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 12, 2018

(version 5)*

Abstract

The fundamental learning theory behind neural networks remains largely open. What classes of functions can neural networks actually learn? Why doesn't the trained neural networks overfit when the it is overparameterized (namely, having more parameters than statistically needed to overfit training data)?

In this work, we prove that overparameterized neural networks can learn some notable concept classes, including two and three-layer networks with fewer parameters and smooth activations. Moreover, the learning can be simply done by SGD (stochastic gradient descent) or its variants in polynomial time using polynomially many samples. The sample complexity can also be almost independent of the number of parameters in the overparameterized network.

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 [26, 27, 31, 42]. In contrast to the widely accepted empirical success, much less theory is known. Despite a recent boost of theoretical studies, many questions remain largely open, including fundamental ones about the optimization and generalization in learning neural networks.

One key challenge in analyzing neural networks is that the corresponding optimization is non-convex and is theoretically hard in the general case [41, 56]. 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 even on both training and test data. Therefore,

what functions can neural networks provably learn?

Another key challenge is that, in practice, neural networks are heavily overparameterized (e.g., [54]): the number of learnable parameters is much larger than the number of the training samples. It is observed that overparameterization empirically improves both optimization and

[*] V1 appears on this date, V2/V3/V4/V5 polish writing and parameters, V5 adds experiments. Authors sorted in alphabetical order. We would like to thank Greg Yang and Sebastien Bubeck for many enlightening conversations.

generalization, appearing to contradict traditional learning theory.[1] Therefore,

why do overparameterized networks (found by those training algorithms) generalize?

1.1 What Can Neural Networks Provably Learn?

Most existing works analyzing the learnability of neural networks [9, 12, 13, 21, 22, 29, 34, 35, 43, 44, 48, 50, 51, 57] make unrealistic assumptions about the data distribution (such as being random Gaussian), and/or make strong assumptions about the network (such as using linear activations). Li and Liang [33] show that two-layer ReLU networks can learn classification tasks when the data come from mixtures of arbitrary but well-separated distributions.

A theorem *without distributional assumptions* on data is often more desirable. Indeed, how to obtain a result that does not depend on the data distribution, but only on the concept class itself, lies in the center of PAC-learning which is one of the foundations of machine learning theory [49]. Also, studying *non-linear activations* is critical because otherwise one can only learn linear functions, which can also be easily learned via linear models without neural networks.

Brutzkus et al. [14] prove that two-layer networks with ReLU activations can learn linearly-separable data (and thus the class of linear functions) using just SGD. This is an (improper) *PAC-learning* type of result because it makes no assumption on the data distribution. Andoni et al. [5] proves that two-layer networks can learn polynomial functions of degree r over d-dimensional inputs in sample complexity $O(d^r)$. Their learner networks use exponential activation functions, where in practice the rectified linear unit (ReLU) activation has been the dominant choice across vastly different domains.

On a separate note, if one treats all but the last layer of neural networks as generating a random feature map (and these hidden layers' movements are negligible), then training only the last layer is a convex task, so one can learn the class of linear functions in this implicit feature space [16, 17]. This result implies low-degree polynomials and compositional kernels can be learned by neural networks in polynomial time. Empirically, training last layer greatly weakens the power of neural networks (see Figure 1).

Our Result. We prove that an important concept class that contains three-layer (resp. two-layer) neural networks equipped with smooth activations can be efficiently learned by three-layer (resp. two-layer) ReLU neural networks via SGD or its variants.

Specifically, suppose in aforementioned class the best network (called the target function or target network) achieves a population risk OPT with respect to some convex loss function. We show that one can learn up to population risk $\mathsf{OPT} + \varepsilon$, using three-layer (resp. two-layer) ReLU networks of size greater than a fixed *polynomial* in the size of the target network, in $1/\varepsilon$, and in the "complexity" of the activation function used in the target network. Furthermore, the sample complexity is also *polynomial* in these parameters, and only poly-logarithmic in the size of the learner ReLU network.

We stress here that this is *agnostic* PAC-learning because we allow the target function to have error (e.g., OPT can be positive for regression), and is *improper* learning because the concept class consists of smaller neural networks comparing to the networks being trained.

Our Contributions. We believe our result gives further insights to the fundamental questions about the learning theory of neural networks.

[1] For example, Livni et al. [37] observed that on synthetic data generated from a target network, SGD converges faster when the learned network has more parameters than the target. Perhaps more interestingly, Arora et al. [6] found that overparameterized networks learned 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. We also have experiments in Figure 1(a).

- To the best of our knowledge, this is the first result showing that using *hidden layers* of neural networks one can provably learn the concept class containing two (or even three) layer neural networks with non-trivial activation functions.[2] In contrast, Daniely [16] focuses on training essentially only the last layer (and the hidden-layer movement is negligible).
- Unlike the two-layer case where the optimization landscape with overparameterization is almost convex [18, 33], our three-layer result gives the first theoretical proof that learning neural networks, even with sophisticated *non-convex* interactions across layers, can still be plausible.
- To some extent we explain the reason why overparameterization improves testing accuracy: with larger overparameterization, one can hope to learn better target functions with possibly larger size, more complex activations, smaller risk OPT, and to a smaller error ε.
- We establish new tools to tackle the learning process of neural networks in general, which can be useful for studying other network architectures and learning tasks. (E.g., the new tools here have allowed researchers to study also the learning of recurrent neural networks [2].)

Other Related Works. We acknowledge a different line of research using *kernels* as improper learners to learn the concept class of neural networks [23, 24, 37, 55]. This is very different from us because we use "neural networks" as learners. In other words, we study the question of "what can neural networks learn" but they study "what alternative methods can replace neural networks." There is also a line of work studying interesting relationship between neural networks and neural tangent kernels [7, 28, 32, 52]. However, in practice, there is big difference between the power of kernel methods and that of neural networks. For instance, ResNet achieves 96% test error on the CIFAR-10 data set but NTK achieves 81% [7]. We also illustrate this in Figure 1.

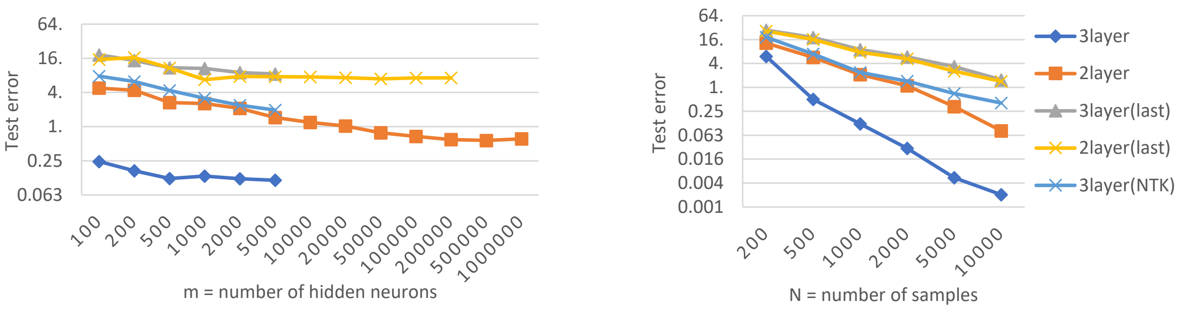
With sufficient overparameterization, for two-layer neural networks Li and Liang [33] and Du et al. [18] show that gradient descent can perfectly fit the training samples when the data is not degenerated. Allen-Zhu et al. [4] show that deep neural networks can also provably fit the training samples in polynomial time. These results do not cover generalization so are not relevant in studying the learnability of neural networks.

1.2 Why Do Overparameterized Networks Generalize?

Our result above assumes that the learner network is sufficiently overparameterized. So, why does it generalize to the population risk and give small test error? More importantly, why does it generalize with 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 the VC dimension grows with the number of parameters. Several works [6, 11, 25, 40] explain generalization by studying some other "complexity" of the learned networks. Most related to the discussion here is [11] where the authors prove a generalization bound in the norms (of weight matrices) of each layer, as opposed to the number of parameters. However, their norms are "sparsity induced norms": 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 those results how a network with such low "complexity" and a good training loss can be found by the training method.

[2] After this paper has appeared online, Arora et al. [8] showed that neural networks can provably learn two-layer networks with a slightly weaker class of smooth activation functions. Namely, the activation functions that are either linear functions or even functions.



(a) $N = 1000$ and vary m

(b) $m = 2000$ and vary N

Figure 1: Performance comparison. `3layer/2layer` stands for training (hidden weights) in three and two-layer neural networks. `(last)` stands for *conjugate kernel* [16], meaning training only the output layer. `(NTK)` stands for *neural tangent kernel* [28] with finite width. We also implemented other direct kernels such as [55] but they perform much worse.

Setup. We consider ℓ_2 regression task on synthetic data where feature vectors $x \in \mathbb{R}^4$ are generated as normalized random Gaussian, and label is generated by target function $F^*(x) = (\sin(3x_1) + \sin(3x_2) + \sin(3x_3) - 2)^2 \cdot \cos(7x_4)$. We use N training samples, and SGD with mini-batch size 50 and best tune learning rates and weight decay parameters. See Appendix 7 for our experiment setup, how we choose such target function, and more experiments.

At a high level, our generalization is made possible with the following sequence of conceptual steps.

- Good networks with small risks are *plentiful*: thanks to overparameterization, with high probability over random initialization, there exists a good network in the close neighborhood of *any* point on the SGD training trajectory. (This corresponds to Section 6.2 and 6.3.)
- The optimization in overparameterized neural networks has benign properties: essentially along the training trajectory, there is no second-order critical points for learning three-layer networks, and no first-order critical points for two-layer. (This corresponds to Section 6.4.)
- In the learned networks, information is also *evenly distributed* among neurons, by utilizing either implicit or explicit regularization. This structure allows a new generalization bound that is (almost) independent of the number of neurons. (This corresponds to Section 6.5 and 6.6, and we also empirically verify it in Section 7.1.)

Since practical neural networks are typically overparameterized, we genuinely hope that our results can provide theoretical insights to networks used in various applications.

Roadmap. We introduce notations in Section 2. In Section 3 and Section 4 we respectively present main results for two and three-layer networks, and discuss what our main contributions are. Our Section 6 summarizes main proof ideas for our three-layer network results, and details are in Appendix C. Our two-layer proof is much easier and in Appendix B.

2 Notations

We use $x = y \pm z$ to denote that $x \in [y - z, y + z]$. We use $\mathcal{N}(\mu, \sigma)$ to denote Gaussian distribution with mean μ and variance σ; or $\mathcal{N}(\mu, \Sigma)$ to denote Gaussian vector with mean μ and covariance Σ. We use $\mathbb{I}_E$ or $\mathbb{I}[E]$ to denote the indicator function of the event E. We use $\sigma(\cdot)$ to denote the

ReLU function $\sigma(x) = \max\{x, 0\}$. Given function $f\colon \mathbb{R} \to \mathbb{R}$, we use $f(x) = (f(x_1), \dots, f(x_m))$ to denote the same function over vectors.

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 also abbreviate $\|w\| = \|w\|_2$ when it is clear from the context. We denote the row ℓ_p norm for $W \in \mathbb{R}^{m\times d}$ (for $p \geq 1$) as

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

By definition, $\|W\|_{2,2} = \|W\|_F$ is the Frobenius norm of W. We use $\|W\|_2$ to denote the matrix spectral norm. 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.

We say $f\colon \mathbb{R}^d \to \mathbb{R}$ is L-Lipschitz continuous if $|f(x) - f(y)| \leq L\|x - y\|_2$; is L-Lipschitz smooth if $\|\nabla f(x) - \nabla f(y)\|_2 \leq L\|x-y\|_2$; and is L-second-order smooth if $\|\nabla^2 f(x) - \nabla^2 f(y)\|_2 \leq L\|x-y\|_2$.

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 where m is the number of hidden neurons, and $1 - e^{-c\log^2(m_1 m_2)}$ for three-layer network where m_1, m_2 are the number of hidden neurons for the first and second layer (to be precisely defined later). In this paper, $\widetilde{O}$ hides factors of $\mathsf{polylog}(m)$ for two-layer networks, or $\mathsf{polylog}(m_1, m_2)$ for three-layer networks.

Function complexity. The following notion measures the complexity of any smooth activation function $\phi(z)$. Suppose $\phi(z) = \sum_{i=0}^{\infty} c_i z^i$. Given a non-negative R, the complexity

$$\mathfrak{C}_\varepsilon(\phi, R) \stackrel{\text{def}}{=} \sum_{i=0}^{\infty} \Big((C^* R)^i + \big(\frac{\sqrt{\log(1/\varepsilon)}}{\sqrt{i}} C^* R\big)^i\Big)|c_i|, \quad \mathfrak{C}_\mathfrak{s}(\phi, R) \stackrel{\text{def}}{=} C^* \sum_{i=0}^{\infty} (i+1)^{1.75} R^i |c_i|$$

where C^* is a sufficiently large constant (e.g., 10^4). Intuitively, $\mathfrak{C}_\mathfrak{s}$ measures the *sample* complexity: how many samples are required to learn ϕ correctly; while $\mathfrak{C}_\varepsilon$ bounds the *network* size: how much over-parameterization is needed for the algorithm to efficiently learn ϕ up to ε error. It is always true that $\mathfrak{C}_\mathfrak{s}(\phi, R) \leq \mathfrak{C}_\varepsilon(\phi, R) \leq \mathfrak{C}_\mathfrak{s}(\phi, O(R)) \times \mathsf{poly}(1/\varepsilon)$.[3] While for $\sin z, \exp(z)$ or low degree polynomials, $\mathfrak{C}_\mathfrak{s}(\phi, O(R))$ and $\mathfrak{C}_\varepsilon(\phi, R)$ only differ by $o(1/\varepsilon)$.

Example 2.1. If $\phi(z) = e^{c\cdot z} - 1$, $\phi(z) = \sin(c\cdot z)$, $\phi(z) = \cos(c \cdot z)$ for constant c or $\phi(z)$ is low degree polynomial, then $\mathfrak{C}_\varepsilon(\phi, 1) = o(1/\varepsilon)$ and $\mathfrak{C}_\mathfrak{s}(\phi, 1) = O(1)$. If $\phi(z) = \mathrm{sigmoid}(z)$ or $\tanh(z)$, we can truncate their Taylor series at degree $\Theta(\log \frac{1}{\varepsilon})$ to get ε approximation. One can verify this gives $\mathfrak{C}_\varepsilon(\phi, 1) \leq \mathsf{poly}(1/\varepsilon)$ and $\mathfrak{C}_\mathfrak{s}(\phi, 1) \leq O(1)$.

3 Result for Two-Layer Networks

We consider learning some unknown distribution $\mathcal{D}$ of data points $z = (x, y) \in \mathbb{R}^d \times \mathcal{Y}$, where x is the input point and y is the label. Without loss of generality, assume $\|x\|_2 = 1$ and $x_d = \frac{1}{2}$.[4] Consider a loss function $L\colon \mathbb{R}^k \times \mathbb{R} \to \mathcal{Y}$ such that for every $y \in \mathcal{Y}$, the function $L(\cdot, y)$ is non-negative, convex, 1-Lipschitz continuous and 1-Lipschitz smooth and $L(0, y) \in [0, 1]$. This includes both the cross-entropy loss and the ℓ_2-regression loss (for bounded $\mathcal{Y}$). [5]

[3] Recall $\big(\frac{\sqrt{\log(1/\varepsilon)}}{\sqrt{i}} C^*\big)^i \leq e^{O(\log(1/\varepsilon))} = \frac{1}{\mathsf{poly}(\varepsilon)}$ for every $i \geq 1$.

[4] $\frac{1}{2}$ can always be padded to the last coordinate, and $\|x\|_2 = 1$ can always be ensured from $\|x\|_2 \leq 1$ by padding $\sqrt{1 - \|x\|_2^2}$. This assumption is for simplifying the presentation. For instance, $x_d = \frac{1}{2}$ allows us to focus only on target networks without bias.

[5] In fact, the non-negativity assumption and the 1-Lipschitz smoothness assumption are not needed for our two-layer result, but we state all of them here for consistency.

Concept class and target function $\boldsymbol{F^*(x)}$. Consider target functions $F^* \colon \mathbb{R}^d \to \mathbb{R}^k$ of

$$F^* = (f_1^*, \dots, f_k^*) \quad \text{and} \quad f_r^*(x) = \sum_{i=1}^{p} a_{r,i}^* \phi_i(\langle w_{1,i}^*, x\rangle)\langle w_{2,i}^*, x\rangle \tag{3.1}$$

where each $\phi_i \colon \mathbb{R} \to \mathbb{R}$ is infinite-order smooth and the weights $w_{1,i}^* \in \mathbb{R}^d, w_{2,i}^* \in \mathbb{R}^d$ and $a_{r,i}^* \in \mathbb{R}$. We assume for simplicity $\|w_{1,i}^*\|_2 = \|w_{2,i}^*\|_2 = 1$ and $|a_{r,i}^*| \leq 1$.[6] We denote by

$$\mathfrak{C}_\varepsilon(\phi, R) := \max_{j\in[p]}\{\mathfrak{C}_\varepsilon(\phi_j, R)\} \quad \text{and} \quad \mathfrak{C}_\mathfrak{s}(\phi, R) := \max_{j\in[p]}\{\mathfrak{C}_\mathfrak{s}(\phi_j, R)\}$$

the complexity of F^* and assume they are bounded.

In the *agnostic PAC-learning language*, our concept class consists of all functions F^* in the form of (3.1) with complexity bounded by threshold C and parameter p bounded by threshold p_0. Let $\mathsf{OPT} = \mathbb{E}[L(F^*(x), y)]$ be the population risk achieved by the *best* target function in this concept class. Then, our goal is to learn this concept class with population risk $\mathsf{OPT} + \varepsilon$ using sample and time complexity *polynomial* in C, p_0 and $1/\varepsilon$. In the remainder of this paper, to simplify notations, we do not explicitly define this concept class parameterized by C and p. Instead, we equivalently state our theorem with respect to any (unknown) target function F^* with specific parameters C and p satisfying $\mathsf{OPT} = \mathbb{E}[L(F^*(x), y)]$. We assume $\mathsf{OPT} \in [0, 1]$ for simplicity.

Remark. Standard two-layer networks $f_r^*(x) = \sum_{i=1}^p a_{r,i}^* \phi(\langle w_{1,i}^*, x\rangle)$ are special cases of (3.1) (by setting $w_{2,i}^* = (0, \dots, 0, 1)$ and $\phi_i = \phi$). Our formulation (3.1) additionally captures combinations of correlations between *non-linear* and linear measurements of different directions of x.

Learner network $\boldsymbol{F(x; W)}$. Using a data set $\mathcal{Z} = \{z_1, \dots, z_N\}$ of N i.i.d. samples from $\mathcal{D}$, we train a network $F = (f_1, \cdots, f_k) \colon \mathbb{R}^d \to \mathbb{R}^k$ with

$$f_r(x) \stackrel{\text{def}}{=} \sum_{i=1}^{m} a_{r,i}\sigma(\langle w_i, x\rangle + b_i) = a_r^\top \sigma(Wx + b) \tag{3.2}$$

where σ is the ReLU activation, $W = (w_1, \dots, w_m) \in \mathbb{R}^{m\times d}$ is the hidden weight matrix, $b \in \mathbb{R}^m$ is the bias vector, and $a_r \in \mathbb{R}^m$ is the output weight vector. To simplify analysis, we only update W and keep b and a_r at initialization values. For such reason, we write the learner network as $f_r(x; W)$ and $F(x; W)$. We sometimes use $b^{(0)} = b$ and $a_r^{(0)} = a_r$ to emphasize they are randomly initialized. Our goal is to learn a weight matrix W with population risk $\mathbb{E}\big[L(F(x; W), y)\big] \leq \mathsf{OPT} + \varepsilon$.

Learning Process. Let $W^{(0)}$ denote the initial value of the hidden weight matrix, and let $W^{(0)} + W_t$ denote the value at time t. (Note that W_t is the matrix of *increments*.) The weights are initialized with Gaussians and then W is updated by the vanilla SGD. More precisely,

- entries of $W^{(0)}$ and $b^{(0)}$ are i.i.d. random Gaussians from $\mathcal{N}(0, 1/m)$,
- entries of each $a_r^{(0)}$ are i.i.d. random Gaussians from $\mathcal{N}(0, \varepsilon_a^2)$ for some fixed $\varepsilon_a \in (0, 1]$.[7]

At time t, SGD samples $z = (x, y) \sim \mathcal{Z}$ and updates $W_{t+1} = W_t - \eta\nabla L(F(x; W^{(0)} + W_t), y)$. [8] The

[6] For general $\|w_{1,i}^*\|_2 \leq B, \|w_{2,i}^*\|_2 \leq B$, $|a_{r,i}^*| \leq B$, the scaling factor B can be absorbed into the activation function $\phi'(x) = \phi(Bx)$. Our results then hold by replacing the complexity of ϕ with ϕ'. As for how much this affects the complexity (i.e., polynomial in B or exponential in B), it varies across different activation functions. The complexity sometimes stays the same when B is a constant, see Example 2.1. In general, it is impossible to obtain a polynomial dependency on B unless well-known hardness assumptions break. For example, learning a single sigmoid with polynomially large weights agnostically is equivalent to agnostically learning a half space to a polynomially small error, which has no polynomial time algorithms under the random K-XOR assumption [15].

[7] We shall choose $\varepsilon_a = \widetilde{\Theta}(\varepsilon)$ in the proof due to technical reason. As we shall see in the three-layer case, if weight decay is used, one can relax this to $\varepsilon_a = 1$.

[8] Strictly speaking, the objective does not have gradient everywhere due to the non-differentiability of ReLU. Throughout the paper, we use ∇ to denote the value computed by setting $\nabla\sigma(x) = \mathbb{I}[x \geq 0]$, which is what is used in practical auto-differentiation softwares.

Algorithm 1 SGD for two-layer networks

```
Input: Data set $\mathcal{Z}$, initialization $W^{(0)}$, step size $\eta$
$W_0 = 0$
for $t = 1, 2, \dots$ do
    Randomly sample $z = (x, y)$ from the data set $\mathcal{Z}$
    Update: $W_t = W_{t-1} - \eta \frac{\partial}{\partial W} L(F(x; W_{t-1} + W^{(0)}), y)$
end for
```

algorithm is in Algorithm 1.

3.1 Main Theorem

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, and $\widetilde{O}$ hides factors of $\mathsf{polylog}(m)$.

Theorem 1 (two-layer)**.** *For every $\varepsilon \in \big(0, \frac{1}{pk\mathfrak{C}_\mathfrak{s}(\phi,1)}\big)$, there exists*

$$M_0 = \mathsf{poly}(\mathfrak{C}_\varepsilon(\phi, 1), 1/\varepsilon) \quad \textit{and} \quad N_0 = \mathsf{poly}(\mathfrak{C}_\mathfrak{s}(\phi, 1), 1/\varepsilon)$$

such that for every $m \geq M_0$ and every $N \geq \widetilde{\Omega}(N_0)$, choosing $\varepsilon_a = \varepsilon / \widetilde{\Theta}(1)$ for the initialization, choosing learning rate $\eta = \widetilde{\Theta}\big(\frac{1}{\varepsilon k m}\big)$ and

$$T = \widetilde{\Theta}\left(\frac{(\mathfrak{C}_\mathfrak{s}(\phi, 1))^2 \cdot k^3 p^2}{\varepsilon^2}\right) \quad ,$$

with high probability over the random initialization, SGD after T iteration satisfies

$$\mathbb{E}_{sgd}\Big[\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}_{(x,y)\sim\mathcal{D}} L(F(x; W^{(0)} + W_t), y)\Big] \leq \mathsf{OPT} + \varepsilon.$$

(Above, $\mathbb{E}_{sgd}$ takes expectation with the randomness of the SGD.)

Remark. SGD only takes one example per iteration. Thus, sample complexity is also at most T.

Example 3.1. For functions such as $\phi(z) = e^z, \sin z, \mathrm{sigmoid}(z), \tanh(z)$ or low degree polynomials, using Example 2.1, our theorem indicates that for target networks with such activation functions, we can learn them using two-layer ReLU networks with

$$\text{size } m = \frac{\mathsf{poly}(k, p)}{\mathsf{poly}(\varepsilon)} \text{ and sample complexity } \min\{N, T\} = \frac{\mathsf{poly}(k, p, \log m)}{\varepsilon^2}$$

We note sample complexity T is (almost) independent of m, the amount of overparametrization.

3.2 Our Interpretations

Overparameterization improves generalization. By increasing m, Theorem 1 supports more target functions with possibly larger size, more complex activations, and smaller population risk OPT. In other words, when m is fixed, among the class of target functions whose complexities are captured by m, SGD can learn the best function approximator of the data, with the smallest population risk. This gives intuition how overparameterization improves test error, see Figure 1(a).

Large margin non-linear classifier. Theorem 1 is a nonlinear analogue of the margin theory for linear classifiers. The target function with a small population risk (and of bounded norm) can be viewed as a “large margin non-linear classifier.” In this view, Theorem 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.

Inductive bias. Recent works (e.g., [4, 33]) show that when the network is heavily overparameterized (that is, m is polynomial in the number of training samples) and no two training samples are identical, then SGD can find a global optimum with 0 classification error (or find a solution with ε training loss) in polynomial time. This does not come with generalization, since it can even fit random labels. Our theorem, combined with [4], confirms the inductive bias of SGD for two-layer networks: when the labels are random, SGD finds a network that memorizes the training data; when the labels are (even only approximately) realizable by some target network, then SGD learns and generalizes. This gives an explanation towards the well-known empirical observations of such inductive bias (e.g., [54]) in the two-layer setting, and is more general than Brutzkus et al. [14] in which the target network is only linear.

4 Result for Three-Layer Networks

Concept class and target function $F^*(x)$. This time we consider more powerful target functions $F^* = (f_1^*, \cdots, f_k^*)$ of the form

$$f_r^*(x) \overset{\text{def}}{=} \sum_{i\in[p_1]} a_{r,i}^* \Phi_i \left(\sum_{j\in[p_2]} v_{1,i,j}^* \phi_{1,j}(\langle w_{1,j}^*, x\rangle) \right) \left(\sum_{j\in[p_2]} v_{2,i,j}^* \phi_{2,j}(\langle w_{2,j}^*, x\rangle) \right) \tag{4.1}$$

where each $\phi_{1,j}, \phi_{2,j}, \Phi_i \colon \mathbb{R} \to \mathbb{R}$ is infinite-order smooth, and the weights $w_{1,i}^*, w_{2,i}^* \in \mathbb{R}^d$, $v_{1,i}^*, v_{2,i}^* \in \mathbb{R}^{p_2}$ and $a_{r,i}^* \in \mathbb{R}$ satisfy $\|w_{1,j}^*\|_2 = \|w_{2,j}^*\|_2 = \|v_{1,i}^*\|_2 = \|v_{2,i}^*\|_2 = 1$ and $|a_{r,i}^*| \leq 1$. Let

$$\mathfrak{C}_\varepsilon(\phi, R) = \max_{j\in[p_2], s\in[1,2]}\{\mathfrak{C}_\varepsilon(\phi_{s,j}, R)\}, \qquad \mathfrak{C}_\varepsilon(\Phi, R) = \max_{j\in[p_1]}\{\mathfrak{C}_\varepsilon(\Phi_j, R)\}$$
$$\mathfrak{C}_\mathfrak{s}(\phi, R) = \max_{j\in[p_2], s\in[1,2]}\{\mathfrak{C}_\mathfrak{s}(\phi_{s,j}, R)\}, \qquad \mathfrak{C}_\mathfrak{s}(\Phi, R) = \max_{j\in[p_1]}\{\mathfrak{C}_\mathfrak{s}(\Phi_j, R)\}$$

to denote the complexity of the two layers, and assume they are bounded.

Our concept class contains measures of correlations between *composite non-linear* functions and *non-linear* functions of the input, there are plenty of functions in this new concept class that may not necessarily have small-complexity representation in the previous formulation (3.1), and as we shall see in Figure 1(a), this is the **critical advantage** of using three-layer networks compared to two-layer ones or their NTKs. The learnability of this correlation is due to the *non-convex* interactions between hidden layers. As a comparison, [16] studies the regime where the changes in hidden layers are negligible thus *can not* show how to learn this concept class with a three-layer network.

Remark 4.1. Standard three-layer networks

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

are only special cases of (4.1). Also, even in the special case of $\Phi_i(z) = z$, the target

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

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

$$f_r^*(x) = \textstyle\sum_{i\in[p_1]} a^*_{r,i}\Phi_i\left(\sum_{j\in[p_2]} v^*_{1,i,j}\phi_{1,j}(\langle w^*_{1,j}, x\rangle)\langle w^*_{3,j}, x\rangle\right)\left(\sum_{j\in[p_2]} v^*_{2,i,j}\phi_{2,j}(\langle w^*_{2,j}, x\rangle)\langle w^*_{4,j}, x\rangle\right)$$

with the mild requirement that $\langle w^*_{1,j}, w^*_{3,j}\rangle = 0$ for each $j \in [p_2]$. We choose to present the slightly weaker formulation (4.1) for cleaner proofs.

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

$$f_r(x) = \sum_{i\in[m_2]} a_{r,i}\sigma(n_i(x) + b_{2,i}) \quad \text{where each } n_i(x) = \sum_{j\in[m_1]} v_{i,j}\sigma\left(\langle w_j, x\rangle + b_{1,j}\right)$$

The first and second layers have m_1 and m_2 hidden neurons. Let $W \in \mathbb{R}^{m_1\times d}$ and $V \in \mathbb{R}^{m_2\times m_1}$ represent the weights of the first and second hidden layers respectively, and $b_1 \in \mathbb{R}^{m_1}$ and $b_2 \in \mathbb{R}^{m_2}$ represent the corresponding bias vectors, $a_r \in \mathbb{R}^{m_2}$ represent the output weight vector.

4.1 Learning Process

Again for simplicity, we only update W and V. The weights are randomly initialized as:

- entries of $W^{(0)}$ and $b_1 = b_1^{(0)}$ are i.i.d. from $\mathcal{N}(0, 1/m_1)$,
- entries of $V^{(0)}$ and $b_2 = b_2^{(0)}$ are i.i.d. from $\mathcal{N}(0, 1/m_2)$,
- entries of each $a_r = a_r^{(0)}$ are i.i.d. from $\mathcal{N}(0, \varepsilon_a^2)$ for $\varepsilon_a = 1$. [9]

As for the optimization algorithm, we use SGD with weight decay and an explicit regularizer.

For some $\lambda \in (0, 1]$, we will use $\lambda F(x; W, V)$ as the learner network, i.e., linearly scale F down by λ. This is equivalent to replacing W, V with $\sqrt{\lambda}W$, $\sqrt{\lambda}V$, since a ReLU network is positive homogenous. The SGD will start with $\lambda = 1$ and slowly decrease it, similar to weight decay.[10]

We also use an explicit regularizer for some $\lambda_w, \lambda_v > 0$ with[11]

$$R(\sqrt{\lambda}W, \sqrt{\lambda}V) \stackrel{\text{def}}{=} \lambda_v\|\sqrt{\lambda}V\|_F^2 + \lambda_w\|\sqrt{\lambda}W\|_{2,4}^4 \ .$$

Now, in each round $t = 1, 2, \dots, T$, we use (noisy) SGD to minimize the following stochastic objective for some fixed λ_{t-1}:

$$\begin{aligned} L_2(\lambda_{t-1}; W', V') \stackrel{\text{def}}{=} L\left(\lambda_{t-1}F\big(x; W^{(0)} + W^\rho + \mathbf{\Sigma}W', V^{(0)} + V^\rho + V'\mathbf{\Sigma}\big)\right) \\ + R(\sqrt{\lambda_{t-1}}W', \sqrt{\lambda_{t-1}}V') \end{aligned} \tag{4.2}$$

Above, the objective is stochastic because (1) $z \sim \mathcal{Z}$ is a random sample from the training set, (2) W^ρ and V^ρ are two small perturbation random matrices with entries i.i.d. drawn from $\mathcal{N}(0, \sigma_w^2)$ and $\mathcal{N}(0, \sigma_v^2)$ respectively, and (3) $\mathbf{\Sigma} \in \mathbb{R}^{m_1\times m_1}$ is a random diagonal matrix with diagonals i.i.d. uniformly drawn from $\{+1, -1\}$. We note that the use of W^ρ and V^ρ is standard for Gaussian smoothing on the objective (and not needed in practice).[12] The use of Σ may be reminiscent of the

[9] Recall in our two-layer result we have chosen $\varepsilon_a < 1$ due to technical reasons; thanks to weight decay, we can simply select $\varepsilon_a = 1$ in our three-layer case.

[10] We illustrate the technical necessity of adding weight decay. During training, it is *easy* to add new information to the current network, but *hard* to forget "false" information that is already in the network. Such false information can be accumulated from randomness of SGD, non-convex landscapes, and so on. Thus, by scaling down the network we can effectively forget false information.

[11] This $\|\cdot\|_{2,4}$ norm on W encourages weights to be more evenly distributed across neurons. It can be replaced with $\|\sqrt{\lambda_{t-1}}W_{t-1}\|_{2,2+\alpha}^{2+\alpha}$ for any constant $\alpha > 0$ for our theoretical purpose. We choose $\alpha = 2$ for simplicity, and observe that in practice, weights are automatically spread out due to data randomness, so this explicit regularization may not be needed. See Section 7.1 for an experiment.

[12] Similar to known non-convex literature [20] or smooth analysis, we introduce Gaussian perturbation W^ρ and V^ρ for theoretical purpose and it is not needed in practice. Also, we apply noisy SGD which is the vanilla SGD plus Gaussian perturbation, which again is needed in theory but believed unnecessary for practice [20].

Algorithm 2 SGD for three-layer networks (second variant (4.2))

```
Input: Data set $\mathcal{Z}$, initialization $W^{(0)}, V^{(0)}$, step size $\eta$, number of inner steps $T_w$, $\sigma_w, \sigma_v, \lambda_w, \lambda_v$.
$W_0 = 0, V_0 = 0, \lambda_1 = 1$, $T = \Theta\big(\eta^{-1} \log \frac{\log(m_1 m_2)}{\varepsilon_0}\big)$.
for $t = 1, 2, \dots, T$ do
    Apply noisy SGD with step size $\eta$ on the stochastic objective $L_2(\lambda_{t-1}; W, V)$ for $T_w$ steps; the starting
    point is $W = W_{t-1}, V = V_{t-1}$ and suppose it reaches $W_t, V_t$.                    ◇ see Lemma A.9
    $\lambda_{t+1} = (1 - \eta)\lambda_t$.                                                  ◇ weight decay
end for
Randomly sample $\widehat{\mathbf{\Sigma}}$ with diagonal entries i.i.d. uniform on $\{1, -1\}$
Randomly sample $\widetilde{\Theta}(1/\varepsilon_0^2)$ many noise matrices $\big\{W^{\rho,j}, V^{\rho,j}\big\}$. Let
    $$j^* = \arg\min_j \Big\{\mathbb{E}_{z \in \mathcal{Z}} L\Big(\lambda_T F\big(x; W^{(0)} + W^{\rho,j} + \widehat{\mathbf{\Sigma}} W_T, V^{(0)} + V^{\rho,j} + V_T \widehat{\mathbf{\Sigma}}\big)\Big)\Big\}$$
Output $W_T^{(out)} = W^{(0)} + W^{\rho,j^*} + \widehat{\mathbf{\Sigma}} W_T$, $V_T^{(out)} = V^{(0)} + V^{\rho,j^*} + V_T \widehat{\mathbf{\Sigma}}$.
```

Dropout technique [47] in practice which randomly masks out neurons, and can also be removed.[13]

Algorithm 2 presents the details. Specifically, in each round t, Algorithm 2 starts with weight matrices W_{t-1}, V_{t-1} and performs T_w iterations. In each iteration it goes in the negative direction of the gradient $\nabla_{W',V'} L_2(\lambda_t; W', V')$ (with respect to a stochastic choice of $z, W^\rho, V^\rho, \mathbf{\Sigma}$). Let the final matrices be W_t, V_t. At the end of this round t, Algorithm 2 performs weight decay by setting $\lambda_t = (1-\eta)\lambda_{t-1}$ for some $\eta > 0$.

4.2 Main Theorems

For notation simplicity, with high probability (or w.h.p.) means with probability $1 - e^{-c \log^2(m_1 m_2)}$ and $\widetilde{O}$ hides factors of $\mathsf{polylog}(m_1, m_2)$.

Theorem 2 (three-layer, second variant)**.** *Consider Algorithm 2. For every* $\mathsf{constant}$ $\gamma \in (0, 1/4]$*, every* $\varepsilon_0 \in (0, 1/100]$*, every* $\varepsilon = \frac{\varepsilon_0}{k p_1 p_2^2 \mathfrak{C}_\mathfrak{s}(\Phi, p_2 \mathfrak{C}_\mathfrak{s}(\phi,1)) \mathfrak{C}_\mathfrak{s}(\phi,1)^2}$*, there exists*

$$M = \mathsf{poly}\Big(\mathfrak{C}_\varepsilon(\Phi, \sqrt{p_2}\mathfrak{C}_\varepsilon(\phi, 1)), \frac{1}{\varepsilon}\Big)$$

such that for every $m_2 = m_1 = m \geq M$*, and properly set* $\lambda_w, \lambda_v, \sigma_w, \sigma_v$ *in Table 1, as long as*

$$N \geq \widetilde{\Omega}\left(\Big(\frac{\mathfrak{C}_\varepsilon(\Phi, \sqrt{p_2}\mathfrak{C}_\varepsilon(\phi, 1)) \cdot \mathfrak{C}_\varepsilon(\phi, 1) \cdot \sqrt{p_2} p_1 k^2}{\varepsilon_0}\Big)^2\right)$$

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

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

Remark 4.3. For general Φ, ignoring p_2, the complexity of three-layer networks is essentially $\mathfrak{C}_\varepsilon(\Phi, \mathfrak{C}_\varepsilon(\phi, 1))$. 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 $\mathfrak{C}_\varepsilon(\Phi', 1) = \mathfrak{C}_\varepsilon(\Phi, C) \approx \mathfrak{C}_\varepsilon(\Phi, \Theta(\mathfrak{C}_\varepsilon(\phi, 1)))$.

[13] In the full paper we study two variants of SGD. This present version is the "second variant," and the first variant $L_1(\lambda_{t-1}; W', V')$ is the same as (4.2) by removing $\mathbf{\Sigma}$. Due to technical difficulty, the best sample complexity we can prove for L_1 is a bit higher.

