On the Neural Tangent Kernel Analysis of Randomly Pruned Neural Networks

Hongru Yang, Zhangyang Wang
VITA Group, The University of Texas at Austin

Abstract

Motivated by both theory and practice, we study how random pruning of the weights affects a neural network's neural tangent kernel (NTK). In particular, this work establishes an equivalence of the NTKs between a fully-connected neural network and its randomly pruned version. The equivalence is established under two cases. The first main result studies the infinite-width asymptotic. It is shown that given a pruning probability, for fully-connected neural networks with the weights randomly pruned at the initialization, as the width of each layer grows to infinity sequentially, the NTK of the pruned neural network converges to the limiting NTK of the original network with some extra scaling. If the network weights are rescaled appropriately after pruning, this extra scaling can be removed. The second main result considers the finite-width case. It is shown that to ensure the NTK's closeness to the limit, the dependence of width on the sparsity parameter is asymptotically linear, as the NTK's gap to its limit goes down to zero. Moreover, if the pruning probability is set to zero (i.e., no pruning), the bound on the required width matches the bound for fully-connected neural networks in previous works up to logarithmic factors. The proof of this result requires developing a novel analysis of a network structure which we called *mask-induced pseudo-networks*. Experiments are provided to evaluate our results.

1 INTRODUCTION

Can a sparse neural network achieve competitive performance as a dense network? The answer to this question can be traced back to the early work of (LeCun et al., 1990) which showed that pruning a fully-trained neural network can preserve the original network's performance while reducing the inference cost. This led to many further developments in post-training pruning such as (Han et al., 2015).

However, such gain seems hard to be transferred to the *training* phase until the discovery of the lottery ticket hypothesis (LTH) (Frankle and Carbin, 2018). The LTH states that there exists a sparse subnetwork inside a dense network at the initialization stage such that when trained in isolation, it can achieve almost matching performance with the original dense network. However, the method they used to find such a network is computationally expensive: they proposed iterative magnitude-based pruning (IMP) with rewinding which requires multiple rounds of pruning and re-training (Frankle and Carbin, 2018; Frankle et al., 2019; Chen et al., 2020). Subsequent work has been making effort in finding good sparse subnetworks at initialization with little or no training (Lee et al., 2018; Wang et al., 2019; Tanaka et al., 2020; Frankle et al., 2020; Sreenivasan et al., 2022b). Nonetheless, these methods suffer a degenerate performance than IMP. Surprisingly, even random pruning, albeit the most naive approach, has been observed to be competitive for sparse training in practice (Su et al., 2020; Frankle et al., 2020; Liu et al., 2022a).

On the theory side, a recent line of work (Malach et al., 2020; Pensia et al., 2020; Sreenivasan et al., 2022a) proves that there exists a subnetwork in a larger network at the random initialization, that can match the performance of a smaller trained network without further training. However, finding such a subnetwork is computationally hard. Other than that, little theoretical understanding of the aforementioned practical pruning method is established. Now, since random pruning is the simplest (and cheapest) avenue towards sparsity, if we can understand how good a random pruned subnetwork could be, compared to the original unpruned network, then we can establish a "lower bound"-type understanding on the effectiveness of neural network pruning, compared to other sophisticated pruning options.

To understand the success of deep networks theoretically, people have proved that running (stochastic) gradient descent on a sufficiently overparameterized deep neural network can rapidly drive the training error toward zero (Du et al., 2018; Allen-Zhu et al., 2019; Du et al., 2019; Ji and

Proceedings of the 26th International Conference on Artificial Intelligence and Statistics (AISTATS) 2023, Valencia, Spain. PMLR: Volume 206.

Telgarsky, 2019; Lee et al., 2019; Zou et al., 2020), and further, under some conditions, those networks are able to generalize (Arora et al., 2019a; Cao and Gu, 2019). All the aforementioned works either explicitly or implicitly establish that the neural network is close to its neural tangent kernel (NTK) (Jacot et al., 2018), provided that the neural network is sufficiently overparameterized. Further, if the network width grows to infinity, this matrix converges to some deterministic matrix under Gaussian initialization. In addition, it is shown that the convergence and generalization of the networks heavily depend on the condition number and the smallest eigenvalue of the NTK (Du et al., 2018, 2019; Arora et al., 2019a; Cao and Gu, 2019).

Motivated by the established theory on NTK and the recent empirical observation that random pruning becomes particularly effective if the original network is wide and deep (Liu et al., 2022a), we study the effect of randomly pruning an overparameterized neural network in the NTK regime by asking the following question:

> *How does random pruning affect the wide neural network's tangent kernel?*

If we can understand and bound the difference between the pruned network's NTK and its unpruned version, then we can hope for formalized results suggesting that the pruned neural network can achieve *fast convergence* to zero training error and yield *good generalization* after training. For practitioners, this perhaps surprising result is likely to bring random pruning back to the spotlight of model compression and efficient training, in our era when neural networks are practically scaled a lot wider and deeper, say those gigantic "foundational models" (Bommasani and et al., 2021).

Interestingly, we show that random pruning only incur limited changes to the neural network's tangent kernel. We now summarize the main contributions of this work:

- **Asymptotic limit.** The first result shows that pruning does not change the NTK much, asymptotically. More specifically, Theorem 3.1 states that given a pruning probability, the NTK of the pruned network converges to the limiting NTK of the original network at the initialization with some extra scaling factors depending on the pruning probability, as the network width grows to infinity sequentially. As a simple corollary, this scaling can be removed by rescaling the weights after pruning. Further, this sequential limit can be indeed approached by increasing the width of the network.
- **Non-asymptotic bound.** The second main result studies how large the network width needs to be to ensure that the pruned network's NTK is close to its infinite-width limit. Theorem 3.5 shows an asymptotically linear dependence of the network width on the sparsity parameter, as the gap between the pruned network's NTK and its limit goes down to zero. Further, if the pruning probability is set to zero, our width lower bound recovers the bound in (Arora et al., 2019b) for fully-connected neural networks up to some logarithmic factors. The proof of Theorem 3.5 requires developing novel analysis of a network structure that is closely related to the pruned network which we called **mask-induced pseudo-networks**. We give a detailed explanation in Section 5.2. We further validate our theory experimentally in Section 6.1.

Although our result is about the networks at the initialization, Du et al. (2018); Arora et al. (2019b); Allen-Zhu et al. (2019) suggested that the network is still closely related to the NTK after training, provided that the network is sufficiently overparameterized. Therefore, by further applying the established analysis in the previous work, the equivalence can still hold after training.

1.1 Related Work

Sparse Neural Networks in Practice. Since the discovery of the Lottery Ticket Hypothesis (Frankle and Carbin, 2018), many efforts have been made to develop methods to find good sparse networks with little overhead. Those methodologies can be divided into two groups: static sparse training and dynamic sparse training (Liu and Wang, 2023).

Static sparse training can be based on either random pruning and non-random pruning. As for random pruning, every layer can be uniformly pruned with the same pre-defined pruning ratio (Mariet and Sra, 2015; He et al., 2017; Gale et al., 2019) or the pruning ratio can be varied for different layers such as Erdö-Rényi (Mocanu et al., 2018) and Erdö-Rényi Kernel (Evci et al., 2020). For non-random pruning, those methods usually prune network weights according to some proposed saliency criteria such as SNIP (Lee et al., 2018), GraSP Wang et al. (2019), SynFlow Tanaka et al. (2020) and NTK-based score Liu and Zenke (2020). On the other hand, dynamic sparse training (Mocanu et al., 2018; Liu et al., 2021a) explores the sparsity pattern in a prune-and-grow scheme according to some criteria (Mocanu et al., 2018; Mostafa and Wang, 2019; Dettmers and Zettlemoyer, 2019; Evci et al., 2020; Ye et al., 2020; Jayakumar et al., 2020; Liu et al., 2021b). Further, the sparsity pattern can be learned by using sparsity-inducing regularizer (Yang et al., 2020). Other ways of reducing the computational cost include finding a good subnetwork and then fine-tuning (Sreenivasan et al., 2022b), and transferring lottery tickets (Morcos et al., 2019; Chen et al., 2021c). To understand the transferability of lottery tickets, Redman et al. (2021) studied IMP via the renormalization group theory in physics. Based on this development, people in practice use sparsity to improve robustness (Chen et al., 2021b; Liu et al., 2022b; Ding et al., 2021) and data efficiency (Chen et al., 2021a; Zhang et al., 2021).

Theoretical Study of The Lottery Ticket Hypothesis. On the theory side, there are works proving that a small dense

network can indeed be approximated by pruning a larger network. Malach et al. (2020) proved that a target network of width d and depth l can be indeed approximated by pruning a randomly initialized network that is of a polynomial factor (in d, l) wider and twice deeper even without further training. Ramanujan et al. (2020) empirically verified this stronger version of LTH. Later, Pensia et al. (2020) improved the widening factor to a logarithmic bound, and Sreenivasan et al. (2022a) proves that with a polylogarithmic widening factor, such a result holds even if the network weights are binary. Unsurprisingly, all of the above results are computationally hard to achieve. In addition, all these works are based on a functional approximation argument and don't consider how pruning affects the training process (and, subsequently, generalization).

Neural Tangent Kernels. Over the past few years, there is tremendous progress on understanding training overparameterized deep neural networks. A series of works (Du et al., 2018; Allen-Zhu et al., 2019; Du et al., 2019; Ji and Telgarsky, 2019; Lee et al., 2019; Zou et al., 2020) have established gradient descent convergence guarantee based on NTK (Jacot et al., 2018). Further, under some conditions, these networks are able to generalize (Arora et al., 2019a; Cao and Gu, 2019). Yang (2019); Arora et al. (2019b) provided asymptotic and non-asymptotic proofs on the limiting NTK. Further, algorithms for computing the tangent kernels are developed for various architectures (Lee et al., 2019; Arora et al., 2019b; Han et al., 2022). Other related works include studying how depth affects the diagonal of NTK (Hanin and Nica, 2019) and the smallest eigenvalue of NTK under certain data distribution assumption (Nguyen et al., 2021). Overall, the neural tangent kernel provides valuable, yet oversimplified, explanation on the neural network's success (Chizat et al., 2019).

One work in a similar spirit to ours is (Liao and Kyrillidis, 2022) which studies the convergence of training an over-parameterized one-hidden-layer neural network with sparse activation by gradient descent. Although both works consider random pruning (or masking), our work is different from theirs in a sense that the sparsity in our setting is from pruning the weights instead of neurons whereas their sparsity is obtained from masking neurons at the each step of gradient descent. Further, we consider neural networks of arbitrary depth and their work is focusing on the one-hidden-layer neural networks. Note that for the problem considered in this work, pruning (masking) neurons will be trivial since it merely incur changes to the network width.

2 PRELIMINARIES

Notations. We use lowercase letters to denote scalars and boldface letters and symbols (e.g. $\mathbf{x}$) to denote vectors and matrices. Element-wise product is denoted by $\odot$ and $\otimes$ denotes the Kronecker product. $\Pi_{\mathbf{x}}$ denotes the orthogonal projection onto the vector space generated by $\mathbf{x}$ and $\Pi_{\mathbf{A}}$ denote the orthogonal projection onto the column space of $\mathbf{A}$. We use $\mathrm{diag}(\mathbf{x})$ to denote a diagonal matrix where its diagonals are elements from the vector $\mathbf{x}$. Further, $\widetilde{O}, \widetilde{\Theta}, \widetilde{\Omega}$ are used to suppress logarithmic factors in O, Θ, Ω.

2.1 Problem Formulation

Here we want to study the training dynamics of a sparse sub-network in an ultra-wide neural network. For simplicity, we first apply our analysis on fully-connected neural networks. We denote by $f(\mathbf{x}) = f(\boldsymbol{\theta}, \mathbf{x})$ the output of the full network, $\widetilde{f}(\mathbf{x}) = f(\boldsymbol{\theta} \odot \mathbf{m}, \mathbf{x}) \in \mathbb{R}$ the output of a sparse sub-network obtained by random pruning where $\boldsymbol{\theta} \in \mathbb{R}^N$ denotes the network parameters, $\mathbf{m} \in \mathbb{R}^N$ is the sparse mask and $\mathbf{x} \in \mathbb{R}^d$ is the input. We distinguish the output of each layer of the original full networks from the sparse sub-networks by adding tilde to the symbols. For simplicity, we assume the network outputs a scalar [1]. We assume that the sparse mask is obtained from sampling each individual weight i.i.d. from a Bernoulli distribution with probability α. Formally, let $\mathbf{x} \in \mathbb{R}^d$ be the input, and denote $\widetilde{\mathbf{g}}^{(0)}(\mathbf{x}) = \mathbf{x}$ and $d_0 = d$. An L-hidden-layer fully connected network can be defined recursively as:

$$\widetilde{\mathbf{f}}^{(h)}(\mathbf{x}) = \left(\mathbf{W}^{(h)} \odot \mathbf{m}^{(h)}\right) \widetilde{\mathbf{g}}^{(h-1)}(\mathbf{x}) \in \mathbb{R}^{d_h},$$
$$\widetilde{\mathbf{g}}^{(h)}(\mathbf{x}) = \sqrt{\frac{c_\sigma}{d_h}} \sigma\left(\widetilde{\mathbf{f}}^{(h)}(\mathbf{x})\right) \in \mathbb{R}^{d_h}, \quad h = 1, 2, \ldots, L,$$

where $\mathbf{W}^{(h)} \in \mathbb{R}^{d_h \times d_{h-1}}$ is the weight matrix in the h-th layer, $\mathbf{m}^{(h)} \in \mathbb{R}^{d_h \times d_{h-1}}$ is the sparse mask for the h-th layer $\sigma : \mathbb{R} \to \mathbb{R}$ is a coordinate-wise activation function which we only consider ReLU activation in this work and $c_\sigma = \left(\mathbb{E}_{z \sim \mathcal{N}(0,1)}\left[\sigma(z)^2\right]\right)^{-1}$ is used to normalize the output of the activation. For ReLU, a simple calculation shows $c_\sigma = 2$. Let $\mathbf{m} = (\mathbf{m}^{(1)}, \ldots, \mathbf{m}^{(L+1)})$ and $\boldsymbol{\theta} = (\mathbf{W}^{(1)}, \ldots, \mathbf{W}^{(L+1)})$ represents the masks and weights in the network, respectively. All the weights $\mathbf{W}_{ij}^{(h)}$ are initialized i.i.d. from $\mathcal{N}(0, 1)$ and the masks $\mathbf{m}_{ij}^{(h)}$ are sampled i.i.d. from Bernoulli(α).

The NTK of the pruned network is given by

$$\widetilde{\boldsymbol{\Theta}}(\mathbf{x}, \mathbf{x}') = \left\langle \frac{\partial \widetilde{f}(\mathbf{x})}{\partial \boldsymbol{\theta}}, \frac{\partial \widetilde{f}(\mathbf{x}')}{\partial \boldsymbol{\theta}} \right\rangle = \sum_{h=1}^{L+1} \left\langle \frac{\partial \widetilde{f}(\mathbf{x})}{\partial \mathbf{W}^{(h)}}, \frac{\partial \widetilde{f}(\mathbf{x}')}{\partial \mathbf{W}^{(h)}} \right\rangle. \tag{1}$$

We now compute the gradient of the pruned network. Note that since the weights being pruned are staying at zero always during the training process, the gradient of the pruned network is simply the masked gradient of the unpruned net-

[1] Without loss of generality, our analysis can be extended to the vector-output case.

work. Thus, its gradient is given by

$$\frac{\partial \widetilde{f}(\mathbf{x})}{\partial \mathbf{W}^{(h)}} = \left(\widetilde{\mathbf{b}}^{(h)}(\mathbf{x})\left(\widetilde{\mathbf{g}}^{(h-1)}(\mathbf{x})\right)^\top\right) \odot \mathbf{m}^{(h)}, \quad (2)$$

where $\widetilde{\mathbf{b}}^{(h)}$ is given by

$$\widetilde{\mathbf{b}}^{(L+1)}(\mathbf{x}) = 1 \in \mathbb{R},$$
$$\widetilde{\mathbf{b}}^{(h)}(\mathbf{x}) = \sqrt{\frac{c_\sigma}{d_h}}\widetilde{\mathbf{D}}^{(h)}(\mathbf{x})(\mathbf{W}^{(h+1)} \odot \mathbf{m}^{(h+1)})^\top \widetilde{\mathbf{b}}^{(h+1)}(\mathbf{x}) \in \mathbb{R}^{d_h}, \quad (3)$$

and

$$\widetilde{\mathbf{D}}^{(h)}(\mathbf{x}) = \text{diag}\left(\dot{\sigma}\left(\widetilde{\mathbf{f}}^{(h)}(\mathbf{x})\right)\right) \in \mathbb{R}^{d_h \times d_h}, \quad h = 1, \ldots, L.$$

Further, in order to give the infinite-width limit of the NTK for the fully-connected neural networks we need to define the following quantities: for $h \in [L]$, define

$$\Sigma^{(0)}(\mathbf{x}, \mathbf{x}') = \mathbf{x}^\top \mathbf{x}',$$
$$\mathbf{\Lambda}^{(h)}(\mathbf{x}, \mathbf{x}') = \begin{bmatrix} \Sigma^{(h-1)}(\mathbf{x}, \mathbf{x}) & \Sigma^{(h-1)}(\mathbf{x}, \mathbf{x}') \\ \Sigma^{(h-1)}(\mathbf{x}', \mathbf{x}) & \Sigma^{(h-1)}(\mathbf{x}', \mathbf{x}') \end{bmatrix} \in \mathbb{R}^{2\times 2},$$
$$\Sigma^{(h)}(\mathbf{x}, \mathbf{x}') = c_\sigma \mathop{\mathbb{E}}_{(u,v)\sim\mathcal{N}(\mathbf{0}, \mathbf{\Lambda}^{(h)})}[\sigma(u)\sigma(v)],$$

and

$$\dot{\Sigma}^{(h)}(\mathbf{x}, \mathbf{x}') = c_\sigma \mathop{\mathbb{E}}_{(u,v)\sim\mathcal{N}(\mathbf{0}, \mathbf{\Lambda}^{(h)})}[\dot{\sigma}(u)\dot{\sigma}(v)],$$

where $\dot{\sigma}$ denotes the derivative of ReLU: $\dot{\sigma}(x) = \mathbb{I}(x > 0)$. We define similar quantities of $\Sigma^{(h)}$ for randomly pruned neural networks in Section 4. It can be shown that

$$\mathbf{\Theta}_\infty(\mathbf{x}, \mathbf{x}') = \lim_{d_1, d_2, \ldots, d_L \to \infty} \sum_{h=1}^{L+1} \left\langle \frac{\partial f(\boldsymbol{\theta}, \mathbf{x})}{\partial \mathbf{W}^{(h)}}, \frac{\partial f(\boldsymbol{\theta}, \mathbf{x}')}{\partial \mathbf{W}^{(h)}} \right\rangle = \sum_{h=1}^{L+1} \left(\Sigma^{(h-1)}(\mathbf{x}, \mathbf{x}') \prod_{h'=h}^{L+1} \dot{\Sigma}^{(h')}(\mathbf{x}, \mathbf{x}') \right).$$

3 MAIN RESULTS

In this section, we present the main results of our work. We show that given the pruning probability, the NTK of the pruned network is closely related to the limiting NTK of the unpruned network, if the network is sufficiently wide.

Asymptotic Limit. We first present the asymptotic limit of the pruned network as width grows to infinity.

Theorem 3.1 (The limiting NTK of randomly pruned networks)**.** *Consider an L-hidden-layer fully-connected ReLU neural network. Suppose the network weights are initialized from an i.i.d. standard Gaussian distribution and the weights except the input layer are pruned independently with probability $1 - \alpha$ at the initialization. Assume the backpropagation is computed by sampling an independent copy of weights. Then, as the width of each layer goes to infinity sequentially,*

$$\lim_{d_1, d_2, \ldots, d_L \to \infty} \widetilde{\mathbf{\Theta}}(\mathbf{x}, \mathbf{x}') = \alpha^L \mathbf{\Theta}_\infty(\mathbf{x}, \mathbf{x}'),$$

where $\widetilde{\mathbf{\Theta}}$ denotes the NTK of the pruned network and $\mathbf{\Theta}_\infty$ denotes the limiting NTK of the unpruned network.

The theorem suggests that given a pruning probability, asymptotically as the network width grows to infinity, the NTK of the randomly pruned network will converges to the limiting NTK of the full network up to some scaling depending on the pruning probability. Although we assume an independent copy of weights for the backward propagation, we will remove this assumption in Theorem 3.5.

