\right] \leq \varepsilon, \tag{42}$$

and hence,

$$\mathbb{E}_{(x,y)\sim\mathcal{D}}\left[L(G^{(b,b)}(x, W^*, V^*), y)\right] \leq \varepsilon_0 + \varepsilon. \tag{43}$$

This lemma shows that we can always find a good network after *fixing* the current sign pattern of the ReLU's, and these matrices will be used as the descent direction in Lemma 5.3.

The next lemma shows that when doing a small update to the current weight, we can safely view the weights as fixed, as long as sufficient amount of smoothing is given.

For every input x and every $W = W^{(0)} + W', V = V^{(0)} + V'$, let $D_{w,x,\rho}$ denote the diagonal matrix with diagonals being the signs of the first layer at weights $W + W^\rho$. For a random diagonal matrix $\mathbf{\Sigma}$, let $D_{w,x,\rho,\eta}$ denote that at weights $W + W^\rho + \eta\mathbf{\Sigma}W''$. Let $D_{v,x,\rho}$ denote that at the second layer at weights $W + W^\rho$ and $V + V^\rho$, and $D_{v,x,\rho,\eta}$ denote that at weights $W + W^\rho + \eta\mathbf{\Sigma}W''$ and $V + V^\rho + \eta\mathbf{\Sigma}V''$.

For a fixed $r \in [k]$, define the pseudo network $P_{\rho,\eta}$ and the smoothed pseudo network $P''_{\rho,\eta}$ as:

$$P_{\rho,\eta} \stackrel{\text{def}}{=} a_r D_{v,x,\rho,\eta}\left((V + V^\rho + \eta V'')D_{w,x,\rho,\eta}\left((W + W^\rho + \eta W'')x + b_1\right) + b_2\right) \tag{44}$$

$$P'_{\rho,\eta} \stackrel{\text{def}}{=} a_r D_{v,x,\rho}\left((V + V^\rho + \eta V'')D_{w,x,\rho}\left((W + W^\rho + \eta W'')x + b_1\right) + b_2\right). \tag{45}$$

We now prove the following lemma:

Lemma 5.7 (Smoothed pseudo vs pseudo)**.** *For every $\eta \leq \frac{1}{\mathbf{poly}(m_1, m_2)}$, for every fixed x with norm one, for every $\|W''\|_{2,\infty} \leq \tau_{w,\infty}, \|W'\|_{2,4} \leq \tau_w, \|V'\|_{2,2} \leq \tau_v$ and $\|V''\|_{2,\infty} \leq \tau_{v,\infty}$, and $m_2 \geq m_1$, we have:*

$$\mathbb{E}_{W^\rho, V^\rho}\left[\frac{|P_{\rho,\eta} - P'_{\rho,\eta}|}{\eta^2}\right] = \widetilde{O}\left(m_1\frac{\tau_{w,\infty}^2}{\sigma_w} + \frac{m_2\tau_{w,\infty}^2}{\sigma_v} + \frac{m_2}{m_1}\frac{\tau_{v,\infty}^2}{\sigma_v}\right) + O_p(\eta). \tag{46}$$

where O_p hides polynomial factor of m_1, m_2.

We will also need the following coupling lemma when Σ is used (Algorithm 2 with $L_R = L_2$).

Lemma 5.8 (Stronger Coupling for random diagonal matrix)**.** *For every W', V' such that $\|V'\|_2 \leq \tau_v, \|W'\|_{2,4} \leq \tau_w$ with $\tau_v \in \left[\frac{1}{\sqrt{m_2}}, \sqrt{\frac{m_1}{m_2}}\right]$, $\tau_w \in \left[\frac{1}{m_1^{3/4}}, \frac{1}{m_1^{1/2}}\right]$, we have for a random diagonal matrix*

$\mathbf{\Sigma}$ *with diagonal entries i.i.d. uniform on* $\{-1,1\}$*, with high probability*

$$f_r(x; W^{(0)} + \mathbf{\Sigma}W', V^{(0)} + V'\mathbf{\Sigma}) = a_r D_{v,x}(V^{(0)} D_{w,x}(W^{(0)}x + b_1) + b_2) + a_r D_{v,x} V' D_{w,x} W' x \tag{47}$$

$$\pm \widetilde{O}\left(\tau_w^{16/5} m_1^{4/5} m_2^{3/2} + \frac{\sqrt{m_2}}{\sqrt{m_1}}\tau_v\right). \tag{48}$$

5.4 Generalization

With the existence and the optimization lemmas, we can bound the training loss. The final piece of our analysis is the Rademacher complexity bound.

Lemma 5.9 (Rademacher complexity for three layer)**.** *For every* $\tau_v, \tau_w \geq 0$*, w.h.p. for every* $r \in [k]$ *and every* $N \geq 1$*, we have by our choice of parameters in Lemma 5.3, the empirical Rademacher complexity is bounded by*

$$\frac{1}{N}\mathbb{E}_{\xi\in\{\pm 1\}^N}\left[\sup_{\|V'\|_F \leq \tau_v, \|W'\|_{2,4}\leq \tau_w}\left|\sum_{i\in[N]} \xi_i \mathbb{E}_{\mathbf{\Sigma}}[f_r(x_i, W^{(0)} + W^\rho + \mathbf{\Sigma}W', V^{(0)} + V^\rho + V'\mathbf{\Sigma})]\right|\right] \tag{49}$$

$$\leq \widetilde{O}\left(\frac{\tau_w \tau_v m_1^{1/4}\sqrt{m_2}}{\sqrt{N}} + \frac{1}{\sqrt{N}}\right). \tag{50}$$

6 Discussions

This paper shows that using an overparameterized neural network as an *improper* learner for some simpler hypothesis class might not be difficult. Especially, using an overparameterized two/three-layer neural network with ReLU activations as an improper learner for two/three-layer networks of smaller size with smooth activations, is computationally tractable.

The analysis can also be applied to convolutional neural networks, a special kind of neural networks that are widely used in practice. In particular, convolutional neural networks consist of multiple convolution layers. Suppose the input is a two dimensional matrix x of size $d \times s$, then a convolutional layer on top of x is defined as follows. There are d' fixed subsets $\{S_1, S_2, \ldots, S_{d'}\}$ of d of size k. The output of the convolution layer is a matrix of size $d' \times m$, whose (i,j)-th entry is $\phi(\langle w_j, x_{S_i}\rangle)$, where x_{S_i} is the submatrix of rows indexed by S_i, w_j is a matrix of proper size (called the weight matrix for the j-th channel), and ϕ is the activation function. Overparameterization then means a larger number of channels m in our learnt network than that in the ground truth. Our analysis can be used to show a similar result for this type of networks.

There are plenty of open directions following our work, especially how to extend our result to larger number of layers. Even for three layers, our algorithm currently uses an explicit regularizer $R(W_t, V_t)$ on the weights to control the generalization error. However, in practice, it is known that neural networks actually *implicitly* regularizes: even without any restriction on the weights, the result learnt by an *unconstraint* overparameterized neural network still generalizes. It is an interesting direction to explain this implicit regularization for three layers and beyond. As to a third direction, currently our result does not directly apply to ground-truth networks with ReLU activations. While there are functions ϕ with $C(\phi, 1) = 2^{O(1/\varepsilon)}$ that approximate a ReLU function in $[-1,1]$ with ε error, obtaining an polynomial, or sub-exponential bound on ground-truth networks with ReLU would be of great interest.

In summary, we believe that our work *opens up* a new direction in both the algorithmic and generalization perspectives of overparameterized neural networks, and pushing forward these ideas can lead to more understanding about neural network learning.

Appendix

A Technical Lemmas

A.1 Probability

Lemma A.1 (Gaussian indicator concentration). *Let $(n_1, \alpha_1, a_{1,1}, a_{2,1}), \cdots, (n_m, \alpha_m, a_{1,m}, a_{2,m})$ be m 4-tuples of i.i.d. standard Gaussian random variables (within a tuple n_i, α_i are not necessarily independent, but they are independent of $a_{1,i}, a_{2,i}$, and $a_{1,i}, a_{2,i}$ are independent).*

For a fixed $L > 0$, let $h : \mathbb{R} \to [-L, L]$, we have: for every $B \geq 1$:

$$\mathbf{Pr}\left[\left|\left(\sum_{i\in[m]} a_{1,i}^2 \mathbb{I}[n_i \geq 0] h(\alpha_i)\right) - m\mathbb{E}[a_{1,1}^2 \mathbb{I}[n_1 \geq 0] h(\alpha_1)]\right| \geq BL(\sqrt{m} + B)\right] \leq 4e^{-B^2/2} \tag{51}$$

And

$$\mathbf{Pr}\left[\left|\left(\sum_{i\in[m]} a_{1,i} a_{2,i} \mathbb{I}[n_i \geq 0] h(\alpha_i)\right)\right| \geq BL(\sqrt{m} + B)\right] \leq 4e^{-B^2/2} \tag{52}$$

Proof of Lemma A.1. Let us consider a fixed $n_1, \alpha_1, \cdots, n_m, \alpha_m$, then since each $\mathbb{I}[n_i \geq 0]h(\alpha_i) \leq L$, by Gaussian chaos variables concentration bound (e.g., Example 2.15 in [Mar15]) we have that

$$\mathbf{Pr}\left[\left|\left(\sum_{i\in[m]} a_{1,i}^2 \mathbb{I}[n_i \geq 0] h(\alpha_i)\right) - m\mathbb{E}[a_{1,1}^2 \mathbb{I}[n_1 \geq 0] h(\alpha_1)]\right| \geq BL(\sqrt{m} + B) \;\middle|\; \{n_i, \alpha_i\}_{i\in[m]}\right] \leq 4e^{-B^2/2} \tag{53}$$

Since this holds for every choice of $\{n_i, \alpha_i\}_{i\in[m]}$ we can complete the proof. The proof of the second inequality is identical. ☐

Lemma A.2 (Gaussian Lipschitz concentration). *Let h be a L-Lipschitz function with $h(0) = 0$, let $a_1, \cdots, a_m$ and $v_1, \cdots, v_m$ be i.i.d. standard Gaussian random variables. We have: for every $B \geq 1$:*

$$\mathbf{Pr}\left[\left|\sum_{i\in[m]} a_i h(v_i)\right| \geq BL(\sqrt{m} + B)\right] \leq 4e^{-B^2/2} \tag{54}$$

Proof of Lemma A.2. For notation simplicity let us call $z = \sum_{i\in[m]} a_i h(v_i)$. Conditional on $\{v_i\}_{i\in[m]}$, $a_i h(v_i)$ is a Gaussian variable with variance $\sum_{i\in[m]} h(v_i)^2$. Thus, we have tail bound:

$$\mathbf{Pr}\left[z \geq B\sqrt{\sum_{i\in[m]} h(v_i)^2} \;\middle|\; \{v_i\}_{i\in[m]}\right] \leq 2e^{-B^2/2} \tag{55}$$

Now, we consider $z' = \sum_{i\in[m]} h(v_i)^2$. Let us define a function

$$f(v_1, \cdots, v_m) = \sqrt{\sum_{i\in[m]} h(v_i)^2} \tag{56}$$

It satisfies that

$$\frac{\partial}{\partial v_i} f(v_1, \cdots, v_m) = \frac{h(v_i)h'(v_i)}{\sqrt{\sum_{i\in[m]} h(v_i)^2}} \tag{57}$$

Thus,

$$\|\nabla f(v_1, \cdots, v_m)\|_2 \le L \tag{58}$$

By the property of Lipschitz functions of Gaussian variables (e.g., Theorem 2.4. in [Mar15]), we know that

$$\mathbf{Pr}\left[f(v_1, \cdots, v_m) \ge \mathbb{E}[f(v_1, \cdots, v_m)] + BL\right] \le 2e^{-B^2/2} \tag{59}$$

Now let us consider $\mathbb{E}[f(v_1, \cdots, v_m)]$, using Jensen's inequality we know that

$$\mathbb{E}[f(v_1, \cdots, v_m)] \le \sqrt{\mathbb{E}[z']} \le \sqrt{m\mathbb{E}[h(v_i)^2]} \le \sqrt{mL^2\mathbb{E}[v_i^2]} = L\sqrt{m} \tag{60}$$

Thus, we have that

$$\mathbf{Pr}\left[z' \ge L^2(\sqrt{m} + B)^2\right] \le 2e^{-B^2/2} \tag{61}$$

Putting together Eq (55) and Eq (61) we complete the proof. $\square$

Proposition A.3. *If X_1, X_2 are independent, and X_1, X_3 are independent conditional on X_2, then X_1 and X_3 are independent.*

Proof. For every x_1, x_2, x_3:

$$\begin{aligned} p[X_1 = x_1, X_3 = x_3 \mid X_2 = x_2] &= p[X_1 = x_1 \mid X_2 = x_2]p[X_3 = x_3 \mid X_2 = x_2] && (62)\\ &= p[X_1 = x_1]p[X_3 = x_3 \mid X_2 = x_2]. && (63)\end{aligned}$$

Multiplying $p[X_2 = x_2]$ on both side leads to:

$$p[X_1 = x_1, X_3 = x_3, X_2 = x_2] = p[X_1 = x_1]p[X_3 = x_3, X_2 = x_2]. \tag{64}$$

Marginalizing away X_2 gives $\mathbf{Pr}[X_1 = x_1, X_3 = x_3] = \mathbf{Pr}[X_1 = x_1]\,\mathbf{Pr}[X_3 = x_3]$, so X_1 and X_3 are independent. $\square$

A.2 Interval Partition

Lemma A.4 (Interval Partition)**.** *For every $\tau \le \frac{1}{100}$, there exists a function $\mathfrak{s}\colon [-1,1] \times \mathbb{R} \to \{-1, 0, 1\}$ and a set $I(y) \subset \mathbb{R}$ for every $y \in [-1,1]$ such that, for every $y \in [-1,1]$,*

1. *(Indicator). $\mathfrak{s}(y, g) = 0$ if $g \notin I(y)$, and $\mathfrak{s}(y, g) \in \{-1, 1\}$ otherwise.*
2. *(Balanced). $\mathbf{Pr}_{g\sim\mathcal{N}(0,1)}[g \in I(y)] = \tau$ for every $y \in [-1, 1]$.*
3. *(Symmetric). $\mathbf{Pr}_{g\sim\mathcal{N}(0,1)}[\mathfrak{s}(y, g) = 1] = \mathbf{Pr}_{g\sim\mathcal{N}(0,1)}[\mathfrak{s}(y, g) = -1]$.*
4. *(Unbiased). $\mathbb{E}_{g\sim\mathcal{N}(0,1)}[\mathfrak{s}(y, g)g \mid g \in I(y)] = y$.*
5. *(Bounded). $\max_{x\in I(y)}\{x\} - \min_{x\in I(y)}\{x\} \le 10\tau$.*
6. *(Lipschitz). $|I(y_1)\triangle I(y_2)| \le O(|y_2 - y_1|)$, where $|I| \stackrel{\text{def}}{=} \int_{x\in I} dx$ is the measure of set $I \subseteq \mathbb{R}$.*

We refer to $I(y)$ as an "interval" although it may actually consist of two disjoint closed intervals.

Proof of Lemma A.4. Let us just prove the case when $y \ge 0$ and the other case is by symmetry. It is clear that, since there are only two degrees of freedom, there is a unique interval $I_1(y) = [y - a(y), y + b(y)]$ with $a(y), b(y) \ge 0$ such that

1. (Half probability). $\mathbf{Pr}_{g\sim\mathcal{N}(0,1)}[g \in I_1(y)] = \frac{\tau}{2}$.
2. (Unbiased). $\mathbb{E}_{g\sim\mathcal{N}(0,1)}[g \mid g \in I_1(y)] = y$.

Next, consider two cases:

1. Suppose $[y - a(y), y + b(y)]$ and $[-y - b(y), -y + a(y)]$ are disjoint. In this case, we just define $I(y) \stackrel{\text{def}}{=} [y - a(y), y + b(y)] \cup [-y - b(y), -y + a(y)]$ and define

$$\mathfrak{s}(y, g) \stackrel{\text{def}}{=} \begin{cases} 1 & \text{if } g \in [y - a(y), y + b(y)]; \\ -1 & \text{if } g \in [-y - b(y), -y + a(y)]; \\ 0 & \text{otherwise.} \end{cases} \tag{65}$$

2. $[y-a(y), y+b(y)]$ and $[-y-b(y), -y+a(y)]$ intersect, in this case, consider the unique interval

$$I_2(y) = [-e(y), e(y)]$$

where $e(y) \geq 0$ is defined so that

$$\mathbb{E}_{g\sim\mathcal{N}(0,1)}[g \mid g \in I_2(y) \ \wedge \ g > 0] = \mathbb{E}_{g\sim\mathcal{N}(0,1)}[|g| \mid g \in I_2(y)] = y \ .$$

It must satisfy $\mathbf{Pr}_{g\sim\mathcal{N}(0,1)}[g \in I_2(y) \wedge g > 0] < \tau/2$, because otherwise we must have $y - a(y) \geq 0$ and the two intervals should not have intersected.

Define $\tau'(y) = \mathbf{Pr}_{g\sim\mathcal{N}(0,1)}[g \in I_2(y)] < \tau$. Let $c(y) > e(y)$ be the unique positive real such that

$$\mathop{\mathbf{Pr}}_{g\sim\mathcal{N}(0,1)}[g \in [e(y), c(y)]] = \frac{\tau - \tau'(y)}{2} \ . \tag{66}$$

Let $d(y) \in [e(y), c(y)]$ be the unique real such that

$$\mathop{\mathbf{Pr}}_{g\sim\mathcal{N}(0,1)}[g \in [e(y), d(y)]] = \mathop{\mathbf{Pr}}_{g\sim\mathcal{N}(0,1)}[g \in [d(y), c(y)]] \ . \tag{67}$$

Finally, we define $I(y) = [-c(y), c(y)]$ and

$$\mathfrak{s}(y, g) \stackrel{\text{def}}{=} \begin{cases} 1 & \text{if } g \in [0, e(y)] \cup [e(y), d(y)] \cup [-d(y), -e(y)]; \\ -1 & \text{if } g \in [-e(y), 0] \cup [d(y), c(y)] \cup [-c(y), -d(y)]; \\ 0 & \text{otherwise.} \end{cases} \tag{68}$$

In both cases, one can carefully verify that properties 1, 2, 3, 4 hold. Property 5 follows from the standard property of Gaussian random variable under condition $\tau \leq 1/100$ and $y \in [-1, 1]$.

To check the final Lipschitz continuity property, recall for a standard Gaussian distribution, inside interval $[-\frac{1}{10}, \frac{1}{10}]$ it behaves, up to multiplicative constant factor, similar to a uniform distribution. Therefore, the above defined functions $a(y)$ and $b(y)$ are $O(1)$-Lipschitz continuous in y. Let $y_0 \geq 0$ be the unique constant such that $y - a(y) = 0$ (it is unique because $y - a(y)$ monotonically decreases as $y \to 0+$. It is clear that for $y_0 \leq y_1 \leq y_2$ it satisfies

$$|I(y_1) \triangle I(y_2)| \leq O(y_2 - y_1) \ .$$

As for the turning point of $y = y_0$, it is clear that

$$\lim_{y \to y_0+} I(y) = [-y_0 - b(y_0), y_0 + b(y_0)] = [-e(y_0), e(y_0)] = \lim_{y \to y_0-} I(y)$$

so the function $I(\cdot)$ is continuous at point $y = y_0$. Finally, consider $y \in [-y_0, y_0]$. One can verify that $e(y)$ is $O(1)$-Lipschitz continuous in y, and therefore the above defined $\tau'(y)$, $d(y)$ and $c(y)$ are also $O(1)$-Lipschitz in y. This means, for $-y_0 \leq y_1 \leq y_2 \leq y_0$, it also satisfies

$$|I(y_1) \triangle I(y_2)| \leq O(y_2 - y_1) \ .$$

This proves the Lipschitz continuity of $I(y)$. □

A.3 Hermite polynomials

Definition A.5. *Let $h_i (i \geq 0)$ denote the degree-i (probabilists') Hermite polynomial*

$$h_i(x) \stackrel{\text{def}}{=} i! \sum_{m=0}^{\lfloor i/2 \rfloor} \frac{(-1)^m}{m!(i-2m)!} \frac{x^{i-2m}}{2^m}$$

satisfying the orthogonality constraint

$$\mathbb{E}_{x \sim \mathcal{N}(0,1)}[h_i(x) h_j(x)] = \sqrt{2\pi} j! \delta_{i,j}.$$

Lemma A.6. *(a) For even $i > 0$, for any $x_1 \in [0,1]$ and b,*

$$\mathbb{E}_{\alpha,\beta \sim \mathcal{N}(0,1)} \left[h_i \left(\alpha x_1 + \beta \sqrt{1 - x_1^2} \right) \mathbb{I}[\alpha \geq b] \right] = p_i x_1^i, \ \textit{where} \tag{69}$$

$$p_i = (i-1)!! \frac{\exp(-b^2/2)}{\sqrt{2\pi}} \sum_{r=1, r \ odd}^{i-1} \frac{(-1)^{\frac{i-1-r}{2}}}{r!!} \binom{i/2 - 1}{(r-1)/2} b^r. \tag{70}$$

(b) For odd i, for any $x_1 \in [0,1]$,

$$\mathbb{E}_{\alpha,\beta \sim \mathcal{N}(0,1)} \left[h_i \left(\alpha x_1 + \beta \sqrt{1 - x_1^2} \right) \mathbb{I}[\alpha \geq 0] \right] = p_i x_1^i, \ \textit{where} \tag{71}$$

$$p_i = \frac{1}{\sqrt{2\pi}} (i-1)!! \sum_{k=0, k \ even}^{i} \binom{i/2}{k/2} (-1)^{k/2}. \tag{72}$$

Proof. Using the summation formula, we have:

$$h_i \left(\alpha x_1 + \beta \sqrt{1 - x_1^2} \right) = \sum_{k=0}^{i} \binom{i}{k} (\alpha x_1)^{i-k} h_k \left(\beta \sqrt{1 - x_1^2} \right). \tag{73}$$

Using the multiplication formula, we have:

$$h_k \left(\beta \sqrt{1 - x_1^2} \right) = \sum_{j=1}^{\lfloor \frac{k}{2} \rfloor} \left(\sqrt{1 - x_1^2} \right)^{k-2j} (-x_1^2)^j \binom{k}{2j} \frac{(2j)!}{j!} 2^{-j} h_{k-2j}(\beta). \tag{74}$$

For even k, since $\mathbb{E}_{\beta \sim \mathcal{N}(0,1)}[h_n(\beta)] = 0$ for $n > 0$, we have

$$\mathbb{E}_{\beta \sim \mathcal{N}(0,1)} \left[h_k \left(\beta \sqrt{1 - x_1^2} \right) \right] = \left(-x_1^2 \right)^{k/2} \frac{k!}{(k/2)!} 2^{-k/2}, \tag{75}$$

and for odd k,

$$\mathbb{E}_{\beta \sim \mathcal{N}(0,1)} \left[h_k \left(\beta \sqrt{1 - x_1^2} \right) \right] = 0. \tag{76}$$

This implies

$$h_i \left(\alpha x_1 + \beta \sqrt{1 - x_1^2} \right) = \sum_{k=0, k \text{ even}}^{i} \binom{i}{k} (\alpha x_1)^{i-k} \left(-x_1^2 \right)^{k/2} \frac{k!}{(k/2)!} 2^{-k/2} \tag{77}$$

$$= x_1^i \sum_{k=0, k \text{ even}}^{i} \binom{i}{k} \alpha^{i-k} \frac{k!}{(k/2)!} (-2)^{-k/2}. \tag{78}$$

Therefore,

$$\mathbb{E}_{\alpha,\beta\sim\mathcal{N}(0,1)}\left[h_i\left(\alpha x_1+\beta\sqrt{1-x_1^2}\right)\mathbb{I}[\alpha\geq b]\right] \tag{79}$$

$$= x_1^i\sum_{k=0,k\text{ even}}^{i}\binom{i}{k}\mathbb{E}_{\alpha\sim\mathcal{N}(0,1)}[\alpha^{i-k}\mathbb{I}[\alpha\geq b]]\frac{k!}{(k/2)!}(-2)^{-k/2}. \tag{80}$$

Define $L_{i,b}$ as:

$$L_{i,b}:=\mathbb{E}_{\alpha\sim\mathcal{N}(0,1)}[\alpha^i\mathbb{I}[\alpha\geq b]]. \tag{81}$$