4.3 Our Contributions

Our sample complexity N scales polynomially with the complexity of the target network, and is *(almost) independent* of m, the amount of overparameterization. This itself can be quite surprising, because recent results on neural network generalization [6, 11, 25, 40] require N to be polynomial in m. Furthermore, Theorem 2 shows three-layer networks can efficiently learn a *bigger* concept class (4.1) comparing to what we know about two-layer networks (3.1).

From a practical standpoint, one can construct target functions of the form (4.1) that cannot be (efficiently) approximated by any two-layer target function in (3.1). If data is generated according to such functions, then it may be necessary to use three-layer networks as learners (see Figure 1).

From a theoretical standpoint, even in the special case of $\Phi(z) = z$, our target function can capture *correlations* between non-linear measurements of the data (recall Remark 4.1). This means $\mathfrak{C}_\varepsilon(\Phi, \mathfrak{C}_\varepsilon(\phi, 1)\sqrt{p_2}) \approx O(\sqrt{p_2}\mathfrak{C}_\varepsilon(\phi, 1))$, so learning it is essentially in the *same complexity* as learning each $\phi_{s,j}$. For example, a three-layer network can learn $\cos(100\langle w_1^*, x\rangle) \cdot e^{100\langle w_2^*, x\rangle}$ up to accuracy ε in complexity $\mathsf{poly}(1/\varepsilon)$, while it is unclear how to do so using two-layer networks.

Technical Contributions. We highlight some technical contributions in the proof of Theorem 2.

In recent results on the training convergence of neural networks for more than two layers [3, 4], the optimization process stays in a close neighborhood of the initialization so that, with heavy overparameterization, the interactions across layers are negligible. In our three-layer case, this means that the matrix W never interacts with V. They then argue that SGD succeeds because the learning process is almost convex: because in this regime SGD more or less simulates a neural tangent kernel [28]. In our analysis, we directly tackle *non-convex* interactions between W and V, and thus provide new proofs techniques that could be useful for future theoretical applications.

Also, for the results [3, 4] and our two-layer Theorem 1 to hold, it suffices to analyze a regime where the "sign pattern" of ReLUs can be replaced with that of the random initialization. (Recall $\sigma(x) = \mathbb{I}_{x\geq 0} \cdot x$ and we call $\mathbb{I}_{x\geq 0}$ the "sign pattern.") In our three-layer analysis, the optimization process has moved *sufficiently away from initialization*, so that the sign pattern change can significantly affect output. This brings in additional technical challenge because we have to tackle non-convex interactions between W and V together with changing sign patterns.[14]

Comparison to Daniely [16]. Daniely [16] studies the learnability of multi-layer networks when (essentially) only the *output layer* is trained, which reduces to a convex task. He shows that multi-layer networks can learn a compositional kernel space, which implies two/three-layer networks can efficiently learn low-degree polynomials. He did not derive the general sample/time complexity bounds for more complex functions such as those in our concept classes (3.1) and (4.1), but showed that they are finite.

In contrast, our learnability result of concept class (4.1) is due to the *non-convex* interaction between hidden layers. Since Daniely [16] studies the regime when the changes in hidden layers are negligible, if three layer networks are used, to the best of our knowledge, their theorem *cannot* lead to similar sample complexity bounds comparing to Theorem 2 by only training the last layer of a three-layer network. Empirically, one can also observe that training hidden layers is better than training the last layer (see Figure 1).

[14] For instance, the number of sign changes can be $m^{0.999}$ for the second hidden layer (see Lemma 6.4). In this region, the network output can be affected by $m^{0.499}$ since each neuron is of value roughly $m^{-1/2}$. Therefore, if after training we replace the sign pattern with random initialization, the output will be meaningless.

5 Conclusion and Discussion

We show by training the *hidden layers* of two-layer (resp. three-layer) overparameterized neural networks, one can efficiently learn some important concept classes including two-layer (resp. three-layer) networks equipped with smooth activation functions. Our result is in the agnostic PAC-learning language thus is *distribution-free*. We believe our work *opens up* a new direction in both algorithmic and generalization perspectives of overparameterized neural networks, and pushing forward can possibly lead to more understanding about deep learning.

Our results apply to other more structured neural networks. As a concrete example, consider convolutional neural networks (CNN). Suppose the input is a two dimensional matrix $x \in \mathbb{R}^{d\times s}$ which can be viewed as d-dimensional vectors in s *channels*, then a convolutional layer on top of x is defined as follows. There are d' fixed subsets $\{S_1, S_2, \dots, S_{d'}\}$ of $[d]$ each 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 $x_{S_i} \in \mathbb{R}^{k'\times s}$ is the submatrix of x with rows indexed by S_i; $w_j \in \mathbb{R}^{k'\times s}$ is the weight matrix of the j-th channel; and ϕ is the activation function. Overparameterization then means a larger number of channels m in our learned network comparing to the target. Our analysis can be adapted to show a similar result for this type of networks.

One can also combine this paper with properties of recurrent neural networks (RNNs) [3] to derive PAC-learning results for RNNs [2], or use the existential tools of this paper to derive PAC-learning results for three-layer residual networks (ResNet) [1]. The latter gives a provable separation between neural networks and kernels in the efficient PAC-learning regime.

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 $R(W', V')$ on the weights to control the generalization. In contrast, in practice, it is observed that neural networks actually *implicitly* regularizes: even without any restriction to weights, the network learned by an *unconstraint* overparameterized neural network still generalizes. It is an interesting direction to explain this inductive bias for three layers and beyond. (This present paper only explains inductive bias for two-layer networks.) As for a third direction, currently our result does not directly apply to target networks with ReLU activations. While there are functions ϕ with $\mathfrak{C}_\varepsilon(\phi, 1) = 2^{O(1/\varepsilon)}$ that approximate ReLU function in $[-1, 1]$ with ε error, obtaining an polynomial or sub-exponential bound on the sampling complexity would be of great interest. On the other hand, the result by [50] indicates that learning ReLU could also be hard for algorithms such as SGD.

Proof Sketch

We present key technical lemmas we used for proving the three-layer network theorems in Section 6 so that interested readers do not need to go through the appendix. Many of them can be of independent interests and have found further applications beyond this paper (such as for residual networks [1] and for recurrent networks [2]). The full proof is in Appendix C.

The two-layer result is based on similar ideas but simpler, and the full proof is in Appendix B.

We give more experiments in Section 7 on Page 21. Our appendix starts on page 24.

6 Main Lemmas for Three Layer Networks

In Section 6.1, we state the main theorem for the first variant of the SGD, which we excluded from the main body due to space limitation. In Section 6.2, we show the existence of some good "pseudo network" that can approximate the target. In Section 6.3, we present our coupling technique between a real network and a pseudo network. In Section 6.4, we present the key lemma about the optimization procedure. In Section 6.5, we state a simple generalization bound that is compatible with our algorithm. These techniques together give rise to the proof of Theorem 3. In Section 6.6, we present additional techniques needed to show Theorem 2.

6.1 First Variant of SGD

In the first variant of SGD, in each round $t = 1, 2, \dots, T$, we use (noisy) SGD to minimize the following stochastic objective for some fixed λ_{t-1}:

$$L_1(\lambda_{t-1}; W', V') \stackrel{\text{def}}{=} L\left(\lambda_{t-1} F\big(x; W^{(0)} + W^\rho + W', V^{(0)} + V^\rho + V'\big)\right) + R(\sqrt{\lambda_{t-1}} W', \sqrt{\lambda_{t-1}} V') \tag{6.1}$$

Above, the objective is stochastic because (1) $z \sim \mathcal{Z}$ is a random sample from the training set, and (2) W^ρ and V^ρ are two small perturbation matrices. This is only different from (4.2) by removing $\boldsymbol{\Sigma}$. We include the pseudocode in Algorithm 3.

Algorithm 3 SGD for three-layer networks (first variant (6.1))

```
Input: Data set $\mathcal{Z}$, initialization $W^{(0)}, V^{(0)}$, step size $\eta$, number of inner steps $T_w$, $\sigma_w, \sigma_v, \lambda_w, \lambda_v$.
$W_0 = 0, V_0 = 0, \lambda_1 = 1, T = \Theta\left(\eta^{-1} \log \frac{\log(m_1 m_2)}{\varepsilon_0}\right)$.
for $t = 1, 2, \dots, T$ do
    Apply noisy SGD with step size $\eta$ on the stochastic objective $L_1(\lambda_{t-1}; W, V)$ for $T_w$ steps;
    the starting point is $W = W_{t-1}, V = V_{t-1}$ and suppose it reaches $W_t, V_t$.   ◇ see Lemma A.9
    $\lambda_{t+1} = (1 - \eta)\lambda_t$.   ◇ weight decay
end for
Randomly sample $\widetilde{\Theta}(1/\varepsilon_0^2)$ many noise matrices $\{W^{\rho,j}, V^{\rho,j}\}$. Let
    $$j^* = \arg\min_j \left\{ \mathbb{E}_{z \in \mathcal{Z}} L\left(\lambda_T F\big(x; W^{(0)} + W^{\rho,j} + W_T, V^{(0)} + V^{\rho,j} + V_T\big)\right)\right\}$$
Output $W_T^{(out)} = W^{(0)} + W^{\rho,j^*} + W_T$, $V_T^{(out)} = V^{(0)} + V^{\rho,j^*} + V_T$.
```

Below is the main theorem for using the first variant of SGD to train three-layer networks.

Theorem 3 (three-layer, first variant)**.** *Consider Algorithm 3. In the same setting as Theorem 2, for every $m_2 = m_1 = m \geq M$, as long as*

$$N \geq \widetilde{\Omega}(M m^{3/2})$$

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

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

As m goes large, this sample complexity $N \approx m^{3/2}$ polynomially scales with m so may not be very efficient (we did not try hard to improve the exponent 3/2). Perhaps interesting enough, this is already *non-trivial*, because N can be much smaller than m^2, the number of parameters of the network or equivalently the naive VC dimension bound. Recall in our second variant of SGD, the sample complexity N only polylogarithmically depend on m.

6.2 Existence

We wish to show the existence of some good "pseudo network" that can approximate the target network. In a pseudo network, each ReLU activation $\sigma(x) = \mathbb{I}_{x \geq 0} x$ is replaced with $\mathbb{I}_{x^{(0)} \geq 0} x$ where $x^{(0)}$ is the value at *random initialization*. Formally, let

- $D_{w,x} \in \{0,1\}^{m_1 \times m_1}$ denote a diagonal sign matrix indicating the sign of the ReLU's for the first layer at *random initialization*, that is, $[D_{w,x}]_{i,i} = \mathbb{I}[\langle w_i^{(0)}, x \rangle + b_{1,i}^{(0)} \geq 0]$, and
- $D_{v,x} \in \{0,1\}^{m_2 \times m_2}$ denote the diagonal sign matrix of the second layer at random initialization.

Consider the output of a three-layer network at randomly initialized sign without bias as

$$g_r^{(0)}(x; W, V) \stackrel{\text{def}}{=} a_r D_{v,x} V D_{w,x} W x$$
$$G^{(0)}(x; W, V) \stackrel{\text{def}}{=} \left(g_1^{(0)}, \cdots, g_k^{(0)}\right)$$

Lemma 6.1 (existence)**.** *For every $\varepsilon \in \left(0, \frac{1}{k p_1 p_2^2 \mathfrak{C}_\mathfrak{s}(\Phi, p_2 \mathfrak{C}_\mathfrak{s}(\phi,1)) \mathfrak{C}_\mathfrak{s}(\phi,1)^2}\right)$, there exists*

$$M = \mathsf{poly}\left(\mathfrak{C}_\varepsilon\big(\Phi, \sqrt{p_2}\mathfrak{C}_\varepsilon(\phi,1)\big), \frac{1}{\varepsilon}\right)$$
$$C_0 = \mathfrak{C}_\varepsilon(\Phi, \sqrt{p_2}\mathfrak{C}_\varepsilon(\phi,1)) \cdot \mathfrak{C}_\varepsilon(\phi,1) \cdot \widetilde{O}(p_1 \sqrt{p_2} k)$$

such that if $m_1, m_2 \geq M$, then with high probability, there exists weights W^*, V^* with

$$\|W^*\|_{2,\infty} = \max_i \|w_i^*\|_2 \leq \frac{C_0}{m_1}, \quad \|V^*\|_{2,\infty} = \max_i \|v_i^*\|_2 \leq \frac{\sqrt{m_1}}{m_2}$$

such that

$$\mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\sum_{r=1}^k \left| f_r^*(x) - g_r^{(0)}(x; W^*, V^*)\right|\right] \leq \varepsilon,$$

and hence,

$$\mathbb{E}_{(x,y)\sim\mathcal{D}}\left[L(G^{(0)}(x; W^*, V^*), y)\right] \leq \mathsf{OPT} + \varepsilon.$$

In other words, at randomly initialized signs, there exist choices of W^* and V^* with small norms so that $G^{(0)}(x; W^*, V^*)$ approximates the target. Later, we will combine this with the coupling lemma Section 6.3 to show a main structural property of overparameterized networks: solutions with good population risks 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.

6.2.1 Technical Ideas

We begin with a simple task to illustrate the main idea. Let $w^* \in \mathbb{R}^d$ be a given vector and suppose we want to approximate function $\phi(\langle w^*, x \rangle)$ (over x) by designing a random function $\mathbb{I}_{\langle w,x \rangle + b_0 \geq 0} h(\langle w, w^* \rangle, b_0)$ (over x) where w is a random Gaussian, b_0 is a random bias, and $h(\cdot,\cdot)$ is a function at our choice. The next lemma says that, we can design h with good property so that the expectation of the random function $\mathbb{I}_{\langle w,x \rangle + b_0 \geq 0} h(\langle w, w^* \rangle, b_0)$ is close to $\phi(\langle w^*, x \rangle)$.

Lemma 6.2 (indicator to function)**.** *For every smooth function ϕ, every $\varepsilon \in \left(0, \frac{1}{\mathfrak{C}_\mathfrak{s}(\phi,1)}\right)$, there exists a function $h : \mathbb{R}^2 \to [-\mathfrak{C}_\varepsilon(\phi,1), \mathfrak{C}_\varepsilon(\phi,1)]$ that is also $\mathfrak{C}_\varepsilon(\phi,1)$-Lipschitz continuous on its first coordinate with the following two (equivalent) properties:*

(a) For every $x_1 \in [-1,1]$:

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

where $\alpha_1,\beta_1,b_0 \sim \mathcal{N}(0,1)$ are independent random variables.

(b) For every $w^*, x \in \mathbb{R}^d$ with $\|w^*\|_2 = \|x\|_2 = 1$:

$$\left|\mathbb{E}\left[\mathbb{I}_{\langle w,x\rangle+b_0\geq 0} h(\langle w,w^*\rangle,b_0)\right] - \phi(\langle w^*,x\rangle)\right| \leq \varepsilon$$

where $w \sim \mathcal{N}(0,\mathbf{I})$ is an d-dimensional Gaussian, $b_0 \sim \mathcal{N}(0,1)$.

Furthermore, h satisfies $\mathbb{E}_{\alpha_1,b_0\sim\mathcal{N}(0,1)}\left[h(\alpha_1,b_0)^2\right] \leq (\mathfrak{C}_\mathfrak{s}(\phi,1))^2$.

If one designs a vector $w^{\divideontimes} = \big(0,\dots,0,2h(\langle w,w^*\rangle,b_0)\big)$, then using $x_d = 1/2$, Lemma 6.2 implies $\left|\mathbb{E}\left[\mathbb{I}_{\langle w,x\rangle+b_0\geq 0}\langle w^{\divideontimes},x\rangle\right] - \phi(\langle w^*,x\rangle)\right| \leq \varepsilon$. Therefore, Lemma 6.2 corresponds to Lemma 6.1 in the special case of a *single* neuron.

Remark. The reason Lemma 6.2b is equivalent to Lemma 6.2a consists of a few thinking steps. Without loss of generality, one can assume $w^* = (1,0,0,\dots,0)$ and write $\phi(\langle w^*,x\rangle) = \phi(x_1)$ and write $h(\langle w,w^*\rangle,b_0) = h(w_1,b_0)$ so it only depends on the first coordinate of w and b_0. Under such simplifications, the second through last coordinates do not make any difference, so we can assume without loss of generality that w^*, w, x are only in 2 dimensions, and write $w = (\alpha_1,\beta_1)$ and $x = (x_1, \sqrt{1-x_1^2})$. In sum, proving Lemma 6.2a suffices in establishing Lemma 6.2b.

Given Lemma 6.2, we can directly apply it to the two-layer case (see Appendix B.1) to show the existence of good pseudo networks. As for the three-layer case, we need to apply Lemma 6.2 twice: once for (each neuron of) the second hidden layer and once for the output.

Consider the the input (without bias) to a single neuron of the second hidden layer at random initialization. Without loss of generality, say the first neuron, given as:

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

Even though $n_1(x)$ is completely random, using Lemma 6.2, we can derive the following lemma which rewrites $n_1(x)$ in the direction of an arbitrary function ϕ.

Lemma 6.3 (information out of randomness)**.** *For every smooth function ϕ, every $w^* \in \mathbb{R}^d$ with $\|w^*\|_2 = 1$, for every $\varepsilon \in \big(0, \frac{1}{\mathfrak{C}_\mathfrak{s}(\phi,1)}\big)$, there exists real-valued functions*

$$\rho(v_1^{(0)},W^{(0)},b_1^{(0)}),\ B(x,v_1^{(0)},W^{(0)},b_1^{(0)}), R(x,v_1^{(0)},W^{(0)},b_1^{(0)}),\ \textit{and}\ \phi_\varepsilon(x)$$

such that for every x:

$$n_1(x) = \rho\left(v_1^{(0)},W^{(0)},b_1^{(0)}\right)\phi_\varepsilon(x) + B\left(x,v_1^{(0)},W^{(0)},b_1^{(0)}\right) + R\left(x,v_1^{(0)},W^{(0)},b_1^{(0)}\right) .$$

Moreover, letting $C = \mathfrak{C}_\varepsilon(\phi,1)$ be the complexity of ϕ, and if $v_{1,i}^{(0)} \sim \mathcal{N}(0,\frac{1}{m_2})$ and $w_{i,j}^{(0)}, b_{1,i}^{(0)} \sim \mathcal{N}(0,\frac{1}{m_1})$ are at random initialization, then 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 $\|x\|_2 = 1$, $|\phi_\varepsilon(x) - \phi(\langle w^*,x\rangle)| \leq \varepsilon$.*
4. *For every fixed x with $\|x\|_2 = 1$, with high probability $\left|R\left(x,v_1^{(0)},W^{(0)},b_1^{(0)}\right)\right| \leq \widetilde{O}\left(\frac{1}{\sqrt{m_1 m_2}}\right)$ and $\left|B\left(x,v_1^{(0)},W^{(0)},b_1^{(0)}\right)\right| \leq \widetilde{O}\left(\frac{1}{\sqrt{m_2}}\right)$.*

Furthermore, there exists real-valued function $\widetilde{\rho}(v_1^{(0)})$ satisfying with high probability:

$$\widetilde{\rho}(v_1^{(0)}) \sim \mathcal{N}\left(0, \frac{1}{100C^2 m_2}\right) \quad \textit{and} \quad \mathcal{W}_2(\rho|_{W^{(0)}, b_1^{(0)}}, \widetilde{\rho}) \le \widetilde{O}\left(\frac{1}{C\sqrt{m_1 m_2}}\right) \quad .$$

Lemma 6.3 shows that, up to some small noise B and R, 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 target network. This allows us to apply Lemma 6.2 again for the output layer, so as to construct for instance a composite function $\Phi(\cdot)$ with ϕ_i as its inputs.[15]

Remark. Lemma 6.3 may sound weird at first look because random initialization cannot carry any information about the target. 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 6.3 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 Lemma 6.3 and applying Lemma 6.2 once more, we can prove Lemma 6.1.

6.3 Coupling Between Real and Pseudo Networks

Suppose we are currently at weights $W^{(0)} + W' + W^\rho$, $V^{(0)} + V' + V^\rho$, where matrices W^ρ, V^ρ are random Gaussian matrices such that:

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

for some $\sigma_v, \sigma_w \in [1/(m_1 m_2), 1]$ to be specified later, and W', V' are matrices with bounded norms that can depend on the randomness of $W^{(0)}, b_1^{(0)}, V^{(0)}, b_2^{(0)}$. Intuitively, W' and V' capture how much the algorithm has moved away from the initialization, while W^ρ, V^ρ are introduced for adding smoothness in the optimization, see Section 6.4.

Let us introduce the notion of pseudo networks at the current weights. Let

- $D_{w,x}$ denote the diagonal sign matrix of the first layer at random initialization $W^{(0)}$,
- $D_{v,x}$ denote the diagonal sign matrix of the second layer at random initialization $W^{(0)}, V^{(0)}$,
- $D_{w,x} + D'_{w,x} \in \{0,1\}^{m_1 \times m_1}$ denote the diagonal sign matrix of the first layer at weights $W^{(0)} + W' + W^\rho$ and $V^{(0)} + V' + V^\rho$, i.e., $[D_{w,x} + D'_{w,x}]_{i,i} = \mathbb{I}[\langle w_i^{(0)} + w_i' + w_i^{(\rho)}, x\rangle + b_{1,i}^{(0)} \ge 0]$.
- $D_{v,x} + D'_{v,x} \in \{0,1\}^{m_2 \times m_2}$ denote the diagonal sign matrix of the second layer at these weights.

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

$$\begin{aligned} g_r(x; W, V) &= a_r(D_{v,x} + D'_{v,x})\left(V(D_{w,x} + D'_{w,x})(Wx + b_1) + b_2\right) \\ g_r^{(b)}(x; W, V) &= a_r(D_{v,x} + D'_{v,x})V(D_{w,x} + D'_{w,x})(Wx + b_1) \\ g_r^{(b,b)}(x; W, V) &= a_r(D_{v,x} + D'_{v,x})V(D_{w,x} + D'_{w,x})Wx \end{aligned}$$

As a sanity check, at $W^{(0)} + W' + W^\rho$, $V^{(0)} + V' + V^\rho$ the pseudo network equals the true one:

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

[15] One technical issue is the following. When considering multiple neurons of the second layer, those Gaussian variables ρ are not independent because they all depend on $W^{(0)}$. The notion of $\widetilde{\rho}$ in Lemma 6.3 removes such dependency across multiple neurons, at the expense of small Wasserstein distance.

We state Lemma 6.4 below.

Lemma 6.4 (coupling)**.** *Suppose $\tau_v \in [0,1]$, $\tau_w \in \big(\frac{1}{m_1^{3/2}}, \frac{1}{m_1^{1/2}}\big]$, $\sigma_w \in \big(\frac{1}{m_1^{3/2}}, \frac{\tau_w}{m_1^{1/4}}\big]$, $\sigma_v \in \big(0, \frac{1}{m_2^{1/2}}\big]$ and $\eta > 0$. Given fixed unit vector x, and perturbation matrices W', V', W'', V'' (that may depend on the randomness of $W^{(0)}, b_1^{(0)}, V^{(0)}, b_2^{(0)}$ and x) satisfying*

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

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

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

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

The first statement "sparse sign change" of Lemma 6.4 says that, if we move from random initialization $W^{(0)}, V^{(0)}$ to $W^{(0)} + W' + W^\rho, V^{(0)} + V' + V^\rho$, then how many signs of the ReLUs (in each layer) will change, as a function of the norms of W' and V'. This calculation is similar to [4] but slightly more involved due to the $\|\cdot\|_{2,4}$ norm that we use here.

The second statement "cross term vanish" of Lemma 6.4 studies, if we are currently at weights $(W, V) = (W^{(0)} + W' + W^\rho, V^{(0)} + V' + V^\rho)$ and want to move to $(W + \eta\mathbf{\Sigma}W'', V + \eta V''\mathbf{\Sigma})$ where $\mathbf{\Sigma}$ is a diagonal matrix with diagonal entries i.i.d. uniformly chosen from $\{\pm 1\}$, then how does the function value change in the pseudo network.

6.3.1 Coupling + Existence

We quickly point out a corollary by applying the coupling and existential lemmas together.

Recall in the existential Lemma 6.1, we have studied a pseudo network $G^{(0)}$ where the signs are determined at the random initialization $W^{(0)}, V^{(0)}$. Now, the coupling Lemma 6.4 says that the amount of sign change from $G^{(0))}$ to $G^{(b,b)}$ can be controlled. Therefore, if parameters are chosen appropriately, the existential Lemma 6.1 should also apply to $G^{(b,b)}$. Formally,

Corollary 6.5 (existence after coupling)**.** *In the same setting as Lemma 6.1, given perturbation matrices W', V' (that may depend on the randomness of the initialization and the data distribution $\mathcal{D}$) with*

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

Using parameter choices from Table 1, w.h.p. there exist W^* and V^* (independent of the randomness of W^ρ, V^ρ) satisfying

$$\|W^*\|_{2,\infty} = \max_i \|w_i^*\|_2 \le \frac{C_0}{m_1}, \quad \|V^*\|_{2,\infty} = \max_i \|v_i^*\|_2 \le \frac{\sqrt{m_1}}{m_2}$$

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

$$\mathbb{E}_{(x,y)\sim\mathcal{D}}\left[L(G^{(b,b)}(x; W^*, V^*), y)\right] \le \mathsf{OPT} + \varepsilon.$$

In the three-layer network results, we choose parameters

$$\tau'_v = \frac{1}{\sqrt{\varepsilon_0}} \frac{m_1^{1/2-0.005}}{m_2^{1/2}}, \quad \tau'_w = \frac{C_0}{\varepsilon_0^{1/4}} \frac{1}{m_1^{3/4-0.005}}, \quad \lambda_v = \frac{2}{\tau_v^2}, \quad \lambda_w = \frac{2}{(\tau'_w)^4},$$

$$\sigma_v = \frac{1}{m_2^{1/2+0.01}}, \quad \sigma_w = \frac{1}{m_1^{1-0.01}}, \quad \tau_v = \frac{m_1^{1/2-0.001}}{m_2^{1/2}} \gg \tau'_v, \quad \tau_w = \frac{1}{m_1^{3/4-0.01}} \gg \tau'_w$$

$$m_2 = m_1, \qquad \varepsilon_a = 1,$$

for $C_0 = \mathfrak{C}_\varepsilon(\Phi, \sqrt{p_2}\mathfrak{C}_\varepsilon(\phi,1)) \cdot \mathfrak{C}_\varepsilon(\phi,1) \cdot \widetilde{O}(p_1\sqrt{p_2}k)$ and $\varepsilon = \frac{\varepsilon_0}{kp_1p_2^2\mathfrak{C}_\mathfrak{s}(\Phi,p_2\mathfrak{C}_\mathfrak{s}(\phi,1))\mathfrak{C}_\mathfrak{s}(\phi,1)^2}$

Table 1: Three-layer parameter choices (the constant inside Θ can depend on γ).
λ_w, λ_v give the weights of the regularizer $\lambda_w\|\cdot\|_{2,4}^4 + \lambda_v\|\cdot\|_{2,2}^2$ in the objective (6.1) and (4.2).
σ_w, σ_v give the amount of Gaussian perturbation we add to the objective (for analysis purpose).
τ'_w, τ'_v are set so that if the regularizer is bounded, it satisfies $\|W'\|_{2,4} \le \tau'_w$ and $\|V'\|_F \le \tau'_v$.
τ_w, τ_v are set so that the coupling lemma works whenever $\|W'\|_{2,4} \le \tau_w$ and $\|V'\|_F \le \tau_v$.

As we shall see later, Corollary 6.5 gives rise to W^* and V^* that shall be used as a descent direction for the objective. (Corollary 6.5 did not use all the parameters inside Table 1, and some of those parameters shall be used in later subsections.)

6.4 Optimization

The na ïvelyapproach is to

- use the property that solutions with good risks are dense in the parameter space (i.e., Corollary 6.5) to show that the optimization landscape of the overparameterized three-layer neural network is benign: it has no spurious local minimal or (more or less) even any second-order critical points; and
- use existing theorems on escaping saddle points (such as [20]) to show that SGD will not be stuck in saddle points and thus converges.

Key issue. Unfortunately, before digging into the details, there is already a big hole in this approach. ReLU networks are not second-order-differentiable: a ReLU activation does not have a well-defined Hessian/sub-Hessian at zero. One may na ïvelythink 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. In other words, the only non-smooth point of ReLU has a Hessian value equal to the Dirac δ-function, but such 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 na ïvelyignore the "Hessian" of ReLU at zeros.*

Smoothing. To fix the na ïveapproach, 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 in x 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. We show that $\mathbb{E}[L(F(x; W+W^\rho, V+V^\rho), y)]$ also has the desired property of essentially having no second-order critical points. Perhaps worth pointing out, the Hessian of this smoothed function is *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 [30]. Smooth analysis [45] 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:

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

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. $R(\sqrt{\lambda_t} W_t, \sqrt{\lambda_t} V_t) = \lambda_v \|\sqrt{\lambda_t} V_t\|_F^2 + \lambda_w \|\sqrt{\lambda_t} W_t\|_{2,4}^4$ and λ_v, λ_w are set such that $\lambda_v \|\sqrt{\lambda_t} V^*\|_F^2 \le \varepsilon_0$ and $\lambda_w \|\sqrt{\lambda_t} W^*\|_{2,4}^4 \le \varepsilon_0$ for every W^* and V^* coming from Lemma 6.1. See Algorithm 3 (first SGD variant) for the details. We prove the following lemma.

Lemma 6.6 (descent direction)**.** *For every $\varepsilon_0 \in (0, 1)$ and $\varepsilon = \frac{\varepsilon_0}{k p_1 p_2^2 \mathfrak{C}_\mathfrak{s}(\Phi, p_2 \mathfrak{C}_\mathfrak{s}(\phi,1)) \mathfrak{C}_\mathfrak{s}(\phi,1)^2}$, for every constant $\gamma \in (0, 1/4]$, consider the parameter choices in Table 1, and consider any λ_t, W_t, V_t (that may depend on the randomness of $W^{(0)}, b^{(0)}, V^{(0)}, b^{(1)}$ and $\mathcal{Z}$) with*

$$\lambda_t \in \left((\varepsilon/\log(m_1 m_2))^{\Theta(1)}, 1\right] \quad \text{and} \quad L'(\lambda_t, W_t, V_t) \in [(1+\gamma)\mathsf{OPT} + \Omega(\varepsilon_0/\gamma), \widetilde{O}(1)]$$

With high probability over the random initialization, there exists W^*, V^* with $\|W^*\|_F, \|V^*\|_F \le 1$ such that for every $\eta \in \left[0, \frac{1}{\mathsf{poly}(m_1, m_2)}\right]$:

$$\begin{aligned} &\min\left\{\mathbb{E}_{\mathbf{\Sigma}}\Big[L'\left(\lambda_t, W_t + \sqrt{\eta}\mathbf{\Sigma} W^*, V_t + \sqrt{\eta} V^* \mathbf{\Sigma}\right)\Big], L'\big((1-\eta)\lambda_t, W_t, V_t\big)\right\} \\ &\le (1 - \eta\gamma/4)(L'(\lambda_t, W_t, V_t)) \ , \end{aligned}$$

where $\mathbf{\Sigma} \in \mathbb{R}^{m_1\times m_1}$ is a diagonal matrix with each diagonal entry i.i.d. uniformly drawn from $\{\pm 1\}$.

Lemma 6.6 says one of the following two scenarios will happen. Either (1) there exist W^*, V^* so that updating in a random direction $(\mathbf{\Sigma} W^*, V^* \mathbf{\Sigma})$ decreases the objective, or (2) performing weight decay decreases the objective. It is a simple exercise to check that, if (1) happens, then $F'(\lambda, \cdot, \cdot)$ has a very negative curvature in the Hessian (see Fact A.8) at the current point. Therefore, Lemma 6.6 essentially says that after appropriate regularization, every second-order critical point of L' is approximately global minimum.

More interestingly, since (noisy) stochastic gradient descent is capable of finding approximate second-order critical points with a global convergence rate, one can show the following final convergence rate for minimizing L' for Algorithm 3.

Lemma 6.7 (convergence)**.** *In the setting of Theorem 3, with probability at least 99/100, Algorithm 3 (the first SGD variant) converges in $TT_w = \mathsf{poly}\,(m_1, m_2)$ iterations to a point*

$$L'(\lambda_T, W_T, V_T) \le (1+\gamma)\mathsf{OPT} + \varepsilon_0.$$

6.4.1 Details of Smoothing

The next lemma shows that when doing a small update to the current weight, one can view the sign pattern as fixed for the smoothed objective, up to a small error. Specifically, for every input x and every $W = W^{(0)} + W', V = V^{(0)} + V'$, let $\mathbf{\Sigma}$ be a random diagonal matrix with $\pm$ entries, and let

- $D_{w,x,\rho}$ denote the diagonal matrix with diagonals being 0-1 signs of the first layer at $W + W^\rho$;
- $D_{w,x,\rho,\eta}$ denote of the first layer at weights $W + W^\rho + \eta \mathbf{\Sigma} W''$;
- $D_{v,x,\rho}$ denote that of the second layer at weights $W + W^\rho$ and $V + V^\rho$; and
- $D_{v,x,\rho,\eta}$ denote that of the second layer at weights $W + W^\rho + \eta \mathbf{\Sigma} W''$ and $V + V^\rho + \eta \mathbf{\Sigma} V''$.

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

$$\begin{aligned} P_{\rho,\eta} &\stackrel{\text{def}}{=} f_r(x; W + W^\rho + \eta \mathbf{\Sigma} W'', V + V^\rho + \eta V'' \mathbf{\Sigma}) \\ &= a_r D_{v,x,\rho,\eta} \Big((V + V^\rho + \eta V'' \mathbf{\Sigma}) D_{w,x,\rho,\eta} \left((W + W^\rho + \eta \mathbf{\Sigma} W'') x + b_1 \right) + b_2 \Big) \\ P'_{\rho,\eta} &\stackrel{\text{def}}{=} g_r(x; W + W^\rho + \eta \mathbf{\Sigma} W'', V + V^\rho + \eta V'' \mathbf{\Sigma}) \\ &= a_r D_{v,x,\rho} \Big((V + V^\rho + \eta V'' \mathbf{\Sigma}) D_{w,x,\rho} \left((W + W^\rho + \eta \mathbf{\Sigma} W'') x + b_1 \right) + b_2 \Big). \end{aligned}$$

We prove the following lemma:

Lemma 6.8 (smoothed real vs pseudo)**.** *There exists $\eta_0 = \frac{1}{\mathsf{poly}(m_1, m_2)}$ such that, for every $\eta \leq \eta_0$, for every fixed x with $\|x\|_2 = 1$, for every W', V', W'', V'' that may depend on the randomness of the initialization and*

$$\|W'\|_{2,4} \leq \tau_w, \quad \|V'\|_{2,2} \leq \tau_v, \quad \|W''\|_{2,\infty} \leq \tau_{w,\infty}, \quad \|V''\|_{2,\infty} \leq \tau_{v,\infty}$$

we have with high probability:

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

where O_p hides polynomial factor of m_1, m_2.

In other words, when working with a Gaussian smoothed network (with output $P_{\rho,\eta}$), it suffices to study a pseudo network (with output $P'_{\rho,\eta}$). In the proof of Lemma 6.6, this allows us to go from the real (smoothed) network to the pseudo (smoothed) network, and then apply Lemma 6.4.

6.5 Generalization

In our first variant of SGD Algorithm 3, we show a very crude Rademacher complexity bound that can be derived from the contraction lemma (a.k.a. Talagrand's concentration inequality).

Lemma 6.9 (generalization for $L_R = L_1$)**.** *For every $\tau'_v, \tau'_w \geq 0$, every $\sigma_v \in (0, 1/\sqrt{m_2}]$, w.h.p. for every $r \in [k]$ and every $N \geq 1$, the empirical Rademacher complexity is bounded by*

$$\begin{aligned} &\frac{1}{N} \mathbb{E}_{\xi \in \{\pm 1\}^N} \left[\sup_{\|V'\|_F \leq \tau'_v, \|W'\|_{2,4} \leq \tau'_w} \sum_{i \in [N]} \xi_i f_r(x_i; W^{(0)} + W^\rho + W', V^{(0)} + V^\rho + V') \right] \\ &\leq \widetilde{O} \left(\frac{\tau'_w m_1 \sqrt{m_2} + \tau'_v m_2}{\sqrt{N}} + \frac{\tau'_v \sqrt{m_1 m_2 \tau'_w (1/\sqrt{m_1} + \tau'_w)}}{N^{1/4}} \right) \quad . \end{aligned}$$

Since the population risk is bounded by the Rademacher complexity, combining this with Lemma 6.7, one can easily prove Theorem 3.

6.6 Second Variant of SGD

In our second variant of SGD Algorithm 2, we have added the Dropout-type noise matrix $\mathbf{\Sigma}$ directly into the objective L_2 (see (4.2)). Our new stochastic objective is the following.

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

To show that this gives a better sampling complexity bound, we need the following stronger coupling lemma. It gives a somewhat better bound on the "cross term vanish" part comparing to the old coupling Lemma 6.4.

Lemma 6.10 (stronger coupling). *With high probability over the random initialization and over a random diagonal matrix $\mathbf{\Sigma}$ with diagonal entries i.i.d. generated from $\{-1, 1\}$, it satisfies that for every W', V' with $\|V'\|_2 \leq \tau_v$, $\|W'\|_{2,4} \leq \tau_w$ for $\tau_v \in [0, 1]$ and $\tau_w \in \left[\frac{1}{m_1^{3/4}}, \frac{1}{m_1^{9/16}}\right]$, we have*

$$f_r(x; W^{(0)} + \mathbf{\Sigma} W', V^{(0)} + V'\mathbf{\Sigma}) = a_r D_{v,x}(V^{(0)} D_{w,x}(W^{(0)} x + b_1) + b_2) + a_r D_{v,x} V' D_{w,x} W' x$$
$$\pm \widetilde{O}\left(\tau_w^{8/5} m_1^{9/10} + \tau_w^{16/5} m_1^{9/5} \sqrt{m_2} + \frac{\sqrt{m_2}}{\sqrt{m_1}} \tau_v\right) .$$

Under parameter choices Table 1, the last error term is at most ε/k.

For this SGD variant, we also have the following the stronger Rademacher complexity bound. It relies on Lemma 6.10 to reduce the function class to pseudo networks, which are only *linear* functions in W' and V', and then computes its Rademacher complexity.