Remark 3.2. *From (Arora et al., 2019b), if the training dataset of size n is given by $(\mathbf{X}, \mathbf{y})$, the function induced by the NTK $\mathbf{\Theta}(\mathbf{X}, \mathbf{X}) \in \mathbb{R}^{n\times n}$ is given as*

$$f_{\text{ntk}}(\mathbf{x}) = \mathbf{\Theta}(\mathbf{x}, \mathbf{X})^\top \mathbf{\Theta}(\mathbf{X}, \mathbf{X})^{-1} \mathbf{y},$$

where $\mathbf{\Theta}(\mathbf{x}, \mathbf{X}) \in \mathbb{R}^n$. Thus, any scaling factor in front of the NTK is cancelled and the actual function induced by the NTK is the same.

On the other hand, this scaling factor can be removed simply by rescaling the weights according to the pruning probability which is given in the following corollary.

Corollary 3.3. *Consider the same setting as in Theorem 3.1 except now we rescale the mask by $1/\sqrt{\alpha}$. Then, the neural tangent kernel after rescaling $\widetilde{\mathbf{\Theta}}_\alpha$ satisfies*

$$\lim_{d_1, d_2, \ldots, d_L \to \infty} \widetilde{\mathbf{\Theta}}_\alpha(\mathbf{x}, \mathbf{x}') = \mathbf{\Theta}_\infty(\mathbf{x}, \mathbf{x}').$$

Proof. Let $\widetilde{f}_\alpha$ be the network after rescaling and $\mathbf{m}_\alpha$ denote the rescaled mask, i.e., $\mathbf{m}_\alpha = \mathbf{m} \cdot (1/\sqrt{\alpha})$. Based the definition of $\widetilde{\mathbf{b}}^{(h)}(\mathbf{x})$ in Equation (3), we define $\widetilde{\mathbf{b}}_\alpha^{(L+1)} = 1$ and for $h = 1, 2, \ldots, L$,

$$\widetilde{\mathbf{b}}_\alpha^{(h)}(\mathbf{x}) := \sqrt{\frac{c_\sigma}{d_h}}\widetilde{\mathbf{D}}^{(h)}(\mathbf{x})(\mathbf{W}^{(h+1)} \odot \mathbf{m}_\alpha^{(h+1)})^\top \widetilde{\mathbf{b}}_\alpha^{(h+1)}(\mathbf{x}).$$

Based on this definition we have $\widetilde{\mathbf{b}}_\alpha^{(h)}(\mathbf{x}) = (1/\sqrt{\alpha})^{L+1-h}\widetilde{\mathbf{b}}^{(h)}(\mathbf{x})$. Similarly, define the rescaled activation output: $\widetilde{\mathbf{g}}_\alpha^{(1)} = \sqrt{\frac{c_\sigma}{d_h}}\sigma(\mathbf{W}^{(h)}\mathbf{x})$ and for $h = 2, \ldots, L$,

$$\widetilde{\mathbf{g}}_\alpha^{(h)}(\mathbf{x}) = \sqrt{\frac{c_\sigma}{d_h}}\sigma\left(\left(\mathbf{W}^{(h)} \odot \mathbf{m}^{(h)}\right)\widetilde{\mathbf{g}}_\alpha^{(h-1)}(\mathbf{x})\right) \in \mathbb{R}^{d_h}.$$

Since ReLU is positively homogeneous, i.e., $\sigma(cx) = c \cdot \sigma(x)$ for $c > 0$, we have $\widetilde{\mathbf{g}}_\alpha^{(h)}(\mathbf{x}) = (1/\sqrt{\alpha})^{h-1}\widetilde{\mathbf{g}}^{(h)}$. Thus, by Equation (2), for all $h \in [L+1]$ we have

$$\frac{\partial \widetilde{f}_\alpha(\mathbf{x})}{\partial \mathbf{W}^{(h)}} = \left(\frac{1}{\sqrt{\alpha}}\right)^L \frac{\partial \widetilde{f}(\mathbf{x})}{\partial \mathbf{W}^{(h)}}.$$

Plugging this in Equation (1) finishes our proof. □

From Asymptotic to Non-asymptotic. Since Theorem 3.1 considers sequential limits which assumes all the previous layers are already at the limit distribution when we analyze with a given layer. However, a typical drawback of such analysis is that the limit of expectation (as the previous layer's width grows to infinity) is not necessarily the same as the expectation of limit (the previous layer's width is exactly infinite). Thus, we need to justify that the network is indeed able to approach the limit by increasing width. In mathematical language, this is the same as justifying the exchange of limit for $\mathbb{E}\,\sigma(\cdot)$ and $\mathbb{E}\,\dot{\sigma}(\cdot)$. Fortunately, ReLU (and its derivative) are nice enough and we can justify this by leveraging the tools in measure-theoretic probability.

Lemma 3.4. *Conditioned on $\mathbf{g}^{(h-1)}(\mathbf{x}), \mathbf{g}^{(h-1)}(\mathbf{x}')$. Consider a fixed $i \in [d_{h+1}]$. Let*

$$X_{d_h} = \begin{bmatrix} \sqrt{\frac{c_\sigma}{d_h}} \sum_{j=1}^{d_h} \mathbf{W}_{ij}^{(h+1)} \mathbf{m}_{ij}^{(h+1)} \sigma(\widetilde{\mathbf{f}}_j^{(h)}(\mathbf{x})) \\ \sqrt{\frac{c_\sigma}{d_h}} \sum_{j=1}^{d_h} \mathbf{W}_{ij}^{(h+1)} \mathbf{m}_{ij}^{(h+1)} \sigma(\widetilde{\mathbf{f}}_j^{(h)}(\mathbf{x}')) \end{bmatrix} \in \mathbb{R}^2,$$

and let $g : \mathbb{R}^2 \to \mathbb{R}$ to be $g(x, y) \in \{\sigma(x)\sigma(y), \dot{\sigma}(x)\dot{\sigma}(y)\}$. Then,

$$\lim_{d_h \to \infty} \mathbb{E}[g(X_{d_h})] = \mathbb{E}[g(\lim_{d_h \to \infty} X_{d_h})].$$

The proof can be found in Section 8.1 in the Appendix.

Non-Asymptotic Bound. Building upon the asymptotic result in Theorem 3.1, given the pruning probability, we study how wide the neural network needs to be in order for its NTK to be close to the limiting NTK.

Theorem 3.5 (Non-asymptotic Bound of Randomly Pruned Network's NTK, Simplified Version of Theorem 9.8)**.** *Consider an L-hidden-layer fully-connected ReLU neural network with the h-th layer of width d_h. Suppose $d_1 = d_2 = \ldots = d_L = d$. Let the weights be initialized i.i.d. by standard Gaussian distribution. Suppose all the weights except the input layer are pruned independently with probability $1 - \alpha$ at the initialization and rescaled by $1/\sqrt{\alpha}$ after pruning. For $\delta \in (0, 1)$ and sufficiently small $\epsilon > 0$, if*

$$d \geq \widetilde{\Omega}\left(\max\left(\frac{1}{\alpha}\frac{L^6}{\epsilon^4}, \frac{1}{\alpha^2}\frac{L^2}{\epsilon^2}\right)\right), \tag{4}$$

then for any inputs $\mathbf{x}, \mathbf{x}' \in \mathbb{R}^{d_0}$ such that $\|\mathbf{x}\|_2 \leq 1$, $\|\mathbf{x}'\|_2 \leq 1$, with probability at least $1 - \delta$ over the randomness in the initialization and pruning, we have

$$\left|\widetilde{\mathbf{\Theta}}(\mathbf{x}, \mathbf{x}') - \mathbf{\Theta}_\infty(\mathbf{x}, \mathbf{x}')\right| \leq (L + 1)\epsilon.$$

Note that the two terms in Equation (4) has different dependence on $1/\alpha$: only $1/\alpha$ is needed for the forward propagation and $1/\alpha^2$ is needed for the backward pass, which we show in Section 5. If we let $\epsilon \to 0$, the first term in Equation (4) will dominate and the required width d only needs to scale linearly with $1/\alpha$ in this asymptotic case. We validate our theory by comparing the Monte Carlo estimate of NTK value to the limiting NTK value in Section 6.1.

Remark 3.6. *By setting the probability of pruning a given weight to be zero, our result matches the bound for fully-connected neural networks in (Arora et al., 2019b) up to logarithmic factors.*

4 THE ASYMPTOTIC LIMIT

In this section, we show how to derive the asymptotic limit of the NTK of the pruned networks, which gives a proof outline of Theorem 3.1. We give an outline of our analysis in this section and we defer the complete proof to Section 8 in Appendix.

We first introduce two quantities for randomly pruned neural networks analogous to the fully-connected networks.

Definition 4.1. *Define*

$$\widetilde{\Sigma}^{(h)}(\mathbf{x}, \mathbf{x}') := \lim_{d_1, \ldots, d_h \to \infty} \left\langle \widetilde{\mathbf{g}}^{(h)}(\mathbf{x}), \widetilde{\mathbf{g}}^{(h)}(\mathbf{x}') \right\rangle,$$

where the limit is taken sequentially from d_1 to d_h.

As a simple consequence of the law of large numbers, $\widetilde{\Sigma}^{(h)}$ is well-defined. Based on Equation (1), we compute

$$\left\langle \frac{\partial \widetilde{f}(\mathbf{x})}{\partial \mathbf{W}^{(h)}}, \frac{\partial \widetilde{f}(\mathbf{x}')}{\partial \mathbf{W}^{(h)}} \right\rangle = \left(\widetilde{\mathbf{b}}^{(h)}(\mathbf{x})\right)^\top \mathbf{G}^{(h-1)} \widetilde{\mathbf{b}}^{(h)}(\mathbf{x}'),$$

where $\mathbf{G}^{(h-1)}$ is a diagonal matrix and $\mathbf{G}_{ii}^{(h-1)} = \left\langle \widetilde{\mathbf{g}}^{(h-1)}(\mathbf{x}) \odot \mathbf{m}_i^{(h)}, \widetilde{\mathbf{g}}^{(h-1)}(\mathbf{x}') \odot \mathbf{m}_i^{(h)} \right\rangle$. Notice that under the sequential limit, as $d_{h-1} \to \infty$, $\mathbf{G}_{ii}^{(h-1)} \to \alpha \widetilde{\Sigma}^{(h-1)}(\mathbf{x}, \mathbf{x}')$. Thus, the NTK depends on analyzing both the forward propagation and the backward propagation of the pruned neural network. We show the results in the following two simple lemmas.

Lemma 4.2. *Suppose a fully-connected neural network uses ReLU as its activation and $d_1, d_2, \ldots, d_L \to \infty$ sequentially, then*

$$\widetilde{\Sigma}^{(h)}(\mathbf{x}, \mathbf{x}') = \alpha^{h-1} \Sigma^{(h)}(\mathbf{x}, \mathbf{x}'),$$

for $h = 1, 2, \ldots, L$.

Lemma 4.3. *Assume we use a fresh sample of weights in the backward pass, then*

$$\lim_{d_1, \ldots, d_L \to \infty} \left\langle \widetilde{\mathbf{b}}^{(h)}(\mathbf{x}), \widetilde{\mathbf{b}}^{(h)}(\mathbf{x}') \right\rangle = \alpha^{L+1-h} \prod_{h'=h}^{L} \dot{\Sigma}^{(h')}(\mathbf{x}, \mathbf{x}').$$

The proof of Lemma 4.3 assumes that we use an independent Gaussian copy in the backward propagation which can be removed in the next section. Combining the two lemmas provided above, we can prove Theorem 3.1.

Note that pruning the input layer creates additional difficulties since the input dimension is fixed. The NTK of the full network depends on $\Sigma^{(0)}(\mathbf{x},\mathbf{x}') = \mathbf{x}^\top\mathbf{x}'$. If we prune the input layer then $\widetilde{\Sigma}^{(0)}(\mathbf{x},\mathbf{x}') = (\mathbf{m}\odot\mathbf{x})^\top(\mathbf{m}\odot\mathbf{x}')$ which is random. In this case, it seems hard to relate $\widetilde{\Sigma}^{(1)}(\mathbf{x},\mathbf{x}')$ to $\Sigma^{(1)}(\mathbf{x},\mathbf{x}')$ in this asymptotic regime.

5 THE NON-ASYMPTOTIC BOUND

In this section, we give a proof outline of Theorem 3.5. Since in this section we are only talking about the pruned network, there is no longer ambiguity in distinguishing pruned and unpruned networks. For notation ease, we remove the tilde above all the symbols of the quantities in the pruned network. In addition, we use $\mathbf{m}_i^{(h)}$ to denote the i-th row of $\mathbf{m}^{(h)}$ and similar for $\mathbf{w}_i^{(h)}$. From a high level, the proof consists of analyzing the forward propagation and the backward propagation. We give a complete treatment in Section 9 in the Appendix.

5.1 Analyzing the Forward Propagation

We first present our result on the forward propagation.

Theorem 5.1 (Simplified Version of Theorem 9.11)**.** *Consider the same setting as in Theorem 3.5. There exist constants c such that if $\epsilon \le \min(c, \frac{1}{L})$ and*

$$d \ge \widetilde{\Omega}\left(\frac{1}{\alpha}\frac{L^2}{\epsilon^2}\right),$$

then with probability $1-\delta$ over the randomness in the initialization of all the weights and masks, for all $h\in[L]$, $i\in[d_{h+1}]$, $(\mathbf{x}^{(1)},\mathbf{x}^{(2)})\in\{(\mathbf{x},\mathbf{x}),(\mathbf{x},\mathbf{x}'),(\mathbf{x}',\mathbf{x}')\}$,

$$\left|\left(\mathbf{g}^{(h)}(\mathbf{x}^{(1)})\odot\mathbf{m}_i^{(h+1)}\right)^\top\left(\mathbf{g}^{(h)}(\mathbf{x}^{(2)})\odot\mathbf{m}_i^{(h+1)}\right) - \Sigma^{(h)}(\mathbf{x}^{(1)},\mathbf{x}^{(2)})\right| \le \epsilon.$$

Our result provides the required width to ensure the activation of each layer is close to its limit. The dependence on $1/\alpha$ is precisely due to the presence of random masks and notice that each mask is a sub-Gaussian random variable with variance proxy $1/\alpha$.

5.2 Analyzing the Backward Propagation

In this section, we show that $\left\langle \mathbf{b}^{(h)}(\mathbf{x}^{(1)}), \mathbf{b}^{(h)}(\mathbf{x}^{(2)})\right\rangle \approx \prod_{h'=h}^{L}\dot{\Sigma}^{(h')}(\mathbf{x}^{(1)},\mathbf{x}^{(2)})$ under the assumption that the event in Theorem 5.1 occurs. This is where we formally justify the fresh Gaussian copy trick. We consider a fixed pair $(\mathbf{x}^{(1)},\mathbf{x}^{(2)})$ and suppress the dependence on inputs when there is no confusion. We do this by induction: assume $\mathbf{b}^{(h+1)}(\mathbf{x}^{(1)})^\top\mathbf{b}^{(h+1)}(\mathbf{x}^{(2)}) \approx \prod_{h'=h+1}^{L}\dot{\Sigma}(\mathbf{x}^{(1)},\mathbf{x}^{(2)})$. Define $\mathbf{G}_i^{(h)} := [(\mathbf{g}^{(h)}(\mathbf{x})\odot\mathbf{m}_i^{(h+1)}),\ (\mathbf{g}^{(h)}(\mathbf{x}')\odot\mathbf{m}_i^{(h+1)})]$ and $\mathbf{F}_i^{(h+1)} := (\mathbf{W}^{(h+1)}\odot\mathbf{m}^{(h+1)})\mathbf{G}_i^{(h)}$. Notice that the dependence of $\mathbf{b}^{(h+1)}$ on $\mathbf{W}^{(h+1)}$ is by $\mathbf{F}_i^{(h+1)}$. If $\mathbf{W}^{(h+1)}$ is independent to $\mathbf{b}^{(h+1)}$ (which it isn't), then

$$\mathcal{E} := \mathop{\mathbb{E}}_{\mathbf{W}^{(h+1)}}\left[(\mathbf{b}^{(h)}(\mathbf{x}^{(1)}))^\top\mathbf{b}^{(h)}(\mathbf{x}^{(2)})\right] \tag{5}$$
$$= \frac{2}{d_h}\sum_i \mathbf{b}_i^{(h+1)}(\mathbf{x}^{(1)})\mathbf{b}_i^{(h+1)}(\mathbf{x}^{(2)})\mathrm{Tr}(\mathbf{M}_i^{(h+1)}\mathbf{D}\mathbf{M}_i^{(h+1)}).$$

It is easy to show that $\mathrm{Tr}(\mathbf{M}_i^{(h+1)}\mathbf{D}\mathbf{M}_i^{(h+1)}) \approx \dot{\Sigma}$ and $(\mathbf{b}^{(h)}(\mathbf{x}^{(1)}))^\top\mathbf{b}^{(h)}(\mathbf{x}^{(2)})$ is close to its expectation. Then by induction hypothesis we are done. Now we show that $\mathbf{W}^{(h+1)}$ is nearly independent to $\mathbf{b}^{(h+1)}$. Recall a **special property of the standard Gaussian**: given $\mathbf{w}\sim\mathcal{N}(\mathbf{0},\mathbf{I})$ and two fixed vectors $\mathbf{x},\mathbf{y}$, if $\mathbf{x}^\top\mathbf{y}=0$, then $\mathbf{w}^\top\mathbf{x}$ and $\mathbf{w}^\top\mathbf{y}$ are independent. Thus, conditioned on $\mathbf{b}^{(h+1)}$, $\mathbf{G}_i^{(h)},\mathbf{F}_i^{(h+1)},\mathbf{m}^{(h+1)}$, we have $\mathbf{w}_i^{(h+1)}\Pi_{\mathbf{G}_i}^{\perp} \overset{\mathcal{D}}{=} \widetilde{\mathbf{w}}_i^{(h+1)}\Pi_{\mathbf{G}_i}^{\perp}$ where $\widetilde{\mathbf{w}}_i^{(h+1)}$ is an i.i.d. copy of $\mathbf{w}_i^{(h+1)}$. Let

$$\mathbf{b}_\perp^{(h)} := \left(\mathbf{b}^{(h+1)}\right)^\top\begin{bmatrix}((\widetilde{\mathbf{w}}_1^{(h+1)})^\top\Pi_{\mathbf{G}_1}^{\perp})\odot\mathbf{m}_1^{(h+1)}\\ \vdots\\ ((\widetilde{\mathbf{w}}_{d_{h+1}}^{(h+1)})^\top\Pi_{\mathbf{G}_{d_{h+1}}}^{\perp})\odot\mathbf{m}_{d_{h+1}}^{(h+1)}\end{bmatrix}\mathbf{D},$$

$$\mathbf{b}_\parallel^{(h)} := \left(\mathbf{b}^{(h+1)}\right)^\top\begin{bmatrix}((\mathbf{w}_1^{(h+1)})^\top\Pi_{\mathbf{G}_1})\odot\mathbf{m}_1^{(h+1)}\\ \vdots\\ ((\mathbf{w}_{d_{h+1}}^{(h+1)})^\top\Pi_{\mathbf{G}_{d_{h+1}}})\odot\mathbf{m}_{d_{h+1}}^{(h+1)}\end{bmatrix}\mathbf{D}.$$

Notice that $\mathbf{b}^{(h)} = \mathbf{b}_\perp^{(h)} + \mathbf{b}_\parallel^{(h)}$. Next, we are going to show that the main contribution of $\left\langle\mathbf{b}^{(h)}(\mathbf{x}),\mathbf{b}^{(h)}(\mathbf{x}')\right\rangle$ is from $\mathbf{b}_\perp^{(h)}$ and $(\mathbf{b}_\perp^{(h)})^\top\mathbf{b}_\perp^{(h)} \approx \mathcal{E}$ whereas the contribution from the dependent part $\mathbf{b}_\parallel^{(h)}$ is small. We show these two results in Proposition 5.2 and Proposition 5.3.