(a) Now consider even $i>0$. by Lemma A.7, we have for even $i\geq 0$:

$$L_{i,b}=(i-1)!!\Phi(0,1;b)+\phi(0,1;b)\sum_{j=1,j\text{ odd}}^{i-1}\frac{(i-1)!!}{j!!}b^j. \tag{82}$$

So

$$\mathbb{E}_{\alpha,\beta\sim\mathcal{N}(0,1)}\left[h_i\left(\alpha x_1+\beta\sqrt{1-x_1^2}\right)\mathbb{I}[\alpha\geq b]\right] \tag{83}$$

$$= x_1^i\left(\sum_{k=0,k\text{ even}}^{i}\binom{i}{k}L_{i-k,b}\frac{k!}{(k/2)!}(-2)^{-k/2}\right) \tag{84}$$

$$= x_1^i\left(\sum_{k=0,k\text{ even}}^{i}\binom{i}{k}(i-k-1)!!\Phi(0,1;b)\frac{k!}{(k/2)!}(-2)^{-k/2}\right) \tag{85}$$

$$+ x_1^i\phi(0,1;b)\left(\sum_{k=0,k\text{ even}}^{i}\binom{i}{k}\left(\sum_{j=1,j\text{ odd}}^{i-k-1}\frac{(i-k-1)!!}{j!!}b^j\right)\frac{k!}{(k/2)!}(-2)^{-k/2}\right). \tag{86}$$

Since

$$\sum_{k=0,k\text{ even}}^{i}\binom{i}{k}(i-k-1)!!\frac{k!}{(k/2)!}(-2)^{-k/2}=(i-1)!!\sum_{k=0,k\text{ even}}^{i}\binom{i/2}{k/2}(-1)^{k/2}=0, \tag{87}$$

we know that

$$\mathbb{E}_{\alpha,\beta\sim\mathcal{N}(0,1)}\left[h_i\left(\alpha x_1+\beta\sqrt{1-x_1^2}\right)\mathbb{I}[\alpha\geq b]\right]=x_1^i(i-1)!!\phi(0,1;b)\sum_{r=1,r\text{ odd}}^{i-1}c_rb^r \tag{88}$$

where c_r is given by:

$$c_r:=\frac{1}{(i-1)!!}\sum_{k=0,k\text{ even}}^{i-1-r}\binom{i}{k}\frac{(i-k-1)!!}{r!!}\frac{k!}{(k/2)!}(-2)^{-k/2} \tag{89}$$

$$=\sum_{k=0,k\text{ even}}^{i-1-r}\binom{i/2}{k/2}\frac{(-1)^{k/2}}{r!!}=\frac{(-1)^{(i-1-r)/2}}{r!!}\binom{i/2-1}{(i-1-r)/2} \tag{90}$$

$$=\frac{(-1)^{\frac{i-1-r}{2}}}{r!!}\binom{i/2-1}{(r-1)/2}. \tag{91}$$

(b) Now consider odd i. Taking $b=0$, we have $L_{i,0}=\phi(0,1;0)(i-1)!!=\frac{(i-1)!!}{\sqrt{2\pi}}$. Plugging into

(79) leads to:

$$\mathbb{E}_{\alpha,\beta\sim\mathcal{N}(0,1)}\left[h_i\left(\alpha x_1+\beta\sqrt{1-x_1^2}\right)\mathbb{I}[\alpha\geq 0]\right] \tag{92}$$

$$=\frac{x_1^i}{\sqrt{2\pi}}\left(\sum_{k=0,k\text{ even}}^{i}\binom{i}{k}(i-k-1)!!\frac{k!}{(k/2)!}(-2)^{-k/2}\right) \tag{93}$$

$$=\frac{x_1^i}{\sqrt{2\pi}}(i-1)!!\sum_{k=0,k\text{ even}}^{i}\binom{i/2}{k/2}(-1)^{k/2}. \tag{94}$$

The proof is completed. □

Lemma A.7. *Define $L_{i,b}$ as:*

$$L_{i,b}:=\mathbb{E}_{\alpha\sim\mathcal{N}(0,1)}[\alpha^i\mathbb{I}[\alpha\geq b]]. \tag{95}$$

Then $L_{i,b}$'s are given by the recursive formula:

$$L_{0,b}=\Phi(0,1;b):=\Pr_{\alpha\sim\mathcal{N}(0,1)}[\alpha\geq b], \tag{96}$$

$$L_{1,b}=\phi(0,1;b):=\mathbb{E}_{\alpha\sim\mathcal{N}(0,1)}[\alpha\mathbb{I}[\alpha\geq b]]=\frac{\exp(-b^2/2)}{\sqrt{2\pi}}, \tag{97}$$

$$L_{i,b}=b^{i-1}\phi(0,1;b)+(i-1)L_{i-2,b}. \tag{98}$$

As a result:

$$L_{i,b}=(i-1)!!\Phi(0,1;b)+\phi(0,1;b)\sum_{j=1,j\ \ odd}^{i-1}\frac{(i-1)!!}{j!!}b^j \tag{99}$$

Proof. The base cases $L_{0,b}$ and $L_{1,b}$ are easy to verify. Then the lemma comes from induction. □

A.4 Optimization

Lemma A.8 (Hessian or Gradient)**.** *For every $m_3>0$, every 3-order differentiable function $f:\mathbb{R}^{m_3}\to\mathbb{R}$. Suppose for value $\varepsilon>0$, fixed vectors $x,x_1\in\mathbb{R}^{m_3}$ and a random variable $x_2\in\mathbb{R}^{m_3}$ with $\mathbb{E}[x_2]=0$ and $\|x_2\|_2=1$ we have: for every sufficiently small $\eta>0$:*

$$\mathbb{E}_{x_2}[f(x+\eta x_1+\sqrt{\eta}x_2)]\leq f(x)-\eta\varepsilon \tag{100}$$

Then either $\|\nabla f(x)\|_2\geq\frac{\varepsilon}{2\|x_1\|_2}$ or $\lambda_{\min}(\nabla^2 f(x))\leq-\varepsilon$, where $\lambda_{\min}$ is the minimal eigenvalue.

Proof of Lemma A.8. We know that

$$f(x+\eta x_1+\sqrt{\eta}x_2) \tag{101}$$

$$=f(x)+\langle\nabla f(x),\eta x_1+\sqrt{\eta}x_2\rangle+\frac{1}{2}(\eta x_1+\sqrt{\eta}x_2)^\top\nabla^2 f(x)(\eta x_1+\sqrt{\eta}x_2)\pm O(\eta^{1.5}). \tag{102}$$

Taking expectation, we know that

$$\mathbb{E}[f(x+\eta x_1+\sqrt{\eta}x_2)]=f(x)+\eta\langle\nabla f(x),x_1\rangle+\eta\frac{1}{2}\mathbb{E}\left[x_2^\top\nabla^2 f(x)x_2\right]\pm O(\eta^{1.5}) \tag{103}$$

Thus, either $\langle\nabla f(x),x_1\rangle\leq-\varepsilon/2$ or $\mathbb{E}\left[x_2^\top\nabla^2 f(x)x_2\right]\leq-\varepsilon$, which completes the proof. □

Lemma A.9 (Escape saddle points, Theorem 6 of [GHJY15])**.** *Suppose a function $L:\mathbb{R}^{m_3}\to\mathbb{R}$ has a stochastic gradient with Euclidean norm bounded by Q, further more, suppose the function is* *bounded by $|L(x)| \le B$ and it is β-smooth and has ρ-Lipschitz Hessian, then for every $\varepsilon > 0$, noisy stochastic gradient descent outputs a point x_t after $\mathbf{poly}(B, Q, \beta, \rho, 1/\varepsilon)$ iterations such that*

$$\|\nabla L(x_t)\|_2 \le \varepsilon, \quad \text{and } \nabla^2 L(x_t) \succeq -\varepsilon \mathbf{I} \tag{104}$$

B Proof of Main Lemma for Three Layers: Existential Results

B.1 Proof of Lemma 5.1

Lemma 5.1 (from indicator to functions)**.** *For every smooth function ϕ, every $\varepsilon \in (0, 1/\mathfrak{C}(\phi, 1))$, we have that there exists a function $h : \mathbb{R}^2 \to \left[-\mathfrak{C}(\phi, \sqrt{\log(1/\varepsilon)}), \mathfrak{C}(\phi, \sqrt{\log(1/\varepsilon)})\right]$ such that for every $x_1 \in [-1, 1]$:*

$$\left|\mathbb{E}\left[\mathbb{I}_{\alpha_1 x_1 + \beta_1\sqrt{1-x_1^2} + b_0 \ge 0} h(\alpha_1, b_0)\right] - \phi(x_1)\right| \le 2\varepsilon \tag{105}$$

where $\alpha_1, \beta_1 \sim \mathcal{N}(0, 2)$ and $b_0 = 0$ w.p. $1/2$, $b_0 \sim \mathcal{N}(0, 2)$ otherwise, are independent random variables. Moreover, h is $\mathfrak{C}(\phi, \sqrt{\log(1/\varepsilon)})$-Lipschitz on the first coordinate.

For notation simplicity, let us denote $w_0 = (\alpha_1, \beta_1)$ and $x = (x_1, \sqrt{1 - x_1^2})$. Reversely, we can also write $\langle w_0, x\rangle = \alpha$ and $\alpha_1 = \alpha x_1 + \sqrt{1 - x_1^2}\beta$ for two independent $\alpha, \beta \sim \mathcal{N}(0, 1)$. Let $\mathcal{D}$ denote the distribution of being 0 with half probability, and $\mathcal{N}(0, 1)$ with the other half probability.

We first make a technical claim involving in fitting monomials in x_1. We shall then use it to fit arbitrary functions $\phi(x_1)$. Its proof is in Section B.1.1.

Claim B.1. *For odd $i \ge 1$, there exists constant $|p_i'| \ge \frac{(i-2)!!}{20}$ such that*

$$x_1^i = \frac{1}{p_i'} \mathbb{E}_{w_0 \sim \mathcal{N}(0, \mathbf{I}), b_0 \sim \mathcal{D}}\left[h_i(\alpha_1) \cdot \mathbb{I}[b_0 = 0] \cdot \mathbb{I}[\langle x, w_0\rangle + b_0 \ge 0]\right].$$

For even $i \ge 2$, there exists constant $|p_i'| \ge \frac{(i-1)!!}{200 i^2}$ such that

$$x_1^i = \frac{1}{p_i'} \mathbb{E}_{w_0 \sim \mathcal{N}(0, \mathbf{I}), b_0 \sim \mathcal{D}}\left[h_i(\alpha_1) \cdot \mathbb{I}[0 < -b_0 \le 1/(2i)] \cdot \mathbb{I}[\langle x, w_0\rangle + b_0 \ge 0]\right].$$

Using Claim B.1, we have

$$\begin{aligned}
\phi(x_1) &= c_0 + \sum_{i=1}^{\infty} c_{2i} x_1^{2i} + \sum_{i=0}^{\infty} c_{2i+1} x_1^{2i+1} &(106)\\
&= c_0 + \sum_{i=1}^{\infty} c_{2i}' \cdot \mathbb{E}_{\alpha, \beta \sim \mathcal{N}(0,1), b_0 \sim \mathcal{D}}\left[h_i(\alpha_1) \cdot \mathbb{I}[b_0 = 0] \cdot \mathbb{I}[\langle x, w_0\rangle + b_0 \ge 0]\right] &(107)\\
&\quad + \sum_{i=0}^{\infty} c_{2i+1}' \cdot \mathbb{E}_{\alpha, \beta \sim \mathcal{N}(0,1), b_0 \sim \mathcal{D}}\left[h_i(\alpha_1) \cdot \mathbb{I}[0 < -b_0 \le 1/(2i)] \cdot \mathbb{I}[\langle x, w_0\rangle + b_0 \ge 0]\right] &(108)
\end{aligned}$$

where

$$c_{2i+1}' \stackrel{\text{def}}{=} \frac{c_{2i+1}}{p_{2i+1}'}, \quad c_{2i}' \stackrel{\text{def}}{=} \frac{c_{2i}}{p_{2i}'} \quad \text{and} \quad |c_{2i+1}'| \le \frac{20\,|c_{2i+1}|}{(2i-1)!!}, \quad |c_{2i}'| \le \frac{200 i^2\,|c_{2i}|}{(2i-1)!!}. \tag{109}$$

The next technical claim carefully bounds the absolute values of the Hermite polynomials. Its proof is in Section B.1.2.

Claim B.2. *Setting* $B_i \stackrel{\text{def}}{=} 100 i^{1/2}\sqrt{\log\frac{1}{\varepsilon}}$, *we have*

$$\sum_{i=1}^{\infty} |c_i'| \cdot \mathbb{E}_{z\sim\mathcal{N}(0,1)}\big[|h_i(z)|\cdot \mathbb{I}[|z|\geq B_i]\big] \leq \epsilon/8 \tag{110}$$

$$\sum_{i=1}^{\infty} |c_i'| \cdot \mathbb{E}_{z\sim\mathcal{N}(0,1)}\big[|h_i(z)|\cdot \mathbb{I}[|z|\leq B_i]\big] \leq \frac{1}{2}\mathfrak{C}\left(\phi, \sqrt{\log\frac{1}{\varepsilon}}\right) \tag{111}$$

Using Claim B.2, we have

$$\phi(x_1) = c_0 + R'(x_1) \tag{112}$$

$$+\sum_{i=1}^{\infty} c_{2i}' \cdot \mathbb{E}_{\alpha,\beta\sim\mathcal{N}(0,1), b_0\sim\mathcal{D}}\left[h_i(\alpha_1)\cdot\mathbb{I}[|\alpha_1|\leq B_{2i}]\cdot\mathbb{I}[b_0=0]\cdot\mathbb{I}[\langle x, w_0\rangle + b_0 \geq 0]\right] \tag{113}$$

$$+\sum_{i=0}^{\infty} c_{2i+1}' \cdot \mathbb{E}_{\alpha,\beta\sim\mathcal{N}(0,1), b_0\sim\mathcal{D}}\left[h_i(\alpha_1)\cdot\mathbb{I}[|\alpha_1|\leq B_{2i+1}]\cdot\mathbb{I}[0<-b_0\leq 1/(2i)]\cdot\mathbb{I}[\langle x, w_0\rangle + b_0 \geq 0]\right] \tag{114}$$

where $|R'(x_1)| < \epsilon/4$ uses (110). In other words, if we define

$$h(\alpha_1, b_0) \stackrel{\text{def}}{=} 2c_0 \tag{115}$$

$$+\sum_{i=1}^{\infty} c_{2i}' \cdot \left(h_i(\alpha_1)\cdot\mathbb{I}[|\alpha_1|\leq B_{2i}] + h_i(\mathrm{sign}(\alpha_1)B_{2i})\cdot\mathbb{I}[|\alpha_1|>B_{2i}]\right)\cdot\mathbb{I}[b_0=0] \tag{116}$$

$$+\sum_{i=0}^{\infty} c_{2i+1}' \cdot \left(h_i(\alpha_1)\cdot\mathbb{I}[|\alpha_1|\leq B_{2i+1}] + h_i(\mathrm{sign}(\alpha_1))\cdot\mathbb{I}[|\alpha_1|>B_{2i+1}]\right)\cdot\mathbb{I}[0<-b_0\leq 1/(2i)] \ . \tag{117}$$

then we have

$$\left|\mathbb{E}_{\alpha,\beta\sim\mathcal{N}(0,1), b_0\sim\mathcal{D}}\big[\mathbb{I}[\langle x, w_0\rangle + b_0\geq 0]\cdot h(\alpha_1, b_0)\big] - \phi(x_1)\right| = \left|R'(x_1)\right| \leq \varepsilon \tag{118}$$

and using (111) we have

$$|h(\alpha_1, b_0)| \leq 2c_0 + \frac{1}{2}\mathfrak{C}\left(\phi, \sqrt{\log\frac{1}{\varepsilon}}\right) \leq \mathfrak{C}\left(\phi, \sqrt{\log\frac{1}{\varepsilon}}\right) \ . \tag{119}$$

This finishes the proof of Lemma 5.1. ∎

B.1.1 Proof of Claim B.1

Proof of Claim B.1. We treat the two cases separately.

Odd i. By Lemma A.6, we know that

$$2\mathbb{E}_{w_0\sim\mathcal{N}(0,\mathbf{I}), b_0\sim\mathcal{D}}\left[h_i(\alpha_1)\cdot\mathbb{I}[b_0=0]\cdot\mathbb{I}[\langle x, w_0\rangle + b_0\geq 0]\right] \tag{120}$$

$$= \mathbb{E}_{w_0\sim\mathcal{N}(0,\mathbf{I})}\left[h_i(\alpha_1)\cdot\mathbb{I}[\langle x, w_0\rangle\geq 0]\right] \tag{121}$$

$$= \mathbb{E}_{\alpha,\beta\sim\mathcal{N}(0,1)}\left[h_i\left(\alpha x_1 + \beta\sqrt{1-x_1^2}\right)\mathbb{I}[\alpha\geq 0]\right] = p_i x_1^i \tag{122}$$

where

$$p_i = \frac{1}{\sqrt{2\pi}}(i-1)!! \sum_{k=0, k \text{ even}}^{i} \binom{i/2}{k/2}(-1)^{k/2} \ . \tag{123}$$

We can bound p_i as follows. Since

$$\sum_{k=0,k \text{ even}}^{i} \binom{i/2}{k/2}(-1)^{k/2} = (-1)^{(i-1)/2}\binom{i/2-1}{(i-1)/2}, \tag{124}$$

we have

$$|p_i| = \left|\frac{1}{\sqrt{2\pi}}(i-1)!!(-1)^{(i-1)/2}\binom{i/2-1}{(i-1)/2}\right| \geq \frac{(i-2)!!}{10}. \tag{125}$$

Even i. Again by Lemma A.6, we know that

$$\begin{aligned}
&2\mathbb{E}_{w_0\sim\mathcal{N}(0,\mathbf{I}),b_0\sim\mathcal{D}}\left[h_i(\alpha_1)\cdot\mathbb{I}[0<-b_0\leq 1/(2i)]\cdot\mathbb{I}[\langle x,w_0\rangle+b_0\geq 0]\right] && (126)\\
&=\mathbb{E}_{w_0\sim\mathcal{N}(0,\mathbf{I}),b_0\sim\mathcal{N}(0,1)}\left[h_i(\alpha_1)\cdot\mathbb{I}[0<-b_0\leq 1/(2i)]\cdot\mathbb{I}[\langle x,w_0\rangle+b_0\geq 0]\right] && (127)\\
&=\mathbb{E}_{b_0\sim\mathcal{N}(0,1)}\left[\mathbb{E}_{\alpha,\beta\sim\mathcal{N}(0,1)}\left[h_i\left(\alpha x_1+\beta\sqrt{1-x_1^2}\right)\cdot\mathbb{I}[\alpha\geq -b_0]\right]\cdot\mathbb{I}[0<-b_0\leq 1/(2i)]\right] && (128)\\
&=\mathbb{E}_{b_0\sim\mathcal{N}(0,1)}\left[p_i\cdot\mathbb{I}[0<-b_0\leq 1/(2i)]\right]\times x_1^i && (129)\\
& && (130)
\end{aligned}$$

where

$$p_i = (i-1)!!\frac{\exp(-b_0^2/2)}{\sqrt{2\pi}}\sum_{r=1,r \text{ odd}}^{i-1}\frac{(-1)^{\frac{i-1-r}{2}}}{r!!}\binom{i/2-1}{(r-1)/2}(-b_0)^r\ . \tag{131}$$

We try to bound the coefficient "$\mathbb{E}_{b_0\sim\mathcal{N}(0,1)}\left[p_i\cdot\mathbb{I}[0<-b_0\leq 1/(2i)]\right]$" as follows. Define c_r as:

$$c_r := \frac{(-1)^{\frac{i-1-r}{2}}}{r!!}\binom{i/2-1}{(r-1)/2}. \tag{132}$$

Then, for $0\leq -b_0\leq \frac{1}{2i}$, we know that for all $1<r\leq i-1$, r odd:

$$|c_r(-b_0)^r| \leq \frac{1}{4}|c_{r-2}(-b_0)^{r-2}|, \tag{133}$$

which implies

$$\left|\sum_{r=1,r \text{ odd}}^{i-1} c_r(-b_0)^r\right| \geq \frac{2}{3}|c_1 b_0| = \frac{2}{3}|b_0| \tag{134}$$

and

$$\operatorname{sign}\left(\sum_{r=1,r \text{ odd}}^{i-1} c_r(-b_0)^r\right) = \operatorname{sign}(c_1) \tag{135}$$

is independent of the randomness of b_0. Therefore, using the formula of p_i in (131):

$$\begin{aligned}
&\left|\mathbb{E}_{b_0\sim\mathcal{N}(0,1)}[p_i\cdot\mathbb{I}[0\leq -b_0\leq 1/(2i)]]\right| && (136)\\
&=\left|\mathbb{E}_{b_0\sim\mathcal{N}(0,1)}\left[(i-1)!!\frac{\exp(-b_0^2/2)}{\sqrt{2\pi}}\sum_{r=1,r \text{ odd}}^{i-1}c_r(-b_0)^r\cdot\mathbb{I}[0\leq -b_0\leq 1/(2i)]\right]\right| && (137)\\
&\geq \mathbb{E}_{b_0\sim\mathcal{N}(0,1)}\left[(i-1)!!\frac{\exp(-b_0^2/2)}{\sqrt{2\pi}}\frac{2}{3}|b_0|\cdot\mathbb{I}[0\leq -b_0\leq 1/(2i)]\right] && (138)\\
&\geq \frac{(i-1)!!}{100i^2}. && (139)
\end{aligned}$$

$\square$

B.1.2 Proof of Claim B.2

Proof of Claim B.2. By the definition of Hermite polynomial (see Definition A.5), we have that

$$|h_i(x)| \le \sum_{j=0}^{\lfloor i/2 \rfloor} \frac{i!|x|^{i-2j}}{j!(i-2j)!2^j} \le \sum_{j=0}^{\lfloor i/2 \rfloor} |x|^{i-2j} i^{2j} \tag{140}$$

and this implies for $|z| \ge 1$,

$$|c_i' h_i(z)| \le O(1)|c_i| \frac{i^4}{i!!} \sum_{j=0}^{\lfloor i/2 \rfloor} |z|^{i-2j} i^{2j} \tag{141}$$

As a result, denoting by $b = B_i$, we have

$$|c_i'| \cdot \mathbb{E}_{z\sim\mathcal{N}(0,1)}[|h_i(z)| \cdot \mathbb{I}[|z| \ge b]]| \le 2|c_i'| \cdot \mathbb{E}_{z\sim\mathcal{N}(0,1)}\left[\sum_{j=0}^{\lfloor i/2 \rfloor} |z|^{i-2j} i^{2j} \mathbb{I}[z \ge b]\right] \tag{142}$$

$$= O(1)|c_i| \frac{i^4}{i!!} \sum_{j=0}^{\lfloor i/2 \rfloor} L_{i-2j,b} \cdot i^{2j} \ , \tag{143}$$

where recall from Lemma A.7 that

$$L_{i,b} = (i-1)!!\Phi(0,1;b) + \phi(0,1;b) \sum_{j=1, j \text{ odd}}^{i-1} \frac{(i-1)!!}{j!!} b^j \tag{144}$$

$$\le O(1) e^{-b^2/2} (i-1)!! \cdot i \cdot b^i \tag{145}$$

$$\le O(1) e^{-b^2/2} (i+1)!! \cdot b^i \tag{146}$$

Thus we have that

$$\sum_{j=0}^{\lfloor i/2 \rfloor} L_{i-2j,b} \cdot i^{2j} \le O(1) \sum_{j=0}^{\lfloor i/2 \rfloor} e^{-b^2/2} (i+1-2j)!! \cdot i^{2j} \cdot b^{i-2j} \tag{147}$$

$$\le O(1) \sum_{j=0}^{\lfloor i/2 \rfloor} e^{-b^2/2} \cdot (i+1)!! \cdot (10i)^j \cdot b^{i-2j} \tag{148}$$

$$\overset{①}{\le} O(1) \cdot (i+3)!! \cdot e^{-b^2/2} \cdot b^i \tag{149}$$

$$\overset{②}{\le} O(1) \cdot \left(100 i^{1/2} \sqrt{\log \frac{1}{\varepsilon}}\right)^{2i} \cdot e^{-10^4 i \log \frac{1}{\varepsilon}} \tag{150}$$

$$\overset{③}{\le} \left(100 i^{1/2} \sqrt{\log \frac{1}{\varepsilon}}\right)^{2i} (\varepsilon)^{10^3 i} \frac{1}{|c_i|} \tag{151}$$

$$\le (\varepsilon)^{10i} i^i \frac{1}{|c_i|} \tag{152}$$

Above, inequality ① uses $b^{2j} = B_i^{2j} \ge (10i)^j$; inequality ② uses our definition of $b = B_i$; and inequality ③ uses $\varepsilon|c_i| \le 1$. Putting this back to (143), we have

$$\sum_{i=1}^{\infty} |c_i'| \cdot \mathbb{E}_{z\sim\mathcal{N}(0,1)}[|h_i(z)| \cdot \mathbb{I}[|z| \ge b]]| \le O(1) \sum_{i=1}^{\infty} \frac{i^4}{i!!} (\varepsilon)^{10i}\, i^i \le \epsilon/8. \tag{153}$$

Above, in the last inequality we have used $\frac{i^4}{i!!} i^i \le 40 \cdot 4^i$ for $i \ge 1$.