Lemma 6.11 (generalization for $L_R = L_2$). *For every $\tau'_v \in [0, 1]$, $\tau'_w \in \left[\frac{1}{m_1^{3/4}}, \frac{1}{m_1^{9/16}}\right]$, every $\sigma_w \in [0, 1/\sqrt{m_1}]$ and $\sigma_v \in [0, 1/\sqrt{m_2}]$, w.h.p. for every $r \in [k]$ and every $N \geq 1$, we have by our choice of parameters in Lemma 6.6, the empirical Rademacher complexity is bounded by*

$$\frac{1}{N}\mathbb{E}_{\xi \in \{\pm 1\}^N}\left[\sup_{\|V'\|_F \leq \tau'_v, \|W'\|_{2,4} \leq \tau'_w} \left| \sum_{i \in [N]} \xi_i \mathbb{E}_{\mathbf{\Sigma}}[f_r(x_i; W^{(0)} + W^\rho + \mathbf{\Sigma} W', V^{(0)} + V^\rho + V'\mathbf{\Sigma})]\right|\right]$$
$$\leq \widetilde{O}\left(\frac{\tau'_w \tau'_v m_1^{1/4} \sqrt{m_2}}{\sqrt{N}} + \left((\tau'_w)^{8/5} m_1^{9/10} + (\tau'_w)^{16/5} m_1^{9/5} \sqrt{m_2} + \frac{\sqrt{m_2}}{\sqrt{m_1}} \tau'_v\right)\right).$$

Under parameter choices in Table 1, this is at most $\widetilde{O}\left(\frac{\tau'_w \tau'_v m_1^{1/4} \sqrt{m_2}}{\sqrt{N}}\right) + \varepsilon/k$.

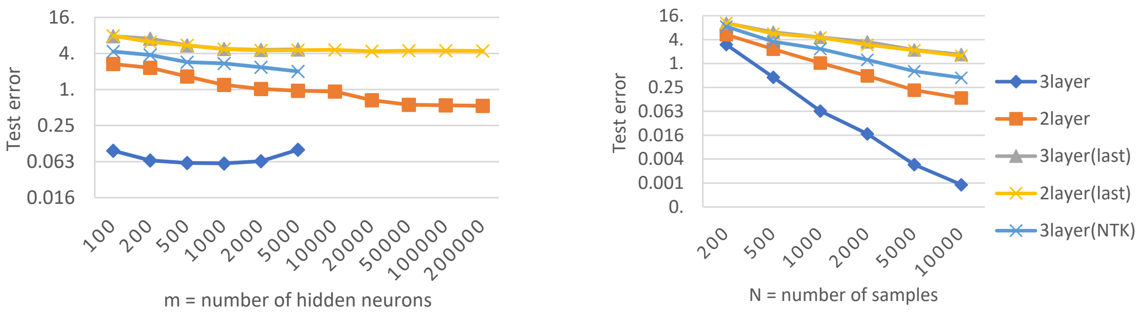
After plugging Lemma 6.10 and Lemma 6.11 into the final proof, we can show Theorem 2.

Remark. In contrast, without the Dropout-type noise matrix $\mathbf{\Sigma}$, the error term in the old coupling Lemma 6.4 is too large so we cannot get better Rademacher complexity bounds.

7 Preliminary Empirical Evaluation

We discuss our experiment details for Figure 1.

Recall that we generate synthetic data where the feature vectors $x \in \mathbb{R}^4$ are generated as random Gaussian then normalized to norm 1, and labels are generated by target function $y = F^*(x) = (\sin(3x_1) + \sin(3x_2) + \sin(3x_3) - 2)^2 \cdot \cos(7x_4)$. Intuitively, the constants 3 and 7 control



(a) $N = 1000$ and vary m

(b) $m = 2000$ and vary N

Figure 2: Performance comparison. `3layer/2layer` stands for training (hidden weights) in three and two-layer neural networks. `(last)` stands for *conjugate kernel* [16], meaning training only the output layer. `(NTK)` stands for *neural tangent kernel* [28] with finite width.

Setup. We consider ℓ_2 regression task on synthetic data where feature vectors $x \in \mathbb{R}^4$ are generated as normalized random Gaussian, and label is generated by target function $y = F^*(x) = (\tanh(8x_1) + \tanh(8x_2) + \tanh(8x_3) - 2)^2 \cdot \tanh(8x_4)$.

the complexity of the activation functions in the target function, and we choose these values to ensure that the two factors in the target function have roughly the same complexity.[16]

To be more consistent with our theorems, we implement fully connected neural networks and train only hidden weights (namely, W in the two-layer case and W, V in the two-layer case). We also implement NTK with respect to only hidden weights. For conjugate kernel, we only train the last (output) layer, that is, the weights $a_r \in \mathbb{R}^m$ for $r \in [k]$ in the language of this paper.

To be consistent with our theorems, we choose random initialization as follows. Entries of a_r are i.i.d. from $\mathcal{N}(0, 1)$, and entries of W, V, b_1, b_2 are i.i.d. from $\mathcal{N}(0, \frac{1}{m})$. This ensures that the output at random initialization is $\Theta(1)$.

We use the default SGD optimizer of pytorch, with momentum 0.9, mini-batch size 50, learning rate `lr` and weight decay parameter `wd`. We carefully run each algorithm with respect to `lr` and `wd` in the set $\{10^{-k}, 2 \cdot 10^{-k}, 5 \cdot 10^{-k} \colon k \in \mathbb{Z}\}$, and presents the best one in terms of testing accuracy. In each parameter setting, we run SGD for 800 epochs, and decrease `lr` by 10 on epoch 400.

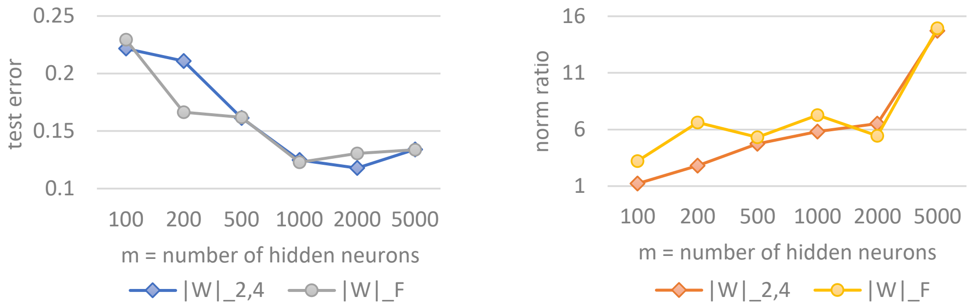
In Figure 2, we provide an additional experiment for target function $y = F^*(x) = (\tanh(8x_1) + \tanh(8x_2) + \tanh(8x_3) - 2)^2 \cdot \tanh(8x_4)$.

7.1 Justification of Our $\|W\|_{2,4}$ Regularizer

Recall our three-layer theorem requires a slightly unconventional regularizer, namely, the $\|W\|_{2,4}$ norm of the first layer to encourage weights to be more evenly distributed across neurons. Is this really necessary? To get some preliminary idea we run our aforementioned three-layer experiment on the first data set,

- once with the traditional weight decay on W (which corresponds to minimizing $\|W\|_F$), and
- once with the $\|W\|_{2,4}$ norm without weight decay.

[16] Recall that our sample complexity in Theorem 2 has the form $\mathfrak{C}(\Phi, \mathfrak{C}(\phi))$ ignoring other parameters. But in fact, one can be more careful on this target function and derive a sample complexity of the form $\mathfrak{C}(\Phi, \mathfrak{C}(\phi_1)) \cdot \mathfrak{C}(\phi_2)$, where $\Phi(x) = (x-2)^2$, $\phi_1(x) = \sin(3x)$ and $\phi_2(x) = \cos(7x)$. Because of this, to show the best contrast between two and three-layer networks, we choose constant 7 on ϕ_2 so that $\mathfrak{C}(\phi_2) \approx \mathfrak{C}(\phi_1)^2$.



(a)

(b)

Figure 3: Empirical comparison between regularizers $\|W\|_F$ and $\|W\|_{2,4}$.

In both cases we best tune the learning rates as well as the weight decay parameter (or the weight of the $\|W\|_{2,4}^4$ regularizer). We present our findings in Figure 3.

In Figure 3(a), we observe that there is no real difference between the two regularizers in terms of *test error*. In one case ($m = 200$) using $\|W\|_{2,4}$ even gives slightly better test accuracy (but we do not want to claim this as a general phenomenon).

More importantly, since by Cauchy-Schwarz we have $\|W\|_F \geq \|W\|_{2,4} \geq m^{-1/4}\|W\|_F$. Let us consider the following norm ratio:

$$ratio \stackrel{\text{def}}{=} \frac{\|W\|_{2,4}^4 \cdot m}{\|W\|_F^4} \in [1, m]$$

If the *ratio* is close to 1, then one can argue that the more *evenly distributed* the rows (i.e., neurons) of $W \in \mathbb{R}^{m \times d}$ are. In contrast, if *ratio* is close to m then the weights in W are concentrated to a single row. (Of course, we cannot expect *ratio* to be really close to 1 since there may not exist such solutions with good accuracy to begin with.)

In Figure 3(b), we see that *ratio* is roughly the same between two types of regularizers. This means, even if we regularize only the Frobenius norm, weights are still sufficiently distributed across neurons comparing to what we can do by regularizing $\|W\|_{2,4}$. We leave it a future work to study why SGD encourages such implicit regularization.

Appendix

We provide technical preliminaries in Appendix A, give our two-layer proofs in Appendix B, and three-layer proofs in Appendix C.

A Technical Preliminaries

Wasserstein distance. The ℓ_2 Wasserstein distance between random variables A, B is

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

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).

Slightly abusing notation, in this paper, we say a random variable X satisfies $|X| \leq B$ with high probability if (1) $|X| \leq B$ w.h.p. and (2) $\mathcal{W}_2(X,0) \leq B$. For instance, if $g = \mathcal{N}(0,1/m)$, then $|g| \leq \widetilde{O}(1/\sqrt{m})$ with high probability.

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 i.i.d. samples from some distribution, where within a 4-tuples:*

- *the marginal distribution of $a_{1,i}$ and $a_{2,i}$ is standard Gaussian $\mathcal{N}(0,1)$;*
- *n_i and α_i are not necessarily independent;*
- *$a_{1,i}$ and $a_{2,i}$ are independent; and*
- *n_i and α_i are independent of $a_{1,i}$ and $a_{2,i}$.*

Suppose $h\colon \mathbb{R} \to [-L, L]$ is a fixed function. Then, for every $B \geq 1$:

$$\mathbf{Pr}\left[\left|\left(\sum_{i\in[m]} a_{1,i}a_{2,i}\mathbb{I}[n_i \geq 0]h(\alpha_i)\right)\right| \geq BL(\sqrt{m}+B)\right] \leq 4e^{-B^2/8}$$

and

$$\mathbf{Pr}\left[\left|\left(\sum_{i\in[m]} a_{1,i}^2\mathbb{I}[n_i \geq 0]h(\alpha_i)\right) - m\mathbb{E}[a_{1,1}^2\mathbb{I}[n_1 \geq 0]h(\alpha_1)]\right| \geq BL(\sqrt{m}+B)\right] \leq 4e^{-B^2/8}.$$

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 [38]) we have that

$$\mathbf{Pr}\left[\left|\left(\sum_{i\in[m]} a_{1,i}a_{2,i}\mathbb{I}[n_i \geq 0]h(\alpha_i)\right)\right| \geq BL(\sqrt{m}+B)\middle|\{n_i,\alpha_i\}_{i\in[m]}\right] \leq 4e^{-B^2/8}.$$

Since this holds for every choice of $\{n_i,\alpha_i\}_{i\in[m]}$ we can complete the proof. The second inequality follows from sub-exponential concentration bounds. □

Proposition A.2. *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:

$$\begin{aligned}\mathbf{Pr}[X_1 = x_1, X_3 = x_3 \mid X_2 = x_2] &= \mathbf{Pr}[X_1 = x_1 \mid X_2 = x_2]\,\mathbf{Pr}[X_3 = x_3 \mid X_2 = x_2]\\ &= \mathbf{Pr}[X_1 = x_1]\,\mathbf{Pr}[X_3 = x_3 \mid X_2 = x_2].\end{aligned}$$

Multiplying $\mathbf{Pr}[X_2 = x_2]$ on both side leads to:

$$\mathbf{Pr}[X_1 = x_1, X_3 = x_3, X_2 = x_2] = \mathbf{Pr}[X_1 = x_1]\,\mathbf{Pr}[X_3 = x_3, X_2 = x_2].$$

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. □

A.2 Central Limit Theorem

The following Wasserstein distance bound of central limit theorem is easy to derive from known results:

Lemma A.3 (new CLT). *Let $X_1, \cdots, X_m \in \mathbb{R}$ be m independent zero-mean random variables with each $|X_i| \le C$, $\sum_{i\in[m]} \mathbb{E}[X_i^2] = V$, then there exists $Z \sim \mathcal{N}(0, V)$ such that*

$$\mathcal{W}_2\left(\textstyle\sum_{i=1}^m X_i, Z\right) = O\left(C \log m\right) \ .$$

Proof. Define $\Sigma_i = \mathbb{E}[X_i^2]$ and without loss of generality assume $\Sigma_1 \ge \Sigma_2 \ge \cdots \ge \Sigma_m$. Let us apply [53, Lemma 1.6] on each i, then for every t such that $t \ge \frac{5C^2}{\Sigma_i}$, let $Y_t \ \sim \mathcal{N}(0, t\Sigma_i)$ be independent of X_i, we have

$$\mathcal{W}_2(Y_t, Y_{t-1} + X_i) \le \frac{5C}{t} \ .$$

In other words—by choosing $t = R_i/\Sigma_i$— we have for every $R_i \ge 5C^2$, letting $Z_i \sim \mathcal{N}(0, R_i)$, $Z_{i+1} \sim \mathcal{N}(0, R_i + \Sigma_i)$ be independent gaussian (also independent of X_i), it satisfies

$$\mathcal{W}_2(Z_{i+1}, Z_i + X_i) \le \frac{5C\Sigma_i}{R_i} \ .$$

Repeatedly applying the above inequality and starting with $Z_1 \sim \mathcal{N}(0, 5C^2)$ and choosing $R_i = 5C^2 + \sum_{j=1}^{i-1} \Sigma_j$, we have $Z_i \sim \mathcal{N}(0, 5C^2 + \sum_{j=1}^{i-1} \Sigma_j)$ and $Z_{m+1} \sim \mathcal{N}(0, 5C^2 + V)$. Using $\Sigma_1 \ge \Sigma_2 \ge \cdots \ge \Sigma_m$, we know that $\frac{5C\Sigma_i}{R_i} \le \frac{5C}{i-1}$ for $i \ge 2$. This implies that

$$\mathcal{W}_2\left(Z_{m+1}, \sum_{i=1}^m X_i\right) \le \sum_{i=2}^m \frac{5C}{i-1} + \mathcal{W}_2\left(Z_2, Z_1 + X_1\right) = O\left(C \log m\right) + \mathcal{W}_2\left(Z_2, Z_1 + X_1\right)$$

Finally, since $|X_1| \le C$ we have $\mathcal{W}_2(X_1, 0) \le C$ and $\Sigma_1^2 \le C^2$. By triangle inequality we have

$$\mathcal{W}_2\left(Z_2, Z_1 + X_1\right) \le \mathcal{W}_2(X_1, 0) + \mathcal{W}_2(Z_1, 0) + \mathcal{W}_2(Z_2, 0) = O(C)$$

There exists $Z \sim \mathcal{N}(0, V)$ such that $\mathcal{W}_2(Z_{m+1}, Z) = O(C)$. All of these together imply

$$\mathcal{W}_2\left(Z, \sum_{i=1}^m X_i\right) = O\left(C \log m\right) \ .$$

□

A.3 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 $I(y) \subset [-2, 2]$ 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). $\max_{x\in I(y)}\{\mathfrak{s}(y,x)x\} - \min_{x\in I(y)}\{\mathfrak{s}(y,x)x\} \leq 10\tau$.

6. (Lipschitz). $|I(y_1)\triangle I(y_2)| \leq O(|y_2 - y_1|)$ *, 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 $y \geq 0$ 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 $a(y), b(y) \geq 0$ 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
$$\mathfrak{s}(y,g) \stackrel{\text{def}}{=} \begin{cases} 1 & \text{if } g \in [y - a(y), y + b(y)]; \\ -1 & \text{if } g \in [-y - b(y), -y + a(y)]; \\ 0 & \text{otherwise.} \end{cases}$$

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
$$\mathbb{E}_{g\sim\mathcal{N}(0,1)}[g \mid g \in I_2(y) \ \wedge \ g > 0] = \mathbb{E}_{g\sim\mathcal{N}(0,1)}[|g| \mid g \in I_2(y)] = y \enspace .$$
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
$$\mathop{\mathbf{Pr}}_{g\sim\mathcal{N}(0,1)}[g \in [e(y), c(y)]] = \frac{\tau - \tau'(y)}{2} \enspace .$$
Let $d(y) \in [e(y), c(y)]$ be the unique real such that
$$\mathop{\mathbf{Pr}}_{g\sim\mathcal{N}(0,1)}[g \in [e(y), d(y)]] = \mathop{\mathbf{Pr}}_{g\sim\mathcal{N}(0,1)}[g \in [d(y), c(y)]] \enspace .$$
Finally, we define $I(y) = [-c(y), c(y)]$ and
$$\mathfrak{s}(y,g) \stackrel{\text{def}}{=} \begin{cases} 1 & \text{if } g \in [0, e(y)] \cup [e(y), d(y)] \cup [-d(y), -e(y)]; \\ -1 & \text{if } g \in [-e(y), 0] \cup [d(y), c(y)] \cup [-c(y), -d(y)]; \\ 0 & \text{otherwise.} \end{cases}$$

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.4 Hermite polynomials

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

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

satisfying the orthogonality constraint

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

where $\delta_{i,j} = 1$ if $i = j$ and $\delta_{i,j} = 0$ otherwise. They have the following summation and multiplication formulas.

$$h_i(x+y) = \sum_{k=0}^{i} \binom{i}{k} x^{i-k} h_k(y),$$

$$h_i(\gamma x) = \sum_{k=0}^{\lfloor \frac{i}{2} \rfloor} \gamma^{i-2k} (\gamma^2 - 1)^k \binom{i}{2k} \frac{(2k)!}{k!} 2^{-k} h_{i-2k}(x).$$

Lemma A.6.

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

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

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

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

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

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

(Throughout this paper, the binomial $\binom{n}{m}$ is defined as $\frac{\Gamma(n+1)}{\Gamma(m+1)\Gamma(n+1-m)}$, and this allows us to write for instance $\binom{5/2}{-1/2}$ without notation change.)

Proof. Using the summation formula of Hermite polynomial, we have:

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

Using the multiplication formula of Hermite polynomial, we have:

$$h_k\left(\beta\sqrt{1-x_1^2}\right) = \sum_{j=0}^{\lfloor \frac{k}{2}\rfloor}\left(\sqrt{1-x_1^2}\right)^{k-2j}\left(-x_1^2\right)^j\binom{k}{2j}\frac{(2j)!}{j!}2^{-j}h_{k-2j}(\beta).$$

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

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

and for odd k,

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

This implies

$$\begin{aligned}\mathbb{E}_{\beta\sim\mathcal{N}(0,1)}\left[h_i\left(\alpha x_1 + \beta\sqrt{1-x_1^2}\right)\right] &= \sum_{k=0,k\text{ even}}^{i}\binom{i}{k}(\alpha x_1)^{i-k}\left(-x_1^2\right)^{k/2}\frac{k!}{(k/2)!}2^{-k/2}\\ &= x_1^i\sum_{k=0,k\text{ even}}^{i}\binom{i}{k}\alpha^{i-k}\frac{k!}{(k/2)!}(-2)^{-k/2}.\end{aligned}$$

Therefore,

$$\begin{aligned}&\mathbb{E}_{\alpha,\beta\sim\mathcal{N}(0,1)}\left[h_i\left(\alpha x_1 + \beta\sqrt{1-x_1^2}\right)\mathbb{I}[\alpha\geq b]\right]\\ &= x_1^i\sum_{k=0,k\text{ even}}^{i}\binom{i}{k}\mathbb{E}_{\alpha\sim\mathcal{N}(0,1)}[\alpha^{i-k}\mathbb{I}[\alpha\geq b]]\frac{k!}{(k/2)!}(-2)^{-k/2}\quad.\end{aligned}\tag{A.1}$$

Define

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

(a) Consider even $i > 0$. By Lemma A.7, we have for even $i \geq 0$:

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

So

$$
\begin{aligned}
&\mathbb{E}_{\alpha,\beta\sim\mathcal{N}(0,1)}\left[h_i\left(\alpha x_1+\beta\sqrt{1-x_1^2}\right)\mathbb{I}[\alpha\geq b]\right]\\
&= x_1^i\left(\sum_{k=0,k\text{ even}}^{i}\binom{i}{k}L_{i-k,b}\frac{k!}{(k/2)!}(-2)^{-k/2}\right)\\
&= x_1^i\left(\sum_{k=0,k\text{ even}}^{i}\binom{i}{k}(i-k-1)!!\Phi(0,1;b)\frac{k!}{(k/2)!}(-2)^{-k/2}\right)\\
&\quad+x_1^i\phi(0,1;b)\left(\sum_{k=0,k\text{ even}}^{i}\binom{i}{k}\left(\sum_{j=1,j\text{ odd}}^{i-k-1}\frac{(i-k-1)!!}{j!!}b^j\right)\frac{k!}{(k/2)!}(-2)^{-k/2}\right).
\end{aligned}
$$

Since

$$
\begin{aligned}
\sum_{k=0,k\text{ even}}^{i}\binom{i}{k}(i-k-1)!!\frac{k!}{(k/2)!}(-2)^{-k/2} &= \sum_{k=0,k\text{ even}}^{i}\frac{i!(i-k-1)!!}{(i-k)!(k/2)!}\frac{(-1)^{k/2}}{2^{k/2}}\\
&= \sum_{k=0,k\text{ even}}^{i}\frac{i!}{(i-k)!!(k/2)!}\frac{(-1)^{k/2}}{2^{k/2}}\\
&= (i-1)!!\sum_{k=0,k\text{ even}}^{i}\frac{i!!}{(i-k)!!(k/2)!}\frac{(-1)^{k/2}}{2^{k/2}}\\
&= (i-1)!!\sum_{k=0,k\text{ even}}^{i}\binom{i/2}{k/2}(-1)^{k/2}\\
&= 0,
\end{aligned}
$$

we know that

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

where c_r is given by:

$$\begin{aligned}
c_r &:= \frac{1}{(i-1)!!} \sum_{k=0,k \text{ even}}^{i-1-r} \binom{i}{k} \frac{(i-k-1)!!}{r!!} \frac{k!}{(k/2)!}(-2)^{-k/2} \\
&= \sum_{k=0,k \text{ even}}^{i-1-r} \binom{i/2}{k/2} \frac{(-1)^{k/2}}{r!!} \\
&= \sum_{j=0}^{(i-1-r)/2} \binom{i/2}{j} \frac{(-1)^{j}}{r!!} \\
&= \sum_{j=0}^{(i-1-r)/2} \binom{j-i/2-1}{j} \frac{1}{r!!} \\
&= \frac{1}{r!!}\binom{-i/2+(i-1-r)/2}{(i-1-r)/2} \\
&= \frac{(-1)^{(i-1-r)/2}}{r!!}\binom{i/2-1}{(i-1-r)/2} \\
&= \frac{(-1)^{\frac{i-1-r}{2}}}{r!!}\binom{i/2-1}{(r-1)/2}.
\end{aligned}$$

(b) Consider odd $i > 0$. By Lemma A.7, we have for odd $i > 0$:

$$L_{i,b} = \phi(0,1;b) \sum_{j=0,j \text{ even}}^{i-1} \frac{(i-1)!!}{j!!} b^j.$$

So

$$\begin{aligned}
&\mathbb{E}_{\alpha,\beta\sim\mathcal{N}(0,1)}\left[h_i\left(\alpha x_1 + \beta\sqrt{1-x_1^2}\right)\mathbb{I}[\alpha \geq b]\right] \\
&= x_1^i \left(\sum_{k=0,k \text{ even}}^{i} \binom{i}{k} L_{i-k,b} \frac{k!}{(k/2)!}(-2)^{-k/2}\right) \\
&= x_1^i \phi(0,1;b) \left(\sum_{k=0,k \text{ even}}^{i} \binom{i}{k} \left(\sum_{j=0,j \text{ even}}^{i-k-1} \frac{(i-k-1)!!}{j!!} b^j\right) \frac{k!}{(k/2)!}(-2)^{-k/2}\right) \\
&= x_1^i (i-1)!! \phi(0,1;b) \sum_{r=0,r \text{ even}}^{i-1} c_r b^r
\end{aligned}$$

where c_r is given by:

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

by a similar calculation as in the even i case.

The proof is completed. □

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

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

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

$$\begin{aligned}
L_{0,b} &= \Phi(0,1;b) := \mathop{\mathbf{Pr}}_{\alpha \sim \mathcal{N}(0,1)}[\alpha \geq b], \\
L_{1,b} &= \phi(0,1;b) := \mathbb{E}_{\alpha \sim \mathcal{N}(0,1)}[\alpha \mathbb{I}[\alpha \geq b]] = \frac{\exp(-b^2/2)}{\sqrt{2\pi}}, \\
L_{i,b} &= b^{i-1}\phi(0,1;b) + (i-1)L_{i-2,b}.
\end{aligned}$$

As a result (with the convention that $0!! = 1$ and $(-1)!! = 1$)

$$\text{for even } i \geq 0: \qquad L_{i,b} = (i-1)!!\Phi(0,1;b) + \phi(0,1;b) \sum_{j=1, j \text{ odd}}^{i-1} \frac{(i-1)!!}{j!!} b^j$$

$$\text{for odd } i > 0: \qquad L_{i,b} = \phi(0,1;b) \sum_{j=0, j \text{ even}}^{i-1} \frac{(i-1)!!}{j!!} b^j \quad .$$

One can verify that for $b \geq 0$,

$$L_{i,b} \leq O(1)e^{-b^2/2} \cdot \sum_{j=0}^{i-1} \frac{(i-1)!!}{j!!} b^j$$

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

A.5 Optimization

Fact A.8. *For every B-second-order smooth function $f : \mathbb{R}^d \to \mathbb{R}$, every $\varepsilon > 0$, every $\eta \in \big(0, O(\frac{\varepsilon^2}{B^2})\big]$, every fixed vector $x \in \mathbb{R}^d$, suppose there is a random vector $x_2 \in \mathbb{R}^d$ with $\mathbb{E}[x_2] = 0$ and $\|x_2\|_2 = 1$ satisfying*

$$\mathbb{E}_{x_2}[f\left(x + \sqrt{\eta}x_2\right)] \leq f(x) - \eta\varepsilon \quad .$$

Then, $\lambda_{\min}(\nabla^2 f(x)) \leq -\varepsilon$, where $\lambda_{\min}$ is the minimal eigenvalue.

Proof of Fact A.8. We know that

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

Taking expectation, we know that

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

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

We also recall the following convergence theorem of SGD for escaping saddle point.[17]

[17] The original proof of [20] was for constant probability but it is easy to change it to $1-\delta$ at the expense of paying a polynomial factor in $1/\delta$. The original proof did not state the objective "non-increasing" guarantee but it is easy to show it for mini-batch SGD. Recall $\eta = \mathsf{poly}(d, B, 1/\varepsilon, 1/\delta)$ in [20] so in each iteration, the original SGD may increase the objective by no more than $\frac{\eta^2 B}{b}$ if a batch size b is used. If b is polynomially large, this goal is easily achievable. In practice, however, using a very small batch size suffices.

Lemma A.9 (escape saddle points, Theorem 6 of [20]). *Suppose a function $f : \mathbb{R}^d \to \mathbb{R}$ has its stochastic gradient bounded by B in Euclidean norm, is absolutely bounded $|f(x)| \leq B$, is B-smooth, and is B-second-order smooth, then for every $\delta > 0$, every $p \in (0,1)$, with probability at least $1-p$, noisy SGD outputs a point x_T after $T = \mathsf{poly}(d, B, 1/\delta, 1/p)$ iterations such that*

$$\nabla^2 f(x_T) \succeq -\delta \mathbf{I} \quad \textit{and} \quad f(x_T) \leq f(x_0) + \delta \cdot \mathsf{poly}(d, B, 1/p)$$

A.6 Rademacher Complexity

Let $\mathcal{F}$ be a set of functions $\mathbb{R}^d \to \mathbb{R}$ and $\mathcal{X} = (x_1, \dots, x_N)$ be a finite set of samples. Recall the *empirical Rademacher complexity* with respect to $\mathcal{X}$ of $\mathcal{F}$ is

$$\widehat{\mathfrak{R}}(\mathcal{X}; \mathcal{F}) \stackrel{\text{def}}{=} \mathbb{E}_{\xi \sim \{\pm 1\}^N} \Big[\sup_{f \in \mathcal{F}} \frac{1}{N} \sum_{i=1}^{N} \xi_i f(x_i) \Big]$$

Lemma A.10 (Rademacher generalization). *Suppose $\mathcal{X} = (x_1, \dots, x_N)$ where each x_i is generated i.i.d. from a distribution $\mathcal{D}$. If every $f \in \mathcal{F}$ satisfies $|f| \leq b$, for every $\delta \in (0,1)$ with probability at least $1-\delta$ over the randomness of $\mathcal{Z}$, it satisfies*

$$\sup_{f \in \mathcal{F}} \left| \mathbb{E}_{x \sim \mathcal{D}}[f(x)] - \frac{1}{N} \sum_{i=1}^{N} f(x_i) \right| \leq 2\widehat{\mathfrak{R}}(\mathcal{Z}; \mathcal{F}) + O\left(\frac{b\sqrt{\log(1/\delta)}}{\sqrt{N}} \right) \ .$$

Corollary A.11. *If $\mathcal{F}_1, \dots, \mathcal{F}_k$ are k classes of functions $\mathbb{R}^d \to \mathbb{R}$ and $L_x \colon \mathbb{R}^k \to [-b, b]$ is a 1-Lipschitz continuous function for any $x \sim \mathcal{D}$, then*

$$\sup_{f_1 \in \mathcal{F}_1, \dots, f_k \in \mathcal{F}_k} \left| \mathbb{E}_{x \sim \mathcal{D}}[L_x(f_1(x), \dots, f_k(x))] - \frac{1}{N} \sum_{i=1}^{N} L_x(f(x_i)) \right| \leq O\Big(\sum_{r=1}^{k} \widehat{\mathfrak{R}}(\mathcal{Z}; \mathcal{F}_r) \Big) + O\left(\frac{b\sqrt{\log(1/\delta)}}{\sqrt{N}} \right) \ .$$

Proof. Let $\mathcal{F}'$ be the class of functions by composing L with $\mathcal{F}_1, \dots, \mathcal{F}_k$, that is, $\mathcal{F}' = \{L_x \circ (f_1, \dots, f_k) \mid f_1 \in \mathcal{F}_1 \cdots f_k \in \mathcal{F}_k\}$. By the (vector version) of the contraction lemma of Rademacher complexity[18] it satisfies $\widehat{\mathfrak{R}}(\mathcal{Z}; \mathcal{F}') \leq O(1) \cdot \sum_{r=1}^{k} \widehat{\mathfrak{R}}(\mathcal{Z}; \mathcal{F}_r)$. □

Proposition A.12. *We recall some basic properties of the Rademacher complexity. Let $\sigma \colon \mathbb{R} \to \mathbb{R}$ be a fixed 1-Lipschitz function.*

(a) (ℓ_2 linear) Suppose $\|x\|_2 \leq 1$ for all $x \in \mathcal{X}$. The class $\mathcal{F} = \{x \mapsto \langle w, x \rangle \mid \|w\|_2 \leq B\}$ has Rademacher complexity $\widehat{R}(\mathcal{X}; \mathcal{F}) \leq O(\frac{B}{\sqrt{N}})$.

(b) (ℓ_1 linear) Suppose $\|x\|_\infty \leq 1$ for all $x \in \mathcal{X} \subseteq \mathbb{R}^m$. The class $\mathcal{F} = \{x \mapsto \langle w, x \rangle \mid \|w\|_1 \leq B\}$ has Rademacher complexity $\widehat{R}(\mathcal{X}; \mathcal{F}) \leq O(\frac{B\sqrt{\log m}}{\sqrt{N}})$.

(c) (addition) $\widehat{R}(\mathcal{X}; \mathcal{F}_1 + \mathcal{F}_2) = \widehat{R}(\mathcal{X}; \mathcal{F}_1) + \widehat{R}(\mathcal{X}; \mathcal{F}_2)$.

(d) (contraction) $\widehat{R}(\mathcal{X}; \sigma \circ \mathcal{F}) \leq \widehat{R}(\mathcal{X}; \mathcal{F})$.

(e) Given $\mathcal{F}_1, \dots, \mathcal{F}_m$ classes of functions $\mathcal{X} \to \mathbb{R}$ and suppose $w \in \mathbb{R}^m$ is a fixed vector, then

$$\mathcal{F}' = \Big\{ x \mapsto \textstyle\sum_{j=1}^{m} w_j \sigma(f_j(x)) \,\Big|\, f_j \in \mathcal{F}_j \Big\}$$

satisfies $\widehat{R}(\mathcal{X}; \mathcal{F}') \leq 2\|w\|_1 \max_{j \in [m]} \widehat{R}(\mathcal{X}; \mathcal{F}_j)$.

[18] There are slightly different versions of the contraction lemma in the literature. For the scalar case without absolute value, see [36, Section 3.8]; for the scalar case with absolute value, see [10, Theorem 12]; and for the vector case without absolute value, see [39].

(f) Given $\mathcal{F}_1, \dots, \mathcal{F}_m$ classes of functions $\mathcal{X} \to \mathbb{R}$ and suppose for each $j \in [m]$ there exist a function $f_j^{(0)} \in \mathcal{F}_j$ satisfying $\sup_{x\in\mathcal{X}} |\sigma(f_j^{(0)}(x))| \le R$, then

$$\mathcal{F}' = \Big\{ x \mapsto \textstyle\sum_{j=1}^m v_j \sigma(f_j(x)) \,\Big|\, f_j \in \mathcal{F}_j \wedge \|v\|_1 \le B \wedge \|v\|_\infty \le D \Big\}$$

satisfies $\widehat{R}(\mathcal{X};\mathcal{F}') \le 2D \sum_{j\in[m]} \widehat{R}(\mathcal{X};\mathcal{F}_j) + O\big(\frac{BR\log m}{\sqrt{N}}\big)$.

Proof. The first three are trivial, and the contraction lemma is classical (see for instance [46]). The derivations of Lemma A.12e and Lemma A.12f are less standard.

For Lemma A.12e, let us choose an arbitrary $f_j^{(0)}$ from each $\mathcal{F}_j$. We write $f \in \mathcal{F}$ to denote $(f_1, \dots, f_m) \in \mathcal{F}_1 \times \cdots \times \mathcal{F}_m$.

$$\begin{aligned}
\mathbb{E}_\xi\Big[\sup_{f\in\mathcal{F}} \sum_{i\in[N]} \xi_i \sum_{j=1}^m w_j \sigma(f_j(x_i))\Big] &\overset{\text{①}}{=} \mathbb{E}_\xi\Big[\sup_{f\in\mathcal{F}} \sum_{i\in[N]} \xi_i \sum_{j=1}^m w_j \big(\sigma(f_j(x_i)) - \sigma(f_j^{(0)}(x_i))\big)\Big] \\
&\le \mathbb{E}_\xi\Big[\sup_{f\in\mathcal{F}} \sum_{j=1}^m |w_j| \Big| \sum_{i\in[N]} \xi_i \big(\sigma(f_j(x_i)) - \sigma(f_j^{(0)}(x_i))\big)\Big|\Big] \\
&= \mathbb{E}_\xi\Big[\sum_{j=1}^m |w_j| \sup_{f_j\in\mathcal{F}_j} \Big| \sum_{i\in[N]} \xi_i \big(\sigma(f_j(x_i)) - \sigma(f_j^{(0)}(x_i))\big)\Big|\Big] \\
&\overset{\text{②}}{\le} 2\mathbb{E}_\xi\Big[\sum_{j=1}^m |w_j| \sup_{f_j\in\mathcal{F}_j} \sum_{i\in[N]} \xi_i \big(\sigma(f_j(x_i)) - \sigma(f_j^{(0)}(x_i))\big)\Big] \\
&\overset{\text{③}}{=} 2\mathbb{E}_\xi\Big[\sum_{j=1}^m |w_j| \sup_{f_j\in\mathcal{F}_j} \sum_{i\in[N]} \xi_i \sigma(f_j(x_i))\Big] \\
&\le 2\|w\|_1 \cdot N \max_{j\in[m]} \widehat{\mathfrak{R}}(\mathcal{X};\sigma\circ\mathcal{F}_j) \overset{\text{④}}{\le} 2N\|w\|_1 \cdot \max_{j\in[m]} \widehat{\mathfrak{R}}(\mathcal{X};\mathcal{F}_j) \enspace .
\end{aligned}$$

Above, ① is because $\xi_i w_j \sigma(f_j^{(0)}(x_i))$ is independent of f_j and thus zero in expectation; ② uses the non-negativity of $\sup_{f_j\in\mathcal{F}_j} \sum_{i\in[N]} \xi_i \big(\sigma(f_j(x_i)) - \sigma(f_j^{(0)}(x_i))\big)$; ③ is for the same reason as ①; and ④ is by Proposition A.12d.