Proposition 5.2 (Informal Version of Proposition 9.20)**.** *Under some appropriate conditions, with probability at least $1-\delta_2/2$ over the randomness in $\mathbf{W}^{(h+1)}$, for any $(\mathbf{x}^{(1)},\mathbf{x}^{(2)})\in\{(\mathbf{x},\mathbf{x}),(\mathbf{x},\mathbf{x}'),(\mathbf{x}',\mathbf{x}')\}$, we have*

$$\left|\frac{2}{d_h}\left(\mathbf{b}_\perp^{(h)}(\mathbf{x}^{(1)})\right)^\top\mathbf{b}_\perp^{(h)}(\mathbf{x}^{(2)}) - \mathcal{E}\right| \le O\left(\sqrt{\frac{\log\frac{1}{\delta_2}}{\alpha d_h}}\right).$$

Proposition 5.3 (Informal Version of Proposition 9.27)**.** *Under some appropriate conditions, if $d\ge\widetilde{\Omega}(\frac{1}{\alpha}\frac{L^2}{\epsilon^2})$, with probability $1-\delta_2/2$ over the randomness in the initialization of $\mathbf{W}^{(h+1)},\mathbf{m}^{(h+1)},\ldots,\mathbf{W}^{(L+1)},\mathbf{m}^{(L+1)}$,*

$$\sqrt{\frac{1}{d_h}}\left\|\mathbf{b}_\parallel^{(h)}\right\|_2 \le O\left(\sqrt{\frac{1}{\alpha^2 d_h}\log\frac{1}{\delta_2}}\right).$$

The proof of Proposition 5.2 requires some intricate calculation and then applying Gaussian chaos concentration bound which is left in Section 9.6.1 in Appendix. We now give a detailed description of the proof of Proposition 5.3. For the ease of presentation, we omit the dependence on layer and inputs when there is no confusion. First of all, we can decompose $\Pi_{\mathbf{G}_i} = \Pi_{\mathbf{g}(\mathbf{x})\odot\mathbf{m}_i} + \Pi_{\mathbf{G}_i/\mathbf{g}(\mathbf{x})\odot\mathbf{m}_i}$ where $\mathbf{G}_i/\mathbf{g}(\mathbf{x}) \odot \mathbf{m}_i$ denotes the subspace of $\mathbf{G}_i$ orthogonal to $\mathbf{g}(\mathbf{x})\odot\mathbf{m}_i$. Bounding the second part is simple by utilizing the special property of the standard Gaussian. We now focus on bounding the first part. Writing $\mathbf{g}_i$ short for $\mathbf{g}\odot\mathbf{m}_i$,

$$\left(\mathbf{b}^{(h+1)}\right)^\top \begin{bmatrix} ((\mathbf{w}_1^{(h+1)})^\top \Pi_{\mathbf{g}_1}) \odot \mathbf{m}_1^{(h+1)} \\ \vdots \\ ((\mathbf{w}_{d_{h+1}}^{(h+1)})^\top \Pi_{\mathbf{g}_{d_{h+1}}}) \odot \mathbf{m}_{d_{h+1}}^{(h+1)} \end{bmatrix}$$
$$= \frac{1}{\sqrt{\alpha}} \sum_i \mathbf{b}_i^{(h+1)} (\mathbf{w}_i^{(h+1)})^\top \frac{\mathbf{g}_i \mathbf{g}_i^\top}{\|\mathbf{g}_i\|_2^2}. \tag{6}$$

The presence of the mask introduces further difficulties in the analysis. In particular, without the pruning masks, the above vector nicely simplifies to

$$(\mathbf{b}^{(h+1)})^\top \mathbf{W}^{(h+1)} \mathbf{g}^{(h)} \frac{\mathbf{g}^{(h)}(\mathbf{x})}{\left\|\mathbf{g}^{(h)}(\mathbf{x})\right\|_2} = f(\mathbf{x}) \frac{\mathbf{g}^{(h)}(\mathbf{x})}{\left\|\mathbf{g}^{(h)}(\mathbf{x})\right\|_2}. \tag{7}$$

We would like the above relation to also hold for pruned network. However, this is not true since each $\mathbf{g}_i$ is different. A closer examination of the expression in Equation (6) tells us that like the relation in Equation (7), the i-th coordinate of this vector can be written as the product of $\mathbf{g}_i^{(h)}(\mathbf{x})$ and some structure similar to the pruned network, which we call *mask-induced pseudo-networks*.

5.2.1 Mask-Induced Pseudo-Network

Definition 5.4 (Pseudo-network induced by mask)**.** *Define the pseudo-network induced by the h-th layer j-th column of sparse masks $\mathbf{m}^{(h)}$ for all $h \in \{2, \dots, L\}$, $j \in [d_{h-1}]$ and $h' \in \{h+1, h+2, \dots, L\}$ to be*

$$\mathbf{g}^{(h,j,h)} = \sqrt{\frac{c_\sigma}{d_h}} \mathbf{D}^{(h)} \mathrm{diag}_i \left(\frac{\mathbf{m}_{ij}^{(h)} \sqrt{\alpha}}{\left\|\mathbf{g}^{(h-1)} \odot \mathbf{m}_i^{(h)}\right\|_2^2} \right) \mathbf{f}^{(h)},$$
$$\mathbf{f}^{(h,j,h')} = \left(\mathbf{W}^{(h')} \odot \mathbf{m}^{(h')}\right) \mathbf{g}^{(h,j,h'-1)},$$
$$\mathbf{g}^{(h,j,h')} = \sqrt{\frac{c_\sigma}{d_{h'}}} \mathbf{D}^{(h')}(\mathbf{x}) \mathbf{f}^{(h,j,h')}.$$

The output of this pseudo-network is $f^{(h,j,L+1)}$.

Using this definition, we can write

$$\left(\frac{1}{\sqrt{\alpha}} \sum_i \mathbf{b}_i^{(h+1)} (\mathbf{w}_i^{(h+1)})^\top \frac{\mathbf{g}_i^{(h)} \mathbf{g}_i^{(h)\top}}{\left\|\mathbf{g}_i^{(h)}\right\|_2^2} \right)_j$$
$$= \frac{1}{\alpha} \mathbf{g}_j^{(h)} f^{(h+1,j,L+1)}.$$

Now our goal is to show that $|f^{(h+1,j,L+1)}| = \widetilde{O}(1)$ for all h, j. This requires us to analyze the forward propagation of this pseudo-network. Specifically, we need to show that the norm of $\mathbf{g}^{(h,j,h')}$ is $O(1)$ for all $h < h' \le L$. However, whether a neuron turns on depends on the input it receives in the pruned network instead of the pseudo-network. Nonetheless, we show that this doesn't matter when we consider the *norm* of the activation in the pseudo-network, since it has the *same* distribution as the activation in the pruned network.

Proposition 5.5. *For any given nonzero vectors $\mathbf{x}, \mathbf{y}$, the distribution of $(\mathbf{w}^\top \mathbf{x})^2 \mathbb{I}(\mathbf{w}^\top \mathbf{y} > 0)$ is the same as $(\mathbf{w}^\top \mathbf{x})^2 \mathbb{I}(\mathbf{w}^\top \mathbf{x} > 0)$ where $\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$.*

This proposition says we can bound the norm of the activation by ignoring which neurons turn on. Thus, utilizing this result, we can analyze the forward propagation of the pseudo-network just as analyzing the pruned network and show that we indeed have $|f^{(h+1,j,L+1)}| = \widetilde{O}(1)$ for all h, j. This completes the proof outline of Proposition 5.3.

6 EXPERIMENTS

This section presents our empirical results. Our results contain two parts: first we validate our theory; then, we evaluate our theory on real world dataset.

6.1 Validating Our Theory

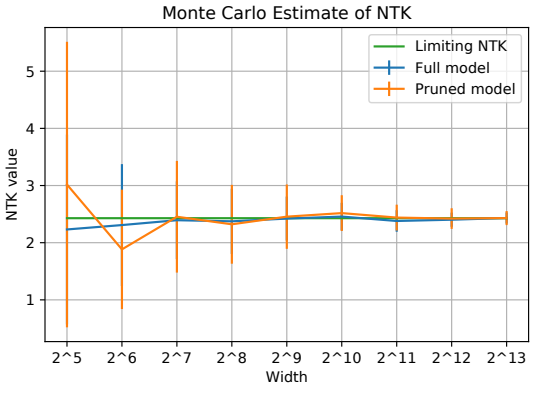


Figure 1: Figure (a) validates Corollary 3.3 which shows the empirical NTK value generated by the full model and pruned model with varying width compared with theoretical NTK limit. The limiting NTK value is computed by a known closed-form formula in (Arora et al., 2019b).

Validation of Corollary 3.3 (and, thus, Theorem 3.1): We show that the empirical NTK value computed from the pruned network converges to the theoretical NTK limit as

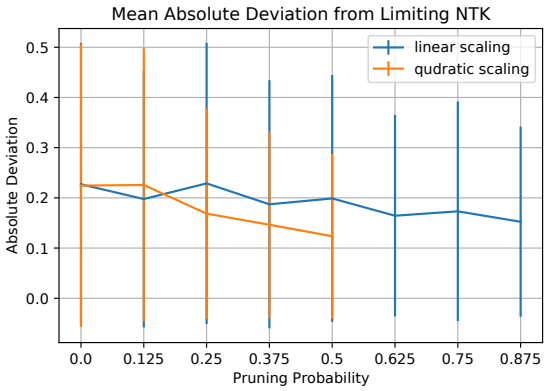


Figure 2: The results of the mean absolute deviation of the empirical NTK value from the limiting NTK. At each pruning probability, the width of the network is scaled quadratically and linearly with respect to $1/\alpha$.

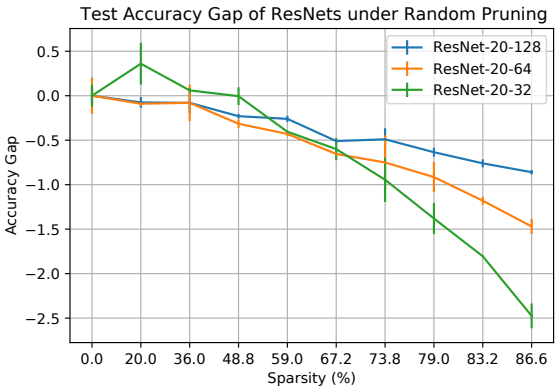


Figure 3: Performance of random pruning without rescaling on ResNet-20 of different widths. Sparsity on the x-axis means the fraction of weights remaining in IMP and pruning probability in random pruning.

the width increases. We use fully-connected neural networks with 3-hidden layers of the same width as our model. We rescale the weights by $1/\sqrt{\alpha}$ after pruning. We first randomly generate two data points $\mathbf{x}, \mathbf{y}$ and then randomly initialize the networks with Gaussian distribution. We fix the pruning probability to be $1/2$ and vary the width from 32 to 8192. For each trial, we create 64 samples of the empirical NTK values generated by the unpruned and pruned networks, and plot their mean. Figure 1 shows that, as the width increase, our empirical estimates from both unpruned and pruned model converge to the limiting NTK value.

Validation of Theorem 3.5: Theorem 3.5 suggests that d_h needs to scale asymptotically linearly with respect to $1/\alpha$ to maintain the gap between the empirical NTK and limiting NTK. To evaluate our Theorem 3.5, we start with a full model of width 1024 and then prune the model with various probability $1-\alpha$ while scaling the width quadratically and linearly with $1/\alpha$. Since quadratically scaling width is expensive, we stop at 0.5 pruning probability. We generate 100 samples for each pruning probability and take their mean absolute deviation from the theoretically computed NTK value. The result is shown in Figure 2. In both cases, the gap to the limiting NTK is non-increasing.

6.2 On the Real World Data

In this section, we further evaluate our theory on real-world data. Our theory suggests that if the network is wide enough, the pruned networks should retain much of the performance of the full networks. We note that here we prune all layers of the neural networks. We adopt the implementation from Chen et al. (2021c).

We extensively test our theory across different neural network architectures and datasets. For pruning methods, in addition to random pruning (with and without rescaling weights after pruning), we also include Iterative Magnitude-based Pruning (IMP) in our experiments. We train fully-connected neural networks on MNIST dataset Deng (2012) and, VGGs and ResNets He et al. (2016) on CIFAR-10 Krizhevsky et al. (2009), and vary the width of these architectures. We generate each data point in the plot by averaging over 2 independent runs. We defer the detailed experiment setup in Section 10.1 in Appendix.

Results. In Figure 3, for random pruning without rescaling, the testing performance gap narrows as the network width is getting larger. For ResNet-20-128, at sparsity 86.6%, the performance of random pruning and the full model is within 1% on CIFAR-10. Similar results have been observed for other pruning methods and other architectures and datasets. Further experiment results are shown in Section 10.2 in Appendix.

7 DISCUSSION AND FUTURE WORK

In this paper, we establish an equivalence between the NTK of a randomly pruned neural network and the limiting NTK of the unpruned network under both asymptotic and finite-width cases. For the finite width case, we establish an asymptotically linear dependence of network width on the sparsity parameter $1/\alpha$. One open problem is whether $1/\alpha^2$ dependence is indeed necessary for the backward propagation so that the width dependence on $1/\alpha$ can be improved to exactly linear instead of asymptotically linear. We leave further investigation on this open problem.

One limitation of our current analysis is that it only applies to random pruning and assumes that the pruning distribution is completely independent from the weight initialization. Therefore, our analysis is not valid for magnitude-based pruning or gradient-based pruning, as the weights being pruned have internal correlations with the magnitude of the weights. Another limitation is that the NTK analysis inherently restricts the neural network's ability to perform

feature learning. We believe that the advantages of pruning, such as improving network generalization, can be demonstrated in a feature learning setting. This direction is left for future research and exploration as well.

Acknowledgements

H. Yang and Z. Wang thank the anonymous reviewers for the helpful feedback and comments. Z. Wang is supported by NSF Scale-MoDL (award number: 2133861).

References

Allen-Zhu, Z., Li, Y., and Song, Z. (2019). A convergence theory for deep learning via over-parameterization. In *International Conference on Machine Learning*, pages 242–252. PMLR.

Arora, S., Du, S., Hu, W., Li, Z., and Wang, R. (2019a). Fine-grained analysis of optimization and generalization for overparameterized two-layer neural networks. In *International Conference on Machine Learning*, pages 322–332. PMLR.

Arora, S., Du, S. S., Hu, W., Li, Z., Salakhutdinov, R. R., and Wang, R. (2019b). On exact computation with an infinitely wide neural net. *Advances in Neural Information Processing Systems*, 32.

Billingsley, P. (1999). Convergence of probability measures. *John Wiley & Sons INC, New York*, 2(2.4).

Bommasani, R. and et al. (2021). On the opportunities and risks of foundation models.

Boucheron, S., Lugosi, G., and Massart, P. (2013). *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press.

Cao, Y. and Gu, Q. (2019). Generalization bounds of stochastic gradient descent for wide and deep neural networks. *Advances in neural information processing systems*, 32.

Chen, T., Cheng, Y., Gan, Z., Liu, J., and Wang, Z. (2021a). Data-efficient gan training beyond (just) augmentations: A lottery ticket perspective. *Advances in Neural Information Processing Systems*, 34.

Chen, T., Frankle, J., Chang, S., Liu, S., Zhang, Y., Wang, Z., and Carbin, M. (2020). The lottery ticket hypothesis for pre-trained bert networks. *Advances in neural information processing systems*, 33:15834–15846.

Chen, T., Zhang, Z., Balachandra, S., Ma, H., Wang, Z., Wang, Z., et al. (2021b). Sparsity winning twice: Better robust generalization from more efficient training. In *International Conference on Learning Representations*.

Chen, X., Cheng, Y., Wang, S., Gan, Z., Liu, J., and Wang, Z. (2021c). The elastic lottery ticket hypothesis. *Advances in Neural Information Processing Systems*, 34.

Chizat, L., Oyallon, E., and Bach, F. (2019). On lazy training in differentiable programming. *Advances in Neural Information Processing Systems*, 32.

Daniely, A., Frostig, R., and Singer, Y. (2016). Toward deeper understanding of neural networks: The power of initialization and a dual view on expressivity. *Advances In Neural Information Processing Systems*, 29:2253–2261.

Deng, L. (2012). The mnist database of handwritten digit images for machine learning research [best of the web]. *IEEE Signal Processing Magazine*, 29(6):141–142.

Dettmers, T. and Zettlemoyer, L. (2019). Sparse networks from scratch: Faster training without losing performance. *arXiv preprint arXiv:1907.04840*.

Ding, S., Chen, T., and Wang, Z. (2021). Audio lottery: Speech recognition made ultra-lightweight, noise-robust, and transferable. In *International Conference on Learning Representations*.

Du, S., Lee, J., Li, H., Wang, L., and Zhai, X. (2019). Gradient descent finds global minima of deep neural networks. In *International Conference on Machine Learning*, pages 1675–1685. PMLR.

Du, S. S., Zhai, X., Poczos, B., and Singh, A. (2018). Gradient descent provably optimizes over-parameterized neural networks. In *International Conference on Learning Representations*.

Evci, U., Gale, T., Menick, J., Castro, P. S., and Elsen, E. (2020). Rigging the lottery: Making all tickets winners. In *International Conference on Machine Learning*, pages 2943–2952. PMLR.

Frankle, J. and Carbin, M. (2018). The lottery ticket hypothesis: Finding sparse, trainable neural networks. In *International Conference on Learning Representations*.

Frankle, J., Dziugaite, G. K., Roy, D., and Carbin, M. (2020). Pruning neural networks at initialization: Why are we missing the mark? In *International Conference on Learning Representations*.

Frankle, J., Dziugaite, G. K., Roy, D. M., and Carbin, M. (2019). Stabilizing the lottery ticket hypothesis. *arXiv preprint arXiv:1903.01611*.

Gale, T., Elsen, E., and Hooker, S. (2019). The state of sparsity in deep neural networks. *arXiv preprint arXiv:1902.09574*.

Grimmett, G. and Stirzaker, D. (2020). *Probability and random processes*. Oxford university press.

Han, I., Zandieh, A., Lee, J., Novak, R., Xiao, L., and Karbasi, A. (2022). Fast neural kernel embeddings for general activations. *arXiv preprint arXiv:2209.04121*.

Han, S., Mao, H., and Dally, W. J. (2015). Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding. *arXiv preprint arXiv:1510.00149*.

Hanin, B. and Nica, M. (2019). Finite depth and width corrections to the neural tangent kernel. In *International Conference on Learning Representations*.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

He, Y., Zhang, X., and Sun, J. (2017). Channel pruning for accelerating very deep neural networks. In *Proceedings of the IEEE international conference on computer vision*, pages 1389–1397.

Jacot, A., Gabriel, F., and Hongler, C. (2018). Neural tangent kernel: Convergence and generalization in neural networks. *Advances in neural information processing systems*, 31.

Jayakumar, S., Pascanu, R., Rae, J., Osindero, S., and Elsen, E. (2020). Top-kast: Top-k always sparse training. *Advances in Neural Information Processing Systems*, 33:20744–20754.

Ji, Z. and Telgarsky, M. (2019). Polylogarithmic width suffices for gradient descent to achieve arbitrarily small test error with shallow relu networks. In *International Conference on Learning Representations*.

Krizhevsky, A., Hinton, G., et al. (2009). Learning multiple layers of features from tiny images.

LeCun, Y., Denker, J. S., and Solla, S. A. (1990). Optimal brain damage. In *Advances in neural information processing systems*, pages 598–605.

Lee, J., Xiao, L., Schoenholz, S., Bahri, Y., Novak, R., Sohl-Dickstein, J., and Pennington, J. (2019). Wide neural networks of any depth evolve as linear models under gradient descent. *Advances in neural information processing systems*, 32:8572–8583.

Lee, N., Ajanthan, T., and Torr, P. (2018). Snip: Single-shot network pruning based on connection sensitivity. In *International Conference on Learning Representations*.

Liao, F. and Kyrillidis, A. (2022). On the convergence of shallow neural network training with randomly masked neurons. *Transactions on Machine Learning Research*.

Liu, S., Chen, T., Chen, X., Shen, L., Mocanu, D. C., Wang, Z., and Pechenizkiy, M. (2022a). The unreasonable effectiveness of random pruning: Return of the most naive baseline for sparse training. In *International Conference on Learning Representations*.

Liu, S., Mocanu, D. C., Matavalam, A. R. R., Pei, Y., and Pechenizkiy, M. (2021a). Sparse evolutionary deep learning with over one million artificial neurons on commodity hardware. *Neural Computing and Applications*, 33(7):2589–2604.

Liu, S. and Wang, Z. (2023). Ten lessons we have learned in the new" sparseland": A short handbook for sparse neural network researchers. *arXiv preprint arXiv:2302.02596*.

Liu, S., Yin, L., Mocanu, D. C., and Pechenizkiy, M. (2021b). Do we actually need dense over-parameterization? in-time over-parameterization in sparse training. In *International Conference on Machine Learning*, pages 6989–7000. PMLR.

Liu, S., Zhu, Z., Qu, Q., and You, C. (2022b). Robust training under label noise by over-parameterization. *arXiv preprint arXiv:2202.14026*.