Moreover, by Eq (160), it holds that

$$\sum_{i=1}^{\infty} |c_i'| \cdot \mathbb{E}_{z\sim\mathcal{N}(0,1)} \left[|h_i(z)| \cdot \mathbb{I}[|z| \le B_i] + |h_i(B_i)| \cdot \mathbb{I}[|z| > B_i] \right] \tag{154}$$

$$\le O(1) \sum_{i=1}^{\infty} |c_i| \frac{i^4}{i!!} \sum_{j=0}^{\lfloor i/2 \rfloor} \frac{B_i^{i-2j} i^{2j}}{j!} \tag{155}$$

$$\le O(1) \sum_{i=1}^{\infty} |c_i| \frac{i^4}{i!!} \sum_{j=0}^{\lfloor i/2 \rfloor} \left(100 i^{1/2} \sqrt{\log \frac{1}{\varepsilon}} \right)^{i-2j} \frac{i^{2j}}{j!} \tag{156}$$

$$\le O(1) \sum_{i=1}^{\infty} |c_i| \frac{i^4}{i!!} \left(100 i^{1/2} \sqrt{\log \frac{1}{\varepsilon}} \right)^{i} \sum_{j=0}^{\lfloor i/2 \rfloor} \frac{i^{j}}{j!} \tag{157}$$

$$\overset{①}{\le} \sum_{i=1}^{\infty} |c_i| \left(O(1) \sqrt{\log \frac{1}{\varepsilon}} \right)^{i} \tag{158}$$

$$\le \frac{1}{2} \mathfrak{C} \left(\phi, \sqrt{\log \frac{1}{\varepsilon}} \right) \ . \tag{159}$$

Here, in ① we use the fact that $\frac{i^j}{j!} \le 10^i$.

In the end, using the fact that

$$|c_i' h_i(z)| \le O(1) |c_i| \frac{i^5}{i!!} \sum_{j=0}^{\lfloor i/2 \rfloor} |z|^{i-2j} i^{2j} \tag{160}$$

We can also bound the Lipschitzness of h on the first coordinate. $\square$

B.2 Proof of Lemma 5.2

Let us consider a single neural of the second layer at initialization, given as:

$$n_1(x) = \sum_{i\in[m_1]} v_{1,i}^{(0)} \sigma \left(\langle w_i^{(0)}, x \rangle + b_{1,i}^{(0)} \right) \tag{161}$$

Let $W^{(0)} = [w_1^{(0)}; \dots; w_{m_1}^{(0)}] \in \mathbb{R}^{m_1 \times d}$, and $v_1^{(0)} = [v_{1,i}^{(0)}]_i, b_1^{(0)} = [b_{1,i}^{(0)}]_i$.

Proof of Lemma 5.2. Without loss of generality we assume $w^* = e_1$. Let $C = \mathfrak{C}(\phi, \sqrt{\log(1/\varepsilon)})$. Recall that

$$n_1(x) = \sum_{i\in[m_1]} v_{1,i}^{(0)} \sigma \left(\langle w_i^{(0)}, x \rangle + b_{1,i}^{(0)} \right) \tag{162}$$

By Lemma 5.1, for every $\varepsilon > 0$, there exists a function h such that for every unit x with $x_d = \frac{1}{2}$

and every $i \in [m_1]$:

$$\mathbb{E}_{w_i^{(0)}}\left[h\left(w_{i,1}^{(0)}, b_{1,i}^{(0)}\right) x_d \mathbb{I}[\langle w_i^{(0)}, x\rangle + b_{1,i}^{(0)} \geq 0]\right] = \frac{\phi_\varepsilon(x_1)}{C} = \frac{\phi_\varepsilon(\langle w^*, x\rangle)}{C} \tag{163}$$

with

$$|\phi_\varepsilon(\langle w^*, x\rangle) - \phi(\langle w^*, x\rangle)| \leq \frac{2\varepsilon}{C} \tag{164}$$

and $\left|h\left(w_{i,1}^{(0)}, b_{1,i}^{(0)}\right)\right| \in [0,1]$ and h is 1-Lipschitz.

Throughout the proof, we fix some parameter τ (that we shall in the end choose $\tau = \frac{1}{100}$). Let us construct the sign function $\mathfrak{s}\colon [-1,1] \times \mathbb{R} \to \{-1,0,1\}$ and the set function $I\colon [-1,1] \ni y \mapsto I(y) \subset \big[-10\tau, 10\tau\big]$ given in Lemma A.4. Now, for every $w_{i,1}^{(0)}$, define

$$I_i \stackrel{\text{def}}{=} I\left(h\left(w_{i,1}^{(0)}, b_{1,i}^{(0)}\right)\right) \subset \big[-10\tau, 10\tau\big] \tag{165}$$

Also define set

$$\mathcal{S} \stackrel{\text{def}}{=} \left\{i \in [m_1]\colon \sqrt{m_2} v_{1,i}^{(0)} \in I_i\right\} \quad . \tag{166}$$

We define what we call "*effective sign* of $v_{1,i}^{(0)}$" to be s_i such that

$$s_i = \mathfrak{s}\left(h\left(w_{i,1}^{(0)}, b_{1,i}^{(0)}\right), \sqrt{m_2} v_{1,i}^{(0)}\right) \tag{167}$$

By the definition of I_i and $\mathfrak{s}$ (see Lemma A.4), we know that $\mathcal{S}$ and the "effective sign" of those $v_{1,i}^{(0)}$ in this set $i \in \mathcal{S}$ are independent of $W^{(0)}$. Indeed, for any fixed choice of $W^{(0)}$, each $i \in [m_1]$ is in set $\mathcal{S}$ with probability τ, and for each $i \in \mathcal{S}$, s_i is ± 1 each with half probability. Therefore, conversely, conditioning on $\mathcal{S} = \mathcal{S}_0$ and $\{s_i\}_{i\in\mathcal{S}} = s$ being fixed, the distribution of $W^{(0)}$ is also unchanged. Define a unit vector $u \in \mathbb{R}^{m_1}$ such that

$$u_i = \begin{cases} \frac{s_i}{\sqrt{|\mathcal{S}|}} & \text{if } i \in \mathcal{S}; \\ 0 & \text{if } i \notin \mathcal{S}. \end{cases} \tag{168}$$

Since the entries of $W^{(0)}$ are i.i.d. generated from $\mathcal{N}(0, 1/m_1)$, we can write $W^{(0)} \in \mathbb{R}^{m_1 \times d}$ as

$$W^{(0)} = \alpha u e_d^\top + \beta \tag{169}$$

where $\alpha \sim \mathcal{N}\left(0, \frac{1}{m_1}\right)$ and $\beta \in \mathbb{R}^{m_1 \times d}$ are two independent Gaussian variables given u (the entries of β are not i.i.d.) This factorizes out the randomness of the last column of $W^{(0)}$ along the direction u, and in particular,

- $\alpha = u^\top W^{(0)} e_d$
- α is independent of u.

A simple observation here is that, although $\alpha \sim \mathcal{N}\left(0, \frac{1}{m_1}\right)$, if we fix $W^{(0)}$ then the distribution of α is not Gaussian. Fortunately, fixing $W^{(0)}$, we still have that the vector $s = (s_1, \ldots, s_{m_1})$ are independent (each $s_i = 0$ with probability $1 - \tau$, and the non-zero ones have random signs). By the Wasserstein distance bound of central limit theorem (see for instance [EMZ18]), there exists some random Gaussian $g \sim \mathcal{N}(0, \frac{1}{m_1})$ that is independent of W_0 such that, letting $\alpha|_{W^{(0)}}$ denote the conditional distribution of α, then with high probability over $W^{(0)}$:

$$\mathcal{W}_2(\alpha|_{W^{(0)}}, g) \leq \widetilde{O}\left(\frac{1}{\sqrt{\tau} m_1}\right) \quad .$$

Therefore, we can write

$$n_1(x) = \sum_{i\in[m_1]} v_{1,i}^{(0)}\sigma\left(\langle w_i^{(0)}, x\rangle + b_{1,i}^{(0)}\right) \tag{170}$$

$$= \underbrace{\sum_{i\notin\mathcal{S}} v_{1,i}^{(0)}\sigma\left(\langle w_i^{(0)}, x\rangle + b_{1,i}^{(0)}\right)}_{\stackrel{\text{def}}{=} B_1\left(x, v_1^{(0)}, W^{(0)}, b_1^{(0)}\right)} + \sum_{i\in\mathcal{S}} v_{1,i}^{(0)}\sigma\left(\langle w_i^{(0)}, x\rangle + b_{1,i}^{(0)}\right), \tag{171}$$

By definition of B_1, conditioning on the randomness of u, we know that B_1 is independent of α. Since u and α are independent, we know that α and B_1 are independent by Proposition A.3. We continue to write

$$n_1(x) - B_1 \tag{172}$$

$$= \sum_{i\in\mathcal{S}} v_{1,i}^{(0)}\sigma\left(\langle w_i^{(0)}, x\rangle + b_{1,i}^{(0)}\right) \tag{173}$$

$$= \sum_{i\in\mathcal{S}} v_{1,i}^{(0)}\mathbb{I}[\langle w_i^{(0)}, x\rangle + b_{1,i}^{(0)} \geq 0]\left(\frac{\alpha s_i}{\sqrt{|S|}}x_d + \langle\beta_i, x\rangle + b_{1,i}^{(0)}\right) \tag{174}$$

$$= \underbrace{\sum_{i\in\mathcal{S}} v_{1,i}^{(0)}\mathbb{I}[\langle w_i^{(0)}, x\rangle + b_{1,i}^{(0)} \geq 0]\frac{\alpha s_i}{\sqrt{|S|}}x_d}_{\stackrel{\text{def}}{=} T_3} + \underbrace{\sum_{i\in\mathcal{S}} v_{1,i}^{(0)}\mathbb{I}[\langle w_i^{(0)}, x\rangle + b_{1,i}^{(0)} \geq 0]\left(\langle\beta_i, x\rangle + b_{1,i}^{(0)}\right)}_{\stackrel{\text{def}}{=} T_4} \tag{175}$$

First consider T_3. For those $i \in \mathcal{S}$, we have (where the second row is by the Unbiased property of the interval (see Lemma A.4):

$$\left|s_i \cdot v_{1,i}^{(0)} - \frac{1}{\sqrt{m_2}}h\left(w_{i,1}^{(0)}, b_{1,i}^{(0)}\right)\right| \leq O\left(\frac{\tau}{\sqrt{m_2}}\right) \quad \text{and} \tag{176}$$

$$\mathbb{E}_{v_{1,i}^{(0)}}\left[s_i \cdot v_{1,i}^{(0)} - \frac{1}{\sqrt{m_2}}h\left(w_{i,1}^{(0)}, b_{1,i}^{(0)}\right) \mid i \in \mathcal{S}\right] = 0 \ . \tag{177}$$

By concentration, for fixed vector x, with high probability over the randomness of $V^{(0)}$:

$$\left|\sum_{i\in\mathcal{S}}\left(v_{1,i}^{(0)}s_i - \frac{1}{\sqrt{m_2}}h\left(w_{i,1}^{(0)}, b_{1,i}^{(0)}\right)\right)\mathbb{I}[\langle w_i^{(0)}, x\rangle + b_{1,i}^{(0)} \geq 0]\right| \leq \widetilde{O}\left(\frac{\tau\sqrt{|S|}}{\sqrt{m_2}}\right). \tag{178}$$

In other words,

$$T_3 = \sum_{i\in\mathcal{S}}\frac{\alpha}{\sqrt{m_2|\mathcal{S}|}}h\left(w_{i,1}^{(0)}, b_{1,i}^{(0)}\right)x_d\mathbb{I}[\langle w_i^{(0)}, x\rangle + b_{1,i}^{(0)} \geq 0] + R_1 \tag{179}$$

where $R_1 = R_1\left(x, v_1^{(0)}, W^{(0)}, b_1^{(0)}\right)$ satisfies $|R_1| \leq \widetilde{O}\left(\frac{\tau}{\sqrt{m_1 m_2}}\right)$. We write

$$\frac{T_3 - R_1}{\alpha} = \sum_{i\in\mathcal{S}}\frac{1}{\sqrt{m_2|\mathcal{S}|}}h\left(w_{i,1}^{(0)}, b_{1,i}^{(0)}\right)x_d\mathbb{I}[\langle w_i^{(0)}, x\rangle + b_{1,i}^{(0)} \geq 0] =: T_5. \tag{180}$$

By (163) (i.e., the property of h), we know that for every fixed x, using concentration bound, with high probability over $W^{(0)}$:

$$\left|T_5 - \frac{\sqrt{|\mathcal{S}|}}{\sqrt{m_2}C}\phi_\varepsilon(\langle w^*, x\rangle)\right| \leq \widetilde{O}\left(\frac{1}{\sqrt{m_2}}\right). \tag{181}$$

Since each $i \in \mathcal{S}$ with probability τ (because $\sqrt{m_1} v_{1,i}^{(0)} \sim \mathcal{N}(0,1)$), we know with high probability:

$$||\mathcal{S}| - \tau m_1| \le O(\sqrt{\tau m_1}) \ , \tag{182}$$

and thus

$$\left| \frac{C\sqrt{m_2}}{\sqrt{\tau m_1}} T_5 - \phi_\varepsilon(\langle w^*, x\rangle) \right| \le \widetilde{O}\left(\frac{C}{\sqrt{\tau m_1}} \right). \tag{183}$$

Let us define

$$\rho\left(v_1^{(0)}, W^{(0)}, b_1^{(0)}\right) \stackrel{\text{def}}{=} \frac{\sqrt{\tau m_1}}{C\sqrt{m_2}} \alpha \sim \mathcal{N}\left(0, \frac{\tau}{C^2 m_2}\right). \tag{184}$$

Then

$$T_1 = \rho\left(v_1^{(0)}, W^{(0)}, b_1^{(0)}\right) \cdot \phi_\varepsilon(\langle w^*, x\rangle) + R_1 + R_2\left(x, v_1^{(0)}, W^{(0)}, b_1^{(0)}\right), \tag{185}$$

$$\text{where } |R_2| \le \widetilde{O}\left(\frac{C}{\sqrt{\tau m_1}} \right) \times \frac{\tau}{C\sqrt{m_2}} = \widetilde{O}\left(\frac{\sqrt{\tau}}{\sqrt{m_1 m_2}} \right) \tag{186}$$

Note that α is independent of u so ρ is also independent of u.

Next consider T_4. For fixed unit vector x, with high probability, we have that

$$\left| \alpha \frac{s_i}{\sqrt{|S|}} \right| = \left| \frac{\alpha}{\sqrt{|S|}} \right| = \widetilde{O}\left(\frac{1}{\sqrt{|S| m_1}} \right) \tag{187}$$

and therefore

$$\langle w_i^{(0)}, x\rangle + b_{1,i}^{(0)} = \frac{\alpha s_i}{\sqrt{|S|}} x_d + \langle \beta_i, x\rangle + b_{1,i}^{(0)} = \langle \beta_i, x\rangle + b_{1,i}^{(0)} \pm \widetilde{O}\left(\frac{1}{\sqrt{|S| m_1}} \right) \ , \tag{188}$$

By the above formula, for an index $i \in \mathcal{S}$ to have $\mathbb{I}[\langle \beta_i, x\rangle + b_{1,i}^{(0)} \ge 0] \ne \mathbb{I}[\langle w_i^{(0)}, x\rangle + b_{1,i}^{(0)} \ge 0]$, it must satisfy $\left|\left(\langle w_i^{(0)}, x\rangle + b_{1,i}^{(0)}\right)\right| \le \widetilde{O}(\frac{1}{\sqrt{|S| m_1}})$. With high probability over the randomness of $W^{(0)}$, there are at most $\widetilde{O}(\sqrt{|S|})$ many such indices $i \in \mathcal{S}$. Therefore, using $|v_{1,i}^{(0)}| \le \widetilde{O}(m_2^{-1/2})$, we have

$$T_4 = \underbrace{\sum_{i \in \mathcal{S}} v_{1,i}^{(0)} \mathbb{I}[\langle \beta_i, x\rangle + b_{1,i}^{(0)} \ge 0] \left(\langle \beta_i, x\rangle + b_{1,i}^{(0)}\right)}_{\stackrel{\text{def}}{=} T_6} + R_3 \tag{189}$$

with $R_3 = R_3\left(x, v_1^{(0)}, W^{(0)}, b_1^{(0)}\right)$ satisfying $|R_3| \le \widetilde{O}\left(\frac{1}{\sqrt{m_1 m_2}}\right)$.

Finally, to bound T_6, we recall

$$\mathcal{S} \stackrel{\text{def}}{=} \left\{ i \in [m_1] \colon \sqrt{m_2} v_{1,i}^{(0)} \in I\left(h\left(w_{i,1}^{(0)}, b_{1,i}^{(0)}\right)\right) \right\} \ . \tag{190}$$

and define a similar notion

$$\mathcal{S}' \stackrel{\text{def}}{=} \left\{ i \in [m_1] \colon \sqrt{m_2} v_{1,i}^{(0)} \in I\left(h\left(\beta_{i,1}, b_{1,i}^{(0)}\right)\right) \right\} \ . \tag{191}$$

Since h is 1-Lipschitz in the first variable, we further have that

$$h\left(w_{i,1}^{(0)}, b_{1,i}^{(0)}\right) = h\left(\beta_{i,1} + \alpha \frac{s_i}{\sqrt{|S|}}, b_{1,i}^{(0)}\right) = h\left(\beta_{i,1}, b_{1,i}^{(0)}\right) \pm \widetilde{O}\left(\frac{1}{\sqrt{|S| m_1}} \right) \ . \tag{192}$$

By the Lipschitz property of intervals (see Lemma A.4, the measure of the symmetric difference $I\big(h(\beta_{i,1}, b_{1,i}^{(0)})\big) \triangle I\big(h(w_{i,1}^{(0)}, b_{1,i}^{(0)})\big)$ is at most $\widetilde{O}(\frac{1}{\sqrt{|S| m_1}}) \le \widetilde{O}(\frac{1}{\sqrt{\tau} m_1})$. Therefore, we know that

$|\mathcal{S}\triangle\mathcal{S}'| \le \widetilde{O}(1/\sqrt{\tau})$. Since each $\left|\langle \beta_i, x\rangle + b_{1,i}^{(0)}\right|$ is $\widetilde{O}(\frac{1}{\sqrt{m_1}})$ with high probability and each $|v_{1,i}^{(0)}|$ is $\widetilde{O}(\frac{1}{\sqrt{m_2}})$ with high probability, we can write

$$T_6 = \underbrace{\sum_{i\in\mathcal{S}'} v_{1,i}^{(0)} \mathbb{I}[\langle \beta_i, x\rangle + b_{1,i}^{(0)} \ge 0] \left(\langle \beta_i, x\rangle + b_{1,i}^{(0)}\right)}_{\stackrel{\text{def}}{=} B_2(x, v_1^{(0)}, W^{(0)}, b_1^{(0)})} + R_4 \tag{193}$$

with $R_4 \stackrel{\text{def}}{=} R_4\left(x, v_1^{(0)}, W^{(0)}, b_1^{(0)}\right)$ satisfying $|R_4| \le \widetilde{O}\left(\frac{1}{\sqrt{\tau m_2 m_1}}\right)$.

Finally, let $B = B_1 + B_2$. Again, conditioning on the randomness of u, we know that B_2 is independent of α. Since u and α are independent, we know that α and B_2 are also independent (by Proposition A.3). In other words, B and α (and therefore ρ) are independent.

Let $R = R_1 + R_2 + R_3 + R_4$ be the residual term, setting $\tau = \frac{1}{100}$, we have $|R| \le \widetilde{O}\left(\frac{1}{\sqrt{m_1 m_2}}\right)$ with high probability.

As for the norm bound on B, recall

$$n_1(x) = \sum_{i\in[m_1]} v_{1,i}^{(0)} \sigma\left(\langle w_i^{(0)}, x\rangle + b_{1,i}^{(0)}\right) \tag{194}$$

and by our random initialization, $n_i(x) \sim \mathcal{N}\big(0, \frac{1}{m_2}\big\|\sigma\big(W^{(0)}x + b_1^{(0)}\big)\big\|_2^2\big)$. At the same time, with high probability $\big\|\sigma\big(W^{(0)}x + b_1^{(0)}\big)\big\|_2^2 = O(1)$. Therefore, we know $|n_1(x)| \le \widetilde{O}(\frac{1}{\sqrt{m_2}})$, and this implies $|B| \le \widetilde{O}(\frac{1}{\sqrt{m_2}})$.

□

B.3 Proof of Lemma 5.6

Let us prove the lemma for a single term

$$a^* \cdot \Phi\left(\sum_{j\in[p_2]} v_{1,j}^* \phi_{1,j}(\langle w_{1,j}^*, x\rangle)\right)\left(\sum_{j\in[p_2]} v_{2,j}^* \phi_{2,j}(\langle w_{2,j}^*, x\rangle)\right), \tag{195}$$

and extending it to multiple terms and multiple outputs is straightforward (see the proof of Lemma G.1).[9] The proof consists of several steps.

B.3.1 Step 1: Existence in expectation

Let us define $p_2 S$ many chunks of the first layer, each chunk corresponds to a set $\mathcal{S}_{j,l}$ of cardinality $|\mathcal{S}_{j,l}| = \frac{m_1}{p_2 S}$ for $j \in [p_2], l \in [S]$, such that

$$\mathcal{S}_{j,l} = \left\{(j-1)\frac{m_1}{p_2} + (l-1)\frac{m_1}{p_2 S} + k \,\middle|\, k \in [\frac{m_1}{p_2 S}]\right\} \tag{197}$$

Let us then denote $v_i[j,l]$ to be $(v_{i,s})_{s\in\mathcal{S}_{j,l}}$ and $W[j,l]$ to be $(W_s)_{s\in\mathcal{S}_{j,l}}$. Recall that the input

[9] Recall we are interested in fitting a multi-output function $F^* = (f_1^*, \cdots, f_k^*)$ with:

$$f_i^*(x) = \sum_{r\in[p_1]} a_{i,r}^* \Phi_r\left(\sum_{j\in[p_2]} v_{1,r,j}^* \phi_{1,j}(\langle w_{1,j}^*, x\rangle)\right)\left(\sum_{j\in[p_2]} v_{2,r,j}^* \phi_{2,j}(\langle w_{2,j}^*, x\rangle)\right) \tag{196}$$

(without bias) to each neuron in the second hidden layer is

$$n_i(x) = \sum_{r \in [m_1]} v_{1,r}^{(0)} \sigma \left(\langle w_r^{(0)}, x \rangle + b_{1,r}^{(0)} \right). \tag{198}$$

Using Lemma 5.2 on these $p_2 S$ different chunks of the m_1 layer (each scaled down by $\frac{1}{\sqrt{p_2 S}}$ before applying Lemma 5.2), we can write $n_i(x)$ as:

$$n_i(x) = \sum_{j \in [p_2], l \in [S]} \rho_j \left(v_i^{(0)}[j,l], W^{(0)}[j,l], b_1^{(0)}[j,l] \right) \phi_{1,j,\varepsilon}(\langle w_{1,j}^*, x \rangle) \tag{199}$$

$$+ \sum_{j \in [p_2], l \in [S]} B_j \left(x, v_i^{(0)}[j,l], W^{(0)}[j,l], b_1^{(0)}[j,l] \right) + R_j \left(x, v_i^{(0)}[j,l], W^{(0)}[j,l], b_1^{(0)}[j,l] \right) \tag{200}$$

where each $\rho_{j,l} = \rho_j \left(v_i^{(0)}[j,l], W^{(0)}[j,l], b_1^{(0)}[j,l] \right) \sim \mathcal{N}(0, \frac{1}{100 C^2 m_2 (p_2 S)})$ are independent Gaussian random variable, for $C = \mathfrak{C}(\phi, \sqrt{\log(1/\varepsilon)})$. Let us denote $\rho_j = \sum_{l \in [S]} \rho_{j,l}$, we know that

$$\rho_j \sim \mathcal{N}\left(0, \frac{1}{100 C^2 p_2 m_2}\right) \stackrel{\text{def}}{=} \mathcal{N}\left(0, \frac{1}{C'^2 m_2}\right) \tag{201}$$

for $C' = 10 C \sqrt{p_2}$. Moreover, $|\phi_{1,j,\varepsilon}(\langle w_{1,j}^*, x \rangle) - \phi_{1,j}(\langle w_{1,j}^*, x \rangle)| \le 2\varepsilon$.