For Lemma A.12f,

$$\begin{aligned}
&\mathbb{E}_\xi\Big[\sup_{f\in\mathcal{F},\|v\|_1\leq B,\|v\|_\infty\leq D}\sum_{i\in[N]}\xi_i\sum_{j=1}^m v_j\sigma(f_j(x_i))\Big]\\
&=\mathbb{E}_\xi\Big[\sup_{f\in\mathcal{F},\|v\|_1\leq B,\|v\|_\infty\leq D}\sum_{i\in[N]}\xi_i\sum_{j=1}^m v_j\big(\sigma(f_j(x_i))-\sigma(f_j^{(0)}(x_i))\big)+\sum_{i\in[N]}\xi_i\sum_{j=1}^m v_j\sigma(f_j^{(0)}(x_i))\Big]\\
&\leq\mathbb{E}_\xi\Big[\sup_{f\in\mathcal{F},\|v\|_\infty\leq D}\sum_{i\in[N]}\xi_i\sum_{j=1}^m v_j\big(\sigma(f_j(x_i))-\sigma(f_j^{(0)}(x_i))\big)\Big]+\mathbb{E}_\xi\Big[\sup_{\|v\|_1\leq B}\sum_{i\in[N]}\xi_i\sum_{j=1}^m v_j\sigma(f_j^{(0)}(x_i))\Big]\\
&\leq D\mathbb{E}_\xi\Big[\sum_{j\in[m]}\sup_{f_j\in\mathcal{F}_j}\Big|\sum_{i\in[N]}\xi_i\big(\sigma(f_j(x_i))-\sigma(f_j^{(0)}(x_i))\big)\Big|\Big]+\mathbb{E}_\xi\Big[\sup_{\|v\|_1\leq B}\sum_{i\in[N]}\xi_i\sum_{j=1}^m v_j\sigma(f_j^{(0)}(x_i))\Big]\\
&\overset{①}{\leq} 2D\mathbb{E}_\xi\Big[\sum_{j\in[m]}\sup_{f_j\in\mathcal{F}_j}\sum_{i\in[N]}\xi_i\big(\sigma(f_j(x_i))-\sigma(f_j^{(0)}(x_i))\big)\Big]+\mathbb{E}_\xi\Big[\sup_{\|v\|_1\leq B}\sum_{i\in[N]}\xi_i\sum_{j=1}^m v_j\sigma(f_j^{(0)}(x_i))\Big]\\
&=2D\mathbb{E}_\xi\Big[\sum_{j\in[m]}\sup_{f_j\in\mathcal{F}_j}\sum_{i\in[N]}\xi_i\sigma(f_j(x_i))\Big]+\mathbb{E}_\xi\Big[\sup_{\|v\|_1\leq B}\sum_{i\in[N]}\xi_i\sum_{j=1}^m v_j\sigma(f_j^{(0)}(x_i))\Big]\\
&=2D\cdot N\sum_{j\in[m]}\widehat{\mathfrak{R}}(\mathcal{X};\sigma\circ\mathcal{F}_j)+\mathbb{E}_\xi\Big[\sup_{\|v\|_1\leq B}\sum_{i\in[N]}\xi_i\sum_{j=1}^m v_j\sigma(f_j^{(0)}(x_i))\Big]\\
&\overset{②}{=}2D\cdot N\sum_{j\in[m]}\widehat{\mathfrak{R}}(\mathcal{X};\sigma\circ\mathcal{F}_j)+O(BR\sqrt{N}\log m)\ .
\end{aligned}$$

Above, ① is from the same derivation as Proposition A.12e; and ② uses Proposition A.12b by viewing function class $\big\{x\mapsto\sum_{j=1}^m v_j\cdot\sigma(f_j^{(0)}(x))\big\}$ as linear in v. □

B Proofs for Two-Layer Networks

Recall from (3.1) the target $F^*=(f_1^*,\cdots,f_k^*)$ for our two-layer case is

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

We consider another function defined as $G(x;W)=(g_1(x;W),\dots,g_k(x;W))$ for the weight matrix W, where

$$g_r(x;W)\stackrel{\text{def}}{=}\sum_{i=1}^m a_{r,i}^{(0)}(\langle w_i,x\rangle+b_i^{(0)})\mathbb{I}[\langle w_i^{(0)},x\rangle+b_i^{(0)}\geq 0]$$

where w_i is the r-th row of W and $w_i^{(0)}$ is the r-th row of $W^{(0)}$ (our random initialization). We call this G a pseudo network. For convenience, we also define a pseudo network $G^{(b)}(x;W)=(g_1^{(b)}(x;W),\dots,g_k^{(b)}(x;W))$ without bias

$$g_r^{(b)}(x;W)\stackrel{\text{def}}{=}\sum_{i=1}^m a_{r,i}^{(0)}\langle w_i,x\rangle\mathbb{I}[\langle w_i^{(0)},x\rangle+b_i^{(0)}\geq 0],$$

Roadmap. Our analysis begins with showing that w.h.p. over the random initialization, there exists a pseudo network in the neighborhood of the initialization that can approximate the target function (see Section B.1). We then show that, near the initialization, the pseudo network approximates the actual ReLU network F (see Section B.2). Therefore, there exists a ReLU network near the initialization approximating the target. Furthermore, it means that the loss surface of the ReLU network is close to that of the pseudo network, which is convex. This then allows us to show the training converges (see Section B.3). Combined with a generalization bound localized to the initialization, we can prove the final theorem that SGD learns a network with small risk.

The proof is much simpler than the three-layer case. First, the optimization landscape is almost convex, so a standard argument for convex optimization applies. While for three-layer case, the optimization landscape is no longer this nice and needs an escaping-from-saddle-point argument, which in turn requires several technicalities (smoothing, explicit regularization, weight decay, and Dropout-like noise). Second, the main technical part of the analysis, the proof of the existential result, only needs to deal with approximating one layer of neurons in the target, while in the three-layer case, a composition of the neurons needs to be approximating, requiring additional delicate arguments. It is also similar with the generalization bounds. However, we believe that the analysis in the two-layer case already shows some of the key ideas, and suggest the readers to read it before reading that for the three-layer case.

B.1 Existential Result

The main focus of this subsection is to show that there exists a good pseudo network near the initialization. (Combining with the coupling result of the next subsection, this translates to the *real* network near the initialization.)

Lemma B.1. *For every $\varepsilon \in (0, \frac{1}{pk\mathfrak{C}_\mathfrak{s}(\phi,1)})$, letting $\varepsilon_a = \varepsilon/\widetilde{\Theta}(1)$, there exists*

$$M = \mathsf{poly}(\mathfrak{C}_\varepsilon(\phi,1), 1/\varepsilon) \quad \textit{and} \quad C_0 = \widetilde{\Theta}\big(\mathfrak{C}_\varepsilon(\phi,1)\big)$$

such that if $m \geq M$, then with high probability there exists $W^{*} = (w_1^{*}, \dots, w_m^{*})$ with $\|W^{*}\|_{2,\infty} \leq \frac{kpC_0}{\varepsilon_a m}$ and $\|W^{*}\|_F \leq \widetilde{O}(\frac{kp\mathfrak{C}_\mathfrak{s}(\phi,1)}{\varepsilon_a\sqrt{m}})$ and

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

and consequently,

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

Corollary B.2. *In the same setting as Lemma B.1, we have that w.h.p.*

$$\mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\sum_{r=1}^{k}\left|f_r^*(x) - g_r(x; W^{(0)} + W^{*})\right|\right] \leq \varepsilon,$$

and consequently,

$$\mathbb{E}_{(x,y)\sim\mathcal{D}}\left[L(G(x; W^{(0)} + W^{*}), y)\right] \leq \mathsf{OPT} + \varepsilon.$$

Proof of Lemma B.1. Recall the pseudo network without bias is given by

$$g_r^{(b)}(x; W) = \sum_{i=1}^{m} a_{r,i}^{(0)}(\langle w_i, x\rangle + b_i^{(0)})\mathbb{I}[\langle w_i^{(0)}, x\rangle + b_i^{(0)} \geq 0].$$

Also recall from Lemma 6.2 that, for each $i \in [p]$, there is function $h^{(i)} \colon \mathbb{R}^2 \to \mathbb{R}$ with $|h^{(i)}| \leq \mathfrak{C}_\varepsilon(\phi, 1)$, satisfying

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

where $\alpha_1, \beta_1, b_0 \sim \mathcal{N}(0,1)$ are independent random Gaussians.

Fit a single function $a^*_{r,i}\phi_i(\langle w^*_{1,i}, x\rangle)\langle w^*_{2,i}, x\rangle$. We first fix some $r \in [k]$ and $i \in [p]$ and construct weights $w^\divideontimes_j \in \mathbb{R}^d$. Define

$$w^\divideontimes_j \stackrel{\text{def}}{=} \frac{1}{\varepsilon_a^2} a^{(0)}_{r,j} a^*_{r,i} h^{(i)}\left(\sqrt{m}\langle w^{(0)}_j, w^*_{1,i}\rangle, \sqrt{m} b^{(0)}_j \right) w^*_{2,i} \tag{B.1}$$

where $\sqrt{m}\langle w^{(0)}_j, w^*_{1,i}\rangle, \sqrt{m} b^{(0)}_j$ has the same distribution with α_1, b_0 in Lemma 6.2. By Lemma 6.2, we have that

$$\begin{aligned}
&\mathbb{E}_{w^{(0)}_j, b^{(0)}_j, a^{(0)}_{r,j}}\left[a^{(0)}_{r,j} \mathbb{I}_{\langle w^{(0)}_j, x\rangle + b^{(0)}_j \geq 0} \langle w^\divideontimes_j, x\rangle \right] \\
&= \mathbb{E}_{w^{(0)}_j, b^{(0)}_j}\left[a^*_{r,i} \mathbb{I}_{\langle w^{(0)}_j, x\rangle + b^{(0)}_j \geq 0} h^{(i)}\left(\sqrt{m}\langle w^{(0)}_j, w^*_{1,i}\rangle, \sqrt{m} b^{(0)}_j \right) \langle w^*_{2,i}, x\rangle \right] \\
&= a^*_{r,i}\phi_i(\langle w^*_{1,i}, x\rangle)\langle w^*_{2,i}, x\rangle \pm \varepsilon .
\end{aligned}$$

Fit a combination $\sum_{i\in[p]} a^*_{r,i}\phi_i(\langle w^*_{1,i}, x\rangle)\langle w^*_{2,i}, x\rangle$. We can re-define (the norm grows by a maximum factor of p)

$$w^\divideontimes_j = \frac{1}{\varepsilon_a^2} a^{(0)}_{r,j} \sum_{i\in[p]} a^*_{r,i} h^{(i)}\left(\sqrt{m}\langle w^{(0)}_j, w^*_{1,i}\rangle, \sqrt{m} b^{(0)}_j \right) w^*_{2,i}$$

and the same above argument gives

$$\mathbb{E}_{w^{(0)}_j, b^{(0)}_j, a^{(0)}_{r,j}}\left[a^{(0)}_{r,j} \mathbb{I}_{\langle w^{(0)}_j, x\rangle + b^{(0)}_j \geq 0} \langle w^\divideontimes_j, x\rangle \right] = \sum_{i\in[p]} a^*_{r,i}\phi_i(\langle w^*_{1,i}, x\rangle)\langle w^*_{2,i}, x\rangle \pm \varepsilon p .$$

Fit multiple outputs. If there are k outputs let us re-define (the norm grows by a maximum factor of k)

$$w^\divideontimes_j = \frac{1}{\varepsilon_a^2} \sum_{r\in[k]} a^{(0)}_{r,j} \sum_{i\in[p]} a^*_{r,i} h^{(i)}\left(\sqrt{m}\langle w^{(0)}_j, w^*_{1,i}\rangle, \sqrt{m} b^{(0)}_j \right) w^*_{2,i} . \tag{B.2}$$

and consider the quantity

$$\Xi_{r,j} \stackrel{\text{def}}{=} a^{(0)}_{r,j} \langle w^\divideontimes_j, x\rangle \mathbb{I}[\langle w^{(0)}_j, x\rangle + b^{(0)}_j \geq 0] \ .$$

By definition of the initialization, we know that for $r' \neq r$, $\mathbb{E}[a^{(0)}_{r,j} a^{(0)}_{r',j}] = 0$. Thus, for every $r \in [k]$, it satisfies

$$
\begin{aligned}
&\mathbb{E}_{w_j^{(0)}, b_j^{(0)}, a_{1,j}^{(0)}, \dots, a_{k,j}^{(0)}} \left[\Xi_{r,j} \right] \\
&= \mathbb{E}_{w_j^{(0)}, b_j^{(0)}, a_{1,j}^{(0)}, \dots, a_{k,j}^{(0)}} \left[\sum_{r' \in [k]} \frac{a_{r,j}^{(0)} a_{r',j}^{(0)}}{\varepsilon_a^2} \sum_{i \in [p]} \mathbb{I}_{\langle w_j^{(0)}, x \rangle + b_j^{(0)} \geq 0} a_{r',i}^* h^{(i)} \left(\sqrt{m} \langle w_j^{(0)}, w_{1,i}^* \rangle, \sqrt{m} b_j^{(0)} \right) \langle w_{2,i}^*, x \rangle \right] \\
&= \mathbb{E}_{w_j^{(0)}, b_j^{(0)}} \left[\sum_{i \in [p]} \mathbb{I}_{\langle w_j^{(0)}, x \rangle + b_j^{(0)} \geq 0} a_{r,i}^* h^{(i)} \left(\sqrt{m} \langle w_j^{(0)}, w_{1,i}^* \rangle, \sqrt{m} b_j^{(0)} \right) \langle w_{2,i}^*, x \rangle \right] \\
&= \sum_{i \in [p]} a_{r,i}^* \phi_i(\langle w_{1,i}^*, x \rangle)) \langle w_{2,i}^*, x \rangle \pm p\varepsilon = f_r^*(x) \pm p\varepsilon \quad .
\end{aligned}
$$

Now, re-scaling each w_j^* by a factor of $\frac{1}{m}$ and re-scaling ε by $\frac{1}{2pk}$, we can write

$$
g_r^{(b)}(x; W^*) = \sum_{j=1}^{m} \Xi_{r,j} \quad \text{and} \quad \mathbb{E}\left[g_r^{(b)}(x; W^*) \right] = f_r^*(x) \pm \frac{\varepsilon}{2k} \quad .
$$

Now, we apply the concentration from Lemma A.1, which implies for our parameter choice of m, with high probability

$$
|g_r^{(b)}(x; W^*) - f_r^*(x)| \leq \frac{\varepsilon}{k} \quad .
$$

The above concentration holds for every fixed x with high probability, and thus also holds in expectation with respect to $(x, y) \sim \mathcal{D}$. This proves the first statement. As for the second statement on $L(G^{(b)}(x; W^*), y)$, it follows from the Lipschitz continuity of L.

Norm on W^*. According to its definition in (B.2), we have for each $j \in [m]$, with high probability $\|w_j^*\|_2 \leq \widetilde{O}(\frac{kpC_0}{\varepsilon_a m})$ (here the additional $\frac{1}{m}$ is because we have re-scaled w_j^* by $\frac{1}{m}$). This means $\|W^*\|_{2,\infty} \leq \widetilde{O}(\frac{kpC_0}{\varepsilon_a m})$. As for the Frobenius norm,

$$
\|W^*\|_F^2 = \sum_{j \in [m]} \|w_j^*\|^2 \leq \sum_{j \in [m]} \widetilde{O}(\frac{k^2 p}{\varepsilon_a^2 m^2}) \cdot \sum_{i \in [p]} h^{(i)} \left(\sqrt{m} \langle w_j^{(0)}, w_{1,i}^* \rangle, \sqrt{m} b_j^{(0)} \right)^2 \tag{B.3}
$$

Now, for each $i \in [p]$, we know that $\sum_{j \in [m]} h^{(i)} \left(\sqrt{m} \langle w_j^{(0)}, w_{1,i}^* \rangle, \sqrt{m} b_j^{(0)} \right)^2$ is a summation of i.i.d. random variables, each with expectation at most $\mathfrak{C}_\mathfrak{s}(\phi, 1)^2$ by Lemma 6.2. Applying concentration, we have with high probability

$$
\sum_{j \in [m]} h^{(i)} \left(\sqrt{m} \langle w_j^{(0)}, w_{1,i}^* \rangle, \sqrt{m} b_j^{(0)} \right)^2 \leq m \cdot \mathfrak{C}_\mathfrak{s}(\phi, 1)^2 + \sqrt{m} \cdot C_0^2 \leq 2m \mathfrak{C}_\mathfrak{s}(\phi, 1)^2
$$

Putting this back to (B.3) we have $\|W^*\|_F^2 \leq \widetilde{O}(\frac{k^2 p^2 \mathfrak{C}_\mathfrak{s}(\phi, 1)^2}{\varepsilon_a^2 m})$. □

Proof of Corollary B.2. Let W^* be the weights constructed in Lemma B.1 to approximate up to error $\varepsilon/2$. Then

$$
|g_r(x; W^{(0)} + W^*) - g_r^{(b)}(x; W^*)| = \left| \sum_{i=1}^{m} a_{r,i}^{(0)} (\langle w_i^{(0)}, x \rangle + b_i^{(0)}) \mathbb{I}[\langle w_i^{(0)}, x \rangle + b_i^{(0)} \geq 0] \right| .
$$

By standard concentration (which uses the randomness of $w^{(0)}$ together with the randomness of $a_{r,1}^{(0)}, \dots, a_{r,m}^{(0)}$), the above quantity is with high probability bounded by $\widetilde{O}(\varepsilon_a) = \varepsilon/2$. This is the only place we need parameter choice ε_a. □

B.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 B.3 (Coupling). *For every unit vector x, w.h.p. over the random initialization, for every time step $t \geq 1$, we have the following. Denote $\tau = \varepsilon_a \eta t$.*

(a) For at most $\widetilde{O}(\tau\sqrt{km})$ fraction of $i \in [m]$:

$$\mathbb{I}[\langle w_i^{(0)}, x\rangle + b_i^{(0)} \geq 0] \neq \mathbb{I}[\langle w_i^{(t)}, x\rangle + b_i^{(0)} \geq 0].$$

(b) For every $r \in [k]$,

$$\left| f_r(x; W^{(0)} + W_t) - g_r(x; W^{(0)} + W_t) \right| = \widetilde{O}(\varepsilon_a k \tau^2 m^{3/2}).$$

(c) For every y:

$$\begin{aligned}&\left\| \frac{\partial}{\partial W} L(F(x; W^{(0)} + W_t), y) - \frac{\partial}{\partial W} L(G(x; W^{(0)} + W_t), y) \right\|_{2,1} \\ &\leq \widetilde{O}(\varepsilon_a k \tau m^{3/2} + \varepsilon_a^2 k^2 \tau^2 m^{5/2}).\end{aligned}$$

Proof of Lemma B.3. Let us recall

$$\begin{aligned} f_r(x; W) &= \sum_{i=1}^m a_{r,i}^{(0)} (\langle w_i, x\rangle + b_i^{(0)}) \mathbb{I}[\langle w_i, x\rangle + b_i^{(0)} \geq 0] \\ g_r(x; W) &= \sum_{i=1}^m a_{r,i}^{(0)} (\langle w_i, x\rangle + b_i^{(0)}) \mathbb{I}[\langle w_i^{(0)}, x\rangle + b_i^{(0)} \geq 0]. \end{aligned} \tag{B.4}$$

(a) W.h.p. over the random initialization, there is $B = \widetilde{O}(1)$ so that every $|a_{r,i}^{(0)}| \leq \varepsilon_a B$. Thus, by the 1-Lipschitz continuity of L, for every $i \in [m]$ and every $t \geq 0$,

$$\left\| \frac{\partial f_r(x; W^{(0)} + W_t)}{\partial w_i} \right\|_2 \leq \varepsilon_a B \quad \text{and} \quad \left\| \frac{\partial L(F(x; W^{(0)} + W_t), y)}{\partial w_i} \right\|_2 \leq \sqrt{k} \varepsilon_a B \tag{B.5}$$

which implies that $\left\| w_i^{(t)} - w_i^{(0)} \right\|_2 \leq \sqrt{k} B \varepsilon_a \eta t = \sqrt{k} B \tau$. Accordingly, define $\mathcal{H}$

$$\mathcal{H} \stackrel{\text{def}}{=} \left\{ i \in [m] \ \Big| \ \left| \left\langle w_i^{(0)}, x \right\rangle + b_i^{(0)} \right| \geq 2\sqrt{k} B \tau \right\}.$$

so it satisfies 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 $\langle w_i^{(0)}, x\rangle \sim \mathcal{N}(0, 1/m)$, and $b_i^{(0)} \sim \mathcal{N}(0, 1/m)$, by standard property of Gaussian one can derive $|\mathcal{H}| \geq m\big(1 - \widetilde{O}(\sqrt{k} B \tau \sqrt{m})\big) = m\big(1 - \widetilde{O}(\tau\sqrt{km})\big)$ with high probability.

(b) It is clear from (B.4) that f_r and g_r only differ on the indices $i \notin \mathcal{H}$. For such an index $i \notin \mathcal{H}$, the sign of $\langle w_i^{(t)}, x\rangle + b_i^{(0)}$ is different from that of $\langle w_i^{(0)}, x\rangle + b_i^{(0)}$ and their difference is at most $\sqrt{k} B \tau$. This contributes at most $\varepsilon_a B \cdot \sqrt{k} B \tau$ difference between f_r and g_r. Then the bound follows from that there are only $\widetilde{O}\big(\tau m \sqrt{km}\big)$ many $i \notin \mathcal{H}$.

(c) Note that

$$\frac{\partial}{\partial w_i} L(F(x; W^{(0)} + W_t), y) = \nabla L(F(x; W^{(0)} + W_t), y) \frac{\partial}{\partial w_i} F(x; W^{(0)} + W_t)$$

By the Lipschitz smoothness assumption on L and Lemma B.3b, we have

$$\|\nabla L(F(x; W^{(0)} + W_t), y) - \nabla L(G(x; W^{(0)} + W_t), y)\|_2 \le \widetilde{O}(\varepsilon_a k^{3/2} \tau^2 m^{3/2}) \ . \tag{B.6}$$

For $i \in \mathcal{H}$ we have $\mathbb{I}[\langle w_i^{(0)}, x\rangle + b_i^{(0)} \ge 0] = \mathbb{I}[\langle w_i^{(t)}, x\rangle + b_i^{(0)} \ge 0]$, so $\frac{\partial F(x; W^{(0)} + W_t)}{\partial w_i} = \frac{\partial G(x; W^{(0)} + W_t)}{\partial w_i}$, and thus the difference is only caused by (B.6). Using (B.5), each such i contributes at most $\widetilde{O}(\varepsilon_a^2 k^2 \tau^2 m^{3/2})$, totaling $\widetilde{O}(\varepsilon_a^2 k^2 \tau^2 m^{5/2})$.

For $i \notin \mathcal{H}$, it contributes at most $\sqrt{k}\varepsilon_a B$ because of (B.5), and there are $\widetilde{O}(\tau m \sqrt{km})$ many such i's, totaling $\widetilde{O}(\varepsilon_a k \tau m^{3/2})$.

□

B.3 Optimization

Recall for $z = (x, y)$

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

For the set of samples $\mathcal{Z}$, define

$$L_F(\mathcal{Z}; W) \stackrel{\text{def}}{=} \frac{1}{|\mathcal{Z}|} \sum_{(x,y)\in\mathcal{Z}} L(F(x; W + W^{(0)}), y) \ ,$$
$$L_G(\mathcal{Z}; W) \stackrel{\text{def}}{=} \frac{1}{|\mathcal{Z}|} \sum_{(x,y)\in\mathcal{Z}} L(G(x; W + W^{(0)}), y) \ .$$

We show the following lemma:

Lemma B.4. *For every $\varepsilon \in (0, \frac{1}{pk\mathfrak{C}_{\mathfrak{s}}(\phi,1)})$, letting $\varepsilon_a = \varepsilon/\widetilde{\Theta}(1)$ and $\eta = \widetilde{\Theta}\big(\frac{1}{\varepsilon k m}\big)$, there exists*

$$M = \mathsf{poly}(\mathfrak{C}_\varepsilon(\phi, 1), 1/\varepsilon) \quad \textit{and} \quad T = \Theta\Big(\frac{k^3 p^2 \cdot \mathfrak{C}_{\mathfrak{s}}(\phi, 1)^2}{\varepsilon^2}\Big)$$

such that, w.h.p., if $m \ge M$ then

$$\frac{1}{T} \sum_{t=0}^{T-1} L_F(\mathcal{Z}; W_t) \le \mathsf{OPT} + \varepsilon.$$

Proof of Lemma B.4. Let W^* be constructed in Corollary B.2. Recall $L(\cdot, y)$ is convex and $G(x; W)$ is linear in W so $L_G(z; W)$ is convex in W. By such convexity, we have

$$\begin{aligned} L_G(\mathcal{Z}; W_t) - L_G(\mathcal{Z}; W^*) &\le \langle \nabla L_G(\mathcal{Z}; W_t), W_t - W^*\rangle \\ &\le \|\nabla L_G(\mathcal{Z}; W_t) - \nabla L_F(\mathcal{Z}; W_t)\|_{2,1} \|W_t - W^*\|_{2,\infty} \\ &\quad + \langle \nabla L_F(\mathcal{Z}; W_t), W_t - W^*\rangle. \end{aligned}$$

We also have

$$\begin{aligned} \|W_{t+1} - W^*\|_F^2 &= \|W_t - \eta \nabla L_F(z^{(t)}, W_t) - W^*\|_F^2 \\ &= \|W_t - W^*\|_F^2 - 2\eta \langle \nabla L_F(z^{(t)}, W_t), W_t - W^*\rangle \\ &\quad + \eta^2 \|\nabla L_F(z^{(t)}, W_t)\|_F^2, \end{aligned}$$

so

$$\begin{aligned}L_G(\mathcal{Z};W_t) - L_G(\mathcal{Z};W^*) \le \|\nabla L_G(\mathcal{Z};W_t) - \nabla L_F(\mathcal{Z};W_t)\|_{2,1}\|W_t - W^*\|_{2,\infty} \\ + \frac{\|W_t - W^*\|_F^2 - \mathbb{E}_{z^{(t)}}[\|W_{t+1} - W^*\|_F^2]}{2\eta} \\ + \frac{\eta}{2}\|\nabla L_F(W_t, z^{(t)})\|_F^2. \qquad \text{(B.7)}\end{aligned}$$

Recall from (B.5) that with high probability over the random initialization,

$$\begin{aligned}\|W_t\|_{2,\infty} = \widetilde{O}(\sqrt{k}\varepsilon_a\eta t) \ , \quad \|W_t - W^*\|_{2,\infty} = \widetilde{O}(\sqrt{k}\varepsilon_a\eta t + \frac{kpC_0}{\varepsilon_a m}) \qquad \text{(B.8)} \\ \text{and} \quad \|\nabla L_F(W_t, z^{(t)})\|_F^2 = \widetilde{O}(\varepsilon_a^2 km) \ ,\end{aligned}$$

where C_0 is as in Lemma B.1. By Lemma B.3c, w.h.p. we know

$$\|\nabla L_G(\mathcal{Z};W_t) - \nabla L_F(\mathcal{Z};W_t)\|_{2,1} \le \Delta = \widetilde{O}(\varepsilon_a^2 k\eta T m^{3/2} + \varepsilon_a^4 k^2(\eta T)^2 m^{5/2}).$$

Therefore, averaging up (B.7) from $t = 0$ to $T - 1$ we have that

$$\begin{aligned}\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}_{sgd}[L_G(\mathcal{Z};W_t)] - L_G(\mathcal{Z};W^*) \le \widetilde{O}\left(\sqrt{k}\varepsilon_a\eta T\Delta + \frac{kpC_0}{\varepsilon_a m}\Delta\right) + \frac{\|W_0 - W^*\|_F^2}{2\eta T} \\ + \widetilde{O}(k\varepsilon_a^2\eta m).\end{aligned}$$

Note that $\|W_0 - W^*\|_F^2 = \|W^*\|_F^2 \le \widetilde{O}\big(\frac{k^2p^2\mathfrak{C}_\mathfrak{s}(\phi,1)^2}{\varepsilon_a^2 m}\big)$ from Lemma B.1. Also recall $\varepsilon_a = \Theta(\varepsilon)$. By choosing $\eta = \widetilde{\Theta}\big(\frac{\varepsilon}{km\varepsilon_a^2}\big)$ and $T = \widetilde{\Theta}\left(k^3p^2\mathfrak{C}_\mathfrak{s}(\phi,1)^2/\varepsilon^2\right)$ we have $\Delta = \widetilde{O}(k^6p^4\mathfrak{C}_\mathfrak{s}(\phi,1)^4m^{1/2}/\varepsilon^2)$. Thus when m is large enough we have:

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}_{sgd}[L_G(\mathcal{Z};W_t)] - L_G(\mathcal{Z};W^*) \le O(\varepsilon).$$

By the coupling in Lemma B.3b, we know that $L_F(\mathcal{Z};W_t)$ is $o(\varepsilon)$-close to $L_G(\mathcal{Z};W_t)$; by applying Corollary B.2, we know that $L_G(\mathcal{Z};W^*)$ is ε-close to OPT. This finishes the proof. □

B.4 Generalization

The generalization can be bounded via known Rademacher complexity results. Recall

$$f_r(x;W^{(0)} + W') \stackrel{\text{def}}{=} \sum_{i=1}^m a_{r,i}^{(0)}\sigma(\langle w_i^{(0)} + w_i', x\rangle + b_i^{(0)}) \ .$$

We have the following simple lemma (see also [36, Theorem 43])[19]

Lemma B.5 (two-layer network Rademacher complexity)**.** *For every $\tau_{w,\infty} \ge 0$, w.h.p. for every $r \in [k]$ and every $N \ge 1$, we have the empirical Rademacher complexity bounded by*

$$\frac{1}{N}\mathbb{E}_{\xi\in\{\pm1\}^N}\left[\sup_{\|W'\|_{2,\infty}\le\tau_{w,\infty}}\sum_{i\in[N]}\xi_i f_r(x_i;W^{(0)} + W')\right] \le \widetilde{O}\left(\frac{\varepsilon_a m\tau_{w,\infty}}{\sqrt{N}}\right).$$

Proof. The proof consists of the following simple steps.

- $\{x \mapsto \langle w_j', x\rangle \mid \|w_j'\|_2 \le \tau_{w,\infty}\}$ has Rademacher complexity $O(\frac{\tau_{w,\infty}}{\sqrt{N}})$ by Proposition A.12a.

[19] We note that [36, Theorem 43] did not have $W^{(0)}$ but the same result holds with the introduction of $W^{(0)}$.

- $\{x \mapsto \langle w_j^{(0)} + w'_j, x\rangle + b_j \mid \|w'_j\|_2 \leq \tau_{w,\infty}\}$ has Rademacher complexity $O(\frac{\tau_{w,\infty}}{\sqrt{N}})$ because singleton class has zero complexity and adding it does not affect complexity by Proposition A.12c.
- $\{x \mapsto f_r(x; W^{(0)} + W') \mid \|W'\|_{2,\infty} \leq \tau_{w,\infty}\}$ has Rademacher complexity $\widetilde{O}(\frac{\varepsilon_a m \tau_{w,\infty}}{\sqrt{N}})$ because w.h.p. $\|a_r^{(0)}\|_1 \leq \widetilde{O}(\varepsilon_a m)$ and Proposition A.12e. □

B.5 Theorem 1: Two-Layer

Proof of Theorem 1. First, we can apply Lemma B.4 to bound the training loss. That is

$$\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}_{(x,y)\in\mathcal{Z}} L(F(x; W_t + W^{(0)}), y) \leq \mathsf{OPT} + \varepsilon .$$

Also recall from (B.8) and our parameter choices for η, T that

$$\|W_t\|_{2,\infty} \leq O\left(\frac{\mathsf{poly}(k,p,\log m)\cdot \mathfrak{C}_{\mathfrak{s}}(\phi,1)^2}{\varepsilon^2 m}\right) .$$

so we can choose $\tau_{w,\infty} = O(\frac{\mathsf{poly}(k,p,\log m)\cdot \mathfrak{C}_{\mathfrak{s}}(\phi,1)^2}{\varepsilon^2 m})$. For each $(x,y) \sim \mathcal{D}$, it is a simple exercise to verify that $|L(F(x; W_t + W^{(0)}), y)| \leq O(\frac{\mathsf{poly}(k,p,\log m)\cdot \mathfrak{C}_{\mathfrak{s}}(\phi,1)^2}{\varepsilon})$ with high probability.[20] Thus, we can plug the Rademacher complexity Lemma B.5 together with $b = O(\frac{\mathsf{poly}(k,p,\log m)\cdot \mathfrak{C}_{\mathfrak{s}}(\phi,1)^2}{\varepsilon})$ into standard generalization statement Corollary A.11. It gives

$$\begin{aligned}&\left|\mathbb{E}_{(x,y)\in\mathcal{D}} L(F(x; W_t + W^{(0)}), y) - \mathbb{E}_{(x,y)\in\mathcal{Z}} L(F(x; W_t + W^{(0)}), y)\right| \\ &\leq O(\frac{\mathsf{poly}(k,p,\log m)\cdot \mathfrak{C}_{\mathfrak{s}}(\phi,1)^2}{\varepsilon\sqrt{N}}) \ . \end{aligned} \tag{B.9}$$

This completes the proof with large enough N. □

Remark B.6. Strictly speaking, $|g_r(x; W)| \leq \widetilde{O}(\varepsilon_a)$ does not hold for every x in $\mathcal{D}$, thus the loss function L is not absolutely bounded so one cannot apply Corollary A.11 directly.[21] We only have the statement that for each sample x, the loss function $|L(F(x; W_t + W^{(0)}), y)| \leq b$ is bounded by some parameter b *with high probability.* By union bound, with high probability this can hold for all the training samples (but possibly not all the testing samples). A simple fix here is to apply a truncation (for analysis purpose only) on the loss function L to make it always bounded by b. Then, we can apply Corollary A.11: the population risk "$\mathbb{E}_{(x,y)\in\mathcal{D}} L(\cdots)$" in (B.9) becomes truncated but the empirical risk "$\mathbb{E}_{(x,y)\in\mathcal{Z}} L(\cdots)$" in (B.9) stays unchanged. In other words, the *truncated* population risk must be small according to Corollary A.11. Finally, we can remove truncation from the population risk, because in the rare event that $|L(F(x; W_t + W^{(0)}), y)|$ exceeds b, it is at most $\mathsf{poly}(m)$ so becomes negligible when evaluating the expectation $\mathbb{E}_{(x,y)\in\mathcal{D}} L(\cdots)$.

Remark B.7. In the above proof, it appears that N scales with ε^{-4} which may seemingly be larger than T which only scales with ε^{-2}. We are aware of a proof that tightens N to be on the order of ε^{-2}. It uses standard (but complicated) martingale analysis and creates extra difficulty that is irrelevant to neural networks in general. We choose not to present it for simplicity.

[20] Indeed, with high probability $|g_r(x; W)| \leq \widetilde{O}(\varepsilon_a)$ and since $\|W_t\|_{2,\infty} \leq \tau_{w,\infty}$ we have $|g_r^{(b)}(x; W_t)| \leq \widetilde{O}(\varepsilon_a \tau_{w,\infty} m)$. Together we have $|g_r(x; W + W_t)| \leq \widetilde{O}(\varepsilon_a \tau_{w,\infty} m)$. By the coupling Lemma B.3b, this implies $|f_r(x; W + W_t)| \leq \widetilde{O}(\varepsilon_a \tau_{w,\infty} m)$ as well. Using $L(0,y) \in [0,1]$ and the 1-Lipschitz continuity finishes the proof.

[21] In some literature this issue was simply ignored or an absolute bound on L is imposed; however, the only globally absolutely bounded convex function is constant.

C Proofs for Three-Layer Networks

Our three-layer proofs follow the same structure as our proof overview in Section 6.

C.1 Existential Results

C.1.1 Lemma 6.2: Indicator to Function

Recall without loss of generality it suffices to prove Lemma 6.2a.

Lemma 6.2a (indicator to function)**.** *For every smooth function ϕ, every $\varepsilon \in \big(0, \frac{1}{\mathfrak{C}_{\mathfrak{s}}(\phi,1)}\big)$, we have that there exists a function $h : \mathbb{R}^2 \to [-\mathfrak{C}_\varepsilon(\phi,1), \mathfrak{C}_\varepsilon(\phi,1)]$ such that for every $x_1 \in [-1,1]$:*

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

where $\alpha_1, \beta_1 \sim \mathcal{N}(0,1)$ and $b_0 \sim \mathcal{N}(0,1)$ are independent random variables. Furthermore:

- *h is $\mathfrak{C}_\varepsilon(\phi,1)$-Lipschitz on the first coordinate.*
- $\mathbb{E}_{\alpha_1, b_0 \sim \mathcal{N}(0,1)}\big[h(\alpha_1,b_0)^2\big] \leq (\mathfrak{C}_{\mathfrak{s}}(\phi,1))^2$.

For notation simplicity, let us denote $w_0 = (\alpha_1, \beta_1)$ and $x = (x_1, \sqrt{1-x_1^2})$ where α_1, β_1 are two independent random standard Gaussians.

Throughout the proof, we also take an alternative view of the randomness. We write $\langle w_0, x\rangle = \alpha$ and $\alpha_1 = \alpha x_1 + \sqrt{1-x_1^2}\beta$ for two independent $\alpha, \beta \sim \mathcal{N}(0,1)$.[22]

We first make a technical claim involving in fitting monomials in x_1. Its proof is in Section C.1.2.

Claim C.1. *Recall $h_i(x)$ is the degree-i Hermite polynomial (see Definition A.5). For every integer $i \geq 1$ there exists constant p'_i with $|p'_i| \geq \frac{(i-1)!!}{200 i^2}$ such that*

$$\begin{aligned}
\textit{for even } i: \quad & x_1^i = \frac{1}{p'_i}\mathbb{E}_{w_0 \sim \mathcal{N}(0,\mathbf{I}), b_0 \sim \mathcal{N}(0,1)}\left[h_i(\alpha_1)\cdot \mathbb{I}[0 < -b_0 \leq 1/(2i)] \cdot \mathbb{I}[\langle x, w_0\rangle + b_0 \geq 0]\right] \\
\textit{for odd } i: \quad & x_1^i = \frac{1}{p'_i}\mathbb{E}_{w_0 \sim \mathcal{N}(0,\mathbf{I}), b_0 \sim \mathcal{N}(0,1)}\left[h_i(\alpha_1)\cdot \mathbb{I}[|b_0| \leq 1/(2i)] \cdot \mathbb{I}[\langle x, w_0\rangle + b_0 \geq 0]\right]
\end{aligned}$$

We next use Claim C.1 to fit arbitrary functions $\phi(x_1)$. By Taylor expansion, we have

$$\begin{aligned}
\phi(x_1) &= c_0 + \sum_{i=1,\text{ odd } i}^{\infty} c_i x_1^i + \sum_{i=2,\text{ even } i}^{\infty} c_i x_1^i \\
&= c_0 + \sum_{i=1}^{\infty} c'_i \cdot \mathbb{E}_{\alpha,\beta,b_0 \sim \mathcal{N}(0,1)}\big[h_i(\alpha_1)\cdot \mathbb{I}[q_i(b_0)] \cdot \mathbb{I}[\langle x, w_0\rangle + b_0 \geq 0]\big]
\end{aligned}$$

where

$$c'_i \stackrel{\text{def}}{=} \frac{c_i}{p'_i} \quad , \quad |c'_i| \leq \frac{200 i^2 |c_i|}{(i-1)!!} \quad \text{and} \quad q_i(b_0) = \begin{cases} |b_0| \leq 1/(2i), & i \text{ is odd;} \\ 0 < -b_0 \leq 1/(2i), & i \text{ is even.} \end{cases} \tag{C.1}$$

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

Claim C.2. *Setting $B_i \stackrel{\text{def}}{=} 100 i^{1/2} + 10\sqrt{\log\frac{1}{\varepsilon}}$, we have*

[22] This is possible for the following reason. Let $x^\perp = (\sqrt{1-x_1^2}, -x_1)$ be unit vector orthogonal to x. We can write $w_0 = \alpha x + \beta x^\perp$ where $\alpha, \beta \sim \mathcal{N}(0,1)$ are two independent Gaussians.