Liu, T. and Zenke, F. (2020). Finding trainable sparse networks through neural tangent transfer. In *International Conference on Machine Learning*, pages 6336–6347. PMLR.

Malach, E., Yehudai, G., Shalev-Schwartz, S., and Shamir, O. (2020). Proving the lottery ticket hypothesis: Pruning is all you need. In *International Conference on Machine Learning*, pages 6682–6691. PMLR.

Mann, H. B. and Wald, A. (1943). On stochastic limit and order relationships. *The Annals of Mathematical Statistics*, 14(3):217–226.

Mariet, Z. and Sra, S. (2015). Diversity networks: Neural network compression using determinantal point processes. *arXiv preprint arXiv:1511.05077*.

Mocanu, D. C., Mocanu, E., Stone, P., Nguyen, P. H., Gibescu, M., and Liotta, A. (2018). Scalable training of artificial neural networks with adaptive sparse connectivity inspired by network science. *Nature communications*, 9(1):1–12.

Morcos, A., Yu, H., Paganini, M., and Tian, Y. (2019). One ticket to win them all: generalizing lottery ticket initializations across datasets and optimizers. *Advances in neural information processing systems*, 32.

Mostafa, H. and Wang, X. (2019). Parameter efficient training of deep convolutional neural networks by dynamic sparse reparameterization. In *International Conference on Machine Learning*, pages 4646–4655. PMLR.

Nguyen, Q., Mondelli, M., and Montufar, G. F. (2021). Tight bounds on the smallest eigenvalue of the neural tangent kernel for deep relu networks. In *International Conference on Machine Learning*, pages 8119–8129. PMLR.

Pensia, A., Rajput, S., Nagle, A., Vishwakarma, H., and Papailiopoulos, D. (2020). Optimal lottery tickets via subset sum: Logarithmic over-parameterization is sufficient. *Advances in Neural Information Processing Systems*, 33:2599–2610.

Ramanujan, V., Wortsman, M., Kembhavi, A., Farhadi, A., and Rastegari, M. (2020). What's hidden in a randomly weighted neural network? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11893–11902.

Redman, W. T., Chen, T., Dogra, A. S., and Wang, Z. (2021). Universality of deep neural network lottery tickets: A renormalization group perspective. *arXiv preprint arXiv:2110.03210*.

Sreenivasan, K., Rajput, S., Sohn, J.-Y., and Papailiopoulos, D. (2022a). Finding nearly everything within random binary networks. In *International Conference on Artificial Intelligence and Statistics*, pages 3531–3541. PMLR.

Sreenivasan, K., Sohn, J.-y., Yang, L., Grinde, M., Nagle, A., Wang, H., Lee, K., and Papailiopoulos, D. (2022b). Rare gems: Finding lottery tickets at initialization. *arXiv preprint arXiv:2202.12002*.

Su, J., Chen, Y., Cai, T., Wu, T., Gao, R., Wang, L., and Lee, J. D. (2020). Sanity-checking pruning methods: Random tickets can win the jackpot. *Advances in Neural Information Processing Systems*, 33:20390–20401.

Tanaka, H., Kunin, D., Yamins, D. L., and Ganguli, S. (2020). Pruning neural networks without any data by iteratively conserving synaptic flow. *Advances in Neural Information Processing Systems*, 33:6377–6389.

Wainwright, M. J. (2019). *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press.

Wang, C., Zhang, G., and Grosse, R. (2019). Picking winning tickets before training by preserving gradient flow. In *International Conference on Learning Representations*.

Yang, G. (2019). Scaling limits of wide neural networks with weight sharing: Gaussian process behavior, gradient independence, and neural tangent kernel derivation. *arXiv preprint arXiv:1902.04760*.

Yang, H., Wen, W., and Li, H. (2020). Deephoyer: Learning sparser neural network with differentiable scale-invariant sparsity measures. In *International Conference on Learning Representations*.

Ye, M., Gong, C., Nie, L., Zhou, D., Klivans, A., and Liu, Q. (2020). Good subnetworks provably exist: Pruning via greedy forward selection. In *International Conference on Machine Learning*, pages 10820–10830. PMLR.

Zhang, Z., Chen, X., Chen, T., and Wang, Z. (2021). Efficient lottery ticket finding: Less data is more. In *International Conference on Machine Learning*, pages 12380–12390. PMLR.

Zou, D., Cao, Y., Zhou, D., and Gu, Q. (2020). Gradient descent optimizes over-parameterized deep relu networks. *Machine Learning*, 109(3):467–492.

Supplementary Materials

Contents

8 ASYMPTOTIC ANALYSIS (Proof of Theorem 3.1)

This section is devoted to prove the asymptotic limit of the pruned networks' NTK. Recall that we use tilde over a symbol to denote the quantity in the *pruned* network and the corresponding symbol without tilde denotes the quantity in the unpruned network.

Theorem 8.1 (The limiting NTK of randomly pruned networks, Restatement of Theorem 3.1)**.** *Consider an L-hidden-layer fully-connected ReLU neural network. Suppose the network weights are initialized from an i.i.d. standard Gaussian distribution and the weights except the input layer are pruned independently with probability $1-\alpha$ at the initialization. Assume the backpropagation is computed by sampling a independent copy of weights. Then, as the width of each layer goes to infinity sequentially,*

$$\lim_{d_1,d_2,\ldots,d_L\to\infty} \widetilde{\mathbf{\Theta}}(\mathbf{x},\mathbf{x}') = \alpha^L \mathbf{\Theta}_\infty(\mathbf{x},\mathbf{x}'),$$

where $\widetilde{\mathbf{\Theta}}$ denotes the NTK of the pruned network and $\mathbf{\Theta}_\infty$ denotes the limiting NTK of the unpruned network.

For the pruned neural networks, its gradient is given by

$$\frac{\partial \widetilde{f}(\mathbf{x})}{\partial \mathbf{W}^{(h)}} = \left(\widetilde{\mathbf{b}}^{(h)}(\mathbf{x})\left(\widetilde{\mathbf{g}}^{(h-1)}(\mathbf{x})\right)^\top\right) \odot \mathbf{m}^{(h)}, \quad h = 2,\ldots,L+1$$

where

$$\widetilde{\mathbf{b}}^{(h)}(\mathbf{x}) = \begin{cases} 1 \in \mathbb{R}, & h = L+1 \\ \sqrt{\frac{c_\sigma}{d_h}}\widetilde{\mathbf{D}}^{(h)}(\mathbf{x})(\mathbf{W}^{(h+1)} \odot \mathbf{m}^{(h+1)})^\top \widetilde{\mathbf{b}}^{(h+1)}(\mathbf{x}) \in \mathbb{R}^{d_h}, & h = 1,\ldots,L, \end{cases} \tag{8}$$

and

$$\widetilde{\mathbf{D}}^{(h)}(\mathbf{x}) = \operatorname{diag}\left(\dot\sigma\left(\widetilde{\mathbf{f}}^{(h)}(\mathbf{x})\right)\right) \in \mathbb{R}^{d_h \times d_h}, \quad h = 1,\ldots,L. \tag{9}$$

Note that since the weights being pruned are staying at zero always during the training process, the gradient of the pruned network is simply the masked gradient of the unpruned network.

Now, we have

$$\left\langle \frac{\partial \widetilde{f}(\mathbf{x})}{\partial \mathbf{W}^{(h)}}, \frac{\partial \widetilde{f}(\mathbf{x}')}{\partial \mathbf{W}^{(h)}} \right\rangle = \left\langle \left(\widetilde{\mathbf{b}}^{(h)}(\mathbf{x})\left(\widetilde{\mathbf{g}}^{(h-1)}(\mathbf{x})\right)^\top\right) \odot \mathbf{m}^{(h)}, \left(\widetilde{\mathbf{b}}^{(h)}(\mathbf{x}')\left(\widetilde{\mathbf{g}}^{(h-1)}(\mathbf{x}')\right)^\top\right) \odot \mathbf{m}^{(h)} \right\rangle.$$

Now we write

$$\left(\widetilde{\mathbf{b}}^{(h)}(\mathbf{x})\left(\widetilde{\mathbf{g}}^{(h-1)}(\mathbf{x})\right)^\top\right) \odot \mathbf{m}^{(h)} = \begin{bmatrix} \widetilde{\mathbf{b}}_1^{(h)}(\mathbf{x})\widetilde{\mathbf{g}}^{(h-1)}(\mathbf{x}) \odot \mathbf{m}_1^{(h)} \\ \widetilde{\mathbf{b}}_2^{(h)}(\mathbf{x})\widetilde{\mathbf{g}}^{(h-1)}(\mathbf{x}) \odot \mathbf{m}_2^{(h)} \\ \vdots \\ \widetilde{\mathbf{b}}_{d_h}^{(h)}(\mathbf{x})\widetilde{\mathbf{g}}^{(h-1)}(\mathbf{x}) \odot \mathbf{m}_{d_h}^{(h)} \end{bmatrix}.$$

Thus,

$$\begin{aligned} \left\langle \frac{\partial \widetilde{f}(\mathbf{x})}{\partial \mathbf{W}^{(h)}}, \frac{\partial \widetilde{f}(\mathbf{x}')}{\partial \mathbf{W}^{(h)}} \right\rangle &= \left\langle \left(\widetilde{\mathbf{b}}^{(h)}(\mathbf{x})\left(\widetilde{\mathbf{g}}^{(h-1)}(\mathbf{x})\right)^\top\right) \odot \mathbf{m}^{(h)}, \left(\widetilde{\mathbf{b}}^{(h)}(\mathbf{x}')\left(\widetilde{\mathbf{g}}^{(h-1)}(\mathbf{x}')\right)^\top\right) \odot \mathbf{m}^{(h)} \right\rangle \\ &= \sum_{i=1}^{d_h} \widetilde{\mathbf{b}}_i^{(h)}(\mathbf{x})\widetilde{\mathbf{b}}_i^{(h)}(\mathbf{x}') \left\langle \widetilde{\mathbf{g}}^{(h-1)}(\mathbf{x}) \odot \mathbf{m}_i^{(h)}, \widetilde{\mathbf{g}}^{(h-1)}(\mathbf{x}') \odot \mathbf{m}_i^{(h)} \right\rangle \\ &= \left(\widetilde{\mathbf{b}}^{(h)}(\mathbf{x})\right)^\top \mathbf{G}^{(h-1)} \widetilde{\mathbf{b}}^{(h)}(\mathbf{x}'), \end{aligned} \tag{10}$$

where we define $\mathbf{G}^{(h-1)}$ as a diagonal matrix and $\mathbf{G}_{ii}^{(h-1)} = \left\langle \widetilde{\mathbf{g}}^{(h-1)}(\mathbf{x}) \odot \mathbf{m}_i^{(h)}, \widetilde{\mathbf{g}}^{(h-1)}(\mathbf{x}') \odot \mathbf{m}_i^{(h)} \right\rangle$. Observe that

$$\lim_{d_{h-1}\to\infty} \left\langle \widetilde{\mathbf{g}}^{(h-1)}(\mathbf{x}) \odot \mathbf{m}_i^{(h)}, \widetilde{\mathbf{g}}^{(h-1)}(\mathbf{x}') \odot \mathbf{m}_i^{(h)} \right\rangle = \lim_{d_{h-1}\to\infty} \frac{c_\sigma}{d_{h-1}} \sum_{j=1}^{d_{h-1}} \sigma\left(\widetilde{\mathbf{f}}_j^{(h-1)}(\mathbf{x})\right) \sigma\left(\widetilde{\mathbf{f}}_j^{(h-1)}(\mathbf{x}')\right) \left(\mathbf{m}_{ij}^{(h)}\right)^2$$

$$
\begin{aligned}
&= \mathbb{E}\left[c_\sigma \sigma\left(\widetilde{\mathbf{f}}_j^{(h-1)}(\mathbf{x})\right)\sigma\left(\widetilde{\mathbf{f}}_j^{(h-1)}(\mathbf{x}')\right)\left(\mathbf{m}_{ij}^{(h)}\right)^2\right] \\
&= \mathbb{E}\left[c_\sigma \sigma\left(\widetilde{\mathbf{f}}_j^{(h-1)}(\mathbf{x})\right)\sigma\left(\widetilde{\mathbf{f}}_j^{(h-1)}(\mathbf{x}')\right)\right]\mathbb{E}\left[\left(\mathbf{m}_{ij}^{(h)}\right)^2\right] \\
&= \alpha\,\mathbb{E}\left[c_\sigma \sigma\left(\widetilde{\mathbf{f}}_j^{(h-1)}(\mathbf{x})\right)\sigma\left(\widetilde{\mathbf{f}}_j^{(h-1)}(\mathbf{x}')\right)\right].
\end{aligned}
$$

This requires us to analyze $\widetilde{\mathbf{f}}^{(h)}(\mathbf{x})$ for $h \in [L]$.

We now analyze the forward dynamics of the pruned neural network:

$$
\begin{aligned}
[\widetilde{\mathbf{f}}^{(h+1)}(\mathbf{x})]_i &= \sum_{j=1}^{d_h}[\mathbf{W}^{(h+1)} \odot \mathbf{m}^{(h+1)}]_{ij}[\widetilde{\mathbf{g}}^{(h)}(\mathbf{x})]_j \\
&= \sqrt{\frac{c_\sigma}{d_h}}\sum_{j=1}^{d_h}[\mathbf{W}^{(h+1)} \odot \mathbf{m}^{(h+1)}]_{ij}\sigma\left(\left[\widetilde{\mathbf{f}}^{(h)}(\mathbf{x})\right]_j\right).
\end{aligned}
$$

Conditioned on $\mathbf{g}^{(h-1)}(\mathbf{x}), \mathbf{g}^{(h-1)}(\mathbf{x}')$, we have $\widetilde{\mathbf{f}}_j^{(h)}(\mathbf{x}), \widetilde{\mathbf{f}}_j^{(h)}(\mathbf{x}')$ are i.i.d. random variables for all $j \in [n]$. However, for $h \in \{1, \ldots, L\}$, as $d_h \to \infty$, by the central limit theorem, $[\widetilde{\mathbf{f}}^{(h+1)}(\mathbf{x})]_i$ converges to a Gaussian random variable. This is certainly not true for the output in the first layer because the input dimension can't go to infinity. Thus, we make assumption that the pruning only starts from the second layer.

Now by i.i.d assumption of the mask and weights, we can compute the covariance of pre-activation as

$$
\begin{aligned}
\mathop{\mathbb{E}}_{\mathbf{W}^{(h+1)}}\left[\left[\widetilde{\mathbf{f}}^{(h+1)}(\mathbf{x})\right]_i\left[\widetilde{\mathbf{f}}^{(h+1)}(\mathbf{x}')\right]_i\middle|\widetilde{\mathbf{f}}^{(h)}, \mathbf{m}^{(h+1)}\right] &= \left\langle \widetilde{\mathbf{g}}^{(h)}(\mathbf{x}) \odot \mathbf{m}_i^{(h+1)}, \widetilde{\mathbf{g}}^{(h)}(\mathbf{x}') \odot \mathbf{m}_i^{(h+1)}\right\rangle \\
&= \frac{c_\sigma}{d_h}\sum_{j=1}^{d_h}\sigma\left(\left[\widetilde{\mathbf{f}}^{(h)}(\mathbf{x})\right]_j\right)\sigma\left(\left[\widetilde{\mathbf{f}}^{(h)}(\mathbf{x}')\right]_j\right)\left(\mathbf{m}_{ij}^{(h+1)}\right)^2 \qquad (11) \\
&\xrightarrow{d_h\to\infty} \alpha c_\sigma\,\mathbb{E}\left[\sigma\left(\left[\widetilde{\mathbf{f}}^{(h)}(\mathbf{x})\right]_j\right)\sigma\left(\left[\widetilde{\mathbf{f}}^{(h)}(\mathbf{x}')\right]_j\right)\right],
\end{aligned}
$$

by the law of large number.

Recall Definition 4.1, we define

$$
\widetilde{\Sigma}^{(h)}(\mathbf{x}, \mathbf{x}') := \lim_{d_1,\ldots,d_h\to\infty}\left\langle \widetilde{\mathbf{g}}^{(h)}(\mathbf{x}), \widetilde{\mathbf{g}}^{(h)}(\mathbf{x}')\right\rangle = \lim_{d_1,\ldots,d_h\to\infty}\frac{c_\sigma}{d_h}\sum_{j=1}^{d_h}\sigma\left(\left[\widetilde{\mathbf{f}}^{(h)}(\mathbf{x})\right]_j\right)\sigma\left(\left[\widetilde{\mathbf{f}}^{(h)}(\mathbf{x}')\right]_j\right).
$$

where the limit is taking sequentially from d_1 to d_h. We further define

$$
\widetilde{\mathbf{\Lambda}}^{(1)} = \begin{bmatrix}\widetilde{\Sigma}^{(0)}(\mathbf{x}, \mathbf{x}) & \widetilde{\Sigma}^{(0)}(\mathbf{x}, \mathbf{x}') \\ \widetilde{\Sigma}^{(0)}(\mathbf{x}', \mathbf{x}) & \widetilde{\Sigma}^{(0)}(\mathbf{x}, \mathbf{x})\end{bmatrix},
$$

$$
\widetilde{\mathbf{\Lambda}}^{(h)} = \alpha\begin{bmatrix}\widetilde{\Sigma}^{(h-1)}(\mathbf{x}, \mathbf{x}) & \widetilde{\Sigma}^{(h-1)}(\mathbf{x}, \mathbf{x}') \\ \widetilde{\Sigma}^{(h-1)}(\mathbf{x}', \mathbf{x}) & \widetilde{\Sigma}^{(h-1)}(\mathbf{x}, \mathbf{x})\end{bmatrix},
$$

Lemma 8.2 (Restatement of Lemma 4.2)**.** *Suppose the neural network uses ReLU as its activation and $d_1, d_2, \ldots, d_L \to \infty$ sequentially, then*

$$
\begin{aligned}
\widetilde{\Sigma}^{(h)}(\mathbf{x}, \mathbf{x}') &= c_\sigma \mathop{\mathbb{E}}_{(u,v)\sim\mathcal{N}(\mathbf{0}, \widetilde{\mathbf{\Lambda}}^{(h)})}[\sigma(u)\sigma(v)], \\
\widetilde{\Sigma}^{(h)}(\mathbf{x}, \mathbf{x}') &= \alpha^{h-1}\Sigma^{(h)}(\mathbf{x}, \mathbf{x}'),
\end{aligned}
$$

for $h = 1, 2, \ldots, L$.

Proof. We prove by induction. First, notice that $\widetilde{\Sigma}^{(0)}(\mathbf{x},\mathbf{x}') = \Sigma^{(0)}(\mathbf{x},\mathbf{x}')$. When $h=1$, there is noting to prove. Now, assume the induction hypothesis holds for all h such that $h \leq t$ where $t \geq 1$ and we want to show that $\widetilde{\Sigma}^{(t+1)}(\mathbf{x},\mathbf{x}') = \alpha^t \Sigma^{(t+1)}(\mathbf{x},\mathbf{x}')$. Notice that Equation 11 is true for $h \in \{1,\dots,L\}$. Therefore, as $d_t \to \infty$

$$\widetilde{\Sigma}^{(t+1)}(\mathbf{x},\mathbf{x}') = c_\sigma \mathop{\mathbb{E}}_{\left[\widetilde{\mathbf{f}}^{(t+1)}(\mathbf{x})\right]_1, \left[\widetilde{\mathbf{f}}^{(t+1)}(\mathbf{x}')\right]_1} \left[\sigma\left(\left[\widetilde{\mathbf{f}}^{(t+1)}(\mathbf{x})\right]_1\right)\sigma\left(\left[\widetilde{\mathbf{f}}^{(t+1)}(\mathbf{x}')\right]_1\right)\right].$$

Assume all the previous layers are already at the limit, for $t=1,\dots,L$,

$$\left(\left[\widetilde{\mathbf{f}}^{(t+1)}(\mathbf{x})\right]_1, \left[\widetilde{\mathbf{f}}^{(t+1)}(\mathbf{x}')\right]_1\right) \sim \mathcal{N}\left(\mathbf{0}, \alpha \begin{bmatrix} \widetilde{\Sigma}^{(t)}(\mathbf{x},\mathbf{x}) & \widetilde{\Sigma}^{(t)}(\mathbf{x},\mathbf{x}') \\ \widetilde{\Sigma}^{(t)}(\mathbf{x}',\mathbf{x}) & \widetilde{\Sigma}^{(t)}(\mathbf{x},\mathbf{x}) \end{bmatrix}\right) = \mathcal{N}(\mathbf{0}, \widetilde{\mathbf{\Lambda}}^{(t+1)}).$$

This proves the first equality.