Let us then denote

$$B_j^{\mathsf{s}}(x) \stackrel{\text{def}}{=} \sum_{l \in [S]} B_j \left(x, v_i^{(0)}[j,l], W^{(0)}[j,l], b_1^{(0)}[j,l] \right) \tag{202}$$

$$R_j^{\mathsf{s}}(x) \stackrel{\text{def}}{=} \sum_{l \in [S]} R_j \left(x, v_i^{(0)}[j,l], W^{(0)}[j,l], b_1^{(0)}[j,l] \right) \tag{203}$$

We know that random variables ρ_j are independent of $B_j(x)$, and with high probability

$$|R_j^{\mathsf{s}}(x)| \le S \times \widetilde{O}\left(\frac{1}{\sqrt{m_1 m_2 (p_2 S)}} \right) \le \widetilde{O}\left(\frac{\sqrt{S}}{\sqrt{m_1 m_2 p_2}} \right) \text{ and} \tag{204}$$

$$|B_j \left(x, v_i^{(0)}[j,l], W^{(0)}[j,l], b_1^{(0)}[j,l] \right)| \le \frac{1}{\sqrt{p_2 S}} \times \widetilde{O}\left(\frac{1}{\sqrt{m_2}} \right) = \widetilde{O}\left(\frac{1}{\sqrt{m_2 p_2 S}} \right) \tag{205}$$

Let us apply the Wasserstein distance version of the central limit theorem (see for instance [EMZ18, Theorem 1]) [10]: since $B_j^{\mathsf{s}}(x)$ is the summation of S i.i.d random variables, there is a Gaussian random variable $\beta_j(x)$ only depending on $B_j^{\mathsf{s}}(x)$ such that

$$\mathcal{W}_2(B_j^{\mathsf{s}}(x), \beta_j(x)) \le \widetilde{O}\left(\frac{1}{\sqrt{m_2 p_2 S}} \right) \tag{206}$$

Define $\beta'(x) = \sum_{j \in [p_2]} \beta_j(x)$, we know that $\beta'(x)$ is a Gaussian random variable independent of all the ρ_j with

$$\mathcal{W}_2 \left(n_i(x), \sum_j \rho_j \phi_{1,j,\varepsilon}(\langle w_{1,j}^*, x \rangle) + \beta'(x) \right) \le \widetilde{O}\left(\frac{\sqrt{S p_2}}{\sqrt{m_1 m_2}} + \frac{\sqrt{p_2}}{\sqrt{m_2 S}} \right) \tag{207}$$

[10] All the variables considered in this section is not absolutely bounded, but only with high probability with a gaussian tail. Strictly speaking, when apply this Theorem we should be first replaced B_j^{s} by $B_j^{\mathsf{s}} \mathbb{I}_{\text{all } B_j \le \widetilde{O}(1/\sqrt{m_2 p_2 S})}$. We choose to avoid writing this truncation in the paper to simply the presentation.

Now, let us slightly override notation and denote

$$\phi_{1,j}(x) = \frac{1}{C'}\phi_{1,j,\varepsilon}(\langle w^*_{1,j}, x\rangle), \tag{208}$$

$$\phi_{2,j}(x) = \phi_{2,j}(\langle w^*_{2,j}, x\rangle). \tag{209}$$

We then have that for $\alpha_{i,j} \stackrel{\text{def}}{=} C'\rho_j \sim \mathcal{N}(0, 1/m_2)$ being i.i.d.:

$$\mathcal{W}_2\left(n_i(x), \sum_{j\in[p_2]} \alpha_{i,j}\phi_{1,j}(x) + \beta'(x)\right) \leq \widetilde{O}\left(\frac{\sqrt{Sp_2}}{\sqrt{m_1 m_2}} + \frac{\sqrt{p_2}}{\sqrt{m_2 S}}\right) \tag{210}$$

Since by our random initialization, $n_i(x) \sim \mathcal{N}\left(0, \frac{1}{m_2}\left\|\sigma\left(W^{(0)}x + b_1^{(0)}\right)\right\|_2^2\right)$, and since for every every unit vector x, w.h.p.

$$\left\|\sigma\left(W^{(0)}x + b_1^{(0)}\right)\right\|_2^2 = 1 \pm \widetilde{O}\left(\frac{1}{\sqrt{m_1}}\right) \tag{211}$$

and therefore $\mathcal{W}_2(n_i(x), g) \leq \widetilde{O}(\frac{1}{\sqrt{m_1 m_2}})$ for $g \sim (0, \frac{1}{m_2})$. Since we can write $g = \sum_{j\in[p_2]} \alpha_{i,j}\phi_{1,j}(x) + \beta(x)$ for

$$\beta(x) \sim \mathcal{N}\left(0, \frac{1}{m_2}\left(1 - \sum_{j\in[p_2]} \phi^2_{1,j}(x)\right)\right) \quad , \tag{212}$$

The whole chain of Wasserstein distance arguments give us (where we choose $S = \sqrt{m_1}$)

$$\mathcal{W}_2\left(n_i(x), \sum_{j\in[p_2]} \alpha_{i,j}\phi_{1,j}(x) + \beta(x)\right) \leq \widetilde{O}\left(\frac{1}{\sqrt{m_1 m_2}} + \frac{\sqrt{Sp_2}}{\sqrt{m_1 m_2}} + \frac{\sqrt{p_2}}{\sqrt{m_2 S}}\right) \leq O\left(\frac{\sqrt{p_2}}{m_1^{1/4}\sqrt{m_2}}\right) \tag{213}$$

Now, if we use Lemma 5.1 again on $\Phi'(z) = \Phi(C'x)$ we know that there is a function $h : \mathbb{R}^2 \to [-C'', C'']$ for $C'' = \mathfrak{C}\left(\Phi', \sqrt{\log(1/\varepsilon)}\right) = O\left(\mathfrak{C}\left(\Phi, C'\sqrt{\log(1/\varepsilon)}\right)\right)$ such that

$$\mathbb{E}\left[\mathbb{I}_{\sum_{j\in[p_2]} \alpha_{i,j}\phi_{1,j}(x) + \beta(x) + b^{(0)}_{2,i} \geq 0}\, h\left(\sum_{j\in[p_2]} v^*_{1,j}\alpha_{i,j}, b^{(0)}_{2,i}\right)\left(\sum_{j\in[p_2]} v^*_{2,j}\phi_{2,j}(x)\right)\right] \tag{214}$$

$$= \Phi\left(C'\sum_{j\in[p_2]} v^*_{1,j}\phi_{1,j}(x)\right)\left(\sum_{j\in[p_2]} v^*_{2,j}\phi_{2,j}(x)\right) \pm \varepsilon C''' \tag{215}$$

Where $C''' = p_2 \sup_{x:\|x\|_2\leq 1}\left|\sum_{j\in[p_2]} v^*_{2,j}\phi_{2,j}(x)\right| \leq p_2^2\mathfrak{C}(\phi, 1)$. Now, using that Wasserstein distance bound from (213), notice that $n_i \sim \mathcal{N}\left(0, \frac{1}{m_2}\left\|\sigma\left(W^{(0)}x + b_1^{(0)}\right)\right\|_2^2\right)$ with $\left\|\sigma\left(W^{(0)}x + b_1^{(0)}\right)\right\|_2 = \Theta(1)$ w.h.p., and h is bounded, we can thus replace $\sum_{j\in[p_2]} \alpha_{i,j}\phi_{1,j}(x) + \beta(x)$ with $n_i(x)$ and derive that, as long as $S = m_1^{1/4}$ we have:

$$\mathbb{E}\left[\mathbb{I}_{n_i(x)+b_{2,i}^{(0)}\geq 0} h\left(\sum_{j\in[p_2]} v_{1,j}^*\alpha_{i,j}, b_{2,i}^{(0)}\right)\left(\sum_{j\in[p_2]} v_{2,j}^*\phi_{2,j}(x)\right)\right] \tag{216}$$

$$= \mathbb{E}\left[\mathbb{I}_{\alpha_{i,j}\phi_{1,j}(x)+\beta(x)+b_{2,i}^{(0)}\geq 0} h\left(\sum_{j\in[p_2]} v_{1,j}^*\alpha_{i,j}, b_{2,i}^{(0)}\right)\left(\sum_{j\in[p_2]} v_{2,j}^*\phi_{2,j}(x)\right)\right] \tag{217}$$

$$\pm O\left(\mathcal{W}_2\left(n_i(x), \sum_{j\in[p_2]} \alpha_{i,j}\phi_{1,j}(x) + \beta(x)\right)\sqrt{m_2}C'''C''\right) \tag{218}$$

$$= \Phi\left(C'\sum_{j\in[p_2]} v_{1,j}^*\phi_{1,j}(x)\right)\left(\sum_{j\in[p_2]} v_{2,j}^*\phi_{2,j}(x)\right) \pm O\left(\varepsilon C''' + \frac{C''C'''\sqrt{p_2}}{m_1^{1/4}}\right). \tag{219}$$

Since $C'\phi_{1,j}(x) = \phi_{1,j,\varepsilon}(\langle w_{1,j}^*, x\rangle)$ and $\phi_{2,j}(x) = \phi_{2,j}(\langle w_{2,j}^*, x\rangle)$, we know that

$$\mathbb{E}\left[\mathbb{I}_{n_i(x)+b_{2,i}^{(0)}\geq 0} h\left(\sum_{j\in[p_2]} v_{1,j}^*\alpha_{i,j}, b_{2,i}^{(0)}\right)\left(\sum_{j\in[p_2]} v_{2,j}^*\phi_{2,j}(\langle w_{2,j}^*, x\rangle)\right)\right] \tag{220}$$

$$\overset{①}{=} \Phi\left(\sum_{j\in[p_2]} v_{1,j}^*\phi_{1,j}(\langle w_{1,j}^*, x\rangle)\right)\left(\sum_{j\in[p_2]} v_{2,j}^*\phi_{2,j}(\langle w_{2,j}^*, x\rangle)\right) \pm O\left(\varepsilon(C''' + L_\Phi) + \frac{C''C'''\sqrt{p_2}}{m_1^{1/4}}\right) \tag{221}$$

$$= \Phi\left(\sum_{j\in[p_2]} v_{1,j}^*\phi_{1,j}(\langle w_{1,j}^*, x\rangle)\right)\left(\sum_{j\in[p_2]} v_{2,j}^*\phi_{2,j}(\langle w_{2,j}^*, x\rangle)\right) \pm O\left(p_2^2(\mathfrak{C}(\Phi,1) + \mathfrak{C}(\phi,1))\varepsilon\right) \tag{222}$$

In ①, L_Φ is the Lipschitz continuity parameter of Φ and it is bounded by $\mathfrak{C}(\Phi, 1)$.

B.3.2 Step 2: From expectation to finite neurons

Intuitively, we wish to apply concentration bound on (220) with respect to all neurons $i \in [m_2]$ on the second layer. Unfortunately, across different choices of i, the values of $n_i(x)$ and $\alpha_{i,j}$ can be correlated. In the remainder of this proof, let us try to correct the two terms to make them independent across i.

First treat $\boldsymbol{\alpha_{i,j}}$. Recall $\alpha_{i,j} = C'\rho_j = C'\sum_{l\in[S]}\rho_{j,l}$ where each $\rho_{j,l}$ is a function on $v_i^{(0)}[j,l], W^{(0)}[j,l]$, and $b_1^{(0)}[j,l]$. Now, using the $\widetilde{\rho}$ notion from Lemma 5.2, let us also define $\widetilde{\rho}_{j,l} = \widetilde{\rho}_j\left(v_i^{(0)}[j,l]\right)$ which is in the same distribution as $\rho_{j,l}$ except that it does not depend on $W^{(0)}$ or $b_1^{(0)}$. We can similarly let $\widetilde{\rho}_j = \sum_{l\in[S]}\widetilde{\rho}_{j,l}$. From Lemma 5.2, we know that with high probability over $W^{(0)}$:

$$\widetilde{\rho}_j \sim \mathcal{N}\left(0, \frac{1}{C'^2 m_2}\right) \quad \text{and} \quad \mathcal{W}_2(\rho_j|_{W^{(0)},b^{(0)}}, \widetilde{\rho}_j) \leq \widetilde{O}\left(\frac{S}{C\sqrt{m_1 m_2}}\right) \tag{223}$$

According, we define $\widetilde{\alpha}_{i,j} = C'\widetilde{\rho}_j = C'\sum_{l\in[S]}\widetilde{\rho}_{j,l}$.

Next treat $n_i(x)$. Recall $n_1(x) = \sum_{i\in[m_1]} v_{1,i}^{(0)} \sigma\left(\langle w_i^{(0)}, x\rangle + b_{1,i}^{(0)}\right)$ and accordingly we define

$$\widetilde{n}_j(x) \stackrel{\text{def}}{=} \frac{\sum_{i\in[m_1]} v_{j,i}^{(0)} \sigma\left(\langle w_i^{(0)}, x\rangle + b_{1,i}^{(0)}\right)}{\|u\|_2} \mathbb{E}[\|u\|_2] \tag{224}$$

where vector $u \stackrel{\text{def}}{=} \left(\sigma\left(\langle w_i^{(0)}, x\rangle + b_{1,i}^{(0)}\right)\right)_{i\in[m_1]}$. By definition, we know

$$\widetilde{n}_j(x) \sim \mathcal{N}\left(0, \frac{1}{m_2}\mathbb{E}[\|u\|_2]^2\right) \tag{225}$$

is a Gaussian variable and is independent of u. As a consequence, the quantities $\widetilde{n}_j(x)$ are independent among different choices of j.

Concentration. Recall from (220) of Step 1 we have

$$\mathbb{E}\left[\mathbb{I}_{n_i(x)+b_{2,i}^{(0)}\geq 0} h\left(\sum_{j\in[p_2]} v_{1,j}^* \alpha_{i,j}, b_{2,i}^{(0)}\right)\left(\sum_{j\in[p_2]} v_{2,j}^* \phi_{2,j}(\langle w_{2,j}^*, x\rangle)\right)\right] \tag{226}$$

$$= \Phi\left(\sum_{j\in[p_2]} v_{1,j}^* \phi_{1,j}(\langle w_{1,j}^*, x\rangle)\right)\left(\sum_{j\in[p_2]} v_{2,j}^* \phi_{2,j}(\langle w_{2,j}^*, x\rangle)\right) \pm O\left(p_2^2(\mathfrak{C}(\Phi,1)+\mathfrak{C}(\phi,1))\varepsilon\right) \tag{227}$$

Using the notions of $\widetilde{n}$ and $\widetilde{\alpha}$ and the Wasserstein distance bound, it implies

$$\mathbb{E}\left[\mathbb{I}_{\widetilde{n}_i(x)+b_{2,i}^{(0)}\geq 0} h\left(\sum_{j\in[p_2]} v_{1,j}^* \widetilde{\alpha}_{i,j}, b_{2,i}^{(0)}\right)\left(\sum_{j\in[p_2]} v_{2,j}^* \phi_{2,j}(\langle w_{2,j}^*, x\rangle)\right)\right] \tag{228}$$

$$= \Phi\left(\sum_{j\in[p_2]} v_{1,j}^* \phi_{1,j}(\langle w_{1,j}^*, x\rangle)\right)\left(\sum_{j\in[p_2]} v_{2,j}^* \phi_{2,j}(\langle w_{2,j}^*, x\rangle)\right) \pm O\left(p_2^2(\mathfrak{C}(\Phi,1)+\mathfrak{C}(\phi,1))\varepsilon\right) \tag{229}$$

By standard concentration —and the independence of tuples $(n_i(x), (\alpha_{i,j})_{j\in[p_2]})$ with respect to different choices of i— we know with high probability

$$\frac{1}{m_2}\sum_{i\in[m_2]}\left[\mathbb{I}_{\widetilde{n}_i(x)+b_{2,i}^{(0)}\geq 0} h\left(\sum_{j\in[p_2]} v_{1,j}^* \widetilde{\alpha}_{i,j}, b_{2,i}^{(0)}\right)\left(\sum_{j\in[p_2]} v_{2,j}^* \phi_{2,j}(\langle w_{2,j}^*, x\rangle)\right)\right] \tag{230}$$

$$= \mathbb{E}\left[\mathbb{I}_{\widetilde{n}_i(x)+b_{2,i}^{(0)}\geq 0} h\left(\sum_{j\in[p_2]} v_{1,j}^* \widetilde{\alpha}_{i,j}, b_{2,i}^{(0)}\right)\left(\sum_{j\in[p_2]} v_{2,j}^* \phi_{2,j}(\langle w_{2,j}^*, x\rangle)\right)\right] \pm \widetilde{O}\left(\frac{C''C'''}{\sqrt{m_2}}\right) \tag{231}$$

Using again the Wasserstein distance between $\widetilde{n}$ and n and between $\widetilde{\alpha}$ and α, we can combine the above two equations to derive that w.h.p.

$$\frac{1}{m_2}\sum_{i\in[m_2]}\left[\mathbb{I}_{n_i(x)+b_{2,i}^{(0)}\geq 0} h\left(\sum_{j\in[p_2]} v_{1,j}^* \alpha_{i,j}, b_{2,i}^{(0)}\right)\left(\sum_{j\in[p_2]} v_{2,j}^* \phi_{2,j}(\langle w_{2,j}^*, x\rangle)\right)\right] \tag{232}$$

$$= \Phi\left(\sum_{j\in[p_2]} v_{1,j}^* \phi_{1,j}(\langle w_{1,j}^*, x\rangle)\right)\left(\sum_{j\in[p_2]} v_{2,j}^* \phi_{2,j}(\langle w_{2,j}^*, x\rangle)\right) \pm O\left(p_2^2(\mathfrak{C}(\Phi,1)+\mathfrak{C}(\phi,1))\varepsilon\right) \tag{233}$$

B.3.3 Step 3: From finite neurons to the network

Now we discuss how to construct the network.

Let $v \in \mathbb{R}^{m_1}$ be an arbitrary vector with $v_i \in \{+1, 1\}$. Define $V^* \in \mathbb{R}^{m_2 \times m_1}$ as

$$V^* = \frac{a^* \varepsilon^{16}}{C_0 m_2} \left(a_i h \left(\sum_{j \in [p_2]} v^*_{1,j} \alpha_{i,j}, b^{(0)}_{2,i} \right) v^\top \right)_{i \in [m_2]} \tag{234}$$

and define $W^* \in \mathbb{R}^{m_1 \times d}$ as

$$W^* = \frac{C_0}{\varepsilon_a^2 \varepsilon^{16} m_1} \left(v_i \sum_{j \in [p_2]} v^*_{2,j} h_{\phi,j} \left(\langle w^*_{2,j}, w^{(0)}_i \rangle, b^{(0)}_{1,i} \right) e_d \right)_{i \in [m_1]} \tag{235}$$

where $h_{\phi,j}$ approximates $\phi_{2,j}$ as defined in Lemma 5.1.

With these choices of weights, when the signs of ReLU's are determined by the random initialization (i.e., $W^{(0)}$ and $V^{(0)}$), we can write the network output as

$$\begin{aligned} g^{(0)}(x, W^*, V^*) &\stackrel{\text{def}}{=} \sum_{i \in [m_2]} a_i \mathbb{I}_{n_i(x) + b^{(0)}_{2,i} \geq 0} \sum_{i' \in [m_1]} v^*_{i,i'} \langle w^*_i, x \rangle \mathbb{I}_{\langle w^{(0)}_{i'}, x \rangle + b^{(0)}_{1,i'} \geq 0} && (236) \\ &= \frac{a^*}{m_2} \sum_{i \in [m_2]} \frac{a_i^2}{\varepsilon_a^2} h \left(\sum_{j \in [p_2]} v^*_{1,j} \alpha_{i,j}, b^{(0)}_{2,i} \right) \mathbb{I}_{n_i(x) + b^{(0)}_{2,i} \geq 0} && (237) \\ &\quad \times \sum_{j \in [p_2]} v^*_{2,j} \left(\frac{1}{m_1} \sum_{i' \in [m_1]} h_{\phi,j} \left(\langle w^*_{2,j}, w^{(0)}_{i'} \rangle, b^{(0)}_{1,i'} \right) \mathbb{I}_{\langle w^{(0)}_{i'}, x \rangle + b_{1,i'} \geq 0} \right) && (238) \end{aligned}$$

We have, using Lemma 5.1, that

$$\mathbb{E} \left[h_{\phi,j} \left(\langle w^*_{2,j}, w^{(0)}_{i'} \rangle, b^{(0)}_{1,i'} \right) \mathbb{I}_{\langle w^{(0)}_{i'}, x \rangle + b_{1,i'} \geq 0} \right] = \phi_{2,j}(\langle w^*_{2,j}, x \rangle) \pm 2\varepsilon \tag{239}$$

Since $h_{\phi,j} \in [-C'', C'']$, applying concentration in Lemma A.1 we have w.h.p

$$\frac{1}{m_1} \sum_{i' \in [m_1]} \left(\langle w^*_{2,j}, w^{(0)}_{i'} \rangle, b^{(0)}_{1,i'} \right) \mathbb{I}_{\langle w^{(0)}_{i'}, x \rangle + b_{1,i'} \geq 0} = \phi_{2,j}(\langle w^*_{2,j}, x \rangle) \pm O(\varepsilon) \tag{240}$$

Using the fact that the quantities $\frac{a_i^2}{\varepsilon_a^2}$ are independent and each equals 1 in expectation, and using (232), we have

$$\begin{aligned} &\frac{1}{m_2} \sum_{i \in [m_2]} \left[\frac{a_i^2}{\varepsilon_a^2} \mathbb{I}_{n_i(x) + b^{(0)}_{2,i} \geq 0} h \left(\sum_{j \in [p_2]} v^*_{1,j} \alpha_{i,j}, b^{(0)}_{2,i} \right) \left(\sum_{j \in [p_2]} v^*_{2,j} \phi_{2,j}(\langle w^*_{2,j}, x \rangle) \right) \right] && (241) \\ &= \Phi \left(\sum_{j \in [p_2]} v^*_{1,j} \phi_{1,j}(\langle w^*_{1,j}, x \rangle) \right) \left(\sum_{j \in [p_2]} v^*_{2,j} \phi_{2,j}(\langle w^*_{2,j}, x \rangle) \right) \pm O\left(p_2^2 (\mathfrak{C}(\Phi, 1) + \mathfrak{C}(\phi, 1)) \varepsilon \right) && (242) \end{aligned}$$

Putting this into (238), and using $a^* \in [-1, 1]$ and $h \in [-C'', C'']$, we know that with high probability

$$\begin{aligned} g^{(0)}(x, W^*, V^*) &= a^* \cdot \Phi \left(\sum_{j \in [p_2]} v^*_{1,j} \phi_{1,j}(\langle w^*_{1,j}, x \rangle) \right) \left(\sum_{j \in [p_2]} v^*_{2,j} \phi_{2,j}(\langle w^*_{2,j}, x \rangle) \right) && (243) \\ &\quad \pm O\left(p_2^2 (\mathfrak{C}(\Phi, 1) + \mathfrak{C}(\phi, 1)) \varepsilon \right) && (244) \end{aligned}$$

Finally, recall that (for each output coordinate) $g^{(b,b)}$ is different from $g^{(0)}$ only by the diagonal signs, namely,

$$g^{(b,b)}(x, W^*, V^*) = a(D_{v,x} + D'_{v,x})V^*(D_{w,x} + D'_{w,x})W^*x \tag{245}$$

where $D_{v,x}$ and $D_{w,x}$ are determined at the random initialization (in the same way as $g^{(0)}$). By Lemma 5.5 and our parameter choices, we know

$$\|D'_{w,x}\|_0 \le \widetilde{O}(\tau_w^{4/5} m_1^{6/5}) \ll \widetilde{O}(m_1^{4/5}) \tag{246}$$

$$\|D'_{v,x}\|_0 \le \widetilde{O}\left(\sigma_v m_2^{3/2} + \tau_v^{2/3} m_2 + \tau_w^{2/3} m_1^{1/6} m_2\right) \le \widetilde{O}((\varepsilon/C_0)^{\Theta(1)} m_2) \tag{247}$$

Using $h, h_{\phi,j} \in [-C'', C'']$ again and the above sparsity bounds (for sign changes), we have w.h.p.

$$|g^{(b,b)}(x, W^*, V^*) - g^{(0)}(x, W^*, V^*)| \le O(\varepsilon) \tag{248}$$

This finishes the proof that, with high probability

$$g^{(b,b)}(x, W^*, V^*) = a^* \cdot \Phi\left(\sum_{j\in[p_2]} v^*_{1,j}\phi_{1,j}(\langle w^*_{1,j}, x\rangle)\right)\left(\sum_{j\in[p_2]} v^*_{2,j}\phi_{2,j}(\langle w^*_{2,j}, x\rangle)\right) \tag{249}$$

$$\pm\, O\left(p_2^2(\mathfrak{C}(\Phi,1) + \mathfrak{C}(\phi,1))\varepsilon\right) \tag{250}$$

Finally, reducing the value of ε by factor $p_2^2(\mathfrak{C}(\Phi,1)+\mathfrak{C}(\phi,1))$ we complete the proof of Lemma 5.6 for a single output and for a single Φ.