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

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

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

(d) $\sum_{i=1}^{\infty} |c'_i| \cdot \mathbb{E}_{z\sim\mathcal{N}(0,1)}\big[\big|\frac{d}{dz}h_i(z)\big| \cdot \mathbb{I}[|z| \leq B_i]\big] \leq \frac{1}{2}\mathfrak{C}_{\varepsilon}(\phi, 1)$

Now, let us define $\widehat{h}_i(\alpha_1) \stackrel{\text{def}}{=} h_i(\alpha_1) \cdot \mathbb{I}[|\alpha_1| \leq B_i] + h_i(\text{sign}(\alpha_1)B_i) \cdot \mathbb{I}[|\alpha_1| > B_i]$ as the truncated version of the Hermite polynomial $h_i(\cdot)$.

Using Claim C.2, we have

$$\phi(x_1) = c_0 + R'(x_1) + \sum_{i=1}^{\infty} c'_i \cdot \mathbb{E}_{\alpha,\beta,b_0\sim\mathcal{N}(0,1)}\Big[\widehat{h}_i(\alpha_1) \cdot \mathbb{I}[q_i(b_0)] \cdot \mathbb{I}[\langle x, w_0\rangle + b_0 \geq 0]\Big]$$

where $|R'(x_1)| < \epsilon/4$ uses Claim C.2a and Claim C.2b. In other words, if we define

$$h(\alpha_1, b_0) \stackrel{\text{def}}{=} 2c_0 + \sum_{i=1}^{\infty} c'_i \cdot \widehat{h}_i(\alpha_1) \cdot \mathbb{I}[q_i(b_0)]$$

then we have

$$\big|\mathbb{E}_{\alpha,\beta,b_0\sim\mathcal{N}(0,1)}\big[\mathbb{I}[\langle x, w_0\rangle + b_0 \geq 0] \cdot h(\alpha_1, b_0)\big] - \phi(x_1)\big| = \big|R'(x_1)\big| \leq \varepsilon/4 \enspace .$$

As for the range of h, we use Claim C.2b and Claim C.2c to derive that

$$|h(\alpha_1, b_0)| \leq 2c_0 + \frac{\varepsilon}{8} + \frac{1}{2}\mathfrak{C}_{\varepsilon}(\phi, 1) \leq \mathfrak{C}_{\varepsilon}(\phi, 1) \enspace .$$

As for the Lipschitz continuity of h on its first coordinate α_1, we observe that for each $i > 0$, $\widehat{h}_i(z)$ has zero sub-gradient for all $|z| \geq B_i$. Therefore, it suffices to bound $\big|\frac{d}{dz}h_i(z)\big|$ for $|z| < B_i$. Replacing the use of Claim C.2c by Claim C.2d immediately give us the same bound on the Lipschitz continuity of h with respect to α_1.

As for the expected square $\mathbb{E}_{\alpha_1,b_0\sim\mathcal{N}(0,1)}\big[h(\alpha_1, b_0)^2\big]$, we can write

$$h(\alpha_1, b_0) = 2c_0 + \sum_{i=1}^{\infty} c'_i \cdot \widehat{h}_i(\alpha_1) \cdot \mathbb{I}[q_i(b_0)] \overset{\textcircled{1}}{=} 2c_0 + \sum_{i=1}^{\infty} c'_i \cdot h_i(\alpha_1) \cdot \mathbb{I}[q_i(b_0)] \pm \frac{\varepsilon}{4}$$

Above, ① uses Claim C.2a and Claim C.2b. Using the othogonality condition of Hermite polynomials (that is, $\mathbb{E}_{x\sim\mathcal{N}(0,1)}[h_i(x)h_j(x)] = \sqrt{2\pi}j!\delta_{i,j}$ from Definition A.5), we immediately have

$$\begin{aligned}
\mathbb{E}_{\alpha_1,b_0\sim\mathcal{N}(0,1)}[h(\alpha_1, b_0)^2] &\leq O(\varepsilon^2 + c_0^2) + O(1) \cdot \sum_{i=1}^{\infty}(c'_i)^2(i!) \cdot \mathbb{E}_{b_0}[\mathbb{I}[q_i(b_0)]] \\
&\leq O(\varepsilon^2 + c_0^2) + O(1) \cdot \sum_{i=1}^{\infty} \frac{(c'_i)^2(i!)}{i} \\
&\leq O(\varepsilon^2 + c_0^2) + O(1) \cdot \sum_{i=1}^{\infty} \frac{(i!) \cdot i^3 \cdot |c_i|^2}{((i-1)!!)^2} \\
&\leq O(\varepsilon^2 + c_0^2) + O(1) \cdot \sum_{i=1}^{\infty} i^{3.5} \cdot |c_i|^2 \leq (\mathfrak{C}_{\mathfrak{s}}(\phi, 1))^2 \enspace .
\end{aligned}$$

Above, ① uses inequality $\frac{i!}{((i-1)!!)^2} \leq 2\sqrt{i}$ for all $i \geq 1$.

This finishes the proof of Lemma 6.2a. ■

C.1.2 Proofs of Claim C.1 and Claim C.2

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

Even $\boldsymbol{i}$. By Lemma A.6, we know that

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

where

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

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

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

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

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

which implies

$$\left|\sum_{r=1,r\text{ odd}}^{i-1}c_r(-b_0)^r\right|\ge\frac{2}{3}|c_1b_0|=\frac{2}{3}|b_0|$$

and

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

is independent of the randomness of b_0. Therefore, using the formula of p_i in (C.2):

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

Odd $\boldsymbol{i}$. Similarly, by Lemma A.6, we have

$$\begin{aligned}
&\mathbb{E}_{w_0\sim\mathcal{N}(0,\mathbf{I}),b_0\sim\mathcal{N}(0,1)}\big[h_i(\alpha_1)\cdot\mathbb{I}[|b_0|\le 1/(2i)]\cdot\mathbb{I}[\langle x,w_0\rangle+b_0\ge 0]\big]\\
&=\mathbb{E}_{b_0\sim\mathcal{N}(0,1)}\big[p_i\cdot\mathbb{I}[|b_0|\le 1/(2i)]\big]\times x_1^i
\end{aligned}$$

where

$$p_i = (i-1)!!\frac{\exp(-b_0^2/2)}{\sqrt{2\pi}} \sum_{r=0, r \text{ even}}^{i-1} \frac{(-1)^{\frac{i-1-r}{2}}}{r!!}\binom{i/2-1}{(r-1)/2}(-b_0)^r \ . \tag{C.3}$$

This time we bound the coefficient "$\mathbb{E}_{b_0\sim\mathcal{N}(0,1)}\left[p_i \cdot \mathbb{I}[|b_0| \le 1/(2i)]\right]$" as follows. Define c_r as:

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

Then, for $|b_0| \le \frac{1}{2i}$, we know that for all even r in $1 < r \le i-1$ it satisfies

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

which implies

$$\left|\sum_{r=0, r \text{ even}}^{i-1} c_r(-b_0)^r\right| \ge \frac{2}{3}|c_0| = \frac{2}{3}\left|\binom{i/2-1}{-1/2}\right| > \frac{1}{2i}.$$

and

$$\operatorname{sign}\left(\sum_{r=1, r \text{ odd}}^{i-1} c_r(-b_0)^r\right) = \operatorname{sign}(c_0)$$

is independent of the randomness of b_0. Therefore, using the formula of p_i in (C.3):

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

□

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

$$|h_i(x)| \le \sum_{j=0}^{\lfloor i/2 \rfloor} \frac{i!|x|^{i-2j}}{j!(i-2j)!2^j} \le \sum_{j=0}^{\lfloor i/2 \rfloor} \frac{|x|^{i-2j}i^{2j}}{j!} \tag{C.4}$$

Using our bound on c_i' (see (C.1)), we have

$$|c_i' h_i(z)| \le O(1)|c_i|\frac{i^4}{i!!}\sum_{j=0}^{\lfloor i/2 \rfloor} \frac{|z|^{i-2j}i^{2j}}{j!} \ . \tag{C.5}$$

(a) Denote by $b = B_i = 100i^{1/2}\theta$ for notational simplicity (for some parameter $\theta \ge 1$ that we shall choose later). We have

$$\begin{aligned}|c_i'| \cdot \mathbb{E}_{z\sim\mathcal{N}(0,1)}[|h_i(z)| \cdot \mathbb{I}[|z| \geq b]]| &\leq 2|c_i'| \cdot \mathbb{E}_{z\sim\mathcal{N}(0,1)}\left[\sum_{j=0}^{\lfloor i/2 \rfloor} \frac{|z|^{i-2j} i^{2j} \mathbb{I}[z \geq b]}{j!}\right] \\ &= O(1)|c_i| \frac{i^4}{i!!} \sum_{j=0}^{\lfloor i/2 \rfloor} \frac{L_{i-2j,b} \cdot i^{2j}}{j!} \quad , \qquad \text{(C.6)}\end{aligned}$$

where recall from Lemma A.7 that

$$\begin{aligned}L_{i,b} &\leq O(1)e^{-b^2/2} \cdot \sum_{j=0}^{i-1} \frac{(i-1)!!}{j!!} b^j \\ &\leq O(1)e^{-b^2/2} \cdot \sum_{j=0}^{i-1} \frac{(i-1)!!}{j!!} \left(100 i^{1/2} \theta\right)^j \\ &\leq O(1)e^{-b^2/2} \left(100\theta\right)^i \cdot (i-1)!! \cdot \sum_{j=0}^{i-1} \frac{i^{j/2}}{j!!} \\ &\leq O(1)e^{-b^2/2} \left(200\theta\right)^i \cdot (i-1)!! \qquad \left(\text{using } \textstyle\sum_{j=0}^{i-1} \frac{i^{j/2}}{j!!} \leq 2^i\right)\end{aligned}$$

Thus we have

$$\begin{aligned}\sum_{j=0}^{\lfloor i/2 \rfloor} \frac{L_{i-2j,b} \cdot i^{2j}}{j!} &\leq O(1) \left(200\theta\right)^i e^{-b^2/2} \sum_{j=0}^{\lfloor i/2 \rfloor} \frac{(i+1-2j)!! \cdot i^{2j}}{j!} \\ &\leq O(1) \left(400\theta\right)^i e^{-b^2/2} \sum_{j=0}^{\lfloor i/2 \rfloor} \frac{i^{(i-2j)/2} \cdot i^{2j}}{j!} \\ &= O(1) \left(400\theta\right)^i e^{-b^2/2} i^{i/2} \sum_{j=0}^{\lfloor i/2 \rfloor} \frac{i^j}{j!} \\ &\overset{\textcircled{1}}{\leq} O(1) \left(1200\theta\right)^i e^{-b^2/2} i^{i/2} \qquad \text{(C.7)} \\ &\overset{\textcircled{2}}{\leq} O(i^{i/2}) \cdot 1200^i \cdot \left(\theta \cdot e^{-10^4\theta^2}\right)^i \\ &\overset{\textcircled{3}}{\leq} i^{i/2} \cdot \frac{1}{100^i} \cdot \varepsilon^2 \\ &\overset{\textcircled{4}}{\leq} i^{i/2} \cdot \frac{1}{100^i} \cdot \frac{\varepsilon}{|c_i|} \qquad \text{(C.8)}\end{aligned}$$

Above, inequality ① uses $\sum_{j=0}^{i-1} \frac{i^j}{j!} \leq 3^i$; inequality ② uses our definition of $b = B_i$; inequality ③ uses $(\theta \cdot e^{-10^4\theta^2})^i \leq \frac{\varepsilon^2}{100000^i}$ for $\theta = 1 + \frac{\sqrt{\log(1/\varepsilon)}}{10\sqrt{i}}$; and inequality ④ uses $\varepsilon|c_i| \leq 1$. Putting this back to (C.6), we have

$$\sum_{i=1}^{\infty} |c_i'| \cdot \mathbb{E}_{z\sim\mathcal{N}(0,1)}[|h_i(z)| \cdot \mathbb{I}[|z| \geq b]]| \leq O(1) \sum_{i=1}^{\infty} \frac{i^4}{i!!} \frac{\varepsilon}{100^i} i^{i/2} \leq \frac{\varepsilon}{8} \quad .$$

Above, in the last inequality we have used $\frac{i^4}{i!!} i^{i/2} \leq 40 \cdot 4^i$ for $i \geq 1$.

(b) Similar to the previous case, we calculate that

$$
\begin{aligned}
|c_i'| \cdot \mathbb{E}_{z\sim\mathcal{N}(0,1)}[|h_i(b)| \cdot \mathbb{I}[|z| \ge b]]| &\le |c_i'| \cdot \mathbb{E}_{z\sim\mathcal{N}(0,1)} \left[\sum_{j=0}^{\lfloor i/2 \rfloor} \frac{b^{i-2j} i^{2j}}{j!} \mathbb{I}[|z| \ge b] \right] \\
&\le O(1) |c_i| \frac{i^4}{i!!} \cdot L_{0,b} b^i \sum_{j=0}^{\lfloor i/2 \rfloor} \frac{b^{-2j} \cdot i^{2j}}{j!} \\
&\overset{\textcircled{1}}{\le} O(1) |c_i| \frac{i^4}{i!!} \cdot e^{-b^2/2} b^i \sum_{j=0}^{\lfloor i/2 \rfloor} \frac{i^j}{j!} \\
&\overset{\textcircled{2}}{\le} O(1) |c_i| \frac{i^4}{i!!} \cdot e^{-b^2/2} (3b)^i \\
&\le |c_i| \frac{i^4}{i!!} \cdot e^{-b^2/2} (3b)^i \quad .
\end{aligned}
$$

Above, inequality ① uses $b^{2j} = B_i^{2j} \ge (10i)^j$; and inequality ② uses again $\sum_{j=0}^{i-1} \frac{i^j}{j!} \le 3^i$. Using this and continue from (C.7) of the previous case, we finish the proof.

(c) Again denote by $b = B_i = 100 i^{1/2} \theta$ for notational simplicity. By Eq. (C.5), it holds that

$$
\begin{aligned}
\sum_{i=1}^{\infty} |c_i'| \cdot \mathbb{E}_{z\sim\mathcal{N}(0,1)} \left[|h_i(z)| \cdot \mathbb{I}[|z| \le B_i] \right] &\le O(1) \sum_{i=1}^{\infty} |c_i| \frac{i^4}{i!!} \sum_{j=0}^{\lfloor i/2 \rfloor} \frac{B_i^{i-2j} i^{2j}}{j!} \\
&\le O(1) \sum_{i=1}^{\infty} |c_i| \frac{i^4}{i!!} \sum_{j=0}^{\lfloor i/2 \rfloor} \left(100 i^{1/2} \theta\right)^{i-2j} \frac{i^{2j}}{j!} \\
&\le O(1) \sum_{i=1}^{\infty} |c_i| \frac{i^4}{i!!} \left(100 i^{1/2} \theta\right)^{i} \sum_{j=0}^{\lfloor i/2 \rfloor} \frac{i^{j}}{j!} \\
&\overset{\textcircled{1}}{\le} \sum_{i=1}^{\infty} |c_i| \left(O(1)\theta\right)^i \\
&\overset{\textcircled{2}}{\le} \sum_{i=1}^{\infty} |c_i| \left(O(1) \big(1 + \frac{\sqrt{\log(1/\varepsilon)}}{10\sqrt{i}}\big) \right)^i \\
&\le \frac{1}{2} \mathfrak{C}_{\varepsilon}(\phi, 1) \quad .
\end{aligned}
$$

Here, in ① we use the fact that $\frac{i^j}{j!} \le 10^i$; in ② we use the fact that $\left(\frac{a}{b}\right)^b \le e^a$ for all $b \in [1, a]$.

(d) By the definition of Hermite polynomial (see Definition A.5), we can also bound

$$
\left|\frac{d}{dx} h_i(x)\right| \le \sum_{j=0}^{\lfloor i/2 \rfloor} |x|^{i-2j} i^{2j}
$$

which is the same upper bound comparing to (C.4). Therefore, the same proof of Claim C.2c also applies to $\left|\frac{d}{dx} h_i(x)\right|$. □

C.1.3 Lemma 6.3: Information out of Randomness

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

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

Lemma 6.3 (information out of randomness)**.** *For every smooth function ϕ, every $w^* \in \mathbb{R}^d$ with $\|w^*\|_2 = 1$, for every $\varepsilon \in \left(0, \frac{1}{\mathfrak{C}_\mathfrak{s}(\phi,1)}\right)$, there exists real-valued functions*

$$\rho(v_1^{(0)}, W^{(0)}, b_1^{(0)}),\ B(x, v_1^{(0)}, W^{(0)}, b_1^{(0)}), R(x, v_1^{(0)}, W^{(0)}, b_1^{(0)}),\ \textit{and}\ \phi_\varepsilon(x)$$

such that for every x:

$$n_1(x) = \rho\left(v_1^{(0)}, W^{(0)}, b_1^{(0)}\right)\phi_\varepsilon(x) + B\left(x, v_1^{(0)}, W^{(0)}, b_1^{(0)}\right) + R\left(x, v_1^{(0)}, W^{(0)}, b_1^{(0)}\right).$$

Moreover, letting $C = \mathfrak{C}_\varepsilon(\phi,1)$ be the complexity of ϕ, and if $v_{1,i}^{(0)} \sim \mathcal{N}(0, \frac{1}{m_2})$ and $w_{i,j}^{(0)}, b_{1,i}^{(0)} \sim \mathcal{N}(0, \frac{1}{m_1})$ are at random initialization, then 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 $\|x\|_2 = 1$, $|\phi_\varepsilon(x) - \phi(\langle w^*, x\rangle)| \le \varepsilon$.*
4. *For every fixed x with $\|x\|_2 = 1$, with high probability $\left|R\left(x, v_1^{(0)}, W^{(0)}, b_1^{(0)}\right)\right| \le \widetilde{O}\left(\frac{1}{\sqrt{m_1 m_2}}\right)$ and $\left|B\left(x, v_1^{(0)}, W^{(0)}, b_1^{(0)}\right)\right| \le \widetilde{O}\left(\frac{1}{\sqrt{m_2}}\right)$.*

Furthermore, there exists real-valued function $\widetilde{\rho}(v_1^{(0)})$ satisfying with high probability:

$$\widetilde{\rho}(v_1^{(0)}) \sim \mathcal{N}\left(0, \frac{1}{100C^2 m_2}\right) \quad \textit{and} \quad \mathcal{W}_2(\rho|_{W^{(0)}, b_1^{(0)}}, \widetilde{\rho}) \le \widetilde{O}\left(\frac{1}{C\sqrt{m_1 m_2}}\right) \quad .$$

Before going to proofs, we recall from Section **??** that we have overridden the notion of "with high probability" so the above statement implies $\mathcal{W}_2(R, 0) \le O\left(\frac{1}{\sqrt{m_1 m_2}}\right)$ and $\mathcal{W}_2(B, 0) \le \widetilde{O}\left(\frac{1}{\sqrt{m_2}}\right)$.

Proof of Lemma 6.3. Without loss of generality we assume $w^* = e_1$. Recall that

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

By Lemma 6.2, for every $\varepsilon > 0$, there exists a function h such that for every unit x with $x_d = \frac{1}{2}$ and every $i \in [m_1]$: [23]

$$\mathbb{E}_{w_i^{(0)} \sim \mathcal{N}(0, \frac{\mathbf{I}}{m_1}), b_{1,i}^{(0)} \sim \mathcal{N}(0, \frac{1}{m_1})}\left[h\left(w_{i,1}^{(0)}, b_{1,i}^{(0)}\right) x_d \mathbb{I}[\langle w_i^{(0)}, x\rangle + b_{1,i}^{(0)} \ge 0]\right] = \frac{\phi_\varepsilon(x_1)}{C} = \frac{\phi_\varepsilon(\langle w^*, x\rangle)}{C} \tag{C.9}$$

with

$$|\phi_\varepsilon(\langle w^*, x\rangle) - \phi(\langle w^*, x\rangle)| \le \varepsilon$$

and $\left|h\left(w_{i,1}^{(0)}, b_{1,i}^{(0)}\right)\right| \in [0,1]$. (Here to apply Lemma 6.2, we have re-scaled h in Lemma 6.2 by $\frac{1}{C}$ and re-scaled α_1, β_1, b_0 in Lemma 6.2 by $\frac{1}{\sqrt{m_1}}$.)

[23] If one wants to work with the more general target function in Remark 4.2, we can assume without loss of generality that $w_3^* = e_d$ (because we have $\langle w_1^*, w_3^*\rangle = 0$ in Remark 4.2).

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 $I\colon [-1,1] \ni y \mapsto I(y) \subset \mathbb{R}$ given in Lemma A.4. Now, for every $w_{i,1}^{(0)}$, define

$$I_i \stackrel{\text{def}}{=} I\left(h\left(w_{i,1}^{(0)}, b_{1,i}^{(0)}\right)\right) \subset [-2,2]$$

Also define set

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

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

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

By the definition of I_i and $\mathfrak{s}$ (see Lemma A.4), we claim 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. In other words, the following unit vector $u \in \mathbb{R}^{m_1}$ is independent of $W^{(0)}$:

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

Since each $i \in \mathcal{S}$ with probability τ, we also know with high probability:

$$|\mathcal{S}| = \tau m_1 \pm O(\sqrt{\tau m_1}) \enspace . \tag{C.10}$$

Conversely, conditioning on $\mathcal{S} = \mathcal{S}_0$ and $\{s_i\}_{i \in \mathcal{S}} = s$ being fixed (or equivalently on u being fixed), the distribution of $W^{(0)}$ is also unchanged. 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

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

where $\alpha \stackrel{\text{def}}{=} u^\top W^{(0)} e_d \sim \mathcal{N}\left(0, \frac{1}{m_1}\right)$ and $\beta \in \mathbb{R}^{m_1 \times d}$ are two independent random 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,

- 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 $W^{(0)}$ (or $b_1^{(0)}$) then the distribution of α is not Gaussian. Fortunately, fixing $W^{(0)}$ and $b_1^{(0)}$, we still have that the coordinates of $s = (s_1, \dots, s_{m_1})$ are i.i.d. (each s_i is zero with probability $1-\tau$, and s_i is ± 1 each with probability $\tau/2$).

Let $\alpha|_{W^{(0)}, b_1^{(0)}}$ denote the conditional distribution of α. With high probability over $W^{(0)}$, $\|W^{(0)} e_d\|_\infty \leq \widetilde{O}\left(\frac{1}{\sqrt{m_1}}\right)$. Fixing the support of u to be $\mathcal{S}$, we know that for every $i \in \mathcal{S}$, u_i is i.i.d. $\pm \frac{1}{\sqrt{|\mathcal{S}|}}$. This implies that fixing $W^{(0)}, b_1^{(0)}, \mathcal{S}$, the quantity

$$\alpha = u^\top W^{(0)} e_d = \sum_{i \in \mathcal{S}} u_i [W^{(0)} e_d]_i \tag{C.11}$$

is a sum of $|\mathcal{S}|$ many independent, mean zero random variables with each $|u_i [W^{(0)} e_d]_i| \leq \widetilde{O}\big(\frac{1}{\sqrt{m_1 |\mathcal{S}|}}\big)$ and $\sum_{i \in \mathcal{S}} u_i [W^{(0)} e_d]_i^2 = \frac{1}{m_1} \pm \widetilde{O}\big(\frac{1}{\sqrt{\tau} m_1^{3/2}}\big)$ w.h.p. Applying any Wasserstein distance bound of central limit theorem (see Lemma A.3), we know that there exists some random Gaussian $g \sim \mathcal{N}(0, \frac{1}{m_1})$ that is independent of W_0 or $b_1^{(0)}$ such that w.h.p.

$$\mathcal{W}_2(\alpha|_{W^{(0)}, b_1^{(0)}}, g) \leq \widetilde{O}\left(\frac{1}{\sqrt{\tau} m_1}\right) \quad . \tag{C.12}$$

We can write

$$\begin{aligned} n_1(x) &= \sum_{i \in [m_1]} v_{1,i}^{(0)} \sigma\left(\langle w_i^{(0)}, x\rangle + b_{1,i}^{(0)}\right) \\ &= \underbrace{\sum_{i \notin \mathcal{S}} v_{1,i}^{(0)} \sigma\left(\langle w_i^{(0)}, x\rangle + b_{1,i}^{(0)}\right)}_{\stackrel{\text{def}}{=} B_1\left(x, v_1^{(0)}, W^{(0)}, b_1^{(0)}\right)} + \sum_{i \in \mathcal{S}} v_{1,i}^{(0)} \sigma\left(\langle w_i^{(0)}, x\rangle + b_{1,i}^{(0)}\right) , \end{aligned}$$

By definition of B_1, conditioning on the randomness of u, we know that B_1 is independent of α—because $\langle w_i^{(0)}, x\rangle + b_{1,i}^{(0)} = \langle \beta_i, x\rangle + b_{1,i}^{(0)}$ for $i \notin \mathcal{S}$. Since u and α are independent, we know that α and B_1 are independent by Proposition A.2. We continue to write

$$\begin{aligned} &n_1(x) - B_1 \\ &= \sum_{i \in \mathcal{S}} v_{1,i}^{(0)} \sigma\left(\langle w_i^{(0)}, x\rangle + b_{1,i}^{(0)}\right) \\ &= \sum_{i \in \mathcal{S}} v_{1,i}^{(0)} \mathbb{I}[\langle w_i^{(0)}, x\rangle + b_{1,i}^{(0)} \geq 0] \left(\frac{\alpha s_i}{\sqrt{|\mathcal{S}|}} x_d + \langle \beta_i, x\rangle + b_{1,i}^{(0)}\right) \\ &= \underbrace{\sum_{i \in \mathcal{S}} v_{1,i}^{(0)} \mathbb{I}[\langle w_i^{(0)}, x\rangle + b_{1,i}^{(0)} \geq 0] \frac{\alpha s_i}{\sqrt{|\mathcal{S}|}} x_d}_{\stackrel{\text{def}}{=} T_3} + \underbrace{\sum_{i \in \mathcal{S}} v_{1,i}^{(0)} \mathbb{I}[\langle w_i^{(0)}, x\rangle + b_{1,i}^{(0)} \geq 0] \left(\langle \beta_i, x\rangle + b_{1,i}^{(0)}\right)}_{\stackrel{\text{def}}{=} T_4} \end{aligned}$$

First consider T_3. For each $i \in \mathcal{S}$ we have

$$\begin{aligned} \left| s_i \cdot v_{1,i}^{(0)} - \frac{1}{\sqrt{m_2}} h\left(w_{i,1}^{(0)}, b_{1,i}^{(0)}\right)\right| &\leq O\left(\frac{1}{\sqrt{m_2}}\right) \quad \text{and} \\ \mathbb{E}_{v_{1,i}^{(0)}}\left[s_i \cdot v_{1,i}^{(0)} - \frac{1}{\sqrt{m_2}} h\left(w_{i,1}^{(0)}, b_{1,i}^{(0)}\right) \mid i \in \mathcal{S}\right] &= 0 \ . \end{aligned}$$

Above, the first row is because $\sqrt{m_2} v_{1,i}^{(0)} \in I_i$ and the Bounded-ness property of the interval (see Lemma A.4); and the second row by the Unbiased property of the interval (see Lemma A.4). By concentration, for fixed vector x, with high probability over the randomness of $V^{(0)}$:

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

In other words,

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

where $R_1 = R_1\left(x, v_1^{(0)}, W^{(0)}, b_1^{(0)}\right)$ satisfies $|R_1| \leq \widetilde{O}\left(\frac{|\alpha|}{\sqrt{|\mathcal{S}|}} \cdot \frac{\sqrt{|\mathcal{S}|}}{\sqrt{m_2}}\right) \leq \widetilde{O}\left(\frac{1}{\sqrt{m_1 m_2}}\right)$. We write

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

By (C.9) (i.e., the property of h), we know that for every fixed x, using concentration bound, with high probability over $W^{(0)}$ and $b^{(0)}$:

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

and thus by (C.10)

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

Let us define

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

Then,

$$T_3 = \rho\left(v_1^{(0)}, W^{(0)}, b_1^{(0)}\right) \cdot \phi_\varepsilon(\langle w^*, x\rangle) + R_1 + R_2\left(x, v_1^{(0)}, W^{(0)}, b_1^{(0)}\right) ,$$
$$\text{where } |R_2| \le \widetilde{O}\left(\frac{C}{\sqrt{\tau m_1}} \right) \times \frac{\tau}{C\sqrt{m_2}} = \widetilde{O}\left(\frac{\sqrt{\tau}}{\sqrt{m_1 m_2}} \right)$$

Note that α is independent of u so ρ is also independent of u. We can also define

$$\widetilde{\rho}\left(v_1^{(0)}\right) \overset{\text{def}}{=} \frac{\sqrt{\tau m_1}}{C\sqrt{m_2}} g \sim \mathcal{N}\left(0, \frac{\tau}{C^2 m_2}\right) .$$

and using (C.12) we can derive the desired bound on $\mathcal{W}_2(\rho|_{W^{(0)}, b_1^{(0)}}, \widetilde{\rho})$ in the statement of Lemma 6.3.

Next consider T_4. Recall

$$T_4 \overset{\text{def}}{=} \sum_{i \in \mathcal{S}} v_{1,i}^{(0)} \mathbb{I}[\langle w_i^{(0)}, x\rangle + b_{1,i}^{(0)} \ge 0] \left(\langle \beta_i, x\rangle + b_{1,i}^{(0)} \right)$$

For fixed unit vector x, with high probability, we have that

$$\left| \alpha \tfrac{s_i}{\sqrt{|\mathcal{S}|}} \right| = \left| \tfrac{\alpha}{\sqrt{|\mathcal{S}|}} \right| = \widetilde{O}\left(\tfrac{1}{\sqrt{|\mathcal{S}| m_1}} \right)$$

and therefore

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

By the above formula, for an index $i \in \mathcal{S}$ to have $\mathbb{I}[\langle \beta_i, x\rangle + b_{1,i}^{(0)} \ge 0] \neq \mathbb{I}[\langle w_i^{(0)}, x\rangle + b_{1,i}^{(0)} \ge 0]$, it must satisfy

$$\left| \left(\langle w_i^{(0)}, x\rangle + b_{1,i}^{(0)} \right) \right| \le \widetilde{O}\left(\frac{1}{\sqrt{|\mathcal{S}| m_1}} \right) \quad . \tag{C.13}$$

Thus, for fixed $\mathcal{S}$, since $W^{(0)}$ is independent of $\mathcal{S}$, with high probability over the randomness of $W^{(0)}$, there are at most $\widetilde{O}\left(\sqrt{|\mathcal{S}|}\right)$ many indices $i \in \mathcal{S}$ satisfying (C.13). In other words, using $|v_{1,i}^{(0)}| \le \widetilde{O}(1/\sqrt{m_2})$ with high probability, we have

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

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

To bound T_6. Recall

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

and we can define a similar notion

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

Observe that

$$h\left(w_{i,1}^{(0)}, b_{1,i}^{(0)} \right) = h\left(\beta_{i,1} + \langle \alpha u_i e_d, e_1 \rangle, b_{1,i}^{(0)} \right) = h\left(\beta_{i,1}, b_{1,i}^{(0)} \right) \quad .$$

Therefore, we have $\mathcal{S} = \mathcal{S}'$ and can write

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

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.2). In other words, setting $B = B_1 + B_2$, B and α (and therefore ρ) are independent.

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

As for the norm bound on B, recall

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

and by our random initialization, $n_1(x) \sim \mathcal{N}\big(0, \frac{1}{m_2} \big\| \sigma\big(W^{(0)} x + b_1^{(0)} \big) \big\|_2^2 \big)$. 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 $|n_1(x)| \leq \widetilde{O}(\frac{1}{\sqrt{m_2}})$, and this implies $|B| \leq \widetilde{O}(\frac{1}{\sqrt{m_2}})$ with high probability. □

C.1.4 Lemma 6.1: Existence

Lemma 6.1 (existence). *For every $\varepsilon \in \big(0, \frac{1}{k p_1 p_2^2 \mathfrak{C}_\mathfrak{s}(\Phi, p_2 \mathfrak{C}_\mathfrak{s}(\phi,1)) \mathfrak{C}_\mathfrak{s}(\phi,1)^2}\big)$, there exists*

$$M = \mathsf{poly}\left(\mathfrak{C}_\varepsilon\big(\Phi, \sqrt{p_2} \mathfrak{C}_\varepsilon(\phi, 1)\big), \frac{1}{\varepsilon} \right)$$
$$C_0 = \mathfrak{C}_\varepsilon(\Phi, \sqrt{p_2} \mathfrak{C}_\varepsilon(\phi, 1)) \cdot \mathfrak{C}_\varepsilon(\phi, 1) \cdot \widetilde{O}(p_1 \sqrt{p_2} k)$$

such that if $m_1, m_2 \geq M$, then with high probability, there exists weights W^*, V^* with

$$\|W^*\|_{2,\infty} = \max_i \|w_i^*\|_2 \leq \frac{C_0}{m_1}, \quad \|V^*\|_{2,\infty} = \max_i \|v_i^*\|_2 \leq \frac{\sqrt{m_1}}{m_2}$$

such that

$$\mathbb{E}_{(x,y)\sim\mathcal{D}} \left[\sum_{r=1}^{k} \left| f_r^*(x) - g_r^{(0)}(x; W^*, V^*) \right| \right] \leq \varepsilon,$$

and hence,

$$\mathbb{E}_{(x,y)\sim\mathcal{D}} \left[L(G^{(0)}(x; W^*, V^*), y) \right] \leq \mathsf{OPT} + \varepsilon.$$

Let us mostly focus on proving Lemma 6.1 for a single term

$$a^* \cdot \Phi\left(\sum_{j\in[p_2]} v^*_{1,j}\phi_{1,j}(\langle w^*_{1,j}, x\rangle)\right)\left(\sum_{j\in[p_2]} v^*_{2,j}\phi_{2,j}(\langle w^*_{2,j}, x\rangle)\right) \quad .$$

Extending it to multiple terms and multiple outputs, that is

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

is rather straightforward (and indeed a repetition of the proof of Lemma B.1 in the two-layer case).

The proof consists of several steps.

Step 1: Existence in expectation

Recall that the input (without bias) to each neuron at random initialization in the second hidden layer is

$$n_i(x) = \sum_{r\in[m_1]} v^{(0)}_{i,r}\sigma\left(\langle w^{(0)}_r, x\rangle + b^{(0)}_{1,r}\right)$$

We first use Lemma 6.3 to derive the following claim:

Claim C.3. *For every $\varepsilon \in (0, 1/\mathfrak{C}_\mathfrak{s}(\phi, 1))$, there exists real-value functions $\phi_{1,j,\varepsilon}(\cdot)$ satisfying*

$$|\phi_{1,j,\varepsilon}(\langle w^*_{1,j}, x\rangle) - \phi_{1,j}(\langle w^*_{1,j}, x\rangle)| \le \varepsilon \textit{ for all } j \in [p_2] \textit{ and } \|x\|_2 = 1,$$

and the following holds. Denote by

$$C \overset{\text{def}}{=} \mathfrak{C}_\varepsilon(\phi, 1), \quad C' \overset{\text{def}}{=} 10C\sqrt{p_2}, \quad \phi_{1,j,\varepsilon}(x) \overset{\text{def}}{=} \frac{1}{C'}\phi_{1,j,\varepsilon}(\langle w^*_{1,j}, x\rangle).$$

For every $i \in [m_2]$, there exist independent Gaussians[24]

$$\alpha_{i,j} \sim \mathcal{N}(0, 1/m_2) \textit{ and } \beta_i(x) \sim \mathcal{N}\left(0, \tfrac{1}{m_2}\left(1 - \textstyle\sum_{j\in[p_2]}\phi^2_{1,j,\varepsilon}(x)\right)\right) \quad ,$$

satisfying

$$\mathcal{W}_2\left(n_i(x), \sum_{j\in[p_2]}\alpha_{i,j}\phi_{1,j}(x) + \beta(x)\right) \le O\left(\frac{p_2^{2/3}}{m_1^{1/6}\sqrt{m_2}}\right)$$

Proof of Claim C.3. 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

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

Let us then denote $v_i[j,l]$ to be $(v_{i,r})_{r\in\mathcal{S}_{j,l}}$ and $W[j,l]$ to be $(W_r)_{r\in\mathcal{S}_{j,l}}$. Recall that the input (without bias) to each neuron at random initialization in the second hidden layer is

$$\begin{aligned} n_i(x) &= \sum_{r\in[m_1]} v^{(0)}_{i,r}\sigma\left(\langle w^{(0)}_r, x\rangle + b^{(0)}_{1,r}\right) \\ &= \sum_{j\in[p_2]}\sum_{l\in[S]}\sum_{r\in\mathcal{S}_{j,l}} v^{(0)}_{i,r}\sigma\left(\langle w^{(0)}_r, x\rangle + b^{(0)}_{1,r}\right) \quad . \end{aligned}$$

[24] More specifically, $\alpha_{i,j} = \alpha_{i,j}(v^{(0)}_i, W^{(0)}, b^{(0)}_1)$ and $\beta_i = \beta_i(x, v^{(0)}_i, W^{(0)}, b^{(0)}_1)$ depend on the randomness of $v^{(0)}_i$, $W^{(0)}$ and $b^{(0)}_1$.