By induction hypothesis on $\widetilde{\Sigma}^{(t)}(\mathbf{x},\mathbf{x}')$, we have $\widetilde{\mathbf{\Lambda}}^{(t+1)} = \alpha \cdot \alpha^{t-1}\mathbf{\Lambda}^{(t+1)}$. Hence

$$\begin{aligned}
\widetilde{\Sigma}^{(t+1)}(\mathbf{x},\mathbf{x}') &= c_\sigma \mathop{\mathbb{E}}_{(u,v)\sim\mathcal{N}(\mathbf{0},\alpha^t\mathbf{\Lambda}^{(t+1)})}[\sigma(u)\sigma(v)] \\
&= c_\sigma \mathop{\mathbb{E}}_{(u',v')\sim\mathcal{N}(\mathbf{0},\mathbf{\Lambda}^{(t+1)})}[\sigma(\alpha^{\frac{t}{2}}u')\sigma(\alpha^{\frac{t}{2}}v')] \\
&= \alpha^t c_\sigma \mathop{\mathbb{E}}_{(u',v')\sim\mathcal{N}(\mathbf{0},\mathbf{\Lambda}^{(t+1)})}[\sigma(u')\sigma(v')] \\
&= \alpha^t \Sigma^{(t+1)}(\mathbf{x},\mathbf{x}'),
\end{aligned}$$

where the second last inequality is from our assumption that the activation is ReLU. □

This lemma implies that

$$\widetilde{\Sigma}^{(h)}(\mathbf{x},\mathbf{x}') = \lim_{d_1,\dots,d_h\to\infty} \left\langle \widetilde{\mathbf{g}}^{(h)}(\mathbf{x}), \widetilde{\mathbf{g}}^{(h)}(\mathbf{x}')\right\rangle = \alpha^{h-1}\Sigma^{(h)}(\mathbf{x},\mathbf{x}'). \tag{12}$$

Thus, combining Equation (10) and Equation (12) we have

$$\begin{aligned}
&\left\langle \left(\widetilde{\mathbf{b}}^{(h)}(\mathbf{x})\left(\widetilde{\mathbf{g}}^{(h-1)}(\mathbf{x})\right)^\top\right)\odot\mathbf{m}^{(h)}, \left(\widetilde{\mathbf{b}}^{(h)}(\mathbf{x}')\left(\widetilde{\mathbf{g}}^{(h-1)}(\mathbf{x}')\right)^\top\right)\odot\mathbf{m}^{(h)}\right\rangle \\
&= \left(\widetilde{\mathbf{b}}^{(h)}(\mathbf{x})\right)^\top \mathbf{G}^{(h-1)}\widetilde{\mathbf{b}}^{(h)}(\mathbf{x}') \\
&\xrightarrow{d_1,\dots,d_{h-1}\to\infty} \alpha^{h-1}\Sigma^{(h-1)}(\mathbf{x},\mathbf{x}') \lim_{d_1,\dots,d_{h-1}\to\infty} \left(\widetilde{\mathbf{b}}^{(h)}(\mathbf{x})\right)^\top \widetilde{\mathbf{b}}^{(h)}(\mathbf{x}').
\end{aligned} \tag{13}$$

Lemma 8.3 (Restatement of Lemma 4.3)**.** *Assume we use a fresh sample of weights in the backward pass, then*

$$\lim_{d_1,\dots,d_L\to\infty} \left\langle \widetilde{\mathbf{b}}^{(h)}(\mathbf{x}), \widetilde{\mathbf{b}}^{(h)}(\mathbf{x}')\right\rangle = \alpha^{L+1-h}\prod_{h'=h}^{L}\dot{\Sigma}^{(h')}(\mathbf{x},\mathbf{x}'). \tag{14}$$

Proof. For the factor $\left\langle \widetilde{\mathbf{b}}^{(h)}(\mathbf{x}), \widetilde{\mathbf{b}}^{(h)}(\mathbf{x}')\right\rangle$, we expand using the definition of $\widetilde{\mathbf{b}}^{(h)}(\mathbf{x})$

$$\begin{aligned}
&\left\langle \widetilde{\mathbf{b}}^{(h)}(\mathbf{x}), \widetilde{\mathbf{b}}^{(h)}(\mathbf{x}')\right\rangle \\
&= \left\langle \sqrt{\frac{c_\sigma}{d_h}}\widetilde{\mathbf{D}}^{(h)}(\mathbf{x})\left(\mathbf{W}^{(h+1)}\odot\mathbf{m}^{(h+1)}\right)^\top \widetilde{\mathbf{b}}^{(h+1)}(\mathbf{x}), \sqrt{\frac{c_\sigma}{d_h}}\widetilde{\mathbf{D}}^{(h)}(\mathbf{x}')\left(\mathbf{W}^{(h+1)}\odot\mathbf{m}^{(h+1)}\right)^\top \widetilde{\mathbf{b}}^{(h+1)}(\mathbf{x}')\right\rangle
\end{aligned}$$

First we analyze $\widetilde{\mathbf{D}}^{(h)}(\mathbf{x})$. Since we use ReLU as the activation function, $\dot{\sigma}(x) = \mathbb{I}(x>0)$ and in particular, $\dot{\sigma}(cx) = \mathbb{I}(cx>0) = \mathbb{I}(x>0) = \dot{\sigma}(x)$ for any positive constant c. By Lemma 8.2, we show that under sequential limit, $\widetilde{\mathbf{f}}^{(h)}(\mathbf{x})$ has the same distribution as $\alpha^{h-1}\mathbf{f}^{(h)}(\mathbf{x})$ which implies $\widetilde{\mathbf{D}}^{(h)}(\mathbf{x})$ has the same distribution as $\mathbf{D}^{(h)}(\mathbf{x})$.

Observe that $\mathbf{W}^{(h+1)} \odot \mathbf{m}^{(h+1)}$ and $\widetilde{\mathbf{b}}^{(h+1)}(\mathbf{x})$ are dependent. Now we apply the independent copy trick which is rigorously justified for ReLU network with Gaussian weights by replacing $\mathbf{W}^{(h+1)}$ with a fresh new sample $\widetilde{\mathbf{W}}^{(h+1)}$.

$$
\begin{aligned}
&\left\langle \widetilde{\mathbf{b}}^{(h)}(\mathbf{x}), \widetilde{\mathbf{b}}^{(h)}(\mathbf{x}') \right\rangle \\
&= \left\langle \sqrt{\frac{c_\sigma}{d_h}} \widetilde{\mathbf{D}}^{(h)}(\mathbf{x}) \left(\mathbf{W}^{(h+1)} \odot \mathbf{m}^{(h+1)} \right)^\top \widetilde{\mathbf{b}}^{(h+1)}(\mathbf{x}), \sqrt{\frac{c_\sigma}{d_h}} \widetilde{\mathbf{D}}^{(h)}(\mathbf{x}') \left(\mathbf{W}^{(h+1)} \odot \mathbf{m}^{(h+1)} \right)^\top \widetilde{\mathbf{b}}^{(h+1)}(\mathbf{x}') \right\rangle \\
&\approx \left\langle \sqrt{\frac{c_\sigma}{d_h}} \widetilde{\mathbf{D}}^{(h)}(\mathbf{x}) \left(\widetilde{\mathbf{W}}^{(h+1)} \odot \mathbf{m}^{(h+1)} \right)^\top \widetilde{\mathbf{b}}^{(h+1)}(\mathbf{x}), \sqrt{\frac{c_\sigma}{d_h}} \widetilde{\mathbf{D}}^{(h)}(\mathbf{x}') \left(\widetilde{\mathbf{W}}^{(h+1)} \odot \mathbf{m}^{(h+1)} \right)^\top \widetilde{\mathbf{b}}^{(h+1)}(\mathbf{x}') \right\rangle \\
&\xrightarrow{d_1,\ldots,d_h\to\infty} \alpha \frac{c_\sigma}{d_h} \mathrm{Tr}\left(\widetilde{\mathbf{D}}^{(h)}(\mathbf{x}) \widetilde{\mathbf{D}}^{(h)}(\mathbf{x}') \right) \lim_{d_1,\ldots,d_h\to\infty} \left\langle \widetilde{\mathbf{b}}^{(h+1)}(\mathbf{x}), \widetilde{\mathbf{b}}^{(h+1)}(\mathbf{x}') \right\rangle \\
&\xrightarrow{d_1,\ldots,d_h\to\infty} \alpha \dot{\Sigma}^{(h)}(\mathbf{x}, \mathbf{x}') \lim_{d_1,\ldots,d_h\to\infty} \left\langle \widetilde{\mathbf{b}}^{(h+1)}(\mathbf{x}), \widetilde{\mathbf{b}}^{(h+1)}(\mathbf{x}') \right\rangle . \qquad (15)
\end{aligned}
$$

where we justify the limit as the following: first let $\mathbf{D}$ short for $\widetilde{\mathbf{D}}^{(h)}(\mathbf{x}) \widetilde{\mathbf{D}}^{(h)}(\mathbf{x}')$

$$
\left(\frac{c_\sigma}{d_h} (\widetilde{\mathbf{W}}^{(h+1)} \odot \mathbf{m}^{(h+1)}) \mathbf{D} (\widetilde{\mathbf{W}}^{(h+1)} \odot \mathbf{m}^{(h+1)})^\top \right)_{ij} = \frac{c_\sigma}{d_h} \sum_k \mathbf{D}_{kk} \widetilde{\mathbf{W}}_{ik}^{(h+1)} \mathbf{m}_{ik}^{(h+1)} \widetilde{\mathbf{W}}_{jk}^{(h+1)} \mathbf{m}_{jk}^{(h+1)},
$$

which converges to a diagonal matrix as $d_h \to \infty$. Thus, the inner product is given by

$$
\begin{aligned}
&\frac{c_\sigma}{d_h} \sum_{i,j} \widetilde{\mathbf{b}}_i^{(h+1)}(\mathbf{x}) \widetilde{\mathbf{b}}_j^{(h+1)}(\mathbf{x}') \sum_k \mathbf{D}_{kk} \widetilde{\mathbf{W}}_{ik}^{(h+1)} \mathbf{m}_{ik}^{(h+1)} \widetilde{\mathbf{W}}_{jk}^{(h+1)} \mathbf{m}_{jk}^{(h+1)} \\
&= \frac{c_\sigma}{d_h} \sum_{i,j} \widetilde{\mathbf{b}}_i^{(h+1)}(\mathbf{x}) \widetilde{\mathbf{b}}_j^{(h+1)}(\mathbf{x}') (\widetilde{\mathbf{w}}_i^{(h+1)} \odot \mathbf{m}_i^{(h+1)})^\top \mathbf{D} (\widetilde{\mathbf{w}}_j^{(h+1)} \odot \mathbf{m}_j^{(h+1)}) \\
&= \frac{c_\sigma}{d_h} \sum_{i,j} \widetilde{\mathbf{b}}_i^{(h+1)}(\mathbf{x}) \widetilde{\mathbf{b}}_j^{(h+1)}(\mathbf{x}') \left(\widetilde{\mathbf{w}}_i^{(h+1)} \right)^\top \mathbf{M}_i^{(h+1)} \mathbf{D} \mathbf{M}_j^{(h+1)} \widetilde{\mathbf{w}}_j^{(h+1)} \\
&\xrightarrow{d_1,\ldots,d_h\to\infty} \frac{c_\sigma}{d_h} \sum_i \widetilde{\mathbf{b}}_i^{(h+1)}(\mathbf{x}) \widetilde{\mathbf{b}}_i^{(h+1)}(\mathbf{x}') \mathrm{Tr}(\mathbf{M}_i^{(h+1)} \mathbf{D} \mathbf{M}_i^{(h+1)}) \\
&\xrightarrow{d_1,\ldots,d_h\to\infty} \alpha \dot{\Sigma}^{(h)}(\mathbf{x}, \mathbf{x}') \lim_{d_1,\ldots,d_h\to\infty} \left\langle \widetilde{\mathbf{b}}^{(h+1)}(\mathbf{x}), \widetilde{\mathbf{b}}^{(h+1)}(\mathbf{x}') \right\rangle ,
\end{aligned}
$$

where $\mathbf{M}_i = \mathrm{diag}(\mathbf{m}_i)$ and $\widetilde{\mathbf{w}}_i$ is the i-th row of $\widetilde{\mathbf{W}}$ and $\lim_{d_h\to\infty} \frac{c_\sigma}{d_h} \mathrm{Tr}(\mathbf{M}_i^{(h+1)} \mathbf{D} \mathbf{M}_i^{(h+1)}) = \alpha \dot{\Sigma}^{(h)}(\mathbf{x}, \mathbf{x}')$. Now, we can unroll the formula of $\left\langle \widetilde{\mathbf{b}}^{(h)}(\mathbf{x}), \widetilde{\mathbf{b}}^{(h)}(\mathbf{x}') \right\rangle$ in Equation (15), we have

$$
\lim_{d_1,\ldots,d_L\to\infty} \left\langle \widetilde{\mathbf{b}}^{(h)}(\mathbf{x}), \widetilde{\mathbf{b}}^{(h)}(\mathbf{x}') \right\rangle = \alpha^{L+1-h} \prod_{h'=h}^{L} \dot{\Sigma}^{(h')}(\mathbf{x}, \mathbf{x}').
$$

□

Proof of Theorem 8.1. Combining the result in Equation (13) and Equation (14), we have

$$
\begin{aligned}
\left\langle \frac{\partial \widetilde{f}(\mathbf{x})}{\partial \mathbf{W}^{(h)}}, \frac{\partial \widetilde{f}(\mathbf{x}')}{\partial \mathbf{W}^{(h)}} \right\rangle &= \left\langle \left(\widetilde{\mathbf{b}}^{(h)}(\mathbf{x}) \left(\widetilde{\mathbf{g}}^{(h-1)}(\mathbf{x}) \right)^\top \right) \odot \mathbf{m}^{(h)}, \left(\widetilde{\mathbf{b}}^{(h)}(\mathbf{x}') \left(\widetilde{\mathbf{g}}^{(h-1)}(\mathbf{x}') \right)^\top \right) \odot \mathbf{m}^{(h)} \right\rangle \\
&\xrightarrow{d_1,\ldots,d_L\to\infty} \alpha^L \Sigma^{(h-1)}(\mathbf{x}, \mathbf{x}') \prod_{h'=h}^{L+1} \dot{\Sigma}^{(h')}(\mathbf{x}, \mathbf{x}').
\end{aligned}
$$

We conclude

$$
\widetilde{\mathbf{\Theta}}_\infty(\mathbf{x}, \mathbf{x}') := \lim_{d_1,d_2,\ldots,d_L\to\infty} \widetilde{\mathbf{\Theta}}(\mathbf{x}, \mathbf{x}') = \alpha^L \mathbf{\Theta}_\infty(\mathbf{x}, \mathbf{x}'), \qquad (16)
$$

which proves Theorem 8.1. □

8.1 Proof of Lemma 3.4: Going from Asymptotic Regime to Non-Asymptotic Regime

Before we give proof for our non-asymptotic result, we note that our asymptotic result is obtained from taking sequential limits of all the hidden layers which is a somewhat a limited notion of limits since we assume all the layer before is already at the limit when we deal with a given layer. Non-asymptotic analysis, on the other hand, consider using a large but finite amount of samples to get close to (but not exactly at) the limit. Thus, we need to justify that the networks are indeed able to approach by increasing width. In mathematical language, this is the same as justifying taking the limit outside of $\mathbb{E}\,\sigma(\cdot), \mathbb{E}\,\dot{\sigma}(\cdot)$.

We invoke several results from measure-theoretic probability theory.

Definition 8.4 (Uniformly integrable)**.** *A sequence of random variables $\{X_n\}$ is called* uniformly integrable *if*

$$\lim_{a\to\infty} \sup_n \mathbb{E}[|X_n|\mathbb{I}(|X_n| \geq a)] \to 0.$$

Lemma 8.5 (Theorem 3, Chapter 7.10 in (Grimmett and Stirzaker, 2020))**.** *Suppose that $\{X_n\}$ is a sequence of random variables satisfying $X_n \to X$ in probability. The following statements are equivalent:*

1. *The family $\{X_n\}$ is uniformly integrable.*
2. *$\mathbb{E}\,|X_n| < \infty$ for all n and $\mathbb{E}\,|X_n| \to \mathbb{E}\,|X| < \infty$.*

Theorem 8.6 (Skorokhod's Representation Theorem, (Billingsley, 1999))**.** *Let $\{\mu_n\}$ be a sequence of probability measure defined on a metric space S such that μ_n converges weakly to some probability measure μ_∞ on S as $n \to \infty$. Suppose that the support of μ_∞ is separable. Then there exists S-valued random variables X_n defined on a common probability space $(\Omega, \mathcal{F}, \mathbb{P})$ such that the law of X_n is is μ_n for all n (including $n = \infty$) and such that $(X_n)_{n\in\mathbb{N}}$ converges to X_∞, $\mathbb{P}$-almost surely.*

Theorem 8.7 (Continuous Mapping Theorem (Mann and Wald, 1943))**.** *Let $\{X_n\}, X$ be random variables defined on a metric space S. Suppose a function $g : S \to S'$ (where S' is another metric space) has the set of discontinuities of measure zero. Then*

$$X_n \xrightarrow{\mathcal{D}} X \quad \Rightarrow \quad g(X_n) \xrightarrow{\mathcal{D}} g(X),$$

where $\xrightarrow{\mathcal{D}}$ represents convergence in distribution.

Lemma 8.8 (Restatement of Lemma 3.4)**.** *Conditioned on $\mathbf{g}^{(h-1)}(\mathbf{x}), \mathbf{g}^{(h-1)}(\mathbf{x}')$. Fix $i \in [d_{h+1}]$. Let*

$$X_n = \begin{bmatrix} \sqrt{\frac{c_\sigma}{n}} \sum_{j=1}^n \mathbf{W}_{ij}^{(h+1)} \mathbf{m}_{ij}^{(h+1)} \sigma(\widetilde{\mathbf{f}}_j^{(h)}(\mathbf{x})) \\ \sqrt{\frac{c_\sigma}{n}} \sum_{j=1}^n \mathbf{W}_{ij}^{(h+1)} \mathbf{m}_{ij}^{(h+1)} \sigma(\widetilde{\mathbf{f}}_j^{(h)}(\mathbf{x}')) \end{bmatrix} \in \mathbb{R}^2,$$

and define let $g : \mathbb{R}^2 \to \mathbb{R}$ to be $g(x, y) \in \{\sigma(x)\sigma(y), \dot{\sigma}(x)\dot{\sigma}(y)\}$. Then,

$$\lim_{n\to\infty} \mathbb{E}[g(X_n)] = \mathbb{E}[g(\lim_{n\to\infty} X_n)].$$

Proof. First of all, conditioned on $\mathbf{g}^{(h-1)}(\mathbf{x}), \mathbf{g}^{(h-1)}(\mathbf{x}')$, we have $\widetilde{\mathbf{f}}_j^{(h)}(\mathbf{x}), \widetilde{\mathbf{f}}_j^{(h)}(\mathbf{x}')$ are i.i.d. random variables for all $j \in [n]$.