The proof generalizes to multiple functions Φ_r (in the same way as Lemma G.1) if we choose the vector $v \in \{-1,1\}^{m_1}$ randomly across different values of r and apply concentration.

The proof generalizes to multiple outputs in the same way as Lemma G.1. ■

C Proof of Main Lemmas for Three Layers: Coupling

C.1 Proof of Lemma 5.5

C.1.1 Proof of Lemma 5.5: Part I, Sparsity

For notation simplicity, we only do the proof when there is no bias term. The proof with bias term is analogous (but more notationally involved). Let us denote

$$z_0 \stackrel{\text{def}}{=} D_{w,x}W^{(0)}x \tag{251}$$

$$z_1 \stackrel{\text{def}}{=} D'_{w,x}W^{(0)}x \tag{252}$$

$$z_2 \stackrel{\text{def}}{=} (D_{w,x} + D'_{w,x})(W^{(0)} + W^\rho + W')x - D_{w,x}W^{(0)}x\ . \tag{253}$$

Let us break $D'_{w,x}$ into two parts:

- diagonal matrix $D''_{w,x}$ denotes the sign change of ReLU's from $W^{(0)}$ to $W^{(0)} + W^\rho$, and
- diagonal matrix $D'_{w,x} - D''_{w,x}$ denotes the sign change from $W^{(0)} + W^\rho$ to $W^{(0)} + W^\rho + \eta\mathbf{\Sigma}W'$.

Sign change in $\boldsymbol{D''_{w,x}}$. By definition, each coordinate of $W^{(0)}x \sim \mathcal{N}\left(0, \frac{1}{m_1}\right)$ and each coordinate of $W^\rho x \sim \mathcal{N}(0, \sigma_w^2)$. Thus, a standard property of Gaussian, for each i, we have $\mathbf{Pr}[|W_i^\rho x| \ge |W_i^{(0)}x|] \le \widetilde{O}\left(\sigma_w\sqrt{m_1}\right)$. By concentration bound, with high probability, the number of sign changes of the ReLU activations in the first hidden layer caused by adding W^ρ is no more than $\sigma_w m_1^{3/2}$, or in symbols:

$$\|D''_{w,x}\|_0 \le \widetilde{O}(\sigma_w m_1^{3/2})\ . \tag{254}$$

Moreover, for each coordinate i with $[D''_{w,x}]_{i,i} \neq 0$, we must have $|(D''_{w,x}W^{(0)}x)_i| \leq |(W^\rho x)_i| \leq \widetilde{O}(\sigma_w)$ with high probability, and thus

$$\|D''_{w,x}W^{(0)}x\|_2 \leq \widetilde{O}\left(\sigma_w\sqrt{\sigma_w m_1^{3/2}}\right) = \widetilde{O}\left(\sigma_w^{3/2}m_1^{3/4}\right) \tag{255}$$

By our choice of parameter $\sigma_w \leq \tau_w/m_1^{1/4}$, we have

$$\|D''_{w,x}\|_0 \leq \tau_w m_1^{5/4} \quad \text{and} \quad \|D''_{w,x}W^{(0)}x\|_2 \leq \tau_w^{3/2}m_1^{3/8} \tag{256}$$

Sign change in $D'_{w,x} - D''_{w,x}$. Let $s = \|D'_{w,x} - D''_{w,x}\|_0$ be the total number of sign changes of the ReLU activations in the first hidden layer caused by further adding $W'x$. Observe that, the total number of coordinates i where $|((W^{(0)}+W^\rho)x_i| \leq s'' \stackrel{\text{def}}{=} \frac{2\tau_w}{s^{1/4}}$ is at most $\widetilde{O}\left(s''m_1^{3/2}\right)$ with high probability. Now, if $s \geq \widetilde{\Omega}\left(s''m_1^{3/2}\right)$, then W' must have caused the sign change of $\frac{s}{2}$ coordinates each by absolute value at least s''. Since $\|W'\|_{2,4} \leq \tau_w$, this is impossible because $\frac{s}{2} \times (s'')^2 > \tau_w^4$. Therefore, we must have

$$s \leq \widetilde{O}(s''m^{3/2}) = \widetilde{O}\left(\frac{\tau_w}{s^{1/4}}m_1^{3/2}\right) \tag{257}$$

After rearranging,

$$\|D'_{w,x} - D''_{w,x}\|_0 \leq s = \widetilde{O}\left(\tau_w^{\frac{4}{5}}m_1^{\frac{6}{5}}\right) \tag{258}$$

Next, for each coordinate i where $(D'_{w,x}-D''_{w,x})_{i,i} \neq 0$, we must have $|((W^{(0)}+W^\rho)x)_i| \leq |(W'x)_i|$, and since $(W'x)_i^4$ must sum up to at most τ_w for those s coordinates, we have

$$\|\left(D'_{w,x} - D''_{w,x}\right)(W^{(0)} + W^\rho)x\|_2 \leq \sqrt{\sum_{i,(D'_{w,x}-D''_{w,x})_{i,i}\neq 0}(W'x)_i^2} \tag{259}$$

$$\leq \sqrt{\sqrt{s \cdot \sum_{i,(D'_{w,x}-D''_{w,x})_{i,i}\neq 0}(W'x)_i^4}} \leq O\left(s^{1/4}\tau_w\right) = \widetilde{O}\left(\tau_w^{\frac{6}{5}}m_1^{\frac{3}{10}}\right) \tag{260}$$

Sum up: First Layer Sign Change. Using (256) and (258) from the two cases above and $\tau_w \leq m_1^{-1/4}$, we have

$$\|D'_{w,x}\|_0 \leq \widetilde{O}\left(\tau_w^{\frac{4}{5}}m_1^{\frac{6}{5}} + \tau_w m_1^{5/4}\right) \leq \widetilde{O}\left(\tau_w^{\frac{4}{5}}m_1^{\frac{6}{5}}\right). \tag{261}$$

By triangle inequality, we have

$$\|z_1\|_2 = \|D'_{w,x}W^{(0)}x\|_2 \tag{262}$$

$$\leq \|\left(D'_{w,x} - D''_{w,x}\right)W^\rho x\|_2 + \|\left(D'_{w,x} - D''_{w,x}\right)(W^{(0)} + W^\rho)x\|_2 + \|D''_{w,x}W^{(0)}x\|_2 \tag{263}$$

$$\overset{①}{\leq} \widetilde{O}\left(\tau_w^{3/2}m_1^{3/8} + \tau_w^{\frac{6}{5}}m_1^{\frac{3}{10}} + \sigma_w\tau_w^{\frac{2}{5}}m_1^{\frac{3}{5}}\right) \leq \widetilde{O}\left(\tau_w^{\frac{6}{5}}m_1^{\frac{3}{10}}\right) \tag{264}$$

Above, ① uses (256), (260) and the fact that with high probability $\|\left(D'_{w,x} - D''_{w,x}\right)W^\rho x\|_2 \leq \|D'_{w,x} - D''_{w,x}\|_0 \cdot \|W^\rho x\|_\infty \leq \widetilde{O}(\sigma_w s^{1/2})$.

Second Layer Sign Change. By 1-Lipshitzness of ReLU, we know that w.h.p.

$$\|z_2\|_2 = \left\|(D_{w,x} + D'_{w,x})(W^{(0)} + W^\rho + W')x - D_{w,x}W^{(0)}x\right\|_2 \tag{265}$$

$$\leq \|(W^\rho + W')x\|_2 \leq \widetilde{O}\left(\tau_w m_1^{1/4} + \sigma_w m_1^{1/2}\right) \leq \widetilde{O}\left(\tau_w m_1^{1/4}\right) \tag{266}$$

where we have used our choice $\sigma_w \leq \tau_w m_1^{-1/4}$.

Now, notice that the sign change in the second layer is caused by input vector

$$\text{changing from} \quad V^{(0)}z_0 \quad \text{to} \quad V^{(0)}z_0 + V^{(0)}z_2 + V^\rho(z_0 + z_2) + V'(z_0 + z_2). \tag{267}$$

Here, we have

$$\|V^\rho(z_0 + z_2)\|_\infty \leq \widetilde{O}(\sigma_v) \tag{268}$$

$$\|V^{(0)}z_2 + V'(z_0 + z_2)\|_2 \leq \widetilde{O}\left(\tau_v + \|z_2\|_2\right) \leq \widetilde{O}\left(\tau_v + \tau_w m_1^{1/4}\right) \tag{269}$$

In comparison (at random initialization) we have $V^{(0)}z_0 \sim \mathcal{N}\left(0, \frac{\|z_0\|_2^2}{m_2} I\right)$ with $\|z_0\|_2 = \widetilde{\Omega}(1)$. Using a careful two-step argument (see Claim C.1), we can bound

$$\|D'_{v,x}\|_0 \leq \widetilde{O}\left(\left(\tau_v + \tau_w m_1^{1/4}\right)^{2/3} m_2 + \sigma_v m_2^{3/2}\right) \tag{270}$$

■

C.1.2 Proof of Lemma 5.5: Part II, Diagonal Cross Term

One can carefully check that

$$g_r(x, W^{(0)} + W^\rho + W' + \eta\mathbf{\Sigma}W'', V^{(0)} + V^\rho + V' + \eta V''\mathbf{\Sigma}) \tag{271}$$

$$= g_r\left(x, W^{(0)} + W^\rho + W', V^{(0)} + V^\rho + V'\right) + g_r^{(b,b)}(\eta\mathbf{\Sigma}W'', \eta V''\mathbf{\Sigma}) \tag{272}$$

$$+ g_r^{(b)}(x, W^{(0)} + W^\rho + W', \eta V''\mathbf{\Sigma}) + g_r^{(b,b)}(x, \eta\mathbf{\Sigma}W'', V^{(0)} + V^\rho + V') \tag{273}$$

We consider the last two error terms.

First error term. The first of these two terms is

$$g_r^{(b)}(x, W^{(0)} + W^\rho + W', \eta V''\mathbf{\Sigma}) \tag{274}$$

$$= \eta a_r(D_{v,x} + D'_{v,x})V''\mathbf{\Sigma}(D_{w,x} + D'_{w,x})((W^{(0)} + W^\rho + W')x + b_1) \tag{275}$$

Clearly, it has zero expectation with respect to $\mathbf{\Sigma}$. With high probability, we have

$$\|(D_{w,x} + D'_{w,x})((W^{(0)} + W^\rho + W')x + b_1)\|_\infty \tag{276}$$

$$\leq \|(W^{(0)} + W^\rho + W')x + b_1\|_\infty \leq \widetilde{O}\left(\tau_w + \sigma_w + \frac{1}{m_1^{1/2}}\right) \leq \widetilde{O}\left(\frac{1}{m_1^{1/2}}\right) \tag{277}$$

and we have $\|a_r(D_{v,x} + D'_{v,x})V''\|_2 \leq \widetilde{O}\left(\tau_v m_2^{1/2}\right)$. By Fact C.2, using the randomness of $\mathbf{\Sigma}$, with high probability

$$|g_r^{(b)}(x, W^{(0)} + W^\rho + W', \eta V''\mathbf{\Sigma})| = \eta\widetilde{O}\left(\frac{\tau_v m_2^{1/2}}{\sqrt{m_1}}\right) \tag{278}$$

First second term. We carefully write down the second error term

$$g_r^{(b,b)}(x, \eta\mathbf{\Sigma}W'', V^{(0)} + V^\rho + V') = \eta a_r(D_{v,x} + D'_{v,x})(V^{(0)} + V^\rho + V')(D_{w,x} + D'_{w,x})\mathbf{\Sigma}W''x \quad (279)$$

$$= \eta a_r(D_{v,x} + D'_{v,x})V'(D_{w,x} + D'_{w,x})\mathbf{\Sigma}W''x \quad (280)$$

$$+ \eta a_r D_{v,x}(V^{(0)} + V^\rho)(D_{w,x} + D'_{w,x})\mathbf{\Sigma}W''x \quad (281)$$

$$+ \eta a_r D'_{v,x}(V^{(0)} + V^\rho)(D_{w,x} + D'_{w,x})\mathbf{\Sigma}W''x \quad (282)$$

Obviously all the three terms on the right hand side have zero expectation with respect to $\mathbf{\Sigma}$.

- For the first term, since w.h.p. $\|a_r(D_{v,x}+D'_{v,x})V'(D_{w,x}+D'_{w,x})\|_2 = \widetilde{O}(\tau_v m_2^{1/2})$ and $\|W''x\|_\infty \leq \tau_w$, by Fact C.2, using the randomness of $\mathbf{\Sigma}$ we know that w.h.p.

$$|a_r(D_{v,x} + D'_{v,x})V'(D_{w,x} + D'_{w,x})\mathbf{\Sigma}W''x| \leq \widetilde{O}(\tau_v m_2^{1/2}\tau_w) \quad (283)$$

- For the second term, since $\|W''x\|_2 \leq \tau_w m_1^{1/4}$ and w.h.p. $\|a_r D_{v,x}(V^{(0)}+V^\rho)(D_{w,x}+D'_{w,x})\|_\infty \leq \|a_r D_{v,x}(V^{(0)} + V^\rho)\|_\infty \leq \widetilde{O}(1)$, by Fact C.2, using the randomness of $\mathbf{\Sigma}$ we know that w.h.p.

$$|a_r D_{v,x}(V^{(0)} + V^\rho)(D_{w,x} + D'_{w,x})\mathbf{\Sigma}W''x| = \widetilde{O}(\tau_w m_1^{1/4}) \quad (284)$$

- For the third term, again by $\|W''x\|_\infty \leq \tau_w$ and Fact C.2, we have: w.h.p.

$$|a_r D'_{v,x}(V^{(0)} + V^\rho)\mathbf{\Sigma}W''x| \leq \widetilde{O}\left(\|a_r D'_{v,x}(V^{(0)} + V^\rho)\|_2\tau_w\right) \quad (285)$$

$$\leq \widetilde{O}\left(m_2^{1/2}\tau_w\right) \quad (286)$$

■

C.1.3 Tool

Claim C.1. *Suppose $V \in \mathbb{R}^{m_2\times m_1}$ is a random matrix with entries drawn i.i.d. from $\mathcal{N}\big(0, \frac{1}{m_2}\big)$, For all unit vector $h \in \mathbb{R}^{m_1}$, and for all $g' \in \mathbb{R}^{m_2}$ that can be written as*

$$g' = g'_1 + g'_2 \text{ where } \|g'_1\| \leq 1 \text{ and } \|g'_2\|_\infty \leq \frac{1}{4\sqrt{m_2}}.$$

Let D' be the diagonal matrix where $(D')_{k,k} = \mathbb{I}_{(Vh+g')_k\geq 0} - \mathbb{I}_{(Vh)_k\geq 0}$. Then, letting $x = D'(Vh+g')$, we have

$$\|x\|_0 \leq O(m_2\|g'_1\|^{2/3} + m_2^{3/2}\|g'_2\|_\infty) \quad \text{and} \quad (287)$$

$$\|x\|_1 \leq O\big(m_2^{1/2}\|g'_1\|^{4/3} + m_2^{3/2}\|g'_2\|_\infty^2\big) \ . \quad (288)$$

Proof of Claim C.1. We first observe $g = Vh$ follows from $\mathcal{N}\big(0, \frac{\mathbf{I}}{m_2}\big)$ regardless of the choice of h. Therefore, in the remainder of the proof, we just focus on the randomness of g.

We also observe that $(D')_{j,j}$ is non-zero for some diagonal $j \in [m_2]$ only if

$$|(g'_1 + g'_2)_j| > |(g)_j| \ . \quad (289)$$

Let $\xi \leq \frac{1}{2\sqrt{m_2}}$ be a parameter to be chosen later. We shall make sure that $\|g'_2\|_\infty \leq \xi/2$.

- We denote by $S_1 \subseteq [m_2]$ the index sets where j satisfies $|(g)_j| \leq \xi$. Since we know $(g)_j \sim \mathcal{N}(0, 1/m_2)$, we have $\mathbf{Pr}[|(g)_j| \leq \xi] \leq O\left(\xi\sqrt{m_2}\right)$ for each $j \in [m_2]$. Using Chernoff bound for all $j \in [m_2]$, we have with high probability

$$|S_1| = |\{i \in [m_2] \colon |(g)_j| \leq \xi\}| \leq O(\xi m_2^{3/2}) \ .$$

Now, for each $j \in S_1$ such that $x_j \neq 0$, we must have $|x_j| = |(g + g'_1 + g'_2)_j| \leq |(g'_1)_j| + 2\xi$ so we can calculate the ℓ_2 norm of x on S_1:

$$\sum_{i \in S_1} |x_j| \leq \sum_{i \in S_1} (|(g'_1)_i| + 2\xi) \leq 2\xi|S_1| + \sqrt{|S_1|}\|g'_1\| \leq O(\|g'_1\|^2 \sqrt{\xi} m_2^{3/4} + \xi^2 m_2^{3/2}) \ .$$

- We denote by $S_2 \subseteq [m_2] \setminus S_1$ the index set of all $j \in [m_2] \setminus S_1$ where $x_j \neq 0$. Using (289), we have for each $j \in S_2$:

$$|(g'_1)_j| \geq |(g)_j| - |(g'_2)_j| \geq \xi - \|g'_2\|_\infty \geq \xi/2$$

This means $|S_2| \leq \frac{4\|g'_1\|^2}{\xi^2}$. Now, for each $j \in S_2$ where $x_j \neq 0$, we know that the signs of $(g + g'_1 + g'_2)_j$ and $(g)_j$ are opposite. Therefore, we must have

$$|x_j| = |(g + g'_1 + g'_2)_j| \leq |(g'_1 + g'_2)_j| \leq |(g'_1)_j| + \xi/2 \leq 2|(g'_1)_j|$$

and therefore

$$\sum_{j \in S_2} |x_j| \leq 2 \sum_{j \in S_2} |(g'_1)_j| \leq 2\sqrt{|S_2|}\|g'_1\| \leq 4\frac{\|g'_1\|^2}{\xi}$$

From above, we have $\|x\|_0 \leq |S_1| + |S_2| \leq O\big(\xi m_2^{3/2} + \frac{\|g'_1\|^2}{\xi^2}\big)$. Choosing $\xi = \max\big\{2\|g'_2\|_\infty, \Theta(\frac{\|g'_1\|^{2/3}}{m_2^{1/2}})\big\}$ we have the desired result on sparsity.

Combining the two cases, we have

$$\|x\|_1 \leq O\Big(\frac{\|g'_1\|^2}{\xi} + \|g'_1\|^2 \sqrt{\xi} m_2^{3/4} + \xi^2 m_2^{3/2}\Big) \leq O\Big(\frac{\|g'_1\|^2}{\xi} + \xi^2 m_2^{3/2}\Big) \ . \tag{290}$$

Choosing $\xi = \max\big\{2\|g'_2\|_\infty, \Theta(\frac{\|g'_1\|^{2/3}}{m_2^{1/2}})\big\}$, we have the desired bound on Euclidean norm. $\square$

Fact C.2. *If $\mathbf{\Sigma}$ is a diagonal matrix with diagonal entries randomly drawn from $\{-1, 1\}$. Then, given vectors x, y, with high probability*

$$|x^\top \mathbf{\Sigma} y| \leq \widetilde{O}(\|x\|_2 \cdot \|y\|_\infty)$$

C.2 Proof of Lemma 5.7

Recall

$$P_{\rho,\eta} \stackrel{\text{def}}{=} a_r D_{v,x,\rho,\eta}\left((V + V^\rho + \eta V'')D_{w,x,\rho,\eta}\left((W + W^\rho + \eta W'')x + b_1\right) + b_2\right) \tag{291}$$

$$P'_{\rho,\eta} \stackrel{\text{def}}{=} a_r D_{v,x,\rho}\left((V + V^\rho + \eta V'')D_{w,x,\rho}\left((W + W^\rho + \eta W'')x + b_1\right) + b_2\right) . \tag{292}$$

Proof of Lemma 5.7. Since $P_{\rho,\eta}$ and $P'_{\rho,\eta}$ only differ in the sign pattern, throughout this proof, we try to bound the (expected) output difference $P_{\rho,\eta} - P'_{\rho,\eta}$ by analyzing these sign changes one by one. We use the same proof structure as Lemma 5.5, that is to first bound the sign changes in the first layer, and then the second layer.

Sign Change of First Layer. We write

$$z = D_{w,x,\rho}(W + W^\rho + \eta \mathbf{\Sigma} W'')x \tag{293}$$

$$z + z' = D_{w,x,\rho,\eta}(W + W^\rho + \eta \mathbf{\Sigma} W'')x \ . \tag{294}$$

The first observation here is that, since $\|\eta \mathbf{\Sigma} W'' x\|_\infty \leq \eta \tau_{w,\infty}$, when a coordinate i has $z'_i \neq 0$, it has value at most $|z'_i| \leq \eta \tau_{w,\infty}$. In other words

$$\|z'\|_\infty \leq \eta \tau_{w,\infty} \ . \tag{295}$$

Since $\|\eta \mathbf{\Sigma} W'' x\|_\infty \leq \eta \tau_{w,\infty}$, and since each coordinate of $W^\rho x$ is i.i.d. from $\mathcal{N}(0, \sigma_w^2)$, we know

$$\forall i \in [m_1]: \quad \mathbf{Pr}_{W^\rho}[z'_i \neq 0] = \widetilde{O}\left(\eta \frac{\tau_{w,\infty}}{\sigma_w}\right). \tag{296}$$

One consequence of (296) is that, $\mathbf{Pr}[\|z'\|_0 \geq 2] \leq O_p(\eta^2)$. When $\|z'\|_0 \geq 2$, the contribution to the output difference $P_{\rho,\eta} - P'_{\rho,\eta}$ is $O_p(\eta)$. In other words, the total contribution to the output difference *in expectation* is at most $O_p(\eta^3)$.