For each $j \in [p_2]$ and $l \in [S]$, let us apply Lemma 6.3 to the summation $\sum_{r \in \mathcal{S}_{j,l}} \cdots$ in the above formula, to approximate $\phi_{1,j}(\langle w^*_{1,j}, x\rangle)$. (We need to replace m_1 with $\frac{m_1}{p_2 S}$ and scale up $W^{(0)}$ and $b_1^{(0)}$ by $\sqrt{p_2 S}$ before applying Lemma 6.3)). It tells us we can write $n_i(x)$ as:

$$\begin{aligned} n_i(x) = & \sum_{j\in[p_2], l \in [S]} \rho_j \left(v_i^{(0)}[j,l], W^{(0)}[j,l], b_1^{(0)}[j,l] \right) \phi_{1,j,\varepsilon}(\langle w^*_{1,j}, x\rangle) \\ & + \sum_{j\in[p_2], l \in [S]} B_j \left(x, v_i^{(0)}[j,l], W^{(0)}[j,l], b_1^{(0)}[j,l] \right) + R_j \left(x, v_i^{(0)}[j,l], W^{(0)}[j,l], b_1^{(0)}[j,l] \right) \end{aligned}$$

where random variables $\rho_{j,l} \stackrel{\text{def}}{=} \rho_j \left(v_i^{(0)}[j,l], W^{(0)}[j,l], b_1^{(0)}[j,l] \right) \sim \mathcal{N}(0, \frac{1}{100C^2 m_2 (p_2 S)})$ are independent Gaussian for different j and l. Let $C \stackrel{\text{def}}{=} \mathfrak{C}_\varepsilon(\phi, 1)$. We know that

$$\rho_j \stackrel{\text{def}}{=} \sum_{l\in[S]} \rho_{j,l} \sim \mathcal{N}\left(0, \frac{1}{100C^2 p_2 m_2}\right) = \mathcal{N}\left(0, \frac{1}{C'^2 m_2}\right)$$

for

$$C' \stackrel{\text{def}}{=} 10C\sqrt{p_2} \ .$$

Moreover, $|\phi_{1,j,\varepsilon}(\langle w^*_{1,j}, x\rangle) - \phi_{1,j}(\langle w^*_{1,j}, x\rangle)| \le \varepsilon$ for each $j \in [p_2]$.

Let us then denote

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

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

Lemma 6.3 tells us that random variables ρ_j are independent of $B_j(x)$, and with high probability

$$\begin{aligned} & |R^{\mathsf{s}}_j(x)| \le S \times \widetilde{O}\left(\frac{1}{\sqrt{(m_1/(p_2 S)) m_2}} \frac{1}{\sqrt{p_2 S}} \right) \le \widetilde{O}\left(\frac{S}{\sqrt{m_1 m_2}} \right) \text{ and} \\ & |B_j \left(x, v_i^{(0)}[j,l], W^{(0)}[j,l], b_1^{(0)}[j,l] \right)| \le \frac{1}{\sqrt{p_2 S}} \times \widetilde{O}\left(\frac{1}{\sqrt{m_2}} \right) = \widetilde{O}\left(\frac{1}{\sqrt{m_2 p_2 S}} \right) \end{aligned}$$

Let us apply the Wasserstein distance version of the central limit theorem (see for instance [19, Theorem 1]) [25]: since $B^{\mathsf{s}}_j(x)$ is the summation of S i.i.d random variables, there is a Gaussian random variable $\beta_j(x)$ only depending on the randomness of $B^{\mathsf{s}}_j(x)$ such that

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

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

$$\mathcal{W}_2 \left(n_i(x), \sum_{j\in[p_2]} \rho_j \phi_{1,j,\varepsilon}(\langle w^*_{1,j}, x\rangle) + \beta'(x) \right) \le \widetilde{O}\left(\frac{S p_2}{\sqrt{m_1 m_2}} + \frac{\sqrt{p_2}}{\sqrt{m_2 S}} \right)$$

[25] 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^{s}_j by $B^{\mathsf{s}}_j \mathbb{I}_{\text{all } B_j \le \widetilde{O}(1/\sqrt{m_2 p_2 S})}$. We choose to avoid writing this truncation in the paper to simply the presentation.

Let us slightly override notation and denote

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

We then have that variables $\alpha_{i,j} \overset{\text{def}}{=} C'\rho_j \sim \mathcal{N}(0, 1/m_2)$ are i.i.d. and

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

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, with high probability $\left\|\sigma\left(W^{(0)}x + b_1^{(0)}\right)\right\|_2^2 = 1 \pm \widetilde{O}\left(\frac{1}{\sqrt{m_1}}\right)$, we can write

$$\mathcal{W}_2(n_i(x), g) \le \widetilde{O}(\frac{1}{\sqrt{m_1 m_2}}) \quad \text{for } g \sim \mathcal{N}(0, \frac{1}{m_2}). \tag{C.14}$$

Since we can write $g = \sum_{j\in[p_2]} \alpha_{i,j}\phi_{1,j,\varepsilon}(x) + \beta_i(x)$ for

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

being an independent from $\alpha_{i,1}, \ldots, \alpha_{i,p_2}$, we conclude that (by choosing $S = (m_1/p_2)^{1/3}$)

$$\mathcal{W}_2\left(n_i(x), \sum_{j\in[p_2]} \alpha_{i,j}\phi_{1,j,\varepsilon}(x) + \beta_i(x)\right) \le \widetilde{O}\left(\frac{1}{\sqrt{m_1 m_2}} + \frac{Sp_2}{\sqrt{m_1 m_2}} + \frac{\sqrt{p_2}}{\sqrt{m_2 S}}\right) \le O\left(\frac{p_2^{2/3}}{m_1^{1/6}\sqrt{m_2}}\right) \quad . \tag{C.15}$$

This finishes the proof of Claim C.3. □

Claim C.4. *In the same notations as Claim C.3, there exists function $h\colon \mathbb{R}^2 \to [-C'', C'']$ for $C'' = \mathfrak{C}_\varepsilon(\Phi, C')$ such that for every $i \in [m_2]$,*

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

Proof of Claim C.4. Let us slightly override notation and denote

$$\phi_{2,j}(x) \overset{\text{def}}{=} \phi_{2,j}(\langle w^*_{2,j}, x\rangle).$$

We apply Lemma 6.2 again with ϕ chosen as $\Phi'(z) = \Phi(C'x)$.[26] We know there exists a function

[26] More specifically, we can choose w from Lemma 6.2 as $\left(\alpha_{i,1}, \ldots, \alpha_{i,p_2}, \beta_i/\sqrt{1 - \sum_{j\in[p_2]} \phi^2_{1,j,\varepsilon}(x)}\right)$, choose x from Lemma 6.2 as $\left(\phi_{1,1,\varepsilon}(x), \ldots, \phi_{1,p_2,\varepsilon}(x), \sqrt{1 - \sum_{j\in[p_2]} \phi^2_{1,j,\varepsilon}(x)}\right)$, choose w^* from Lemma 6.2 as $(v^*_{1,1}, \ldots, v^*_{1,p_2}, 0)$, and choose b_0 from Lemma 6.2 as $b^{(0)}_{2,i}$.

$h\colon \mathbb{R}^2 \to [-C'', C'']$ for $C'' = \mathfrak{C}_\varepsilon(\Phi', 1) = \mathfrak{C}_\varepsilon(\Phi, C')$ such that

$$\begin{aligned}
&\mathbb{E}_{\alpha_i,\beta_i}\left[\mathbb{I}_{\sum_{j\in[p_2]}\alpha_{i,j}\phi_{1,j,\varepsilon}(x)+\beta_i(x)+b^{(0)}_{2,i}\geq 0} h\left(\sum_{j\in[p_2]} v^*_{1,j}\alpha_{i,j}, b^{(0)}_{2,i}\right)\left(\sum_{j\in[p_2]} v^*_{2,j}\phi_{2,j}(x)\right)\right] \\
&\quad = \Phi\left(C'\sum_{j\in[p_2]} v^*_{1,j}\phi_{1,j,\varepsilon}(x)\right)\left(\sum_{j\in[p_2]} v^*_{2,j}\phi_{2,j}(x)\right) \pm \varepsilon C''' \qquad \text{(C.16)}
\end{aligned}$$

where

$$C''' = \sup_{x : \|x\|_2 \leq 1}\left|\sum_{j\in[p_2]} v^*_{2,j}\phi_{2,j}(x)\right| \leq p_2\mathfrak{C}_\mathfrak{s}(\phi, 1) \ .$$

Next, we wish to use the Wasserstein bound from Claim C.3 to replace $\sum_{j\in[p_2]}\alpha_{i,j}\phi_{1,j,\varepsilon}(x) + \beta_i(x)$ with $n_i(x)$. We derive that

$$\begin{aligned}
&\mathbb{E}\left[\mathbb{I}_{n_i(x)+b^{(0)}_{2,i}\geq 0} h\left(\sum_{j\in[p_2]} v^*_{1,j}\alpha_{i,j}, b^{(0)}_{2,i}\right)\left(\sum_{j\in[p_2]} v^*_{2,j}\phi_{2,j}(x)\right)\right] \\
&\overset{①}{=} \mathbb{E}\left[\mathbb{I}_{\alpha_{i,j}\phi_{1,j,\varepsilon}(x)+\beta_i(x)+b^{(0)}_{2,i}\geq 0} h\left(\sum_{j\in[p_2]} v^*_{1,j}\alpha_{i,j}, b^{(0)}_{2,i}\right)\left(\sum_{j\in[p_2]} v^*_{2,j}\phi_{2,j}(x)\right)\right] \\
&\qquad \pm O\left(\mathcal{W}_2\left(n_i(x), \sum_{j\in[p_2]}\alpha_{i,j}\phi_{1,j,\varepsilon}(x) + \beta_i(x)\right)\sqrt{m_2}C'''C''\right) \qquad \text{(C.17)} \\
&\overset{②}{=} \Phi\left(C'\sum_{j\in[p_2]} v^*_{1,j}\phi_{1,j,\varepsilon}(x)\right)\left(\sum_{j\in[p_2]} v^*_{2,j}\phi_{2,j}(x)\right) \pm O\left(\varepsilon C''' + \frac{C''C'''p_2^{2/3}}{m_1^{1/6}}\right) \\
&\overset{③}{=} \Phi\left(\sum_{j\in[p_2]} v^*_{1,j}\phi_{1,j}(\langle w^*_{1,j}, x\rangle)\right)\left(\sum_{j\in[p_2]} v^*_{2,j}\phi_{2,j}(\langle w^*_{2,j}, x\rangle)\right) \pm O\left(\varepsilon C''' + C'''\cdot(\varepsilon p_2)L_\Phi + \frac{C''C'''p_2^{2/3}}{m_1^{1/6}}\right) \\
&\overset{④}{=} \Phi\left(\sum_{j\in[p_2]} v^*_{1,j}\phi_{1,j}(\langle w^*_{1,j}, x\rangle)\right)\left(\sum_{j\in[p_2]} v^*_{2,j}\phi_{2,j}(\langle w^*_{2,j}, x\rangle)\right) \pm O\left(p_2^2\mathfrak{C}_\mathfrak{s}(\Phi, p_2\mathfrak{C}_\mathfrak{s}(\phi, 1))\mathfrak{C}_\mathfrak{s}(\phi, 1)\varepsilon\right) \ .
\end{aligned}$$

Above, ① uses the fact that $n_i \sim \mathcal{N}\left(0, \frac{1}{m_2}\left\|\sigma\left(W^{(0)}x + b^{(0)}_1\right)\right\|_2^2\right)$ with $\left\|\sigma\left(W^{(0)}x + b^{(0)}_1\right)\right\|_2 \leq 2$ w.h.p. and $|h| \leq C''$ is bounded. ② uses Claim C.3 and (C.16). ③ uses $C'\phi_{1,j,\varepsilon}(x) = \phi_{1,j,\varepsilon}(\langle w^*_{1,j}, x\rangle)$, $\phi_{2,j}(x) = \phi_{2,j}(\langle w^*_{2,j}, x\rangle)$, and denote by L_Φ the Lipschitz continuity parameter of Φ (namely, $|\Phi(x) - \Phi(y)| \leq L_\Phi|x - y|$ for all $x, y \in \big[-p_2\mathfrak{C}_\mathfrak{s}(\phi, 1), p_2\mathfrak{C}_\mathfrak{s}(\phi, 1)\big]$)). ④ uses $L_\Phi \leq \mathfrak{C}_\mathfrak{s}(\Phi, p_2\mathfrak{C}_\mathfrak{s}(\phi, 1))$ and our assumption $m_1 \geq M$. This proves Claim C.4. □

Step 2: From expectation to finite neurons

Intuitively, we wish to apply concentration bound on Claim C.4 with respect to all neurons $i \in [m_2]$ on the second layer. Recall $a_i \sim \mathcal{N}(0, \varepsilon_a)$ is the weight of the i-th neuron at the output layer. Our main result of Step 2 is the following claim.

Claim C.5. *In the same notation as Claim C.4,*

$$\frac{1}{m_2}\sum_{i\in[m_2]}\left[\frac{a_i^2}{\varepsilon_a^2}\mathbb{I}_{n_i(x)+b_{2,i}^{(0)}\geq 0}h\left(\sum_{j\in[p_2]}v_{1,j}^*\alpha_{i,j},b_{2,i}^{(0)}\right)\left(\sum_{j\in[p_2]}v_{2,j}^*\phi_{2,j}(\langle w_{2,j}^*,x\rangle)\right)\right]$$
$$=\Phi\left(\sum_{j\in[p_2]}v_{1,j}^*\phi_{1,j}(\langle w_{1,j}^*,x\rangle)\right)\left(\sum_{j\in[p_2]}v_{2,j}^*\phi_{2,j}(\langle w_{2,j}^*,x\rangle)\right)\pm O\left(p_2^2\mathfrak{C}_\mathfrak{s}(\Phi,p_2\mathfrak{C}_\mathfrak{s}(\phi,1))\mathfrak{C}_\mathfrak{s}(\phi,1)\varepsilon\right)$$

Proof of Claim C.5. Across different choices of i, the values of $n_i(x)$ and $\alpha_{i,j}$ can be correlated. This makes it not a trivial thing to apply concentration. In the remainder of this proof, let us try to modify the two quantities to make them independent across i.

First modify $\boldsymbol{\alpha_{i,j}}$. 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 6.3, 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 6.3, we know that with high probability over $W^{(0)}$:[27]

$$\widetilde{\rho}_j\sim\mathcal{N}\left(0,\frac{1}{C'^2m_2}\right)\quad\text{and}\quad\mathcal{W}_2(\rho_{j,l}|_{W^{(0)},b^{(0)}},\widetilde{\rho}_{j,l})\leq\widetilde{O}\left(\frac{1}{C\sqrt{m_1m_2}}\right)$$
$$\Longrightarrow\mathcal{W}_2(\rho_j|_{W^{(0)},b^{(0)}},\widetilde{\rho}_j)\leq\widetilde{O}\left(\frac{S}{C\sqrt{m_1m_2}}\right)\tag{C.18}$$

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

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

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

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

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

is a Gaussian variable and is independent of u. As a consequence, the quantities $\widetilde{n}_i(x)$ are independent among different choices of $i\in[m_1]$. Using standard concentration on $\|u\|_2$ (see the line above (C.14)), we have for every x with $\|x\|_2=1$,

$$\mathcal{W}_2(n_i(x),\widetilde{n}_i(x))\leq\widetilde{O}\left(\frac{1}{\sqrt{m_1m_2}}\right)\tag{C.19}$$

Concentration. Using Claim C.4 of Step 1 we have

$$\mathbb{E}\left[\mathbb{I}_{n_i(x)+b_{2,i}^{(0)}\geq 0}h\left(\sum_{j\in[p_2]}v_{1,j}^*\alpha_{i,j},b_{2,i}^{(0)}\right)\left(\sum_{j\in[p_2]}v_{2,j}^*\phi_{2,j}(\langle w_{2,j}^*,x\rangle)\right)\right]$$
$$=\Phi\left(\sum_{j\in[p_2]}v_{1,j}^*\phi_{1,j}(\langle w_{1,j}^*,x\rangle)\right)\left(\sum_{j\in[p_2]}v_{2,j}^*\phi_{2,j}(\langle w_{2,j}^*,x\rangle)\right)\pm O\left(p_2^2\mathfrak{C}_\mathfrak{s}(\Phi,p_2\mathfrak{C}_\mathfrak{s}(\phi,1))\mathfrak{C}_\mathfrak{s}(\phi,1)\varepsilon\right)$$

[27] Recall from the proof of Claim C.3 that, we need to replace m_1 with $\frac{m_1}{p_2S}$ and scale up $W^{(0)}$ and $b_1^{(0)}$ by $\sqrt{p_2S}$ before applying Lemma 6.3.

Using the notions of $\widetilde{n}$ and $\widetilde{\alpha}$ and the Wasserstein distance bounds (C.18) and (C.19), it implies[28]

$$\begin{aligned}&\mathbb{E}\left[\mathbb{I}_{\widetilde{n}_i(x)+b_{2,i}^{(0)}\geq 0} h\left(\sum_{j\in[p_2]} v_{1,j}^*\widetilde{\alpha}_{i,j}, b_{2,i}^{(0)}\right)\left(\sum_{j\in[p_2]} v_{2,j}^*\phi_{2,j}(\langle w_{2,j}^*, x\rangle)\right)\right]\\&= \Phi\left(\sum_{j\in[p_2]} v_{1,j}^*\phi_{1,j}(\langle w_{1,j}^*, x\rangle)\right)\left(\sum_{j\in[p_2]} v_{2,j}^*\phi_{2,j}(\langle w_{2,j}^*, x\rangle)\right) \pm O\left(p_2^2\mathfrak{C}_\mathfrak{s}(\Phi, p_2\mathfrak{C}_\mathfrak{s}(\phi,1))\mathfrak{C}_\mathfrak{s}(\phi,1)\varepsilon\right)\end{aligned}$$

Using $\mathbb{E}_{a_i}[\frac{a_i^2}{\varepsilon_a^2}] = 1$ and applying standard concentration—and the independence of tuples $(a_i, \widetilde{n}_i(x), (\widetilde{\alpha}_{i,j})_{j\in[p_2]})$ with respect to different choices of i— we know with high probability

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

Using again the Wasserstein distance bounds bounds (C.18) and (C.19), we can combine the above two equations to derive that w.h.p.

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

□

Step 3: From finite neurons to the network

We now construct the network and prove Lemma 6.1. Recall we focus on constructing

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

We shall explain towards the end how to extend it to multiple terms and multiple outputs.

For each $\phi_{2,j}$, applying Lemma 6.2b, we can construct $h_{\phi,j}\colon \mathbb{R}^2 \to [-C, C]$ satisfying

$$\forall i' \in [m_1]\colon \quad \mathbb{E}\left[h_{\phi,j}\left(\langle w_{2,j}^*, w_{i'}^{(0)}\rangle, b_{1,i'}^{(0)}\right)\mathbb{I}_{\langle w_{i'}^{(0)}, x\rangle + b_{1,i'}^{(0)}\geq 0}\right] = \phi_{2,j}(\langle w_{2,j}^*, x\rangle) \pm \varepsilon \quad . \qquad \text{(C.20)}$$

Now consider an arbitrary vector $v \in \mathbb{R}^{m_1}$ with $v_i \in \{-1, 1\}$.

- Define $V^\divideontimes \in \mathbb{R}^{m_2\times m_1}$ as

$$V^\divideontimes = (C_0\cdot C''/C)^{-1/2}\frac{a^*}{m_2}\left(a_i h\left(\sum_{j\in[p_2]} v_{1,j}^*\alpha_{i,j}, b_{2,i}^{(0)}\right)v^\top\right)_{i\in[m_2]}$$

[28] We have skipped the details since it is analogous to (C.17).

where the notations of $h\colon \mathbb{R}^2 \to [-C'', C'']$ and $\alpha_{i,j}$ come from Claim C.4. We have

$$\|V^\divideontimes\|_{2,\infty} \leq \frac{\widetilde{O}(\sqrt{C''C})}{\sqrt{C_0}} \cdot \frac{\sqrt{m_1}}{m_2}$$

- Define $W^\divideontimes \in \mathbb{R}^{m_1 \times d}$ as

$$W^\divideontimes = (C_0 \cdot C''/C)^{1/2} \frac{2}{\varepsilon_a^2 m_1} \left(v_i \sum_{j \in [p_2]} v_{2,j}^* h_{\phi,j} \left(\langle w_{2,j}^*, w_i^{(0)} \rangle, b_{1,i}^{(0)} \right) e_d \right)_{i \in [m_1]} .$$

(If one wants the more general target in Remark 4.2, we can replace the use of e_d with w_4^* following the similar treatment as (B.1) in the two-layer existential proof.) We have

$$\|W^\divideontimes\|_{2,\infty} \leq \frac{2\sqrt{p_2}\sqrt{C_0 \cdot C''C}}{m_1},$$

Given these weights, suppose the signs of ReLU's are determined by the random initialization (i.e., by $W^{(0)}$ and $V^{(0)}$). We can consider the network output

$$\begin{aligned}
g^{(0)}(x; W^\divideontimes, V^\divideontimes) &\overset{\text{def}}{=} a D_{v,x} V^\divideontimes D_{w,x} W^\divideontimes x \\
&= \sum_{i \in [m_2]} a_i \mathbb{I}_{n_i(x) + b_{2,i}^{(0)} \geq 0} \sum_{i' \in [m_1]} v_{i,i'}^* \langle w_i^\divideontimes, x \rangle \mathbb{I}_{\langle w_{i'}^{(0)}, x \rangle + b_{1,i'}^{(0)} \geq 0} \\
&= \frac{a^*}{m_2} \sum_{i \in [m_2]} \frac{a_i^2}{\varepsilon_a^2} h \left(\sum_{j \in [p_2]} v_{1,j}^* \alpha_{i,j}, b_{2,i}^{(0)} \right) \mathbb{I}_{n_i(x) + b_{2,i}^{(0)} \geq 0} \\
&\quad \times \sum_{j \in [p_2]} v_{2,j}^* \left(\frac{1}{m_1} \sum_{i' \in [m_1]} h_{\phi,j} \left(\langle w_{2,j}^*, w_{i'}^{(0)} \rangle, b_{1,i'}^{(0)} \right) \mathbb{I}_{\langle w_{i'}^{(0)}, x \rangle + b_{1,i'} \geq 0} \right) \qquad \text{(C.21)}
\end{aligned}$$

Above, we have used $\langle e_d, x \rangle = x_d = 1/2$. Since $h_{\phi,j} \in [-C'', C'']$, we can apply concentration on (C.20) and get (recalling m_1 is sufficiently large) w.h.p

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

Using Claim C.5, we have

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

Putting this into (C.21), and using $a^* \in [-1, 1]$ and $h \in [-C'', C'']$, we know that with high probability

$$\begin{aligned}
g^{(0)}(x; W^\divideontimes, V^\divideontimes) = a^* \cdot \Phi \left(\sum_{j \in [p_2]} v_{1,j}^* \phi_{1,j}(\langle w_{1,j}^*, x \rangle) \right) \left(\sum_{j \in [p_2]} v_{2,j}^* \phi_{2,j}(\langle w_{2,j}^*, x \rangle) \right) \\
\pm O\left(p_2^2 \mathfrak{C}_\mathfrak{s}(\Phi, p_2 \mathfrak{C}_\mathfrak{s}(\phi, 1)) \mathfrak{C}_\mathfrak{s}(\phi, 1) \varepsilon \right)
\end{aligned}$$

Finally, scaling down the value of ε by factor $p_2^2 \mathfrak{C}_\mathfrak{s}(\Phi, p_2 \mathfrak{C}_\mathfrak{s}(\phi, 1)) \mathfrak{C}_\mathfrak{s}(\phi, 1)$ we complete the proof of Lemma 6.1 for a single output and for a single Φ.

Remark C.6. The proof generalizes to fitting a combination of multiple functions

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

in the same way as the proof of Lemma B.1, if we choose the vector $v \in \{-1,1\}^{m_1}$ uniformly at random and apply concentration. Note that we have to further scale down ε by $\frac{1}{p_1}$ for the same reason as Lemma B.1, and the norm of V^* shall grow by a factor of p_1.

Remark C.7. The proof generalizes to multiple outputs in the same way as Lemma B.1, using the fact that weights $a_{r,i}$ are independent across different outputs r. Note that we have to further scale down ε by $\frac{1}{k}$ for the same reason as Lemma B.1, and the norm of V^* shall grow by a factor of k.

Finally, applying the two remarks above, we have

$$\|V^*\|_{2,\infty} \le \frac{\widetilde{O}(\sqrt{C''C})p_1 k}{\sqrt{C_0}} \cdot \frac{\sqrt{m_1}}{m_2} \quad \text{and} \quad \|W^*\|_{2,\infty} \le \frac{2\sqrt{p_2}\sqrt{C_0 \cdot C''C}}{m_1},$$

We thus finish the proof of Lemma 6.1 with our choice of C_0. ■

C.2 Coupling

C.2.1 Lemma 6.4: Coupling

Lemma 6.4 (coupling, restated)**.** *Suppose* $\tau_v \in [0,1]$, $\tau_w \in \big(\frac{1}{m_1^{3/2}}, \frac{1}{m_1^{1/2}}\big]$, $\sigma_w \in \big(\frac{1}{m_1^{3/2}}, \frac{\tau_w}{m_1^{1/4}}\big]$, $\sigma_v \in \big(0, \frac{1}{m_2^{1/2}}\big]$ *and* $\eta > 0$. *Given fixed unit vector* x*, and perturbation matrices* W', V', W'', V'' *(that may depend on the randomness of* $W^{(0)}, b_1^{(0)}, V^{(0)}, b_2^{(0)}$ *and* x*) satisfying*

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

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

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

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

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

$$\begin{aligned} z_0 &\stackrel{\text{def}}{=} D_{w,x} W^{(0)} x \\ z_2 &\stackrel{\text{def}}{=} (D_{w,x} + D'_{w,x})(W^{(0)} + W^\rho + W')x - D_{w,x}W^{(0)}x \ . \end{aligned}$$

We let

- diagonal matrix $D_{w,x}$ denotes the sign of ReLU's at weights $W^{(0)}$,

- diagonal matrix $D_{w,x} + D''_{w,x}$ denotes the sign of ReLU's at weights $W^{(0)} + W^\rho$, and
- diagonal matrix $D_{w,x} + D'_{w,x}$ denotes the sign of ReLU's at weights $W^{(0)} + W^\rho + W'$.

Sign change in $\boldsymbol{D''_{w,x}}$. 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, by standard property of Gaussian, for each i, we have $\mathbf{Pr}[|W_i^\rho x| \geq |W_i^{(0)} x|] \leq \widetilde{O}\left(\sigma_w \sqrt{m_1}\right)$. 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

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

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

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

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

$$\|D''_{w,x}\|_0 \leq \tau_w m_1^{5/4} \quad \text{and} \quad \|D''_{w,x} W^{(0)} x\|_2 \leq \tau_w^{3/2} m_1^{3/8} \tag{C.22}$$

Sign change in $\boldsymbol{D'_{w,x} - D''_{w,x}}$. 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 $|((W^{(0)} + W^\rho)x_i| \leq s'' \overset{\text{def}}{=} \frac{2\tau_w}{s^{1/4}}$ 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 $\frac{s}{2} \times (s'')^4 > \tau_w^4$. Therefore, we must have

$$\begin{aligned} & s \leq \widetilde{O}(s'' m^{3/2}) = \widetilde{O}\left(\frac{\tau_w}{s^{1/4}} m_1^{3/2}\right) \\ \Longrightarrow \ & \|D'_{w,x} - D''_{w,x}\|_0 = s \leq \widetilde{O}\left(\tau_w^{4/5} m_1^{6/5}\right) \ . \end{aligned} \tag{C.23}$$

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

$$\begin{aligned} \| \left(D'_{w,x} - D''_{w,x}\right) (W^{(0)} + W^\rho) x\|_2 &\leq \sqrt{\sum_{i, (D'_{w,x} - D''_{w,x})_{i,i} \neq 0} (W'x)_i^2} \\ &\leq \sqrt{\sqrt{s \cdot \sum_{i, (D'_{w,x} - D''_{w,x})_{i,i} \neq 0} (W'x)_i^4}} \leq O\left(s^{1/4} \tau_w\right) = \widetilde{O}\left(\tau_w^{6/5} m_1^{3/10}\right) \end{aligned} \tag{C.24}$$

Sum up: First Layer. Combining (C.22) and (C.23) and using $\tau_w \leq m_1^{-1/4}$ from assumption, we have

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

By the 1-Lipschitz continuity of ReLU, we know that w.h.p.

$$\begin{aligned}\|z_2\|_2 &= \left\|(D_{w,x}+D'_{w,x})(W^{(0)}+W^\rho+W')x - D_{w,x}W^{(0)}x\right\|_2\\ &= \left\|\sigma(W^{(0)}+W^\rho+W')x) - \sigma(W^{(0)}x)\right\|_2\\ &\le \|(W^\rho+W')x\|_2 \le \|W'x\|_2 + \|W^\rho x\|_2 \le \widetilde{O}\left(\tau_w m_1^{1/4} + \sigma_w m_1^{1/2}\right) \le \widetilde{O}\left(\tau_w m_1^{1/4}\right)\end{aligned}$$

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

Second Layer Sign Change. The sign change in the second layer is caused by input vector

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

Here, using w.h.p. $\|z\|_2 \le \widetilde{O}(1)$, we have

$$\begin{aligned}\|V^\rho(z_0+z_2)\|_\infty &\le \widetilde{O}(\sigma_v)\cdot(\|z_0\|_2+\|z_2\|_2) \le \widetilde{O}(\sigma_v)\\ \|V^{(0)}z_2 + V'(z_0+z_2)\|_2 &\le \widetilde{O}\left(\tau_v + \|z_2\|_2\right) \le \widetilde{O}\left(\tau_v + \tau_w m_1^{1/4}\right)\end{aligned}$$

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.8), we can bound

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

■

Part II, Diagonal Cross Term

Recall

$$\begin{aligned}g_r(x;W,V) &= a_r(D_{v,x}+D'_{v,x})\left(V(D_{w,x}+D'_{w,x})\left(Wx+b_1\right)+b_2\right)\\ g_r^{(b)}(x;W,V) &= a_r(D_{v,x}+D'_{v,x})V(D_{w,x}+D'_{w,x})(Wx+b_1)\\ g_r^{(b,b)}(x;W,V) &= a_r(D_{v,x}+D'_{v,x})V(D_{w,x}+D'_{w,x})Wx\end{aligned}$$

and one can carefully check that

$$\begin{aligned}&g_r(x;W^{(0)}+W^\rho+W'+\eta\mathbf{\Sigma}W'', V^{(0)}+V^\rho+V'+\eta V''\mathbf{\Sigma})\\ &= g_r\left(x;W^{(0)}+W^\rho+W', V^{(0)}+V^\rho+V'\right) + g_r^{(b,b)}(\eta\mathbf{\Sigma}W'', \eta V''\mathbf{\Sigma})\\ &\quad + \underbrace{g_r^{(b)}(x;W^{(0)}+W^\rho+W', \eta V''\mathbf{\Sigma}) + g_r^{(b,b)}(x;\eta\mathbf{\Sigma}W'', V^{(0)}+V^\rho+V')}_{\text{error terms}}\end{aligned}$$

We consider the last two error terms.

First error term. The first term is

$$\begin{aligned}&g_r^{(b)}(x;W^{(0)}+W^\rho+W', \eta V''\mathbf{\Sigma})\\ &= \eta a_r(D_{v,x}+D'_{v,x})V''\mathbf{\Sigma}(D_{w,x}+D'_{w,x})((W^{(0)}+W^\rho+W')x+b_1)\end{aligned}$$

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

$$\begin{aligned}&\|(D_{w,x}+D'_{w,x})((W^{(0)}+W^\rho+W')x+b_1)\|_\infty\\ &\le \|(W^{(0)}+W^\rho+W')x+b_1\|_\infty \le \widetilde{O}\left(\tau_w+\sigma_w+\frac{1}{m_1^{1/2}}\right) \le \widetilde{O}\left(\frac{1}{m_1^{1/2}}\right) \qquad \text{(C.25)}\end{aligned}$$

and we have $\|a_r(D_{v,x} + D'_{v,x})V''\|_2 \le \widetilde{O}\left(\tau_v m_2^{1/2}\right)$. By Fact C.9, using the randomness of $\mathbf{\Sigma}$, with high probability

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

Second error term. We write down the second error term

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

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 $\|W''x\|_\infty \le \tau_w$, by Fact C.9, using the randomness of $\mathbf{\Sigma}$ we know that w.h.p.

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

- For the second term, since $\|W''x\|_2 \le \tau_w m_1^{1/4}$ and w.h.p.[29]

$$\|a_r D_{v,x}(V^{(0)} + V^\rho)(D_{w,x} + D'_{w,x})\|_\infty \le \|a_r D_{v,x}(V^{(0)} + V^\rho)\|_\infty \le \widetilde{O}(1)$$

  by Fact C.9, using the randomness of $\mathbf{\Sigma}$ we know that w.h.p.

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

- For the third term, again by $\|W''x\|_\infty \le \tau_w$ and Fact C.9, we have: w.h.p.

$$\begin{aligned} |a_r D'_{v,x}(V^{(0)} + V^\rho)\mathbf{\Sigma}W''x| &\le \widetilde{O}\left(\|a_r D'_{v,x}(V^{(0)} + V^\rho)\|_2\tau_w\right) \\ &\le \widetilde{O}\left(m_2^{1/2}\tau_w\right). \end{aligned}$$

■

Tool

The following are some tools used in the above proofs.

A variant of the following claim has appeared in [4].

Claim C.8. *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*

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

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

[29] We note that the derivation of $\|a_r D_{v,x}(V^{(0)} + V^\rho)\|_\infty \le \widetilde{O}(1)$ may be non-trivial for some readers, because $D_{v,x}$ is dependent of the randomness of $V^{(0)} + V^{(\rho)}$. In fact, for every fixed basis vector e_j, we have w.h.p. $\|D_{v,x}(V^{(0)} + V^\rho)e_j\|_2 \le \widetilde{O}(1)$, and thus by the randomness of a_r it satisfies $|a_r D_{v,x}(V^{(0)} + V^\rho)e_j| \le \widetilde{O}(1)$. Taking a union bound over $j = 1, 2, \ldots, m_2$ gives the bound. Such proof idea was repeatedly used in [4].

we have

$$\|x\|_0 \leq O(m_2\|g_1'\|^{2/3} + m_2^{3/2}\|g_2'\|_\infty) \quad \textit{and}$$
$$\|x\|_1 \leq O\big(m_2^{1/2}\|g_1'\|^{4/3} + m_2^{3/2}\|g_2'\|_\infty^2\big) \ .$$

Proof of Claim C.8. 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

$$|(g_1' + g_2')_j| > |(g)_j| \ . \tag{C.26}$$

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 (C.26), 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

$$\|x\|_1 \leq O\Big(\frac{\|g_1'\|^2}{\xi} + \|g_1'\|^2\sqrt{\xi}m_2^{3/4} + \xi^2 m_2^{3/2}\Big) \leq O\Big(\frac{\|g_1'\|^2}{\xi} + \xi^2 m_2^{3/2}\Big) \ .$$

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. □

Fact C.9. *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.2.2 Corollary 6.5: Existence After Coupling

Corollary 6.5 is a corollary to Lemma 6.1 with $g^{(0)}$ replaced with $g^{(b,b)}$. Recall that $g^{(b,b)}$ is different from $g^{(0)}$ only by the diagonal signs, namely,

$$g_r^{(0)}(x; W^*, V^*) = a_r D_{v,x} V^* D_{w,x} W^* x$$
$$g_r^{(b,b)}(x; W^*, V^*) = a_r (D_{v,x} + D'_{v,x}) V^* (D_{w,x} + D'_{w,x}) W^* x$$

where $D_{v,x} + D'_{v,x}$ and $D_{w,x} + D'_{w,x}$ are the diagonal sign matrices determined at $W^{(0)} + W' + W^\rho$, $V^{(0)} + V' + V^\rho$.

Corollary 6.5 (existence after coupling)**.** *In the same setting as Lemma 6.1, perturbation matrices W', V' (that may depend on the randomness of the initialization and $\mathcal{D}$) with*

$$\|W'\|_{2,4} \le \tau_w, \|V'\|_F \le \tau_v \enspace .$$

Using parameter choices from Table 1, w.h.p. there exist W^* and V^* (independent of the randomness of W^ρ, V^ρ) satisfying

$$\|W^*\|_{2,\infty} = \max_i \|w_i^*\|_2 \le \frac{C_0}{m_1}, \quad \|V^*\|_{2,\infty} = \max_i \|v_i^*\|_2 \le \frac{\sqrt{m_1}}{m_2}$$

$$\mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\sum_{r=1}^k \left| f_r^*(x) - g_r^{(b,b)}(x; W^*, V^*)\right|\right] \le \varepsilon,$$
$$\mathbb{E}_{(x,y)\sim\mathcal{D}}\left[L(G^{(b,b)}(x; W^*, V^*), y)\right] \le \mathsf{OPT} + \varepsilon.$$

Proof of Corollary 6.5. The idea to prove Corollary 6.5 is simple. First construct W^* and V^* from Lemma 6.1, and then show that $g^{(0)}$ and $g^{(b,b)}$ are close using Lemma 6.4. By Lemma 6.4 and our parameter choices Table 1, we know

$$\|D'_{w,x}\|_0 \le \widetilde{O}(\tau_w^{4/5} m_1^{6/5}) \ll O(m_1)$$
$$\|D'_{v,x}\|_0 \le \widetilde{O}\left(\sigma_v m_2^{3/2} + \tau_v^{2/3} m_2 + \tau_w^{2/3} m_1^{1/6} m_2\right) \le \widetilde{O}((\varepsilon/C_0)^{\Theta(1)} m_2)$$

Now, recall that Lemma 6.1 says $\|W^*\|_{2,\infty} \le \tau_{w,\infty}$ and $\|V^*\|_{2,\infty} \le \tau_{v,\infty}$ for $\tau_{w,\infty}\tau_{v,\infty} = \frac{C_0}{\sqrt{m_1} m_2}$. Therefore,

$$|a_r D_{v,x} V^* D'_{w,x} W^* x| = \left| \sum_{j\in[m_2]} a_{r,j} (D_{v,x})_{j,j} \langle v_j^*, D'_{w,x} W^* x\rangle \right| \le \widetilde{O}(m_2 \tau_{v,\infty}) \|D'_{w,x} W^* x\|_2$$
$$\le \widetilde{O}(m_2\tau_{v,\infty}) \sqrt{\|D'_{w,x}\|_0} \tau_{w,\infty} \ll \varepsilon$$

$$|a_r D'_{v,x} V^* D_{w,x} W^* x| = \left| \sum_{j\in[m_2]} a_{r,j} (D'_{v,x})_{j,j} \langle v_j^*, D_{w,x} W^* x\rangle \right| \le \widetilde{O}(\|D'_{v,x}\|_0 \tau_{v,\infty}) \|D_{w,x} W^* x\|_2$$
$$\le \widetilde{O}(\|D'_{v,x}\|_0 \tau_{v,\infty}) \cdot O(\sqrt{m_1} \tau_{w,\infty}) \ll \varepsilon$$

In other words,

$$|g^{(b,b)}(x; W^*, V^*) - g^{(0)}(x; W^*, V^*)| \le 2\varepsilon \enspace . \qquad \square$$

C.2.3 Lemma 6.8: Smoothed Real vs Pseudo Networks

Recall

$$\begin{aligned}P_{\rho,\eta} &\stackrel{\text{def}}{=} f_r(x; W + W^\rho + \eta\mathbf{\Sigma}W'', V + V^\rho + \eta V''\mathbf{\Sigma})\\ &= a_r D_{v,x,\rho,\eta}\Big((V + V^\rho + \eta V''\mathbf{\Sigma})D_{w,x,\rho,\eta}\left((W + W^\rho + \eta\mathbf{\Sigma}W'')x + b_1\right) + b_2\Big)\\ P'_{\rho,\eta} &\stackrel{\text{def}}{=} g_r(x; W + W^\rho + \eta\mathbf{\Sigma}W'', V + V^\rho + \eta V''\mathbf{\Sigma})\\ &= a_r D_{v,x,\rho}\Big((V + V^\rho + \eta V''\mathbf{\Sigma})D_{w,x,\rho}\left((W + W^\rho + \eta\mathbf{\Sigma}W'')x + b_1\right) + b_2\Big).\end{aligned}$$

Lemma 6.8 (smoothed real vs pseudo)**.** *There exists $\eta_0 = \frac{1}{\mathsf{poly}(m_1,m_2)}$ such that, for every $\eta \le \eta_0$, for every fixed x with $\|x\|_2 = 1$, for every W', V', W'', V'' that may depend on the randomness of the initialization and x, with*

$$\|W'\|_{2,4} \le \tau_w, \quad \|V'\|_{2,2} \le \tau_v, \quad \|W''\|_{2,\infty} \le \tau_{w,\infty}, \quad \|V''\|_{2,\infty} \le \tau_{v,\infty}$$

we have with high probability:

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

where O_p hides polynomial factor of m_1, m_2.