We first prove the exchange of limit for $g(x, y) = \sigma(x)\sigma(y)$ since this function is continuous. By the Central Limit Theorem, $X_n \xrightarrow{\mathcal{D}} X_\infty \sim \mathcal{N}(\mathbf{0}, \widetilde{\mathbf{\Lambda}}^{(h+1)})$. By the Continuous Mapping Theorem, $g(X_n) \xrightarrow{\mathcal{D}} g(X_\infty)$. Then by the Skorokhod's Representation Theorem in Theorem 8.6, there exists another sequence $\{X_n'\}$ and X_∞' such that $g(X_n) \overset{\mathcal{D}}{=} X_n'$ and $g(X_\infty) \overset{\mathcal{D}}{=} X_\infty'$ and $X_n' \xrightarrow{a.s.} X_\infty'$. Now we use the fact that the sequence $\{X_n'\}$ is uniformly integrable (see Definition 8.4). By Lemma 8.5, this implies convergence in L^1 (and notice that $g(x, y)$ only outputs non-negative values)

$$\lim_{n\to\infty} \mathbb{E}[X_n'] = \mathbb{E}[X_\infty'].$$

Since

$$\mathbb{E}[g(X_n)] = \mathbb{E}[X_n'],$$

$$\mathbb{E}[g(X_\infty)] = \mathbb{E}[X'_\infty],$$

we have

$$\lim_{n\to\infty} \mathbb{E}[g(X_n)] = \mathbb{E}[g(X_\infty)] = \mathbb{E}[g(\mathcal{N}(\mathbf{0}, \widetilde{\mathbf{\Lambda}}^{(h+1)}))].$$

Now we prove the result for $g(x,y) = \dot{\sigma}(x)\dot{\sigma(y)} = \mathbb{I}(x \geq 0, y \geq 0)$. Again, apply Skorokhod's Representation Theorem, there exists a sequence of random variables $\{X''_n\}$ and another random variable X''_∞ such that $X_n \overset{\mathcal{D}}{=} X''_n$ and $X_\infty \overset{\mathcal{D}}{=} X''_\infty$ and $X''_n \xrightarrow{a.s.} X''_\infty$. Since convergence almost surely implies convergence in probability, we have

$$\lim_{n\to\infty} \mathbb{E}[g(X''_n)] = \mathbb{E}[g(X''_\infty)],$$

which implies

$$\lim_{n\to\infty} \mathbb{E}[g(X_n)] = \mathbb{E}[g(X_\infty)].$$

□

9 NON-ASYMPTOTIC ANALYSIS (Proof of Theorem 3.5)

9.1 Probability

Theorem 9.1 (Multiplicative Chernoff Bound)**.** *If $X_1, X_2, \ldots, X_m$ are i.i.d. Bernoulli random variables with probability p, then*

$$\mathbb{P}\left[\left|\sum_{i=1}^m X_i - pm\right| \geq \epsilon pm\right] \leq 2\exp(-\min(\epsilon^2, \epsilon)pm).$$

Theorem 9.2. *Assume $X_1, \ldots, X_m$ are i.i.d. Sub-Gaussian random variables with variance proxy σ^2 and $Y_1, \ldots, Y_m$ are i.i.d. Bernoulli random variables with probability p. For $\epsilon \in (0, 1/2),\ t > 0$,*

$$\mathbb{P}\left[\left|\frac{1}{pm}\sum_{i=1}^m X_i Y_i - \mathbb{E}[X]\right| \geq \epsilon(|\,\mathbb{E}[X]| + t) + t\right] \leq 2\exp(-(1-\epsilon)pmt^2/(2\sigma^2)) + 2\exp(-\min(\epsilon^2, \epsilon)pm).$$

Proof. Let $\widehat{p} = \frac{\sum_{i=1}^m Y_i}{m}$. By the concentration of Sub-Gaussian random variable with variance proxy σ^2, we have

$$\mathbb{P}\left[\left|\frac{1}{\widehat{p}m}\sum_{i=1}^m X_i Y_i - \mathbb{E}[X]\right| \geq t\right] \leq 2\exp(-\widehat{p}mt^2/(2\sigma^2)) + 2\exp(-\min(\epsilon^2, \epsilon)pm).$$

By Theorem 9.1, we have with probability at least $1 - 2\exp(-\min(\epsilon^2, \epsilon)pm)$, $\widehat{p} = (1 \pm \epsilon)p$. Thus, with probability at least $1 - 2\exp(-\widehat{p}mt^2/(2\sigma^2)) - 2\exp(-\min(\epsilon^2, \epsilon)pm)$,

$$\frac{1}{pm}\sum_{i=1}^m X_i Y_i = \frac{\widehat{p}}{p}\frac{1}{\widehat{p}m}\sum_{i=1}^m X_i Y_i = (1 \pm \epsilon)(\mathbb{E}[X] \pm t).$$

□

Theorem 9.3. *Assume $X_1, \ldots, X_m$ are i.i.d. Sub-Gamma random variables with parameters (σ^2, c) and $Y_1, \ldots, Y_m$ are i.i.d. Bernoulli random variables with probability p. For $\epsilon \in (0, 1/2),\ t > 0$,*

$$\mathbb{P}\left[\left|\frac{1}{pm}\sum_{i=1}^m X_i Y_i - \mathbb{E}[X]\right| \geq \epsilon(|\,\mathbb{E}[X]| + t) + t\right] \leq 2\exp(-(1-\epsilon)pm\min(t^2/(2\sigma^2), t/c)) + 2\exp(-\min(\epsilon^2, \epsilon)pm).$$

Proof. By the concentration of Sub-Gamma random variables, we have

$$\mathbb{P}\left[\left|\frac{1}{\widehat{p}m}\sum_{i=1}^{m}X_iY_i-\mathbb{E}[X]\right|\geq t\right]\leq 2\exp(-\widehat{p}m\min(t^2/(2\sigma^2),t/c)).$$

The rest of proof follows from the proof of Theorem 9.2. □

Lemma 9.4 (Gaussian Chaos of Order 2 (Boucheron et al., 2013)). *Let $\boldsymbol{\xi}\sim\mathcal{N}(\mathbf{0},\mathbf{I}_n)$ be an n-dimensional unit Gaussian random vector, $\mathbf{A}\in\mathbb{R}^{n\times n}$ be a symmetric matrix, then for any $t>0$,*

$$\mathbb{P}\left[|\boldsymbol{\xi}^\top\mathbf{A}\boldsymbol{\xi}-\mathbb{E}[\boldsymbol{\xi}^\top\mathbf{A}\boldsymbol{\xi}]|>2\left\|\mathbf{A}\right\|_F\sqrt{t}+2\left\|\mathbf{A}\right\|_2 t\right]\leq 2\exp(-t).$$

Equivalently,

$$\mathbb{P}\left[|\boldsymbol{\xi}^\top\mathbf{A}\boldsymbol{\xi}-\mathbb{E}[\boldsymbol{\xi}^\top\mathbf{A}\boldsymbol{\xi}]|>t\right]\leq 2\exp\left(-\frac{t^2}{4\left\|\mathbf{A}\right\|_F^2+\left\|\mathbf{A}\right\|_2 t}\right).$$

Lemma 9.5 (Example 2.30 in (Wainwright, 2019)). *Let $\mathbf{w}\sim\mathcal{N}(\mathbf{0},\mathbf{I}_d)$ and $\mathcal{A}$ be a set in $\mathbb{R}^d$. Then $\sup_{\mathbf{a}\in\mathcal{A}}\langle\mathbf{a},\mathbf{w}\rangle$ is a sub-Gaussian random variable with variance proxy $\sup_{\mathbf{a}\in\mathcal{A}}\left\|\mathbf{a}\right\|_2^2$.*

9.2 Other Auxiliary Results

Lemma 9.6 (Lemma E.2 in (Arora et al., 2019b)). *For events $\mathcal{A},\mathcal{B}$, define the event $\mathcal{A}\Rightarrow\mathcal{B}$ as $\neg\mathcal{A}\vee\mathcal{B}$. Then $\mathbb{P}[\mathcal{A}\Rightarrow\mathcal{B}]\geq\mathbb{P}[\mathcal{B}|\mathcal{A}]$.*

Lemma 9.7 (Lemma E.3 in (Arora et al., 2019b)). *Let $\mathbf{w}\sim\mathcal{N}(\mathbf{0},\mathbf{I}_d)$, $\mathbf{G}\in\mathbb{R}^{d\times k}$ be some fixed matrix, and random vector $\mathbf{F}=\mathbf{w}^\top\mathbf{G}$, then conditioned on the value of $\mathbf{F}$, $\mathbf{w}$ remains Gaussian in the null space of the column space of $\mathbf{G}$, i.e.,*

$$\Pi_{\mathbf{G}}^{\perp}\mathbf{w}\overset{\mathcal{D}}{=}_{\mathbf{F}=\mathbf{w}^\top\mathbf{G}}\Pi_{\mathbf{G}}^{\perp}\widetilde{\mathbf{w}}.$$

where $\widetilde{\mathbf{w}}\sim\mathcal{N}(\mathbf{0},\mathbf{I}_d)$ is a fresh i.i.d. copy of $\mathbf{w}$.

9.3 Proof of the Main Result

Now we prove our main result. Notice that we rescale the mask $\mathbf{m}_{ij}^{(h)}\sim\sqrt{\frac{1}{\alpha}}\text{Bernoulli}(\alpha)$ so that $\mathbb{E}(\mathbf{m}_{ij}^{(h)})^2=1$. From a high level, our proof follows the proof outline of our asymptotic result.

Theorem 9.8 (Non-Asymptotic Bound, Full Version of Theorem 3.5). *Consider an L-hidden-layer fully-connected ReLU neural network with all the weights initialized with i.i.d. standard Gaussian distribution. Suppose all the weights except the input layer are pruned with probability $1-\alpha$ at the initialization and after pruning we rescale the weights by $1/\sqrt{\alpha}$. For $\delta\in(0,1)$ and sufficiently small $\epsilon>0$, if*

$$d_h\geq\Omega\left(\max(\frac{1}{\alpha}\frac{L^6}{\epsilon^4}\log\frac{Ld_{h+1}}{\delta},\frac{1}{\alpha^2}\frac{L^2}{\epsilon^2}\log\frac{Ld_{h+1}\sum_{h'=1}^{L-1}d_{h'}}{\delta},\frac{1}{\alpha}\frac{L^4}{\epsilon^2}\log\frac{2Ld_{h+1}\sum_{h'=1}^{h-1}d'_h}{\delta_3})\right),\ \forall h\in[L].$$

Then for any inputs $\mathbf{x},\mathbf{x}'\in\mathbb{R}^{d_0}$ such that $\left\|\mathbf{x}\right\|_2\leq 1,\ \left\|\mathbf{x}'\right\|_2\leq 1$, with probability at least $1-\delta$ we have

$$\left|\left\langle\frac{\partial f(\boldsymbol{\theta},\mathbf{x})}{\partial\boldsymbol{\theta}},\frac{\partial f(\boldsymbol{\theta},\mathbf{x}')}{\partial\boldsymbol{\theta}}\right\rangle-\boldsymbol{\Theta}^{(L)}(\mathbf{x},\mathbf{x}')\right|\leq(L+1)\epsilon.$$

Our analysis conditions on the following event occur.

Lemma 9.9. *For $\epsilon\in(0,1/2),\delta\in(0,1)$, if $d_h\geq\Omega(\frac{1}{\alpha\epsilon^2}\cdot\log(\frac{2d_{h+1}L}{\delta}))$, then*

$$\mathbb{P}\left[\forall i\in[d_{h+1}],\ h\in[L]:\left|\sum_{j=1}^{d_h}\mathbb{I}(\mathbf{m}_{ij}^{(h)}\neq 0)-\alpha d_h\right|\geq\epsilon\alpha d_h\right]\leq\delta$$

Proof. The proof is by applying Theorem 9.1 and then take a union bound over $i \in [d_{h+1}], h \in [L]$. □

Let $\mathbf{m}_i^{(h)}$ denote the i-th row of the mask in h-th layer. We first define the following events:

- $\mathcal{A}_i^h(\mathbf{x}, \mathbf{x}', \epsilon_1) := \left\{ \left| \left(\mathbf{g}^{(h)}(\mathbf{x}) \odot \mathbf{m}_i^{(h+1)}\right)^\top \left(\mathbf{g}^{(h)}(\mathbf{x}') \odot \mathbf{m}_i^{(h+1)}\right) - \mathbf{\Sigma}^{(h)}(\mathbf{x}, \mathbf{x}') \right| \le \epsilon_1 \right\}$.
- $\mathcal{A}^h(\mathbf{x}, \mathbf{x}', \epsilon_1) = \bigcap_{i=1}^{d_{h+1}} \mathcal{A}_i^h(\mathbf{x}, \mathbf{x}', \epsilon_1) \bigcap \left\{ \left| \left(\mathbf{g}^{(h)}(\mathbf{x})\right)^\top \mathbf{g}^{(h)}(\mathbf{x}') - \mathbf{\Sigma}^{(h)}(\mathbf{x}, \mathbf{x}') \right| \le \epsilon_1 \right\}$.
- $\overline{\mathcal{A}}^h(\epsilon_1) = \mathcal{A}^h(\mathbf{x}, \mathbf{x}, \epsilon_1) \cap \mathcal{A}^h(\mathbf{x}, \mathbf{x}', \epsilon_1) \cap \mathcal{A}^h(\mathbf{x}', \mathbf{x}', \epsilon_1)$.
- $\overline{\mathcal{A}}(\epsilon_1) = \bigcap_{h=0}^{L} \overline{\mathcal{A}}^h(\mathbf{x}, \mathbf{x}', \epsilon_1)$.
- $\mathcal{B}^h(\mathbf{x}, \mathbf{x}', \epsilon_2) = \left\{ \left| \left\langle \mathbf{b}^{(h)}(\mathbf{x}), \mathbf{b}^{(h)}(\mathbf{x}') \right\rangle - \prod_{h=h}^{L} \dot{\Sigma}^{(h)}(\mathbf{x}, \mathbf{x}') \right| < \epsilon_2 \right\}$.
- $\overline{\mathcal{B}}^h(\epsilon_2) = \mathcal{B}^h(\mathbf{x}, \mathbf{x}, \epsilon_2) \cap \mathcal{B}^h(\mathbf{x}, \mathbf{x}', \epsilon_2) \cap \mathcal{B}^h(\mathbf{x}', \mathbf{x}', \epsilon_2)$.
- $\overline{\mathcal{B}}(\epsilon_2) = \bigcap_{h=1}^{L+1} \overline{\mathcal{B}}^h(\mathbf{x}, \mathbf{x}', \epsilon_2)$.
- $\overline{\mathcal{C}}(\epsilon_3)$: a event defined in Definition 9.23.
- $\mathcal{D}_i^h(\mathbf{x}, \mathbf{x}', \epsilon_4) = \left\{ \left| 2\frac{\mathrm{Tr}(\mathbf{M}_i^{(h+1)} \mathbf{D}^{(h)}(\mathbf{x}, \mathbf{x}') \mathbf{M}_i^{(h+1)})}{d_h} - \dot{\Sigma}^{(h)}(\mathbf{x}, \mathbf{x}') \right| < \epsilon_4 \right\}$ where $\mathbf{M}_i^{(h+1)} = \mathrm{diag}(\mathbf{m}_i^{(h+1)})$.
- $\mathcal{D}^h(\mathbf{x}, \mathbf{x}', \epsilon_4) = \bigcap_{i=1}^{d_{h+1}} \mathcal{D}_i^h(\mathbf{x}, \mathbf{x}', \epsilon_4)$.
- $\overline{\mathcal{D}}^h(\epsilon_4) = \mathcal{D}^h(\mathbf{x}, \mathbf{x}, \epsilon_4) \cap \mathcal{D}^h(\mathbf{x}, \mathbf{x}', \epsilon_4) \cap \mathcal{D}^h(\mathbf{x}', \mathbf{x}', \epsilon_4)$.
- $\overline{\mathcal{D}}(\epsilon_4) = \bigcap_{h=1}^{L+1} \overline{\mathcal{D}}^h(\epsilon_4)$.

Proof of Theorem 9.8. Recall that

$$\left\langle \frac{\partial \widetilde{f}(\mathbf{x})}{\partial \mathbf{W}^{(h)}}, \frac{\partial \widetilde{f}(\mathbf{x}')}{\partial \mathbf{W}^{(h)}} \right\rangle = \left(\widetilde{\mathbf{b}}^{(h)}(\mathbf{x})\right)^\top \mathbf{G}^{(h-1)} \widetilde{\mathbf{b}}^{(h)}(\mathbf{x}'),$$

where $\mathbf{G}^{(h-1)}$ is a diagonal matrix and $\mathbf{G}_{ii}^{(h-1)} = \left\langle \widetilde{\mathbf{g}}^{(h-1)}(\mathbf{x}) \odot \mathbf{m}_i^{(h)}, \widetilde{\mathbf{g}}^{(h-1)}(\mathbf{x}') \odot \mathbf{m}_i^{(h)} \right\rangle$ and

$$\lim_{d_1, d_2, \ldots, d_L \to \infty} \left\langle \frac{\partial \widetilde{f}(\mathbf{x})}{\partial \mathbf{W}^{(h)}}, \frac{\partial \widetilde{f}(\mathbf{x}')}{\partial \mathbf{W}^{(h)}} \right\rangle = \Sigma^{(h-1)}(\mathbf{x}^{(1)}, \mathbf{x}^{(2)}) \prod_{h'=h}^{L} \dot{\Sigma}^{(h')}(\mathbf{x}^{(1)}, \mathbf{x}^{(2)}).$$

The rest of proof of our main result is based on letting Theorem 9.10 hold for ϵ' and then take $\epsilon := \epsilon'/L$. □

Theorem 9.10. *Consider the same setting as in Theorem 9.8. If*

$$d_h \ge \Omega\left(\max\left(\frac{1}{\alpha} \frac{L^2}{\epsilon^4} \log \frac{L d_{h+1}}{\delta}, \frac{1}{\alpha^2} \frac{1}{\epsilon^2} \log \frac{L d_{h+1} \sum_{h'=1}^{L-1} d_{h'}}{\delta}, \frac{1}{\alpha} \frac{L^2}{\epsilon^2} \log \frac{2 L d_{h+1} \sum_{h'=1}^{h-1} d'_h}{\delta_3} \right) \right), \ \forall h \in [L],$$

and $\epsilon \le \frac{c}{L}$ for some constant c, then for any fixed $\mathbf{x}, \mathbf{x}' \in \mathbb{R}^{d_0}$, $\|\mathbf{x}\|_2$, $\|\mathbf{x}'\|_2 \le 1$, we have with probability $1 - \delta$, $\forall 0 \le h \le L$, $\forall (\mathbf{x}^{(1)}, \mathbf{x}^{(2)}) \in \{(\mathbf{x}, \mathbf{x}), (\mathbf{x}, \mathbf{x}'), (\mathbf{x}', \mathbf{x}')\}$,

$$\left| \left(\mathbf{g}^{(h)}(\mathbf{x}^{(1)}) \odot \mathbf{m}_i^{(h+1)}\right)^\top \left(\mathbf{g}^{(h)}(\mathbf{x}^{(2)}) \odot \mathbf{m}_i^{(h+1)}\right) - \Sigma^{(h)}(\mathbf{x}^{(1)}, \mathbf{x}^{(2)}) \right| \le \epsilon^2/2, \quad \forall i \in [d_{h+1}],$$

and

$$\left| \left\langle \mathbf{b}^{(h)}(\mathbf{x}^{(1)}), \mathbf{b}^{(h)}(\mathbf{x}^{(2)}) \right\rangle - \prod_{h'=h}^{L} \dot{\Sigma}^{(h')}(\mathbf{x}^{(1)}, \mathbf{x}^{(2)}) \right| < 3L\epsilon.$$

In other words,

$$\mathbb{P}\left[\overline{\mathcal{A}}\left(\frac{\epsilon^2}{2}\right)\bigcap\overline{\mathcal{B}}(3L\epsilon)\right]\geq 1-\delta.$$

The first part of the result of Theorem 9.10 is proved by the following theorem.

Theorem 9.11 (Full Version of Theorem 5.1)**.** *Consider the same setting as in Theorem 9.8. There exist constants c such that if $d_h \geq \Omega(\frac{1}{\alpha}\frac{L^2}{\epsilon^2}\log\frac{18d_{h+1}L}{\delta})$, $\forall h\in\{1,2,\ldots,L\}$ and $\epsilon\leq\min(c,\frac{1}{L})$ then for any fixed $\mathbf{x},\mathbf{x}'\in\mathbb{R}^{d_0}$, $\|\mathbf{x}\|_2,\|\mathbf{x}'\|_2\leq 1$, we have with probability $1-\delta$, $\forall 0\leq h\leq L$, $\forall i\in[d_{h+1}]$, $\forall(\mathbf{x}^{(1)},\mathbf{x}^{(2)})\in\{(\mathbf{x},\mathbf{x}),(\mathbf{x},\mathbf{x}'),(\mathbf{x}',\mathbf{x}')\}$,*

$$\left|\left(\mathbf{g}^{(h)}(\mathbf{x}^{(1)})\odot\mathbf{m}_i^{(h+1)}\right)^\top\left(\mathbf{g}^{(h)}(\mathbf{x}^{(2)})\odot\mathbf{m}_i^{(h+1)}\right)-\Sigma^{(h)}(\mathbf{x}^{(1)},\mathbf{x}^{(2)})\right|\leq\epsilon$$
$$\left|\left(\mathbf{g}^{(h)}(\mathbf{x}^{(1)})\right)^\top\mathbf{g}^{(h)}(\mathbf{x}^{(2)})-\Sigma^{(h)}(\mathbf{x}^{(1)},\mathbf{x}^{(2)})\right|\leq\epsilon.$$

In other words, if $d_h\geq\Omega(\frac{1}{\alpha}\frac{L^2}{\epsilon_1^2}\log\frac{18d_{h+1}L}{\delta_1})$, $\forall h\in\{1,2,\ldots,L\}$ and $\epsilon_1\leq\min(c_2,\frac{1}{L})$ then

$$\mathbb{P}\left[\overline{\mathcal{A}}(\epsilon_1)\right]\geq 1-\delta_1$$

The proof of Theorem 9.11 can be found in Section 9.4.