Thus, we only need to consider the case $\|z\|_0 = 1$. Let i be this coordinate so that $z'_i \neq 0$. This happens with probability at most $O(\eta \tau_{w,\infty}/\sigma_w)$ for each $i \in [m_1]$. The contribution of z' to the output difference $P_{\rho,\eta} - P'_{\rho,\eta}$ is

$$a_r D_{v,x,\rho}(V + V^\rho + \eta V'') z' \tag{297}$$

and let us deal with the three terms separately:

- For the term $a_r D_{v,x,\rho} \eta V'' z'$, it is of absolute value at most $O_p(\eta^2)$. Since $\|z_0\|_0 = 1$ happens with probability $O_p(\eta)$, the total contribution to the expected output difference is only $O_p(\eta^3)$.
- For the term $a_r D_{v,x,\rho}(V + V^\rho) z'$, we first observe that with high probability $\|a_r D_{v,x,\rho}(V + V^\rho)\|_\infty \leq \widetilde{O}(\frac{\|a_r\|_2}{\sqrt{m_2}}) \leq \widetilde{O}(1)$. Therefore, given $\|z'\|_0 = 1$ and $\|z'\|_\infty \leq \eta \tau_{w,\infty}$, we have that $\|a_r D_{v,x,\rho}(V+V^\rho) z'\| \leq \widetilde{O}(\eta \tau_{w,\infty})$. Since this happens with probability at most $O(\eta \tau_{w,\infty}/\sigma_w) \times m_1$ —recall there are m_1 many possible $i \in [m_1]$— the total contribution to the expected output difference is

$$\widetilde{O}\left(\eta^2 m_1 \frac{\tau_w^2}{\sigma_w}\right) + O_p(\eta^3) \tag{298}$$

Sign Change of Second Layer. Recall that the sign of the ReLU of the second layer is changed from $D_{w,x,\rho}$ to $D_{w,x,\rho,\eta}$. Let us compare the vector inputs of these two matrices before ReLU is applied, that is

$$\big((V + V^\rho + \eta V'') D_{w,x,\rho,\eta}\big((W + W^\rho + \eta W'')x + b_1\big) + b_2\big) \tag{299}$$

$$- \left((V + V^\rho) D_{w,x,\rho}\left((W + W^\rho)x + b_1\right) + b_2\right). \tag{300}$$

This difference has the following four terms:

1. $\eta (V + V^\rho) D_{w,x,\rho} \mathbf{\Sigma} W'' x$.

   With $\|W'' x\|_\infty \leq \tau_{w,\infty}$ and $\|V\|_2 \leq \widetilde{O}(1)$, by Fact C.2 we know that w.h.p.

   $$\|\eta (V + V^\rho) D_{w,x,\rho} \mathbf{\Sigma} W'' x\|_\infty \leq \eta \|(V + V^\rho) D_{w,x,\rho}\|_2 \cdot \|W'' x\|_\infty \cdot \widetilde{O}(1) \leq \widetilde{O}(\eta \tau_{w,\infty}) \ .$$

2. $(V + V^\rho + \eta V'' \mathbf{\Sigma}) z'$.

   This is non-zero with probability $O_p(\eta)$, and when it is non-zero, its Euclidean norm is $O_p(\eta)$.

3. $\eta V'' \mathbf{\Sigma} z$.

   Since $\|z\|_\infty \leq \widetilde{O}(\tau_w + \sigma_w + m_1^{-1/2}) = O(m_1^{-1/2})$ owing to (276), by Fact C.2, we know that w.h.p.

   $$\|\eta V'' \mathbf{\Sigma} z\|_\infty \leq \widetilde{O}(\eta) \cdot \max_i \|V''_i\|_2 \cdot \|z\|_\infty \leq \widetilde{O}(\eta \tau_{v,\infty} m_1^{-1/2}) \ .$$

4. $\eta^2 V'' \mathbf{\Sigma} D_{w,x,\rho} \mathbf{\Sigma} W'' x$ which is $O_p(\eta^2)$ in norm.

Thus, since originally each coordinate of $V^\rho z$ follows from $\mathcal{N}(0, \sigma_v \|z\|_2^2)$ with $\|z\|_2 = \widetilde{\Omega}(1)$, using a similar argument as (296)[11] we can bound the contribution to the *expected* output difference $P_{\rho,\eta} - P'_{\rho,\eta}$ (due to sign change in second layer) by:

$$m_2 \times \widetilde{O}\left(\eta\left(\frac{1}{\sqrt{m_1}}\tau_{v,\infty} + \tau_{w,\infty}\right) \times \eta\frac{(\frac{1}{\sqrt{m_1}}\tau_{v,\infty} + \tau_{w,\infty})}{\sigma_v}\right) + O_p(\eta^3) \tag{301}$$

$$= \widetilde{O}\left(\eta^2\frac{m_2\tau_{v,\infty}^2}{\sigma_v m_1} + \eta^2\frac{m_2\tau_{w,\infty}^2}{\sigma_v}\right) + O_p(\eta^3)\ . \tag{302}$$

□

C.3 Proof of Lemma 5.8

Proof of Lemma 5.8. For notation simplicity, let us do the proof without the bias term b_1 and b_2. The proof with them are analogous.

Recall that $D_{w,x}$ and $D_{v,x}$ give the sign matrices at random initialization $W^{(0)}$, $V^{(0)}$, and we let $D_{w,x} + D'_{w,x}$ and $D_{v,x} + D'_{v,x}$ be the sign matrices at $W^{(0)} + \mathbf{\Sigma}W'$, $V^{(0)} + V'\mathbf{\Sigma}$.

Our main technique here is to use the 1-Lipschitzness of the ReLU function. It shows that, for every $(a_i, b_i)_{i\in[m_2]}$:

$$\sum_{i\in[m_2]} a_{i,r}(\sigma(a_i + b_i) - \sigma(a_i)) \le \widetilde{O}(\sqrt{m_2})\|(\sigma(a_i + b_i) - \sigma(a_i))_{i\in[m_2]}\|_2 \tag{303}$$

$$\le \widetilde{O}(\sqrt{m_2})\|b\|_2 \tag{304}$$

Define

$$z = D_{w,x}W^{(0)}x \tag{305}$$

$$z_1 = D_{w,x}\mathbf{\Sigma}W'x \tag{306}$$

$$z_2 = D'_{w,x}(W^{(0)}x + \mathbf{\Sigma}W'x) \tag{307}$$

Since w.h.p. each coordinate of z has $\|z\|_\infty = \widetilde{O}(m_1^{-1/2})$, using Fact C.2 (so using the randomness of $\mathbf{\Sigma}$), we know with high probability

$$\|V'\mathbf{\Sigma}z\|_2^2 = \sum_{i\in[m_2]} \langle v_i, \mathbf{\Sigma}z\rangle^2 \le \sum_{i\in[m_2]} \widetilde{O}\big(\|v_i\|_2^2 \cdot \|z\|_\infty^2\big) \le \widetilde{O}(m_1^{-1}) \sum_{i\in[m_1]} \|v_i\|_2^2 = \widetilde{O}(\tau_v^2 m_1^{-1}) \tag{308}$$

Thus we have: $\|V'\mathbf{\Sigma}z\|_2 \le \widetilde{O}(\tau_v m_1^{-1/2})$.

On the other hand, by Lemma 5.5, $\|z_2\|_0 \le s = \widetilde{O}(\tau_w^{4/5} m_1^{6/5})$. Therefore, using the same

[11] Namely, to first show that each coordinate $i \in [m_2]$ satisfies $(D_{v,x,\rho,\eta} - D_{v,x,\rho})_{i,i} \neq 0$ with probability $\widetilde{O}(\eta\frac{(\frac{1}{\sqrt{m_1}}\tau_{v,\infty}+\tau_{w,\infty})}{\sigma_v})$. Then, since we can ignoring terms of magnitude $O_p(\eta^3)$, it suffices to consider the case of $\|D_{v,x,\rho,\eta} - D_{v,x,\rho}\|_0 = 1$, which occurs with probability at most $m_2 \times \widetilde{O}(\eta\frac{(\frac{1}{\sqrt{m_1}}\tau_{v,\infty}+\tau_{w,\infty})}{\sigma_v})$ by union bound. Finally, each coordinate changes by at most $\widetilde{O}\big(\eta(\frac{1}{\sqrt{m_1}}\tau_{v,\infty} + \tau_{w,\infty})\big)$ by the argument above.

derivation as (260), we can bound its Euclidean norm

$$\|z_2\|_2 = \|D'_{w,x}(W^{(0)} + \mathbf{\Sigma} W')x\|_2 \le \sqrt{\sum_{i,(D'_{w,x})_{i,i}\neq 0} (\mathbf{\Sigma} W'x)_i^2} \tag{309}$$

$$\le \sqrt{\sqrt{s \cdot \sum_{i,(D'_{w,x})_{i,i}\neq 0} (W'x)_i^4}} \le O\left(s^{1/4}\tau_w\right) = \widetilde{O}\left(\tau_w^{\frac{6}{5}} m_1^{\frac{3}{10}}\right) \tag{310}$$

Thus, using Fact C.2, we also have w.h.p

$$\|V'\mathbf{\Sigma} z_2\|_2 \le \widetilde{O}\big(\|V'\|_F \|z_2\|_2\big) \le \tau_v \|z_2\|_2 \ . \tag{311}$$

Together, we can bound that

$$\sum_{i\in[m_2]} a_{i,r}\sigma\left(\langle v_i + \mathbf{\Sigma} v'_i, z + z_1 + z_2\rangle\right) \tag{312}$$

$$= \sum_{i\in[m_2]} a_{i,r}\sigma\left(\langle v_i, z + z_1 + z_2\rangle + \langle \mathbf{\Sigma} v'_i, z_1\rangle\right) \pm \widetilde{O}\left(\frac{\sqrt{m_2}}{\sqrt{m_1}}\tau_v + \sqrt{m_2}\tau_v\|z_2\|_2\right) \tag{313}$$

$$= \sum_{i\in[m_2]} a_{i,r}\sigma\left(\langle v_i, z + z_1 + z_2\rangle + \langle \mathbf{\Sigma} v'_i, z_1\rangle\right) \pm \widetilde{O}\left(\frac{\sqrt{m_2}}{\sqrt{m_1}}\tau_v + \sqrt{m_2}\tau_v\tau_w^{6/5} m_1^{3/10}\right) \tag{314}$$

Next, to bound the first term, we consider the difference between

$$① = a_r(D_{v,x} + D''_{v,x})\left(V^{(0)}(z + z_1 + z_2) + V' D_{w,x} W'x\right) \tag{315}$$

$$\stackrel{\text{def}}{=} \sum_{i\in[m_2]} a_{i,r}\sigma\left(\langle v_i, z + z_1 + z_2\rangle + \langle \mathbf{\Sigma} v'_i, z_1\rangle\right) \tag{316}$$

$$② = a_r D_{v,x}\left(V^{(0)}(z + z_1 + z_2) + V' D_{w,x} W'x\right) \tag{317}$$

where the sign change matrix $D''_{v,x}$ is due to moving input from Vz to $V^{(0)}(z+z_1+z_2)+V'D_{w,x}W'x$.

- For $Vz_1 = V^{(0)} D_{w,x}\mathbf{\Sigma} W'x$, since $\|W'x\|_2 \le \tau_w m_1^{1/4}$ and $\max_i \|V_i^{(0)}\|_\infty \le \widetilde{O}(\frac{1}{\sqrt{m_2}})$, by Fact C.2 (thus using the randomness of $\mathbf{\Sigma}$), we know that w.h.p. $\|V^{(0)} z_1\|_\infty \le \widetilde{O}\left(\tau_w \frac{m_1^{1/4}}{\sqrt{m_2}}\right)$.
- For the Vz_2 term, using the sparsity of z_2 we know that w.h.p. $\|V^{(0)} z_2\|_\infty \le \widetilde{O}(\|z_2\|_2 \sqrt{s} m_1^{-1/2}) \le \widetilde{O}\left(\tau_w^{8/5} m_1^{2/5}\right)$.
- For the final term we also have $\|V' D_{w,x} W'x\|_2 \le \|V'\|_2 \cdot \|W'x\|_2 \le \tau_w m_1^{1/4}\tau_v$.

Together, using $\|a_r\|_\infty \le \widetilde{O}(1)$ and invoking Claim C.1, we can bound it by:

$$|① - ②| \le \widetilde{O}\left(\left(\tau_w \frac{m_1^{1/4}}{\sqrt{m_2}} + \tau_w^{8/5} m_1^{2/5}\right)^2 m_2^{3/2} + (\tau_w m_1^{1/4}\tau_v)^{4/3} m_2^{1/2}\right) \tag{318}$$

Finally, from ② to our desired goal

$$a_r D_{v,x} V^{(0)} D_{w,x} W^{(0)} x + a_r D_{v,x} V' D_{w,x} W'x = a_r D_{v,x}\left(V^{(0)} z + V' D_{w,x} W'x\right) \tag{319}$$

there are still two terms:

- Since w.h.p. $\|a_r D_{v,x} V^{(0)}\|_\infty = \widetilde{O}(1)$ and z_2 is s sparse, we know that w.h.p.

$$|a_r D_{v,x} V^{(0)} z_2| \le \widetilde{O}(\|z_2\|_2 \sqrt{s}) \le \widetilde{O}\left(\tau_w^{8/5} m_1^{9/10}\right) \tag{320}$$

- Since $\|W' x\|_2 \le \tau_w m_1^{1/4}$ and w.h.p. $\|a_r D_{v,x} V^{(0)} D_{w,x}\|_\infty \le \widetilde{O}(1)$, by Fact C.2,

$$|a_r D_{v,x} V^{(0)} z_1| = |a_r D_{v,x} V^{(0)} D_{w,x} \mathbf{\Sigma} W' x| \le \widetilde{O}(\|a_r D_{v,x} V^{(0)} D_{w,x}\|_\infty \cdot \|W' x\|_2) \le \widetilde{O}\left(\tau_w m_1^{1/4}\right) \tag{321}$$

Putting together (314), (318), (320), (321), one can carefully verify that

$$\begin{aligned}
f_r(x, W^{(0)} + \mathbf{\Sigma} W', V^{(0)} + V' \mathbf{\Sigma}) &= a_r D_{v,x} V^{(0)} D_{w,x} W^{(0)} x + a_r D_{v,x} V' D_{w,x} W' x && (322)\\
&\quad \pm \widetilde{O}\left(\tau_w^{8/5} m_1^{9/10}\right) \pm \widetilde{O}\left(\tau_w m_1^{1/4}\right) && (323)\\
&\quad \pm \widetilde{O}\left(\left(\tau_w \frac{m_1^{1/4}}{\sqrt{m_2}} + \tau_w^{8/5} m_1^{2/5}\right)^2 m_2^{3/2} + (\tau_w m_1^{1/4} \tau_v)^{4/3} m_2^{1/2}\right) && (324)\\
&\quad \pm \widetilde{O}\left(\frac{\sqrt{m_2}}{\sqrt{m_1}} \tau_v + \sqrt{m_2} \tau_v \tau_w^{6/5} m_1^{3/10}\right) && (325)\\
&= a_r D_{v,x} V^{(0)} D_{w,x} W^{(0)} x + a_r D_{v,x} V' D_{w,x} W' x && (326)\\
&\quad \pm \widetilde{O}\left(\tau_w^{16/5} m_1^{4/5} m_2^{3/2} + \frac{\sqrt{m_2}}{\sqrt{m_1}} \tau_v\right) && (327)
\end{aligned}$$

Above, we have used our parameter choices of $\tau_v \in [\frac{1}{\sqrt{m_2}}, \frac{\sqrt{m_1}}{\sqrt{m_2}}]$, $\tau_w \in [\frac{1}{m_1^{3/4}}, \frac{1}{\sqrt{m_1}}]$, and $m_2 \ge m_1$. □

D Proof of Main Lemmas for Three Layers: Optimization

D.1 Proof of Lemma 5.3

Recall for $z = (x, y)$, we have

$$L_F(z, W_t) = L(F(x; W^{(0)} + W_t), y).$$

Also, recall

$$L'(\lambda_t, W_t, V_t) = \mathbb{E}_{W^\rho, V^\rho, x, y \sim \mathcal{Z}}\left[L\left(\lambda_t F\left(W^{(0)} + W^\rho + W_t, V^{(0)} + V^\rho + V_t, x\right), y\right)\right] + R(W_t, V_t) \tag{328}$$

Proof of Lemma 5.3. Let $W = W^{(0)}$, $V = V^{(0)}$. Let us define the "pseudo function" G for every W', V' as

$$G(W', V', x) = \lambda_t a^\top D_{v,x,\rho,t}[(V + V^\rho + V') D_{w,x,\rho,t}[(W + W^\rho + W')x + b_1] + b_2] \tag{329}$$

where $D_{v,x,\rho,t}$ and $D_{w,x,\rho,t}$ are the sign at weights $W + W^\rho + W_t, V + V^\rho + V_t$.

Also, recall the real network as

$$F(W', V', x) = a^\top D_{v,x,\rho,V'}[(V + V^\rho + V') D_{w,x,\rho,W'}[(W + W^\rho + W')x + b_1] + b_2] \tag{330}$$

where $D_{v,x,\rho,V'}$ and $D_{w,x,\rho,W'}$ are the sign at weights $W + W^\rho + W', V + V^\rho + V'$.

As a sanity check, we have $G(W_t, V_t, x) = \lambda_t F(W_t, V_t, x)$.

In this proof, let us define $\tau_w \stackrel{\text{def}}{=} \sigma_w m_1^{1/4} > \tau'_w$ and we will apply the coupling lemmas (e.g., Lemma 5.5) with respect to τ_w and τ_v. By our choice of parameters, as long as $L'(W_t, V_t) \leq 2$, we know that $\|W_t\|_{2,4} \leq \tau'_w \leq \tau_w$ and $\|V_t\|_{2,2} \leq \tau_v$.

At the same time, applying Lemma 5.6, we know that there exists W^*, V^* with

$$\lambda_w \|W^*\|_{2,4}^4, \lambda_v \|V^*\|_{2,2}^2 \leq \varepsilon^2 \tag{331}$$

and

$$\|F^*(x) - G^*(x)\|_2 \leq \varepsilon \quad \text{where} \quad G^*(x) \stackrel{\text{def}}{=} a^\top D_{v,x,\rho,t} V^* D_{w,x,\rho,t} W^* x \ . \tag{332}$$

Let us define update direction $W' = W_t + \sqrt{\eta}\mathbf{\Sigma} W^*, V' = V_t + \sqrt{\eta} V^* \mathbf{\Sigma}$.

Change in Regularizer. We first consider the change of the regularizer. We know that

$$\mathbb{E}_{\mathbf{\Sigma}}\left[\|V_t + \sqrt{\eta} V^* \mathbf{\Sigma}\|_F^2\right] = \|V_t\|_F^2 + \eta \|V^*\|_F^2 \ . \tag{333}$$

On the other hand,

$$\mathbb{E}\left[\|W_t + \sqrt{\eta}\mathbf{\Sigma} W^*\|_{2,4}^4\right] = \sum_{i\in[m_1]} \mathbb{E}\left[\|[W_t]_i + \sqrt{\eta}\mathbf{\Sigma} W_i^*\|_2^4\right] \tag{334}$$

For each term $i \in [m_1]$, we can bound

$$\|[W_t]_i + \sqrt{\eta}\mathbf{\Sigma} W_i^*\|_2^2 = \|[W_t]_i\|_2^2 + \eta \|W_i^*\|_2^2 + 2\sqrt{\eta}\langle [W_t]_i, W_i^*\rangle (\mathbf{\Sigma})_{i,i} \tag{335}$$

and therefore

$$\begin{aligned} \mathbb{E}\left[\|[W_t]_i + \sqrt{\eta}\mathbf{\Sigma} W_i^*\|_2^4\right] &= \|[W_t]_i\|_2^4 + 4\eta\langle [W_t]_i, W_i^*\rangle^2 + \eta^2 \|W_i^*\|_2^4 + 2\eta \|[W_t]_i\|_2^2 \|W_i^*\|_2^2 &(336)\\ &\leq \|[W_t]_i\|_2^4 + 6\eta \|[W_t]_i\|_2^2 \|W_i^*\|_2^2 + O_p(\eta^2) \ . &(337)\end{aligned}$$

By Cauchy-Schwarz, we know that

$$\sum_{i\in[m_1]} \|[W_t]_i\|_2^2 \|W_i^*\|_2^2 \leq \sqrt{\left(\sum_{i\in[m_1]} \|[W_t]_i\|_2^4\right)\left(\sum_{i\in[m_1]} \|W_i^*\|_2^4\right)} \leq \|W_t\|_{2,4}^2 \|W^*\|_{2,4}^2 \tag{338}$$

Therefore, we know that

$$\mathbb{E}\left[\|W_t + \sqrt{\eta}\mathbf{\Sigma} W^*\|_{2,4}^4\right] \leq \|W_t\|_2^4 + 6\eta \|W_t\|_{2,4}^2 \|W^*\|_{2,4}^2 + O_p(\eta^2) \tag{339}$$

By $\lambda_v \|V^*\|_F^2 \leq \varepsilon^2$ and $\lambda_w \|W^*\|_{2,4}^4 \leq \varepsilon^2$, we know that

$$\mathbb{E}[R(W', V')] \leq R(W_t, V_t) + 20\eta\varepsilon. \tag{340}$$

Change in Objective. We now consider the change in the objective value. For polynomially small η, by Lemma 5.6, we know that the construction of good network W^*, V^* satisfies $\tau_{v,\infty} \leq \frac{1}{m_1^{999/2000}}$ and $\tau_{w,\infty} \leq \frac{1}{m_1^{999/1000}}$. By Lemma 5.7, we have:

$$\mathbb{E}_{W^\rho, V^\rho}\left|G(W', V', x) - F(W', V', x)\right| \leq O(\varepsilon\eta) + O_p(\eta^{1.5}). \tag{341}$$

First focus on $G(W', V', x)$. By Lemma 5.5, as long as $L'(W_t, V_t) \leq 2$ we must have —by our choice of regularizer— $\|W_t\|_{2,4} \leq \tau_w$ and $\|V_t\|_{2,2} \leq \tau_v$. Therefore,

$$G(W', V', x) = \lambda_t F(W_t, V_t, x) + \sqrt{\eta} G'(x) + \eta G^*(x) \pm O_p(\eta^{1.5}) \tag{342}$$

where $\mathbb{E}_{\mathbf{\Sigma}}[G(x)] = 0$ and w.h.p. $\|G'(x)\|_2 \leq \varepsilon$ and $G^*(x)$ is from (332).

Combining (341) and (342), we know that for every fixed x, y in the support of distribution $\mathcal{Z}$:

$$\mathbb{E}_{W^\rho,V^\rho,\mathbf{\Sigma}}[L(\lambda_t F(W',V',x),y)] \tag{343}$$

$$\leq \mathbb{E}_{W^\rho,V^\rho,\mathbf{\Sigma}}[L(G(W',V',x),y)] + O(\eta\varepsilon) + O_p(\eta^{1.5}). \tag{344}$$

$$\overset{①}{\leq} \mathbb{E}_{W^\rho,V^\rho}\left[L(\mathbb{E}_{\mathbf{\Sigma}}[G(W',V',x),y])\right] \tag{345}$$

$$+ \mathbb{E}_{W^\rho,V^\rho,\mathbf{\Sigma}}\left\|G(W',V',x) - G(W_t,V_t,x)\right\|^2 + O(\eta\varepsilon) + O_p(\eta^{1.5}). \tag{346}$$

$$\overset{②}{\leq} \mathbb{E}_{W^\rho,V^\rho} L\left(G(W_t,V_t,x) + \eta G^*(x), y\right) + O(\eta\varepsilon) + O_p(\eta^{1.5}). \tag{347}$$

$$\overset{③}{\leq} \mathbb{E}_{W^\rho,V^\rho} L\left(G(W_t,V_t,x) + \eta F^*(x), y\right) + O(\eta\varepsilon) + O_p(\eta^{1.5}). \tag{348}$$

$$= \mathbb{E}_{W^\rho,V^\rho} L\left(\lambda_t F(W_t,V_t,x) + \eta F^*(x), y\right) + O(\eta\varepsilon) + O_p(\eta^{1.5}). \tag{349}$$

Above, ① uses the 1-smoothness of L, meaning

$$\mathbb{E}[L(v)] \leq L(\mathbb{E}[v]) + \mathbb{E}[\|v - \mathbb{E}[v]\|^2] \cdot \|\nabla^2 L(\mathbb{E}[v])\|_2$$

and $\mathbb{E}_{\mathbf{\Sigma}}[G(W',V',x)] = G(W_t,V_t,x) + \eta G^*(x)$. Inequality ② also uses $\mathbb{E}_{\mathbf{\Sigma}}[G(W',V',x)] = G(W_t,V_t,x) + \eta G^*(x)$. Inequality ③ uses (332).