Proof of Lemma 6.8. Since $P_{\rho,\eta}$ and $P'_{\rho,\eta}$ only differ in the sign pattern, we try to bound the (expected) output difference $P_{\rho,\eta} - P'_{\rho,\eta}$ by analyzing these sign changes. We use the same proof structure as Lemma 6.4, that is to first bound the sign changes in the first layer, and then the second layer. One can carefully verify

$$\begin{aligned}P_{\rho,\eta} - P'_{\rho,\eta} = &\underbrace{a_r D_{v,x,\rho}(V + V^\rho + \eta V''\mathbf{\Sigma})\left(D_{w,x,\rho,\eta} - D_{w,x,\rho}\right)\left((W + W^\rho + \eta\mathbf{\Sigma}W'')x + b_1\right)}_{\clubsuit}\\ &+ \underbrace{a_r\left(D_{v,x,\rho,\eta} - D_{v,x,\rho}\right)\Big((V + V^\rho + \eta V''\mathbf{\Sigma})D_{w,x,\rho,\eta}\left((W + W^\rho + \eta\mathbf{\Sigma}W'')x + b_1\right) + b_2\Big)}_{\spadesuit}\ .\end{aligned}$$

We call ♣ the output difference caused by sign change of the first layer; and ♠ that caused by the sign change of the second layer.

Sign Change of First Layer. We write

$$\begin{aligned}z &= D_{w,x,\rho}\left((W + W^\rho + \eta\mathbf{\Sigma}W'')x + b_1\right)\\ z + z' &= D_{w,x,\rho,\eta}\left((W + W^\rho + \eta\mathbf{\Sigma}W'')x + b_1\right)\ .\end{aligned}$$

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

$$\|z'\|_\infty \le \eta\tau_{w,\infty}\ .$$

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

$$\forall i \in [m_1]: \quad \Pr_{W^\rho}[z'_i \neq 0] \le \widetilde{O}\left(\frac{\eta\tau_{w,\infty}}{\sigma_w}\right). \tag{C.27}$$

One consequence of (C.27) is $\mathbf{Pr}[\|z'\|_0 \ge 2] \le O_p(\eta^2)$. When $\|z'\|_0 \ge 2$, the contribution to ♣ is $O_p(\eta)$. Multiplying them together, the total contribution to ♣ *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 $\clubsuit$ is

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

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 value of $\clubsuit$ 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 $\|a_r D_{v,x,\rho}(V + V^\rho)\|_\infty \le \widetilde{O}(\frac{\|a_r\|_2}{\sqrt{m_2}}) \le \widetilde{O}(1)$ (for a proof see Footnote 29). Therefore, given $\|z'\|_0 = 1$ and $\|z'\|_\infty \le \eta\tau_{w,\infty}$, we have that $\|a_r D_{v,x,\rho}(V + V^\rho)z'\| \le \widetilde{O}(\eta\tau_{w,\infty})$. 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 value of $\clubsuit$ is $\widetilde{O}\left(\eta^2 m_1 \frac{\tau_w^2}{\sigma_w}\right) + O_p(\eta^3)$.

In sum, we have with high probability

$$\mathbb{E}_{W^\rho, V^\rho}[|\clubsuit|] \le \widetilde{O}\left(\eta^2 m_1 \frac{\tau_w^2}{\sigma_w}\right) + O_p(\eta^3)$$

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

$$\begin{aligned}\delta = &\big((V + V^\rho + \eta V'')D_{w,x,\rho,\eta}\left((W + W^\rho + \eta W'')x + b_1\right) + b_2\big)\\ &- \left((V + V^\rho)D_{w,x,\rho}\left((W + W^\rho)x + b_1\right) + b_2\right).\end{aligned}$$

This difference δ has the following four terms:

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

   With $\|W''x\|_\infty \le \tau_{w,\infty}$ and $\|V\|_2 \le \widetilde{O}(1)$, by Fact C.9 we know that w.h.p.

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

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 $\|z\|_\infty \le \widetilde{O}(\tau_w + \sigma_w + m_1^{-1/2}) = O(m_1^{-1/2})$ owing to (C.25), by Fact C.9, we know that w.h.p.

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

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

In sum,

$$\|\delta\|_2 \le \begin{cases} O_p(\eta), & \text{w.p. } \le O_p(\eta);\\ \widetilde{O}(\frac{\eta\tau_{v,\infty}}{\sqrt{m_1}} + \eta\tau_{w,\infty}) + O_p(\eta^2), & \text{otherwise.}\end{cases}$$

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 (C.27)[30] we can bound the contribution to the *expected* value of $\spadesuit$ by:

$$\begin{aligned}\mathbb{E}_{W^\rho,V^\rho}[|\spadesuit|] &\leq m_2 \times \widetilde{O}\left(\eta\left(\frac{1}{\sqrt{m_1}}\tau_{v,\infty}+\tau_{w,\infty}\right)\times \eta\frac{(\frac{1}{\sqrt{m_1}}\tau_{v,\infty}+\tau_{w,\infty})}{\sigma_v}\right)+O_p(\eta^3)\\ &= \widetilde{O}\left(\eta^2\frac{m_2\tau_{v,\infty}^2}{\sigma_v m_1}+\eta^2\frac{m_2\tau_{w,\infty}^2}{\sigma_v}\right)+O_p(\eta^3)\ .\end{aligned}$$

□

C.2.4 Lemma 6.10: Stronger Coupling

We will need the following coupling lemma when $\mathbf{\Sigma}$ is used (for Algorithm 2).

Lemma 6.10 (stronger coupling)**.** *Given a fixed x, with high probability over the random initialization and over a random diagonal matrix $\mathbf{\Sigma}$ with diagonal entries i.i.d. generated from $\{-1,1\}$, it satisfies that for every W',V' (that can depend on the initialization and x but not $\mathbf{\Sigma}$) with $\|V'\|_2\leq\tau_v$, $\|W'\|_{2,4}\leq\tau_w$ for $\tau_v\in[0,1]$ and $\tau_w\in\big[\frac{1}{m_1^{3/4}},\frac{1}{m_1^{9/16}}\big]$, we have*

$$\begin{aligned}f_r(x;W^{(0)}+\mathbf{\Sigma}W',V^{(0)}+V'\mathbf{\Sigma}) = a_r D_{v,x}(V^{(0)}D_{w,x}(W^{(0)}x+b_1)+b_2)+a_r D_{v,x}V'D_{w,x}W'x\\ \pm\widetilde{O}\left(\tau_w^{8/5}m_1^{9/10}+\tau_w^{16/5}m_1^{9/5}\sqrt{m_2}+\frac{\sqrt{m_2}}{\sqrt{m_1}}\tau_v\right)\ .\end{aligned}$$

Under parameter choices Table 1, the last error term is at most ε/k.

Proof of Lemma 6.10. For notation simplicity, let us do the proof without the bias term b_1 and b_2. The proof with them are analogous.

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

$$\begin{aligned}z &= D_{w,x}^{(0)}W^{(0)}x\\ z_1 &= D_{w,x}^{(0)}\mathbf{\Sigma}W'x\\ z_2 &= D'_{w,x}(W^{(0)}x+\mathbf{\Sigma}W'x).\end{aligned}$$

Since w.h.p. each coordinate of z has $\|z\|_\infty=\widetilde{O}(m_1^{-1/2})$, using Fact C.9 (so using the randomness of $\mathbf{\Sigma}$), we know with high probability

$$\|V'\mathbf{\Sigma}z\|_2^2=\sum_{i\in[m_2]}\langle v'_i,\mathbf{\Sigma}z\rangle^2\leq\sum_{i\in[m_2]}\widetilde{O}\big(\|v'_i\|_2^2\cdot\|z\|_\infty^2\big)\leq\widetilde{O}(m_1^{-1})\sum_{i\in[m_1]}\|v'_i\|_2^2=\widetilde{O}(\tau_v^2 m_1^{-1})\qquad\text{(C.28)}$$

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

On the other hand, applying Lemma 6.4 with $W''=0$ and $V''=0$, we have $\|z_2\|_0\leq s=$

[30] 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 $\widetilde{O}\big(\eta\frac{(\frac{1}{\sqrt{m_1}}\tau_{v,\infty}+\tau_{w,\infty})}{\sigma_v}\big)$. 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 $m_2\times\widetilde{O}\big(\eta\frac{(\frac{1}{\sqrt{m_1}}\tau_{v,\infty}+\tau_{w,\infty})}{\sigma_v}\big)$ 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.

$\widetilde{O}(\tau_w^{4/5} m_1^{6/5})$. Therefore, using the same derivation as (C.24), we can bound its Euclidean norm

$$\begin{aligned}\|z_2\|_2 = \|D'_{w,x}(W^{(0)} + \mathbf{\Sigma} W')x\|_2 &\le \sqrt{\sum_{i,(D'_{w,x})_{i,i}\neq 0} (\mathbf{\Sigma}W'x)_i^2} \\ &\le \sqrt{\sqrt{s \cdot \sum_{i,(D'_{w,x})_{i,i}\neq 0} (W'x)_i^4}} \le O\left(s^{1/4}\tau_w\right) = \widetilde{O}\left(\tau_w^{6/5} m_1^{3/10}\right)\end{aligned}$$

Using the same derivation as (C.28), we also have w.h.p

$$\|V'\mathbf{\Sigma} z_2\|_2 \le \widetilde{O}\big(\|V'\|_F \|z_2\|_\infty\big) \le \widetilde{O}\big(\|V'\|_F \|z_2\|_2\big) \le \widetilde{O}\left(\tau_v \tau_w^{6/5} m_1^{3/10}\right) \quad .$$

Now, recall using Cauchy-Shwartz and the 1-Lipschitz continuity of the ReLU function, we have for every $(x_i, y_i)_{i\in[m_2]}$, with high probability (over a):

$$\sum_{i\in[m_2]} a_{i,r}(\sigma(x_i + y_i) - \sigma(x_i)) \le \widetilde{O}(\sqrt{m_2})\|(\sigma(x_i + y_i) - \sigma(x_i))_{i\in[m_2]}\|_2 \le \widetilde{O}(\sqrt{m_2})\|y\|_2.$$

Therefore, we can bound that

$$\begin{aligned} & f_r(x; W^{(0)} + \mathbf{\Sigma} W', V^{(0)} + V'\mathbf{\Sigma}) \\ = & \sum_{i\in[m_2]} a_{i,r}\sigma\left(\langle v_i^{(0)} + \mathbf{\Sigma} v'_i, z + z_1 + z_2\right) \\ = & \underbrace{\sum_{i\in[m_2]} a_{i,r}\sigma\left(\langle v_i^{(0)}, z + z_1 + z_2\rangle + \langle \mathbf{\Sigma} v'_i, z_1\rangle\right)}_{\text{①}} \pm \widetilde{O}\left(\frac{\sqrt{m_2}}{\sqrt{m_1}}\tau_v + \sqrt{m_2}\tau_v\tau_w^{6/5} m_1^{3/10}\right) \end{aligned} \tag{C.29}$$

To bound ①, we consider the difference between

$$\begin{aligned} \text{①} &= a_r(D_{v,x}^{(0)} + D''_{v,x})\left(V^{(0)}(z + z_1 + z_2) + V'D_{w,x}^{(0)}W'x\right) \\ \text{②} &= a_r D_{v,x}^{(0)}\left(V^{(0)}(z + z_1 + z_2) + V'D_{w,x}^{(0)}W'x\right) \end{aligned}$$

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

- $V^{(0)}z_1 = V^{(0)}D_{w,x}^{(0)}\mathbf{\Sigma} W'x$. 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.9 (thus using the randomness of $\mathbf{\Sigma}$), we know that w.h.p. $\|V^{(0)}z_1\|_\infty \le \widetilde{O}\left(\tau_w \frac{m_1^{1/4}}{\sqrt{m_2}}\right)$.
- $V^{(0)}z_2$. Using the sparsity of z_2 we know w.h.p. $\|V^{(0)}z_2\|_\infty \le \widetilde{O}(\|z_2\|_2\sqrt{s}m_2^{-1/2}) \le \widetilde{O}\left(\frac{\tau_w^{8/5} m_1^{9/10}}{\sqrt{m_2}}\right)$
- $\|V'D_{w,x}^{(0)}W'x\|_2 \le \|V'\|_F \cdot \|W'x\|_2 \le \tau_w m_1^{1/4}\tau_v$.

Together, using $\|a_r\|_\infty \le \widetilde{O}(1)$ and invoking Claim C.8, we can bound it by:

$$|\text{①} - \text{②}| \le \widetilde{O}\left(\left(\tau_w \frac{m_1^{1/4}}{\sqrt{m_2}} + \frac{\tau_w^{8/5} m_1^{9/10}}{\sqrt{m_2}}\right)^2 m_2^{3/2} + (\tau_w m_1^{1/4}\tau_v)^{4/3} m_2^{1/2}\right) \tag{C.30}$$

(When invoking Claim C.8, we need $\tau_w \le m_1^{-9/16}$ and $\tau_w m_1^{1/4}\tau_v \le 1$.)

Finally, from ② to our desired goal

$$③ = a_r D^{(0)}_{v,x} V^{(0)} D^{(0)}_{w,x} W^{(0)} x + a_r D^{(0)}_{v,x} V' D^{(0)}_{w,x} W' x = a_r D^{(0)}_{v,x} \left(V^{(0)} z + V' D^{(0)}_{w,x} W' x \right)$$

there are still two terms:

- Since w.h.p. $\|a_r D^{(0)}_{v,x} V^{(0)}\|_\infty = \widetilde{O}(1)$ and z_2 is s sparse, we know that w.h.p.
$$|a_r D^{(0)}_{v,x} V^{(0)} z_2| \leq \widetilde{O}(\|z_2\|_2 \sqrt{s}) \leq \widetilde{O}\left(\tau_w^{8/5} m_1^{9/10}\right)$$
- Since $\|W'x\|_2 \leq \tau_w m_1^{1/4}$ and w.h.p. $\|a_r D^{(0)}_{v,x} V^{(0)} D^{(0)}_{w,x}\|_\infty \leq \widetilde{O}(1)$ (see Footnote 29), by Fact C.9,
$$|a_r D^{(0)}_{v,x} V^{(0)} z_1| = |a_r D^{(0)}_{v,x} V^{(0)} D^{(0)}_{w,x} \mathbf{\Sigma} W' x| \leq \widetilde{O}(\|a_r D^{(0)}_{v,x} V^{(0)} D^{(0)}_{w,x}\|_\infty \cdot \|W'x\|_2) \leq \widetilde{O}\left(\tau_w m_1^{1/4}\right)$$

In other words

$$|② - ③| \leq \widetilde{O}\left(\tau_w^{8/5} m_1^{9/10} + \tau_w m_1^{1/4}\right) \tag{C.31}$$

Putting together (C.29), (C.30), (C.31), one can carefully verify that

$$\begin{aligned}
f_r(x; W^{(0)} + \mathbf{\Sigma} W', V^{(0)} + V'\mathbf{\Sigma}) &= ① \pm \widetilde{O}\left(\frac{\sqrt{m_2}}{\sqrt{m_1}} \tau_v + \sqrt{m_2} \tau_v \tau_w^{6/5} m_1^{3/10}\right) \\
&= a_r D^{(0)}_{v,x} V^{(0)} D^{(0)}_{w,x} W^{(0)} x + a_r D^{(0)}_{v,x} V' D^{(0)}_{w,x} W' x \\
&\qquad \pm \widetilde{O}\left(\tau_w^{8/5} m_1^{9/10} + \tau_w m_1^{1/4}\right) \\
&\qquad \pm \widetilde{O}\left(\left(\tau_w \frac{m_1^{1/4}}{\sqrt{m_2}} + \frac{\tau_w^{8/5} m_1^{9/10}}{\sqrt{m_2}}\right)^2 m_2^{3/2} + (\tau_w m_1^{1/4} \tau_v)^{4/3} m_2^{1/2}\right) \\
&\qquad \pm \widetilde{O}\left(\frac{\sqrt{m_2}}{\sqrt{m_1}} \tau_v + \sqrt{m_2} \tau_v \tau_w^{6/5} m_1^{3/10}\right) \\
&= a_r D^{(0)}_{v,x} V^{(0)} D^{(0)}_{w,x} W^{(0)} x + a_r D^{(0)}_{v,x} V' D^{(0)}_{w,x} W' x \\
&\qquad \pm \widetilde{O}\left(\tau_w^{8/5} m_1^{9/10} + \tau_w^{16/5} m_1^{9/5} \sqrt{m_2} + \frac{\sqrt{m_2}}{\sqrt{m_1}} \tau_v\right)
\end{aligned}$$

Above, we have used our parameter choices $\tau_v \in [0,1]$ and $\tau_w \in [\frac{1}{m_1^{3/4}}, \frac{1}{m_1^{9/16}}]$. □

C.3 Optimization

Recall in the first variant of SGD,

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

C.3.1 Lemma 6.6: Descent Direction

Lemma 6.6 (descent direction). *For every $\varepsilon_0 \in (0,1)$ and $\varepsilon = \frac{\varepsilon_0}{k p_1 p_2^2 \mathfrak{C}_\mathfrak{s}(\Phi, p_2 \mathfrak{C}_\mathfrak{s}(\phi,1)) \mathfrak{C}_\mathfrak{s}(\phi,1)^2}$, for every constant $\gamma \in (0, 1/4]$, consider the parameter choices in Table 1, and consider any λ_t, W_t, V_t (that may depend on the randomness of $W^{(0)}, b^{(0)}, V^{(0)}, b^{(1)}$ and $\mathcal{Z}$) with*

$$\lambda_t \in \left((\varepsilon / \log(m_1 m_2))^{\Theta(1)}, 1\right] \quad \textit{and} \quad L'(\lambda_t, W_t, V_t) \in [(1+\gamma)\mathsf{OPT} + \Omega(\varepsilon_0/\gamma), \widetilde{O}(1)]$$

With high probability over the random initialization, there exists W^*, V^* with $\|W^*\|_F, \|V^*\|_F \le 1$ such that for every $\eta \in \left[0, \frac{1}{\mathsf{poly}(m_1,m_2)}\right]$:

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

where $\mathbf{\Sigma} \in \mathbb{R}^{m_1 \times m_1}$ is a diagonal matrix with each diagonal entry i.i.d. uniformly drawn from $\{\pm 1\}$.

Proof of Lemma 6.6. For each output $r \in [k]$,

- Define the "pseudo function" for every W', V' as

  $$g_r(x; W', V') = a_r D_{v,x,\rho,t}[(V^{(0)} + V^\rho + V')D_{w,x,\rho,t}[(W^{(0)} + W^\rho + W')x + b_1] + b_2]$$

  where $D_{v,x,\rho,t}$ and $D_{w,x,\rho,t}$ are the diagonal sign matrices at weights $W^{(0)} + W^\rho + W_t, V^{(0)} + V^\rho + V_t$.
- Recall the real network as

  $$f_r(x; W', V') = a_r D_{v,x,\rho,V'}[(V^{(0)} + V^\rho + V')D_{w,x,\rho,W'}[(W^{(0)} + W^\rho + W')x + b_1] + b_2]$$

  where $D_{v,x,\rho,V'}$ and $D_{w,x,\rho,W'}$ are the diagonal sign matrices at weights $W^{(0)} + W^\rho + W', V^{(0)} + V^\rho + V'$.
- $G(x; W', V') = (g_1, \dots, g_k)$ and $F(x; W', V') = (f_1, \dots, f_k)$.

As a sanity check, we have $G(x; W_t, V_t) = F(x; W_t, V_t)$. (Here we slightly override the notation and use $f_r(x; W', V')$ to denote $f_r(x; W^{(0)} + W^\rho + W', V^{(0)} + V^\rho + V')$.)

By our regularizer parameters λ_w, λ_v in Table 1, as long as $L'(\lambda_t, W_t, V_t) \le \widetilde{O}(1)$, we know

$$\begin{aligned} R(\sqrt{\lambda_t}W_t, \sqrt{\lambda_t}V_t) \le \widetilde{O}(1) &\implies \|\sqrt{\lambda_t}W_t\|_{2,4} \le \widetilde{O}(\tau'_w) \text{ and } \|\sqrt{\lambda_t}V_t\|_{2,2} \le \widetilde{O}(\tau'_v) \\ &\implies \|W_t\|_{2,4} \le \tau_w \text{ and } \|V_t\|_{2,2} \le \tau_v \end{aligned} \tag{C.32}$$

Applying Corollary 6.5 (but scaling up the target F^* by $\frac{1}{\lambda_t}$), we know that there exists W^*, V^* with (here we have scaled up W^* and scaled down V^* both by $m_1^{0.005}$)

$$\|\sqrt{\lambda_t}W^*\|_{2,\infty} \le \frac{C_0}{m_1^{1-0.005}}, \quad \|\sqrt{\lambda_t}V^*\|_{2,\infty} \le \frac{m_1^{1/2-0.005}}{m_2} \quad \text{and} \tag{C.33}$$

$$\|\frac{1}{\lambda_t}F^*(x) - G^*(x)\|_2 \le \varepsilon \quad \text{where} \quad G^*(x) \stackrel{\text{def}}{=} \left(a_r^\top D_{v,x,\rho,t}V^* D_{w,x,\rho,t}W^* x\right)_{r\in[k]} \ . \tag{C.34}$$

By our parameter choices in Table 1, this implies

$$\lambda_w\|\sqrt{\lambda_t}W^*\|_{2,4}^4 \le \varepsilon_0 \quad \text{and} \quad \lambda_v\|\sqrt{\lambda_t}V^*\|_F^2 \le \varepsilon_0$$
$$\|W^*\|_F \ll 1 \quad \text{and} \quad \|V^*\|_F \ll 1 \ .$$

Let us study an update direction

$$\widehat{W} = W_t + \sqrt{\eta}\mathbf{\Sigma}W^*, \widehat{V} = V_t + \sqrt{\eta}V^*\mathbf{\Sigma}.$$

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

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

On the other hand,

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

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

$$\|w_{t,i} + \sqrt{\eta}\boldsymbol{\Sigma} w_i^{*}\|_2^2 = \|w_{t,i}\|_2^2 + \eta\|w_i^{*}\|_2^2 + 2\sqrt{\eta}\langle w_{t,i}, w_i^{*}\rangle(\boldsymbol{\Sigma})_{i,i}$$

and therefore

$$\begin{aligned}\mathbb{E}\left[\left\|w_{t,i} + \sqrt{\eta}\boldsymbol{\Sigma} w_i^{*}\right\|_2^4\right] &= \|w_{t,i}\|_2^4 + 4\eta\langle w_{t,i}, w_i^{*}\rangle^2 + \eta^2\|w_i^{*}\|_2^4 + 2\eta\|w_{t,i}\|_2^2\|w_i^{*}\|_2^2 \\ &\le \|w_{t,i}\|_2^4 + 6\eta\|w_{t,i}\|_2^2\|w_i^{*}\|_2^2 + O_p(\eta^2) \ .\end{aligned}$$

(Recall we use $O_p(\cdot)$ to hide polynomial factors in m_1 and m_2.) By Cauchy-Schwarz,

$$\sum_{i\in[m_1]}\|w_{t,i}\|_2^2\|w_i^{*}\|_2^2 \le \sqrt{\left(\sum_{i\in[m_1]}\|w_{t,i}\|_2^4\right)\left(\sum_{i\in[m_1]}\|w_i^{*}\|_2^4\right)} \le \|W_t\|_{2,4}^2\|W^{*}\|_{2,4}^2$$

and therefore

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

By $\lambda_v\|\sqrt{\lambda_t}V^{*}\|_F^2 \le \varepsilon_0$, $\lambda_w\|\sqrt{\lambda_t}W^{*}\|_{2,4}^4 \le \varepsilon_0$, and $\lambda_w\|\sqrt{\lambda_t}W_t\|_{2,4}^4 \le R(\sqrt{\lambda_t}W_t, \sqrt{\lambda_t}V_t)$, we know that

$$\begin{aligned}\mathbb{E}[R(\sqrt{\lambda_t}\widehat{W}, \sqrt{\lambda_t}\widehat{V})] &\le R(\sqrt{\lambda_t}W_t, \sqrt{\lambda_t}V_t) + 4\eta\varepsilon_0 + 6\eta\sqrt{\varepsilon_0}\cdot\sqrt{R(\sqrt{\lambda_t}W_t, \sqrt{\lambda_t}V_t)} \\ &\le R(\sqrt{\lambda_t}W_t, \sqrt{\lambda_t}V_t) + 10\eta\varepsilon_0 + \frac{1}{4}\eta R(\sqrt{\lambda_t}W_t, \sqrt{\lambda_t}V_t) \ . \qquad \text{(C.35)}\end{aligned}$$

Change in Objective. We now consider the change in the objective value. Recall from (C.33) the construction of good network W^{*}, V^{*} satisfies $\tau_{v,\infty} \le \frac{1}{m_1^{999/2000}}$ and $\tau_{w,\infty} \le \frac{1}{m_1^{999/1000}}$. For polynomially small η, by Lemma 6.8 (replacing its η with $\sqrt{\eta}$), we have:

$$\mathbb{E}_{W^\rho,V^\rho}\left|g_r(x;\widehat{W},\widehat{V}) - f_r(x;\widehat{W},\widehat{V})\right| \le O(\varepsilon\eta) + O_p(\eta^{1.5}). \qquad \text{(C.36)}$$

First we focus on $G(x;\widehat{W},\widehat{V})$. Since $\|W_t\|_{2,4} \le \tau_w$ and $\|V_t\|_{2,2} \le \tau_v$ from (C.32), we can apply Lemma 6.4 to get

$$\begin{aligned}G(x;\widehat{W},\widehat{V}) &= G(x;W_t,V_t) + \sqrt{\eta}G'(x) + \eta G^*(x) \\ &= F(x;W_t,V_t) + \sqrt{\eta}G'(x) + \eta G^*(x) \qquad \text{(C.37)}\end{aligned}$$

where $G'(x)$ is from Lemma 6.4 and satisfies $\mathbb{E}_{\boldsymbol{\Sigma}}[G'(x)] = 0$ and w.h.p. $\|G'(x)\|_2 \le \varepsilon$; and $G^*(x)$ is from (C.34).

Combining (C.36) and (C.37), we know that for every fixed x, y in the support of distribution $\mathcal{Z}$:

$$\begin{aligned}&\mathbb{E}_{W^\rho,V^\rho,\boldsymbol{\Sigma}}[L(\lambda_t F(x;\widehat{W},\widehat{V}),y)] \\ &\le \mathbb{E}_{W^\rho,V^\rho,\boldsymbol{\Sigma}}[L(\lambda_t G(x;\widehat{W},\widehat{V}),y)] + O(\eta\varepsilon) + O_p(\eta^{1.5}). \\ &\overset{①}{\le} \mathbb{E}_{W^\rho,V^\rho}\left[L(\lambda_t\mathbb{E}_{\boldsymbol{\Sigma}}[G(x;\widehat{W},\widehat{V}),y])\right] + \mathbb{E}_{W^\rho,V^\rho,\boldsymbol{\Sigma}}\left\|\sqrt{\eta}G'(x)\right\|^2 + O(\eta\varepsilon) + O_p(\eta^{1.5}). \\ &\overset{②}{\le} \mathbb{E}_{W^\rho,V^\rho}L\left(\lambda_t G(x;W_t,V_t) + \lambda_t\eta G^*(x),y\right) + O(\eta\varepsilon) + O_p(\eta^{1.5}). \\ &\overset{③}{\le} \mathbb{E}_{W^\rho,V^\rho}L\left(\lambda_t G(x;W_t,V_t) + \eta F^*(x),y\right) + O(\eta\varepsilon) + O_p(\eta^{1.5}). \\ &= \mathbb{E}_{W^\rho,V^\rho}L\left(\lambda_t F(x;W_t,V_t) + \eta F^*(x),y\right) + O(\eta\varepsilon) + O_p(\eta^{1.5}). \qquad \text{(C.38)}\end{aligned}$$

Above, ① uses the 1-Lipschitz smoothness of L which implies

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

and $\mathbb{E}_{\mathbf{\Sigma}}[G(\widehat{W}, \widehat{V}, x)] = G(W_t, V_t, x) + \eta G^*(x)$. Inequality ② also uses $\mathbb{E}_{\mathbf{\Sigma}}[G(x; \widehat{W}, \widehat{V})] = G(x; W_t, V_t) + \eta G^*(x)$. Inequality ③ uses (C.34) and the 1-Lipschitz continuity of L.

Next, by convexity of the loss function, we have

$$\begin{aligned} L\left(\lambda_t F(x; W_t, V_t) + \eta F^*(x), y\right) &= L\left((1-\eta)(1-\eta)^{-1}\lambda_t F(x; W_t, V_t) + \eta F^*(x), y\right) \\ &\leq (1-\eta)\left(L((1-\eta)^{-1}\lambda_t F(x; W_t, V_t), y)\right) + \eta L(F^*(x), y) \end{aligned} \quad \text{(C.39)}$$

For sufficiently small η, we know that

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

Putting this into (C.39), we have

$$\begin{aligned} & L\left(\lambda_t F(x; W_t, V_t) + \eta F^*(x), y\right) \\ & \leq (1-\eta)\left(2L\left(\lambda_t F(x; W_t, V_t), y\right) - L((1-\eta)\lambda_t F(x; W_t, V_t), y)\right) + \eta L(F^*(x), y) + O_p(\eta^2) \end{aligned} \quad \text{(C.40)}$$

Putting All Together. Let us denote

$$\begin{aligned} c_1 &= \mathbb{E}_{W^\rho, V^\rho, \mathbf{\Sigma}, (x,y)\sim\mathcal{Z}}[L(\lambda_t F(x; \widehat{W}, \widehat{V}), y)] \\ c_2 &= \mathbb{E}_{W^\rho, V^\rho, (x,y)\sim\mathcal{Z}}[L((1-\eta)\lambda_t F(x; W_t, V_t), y)] \\ c_3 &= \mathbb{E}_{W^\rho, V^\rho, (x,y)\sim\mathcal{Z}}[L(\lambda_t F(x; W_t, V_t), y)] \\ c_1' &= \mathbb{E}_{\mathbf{\Sigma}}\left[L'\left(\lambda_t, \widehat{W}, \widehat{V}\right)\right] = c_1 + \mathbb{E}_{\mathbf{\Sigma}}[R(\sqrt{\lambda_t}\widehat{W}, \sqrt{\lambda_t}\widehat{V})] \\ c_2' &= L'\left((1-\eta)\lambda_t, W_t, V_t\right) = c_2 + R(\sqrt{(1-\eta)\lambda_t}W_t, \sqrt{(1-\eta)\lambda_t}V_t) \\ c_3' &= L'\left(\lambda_t, W_t, V_t\right) = c_3 + R(\sqrt{\lambda_t}W_t, \sqrt{\lambda_t}V_t) \end{aligned}$$

The objective growth inequalities (C.38) and (C.40) together imply

$$c_1 \leq (1-\eta)\left(2c_3 - c_2\right) + \eta(\mathsf{OPT} + O(\varepsilon)) + O_p(\eta^{1.5}) \quad \text{(C.41)}$$

The regularizer growth inequality (C.35) implies

$$c_1' - c_1 \leq (1 + \frac{\eta\gamma}{4})(c_3' - c_3) + O(\eta\varepsilon_0/\gamma)$$

and therefore

$$\begin{aligned} c_1' - c_1 - \left((1-\eta)\left(2(c_3' - c_3) - (c_2' - c_2)\right)\right) &\leq c_1' - c_1 - \left((1-\eta)\left(2(c_3' - c_3) - (1-\eta)(c_3' - c_3)\right)\right) \\ &\leq c_1' - c_1 - \left(1 - \eta^2\right)(c_3' - c_3) \\ &\leq \frac{\eta\gamma}{4}c_3' + O(\eta\varepsilon_0/\gamma) + O(\eta^2) \ . \end{aligned} \quad \text{(C.42)}$$

Putting (C.41) and (C.42) together we have

$$c_1' \leq (1-\eta)\left(2c_3' - c_2'\right) + \frac{\eta\gamma}{4}c_3' + \eta(\mathsf{OPT} + O(\varepsilon_0/\gamma)) + O_p(\eta^{1.5})$$

Multiplying $\frac{1}{2(1-\eta)}$ on both sides, we have:

$$\frac{1}{2}(1-\eta)^{-1}c_1' + \frac{1}{2}c_2' \leq c_3' + \frac{\eta\gamma}{8}c_3' + \eta\frac{1}{2}\mathsf{OPT} + O(\eta\varepsilon_0/\gamma) + O_p(\eta^{1.5})$$

Therefore,

$$\left(\frac{1}{2}(1-\eta)^{-1} + \frac{1}{2}\right)\min\{c_1', c_2'\} \leq \left(1 + \frac{\eta\gamma}{8}\right)c_3' + \eta\frac{1}{2}\mathsf{OPT} + O(\eta\varepsilon_0/\gamma) + O_p(\eta^{1.5})$$

and this implies that

$$\min\{c_1', c_2'\} \le \left(1 - \eta\frac{1}{2} + \frac{\eta\gamma}{8}\right) c_3' + \eta\frac{1}{2}\mathsf{OPT} + O(\eta\varepsilon_0/\gamma) + O_p(\eta^{1.5})$$

Therefore, as long as $c_3' \ge (1+\gamma)\mathsf{OPT} + \Omega(\varepsilon_0/\gamma)$ and $\gamma \in [0,1]$, we have:

$$\min\{c_1', c_2'\} \le (1 - \eta\gamma/4)c_3'$$

This completes the proof. □

C.3.2 Lemma 6.7: Convergence

Lemma 6.7 (convergence)**.** *In the setting of Theorem 3, with probability at least* 99/100*, Algorithm 3 (the first SGD variant) converges in* $TT_w = \mathsf{poly}(m_1, m_2)$ *iterations to a point*

$$L'(\lambda_T, W_T, V_T) \le (1+\gamma)\mathsf{OPT} + \varepsilon_0.$$

Proof of Lemma 6.7. For the first variant of SGD, note that there are $T = \Theta(\eta^{-1}\log\frac{\log(m_1 m_2)}{\varepsilon_0})$ rounds of weight decay, which implies that $\lambda_t \ge (\varepsilon/\log(m_1 m_2))^{O(1)}$ is always satisfied (because γ is a constant). By Lemma 6.6, we know that as long as $L' \in [(1+\gamma)\mathsf{OPT} + \Omega(\varepsilon_0/\gamma), \widetilde{O}(1)]$, then there exists $\|W^*\|_F, \|V^*\|_F \le 1$ such that either

$$\mathbb{E}_{\mathbf{\Sigma}}\left[L'\left(\lambda_{t-1}, W_t + \sqrt{\eta}\mathbf{\Sigma}W^*, V_t + \sqrt{\eta}V^*\mathbf{\Sigma}\right)\right] \le (1 - \eta\gamma/4)(L'(\lambda_{t-1}, W_t, V_t))$$

or

$$L'((1-\eta)\lambda_{t-1}, W_t, V_t) \le (1 - \eta\gamma/4)(L'(\lambda_{t-1}, W_t, V_t))$$

In the first case, recall L' is $B = \mathsf{poly}(m_1, m_2)$ second-order smooth,[31] by Fact A.8, it satisfies (λ_{t-1} is fixed and the Hessian is with respect to W and V):

$$\lambda_{\min}\left(\nabla^2 L'(\lambda_{t-1}, W_t, V_t)\right) < -1/(m_1 m_2)^8 \ .$$

On the other hand, for every $t \ge 1$, since W_t is the output of noisy SGD, by the escape saddle point theorem of [20] (stated in Lemma A.9), we know with probability at least $1-p$ it satisfies $\lambda_{\min}\left(\nabla^2 L'(\lambda_{t-1}, W_t, V_t)\right) > -1/(m_1 m_2)^8$. Choosing $p = \frac{1}{1000T}$, we know with probability at least 0.999, this holds for all rounds $t = 1, 2, \ldots, T$. In other words, for all rounds $t = 1, 2, \ldots, T$, the first case cannot happen and therefore as long as $L' \ge (1+\gamma)\mathsf{OPT} + \Omega(\varepsilon_0/\gamma)$,

$$L'((1-\eta)\lambda_{t-1}, W_t, V_t) \le (1 - \eta\gamma/4)(L'(\lambda_{t-1}, W_t, V_t)).$$

On the other hand, for each round $t = 0, 1, \ldots, T-1$, as long as $L' \le \widetilde{O}(1)$, by Lemma A.9, it holds that

$$L'(\lambda_t, W_{t+1}, V_{t+1}) \le L'(\lambda_t, W_t, V_t) + (m_1 m_2)^{-1} \ .$$

Since initially, $L'(\lambda_1, W_0, V_0) \le \widetilde{O}(1)$ w.h.p., we have w.h.p. $L' \le \widetilde{O}(1)$ throughout the process. Since γ is a constant, after $T = \Theta(\eta^{-1}\log\frac{\log m}{\varepsilon_0})$ rounds of weight decay, we have $L' \le (1+\gamma)\mathsf{OPT} + O(\varepsilon_0/\gamma)$. Since γ is a constant, re-scaling ε_0 down by a constant factor finishes the proof. □

Remark C.10. For the second variant of the SGD, note that

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

[31] When convoluted with Gaussian noise $\nabla(f * g) = f * \nabla g$, every bounded function f becomes infinite-order differentiable with parameter B inversely-polynomially dependent on the noise level 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 results from Lemma 6.6 and Lemma 6.7 still apply, if we replace L' with

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

C.4 Generalization

C.4.1 Lemma 6.9: Generalization For $L_R = L_1$

We derive a very crude Rademacher complexity bound for our three-layer neural network. We have not tried to tighten the polynomial dependency in m_1 and m_2.