Lemma 9.12. *If $d_h\geq\Omega(\frac{1}{\alpha}\frac{1}{\epsilon_4^2}\log\frac{12Ld_{h+1}}{\delta_4})$ for all $h\in[L]$, then*

$$\mathbb{P}\left[\overline{\mathcal{A}}(\epsilon_1^2/2)\Rightarrow\overline{\mathcal{D}}\left(\epsilon_1+\epsilon_4\right)\right]\geq 1-\delta_4.$$

The proof of Lemma 9.12 on a single pair can be found in Section 9.5 and then take a union bound over pairs $(\mathbf{x},\mathbf{x}),(\mathbf{x},\mathbf{x}'),(\mathbf{x}',\mathbf{x}')$.

Lemma 9.13. *If $d_h\geq\Omega(\frac{1}{\alpha}\frac{L^2}{\epsilon^2}\log\frac{2Ld_{h+1}\sum_{h'=1}^{h-1}d_h'}{\delta_3})=\widetilde{\Omega}(\frac{1}{\alpha}\frac{L^2}{\epsilon^2})$ for all $h\in[L]$, then*

$$\mathbb{P}\left[\overline{\mathcal{A}}(\epsilon_1)\Rightarrow\overline{\mathcal{C}}\left(2\sqrt{\log\frac{4\sum_{h'=1}^{L-1}d_{h'}}{\delta_3}}\right)\right]\geq 1-\delta_3$$

The proof of Lemma 9.13 can be found in Section 9.6.2.

Lemma 9.14. *Let $\epsilon_3=2\sqrt{\log\frac{4\sum_{h'=1}^{L-1}d_{h'}}{\delta_3}}$. If $d_h\geq\frac{8}{\alpha}\log\frac{6}{\delta_2}$, with probability $1-\delta_2$, the event $\overline{\mathcal{C}}(\epsilon_3)$ holds and, there exists constant C,C' such that for any $\epsilon_2,\epsilon_4\in[0,1]$, we have*

$$\mathbb{P}\left[\overline{\mathcal{A}}^L(\epsilon_1^2/2)\bigcap\overline{\mathcal{B}}^{h+1}(\epsilon_2)\bigcap\overline{\mathcal{C}}(\epsilon_3)\bigcap\overline{\mathcal{D}}^h(\epsilon_4)\Rightarrow\overline{\mathcal{B}}^h\left(\epsilon_2+2\epsilon_4+\frac{48\sqrt{2}}{\sqrt{d_h}}+48\sqrt{\frac{2}{\alpha}\frac{\log\frac{8}{\delta_2}}{d_h}}+\frac{96}{\alpha}\frac{\sqrt{2\log\frac{4\sum_{h'=1}^{L-1}d_{h'}}{\delta_3}}}{\sqrt{d_h}}\right)\right]$$
$$\geq 1-\delta_2/2.$$

The proof of Lemma 9.14 can be found in Section 9.6.

Proof of Theorem 9.10. We prove by induction on Lemma 9.14. We first let the event in Lemma 9.9 holds with ϵ and probability $1-\delta/5$. In the statement of Theorem 9.11, we set $\delta_1=\delta/5$, $\epsilon_1=\frac{\epsilon^2}{8}$, if $d_h\geq\Omega(\frac{1}{\alpha}\frac{L^2}{\epsilon^4}\log\frac{d_{h+1}L}{\delta})=\widetilde{\Omega}(\frac{1}{\alpha}\frac{L^2}{\epsilon^4})$, $\forall h\in\{1,2,\ldots,L\}$, we have

$$\mathbb{P}[\overline{\mathcal{A}}(\epsilon^2/8)]\geq 1-\delta/5.$$

In the statement of Lemma 9.12, we set $\delta_4 = \delta/5$ and $\epsilon_1 = \epsilon/2,\ \epsilon_4 = \epsilon/4$. If $d_h \geq \Omega(\frac{1}{\alpha}\frac{1}{\epsilon^2}\log\frac{Ld_{h+1}}{\delta}) = \widetilde{\Omega}(\frac{1}{\alpha}\frac{1}{\epsilon^2})$ for all $h \in [L]$, then

$$\mathbb{P}\left[\overline{\mathcal{A}}(\epsilon^2/8) \Rightarrow \overline{\mathcal{D}}\left(\epsilon\right)\right] \geq 1 - \delta/5.$$

In the statement of Lemma 9.13, setting $\delta_3 = \delta/5$, if $d_h \geq \Omega(\frac{1}{\alpha}\frac{L^2}{\epsilon^2}\log\frac{2Ld_{h+1}\sum_{h'=1}^{h-1}d'_h}{\delta_3}) = \widetilde{\Omega}(\frac{1}{\alpha}\frac{L^2}{\epsilon^2})$ for all $h \in [L]$, then

$$\mathbb{P}\left[\overline{\mathcal{A}}(\epsilon^2/8) \Rightarrow \overline{\mathcal{C}}\left(2\sqrt{\log\frac{20\sum_{h'=1}^{L-1}d_{h'}}{\delta}}\right)\right] \geq 1 - \delta/5.$$

Take a union bound we have

$$\mathbb{P}\left[\overline{\mathcal{A}}(\epsilon^2/2)\bigcap\overline{\mathcal{C}}\left(2\sqrt{\log\frac{20\sum_{h'=1}^{L-1}d_{h'}}{\delta}}\right)\bigcap\overline{\mathcal{D}}(\epsilon)\right] \geq 1 - \frac{3\delta}{5}.$$

Now we begin the induction. First of all, $\mathbb{P}\left[\overline{\mathcal{B}}^{L+1}(0)\right] = 1$ by definition. For $1 \leq h \leq L$, in the statement of Lemma 9.14, set $\epsilon_2 = 3(L+1-h), \epsilon_3 = 2\sqrt{\log\frac{20\sum_{h'=1}^{L-1}d_{h'}}{\delta}}, \delta_2 = \frac{\delta}{4L}$. If $d_h \geq \Omega(\frac{1}{\alpha^2}\frac{1}{\epsilon^2}\log\frac{L\sum_{h'=1}^{L-1}d_{h'}}{\delta}) = \widetilde{\Omega}(\frac{1}{\alpha^2}\frac{1}{\epsilon^2})$, we have $\frac{48\sqrt{2}}{\sqrt{d_h}} + 48\sqrt{\frac{2}{\alpha}\frac{\log\frac{8}{\delta_2}}{d_h}} + \frac{96}{\alpha}\frac{\sqrt{2\log\frac{4\sum_{h'=1}^{L-1}d_{h'}}{\delta_3}}}{\sqrt{d_h}} < \epsilon/2$. Thus we have

$$\begin{aligned}
&\mathbb{P}\left[\overline{\mathcal{B}}^{(h+1)}((3L-3h)\epsilon)\bigcap\overline{\mathcal{C}}\left(\epsilon_3\right)\bigcap\overline{\mathcal{D}}(\epsilon) \Rightarrow \overline{\mathcal{B}}^h\left((3L+2-3h)\epsilon + \frac{48\sqrt{2}}{\sqrt{d_h}} + 48\sqrt{\frac{2}{\alpha}\frac{\log\frac{8}{\delta_2}}{d_h}} + \frac{96}{\alpha}\frac{\sqrt{2\log\frac{4\sum_{h'=1}^{L-1}d_{h'}}{\delta_3}}}{\sqrt{d_h}}\right)\right] \\
&\geq \mathbb{P}\left[\overline{\mathcal{B}}^{(h+1)}((3L-3h)\epsilon)\bigcap\overline{\mathcal{C}}\left(\epsilon_3\right)\bigcap\overline{\mathcal{D}}(\epsilon) \Rightarrow \overline{\mathcal{B}}^h((3L+3-3h)\epsilon)\right] \\
&\geq 1 - \frac{\delta}{5L}
\end{aligned}$$

Applying union bound for every $h \in [L]$, we have

$$\begin{aligned}
&\mathbb{P}\left[\overline{\mathcal{A}}^L(\epsilon^2/8)\bigcap\overline{\mathcal{B}}(3L\epsilon)\bigcap\overline{\mathcal{C}}(\epsilon_3)\bigcap\overline{\mathcal{D}}(\epsilon)\right] \\
&\geq \mathbb{P}\left[\overline{\mathcal{A}}^L(\epsilon^2/8)\bigcap_{h=1}^{L}\overline{\mathcal{B}}^h(3(L+1-h)\epsilon)\bigcap\overline{\mathcal{C}}(\epsilon_3)\bigcap\overline{\mathcal{D}}(\epsilon)\right] \\
&\geq 1 - \mathbb{P}\left[\neg\left(\overline{\mathcal{A}}(\epsilon^2/8)\bigcap\overline{\mathcal{C}}(\epsilon_3)\bigcap\overline{\mathcal{D}}^h(\epsilon)\right)\right] \\
&\quad - \sum_{h=1}^{L}\mathbb{P}\left[\neg\left(\overline{\mathcal{B}}^{(h+1)}((3L-3h)\epsilon)\bigcap\overline{\mathcal{C}}\left(\epsilon_3\right)\bigcap\overline{\mathcal{D}}(\epsilon) \Rightarrow \overline{\mathcal{B}}^h((3L+3-3h)\epsilon)\right)\right] \\
&\geq 1 - \delta
\end{aligned}$$

□

9.4 Proof of Theorem 9.11: Forward Propagation

In this section, we prove $\overline{\mathcal{A}}(\mathbf{x}, \mathbf{x}', \epsilon)$ holds which is shown in Theorem 9.11 below. The main goal is to obtain bounds on $\left|\left(\mathbf{m}^{(h+1)} \odot \mathbf{g}^{(h)}(\mathbf{x})\right)^\top\left(\mathbf{m}^{(h+1)} \odot \mathbf{g}^{(h)}(\mathbf{x}')\right) - \Sigma^{(h)}(\mathbf{x}, \mathbf{x}')\right|$. We first introduce a result from previous work.

Lemma 9.15 (Lemma 13 in (Daniely et al., 2016)). *Define the function*

$$\overline{\sigma}(\mathbf{\Sigma}) = c_\sigma \mathop{\mathbb{E}}_{(X,Y)\sim\mathcal{N}(\mathbf{0},\mathbf{\Sigma})}\sigma(X)\sigma(Y),$$

and the set

$$\mathcal{M}_+^\gamma := \left\{ \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{12} & \Sigma_{22} \end{bmatrix} \in \mathcal{M}_+ | 1-\gamma \leq \Sigma_{11}, \Sigma_{22} \leq 1+\gamma \right\},$$

where $\mathcal{M}_+$ denote the set of positive semi-definite matrices. Then $\overline{\sigma}$ is $(1+o(\epsilon))$-Lipschitz on $\mathcal{M}_+^\epsilon$ with respect to ∞-norm.

Our analysis follows from the proof of Theorem 14 in (Daniely et al., 2016).

Proof of Theorem 9.11. We prove the first inequality first. Define the quantity $B_d = \sum_{i=1}^d (1+o(\epsilon))^i$.

We begin our proof by saying the h-th layer of a neural network is well-initialized if $\forall i \in [d_{h+1}]$, we have

$$\left| \left(\mathbf{g}^{(h)}(\mathbf{x}^{(1)}) \odot \mathbf{m}_i^{(h+1)} \right)^\top \left(\mathbf{g}^{(h)}(\mathbf{x}^{(2)}) \odot \mathbf{m}_i^{(h+1)} \right) - \Sigma^{(h)}(\mathbf{x}^{(1)}, \mathbf{x}^{(2)}) \right| \leq \epsilon \frac{B_h}{B_L}.$$

We prove the result by induction. Since we don't prune the input layer, the result trivially holds for $h = 0$. Assume all the layers first $h-1$ layers are well-initialized.

Now, conditioned on $\mathbf{g}^{(h-1)}(\mathbf{x}^{(1)}), \mathbf{g}^{(h-1)}(\mathbf{x}^{(2)}), \mathbf{m}^{(h)}$, we have

$$\begin{aligned}
&\mathop{\mathbb{E}}_{\mathbf{W}^{(h)}, \mathbf{m}^{(h+1)}} \left[\left(\mathbf{g}^{(h)}(\mathbf{x}^{(1)}) \odot \mathbf{m}_i^{(h+1)} \right)^\top \left(\mathbf{g}^{(h)}(\mathbf{x}^{(2)}) \odot \mathbf{m}_i^{(h+1)} \right) \right] \\
&= \mathop{\mathbb{E}}_{\mathbf{W}^{(h)}} \left[\left(\mathbf{g}^{(h)}(\mathbf{x}^{(1)}) \right)^\top \mathbf{g}^{(h)}(\mathbf{x}^{(2)}) \right] \\
&= \frac{c_\sigma}{d_h} \sum_{i=1}^{d_h} \mathop{\mathbb{E}}_{\mathbf{W}^{(h)}} \left[\sigma \left(\left\langle \mathbf{W}_i^{(h)}, \mathbf{m}_i^{(h)} \odot \mathbf{g}^{(h-1)}(\mathbf{x}^{(1)}) \right\rangle \right) \sigma \left(\left\langle \mathbf{W}_i^{(h)}, \mathbf{m}_i^{(h)} \odot \mathbf{g}^{(h-1)}(\mathbf{x}^{(2)}) \right\rangle \right) \right].
\end{aligned}$$

where $\mathbf{W}_i^{(h)}$ denotes the i-th row of $\mathbf{W}^{(h)}$. Define

$$\begin{aligned}
\widehat{\Sigma}_i^{(h)}(\mathbf{x}^{(1)}, \mathbf{x}^{(2)}) &= \left(\mathbf{g}^{(h)}(\mathbf{x}^{(1)}) \odot \mathbf{m}_i^{(h+1)} \right)^\top \left(\mathbf{g}^{(h)}(\mathbf{x}^{(2)}) \odot \mathbf{m}_i^{(h+1)} \right), \\
\widehat{\mathbf{\Lambda}}_i^{(h)}(\mathbf{x}^{(1)}, \mathbf{x}^{(2)}) &= \begin{bmatrix} \widehat{\Sigma}_i^{(h)}(\mathbf{x}^{(1)}, \mathbf{x}^{(1)}) & \widehat{\Sigma}_i^{(h)}(\mathbf{x}^{(1)}, \mathbf{x}^{(2)}) \\ \widehat{\Sigma}_i^{(h)}(\mathbf{x}^{(2)}, \mathbf{x}^{(1)}) & \widehat{\Sigma}_i^{(h)}(\mathbf{x}^{(2)}, \mathbf{x}^{(2)}) \end{bmatrix}.
\end{aligned}$$

Notice that for a given j, conditioned on $\mathbf{m}^{(h)}$, $\mathbf{g}^{(h-1)}(\mathbf{x}^{(1)})$ and $\mathbf{g}^{(h-1)}(\mathbf{x}^{(2)})$, and consider the randomness in $\mathbf{W}_j$, $\sigma \left(\left\langle \mathbf{W}_j^{(h)}, \mathbf{m}_j^{(h)} \odot \mathbf{g}^{(h-1)}(\mathbf{x}^{(1)}) \right\rangle \right) \sigma \left(\left\langle \mathbf{W}_j^{(h)}, \mathbf{m}_j^{(h)} \odot \mathbf{g}^{(h-1)}(\mathbf{x}^{(2)}) \right\rangle \right)$ is subgamma with parameters $(O(1), O(1))$ and

$$\begin{aligned}
&\mathbb{E} \left[\sigma \left(\left\langle \mathbf{W}_j^{(h)}, \mathbf{m}_j^{(h)} \odot \mathbf{g}^{(h-1)}(\mathbf{x}^{(1)}) \right\rangle \right) \sigma \left(\left\langle \mathbf{W}_j^{(h)}, \mathbf{m}_j^{(h)} \odot \mathbf{g}^{(h-1)}(\mathbf{x}^{(2)}) \right\rangle \right) \right] \\
&\leq \sqrt{\mathbb{E} \left[\left(\sigma \left(\left\langle \mathbf{W}_j^{(h)}, \mathbf{m}_j^{(h)} \odot \mathbf{g}^{(h-1)}(\mathbf{x}^{(1)}) \right\rangle \right) \right)^2 \right] \mathbb{E} \left[\left(\sigma \left(\left\langle \mathbf{W}_j^{(h)}, \mathbf{m}_j^{(h)} \odot \mathbf{g}^{(h-1)}(\mathbf{x}^{(2)}) \right\rangle \right) \right)^2 \right]} \\
&\leq \sqrt{\mathbb{E} \left[\left(\left\langle \mathbf{W}_j^{(h)}, \mathbf{m}_j^{(h)} \odot \mathbf{g}^{(h-1)}(\mathbf{x}^{(1)}) \right\rangle \right)^2 \right] \mathbb{E} \left[\left(\left\langle \mathbf{W}_j^{(h)}, \mathbf{m}_j^{(h)} \odot \mathbf{g}^{(h-1)}(\mathbf{x}^{(2)}) \right\rangle \right)^2 \right]} \\
&= \left\| \mathbf{m}_j^{(h)} \odot \mathbf{g}^{(h-1)}(\mathbf{x}^{(1)}) \right\|_2 \left\| \mathbf{m}_j^{(h)} \odot \mathbf{g}^{(h-1)}(\mathbf{x}^{(2)}) \right\|_2 \leq 4,
\end{aligned}$$

where the first inequality is by Cauchy-Schwarz inequality.

By Theorem 9.3, we have

$$\mathbb{P} \left[\left| \widehat{\Sigma}_i^{(h)}(\mathbf{x}^{(1)}, \mathbf{x}^{(2)}) - \mathop{\mathbb{E}}_{\mathbf{W}^{(h)}, \mathbf{m}^{(h+1)}} \widehat{\Sigma}_i^{(h)}(\mathbf{x}^{(1)}, \mathbf{x}^{(2)}) \right| > \epsilon \right] \leq 4 \exp \left\{ -\Omega(\alpha d_h \epsilon^2) \right\},$$

for some constant c_2 such that $\epsilon < c_2$.

Taking a union bound over $i \in [d_{h+1}]$, we have if $d_h \geq \Omega(\frac{1}{\alpha}\frac{B_L^2 \log \frac{8d_{h+1}L}{\delta}}{\epsilon^2})$, then with probability $1-\frac{\delta}{L}$ for all $i \in [d_{h+1}]$,

$$\left|\left(\mathbf{g}^{(h)}(\mathbf{x}^{(1)}) \odot \mathbf{m}_i^{(h+1)}\right)^\top \left(\mathbf{g}^{(h)}(\mathbf{x}^{(2)}) \odot \mathbf{m}_i^{(h+1)}\right) - \frac{c_\sigma}{d_h}\sum_{j=1}^{d_h} \mathop{\mathbb{E}}_{(u,v)\sim\mathcal{N}(\mathbf{0},\widehat{\mathbf{\Lambda}}_j^{(h-1)}(\mathbf{x}^{(1)},\mathbf{x}^{(2)}))}[\sigma(u)\sigma(v)]\right| \leq \epsilon/B_L.$$

Now apply triangle inequality

$$\begin{aligned}
&\left|\left(\mathbf{g}^{(h)}(\mathbf{x}^{(1)}) \odot \mathbf{m}_i^{(h+1)}\right)^\top \left(\mathbf{g}^{(h)}(\mathbf{x}^{(2)}) \odot \mathbf{m}_i^{(h+1)}\right) - \Sigma^{(h)}(\mathbf{x}^{(1)},\mathbf{x}^{(2)})\right| \\
&\leq \left|\left(\mathbf{g}^{(h)}(\mathbf{x}^{(1)}) \odot \mathbf{m}_i^{(h+1)}\right)^\top \left(\mathbf{g}^{(h)}(\mathbf{x}^{(2)}) \odot \mathbf{m}_i^{(h+1)}\right) - \frac{c_\sigma}{d_h}\sum_{j=1}^{d_h} \mathop{\mathbb{E}}_{(u,v)\sim\mathcal{N}(\mathbf{0},\widehat{\mathbf{\Lambda}}_j^{(h-1)}(\mathbf{x}^{(1)},\mathbf{x}^{(2)}))}[\sigma(u)\sigma(v)]\right| \\
&\quad + \left|\frac{c_\sigma}{d_h}\sum_{j=1}^{d_h} \mathop{\mathbb{E}}_{(u,v)\sim\mathcal{N}(\mathbf{0},\widehat{\mathbf{\Lambda}}_j^{(h-1)}(\mathbf{x}^{(1)},\mathbf{x}^{(2)}))}[\sigma(u)\sigma(v)] - \Sigma^{(h)}(\mathbf{x}^{(1)},\mathbf{x}^{(2)})\right| \\
&\leq \epsilon/B_L + \frac{1}{d_h}\sum_{i=1}^{d_h}\left|c_\sigma \mathop{\mathbb{E}}_{(u,v)\sim\mathcal{N}(\mathbf{0},\widehat{\mathbf{\Lambda}}_j^{(h-1)}(\mathbf{x}^{(1)},\mathbf{x}^{(2)}))}[\sigma(u)\sigma(v)] - \Sigma^{(h)}(\mathbf{x}^{(1)},\mathbf{x}^{(2)})\right| \\
&\leq \epsilon/B_L + \frac{1}{d_h}\sum_{i=1}^{d_h}(1+o(\epsilon))\epsilon\frac{B_{h-1}}{B_L} = \epsilon\frac{B_h}{B_L},
\end{aligned}$$

where the last inequality applies by the fact that $\overline{\sigma}$ is $(1+o(\epsilon))$-Lipschitz on $\mathcal{M}_+^\gamma$ with respect to the ∞-norm in Lemma 9.15 and the induction hypothesis that the first $h-1$ layers are well-initialized.