Next, by convexity of the loss function, we have

$$L\left(\lambda_t F(W_t,V_t,x) + \eta F^*(x), y\right) = L\left((1-\eta)(1-\eta)^{-1}\lambda_t F(W_t,V_t,x) + \eta F^*(x), y\right) \tag{350}$$

$$\leq (1-\eta)\left(L((1-\eta)^{-1}\lambda_t F(W_t,V_t,x), y)\right) + \eta L(F^*(x), y) \tag{351}$$

For sufficiently small η, we know that

$$L((1-\eta)^{-1}\lambda_t F(W_t,V_t,x), y) + L((1-\eta)\lambda_t F(W_t,V_t,x), y) \leq 2L\left(\lambda_t F(W_t,V_t,x), y\right) + O_p(\eta^2) \tag{352}$$

Putting this into (351), we have

$$L\left(\lambda_t F(W_t,V_t,x) + \eta F^*(x), y\right) \tag{353}$$

$$\leq (1-\eta)\left(2L\left(\lambda_t F(W_t,V_t,x), y\right) - L((1-\eta)\lambda_t F(W_t,V_t,x), y)\right) \tag{354}$$

$$+ \eta L(F^*(x), y) + O_p(\eta^2) \tag{355}$$

Let us denote

$$c_1 = \mathbb{E}_{\mathbf{\Sigma}}\left[L'\left(\lambda_t, W_t + \sqrt{\eta}\mathbf{\Sigma}W^*, V_t + \sqrt{\eta}V^*\mathbf{\Sigma}\right)\right] \tag{356}$$

$$= \mathbb{E}_{W^\rho,V^\rho,\mathbf{\Sigma},(x,y)\sim\mathcal{Z}}[L(\lambda_t F(W',V',x),y)] \tag{357}$$

$$c_2 = L'((1-\eta)\lambda_t, W_t, V_t) \tag{358}$$

$$c_3 = L'(\lambda_t, W_t, V_t) \tag{359}$$

The above two inequalities (349) and (355) together imply

$$c_1 \leq (1-\eta)\left(2c_3 - c_2\right) + \eta(\varepsilon_0 + O(\varepsilon)) + O_p(\eta^{1.5}) \tag{360}$$

Rearranging, we have:

$$\frac{1}{2}(1-\eta)^{-1}c_1 + \frac{1}{2}c_2 \leq c_3 + \eta\frac{1}{2}\varepsilon_0 + O(\eta\varepsilon) + O_p(\eta^{1.5}) \tag{361}$$

Therefore,

$$\left(\frac{1}{2}(1-\eta)^{-1} + \frac{1}{2}\right)\min\{c_1, c_2\} \leq c_3 + \eta\frac{1}{2}\varepsilon_0 + O(\eta\varepsilon) + O_p(\eta^{1.5}) \tag{362}$$

and this implies that

$$\min\{c_1, c_2\} \leq \left(1 - \eta\frac{1}{2}\right) c_3 + \eta\frac{1}{2}\varepsilon_0 + O(\eta\varepsilon) + O_p(\eta^{1.5}) \tag{363}$$

Therefore, as long as $c_3 \geq (1+\gamma)\varepsilon_0 + \Omega(\varepsilon)$ and $\gamma \in [0,1]$, we have:

$$\min\{c_1, c_2\} \leq (1 - \eta\gamma/4)c_3 \tag{364}$$

This completes the proof. □

D.2 Proof of Lemma 5.4

Proof. For the first variant of SGD, note that there are only $\leq T/T_w = \Theta(\log\frac{1}{\varepsilon}/\eta)$ rounds of weight decay, which implies that $\lambda_t \geq (\varepsilon')^{-O(1)}$ is always satisfied. By Lemma 5.3, we know that as long as $L \in [(1+\gamma)\varepsilon_0 + \widetilde{\Omega}(\varepsilon), 2]$, then there exists $\|W^*\|_F, \|V^*\|_F \leq 1$ such that either

$$\mathbb{E}_{\mathbf{\Sigma}}\left[L'\left(\lambda_t, W_t + \sqrt{\eta}\mathbf{\Sigma}W^*, V_t + \sqrt{\eta}V^*\mathbf{\Sigma}\right)\right] \leq (1-\eta\gamma/4)(L'(\lambda_t, W_t, V_t)) \tag{365}$$

or

$$L'((1-\eta)\lambda_t, W_t, V_t) \leq (1-\eta\gamma/4)(L'(\lambda_t, W_t, V_t)) \tag{366}$$

In the first case, since L' is $L_0 = \mathbf{poly}(m_1, m_2)$ Lipschitz and $L_1 = \mathbf{poly}(m_1, m_2)$ smooth[12], by Lemma A.8, either $\nabla^2 L'$ has a negative eigen-direction of $\leq -1/(m_1m_2)^8$, or $\|\nabla L'\|_2 \geq 1/(m_1m_2)^8$. Since L' also has stochastic gradient bounded by $\mathbf{poly}(m_1, m_2)$ w.h.p. Apply known results about escaping from saddle points [GHJY15], we know that for sufficiently large $T_w = \mathbf{poly}(m_1, m_2)$, for at least $(1 - \frac{1}{L_0(m_1m_2)^{10}})T_w$ many iterations in $t \in [t_0 + T_w, t_0 + 2T_w]$ it satisfies:

$$\mathbb{E}_{\mathbf{\Sigma}}\left[L'\left(\lambda_t, W_t + \sqrt{\eta}\mathbf{\Sigma}W^*, V_t + \sqrt{\eta}V^*\mathbf{\Sigma}\right)\right] > (1-\eta\gamma/4)(L'(\lambda_t, W_t, V_t)) \tag{367}$$

Thus, as long as $L' \geq (1+\gamma)\varepsilon_0 + \widetilde{\Omega}(\varepsilon)$, we have for these t:

$$L'((1-\eta)\lambda_t, W_t, V_t) \leq (1-\eta\gamma/4)(L'(\lambda_t, W_t, V_t)). \tag{368}$$

Let us denote the set of these t as $\mathcal{S}$. If we pick $t_0 \in \mathcal{S}$, we will decrease objective to $(1-\eta\gamma/4)L'$, as long as $L' \geq (1+\gamma)\varepsilon_0 + \widetilde{\Omega}(\varepsilon)$. Otherwise, if $t_0 \notin \mathcal{S}$, using the norm bounds of the weights we know that we would increase the objective to at most $L' + \eta L_0(m_1m_2)^4$ w.h.p. Thus, as long as $L' \geq (1+\gamma)\varepsilon_0 + \widetilde{\Omega}(\varepsilon)$, the decrease in expectation is:

$$\mathbb{E}_{t_0}[L'((1-\eta)\lambda_{t_0}, W_{t_0}, V_{t_0}) \mid L'(\lambda_{t_0}, W_{t_0}, V_{t_0}) \geq (1+\gamma)\varepsilon_0 + \widetilde{\Omega}(\varepsilon)] \tag{369}$$

$$\leq (1-\eta\gamma/8)\mathbb{E}_{t_0}[L'(\lambda_{t_0}, W_{t_0}, V_{t_0}) \mid L'(\lambda_{t_0}, W_{t_0}, V_{t_0}) \geq (1+\gamma)\varepsilon_0 + \widetilde{\Omega}(\varepsilon)] \tag{370}$$

On the other hand, as long as $L' \leq 2$ it holds that $L'((1-\eta)\lambda_{t_0}, W_{t_0}, V_{t_0}) = L'(\lambda_{t_0}, W_{t_0}, V_{t_0}) \pm \eta L_0(m_1m_2)^4$.

Note that initially, $L' \leq 1$. Since both SGD or the weight decay will only decrease L' in expectation as long as $L' \geq 3/2$, using a sufficiently small learning rate $\eta = \frac{1}{\mathbf{poly}(m_1,m_2)}$ we have $L' \leq 2$ w.h.p.

Finally, for any constant $\gamma \in (0, 1/4]$, to decrease the objective to $\leq \varepsilon$, using a standard martingale concentration we know that w.p $\geq 99/100$ after $T_2 = T/(4T_w) = \Theta(\log\frac{1}{\varepsilon}/\eta)$ rounds of weight decay, we have $L' \leq (1+\gamma)\varepsilon_0 + \widetilde{\Omega}(\varepsilon)$.

For the second variant of the SGD, note that

$$\mathbf{\Sigma}_1 W_t + \sqrt{\eta}\mathbf{\Sigma}_1\mathbf{\Sigma}W^*, \quad V_t\mathbf{\Sigma}_1 + \sqrt{\eta}V^*\mathbf{\Sigma}_1\mathbf{\Sigma} \tag{371}$$

[12]Due to $\nabla(f * g) = f * \nabla g$ for every bounded function f and differentiable function g

satisfies that $\mathbf{\Sigma}_1\mathbf{\Sigma}$ is still a diagonal matrix with each diagonal entry i.i.d. $\{\pm 1\}$. Thus, the convergence result also holds for loss:

$$L''(\lambda_t, W_t, V_t) = \mathbb{E}_{W^\rho, V^\rho, \mathbf{\Sigma}, x, y \sim \mathcal{Z}} \left[L\left(\lambda_t F\left(W^{(0)} + W^\rho + \mathbf{\Sigma} W_t, V^{(0)} + V^\rho + V_t \mathbf{\Sigma}, x \right), y \right) \right] + R(W_t, V_t). \tag{372}$$

□

E Proof of Main Lemmas for Three Layers: Generalization

Proof of Lemma 5.9. By Lemma 5.8 and Lemma 5.5, we know that by our choice of parameters,

$$f_r(x, W + \mathbf{\Sigma} W', V + V' \mathbf{\Sigma}) = a_r \left(D_{v,x} V D_{w,x} (W x + b_1) + b_2 \right) + a_r D_{v,x} V' D_{w,x} W' x \pm \widetilde{O}(\varepsilon). \tag{373}$$

Now, we focus on the Rademacher complexity of the right hand side. By the contraction Lemma, we have that

$$\mathbb{E}_\xi \sup_{\|V'\|_2 \le \tau_v, \|W'\|_{2,4} \le \tau_w} \left| \sum_{j \in [N]} \xi_j \left(a_r \left(D_{v,x_j} V D_{w,x_j} (W x_j + b_1) + b_2 \right) + a_r D_{v,x_j} V' D_{w,x_j} W' x_j \right) \right| \tag{374}$$

$$\le \mathbb{E}_\xi \sup_{\|V'\|_2 \le \tau_v, \|W'\|_{2,4} \le \tau_w} \left| \sum_{j \in [N]} \xi_j a_r \left(D_{v,x_j} V D_{w,x_j} (W x_j + b_1) + b_2 \right) \right| \tag{375}$$

$$+ \mathbb{E}_\xi \sup_{\|V'\|_2 \le \tau_v, \|W'\|_{2,4} \le \tau_w} \left| \sum_{j \in [N]} \xi_j a_r D_{v,x_j} V' D_{w,x_j} W' x_j \right|. \tag{376}$$

For the first term, it does not depend on V', W'. Since w.h.p. $|a_r \left(D_{v,x_j} V D_{w,x_j} (W x_j + b_1) + b_2 \right)| \le \widetilde{O}(1)$, we can bound it by:

$$\mathbb{E}_\xi \left| \sum_{j \in [N]} \xi_j a_r \left(D_{v,x_j} V D_{w,x_j} (W x_j + b_1) + b_2 \right) \right| \le \widetilde{O}(\sqrt{N}). \tag{377}$$

For the second term we have that

$$\sup_{\|W'\|_{2,4} \le \tau_w, \|V'\|_F \le \tau_v} \left| \sum_{j \in [N]} \xi_j a_r D_{v,x_j} V' D_{w,x_j} W' x_j \right| \tag{378}$$

$$\le \sup_{\|W'\|_{2,4} \le \tau_w, \|V'\|_F \le \tau_v} \left\| \sum_{j \in [N]} \xi_j x_j a_r D_{v,x_j} V' D_{w,x_j} \right\|_F \|W'\|_F \tag{379}$$

$$\le \tau_w m_1^{1/4} \sup_{\|V'\|_F \le \tau_v} \left\| \sum_{j \in [N]} \xi_j x_j a_r D_{v,x_j} V' D_{w,x_j} \right\|_F \tag{380}$$

Let us bound the last term entry by entry. Let $[D_{w,x_j}]_q$ denote the q-th column of D_{w,x_j}, $[V']_q$ the q-th column of V'.

$$\sup_{\|V'\|_F \le \tau_v} \left\| \sum_{j\in[N]} \xi_j x_j a_r D_{v,x_j} V' D_{w,x_j} \right\|_F \tag{381}$$

$$= \sup_{\|V'\|_F \le \tau_v} \sqrt{\sum_{p\in[d],q\in[m_1]} \left(\sum_{j\in[N]} \xi_j x_{j,p} a_r D_{v,x_j} V' [D_{w,x_j}]_q \right)^2} \tag{382}$$

$$= \sup_{\|V'\|_F \le \tau_v} \sqrt{\sum_{p\in[d],q\in[m_1]} \left(\sum_{j\in[N]} \xi_j [D_{w,x_j}]_{q,q} x_{j,p} a_r D_{v,x_j} [V']_q \right)^2} \tag{383}$$

$$\le \sup_{\|V'\|_F \le \tau_v} \sqrt{\sum_{p\in[d],q\in[m_1]} \left\| \sum_{j\in[N]} \xi_j [D_{w,x_j}]_{q,q} x_{j,p} a_r D_{v,x_j} \right\|^2 \|[V']_q\|^2} \tag{384}$$

For random ξ_j, we know that w.h.p.

$$\left\| \sum_{j\in[N]} \xi_j [D_{w,x_j}]_{q,q} x_{j,p} a_r D_{v,x_j} \right\|^2 = \widetilde{O}\left(\|a_r\|_2^2 \sum_{j\in[N]} x_{j,p}^2 \right), \tag{385}$$

Thus,

$$\mathbb{E}_\xi \sup_{\|V'\|_F \le \tau_v} \sqrt{\sum_{p\in[d],q\in[m_1]} \left\| \sum_{j\in[N]} \xi_j [D_{w,x_j}]_{q,q} x_{j,p} a_r D_{v,x_j} \right\|^2 \|[V']_q\|^2} \tag{386}$$

$$\le \widetilde{O}\left(\sup_{\|V'\|_F \le \tau_v} \|a_r\|_2 \sqrt{\sum_{p\in[d],q\in[m_1]} \sum_{j\in[N]} x_{j,p}^2 \|[V']_q\|^2} \right) \tag{387}$$

$$\le \widetilde{O}\left(\|a_r\|_2 \sup_{\|V'\|_F \le \tau_v} \sqrt{\sum_{q\in[m_1]} \sum_{j\in[N]} \|[V']_q\|^2} \right) \tag{388}$$

$$\le \widetilde{O}\left(\tau_v \sqrt{m_2 N} \right). \tag{389}$$

This completes the proof. □

F Proof of the Main Theorems for Three Layers

Proof of Theorem 4.2 and 4.3. We first consider the convergence: Since we know that

$$\mathbb{E}_{(x,y)\sim\mathcal{Z},W^\rho,V^\rho} L(\lambda_T F(x; W^{(0)} + W'_T + W^\rho, V^{(0)} + V'_T + V^\rho), y) \le (1+\gamma)\varepsilon_0 + \varepsilon. \tag{390}$$

Thus since the loss L has range $\in [0,1]$, by randomly sample $\widetilde{O}(1/\varepsilon^2)$ many $W^{\rho,j}, V^{\rho,j}$, we know that w.h.p. there exists one j^* with

$$\mathbb{E}_{z\in\mathcal{Z}} L_F(z, \lambda_T, W^{(0)} + W^{\rho,j^*} + W'_T, V^{(0)} + V^{\rho,j^*} + V'_T) \le (1+\gamma)\varepsilon_0 + 2\varepsilon \tag{391}$$

Finally, the sample complexity bound for the first algorithm can be adapted from [BFT17, Theorem 1.1], by noticing that $\|W_T\|_{2,1} \le m_1^{3/4}\|W_T\|_{2,4} \le m_1^{3/4}\tau'_w$ and $\|V_T\|_{2,1} \le m_2^{1/2}\|V_T\|_{2,2} \le$

$m_2^{1/2}\tau_v$.[13] Using our choices of τ'_w and τ_v from the statement of Lemma 5.3 as well as $m_1 = m_2$, this bound implies as long as $N \geq \widetilde{O}(M(m_2)^2)$, for randomly sampled W^ρ, V^ρ, it holds that w.h.p.[14]

$$\mathbb{E}_{z\in\mathcal{D}} L_F(z, \lambda_T, W^{(0)} + W^\rho + W'_T, V^{(0)} + V^\rho + V'_T) \tag{392}$$
$$\leq \mathbb{E}_{z\in\mathcal{Z}} L_F(z, \lambda_T, W^{(0)} + W^\rho + W'_T, V^{(0)} + V^\rho + V'_T) + 10\varepsilon \tag{393}$$

Thus, w.h.p. for all the $\widetilde{\Theta}(1/\varepsilon^2)$ many samples $W^{\rho,j}, V^{\rho,j}$ it holds:

$$\mathbb{E}_{z\in\mathcal{D}} L_F(z, \lambda_T, W^{(0)} + W^{\rho,j} + W'_T, V^{(0)} + V^{\rho,j} + V'_T) \tag{394}$$
$$\leq \mathbb{E}_{z\in\mathcal{Z}} L_F(z, \lambda_T, W^{(0)} + W^{\rho,j} + W'_T, V^{(0)} + V^{\rho,j} + V'_T) + 10\varepsilon \tag{395}$$

This implies

$$\mathbb{E}_{z\in\mathcal{D}} L_F(z, \lambda_T, W^{(0)} + W^{\rho,j^*} + W'_T, V^{(0)} + V^{\rho,j^*} + V'_T) \leq (1+\gamma)\varepsilon_0 + 10\varepsilon \tag{396}$$

As for the second algorithm, by Lemma 5.8, we know that w.h.p. for each output $r \in [k]$

$$f_r(x, W^{(0)} + \mathbf{\Sigma} W'_T, V^{(0)} + V'_T \mathbf{\Sigma}) = a_r D_{v,x} V^{(0)} D_{w,x} W^{(0)} x + a_r D_{v,x} V' D_{w,x} W' x \tag{397}$$
$$\pm \widetilde{O}\left((\tau'_w)^{16/5} m_1^{4/5} m_2^{3/2} + \frac{\sqrt{m_2}}{\sqrt{m_1}} \tau_v \right) \tag{398}$$
$$= a_r D_{v,x} V^{(0)} D_{w,x} W^{(0)} x + a_r D_{v,x} V' D_{w,x} W' x \pm \varepsilon \tag{399}$$

where the last equality is by our choices of τ_w and τ_v from the statement of Lemma 5.3.

Now, we can view W^ρ, V^ρ as part of the initialization, this implies that w.h.p.

$$f_r(x, W^{(0)} + W^\rho + \mathbf{\Sigma} W', V^{(0)} + V^\rho + V' \mathbf{\Sigma}) \tag{400}$$
$$= a_r D_{v,x,\rho} V^{(0)} D_{w,x,\rho} W^{(0)} x + a_r D_{v,x,\rho} V' D_{w,x,\rho} W' x \pm \varepsilon \tag{401}$$

where $D_{v,x,\rho}$ and $D_{w,x,\rho}$ are the diagonal sign indicator matrices at weights $W^{(0)} + W^\rho$, $V^{(0)} + V^\rho$.

Since $a_r D_{v,x,\rho} V^{(0)} D_{w,x,\rho} W^{(0)} x + a_r D_{v,x,\rho} V' D_{w,x,\rho} W' x$ is independent of the choice of $\mathbf{\Sigma}$, we have w.h.p. over randomly chosen $\mathbf{\Sigma}$, there exists one $W^{\rho,j^*}, V^{\rho,j^*}$ in the random sampled noise matrices such that

$$\mathbb{E}_{z\in\mathcal{Z}} L_F(z, \lambda_T, W^{(0)} + W^{\rho,j^*} + W'_T, V^{(0)} + V^{\rho,j^*} + V'_T) \leq (1+\gamma)\varepsilon_0 + 2\varepsilon\ . \tag{402}$$

As for the sample complexity of this second algorithm, it comes from Lemma 5.9. Since $L \leq 2$, we know that $\|W_T\|_{2,4} \leq \tau'_w$ and $\|V_T\|_{2,2} \leq \tau_v$. By our choice of parameter, $m_2^{1/2} m_1^{1/4} \tau_v \tau'_w \leq M$. Thus, apply Lemma 5.9 together with the fact that L is 1-Lipschitz and 1-smooth we complete the proof (because Rademacher complexity implies generalization).

$\square$

[13] They have defined $\|A\|_{2,1} = \|A_{1,:}\|_2 + \cdots \|A_{d,:}\|_2$ which is at most $d^{3/4}$ times our $\|A\|_{2,4}$ norm, and at most $d^{1/2}$ times our $\|A\|_{2,2} = \|A\|_F$ norm.

[14] More precisely, their Theorem 1.1 states a generalization bound for classification only. However, their proof of Theorem 1.1 in fact calculates the Rademacher complexity that we need. To see how the m_2 factor shows up, let us ignore *small-order terms* and use the big-O notion to only show m dependency. The Rademacher complexity from [BFT17, Theorem 1.1] is essentially

$$O\left(\frac{1}{\sqrt{N}} \|(a_{r,i})_{r,i}\|_\sigma \|W^{(0)} + W_T\|_\sigma \|V^{(0)} + V_T\|_\sigma \cdot \left(\frac{\|W_T\|_{2,1}^{2/3}}{\|W^{(0)} + W_T\|_\sigma^{2/3}} + \frac{\|V_T\|_{2,1}^{2/3}}{\|V^{(0)} + V_T\|_\sigma^{2/3}} \right)^{3/2} \right)\ .$$

In our parameter settings, we have $\|W_T\|_{2,1} \leq O(1)$, $\|V_T\|_{2,1} \leq O(\sqrt{m_2})$, $\|(a_{r,i})_{r\in[k],i\in[m_2]}\|_\sigma \leq O(\sqrt{m_2})$, $\|V^{(0)} + V_T\|_\sigma = \Theta(1)$, $\|W^{(0)} + W_T\|_\sigma = \Theta(1)$, where $\|\cdot\|_\sigma$ is the matrix spectral norm. Therefore, choosing $N \geq \Omega((m_2)^2)$ (together with other polynomial factors that we denote by M) suffices.

G Proof for Two Layer Networks

The proof for two-layer networks are simpler, so we put it in a separate section.

G.1 Existential Result

We consider another function defined as $G(x; W) = (g_1(x; W), \dots, g_k(x; W))$ for the weight matrix W, where

$$g_r(x; W) = \sum_{i=1}^{m} a_{r,i}(\langle w_i, x\rangle + b_i^{(0)})\mathbb{I}[\langle w_i^{(0)}, x\rangle + b_i^{(0)} \geq 0], \tag{403}$$

where w_i is the r-th row of W. We call this G a pseudo network. We show here that there exists a good pseudo network near the initialization.

Though not explicitly written out, combining with the coupling result in the next subsection, we can show the existence of a good real network near the initialization.