Lemma 6.9 (generalization for $L_R = L_1$)**.** *For every $\tau'_v, \tau'_w \ge 0$, every $\sigma_v \in (0, 1/\sqrt{m_2}]$, w.h.p. for every $r \in [k]$ and every $N \ge 1$, the empirical Rademacher complexity is bounded by*

$$\begin{aligned} &\frac{1}{N}\mathbb{E}_{\xi\in\{\pm1\}^N}\left[\sup_{\|V'\|_F \le \tau'_v, \|W'\|_{2,4}\le\tau'_w} \sum_{i\in[N]} \xi_i f_r(x_i; W^{(0)} + W^\rho + W', V^{(0)} + V^\rho + V')\right] \\ &\le \widetilde{O}\left(\frac{\tau'_w m_1\sqrt{m_2} + \tau'_v m_2}{\sqrt{N}} + \frac{\tau'_v\sqrt{m_1 m_2 \tau'_w(1/\sqrt{m_1} + \tau'_w)}}{N^{1/4}}\right) . \end{aligned}$$

Proof. Let $W = W^{(0)} + W^\rho$ and $V = V^{(0)} + V^\rho$ for notation simplicity. Recall the input to the j-th neuron on the second layer is

$$n_j(x; W + W', V + V') = \sum_{i\in[m_1]} (v_{j,i} + v'_{j,i})\sigma\left(\langle w_i + w'_i, x\rangle + b^{(0)}_{1,i}\right) + b^{(0)}_{2,j}$$

For analysis purpose, let us truncate V' by zeroing out all of its large coordinates. Namely, $V'' \in \mathbb{R}^{m_2\times m_1}$ is defined so that $V''_{i,j} = V'_{i,j}$ if $|V'_{i,j}| \le \delta$ and $V''_{i,j} = 0$ otherwise. At most $(\tau'_v)^2/\delta^2$ coordinates will be zeroed out because $\|V'\|_F \le \tau'_v$. Since for each x in the training set, we have with high probability $|\sigma(\langle w_i + w'_i, x\rangle + b^{(0)}_{1,i})| \le \widetilde{O}(\frac{1}{\sqrt{m_1}} + \|w'_i\|_2) \le \widetilde{O}(\frac{1}{\sqrt{m_1}} + \tau'_w)$, and since $\|a_r\|_\infty \le \widetilde{O}(1)$, it satisfies

$$|f_r(x; W + W', V + V') - f_r(x; W + W', V + V'')| \le \widetilde{O}(\frac{1}{\sqrt{m_1}} + \tau'_w) \times \frac{(\tau'_v)^2}{\delta} \ . \tag{C.43}$$

We now bound the Rademacher complexity of $f_r(x_i; W + W', V + V'')$ in the following simple steps.

- $\{x \mapsto \langle w'_i, x\rangle \mid \|w'_i\|_2 \le \tau'_w\}$ has Rademacher complexity $O(\frac{\tau'_w}{\sqrt{N}})$ by Proposition A.12a.
- $\{x \mapsto \langle w_i + w'_i, x\rangle + b_i \mid \|w'_i\|_2 \le \tau'_w\}$ has Rademacher complexity $O(\frac{\tau'_w}{\sqrt{N}})$ because singleton class has zero complexity and adding it does not affect complexity by Proposition A.12c.
- $\{x \mapsto n_j(x; W + W', V + V'') \mid \|W'\|_{2,\infty} \le \tau'_w \wedge \|v''_j\|_1 \le \sqrt{m_1}\tau'_v \wedge \|v''_j\|_\infty \le \delta\}$ has Rademacher complexity $O(\frac{\tau'_w}{\sqrt{N}}) \cdot \widetilde{O}(\frac{m_1}{\sqrt{m_2}} + \delta m_1) + \widetilde{O}(\frac{\tau'_v}{\sqrt{N}})$. This is because w.h.p. $\|v_j\|_1 \le \widetilde{O}(\frac{m_1}{\sqrt{m_2}})$ so we can apply Proposition A.12e, and because $\|v''_j\|_1 \le \sqrt{m_1}\tau'_v$ and $\|v''_j\|_\infty \le \delta$ so we can apply Proposition A.12f by choosing $f^{(0)}_i(x) = \langle w_i, x\rangle + b_i$ which satisfies $|f^{(0)}_i(x)| \le \widetilde{O}(\frac{1}{\sqrt{m_1}})$ w.h.p.
- $\{x \mapsto f_r(x; W + W', V + V'') \mid \|W'\|_{2,\infty} \le \tau'_w \wedge \forall j \in [m_2], \|v''_j\|_1 \le \sqrt{m_1}\tau'_v\}$ has Rademacher complexity $\left(O(\frac{\tau'_w}{\sqrt{N}}) \cdot \widetilde{O}(\frac{m_1}{\sqrt{m_2}} + \delta m_1) + \widetilde{O}(\frac{\tau'_v}{\sqrt{N}})\right) \cdot \widetilde{O}(m_2)$ because w.h.p. $\|a_r\|_1 \le \widetilde{O}(m_2)$ and Proposition A.12e.

Finally, noticing that $\|W'\|_{2,4} \leq \tau'_w$ implies $\|W'\|_{2,\infty} \leq \tau'_w$ and $\|V''\|_F \leq \tau'_v$ implies $\|v''_j\|_1 \leq \sqrt{m_1}\|v''_j\|_2 \leq \sqrt{m_1}\tau'_v$, we finish the proof that the Rademacher complexity of $f_r(x_i; W+W', V+V'')$ is at most

$$\widetilde{O}\left(\frac{\tau'_w m_1\sqrt{m_2} + \tau'_v m_2}{\sqrt{N}} + \frac{\tau'_w m_1 m_2 \delta}{\sqrt{N}}\right) .$$

Combining this with (C.43), and tuning the best choice of δ gives the desired result. □

C.4.2 Lemma 6.11: Generalization For $L_R = L_2$

Lemma 6.11 (generalization for $L_R = L_2$)**.** *For every $\tau'_v \in [0,1]$, $\tau'_w \in \big[\frac{1}{m_1^{3/4}}, \frac{1}{m_1^{9/16}}\big]$, every $\sigma_w \in [0, 1/\sqrt{m_1}]$ and $\sigma_v \in [0, 1/\sqrt{m_2}]$, w.h.p. for every $r \in [k]$ and every $N \geq 1$, we have by our choice of parameters in Lemma 6.6, the empirical Rademacher complexity is bounded by*

$$\frac{1}{N}\mathbb{E}_{\xi\in\{\pm1\}^N}\left[\sup_{\|V'\|_F\leq\tau'_v, \|W'\|_{2,4}\leq\tau'_w}\left|\sum_{i\in[N]}\xi_i\mathbb{E}_{\mathbf{\Sigma}}[f_r(x_i; W^{(0)} + W^\rho + \mathbf{\Sigma}W', V^{(0)} + V^\rho + V'\mathbf{\Sigma})]\right|\right]$$

$$\leq \widetilde{O}\left(\frac{\tau'_w\tau'_v m_1^{1/4}\sqrt{m_2}}{\sqrt{N}} + \left((\tau'_w)^{8/5}m_1^{9/10} + (\tau'_w)^{16/5}m_1^{9/5}\sqrt{m_2} + \frac{\sqrt{m_2}}{\sqrt{m_1}}\tau'_v\right)\right) .$$

Under parameter choices in Table 1, this is at most $\widetilde{O}\big(\frac{\tau'_w\tau'_v m_1^{1/4}\sqrt{m_2}}{\sqrt{N}}\big) + \varepsilon/k$.

Proof of Lemma 6.11. Let $W = W^{(0)} + W^\rho$ and $V = V^{(0)} + V^\rho$ for notation simplicity. Applying Lemma 6.10 (with τ_w chosen as τ'_w), we know that by our choice of parameters,

$$f_r(x; W + \mathbf{\Sigma}W', V + V'\mathbf{\Sigma}) = a_r D_{v,x}\left(V D_{w,x}(Wx + b_1) + b_2\right) + a_r D_{v,x} V' D_{w,x} W' x \pm B$$

for $B = \widetilde{O}\big((\tau'_w)^{8/5}m_1^{9/10} + (\tau'_w)^{16/5}m_1^{9/5}\sqrt{m_2} + \frac{\sqrt{m_2}}{\sqrt{m_1}}\tau'_v\big)$.

We bound the Rademacher complexity of the right hand side. It consists of three terms and the Rademacher complexity is the summation of the three (recall Proposition A.12c).

The first term does not depend on W' or V' so has Rademacher complexity zero.

The third term has Rademacher complexity at most B.

The second term corresponds to the function class

$$\mathcal{F} = \left\{x \mapsto a_r D_{v,x} V' D_{w,x} W' x \mid \|V'\|_F \leq \tau'_v, \|W'\|_{2,4} \leq \tau'_w\right\}$$

We calculate its Rademacher complexity as follows.

$$\begin{aligned}
&\sup_{\|W'\|_{2,4}\leq\tau'_w, \|V'\|_F\leq\tau'_v}\left|\sum_{j\in[N]}\xi_j a_r D_{v,x_j} V' D_{w,x_j} W' x_j\right| \\
=&\sup_{\|W'\|_{2,4}\leq\tau'_w, \|V'\|_F\leq\tau'_v}\left|\mathbf{Tr}\Big(\sum_{j\in[N]}\xi_j x_j a_r D_{v,x_j} V' D_{w,x_j} W'\Big)\right| \\
\leq&\sup_{\|W'\|_{2,4}\leq\tau'_w, \|V'\|_F\leq\tau'_v}\left\|\sum_{j\in[N]}\xi_j x_j a_r D_{v,x_j} V' D_{w,x_j}\right\|_F \|W'\|_F \\
\leq&\, \tau'_w m_1^{1/4}\sup_{\|V'\|_F\leq\tau'_v}\left\|\sum_{j\in[N]}\xi_j x_j a_r D_{v,x_j} V' D_{w,x_j}\right\|_F
\end{aligned}$$

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'.

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

For random ξ_j, we know that w.h.p. over the randomness of ξ_j (notice that we can do so because D_{w,x_j} only depends on $W^{(0)}$ but not on W', so we can take randomness argument on ξ),

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

Thus,

$$\begin{aligned}
&\mathbb{E}_\xi \sup_{\|V'\|_F\le\tau'_v} \sqrt{\sum_{p\in[d],q\in[m_1]} \left\|\sum_{j\in[N]} \xi_j [D_{w,x_j}]_{q,q} x_{j,p} a_r D_{v,x_j}\right\|^2 \|[V']_q\|^2} \\
&\le \widetilde{O}\left(\sup_{\|V'\|_F\le\tau'_v} \|a_r\|_2 \sqrt{\sum_{p\in[d],q\in[m_1]} \sum_{j\in[N]} x_{j,p}^2 \|[V']_q\|^2}\right) \\
&\le \widetilde{O}\left(\|a_r\|_2 \sup_{\|V'\|_F\le\tau'_v} \sqrt{\sum_{q\in[m_1]} \sum_{j\in[N]} \|[V']_q\|^2}\right) \\
&\le \widetilde{O}\left(\tau'_v \sqrt{m_2 N}\right).
\end{aligned}$$

This implies $\widehat{\mathfrak{R}}(\mathcal{X};\mathcal{F}) \le \widetilde{O}\big(\frac{\tau'_w \tau'_v m_1^{1/4} \sqrt{m_2}}{\sqrt{N}}\big)$ and finishes the proof. □

C.5 Final Theorems

C.5.1 Theorem 3: First SGD Variant

Theorem 3. *Consider Algorithm 3. For every* constant *$\gamma \in (0, 1/4]$, every $\varepsilon_0 \in (0, 1/100]$, every $\varepsilon = \frac{\varepsilon_0}{k p_1 p_2^2 \mathfrak{C}_\mathfrak{s}(\Phi, p_2 \mathfrak{C}_\mathfrak{s}(\phi,1)) \mathfrak{C}_\mathfrak{s}(\phi,1)^2}$, there exists*

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

such that for every $m_2 = m_1 = m \ge M$, and properly set $\lambda_w, \lambda_v, \sigma_w, \sigma_v$ in Table 1, as long as

$$N \ge \widetilde{\Omega}(M m^{3/2})$$

there is a choice $\eta = 1/\mathsf{poly}(m_1, m_2)$ and $T = \mathsf{poly}(m_1, m_2)$ such that with probability $\geq 99/100$,

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

Proof of Theorem 3. For notation simplicity let $L_F(z; \lambda, W, V) \overset{\text{def}}{=} L(\lambda F(x; W, V), y)$ for $z = (x, y)$. For the first SGD variant, recall from Lemma 6.7 that

$$\mathbb{E}_{z\sim\mathcal{Z}, W^\rho, V^\rho} L_F(z; \lambda_T, W^{(0)} + W_T + W^\rho, V^{(0)} + V_T + V^\rho) + R(\sqrt{\lambda_T} W_T, \sqrt{\lambda_T} V_T) \leq (1+\gamma)\mathsf{OPT} + \varepsilon_0.$$

Since $L_F(z; \lambda_T, W^{(0)} + W_T + W^\rho, V^{(0)} + V_T + V^\rho) \in [0, \widetilde{O}(1)]$ w.h.p., by randomly sampling $\widetilde{O}(1/\varepsilon_0^2)$ many $W^{\rho,j}, V^{\rho,j}$, we know w.h.p. there exists one j^* with

$$\mathbb{E}_{z\in\mathcal{Z}} L_F(z; \lambda_T, W^{(0)} + W^{\rho,j^*} + W_T, V^{(0)} + V^{\rho,j^*} + V_T) \leq (1+\gamma)\mathsf{OPT} + 2\varepsilon_0$$

Now, recall that

$$\|\sqrt{\lambda_T} W_T\|_{2,4} \leq \tau'_w \text{ and } \|\sqrt{\lambda_T} V_T\|_F \leq \tau'_v$$

due to our regularizer (see (C.32)). By simple spectral norm bound, we also know w.h.p. for every $(x, y) \sim \mathcal{D}$ and j,[32]

$$\left| L_F(z; \lambda_T, W^{(0)} + W^{\rho,j} + W_T, V^{(0)} + V^{\rho,j} + V_T) \right| \leq \widetilde{O}(\sqrt{km_2}) \ .$$

Therefore, we can plug in the Rademacher complexity from Lemma 6.9 and $b = \widetilde{O}(\sqrt{km_2})$ into standard generalization statement Corollary A.11. Using our choices of τ'_w and τ'_v from Table 1 as well as $m_1 = m_2$, this bound implies as long as $N \geq \widetilde{O}(M(m_2)^{3/2})$, w.h.p. for every pair $W^{\rho,j}, V^{\rho,j}$, it holds

$$\begin{aligned}&\mathbb{E}_{z\in\mathcal{D}} L_F(z; \lambda_T, W^{(0)} + W^{\rho,j} + W_T, V^{(0)} + V^{\rho,j} + V_T)\\ &\leq \mathbb{E}_{z\in\mathcal{Z}} L_F(z; \lambda_T, W^{(0)} + W^{\rho,j} + W_T, V^{(0)} + V^{\rho,j} + V_T) + \varepsilon_0\end{aligned}$$

Together, we have

$$\mathbb{E}_{z\in\mathcal{D}} L_F(z; \lambda_T, W^{(0)} + W^{\rho,j^*} + W_T, V^{(0)} + V^{\rho,j^*} + V_T) \leq (1+\gamma)\mathsf{OPT} + 3\varepsilon_0$$

□

C.5.2 Theorem 2: Second SGD Variant

Theorem 2. *Consider Algorithm 2. For every* $\mathsf{constant}$ *$\gamma \in (0, 1/4]$, every $\varepsilon_0 \in (0, 1/100]$, every $\varepsilon = \frac{\varepsilon_0}{kp_1p_2^2\mathfrak{C}_\mathfrak{s}(\Phi, p_2\mathfrak{C}_\mathfrak{s}(\phi,1))\mathfrak{C}_\mathfrak{s}(\phi,1)^2}$, there exists*

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

such that for every $m_2 = m_1 = m \geq M$, and properly set $\lambda_w, \lambda_v, \sigma_w, \sigma_v$ in Table 1, as long as

$$N \geq \widetilde{\Omega}\left(\Big(\frac{\mathfrak{C}_\varepsilon(\Phi, \sqrt{p_2}\mathfrak{C}_\varepsilon(\phi, 1)) \cdot \mathfrak{C}_\varepsilon(\phi, 1) \cdot \sqrt{p_2} p_1 k^2}{\varepsilon_0}\Big)^2\right)$$

there is a choice $\eta = 1/\mathsf{poly}(m_1, m_2)$ and $T = \mathsf{poly}(m_1, m_2)$ such that with probability $\geq 99/100$,

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

[32] Indeed, with our parameter choices in Table 1, the spectral norms $\sqrt{\lambda_T}\|W^{(0)} + W^{\rho,j} + W_T\|_2 \leq O(1) + \|\sqrt{\lambda_T} W_T\|_F \leq O(1 + m_1^{1/4}\tau'_w) \leq O(1)$ and $\sqrt{\lambda_T}\|V^{(0)} + V^{\rho,j} + V_T\|_2 \leq O(1) + \|\sqrt{\lambda_T} V_T\|_F \leq O(1 + \tau'_v) \leq O(1)$. Therefore, the network output $\lambda_T F(x; W^{(0)} + W^{\rho,j} + W_T, V^{(0)} + V^{\rho,j} + V_T)$ must be bounded by $\widetilde{O}(\sqrt{km_2})$ in Euclidean norm. By the assumption that $L(0, y) \in [0, 1]$ and $L(\cdot, y)$ is 1-Lipschitz continuous in the first variable, we have that L_F is bounded as stated.

Proof of Theorem 2. For notation simplicity let $L_F(z;\lambda,W,V) \stackrel{\text{def}}{=} L(\lambda F(x;W,V),y)$ for $z=(x,y)$. Recall from Remark C.10 that Lemma 6.7 still works in this setting, so we have

$$\begin{aligned}\mathbb{E}_{W^\rho,V^\rho,\boldsymbol{\Sigma},z\sim\mathcal{Z}}\left[L_F\left(z;\lambda_T,W^{(0)}+W^\rho+\boldsymbol{\Sigma}W_T,V^{(0)}+V^\rho+V_T\boldsymbol{\Sigma}\right)\right]&\\ +R(\sqrt{\lambda_T}W_T,\sqrt{\lambda_T}V_T)\le(1+\gamma)\mathsf{OPT}+\varepsilon_0&\ .\end{aligned}$$

For the same reason as the proof of Theorem 3, we know w.h.p. among $\widetilde{O}(1/\varepsilon_0^2)$ choices of j,

$$\min_j\left\{\mathbb{E}_{\boldsymbol{\Sigma},z\in\mathcal{Z}}L_F(z;\lambda_T,W^{(0)}+W^{\rho,j}+\boldsymbol{\Sigma}W_T,V^{(0)}+V^{\rho,j}+V_T\boldsymbol{\Sigma})\right\}\le(1+\gamma)\mathsf{OPT}+2\varepsilon_0 \tag{C.44}$$

Without loss of generality, in the remainder of the proof we assume $\mathsf{OPT}\le O(\varepsilon_0)$. This can be done because is ε_0 is too small we can increase it to $\varepsilon_0=\Theta(\mathsf{OPT})$. By our regularizer parameters λ_w,λ_v in Table 1, we know

$$\begin{aligned}&R(\sqrt{\lambda_T}W_T,\sqrt{\lambda_T}V_T)\le(1+\gamma)\mathsf{OPT}+\varepsilon_0\le O(\varepsilon_0)\\ \Longrightarrow\ &\|\sqrt{\lambda_T}W_T\|_{2,4}\le O(\tau_w'\varepsilon_0^{1/4})\text{ and }\|\sqrt{\lambda_T}V_T\|_{2,2}\le O(\tau_v'\varepsilon_0^{1/2})\end{aligned}\tag{C.45}$$

By Lemma 6.10 (but viewing $V^{(0)}+V^\rho$ as $V^{(0)}$ and viewing $W^{(0)}+W^\rho$ as $W^{(0)}$), we know for every (x,y), w.h.p. over $W^{(0)},V^{(0)},W^\rho,V^\rho,\boldsymbol{\Sigma}$

$$\begin{aligned}&f_r(x;W^{(0)}+W^\rho+\boldsymbol{\Sigma}W_T,V^{(0)}+V^\rho+V_T\boldsymbol{\Sigma})\\ &=a_rD_{v,x,\rho}\left((V^{(0)}+V^\rho)D_{w,x,\rho}\big((W^{(0)}+V^\rho)x+b_1\big)+b_2\right)+a_rD_{v,x,\rho}V_TD_{w,x,\rho}W_Tx\pm\varepsilon\\ &=f_r(x;W^{(0)}+W^\rho,V^{(0)}+V^\rho)+g_r^{(b,b)}(x;W_T,V_T)\pm\varepsilon/k\end{aligned}$$

where $D_{v,x,\rho}$ and $D_{w,x,\rho}$ are the diagonal sign indicator matrices at weights $W^{(0)}+W^\rho$, $V^{(0)}+V^\rho$, and we denote by $g_r^{(b,b)}(x;W',V')\stackrel{\text{def}}{=}a_rD_{v,x,\rho}V'D_{w,x,\rho}W'x$ the output of the pseudo network. This immediately implies for every (x,y) and every j, w.h.p. over $W^{(0)},V^{(0)},W^{\rho,j},V^{\rho,j},\boldsymbol{\Sigma}$

$$\begin{aligned}&f_r(x;W^{(0)}+W^{\rho,j}+\boldsymbol{\Sigma}W_T,V^{(0)}+V^{\rho,j}+V_T\boldsymbol{\Sigma})\\ &=f_r(x;W^{(0)}+W^{\rho,j},V^{(0)}+V^{\rho,j})+g_r^{(b,b)}(x;W_T,V_T)\pm\varepsilon/k\end{aligned}\tag{C.46}$$

Using the 1-Lipschitz continuity of L together with (C.45) and (C.46), it is not hard to derive that for every j, with high probability over $(x,y)\sim\mathcal{D}$,$W^{(0)},V^{(0)},W^{\rho,j},V^{\rho,j}$[33]

$$\left|\mathbb{E}_{\boldsymbol{\Sigma}}L_F(z;\lambda_T,W^{(0)}+W^{\rho,j}+\boldsymbol{\Sigma}W_T,V^{(0)}+V^{\rho,j}+V_T\boldsymbol{\Sigma})\right|\le\widetilde{O}(C_0)\ .$$

Therefore, we can plug in the Rademacher complexity from Lemma 6.11 with $b=\widetilde{O}(C_0)$ into standard generalization statement Corollary A.11.[34] Using our choices of τ_w' and τ_v' from Table 1 as well as $m_1=m_2$, the Rademacher complexity is dominated by

$$\lambda_T\frac{m_2^{1/2}m_1^{1/4}\|W_T\|_{2,4}\|V_T\|_{2,2}}{\sqrt{N}}\le\frac{\varepsilon_0^{3/4}m_2^{1/2}m_1^{1/4}\tau_v'\tau_w'}{\sqrt{N}}\le\frac{C_0}{\sqrt{N}}$$

In other words, as long as $N\ge(kC_0/\varepsilon_0)^2$, the Rademacher complexity of a single output is at most $\frac{\varepsilon_0}{k}$, so the generalization error is at most ε_0 by Corollary A.11. Or, in symbols, for every j, w.h.p.

[33] Indeed, $|f_r(x;W^{(0)}+W^{\rho,j},V^{(0)}+V^{\rho,j})|\le\widetilde{O}(1)$ with high probability, and $\lambda_T|g_r^{(b,b)}(x;W_T,V_T)|\le\widetilde{O}(\sqrt{m_2}\|\sqrt{\lambda_T}V_T\|_2\|\sqrt{\lambda_T}W_T\|_2)\le\widetilde{O}(\sqrt{m_2}\|\sqrt{\lambda_T}V_T\|_F\|\sqrt{\lambda_T}W_T\|_F)\le\widetilde{O}(\varepsilon_0^{3/4}\sqrt{m_2}m_1^{1/4}\tau_w'\tau_v')\le\widetilde{O}(C_0)$ by spectral norm bounds.

[34] Strictly speaking, Corollary A.11 requires an absolute value bound b as opposed to a high probability bound. It is a simple exercise to deal with this issue, see for instance Remark B.6 in our two-layer proof.

over $W^{(0)}, V^{(0)}, W^{\rho,j}, V^{\rho,j}$,

$$\begin{aligned} &\mathbb{E}_{\mathbf{\Sigma}, z\in\mathcal{D}} L_F(z;\lambda_T, W^{(0)}+W^{\rho,j}+\mathbf{\Sigma}W_T, V^{(0)}+V^{\rho,j}+V_T\mathbf{\Sigma}) \\ &\leq \mathbb{E}_{\mathbf{\Sigma}, z\in\mathcal{Z}} L_F(z;\lambda_T, W^{(0)}+W^{\rho,j}+\mathbf{\Sigma}W_T, V^{(0)}+V^{\rho,j}+V_T\mathbf{\Sigma}) + \varepsilon_0 \end{aligned} \tag{C.47}$$

Putting this into (C.44), we have

$$\min_j \Big\{ \mathbb{E}_{\mathbf{\Sigma}, z\in\mathcal{D}} L_F(z;\lambda_T, W^{(0)}+W^{\rho,j}+\mathbf{\Sigma}W_T, V^{(0)}+V^{\rho,j}+V_T\mathbf{\Sigma}) \Big\} \leq (1+\gamma)\mathsf{OPT} + 3\varepsilon_0 \tag{C.48}$$

Next, let us take expectation over $z \sim \mathcal{D}$ for (C.46) (strictly speaking, this needs one to carefully deal with the tail bound and apply the 1-Lipschitz continuity of L). We have for every j, w.h.p over $W^{(0)}, V^{(0)}, W^{\rho,j}, V^{\rho,j}, \mathbf{\Sigma}$

$$\begin{aligned} &\mathbb{E}_{(x,y)\sim\mathcal{D}} \Big[L\Big(\lambda_T F\Big(x; W^{(0)}+W^{\rho,j}+\mathbf{\Sigma}W_T, V^{(0)}+V^{\rho,j}+V_T\mathbf{\Sigma}\Big), y\Big)\Big] \\ &= \mathbb{E}_{(x,y)\sim\mathcal{D}} \Big[L\Big(\lambda_T F\Big(x; W^{(0)}+W^{\rho,j}, V^{(0)}+V^{\rho,j}\Big) + \lambda_T G^{(b,b)}(x; W_T, V_T), y\Big)\Big] \pm 2\varepsilon \end{aligned}$$

Since the right hand side (except the $\pm 2\varepsilon$ term) does not depend on the randomness of $\mathbf{\Sigma}$, we know that the left hand side with respect to a random sample $\widehat{\mathbf{\Sigma}}$ must stay close to its expectation with respect to $\mathbf{\Sigma}$. Or, in symbols, for every j, w.h.p over $W^{(0)}, V^{(0)}, W^{\rho,j}, V^{\rho,j}$

$$\begin{aligned} &\mathbb{E}_{(x,y)\sim\mathcal{D}} \Big[L\Big(\lambda_T F\Big(x; W^{(0)}+W^{\rho,j}+\widehat{\mathbf{\Sigma}}W_T, V^{(0)}+V^{\rho,j}+V_T\widehat{\mathbf{\Sigma}}\Big), y\Big)\Big] \\ &= \mathbb{E}_{\mathbf{\Sigma},(x,y)\sim\mathcal{D}} \Big[L\Big(\lambda_T F\Big(x; W^{(0)}+W^{\rho,j}+\mathbf{\Sigma}W_T, V^{(0)}+V^{\rho,j}+V_T\mathbf{\Sigma}\Big), y\Big)\Big] \pm 4\varepsilon \ . \end{aligned} \tag{C.49}$$

For similar reason, replacing $\mathcal{D}$ with $\mathcal{Z}$, we have

$$\begin{aligned} &\mathbb{E}_{(x,y)\sim\mathcal{Z}} \Big[L\Big(\lambda_T F\Big(x; W^{(0)}+W^{\rho,j}+\widehat{\mathbf{\Sigma}}W_T, V^{(0)}+V^{\rho,j}+V_T\widehat{\mathbf{\Sigma}}\Big), y\Big)\Big] \\ &= \mathbb{E}_{\mathbf{\Sigma},(x,y)\sim\mathcal{Z}} \Big[L\Big(\lambda_T F\Big(x; W^{(0)}+W^{\rho,j}+\mathbf{\Sigma}W_T, V^{(0)}+V^{\rho,j}+V_T\mathbf{\Sigma}\Big), y\Big)\Big] \pm 4\varepsilon \ . \end{aligned} \tag{C.50}$$

These imply two things.

- Putting (C.50) into (C.44), we have

$$\begin{aligned} &\mathbb{E}_{z\in\mathcal{Z}} L_F(z;\lambda_T, W^{(0)}+W^{\rho,j^*}+\widehat{\mathbf{\Sigma}}W_T, V^{(0)}+V^{\rho,j^*}+V_T\widehat{\mathbf{\Sigma}}) \qquad \text{(C.51)} \\ &= \min_j \Big\{ \mathbb{E}_{z\in\mathcal{Z}} L_F(z;\lambda_T, W^{(0)}+W^{\rho,j}+\widehat{\mathbf{\Sigma}}W_T, V^{(0)}+V^{\rho,j}+V_T\widehat{\mathbf{\Sigma}}) \Big\} \leq (1+\gamma)\mathsf{OPT} + 3\varepsilon_0 \ . \end{aligned}$$

- Putting (C.50) and (C.49) into (C.47), we have

$$\begin{aligned} &\mathbb{E}_{z\in\mathcal{D}} L_F(z;\lambda_T, W^{(0)}+W^{\rho,j}+\widehat{\mathbf{\Sigma}}W_T, V^{(0)}+V^{\rho,j}+V_T\widehat{\mathbf{\Sigma}}) \\ &\leq \mathbb{E}_{z\in\mathcal{Z}} L_F(z;\lambda_T, W^{(0)}+W^{\rho,j}+\widehat{\mathbf{\Sigma}}W_T, V^{(0)}+V^{\rho,j}+V_T\widehat{\mathbf{\Sigma}}) + 2\varepsilon_0 \end{aligned} \tag{C.52}$$

Combining (C.51) and (C.52), we immediately have

$$\mathbb{E}_{(x,y)\sim\mathcal{D}} \Big[L\Big(\lambda_T F\Big(x; W^{(0)}+W^{\rho,j^*}+\widehat{\mathbf{\Sigma}}W_T, V^{(0)}+V^{\rho,j^*}+V_T\widehat{\mathbf{\Sigma}}\Big), y\Big)\Big] \leq (1+\gamma)\mathsf{OPT} + 5\varepsilon_0$$

as desired. Scaling down ε_0 by constant finishes the proof. □

References

[1] Zeyuan Allen-Zhu and Yuanzhi Li. What Can ResNet Learn Efficiently, Going Beyond Kernels? *arXiv preprint arXiv:1905.10337*, 2019.

[2] Zeyuan Allen-Zhu and Yuanzhi Li. Can SGD Learn Recurrent Neural Networks with Provable Generalization? *ArXiv e-prints*, abs/1902.01028, February 2019.

[3] Zeyuan Allen-Zhu, Yuanzhi Li, and Zhao Song. On the convergence rate of training recurrent neural networks. *arXiv preprint arXiv:1810.12065*, 2018.

[4] Zeyuan Allen-Zhu, Yuanzhi Li, and Zhao Song. A convergence theory for deep learning via over-parameterization. In *ICML*, 2019. Full version available at `http://arxiv.org/abs/1811.03962`.

[5] Alexandr Andoni, Rina Panigrahy, Gregory Valiant, and Li Zhang. Learning polynomials with neural networks. In *International Conference on Machine Learning*, pages 1908–1916, 2014.

[6] 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.

[7] Sanjeev Arora, Simon S Du, Wei Hu, Zhiyuan Li, Ruslan Salakhutdinov, and Ruosong Wang. On exact computation with an infinitely wide neural net. *arXiv preprint arXiv:1904.11955*, 2019.

[8] Sanjeev Arora, Simon S Du, Wei Hu, Zhiyuan Li, and Ruosong Wang. Fine-grained analysis of optimization and generalization for overparameterized two-layer neural networks. *arXiv preprint arXiv:1901.08584*, 2019.

[9] Ainesh Bakshi, Rajesh Jayaram, and David P Woodruff. Learning two layer rectified neural networks in polynomial time. *arXiv preprint arXiv:1811.01885*, 2018.

[10] Peter L. Bartlett and Shahar Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3(Nov):463–482, 2002.

[11] 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.

[12] Digvijay Boob and Guanghui Lan. Theoretical properties of the global optimizer of two layer neural network. *arXiv preprint arXiv:1710.11241*, 2017.

[13] Alon Brutzkus and Amir Globerson. Globally optimal gradient descent for a convnet with gaussian inputs. *arXiv preprint arXiv:1702.07966*, 2017.

[14] Alon Brutzkus, Amir Globerson, Eran Malach, and Shai Shalev-Shwartz. Sgd learns over-parameterized networks that provably generalize on linearly separable data. In *International Conference on Learning Representations*, 2018.

[15] Amit Daniely. Complexity theoretic limitations on learning halfspaces. In *Proceedings of the forty-eighth annual ACM symposium on Theory of Computing*, pages 105–117. ACM, 2016.

[16] Amit Daniely. Sgd learns the conjugate kernel class of the network. In *Advances in Neural Information Processing Systems*, pages 2422–2430, 2017.

[17] Amit Daniely, Roy Frostig, and Yoram Singer. Toward deeper understanding of neural networks: The power of initialization and a dual view on expressivity. In *Advances in Neural Information Processing Systems (NIPS)*, pages 2253–2261, 2016.

[18] Simon S Du, Xiyu Zhai, Barnabas Poczos, and Aarti Singh. Gradient descent provably optimizes over-parameterized neural networks. *arXiv preprint arXiv:1810.02054*, 2018.

[19] Ronen Eldan, Dan Mikulincer, and Alex Zhai. The clt in high dimensions: quantitative bounds via martingale embedding. *arXiv preprint arXiv:1806.09087*, 2018.

[20] 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.

[21] Rong Ge, Jason D Lee, and Tengyu Ma. Learning one-hidden-layer neural networks with landscape design. *arXiv preprint arXiv:1711.00501*, 2017.

[22] Rong Ge, Rohith Kuditipudi, Zhize Li, and Xiang Wang. Learning two-layer neural networks with symmetric inputs. In *International Conference on Learning Representations*, 2019.

[23] Surbhi Goel and Adam Klivans. Learning neural networks with two nonlinear layers in polynomial time. *arXiv preprint arXiv:1709.06010v4*, 2018.

[24] Surbhi Goel, Varun Kanade, Adam Klivans, and Justin Thaler. Reliably learning the relu in polynomial time. In *Conference on Learning Theory*, pages 1004–1042, 2017.

[25] Noah Golowich, Alexander Rakhlin, and Ohad Shamir. Size-independent sample complexity of neural networks. In *Proceedings of the Conference on Learning Theory*, 2018.

[26] 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.

[27] 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.

[28] Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in neural information processing systems*, pages 8571–8580, 2018.

[29] Kenji Kawaguchi. Deep learning without poor local minima. In *Advances in Neural Information Processing Systems*, pages 586–594, 2016.

[30] Robert Kleinberg, Yuanzhi Li, and Yang Yuan. An alternative view: When does sgd escape local minima? *arXiv preprint arXiv:1802.06175*, 2018.

[31] 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.

[32] Jaehoon Lee, Lechao Xiao, Samuel S Schoenholz, Yasaman Bahri, Jascha Sohl-Dickstein, and Jeffrey Pennington. Wide neural networks of any depth evolve as linear models under gradient descent. *arXiv preprint arXiv:1902.06720*, 2019.

[33] Yuanzhi Li and Yingyu Liang. Learning overparameterized neural networks via stochastic gradient descent on structured data. In *Advances in Neural Information Processing Systems*, 2018.

[34] 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.

[35] Yuanzhi Li, Tengyu Ma, and Hongyang Zhang. Algorithmic regularization in over-parameterized matrix recovery. *arXiv preprint arXiv:1712.09203*, 2017.

[36] Percy Liang. CS229T/STAT231: Statistical Learning Theory (Winter 2016). `https://web.stanford.edu/class/cs229t/notes.pdf`, April 2016. accessed January 2019.

[37] 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.

[38] 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.

[39] Andreas Maurer. A vector-contraction inequality for rademacher complexities. In *International Conference on Algorithmic Learning Theory*, pages 3–17. Springer, 2016.

[40] 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.

[41] Ohad Shamir. Distribution-specific hardness of learning neural networks. *Journal of Machine Learning Research*, 19(32), 2018.

[42] 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.

[43] 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.

[44] Daniel Soudry and Yair Carmon. No bad local minima: Data independent training error guarantees for multilayer neural networks. *arXiv preprint arXiv:1605.08361*, 2016.

[45] 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.

[46] Karthik Sridharan. Machine Learning Theory (CS 6783). `http://www.cs.cornell.edu/courses/cs6783/2014fa/lec7.pdf`, 2014. accessed January 2019.

[47] 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.

[48] 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.

[49] Leslie Valiant. A theory of the learnable. *Communications of the ACM*, 1984.

[50] Santosh Vempala and John Wilmes. Polynomial convergence of gradient descent for training one-hidden-layer neural networks. *arXiv preprint arXiv:1805.02677*, 2018.

[51] Bo Xie, Yingyu Liang, and Le Song. Diversity leads to generalization in neural networks. *arXiv preprint Arxiv:1611.03131*, 2016.

[52] Greg Yang. Scaling limits of wide neural networks with weight sharing: Gaussian process behavior, gradient independence, and neural tangent kernel derivation. *arXiv preprint arXiv:1902.04760*, 2019.

[53] Alex Zhai. A high-dimensional CLT in $\mathcal{W}_2$ distance with near optimal convergence rate. *Probability Theory and Related Fields*, 170(3-4):821–845, 2018.

[54] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. In *ICLR*, 2017. arXiv 1611.03530.

[55] 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.

[56] 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.

[57] 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.