Lemma G.1. *For every $\varepsilon \in (0, 1/\mathfrak{C}(\phi, 1))$, let $\varepsilon_a = \varepsilon/\widetilde{\Theta}(1)$, then there exists $M = \mathbf{poly}(\mathfrak{C}(\phi, \sqrt{\log(p/\varepsilon)}), 1/\varepsilon, k, p)$ such that if $m \geq M$, then with high probability, there exists $\widetilde{W} = (\widetilde{w}_1, \dots, \widetilde{w}_m)$ such that $\|\widetilde{w}_i\|_2 \leq \frac{C_0}{m}$ for some $C_0 = \mathbf{poly}(\mathfrak{C}(\phi, \sqrt{\log(p/\varepsilon)}), 1/\varepsilon, k, p, \log m)$, and*

$$\mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\sum_{r=1}^{k}\left|f_r^*(x) - g_r(x; \widetilde{W})\right|\right] \leq \varepsilon, \tag{404}$$

and consequently,

$$\mathbb{E}_{(x,y)\sim\mathcal{D}}\left[L(G(x; \widetilde{W}), y)\right] \leq \varepsilon_0 + \varepsilon. \tag{405}$$

Corollary G.2. *In the same setting as Lemma G.1, we have that w.h.p.*

$$\mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\sum_{r=1}^{k}\left|f_r^*(x) - g_r(x; W^{(0)} + \widetilde{W})\right|\right] \leq \varepsilon, \tag{406}$$

and consequently,

$$\mathbb{E}_{(x,y)\sim\mathcal{D}}\left[L(G(x; W^{(0)} + \widetilde{W}), y)\right] \leq \varepsilon_0 + \varepsilon. \tag{407}$$

Proof of Lemma G.1. Recall the pseudo network is given by

$$g_r(x; W) = \sum_{i=1}^{m} a_{r,i}(\langle w_i, x\rangle + b_i^{(0)})\mathbb{I}[\langle w_i^{(0)}, x\rangle + b_i^{(0)} \geq 0]. \tag{408}$$

By Lemma 5.1, we know that for every $\varepsilon > 0$, there is a function $h^{(i)}$ such that the range and Lipschitzness of $h^{(i)}$ is bounded by $\mathfrak{C}(\phi, \sqrt{\log(1/\varepsilon)})$ with:

$$\left|\mathbb{E}\left[\mathbb{I}_{\alpha_1 x_1 + \beta_1\sqrt{1-x_1^2} + b_0 \geq 0} h^{(i)}(\alpha_1, b_0)\right] - \phi_i(x_1)\right| \leq 2\varepsilon \tag{409}$$

where $\alpha_1, \beta_1 \sim \mathcal{N}(0, 1)$ and $b_0 = 0$ w.p. $1/2$, $b_0 \sim \mathcal{N}(0, 2)$ otherwise, are independent random variables.

Fit a single function $\phi_i(\langle w_{1,i}^*, x\rangle)\langle w_{2,i}^*, x\rangle$: We shall construct the weights $\widetilde{w}_j$. Define

$$\widetilde{w}_j = \frac{1}{\varepsilon_a^2} a_{r,j} h^{(i)}\left(\sqrt{m}\langle w_j^{(0)}, w_{1,i}^*\rangle, \sqrt{m} b_j^{(0)}\right) w_{2,i}^* \tag{410}$$

where $\sqrt{m}\langle w_j^{(0)}, w_{1,i}^*\rangle, \sqrt{m}b_j^{(0)}$ has the same distribution with α_1, b_0.

By the above result, we have that

$$\mathbb{E}_{w_j^{(0)}, b_j^{(0)}, a_{r,j}}\left[a_{r,j}\mathbb{I}_{\langle w_j^{(0)}, x\rangle + b_j^{(0)} \geq 0}\langle \widetilde{w}_j, x\rangle\right] \tag{411}$$

$$= \mathbb{E}_{w_j^{(0)}, b_j^{(0)}}\left[\mathbb{I}_{\langle w_j^{(0)}, x\rangle + b_j^{(0)} \geq 0} h^{(i)}\left(\sqrt{m}\langle w_j^{(0)}, w_{1,i}^*\rangle, \sqrt{m}b_j^{(0)}\right)\langle w_{2,i}^*, x\rangle\right] \tag{412}$$

$$= \phi_i(\langle w_{1,i}^*, x\rangle)\langle w_{2,i}^*, x\rangle \pm 2\varepsilon. \tag{413}$$

Here we use that $\sqrt{m}w_j^{(0)}$ can be factorized into $\alpha_1 w_{1,i}^* + \beta_1 w_\perp$ with $w_\perp$ being the random direction orthogonal to $w_{1,i}^*$, so $\langle w_j^{(0)}, x\rangle$ can be replaced by $\alpha_1\langle w_{1,i}^*, x\rangle + \beta_1\sqrt{1 - \langle w_{1,i}^*, x\rangle^2}$.

Fit a combination To fit $\sum_{i\in[p]} \phi_i(\langle w_{1,i}^*, x\rangle)\langle w_{2,i}^*, x\rangle$, we can just define

$$\widetilde{w}_j = \frac{1}{\varepsilon_a^2} a_{r,j} \sum_{i\in[p]} h^{(i)}\left(\sqrt{m}\langle w_j^{(0)}, w_{1,i}^*\rangle, \sqrt{m}b_j^{(0)}\right) w_{2,i}^* \tag{414}$$

Using the same argument as above we can show the result.

Fit multiple output: Let us define

$$\widetilde{w}_j = \frac{1}{\varepsilon_a^2} \sum_{r\in[k]} a_{r,j} \sum_{i\in[p]} h^{(i)}\left(\sqrt{m}\langle w_j^{(0)}, w_{1,i}^*\rangle, \sqrt{m}b_j^{(0)}\right) w_{2,i}^*. \tag{415}$$

By definition of the initialization, we know that for $r' \neq r$, $\mathbb{E}[a_{r,j}a_{r',j}] = 0$. Thus for every $r \in [k]$:

$$\mathbb{E}_{w_j^{(0)}, b_j^{(0)}, a_{r,j}}\left[a_{r,j}\mathbb{I}_{\langle w_j^{(0)}, x\rangle + b_j^{(0)} \geq 0}\langle \widetilde{w}_j, x\rangle\right] \tag{416}$$

$$= \mathbb{E}_{w_j^{(0)}, b_j^{(0)}}\left[\sum_{i\in[p]} \mathbb{I}_{\langle w_j^{(0)}, x\rangle + b_j^{(0)} \geq 0} h^{(i)}\left(\sqrt{m}\langle w_j^{(0)}, w_{1,i}^*\rangle, \sqrt{m}b_j^{(0)}\right)\langle w_{2,i}^*, x\rangle\right] \tag{417}$$

$$= \sum_{i\in[p]} \phi_i(\langle w_{1,i}^*, x\rangle))\langle w_{2,i}^*, x\rangle \pm 2p\varepsilon. \tag{418}$$

The sample complexity follows from Lemma A.1, by scaling $\widetilde{w}_i$ by a factor of $\frac{1}{m}$ and scaling ε by $1/(2p)$. □

Proof of Corollary G.2. Let $\widetilde{W}$ be the weights constructed in Lemma G.1 to approximate up to error $\varepsilon/2$. Let $w_i^* = \widetilde{w}_i + w_i^{(0)}$. Then

$$|g_r(x; \widetilde{W}) - g_r(x; W^*)| = \left|\sum_{i=1}^m a_{r,i}(\langle w_i^{(0)}, x\rangle + b_i^{(0)})\mathbb{I}[\langle w_i^{(0)}, x\rangle + b_i^{(0)} \geq 0]\right|. \tag{419}$$

By standard concentration, this is w.h.p. $\widetilde{O}(\varepsilon_a) = \varepsilon/2$. □

G.2 Coupling

Here we show that the weights after a properly bounded amount of updates stay close to the initialization, and thus the pseudo network is close to the real network using the same weights.

Lemma G.3 (Coupling). *For every unit vector x, w.h.p. over the random initialization, for every $t > 0$, we have the following. Denote $\tau = \eta t$.*

1. *For at most* $\widetilde{O}(\tau\sqrt{m})$ *fraction of* $i \in [m]$*:*

$$\mathbb{I}[\langle w_i^{(0)}, x\rangle + b_i^{(0)} \geq 0] \neq \mathbb{I}[\langle w_i^{(t)}, x\rangle + b_i^{(0)} \geq 0].$$

2. *For every* $r \in [k]$*,*

$$\left| f_r(x; W^{(0)} + W_t) - g_r(x; W^{(0)} + W_t) \right| = \widetilde{O}(\sqrt{k}\tau^2 m^{3/2}).$$

3. *For every* y*:*

$$\left\| \nabla L(F(x; W^{(0)} + W_t), y) - \nabla L(G(x; W^{(0)} + W_t), y) \right\|_{2,2} \leq \widetilde{O}(k\tau m^{3/2} + k^{3/2}\tau^2 m^{5/2}).$$

Proof of Lemma G.3. (1) W.h.p. over the random initialization, there is $B = \widetilde{O}(1)$ so that every $|a_{r,i}^{(0)}| \leq B$ and $|b_i^{(0)}| \leq B$. Thus, for every $i \in [m]$ and every $t \geq 0$,

$$\left\| \frac{\partial L(F(x; W^{(0)} + W_t), y)}{\partial w}_i \right\|_2 \leq \sqrt{k}B$$

which implies that $\left\| w_i^{(t)} - w_i^{(0)} \right\|_2 \leq \sqrt{k}B\eta t = \sqrt{k}B\tau$.

Now, consider the set $\mathcal{H}$ such that

$$\mathcal{H} = \left\{ i \in [m] \;\middle|\; \left| \left\langle w_i^{(0)}, x \right\rangle + b_i^{(0)} \right| \geq 2\sqrt{k}B\tau \right\}.$$

Note that for every $i \in \mathcal{H}$,

$$\left| \left(\left\langle w_i^{(t)}, x \right\rangle + b_i^{(0)} \right) - \left(\left\langle w_i^{(0)}, x \right\rangle + b_i^{(0)} \right) \right| \leq \sqrt{k}B\tau$$

which implies $\mathbb{I}[\langle w_i^{(0)}, x\rangle + b_i^{(0)} \geq 0] = \mathbb{I}[\langle w_i^{(t)}, x\rangle + b_i^{(0)} \geq 0]$.

Now, we need to bound the size of $\mathcal{H}$. Since $\left\langle w_i^{(0)}, x \right\rangle \sim \mathcal{N}(0, 1/m)$, and $b_i^{(0)} \sim \mathcal{N}(0, 1/m)$ with probability 1/2 and $b_i^{(0)} = 0$ otherwise, by standard property of Gaussian we directly have that for $|\mathcal{H}| \geq 1 - 2\sqrt{k}B\eta t\sqrt{m} = 1 - \widetilde{O}(\tau\sqrt{m})$.

(2) f_r and g_r only differ on the nodes $i \notin \mathcal{H}$. For such an i, $\langle w_i^{(t)}, x\rangle + b_i^{(0)}$ has a sign different from $\langle w_i^{(0)}, x\rangle + b_i^{(0)}$, and their difference is at most $\sqrt{k}B\tau$, so it contributes at most $\sqrt{k}B^2\tau$ difference between f_r and g_r. Then the bound follows from that there are only $2\tau m\sqrt{m}$ many $i \notin \mathcal{H}$.

(3) If $\mathbb{I}[\langle w_i^{(0)}, x\rangle + b_i^{(0)} \geq 0] = \mathbb{I}[\langle w_i^{(t)}, x\rangle + b_i^{(0)} \geq 0]$, then $\frac{\partial F(x; W^{(0)} + W_t)}{\partial w_i} = \frac{\partial G(x; W^{(0)} + W_t)}{\partial w_i}$, and thus the difference is only caused by $L'(\cdot, y)$. By the Lipschitz smoothness assumption on L and the statement (2), each such i contributes at most $k^{3/2}B^2\tau^2 m\sqrt{m}$. Otherwise, it contributes at most kB, and there are $2\tau m^{3/2}$ so many such i's, leading to the bound. $\square$

G.3 Optimization

For the set of samples $\mathcal{Z}$, let

$$L_F(W, \mathcal{Z}) := \frac{1}{|\mathcal{Z}|} \sum_{(x,y) \in \mathcal{Z}} L(F(x, W + W^{(0)}), y), \tag{420}$$

$$L_G(W, \mathcal{Z}) := \frac{1}{|\mathcal{Z}|} \sum_{(x,y) \in \mathcal{Z}} L(G(x, W + W^{(0)}), y). \tag{421}$$

We show the following lemma:

Lemma G.4. *If $m \geq M$, then w.h.p. there exists $T = \mathbf{poly}(\mathfrak{C}(\phi, \sqrt{\log(p/\varepsilon)}), k, p, 1/\varepsilon, \log m)$ and some $t \leq T$ such that*

$$L_F(W_t, \mathcal{Z}) = \varepsilon_0 + \varepsilon.$$

Proof of Lemma G.4. Let $\widetilde{W}$ be as in Corollary G.2. Since $L_G(\cdot, y)$ is convex, we have

$$\begin{aligned}
&L_G(W_t, \mathcal{Z}) - L_G(\widetilde{W}, \mathcal{Z}) && (422)\\
&\leq \langle \nabla L_G(W_t, \mathcal{Z}), W_t - \widetilde{W} \rangle && (423)\\
&\leq \|\nabla L_G(W_t, \mathcal{Z}) - \nabla L_F(W_t, \mathcal{Z})\|_{2,1} \|W_t - \widetilde{W}\|_{2,\infty} && (424)\\
&\quad + \langle \nabla L_F(W_t, \mathcal{Z}), W_t - \widetilde{W} \rangle. && (425)
\end{aligned}$$

We also have

$$\begin{aligned}
\|W_{t+1} - \widetilde{W}\|_F^2 &= \|W_t - \eta \nabla L_F(W_t, z^{(t)}) - \widetilde{W}\|_F^2 && (426)\\
&= \|W_t - \widetilde{W}\|_F^2 - 2\eta \langle \nabla L_F(W_t, z^{(t)}), W_t - \widetilde{W} \rangle + \eta^2 \|\nabla L_F(W_t, z^{(t)})\|_F^2, && (427)
\end{aligned}$$

so

$$\begin{aligned}
L_G(W_t, \mathcal{Z}) - L_G(\widetilde{W}, \mathcal{Z}) \leq & \|\nabla L_G(W_t, \mathcal{Z}) - \nabla L_F(W_t, \mathcal{Z})\|_{2,1} \|W_t - \widetilde{W}\|_{2,\infty} && (428)\\
& + \frac{\|W_t - \widetilde{W}\|_F^2 - \|W_{t+1} - \widetilde{W}\|_F^2}{2\eta} && (429)\\
& + \frac{\eta}{2} \|\nabla L_F(W_{t+1}, z^{(t)})\|_F^2. && (430)
\end{aligned}$$

W.h.p. over the initialization, every $|a_{r,i}^{(0)}| \leq \widetilde{O}(1)$ and $|b_i^{(0)}| \leq \widetilde{O}(1)$, so $\|W_t - \widetilde{W}\|_{2,\infty} = \widetilde{O}(\sqrt{k}\eta t + \frac{C_0}{m})$ and $\|\nabla L_F(W_{t+1}, z^{(t)})\|_F^2 = \widetilde{O}(m)$, where C_0 is as in Lemma G.1. By Lemma G.3, w.h.p. we know

$$\|\nabla L_G(W_t, \mathcal{Z}) - \nabla L_F(W_t, \mathcal{Z})\|_{2,1} \leq \Delta = \widetilde{O}(k \eta t m^{3/2} + k^{3/2} (\eta t)^2 m^{5/2}).$$

Therefore, summing up Eq (430) from $t = 0$ to $T - 1$ we have that

$$\sum_{t=0}^{T-1} L_G(W_t, \mathcal{Z}) - L_G(\widetilde{W}, \mathcal{Z}) \leq \widetilde{O}\left(\eta T^2 \Delta + \frac{C_0}{m} T \Delta\right) + \frac{\|W_0 - \widetilde{W}\|_F^2}{2\eta} + \widetilde{O}(\eta T m). \tag{431}$$

Note that $\|W_0 - \widetilde{W}\|_F^2 = \|\widetilde{W}\|_F^2 = \widetilde{O}(C_0^2/m)$. Also, by choosing $\eta = \widetilde{O}\left(\varepsilon/m\right)$ and $T = \widetilde{O}\left((C')^2/\varepsilon^2\right)$ for some $C' = \mathbf{poly}(\mathfrak{C}(\phi, \sqrt{\log(p/\varepsilon)}), k, p, 1/\varepsilon)$, we have $\Delta = \widetilde{O}(k^{3/2} C'^4 m^{1/2}/\varepsilon^2)$. Thus when m is large enough we have:

$$\frac{1}{T} \sum_{t=0}^{T-1} L_G(W_t, \mathcal{Z}) - L_G(\widetilde{W}, \mathcal{Z}) \leq O(\varepsilon). \tag{432}$$

By the coupling in Lemma G.3, we know that $L_F(W_t, \mathcal{Z})$ is close to $L_G(W_t, \mathcal{Z})$, leading to the lemma. □

G.4 Generalization

The generalization can be bounded via known Rademacher complexity results. Below is an application of Theorem 1 in [NTS15] to our setting.

Lemma G.5 (Two layer network Rademacher complexity)**.** *For every $\tau_w \geq 0$, w.h.p. for every*

$r \in [k]$ *and every* $N \geq 1$*, we have the empirical Rademacher complexity bounded by*

$$\frac{1}{N}\mathbb{E}_{\xi \in \{\pm 1\}^N}\left[\sup_{\|W'\|_{2,\infty} \leq \tau_w} \sum_{i \in [N]} \xi_i f_r(x_i; W^{(0)} + W')\right] \leq \widetilde{O}\left(\frac{m\tau_w}{\sqrt{N}}\right). \tag{433}$$

G.5 Proof of the Main Theorem

Proof of Theorem 4.1. First, we can apply Lemma G.4 to bound the training loss. Since w.h.p. over the initialization, we have $\|\nabla_{w_i}\|_2 \leq \widetilde{O}(1)$, we know that for properly chosen polynomial $C' = \mathbf{poly}(\mathfrak{C}(\phi, \sqrt{\log(p/\varepsilon)}), k, p, 1/\varepsilon)$, we have that $\|W_t\|_{2,\infty} \leq \widetilde{O}(\sqrt{k}\eta T) \leq \frac{C'}{m}$. By the Rademacher complexity bound in Lemma G.5, we can complete the proof with large enough m. $\square$

References

[ADLS17] Jayadev Acharya, Ilias Diakonikolas, Jerry Li, and Ludwig Schmidt. Sample-optimal density estimation in nearly-linear time. In *Proceedings of the Twenty-Eighth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 1278–1289. SIAM, 2017.

[AGNZ18] Sanjeev Arora, Rong Ge, Behnam Neyshabur, and Yi Zhang. Stronger generalization bounds for deep nets via a compression approach. *arXiv preprint arXiv:1802.05296*, 2018.

[APVZ14] Alexandr Andoni, Rina Panigrahy, Gregory Valiant, and Li Zhang. Learning polynomials with neural networks. In *International Conference on Machine Learning*, pages 1908–1916, 2014.

[BFT17] Peter L Bartlett, Dylan J Foster, and Matus J Telgarsky. Spectrally-normalized margin bounds for neural networks. In *Advances in Neural Information Processing Systems*, pages 6241–6250, 2017.

[BG17] Alon Brutzkus and Amir Globerson. Globally optimal gradient descent for a convnet with gaussian inputs. *arXiv preprint arXiv:1702.07966*, 2017.

[BGMS17] Alon Brutzkus, Amir Globerson, Eran Malach, and Shai Shalev-Shwartz. Sgd learns over-parameterized networks that provably generalize on linearly separable data. *arXiv preprint arXiv:1710.10174*, 2017.

[BL17] Digvijay Boob and Guanghui Lan. Theoretical properties of the global optimizer of two layer neural network. *arXiv preprint arXiv:1710.11241*, 2017.

[Dan17] Amit Daniely. Sgd learns the conjugate kernel class of the network. In *Advances in Neural Information Processing Systems*, pages 2422–2430, 2017.

[DZPS18] Simon S Du, Xiyu Zhai, Barnabas Poczos, and Aarti Singh. Gradient descent provably optimizes over-parameterized neural networks. *arXiv preprint arXiv:1810.02054*, 2018.

[EMZ18] Ronen Eldan, Dan Mikulincer, and Alex Zhai. The clt in high dimensions: quantitative bounds via martingale embedding. *arXiv preprint arXiv:1806.09087*, 2018.

[GHJY15] Rong Ge, Furong Huang, Chi Jin, and Yang Yuan. Escaping from saddle pointsonline stochastic gradient for tensor decomposition. In *Conference on Learning Theory*, pages 797–842, 2015.

[GLM17] Rong Ge, Jason D Lee, and Tengyu Ma. Learning one-hidden-layer neural networks with landscape design. *arXiv preprint arXiv:1711.00501*, 2017.

[GMH13] Alex Graves, Abdel-rahman Mohamed, and Geoffrey Hinton. Speech recognition with deep recurrent neural networks. In *Acoustics, speech and signal processing (icassp), 2013 ieee international conference on*, pages 6645–6649. IEEE, 2013.

[GRS18] Noah Golowich, Alexander Rakhlin, and Ohad Shamir. Size-independent sample complexity of neural networks. In *Proceedings of the Conference on Learning Theory*, 2018.

[HZRS16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[Kaw16] Kenji Kawaguchi. Deep learning without poor local minima. In *Advances in Neural Information Processing Systems*, pages 586–594, 2016.

[KLY18] Robert Kleinberg, Yuanzhi Li, and Yang Yuan. An alternative view: When does sgd escape local minima? *arXiv preprint arXiv:1802.06175*, 2018.

[KS04] Adam R Klivans and Rocco A Servedio. Learning dnf in time $\widetilde{O}(n^{1/3})$. *Journal of Computer and System Sciences*, 68(2):303–318, 2004.

[KSH12] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[LL18] Yuanzhi Li and Yingyu Liang. Learning overparameterized neural networks via stochastic gradient descent on structured data. In *Advances in Neural Information Processing Systems*, 2018.

[LSS14] Roi Livni, Shai Shalev-Shwartz, and Ohad Shamir. On the computational efficiency of training neural networks. In *Advances in Neural Information Processing Systems*, pages 855–863, 2014.

[LY17] Yuanzhi Li and Yang Yuan. Convergence analysis of two-layer neural networks with relu activation. In *Advances in Neural Information Processing Systems*, pages 597–607, 2017.

[Mar15] Martin J. Wainwright. Basic tail and concentration bounds. `https://www.stat.berkeley.edu/~mjwain/stat210b/Chap2_TailBounds_Jan22_2015.pdf`, 2015. Online; accessed Oct 2018.

[NBMS17] Behnam Neyshabur, Srinadh Bhojanapalli, David McAllester, and Nathan Srebro. A pac-bayesian approach to spectrally-normalized margin bounds for neural networks. *arXiv preprint arXiv:1707.09564*, 2017.

[NTS15] Behnam Neyshabur, Ryota Tomioka, and Nathan Srebro. Norm-based capacity control in neural networks. In *Conference on Learning Theory*, pages 1376–1401, 2015.

[SC16] Daniel Soudry and Yair Carmon. No bad local minima: Data independent training error guarantees for multilayer neural networks. *arXiv preprint arXiv:1605.08361*, 2016.

[Sha18] Ohad Shamir. Distribution-specific hardness of learning neural networks. *Journal of Machine Learning Research*, 19(32), 2018.

[SHK+14] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958, 2014.

[SHM+16] David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. Mastering the game of go with deep neural networks and tree search. *nature*, 529(7587):484, 2016.

[SJL17] Mahdi Soltanolkotabi, Adel Javanmard, and Jason D Lee. Theoretical insights into the optimization landscape of over-parameterized shallow neural networks. *arXiv preprint arXiv:1707.04926*, 2017.

[ST01] Daniel Spielman and Shang-Hua Teng. Smoothed analysis of algorithms: Why the simplex algorithm usually takes polynomial time. In *Proceedings of the thirty-third annual ACM symposium on Theory of computing*, pages 296–305. ACM, 2001.

[Tia17] Yuandong Tian. An analytical formula of population gradient for two-layered relu network and its applications in convergence and critical point analysis. *arXiv preprint arXiv:1703.00560*, 2017.

[Val84] Leslie Valiant. A theory of the learnable. *Communications of the ACM*, 1984.

[XLS16] Bo Xie, Yingyu Liang, and Le Song. Diversity leads to generalization in neural networks. *arXiv preprint Arxiv:1611.03131*, 2016.

[ZBH+17] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. In *ICLR*, 2017. arXiv 1611.03530.

[ZLJ16] Yuchen Zhang, Jason D Lee, and Michael I Jordan. l1-regularized neural networks are improperly learnable in polynomial time. In *International Conference on Machine Learning*, pages 993–1001, 2016.

[ZLWJ17] Yuchen Zhang, Jason Lee, Martin Wainwright, and Michael Jordan. On the learnability of fully-connected neural networks. In *Artificial Intelligence and Statistics*, pages 83–91, 2017.

[ZSJ+17] Kai Zhong, Zhao Song, Prateek Jain, Peter L Bartlett, and Inderjit S Dhillon. Recovery guarantees for one-hidden-layer neural networks. *arXiv preprint arXiv:1706.03175*, 2017.