Finally we expand $B_d = \sum_{i=1}^d (1+o(\epsilon))^i$ and take $\epsilon = \min(c_2, \frac{1}{L})$, we have

$$B_L = \sum_{i=1}^L (1+o(\epsilon))^i \leq \sum_{i=1}^L e^{o(\epsilon)L} = O(L).$$

The proof for

$$\left|\left(\mathbf{g}^{(h)}(\mathbf{x}^{(1)})\right)^\top \mathbf{g}^{(h)}(\mathbf{x}^{(2)}) - \Sigma^{(h)}(\mathbf{x}^{(1)},\mathbf{x}^{(2)})\right| \leq \epsilon,$$

largely follows the same steps as above since

$$\mathop{\mathbb{E}}_{\mathbf{W}^{(h)}}\left(\mathbf{g}^{(h)}(\mathbf{x}^{(1)})\right)^\top \mathbf{g}^{(h)}(\mathbf{x}^{(2)}) = \mathop{\mathbb{E}}_{\mathbf{W}^{(h)},\mathbf{m}^{(h+1)}} \widehat{\Sigma}_i^{(h)}(\mathbf{x}^{(1)},\mathbf{x}^{(2)}).$$

Now applying the concentration of sub-Gamma random variables we have

$$\mathbb{P}\left[\left|\left(\mathbf{g}^{(h)}(\mathbf{x}^{(1)})\right)^\top \mathbf{g}^{(h)}(\mathbf{x}^{(2)}) - \mathop{\mathbb{E}}_{\mathbf{W}^{(h)}}\left(\mathbf{g}^{(h)}(\mathbf{x}^{(1)})\right)^\top \mathbf{g}^{(h)}(\mathbf{x}^{(2)})\right| \geq \epsilon\right] \leq 2\exp\{-\epsilon^2 d_h\}$$

for sufficiently small ϵ, which requires $d_h \geq \Omega(\frac{1}{\epsilon^2}\log\frac{6L}{\delta})$ by taking a union bound over L. □

Lemma 9.16. *Assume the event $\overline{\mathcal{A}}(\mathbf{x},\mathbf{x}',\epsilon)$ holds for $\epsilon < 1$. Then, with probability at least $1-\delta$ over the randomness of $\mathbf{w}^{(L+1)}$*

$$|f^{(L+1)}(\mathbf{x})| \leq \sqrt{2\log\frac{2}{\delta}}.$$

Proof. By definition, we have $f^{(L+1)}(\mathbf{x}) = \left\langle \mathbf{w}^{(L+1)} \odot \mathbf{m}^{(L+1)}, \mathbf{g}^{(h)}(\mathbf{x})\right\rangle$. By our assumption, $\left\|\mathbf{g}^{(h)}(\mathbf{x}) \odot \mathbf{m}^{(h+1)}\right\|_2^2 \leq 2$. Thus, by apply standard Gaussian tail bound, with probability at least $1-\delta$,

$$|f^{(L+1)}(\mathbf{x})| \leq \sqrt{2\log\frac{2}{\delta}}.$$

□

9.5 Proof of Lemma 9.12: Analyzing the Activation Gradient of a Single Layer

To prove Lemma 9.12, we first introduce a previous result.

Lemma 9.17 (Lemma E.8. (Arora et al., 2019b)). *Define*

$$t_{\dot\sigma}(\mathbf{\Sigma}) = c_\sigma \mathop{\mathbb{E}}_{(u,v)\sim\mathcal{N}(\mathbf{0},\mathbf{\Sigma}')}[\dot\sigma(u)\dot\sigma(v)] \quad \text{with} \quad \mathbf{\Sigma}' = \begin{bmatrix} 1 & \frac{\Sigma_{12}}{\sqrt{\Sigma_{11}\Sigma_{22}}} \\ \frac{\Sigma_{12}}{\sqrt{\Sigma_{11}\Sigma_{22}}} & 1 \end{bmatrix}.$$

Then

$$\left\|\mathbf{G}^{(h)}(\mathbf{x},\mathbf{x}') - \mathbf{\Lambda}^{(h)}(\mathbf{x},\mathbf{x}')\right\|_\infty \le \frac{\epsilon^2}{2} \Rightarrow \left|t_{\dot\sigma}\left(\mathbf{G}^{(h)}(\mathbf{x},\mathbf{x}')\right) - t_{\dot\sigma}\left(\mathbf{\Lambda}^{(h)}(\mathbf{x},\mathbf{x}')\right)\right| \le \epsilon.$$

Proof of Lemma 9.12. Conditioned on $\widehat{\mathbf{\Lambda}}_i^{(h)},\ \forall i \in [d_h]$ and consider the randomness of $\mathbf{W}^{(h)}, \mathbf{m}^{(h+1)}$, we have

$$\begin{aligned}
&\mathop{\mathbb{E}}_{\mathbf{W}^{(h)},\mathbf{m}^{(h+1)}}\left[2\frac{\mathrm{Tr}(\mathbf{M}_i^{(h+1)}\mathbf{D}^{(h)}(\mathbf{x},\mathbf{x}')\mathbf{M}_i^{(h+1)})}{d_h}\right] \\
&= \mathop{\mathbb{E}}_{\mathbf{W}^{(h)}}\left[2\frac{\mathrm{Tr}(\mathbf{D}^{(h)}(\mathbf{x},\mathbf{x}'))}{d_h}\right] \\
&= \frac{1}{d_h}\sum_{i=1}^{d_h}\mathop{\mathbb{E}}_{\mathbf{W}^{(h)}}\left[\dot\sigma\left(\left\langle\mathbf{W}_i^{(h)}, \mathbf{m}_i^{(h)}\odot\mathbf{g}^{(h-1)}(\mathbf{x})\right\rangle\right)\dot\sigma\left(\left\langle\mathbf{W}_i^{(h)}, \mathbf{m}_i^{(h)}\odot\mathbf{g}^{(h-1)}(\mathbf{x}')\right\rangle\right)\right] \\
&= \frac{1}{d_h}\sum_{i=1}^{d_h} t_{\dot\sigma}\left(\widehat{\mathbf{\Lambda}}_i^{(h)}\right).
\end{aligned}$$

Now, by triangle inequality and our assumption on $\widehat{\mathbf{\Lambda}}_i,\ \forall i \in [d_h]$, apply Lemma 9.17

$$\left|t_{\dot\sigma}\left(\mathbf{\Lambda}^{(h)}(\mathbf{x},\mathbf{x}')\right) - \frac{1}{d_h}\sum_{i=1}^{d_h} t_{\dot\sigma}\left(\widehat{\mathbf{\Lambda}}_i^{(h)}\right)\right| \le \frac{1}{d_h}\sum_{i=1}^{d_h}\left|t_{\dot\sigma}\left(\mathbf{\Lambda}^{(h)}(\mathbf{x},\mathbf{x}')\right) - t_{\dot\sigma}\left(\widehat{\mathbf{\Lambda}}_i^{(h)}\right)\right| \le \epsilon_1.$$

Finally, since $\dot\sigma(\mathbf{f}_j^{(h)}(\mathbf{x}))\dot\sigma(\mathbf{f}_j^{(h)}(\mathbf{x}'))$ is a 0-1 random variable, it is sub-Gaussian with variance proxy $\frac{1}{4}$. By Theorem 9.2, for a given i and $t > 0$,

$$\mathbb{P}\left[\left|2\frac{\mathrm{Tr}(\mathbf{M}_i^{(h+1)}\mathbf{D}^{(h)}(\mathbf{x},\mathbf{x}')\mathbf{M}_i^{(h+1)})}{d_h} - \frac{1}{d_h}\sum_{i=1}^{d_h} t_{\dot\sigma}\left(\widehat{\mathbf{\Lambda}}_i^{(h)}\right)\right| > t\right] \le 4\exp\left\{-\Omega(\alpha d_h t^2)\right\}.$$

Finally, by taking a union bound over $h \in [L]$, $i \in [d_{h+1}]$, if $d_h \ge \Omega(\frac{1}{\alpha}\frac{1}{\epsilon_4^2}\log\frac{4Ld_{h+1}}{\delta})$ with probability $1-\delta$ over the randomness of $\mathbf{W}^{(h)}, \mathbf{m}^{(h+1)}$, we have $\forall h \in [L], i \in [d_{h+1}]$,

$$\left|2\frac{\mathrm{Tr}(\mathbf{M}_i^{(h+1)}\mathbf{D}^{(h)}(\mathbf{x},\mathbf{x}')\mathbf{M}_i^{(h+1)})}{d_h} - \frac{1}{d_h}\sum_{i=1}^{d_h} t_{\dot\sigma}\left(\widehat{\mathbf{\Lambda}}_i^{(h)}\right)\right| < \epsilon_4.$$

By triangle inequality we have

$$\left|2\frac{\mathrm{Tr}(\mathbf{M}_i^{(h+1)}\mathbf{D}^{(h)}(\mathbf{x},\mathbf{x}')\mathbf{M}_i^{(h+1)})}{d_h} - t_{\dot\sigma}\left(\mathbf{\Lambda}^{(h)}(\mathbf{x},\mathbf{x}')\right)\right| < \epsilon_1 + \epsilon_4.$$

□

9.6 Proof of Lemma 9.14: The Fresh Gaussian Copy Trick

Proof of Lemma 9.14. The goal is to show that

$$\left|\left(\mathbf{b}^{(h+1)}(\mathbf{x}^{(1)})\right)^\top(\mathbf{W}^{(h+1)}\odot\mathbf{m}^{(h+1)})\mathbf{D}^{(h)}(\mathbf{x}^{(1)})\mathbf{D}^{(h)}(\mathbf{x}^{(2)})(\mathbf{W}^{(h+1)}\odot\mathbf{m}^{(h+1)})^\top\mathbf{b}^{(h+1)}(\mathbf{x}^{(2)}) - \prod_{h'=h}^{L}\dot\Sigma^{(h')}(\mathbf{x}^{(1)},\mathbf{x}^{(2)})\right|$$

is small. We can write

$$\left(\mathbf{b}^{(h+1)}(\mathbf{x}^{(1)})\right)^\top (\mathbf{W}^{(h+1)} \odot \mathbf{m}^{(h+1)})\mathbf{D}^{(h)}(\mathbf{x}^{(1)})\mathbf{D}^{(h)}(\mathbf{x}^{(2)})(\mathbf{W}^{(h+1)} \odot \mathbf{m}^{(h+1)})^\top \mathbf{b}^{(h+1)}(\mathbf{x}^{(2)})$$
$$= \frac{c_\sigma}{d_h} \sum_{i,j} \mathbf{b}_i^{(h+1)}(\mathbf{x}^{(1)})\mathbf{b}_j^{(h+1)}(\mathbf{x}^{(2)}) \left(\mathbf{w}_i^{(h+1)}\right) \mathbf{M}_i^{(h+1)}\mathbf{D}^{(h)}(\mathbf{x}^{(1)})\mathbf{D}^{(h)}(\mathbf{x}^{(2)})\mathbf{M}_j^{(h+1)}\mathbf{w}_j^{(h+1)}$$

We first show that this term is close to

$$\frac{2}{d_h} \sum_i \mathbf{b}_i^{(h+1)}(\mathbf{x}^{(1)})\mathbf{b}_i^{(h+1)}(\mathbf{x}^{(2)})\mathrm{Tr}(\mathbf{M}_i^{(h+1)}\mathbf{D}\mathbf{M}_i^{(h+1)}) \tag{17}$$

We do this by the following.

Let $\mathbf{G}_i^{(h)} = [(\mathbf{g}^{(h)}(\mathbf{x}) \odot \mathbf{m}_i^{(h+1)}) \quad (\mathbf{g}^{(h)}(\mathbf{x}') \odot \mathbf{m}_i^{(h+1)})]$ and $\mathbf{G}^{(h)} = [\mathbf{G}_1^{(h)}\mathbf{G}_2^{(h)} \dots \mathbf{G}_{d_{h+1}}^{(h)}]$ and $\mathbf{F}^{(h+1)} = (\mathbf{W}^{(h+1)} \odot \mathbf{m}^{(h+1)})\mathbf{G}^{(h)}$. We further simplify our notation to let $\mathbf{G}_i = \mathbf{G}_i^{(h)}$ since there is no ambiguity on layers. Notice that conditioned on $\mathbf{F}^{(h+1)}, \mathbf{m}^{(h+1)}, \mathbf{G}^{(h)}$, notice that

$$\left(\mathbf{b}^{(h+1)}(\mathbf{x})\right)^\top \begin{bmatrix} ((\mathbf{w}_1^{(h+1)})^\top \Pi_{\mathbf{G}_1}^\perp) \odot \mathbf{m}_1^{(h+1)} \\ ((\mathbf{w}_2^{(h+1)})^\top \Pi_{\mathbf{G}_2}^\perp) \odot \mathbf{m}_2^{(h+1)} \\ \vdots \\ ((\mathbf{w}_{d_{h+1}}^{(h+1)})^\top \Pi_{\mathbf{G}_{d_{h+1}}}^\perp) \odot \mathbf{m}_{d_{h+1}}^{(h+1)} \end{bmatrix} \in \mathbb{R}^{d_h}$$

has multivariate Gaussian distribution and by Lemma 9.7 it has the same distribution as

$$\left(\mathbf{b}^{(h+1)}(\mathbf{x})\right)^\top \begin{bmatrix} ((\mathbf{w}_1^{(h+1)})^\top \Pi_{\mathbf{G}_1}^\perp) \odot \mathbf{m}_1^{(h+1)} \\ ((\mathbf{w}_2^{(h+1)})^\top \Pi_{\mathbf{G}_2}^\perp) \odot \mathbf{m}_2^{(h+1)} \\ \vdots \\ ((\mathbf{w}_{d_{h+1}}^{(h+1)})^\top \Pi_{\mathbf{G}_{d_{h+1}}}^\perp) \odot \mathbf{m}_{d_{h+1}}^{(h+1)} \end{bmatrix} = \sum_{i=1}^{d_{h+1}} \mathbf{b}_i^{(h+1)}(\mathbf{x})((\widetilde{\mathbf{w}}_i^{(h+1)})^\top \Pi_{\mathbf{G}_i}^\perp) \odot \mathbf{m}_i^{(h+1)},$$

where $\widetilde{\mathbf{w}}_i^{(h+1)}$ is a fresh copy of i.i.d. Gaussian. First of all, let $\mathbf{M}_i^{(h+1)} = \mathrm{diag}(\mathbf{m}_i^{(h+1)})$, and we have

$$\frac{c_\sigma}{d_h} \sum_{i,j} \mathbf{b}_i^{(h+1)}(\mathbf{x}^{(1)})\mathbf{b}_j^{(h+1)}(\mathbf{x}^{(2)}) \left(\mathbf{w}_i^{(h+1)}\right) (\Pi_{\mathbf{G}_i} + \Pi_{\mathbf{G}_i}^\perp)\mathbf{M}_i^{(h+1)}\mathbf{D}^{(h)}(\mathbf{x}^{(1)})\mathbf{D}^{(h)}(\mathbf{x}^{(2)})\mathbf{M}_j^{(h+1)}(\Pi_{\mathbf{G}_j} + \Pi_{\mathbf{G}_j}^\perp)\mathbf{w}_j^{(h+1)}$$
$$= \frac{c_\sigma}{d_h} \sum_{i,j} \mathbf{b}_i^{(h+1)}(\mathbf{x}^{(1)})\mathbf{b}_j^{(h+1)}(\mathbf{x}^{(2)}) \left(\mathbf{w}_i^{(h+1)}\right)^\top \Pi_{\mathbf{G}_i}^\perp \mathbf{M}_i^{(h+1)}\mathbf{D}^{(h)}(\mathbf{x}^{(1)})\mathbf{D}^{(h)}(\mathbf{x}^{(2)})\mathbf{M}_j^{(h+1)}\Pi_{\mathbf{G}_j}^\perp \mathbf{w}_j^{(h+1)}$$
$$+ \frac{c_\sigma}{d_h} \sum_{i,j} \mathbf{b}_i^{(h+1)}(\mathbf{x}^{(1)})\mathbf{b}_j^{(h+1)}(\mathbf{x}^{(2)}) \left(\mathbf{w}_i^{(h+1)}\right)^\top \Pi_{\mathbf{G}_i}^\perp \mathbf{M}_i^{(h+1)}\mathbf{D}^{(h)}(\mathbf{x}^{(1)})\mathbf{D}^{(h)}(\mathbf{x}^{(2)})\mathbf{M}_j^{(h+1)}\Pi_{\mathbf{G}_j} \mathbf{w}_j^{(h+1)}$$
$$+ \frac{c_\sigma}{d_h} \sum_{i,j} \mathbf{b}_i^{(h+1)}(\mathbf{x}^{(1)})\mathbf{b}_j^{(h+1)}(\mathbf{x}^{(2)}) \left(\mathbf{w}_i^{(h+1)}\right)^\top \Pi_{\mathbf{G}_i} \mathbf{M}_i^{(h+1)}\mathbf{D}^{(h)}(\mathbf{x}^{(1)})\mathbf{D}^{(h)}(\mathbf{x}^{(2)})\mathbf{M}_j^{(h+1)}\Pi_{\mathbf{G}_j}^\perp \mathbf{w}_j^{(h+1)}$$
$$+ \frac{c_\sigma}{d_h} \sum_{i,j} \mathbf{b}_i^{(h+1)}(\mathbf{x}^{(1)})\mathbf{b}_j^{(h+1)}(\mathbf{x}^{(2)}) \left(\mathbf{w}_i^{(h+1)}\right)^\top \Pi_{\mathbf{G}_i} \mathbf{M}_i^{(h+1)}\mathbf{D}^{(h)}(\mathbf{x}^{(1)})\mathbf{D}^{(h)}(\mathbf{x}^{(2)})\mathbf{M}_j^{(h+1)}\Pi_{\mathbf{G}_j} \mathbf{w}_j^{(h+1)}. \tag{18}$$

We now show that the main contribution from the above term is from the part that involves $\Pi_{\mathbf{G}_i}^\perp$ and is close to the term in Equation (17), and the part with $\Pi_{\mathbf{G}_i}$ is small. This is done by Proposition 9.20 and Proposition 9.27. The rest of proof is by Proposition 9.18. □

Proposition 9.18. *If $\overline{\mathcal{A}}^L(\epsilon_1^2/2) \bigcap \overline{\mathcal{B}}^{h+1}(\epsilon_2) \bigcap \overline{\mathcal{C}}(\epsilon_3) \bigcap \overline{\mathcal{D}}^h(\epsilon_4)$, then we have*

$$\left| \frac{2}{d_h} \sum_i \mathbf{b}_i^{(h+1)}(\mathbf{x}^{(1)})\mathbf{b}_i^{(h+1)}(\mathbf{x}^{(2)})\mathrm{Tr}(\mathbf{M}_i^{(h+1)}\mathbf{D}\mathbf{M}_i^{(h+1)}) - \prod_{h'=h}^{L} \dot{\Sigma}^{(h')}(\mathbf{x}^{(1)}, \mathbf{x}^{(2)}) \right| \le \epsilon_2 + 2\epsilon_4.$$

Proof.

$$\begin{aligned}
&\left|\frac{2}{d_h}\sum_i \mathbf{b}_i^{(h+1)}(\mathbf{x}^{(1)})\mathbf{b}_i^{(h+1)}(\mathbf{x}^{(2)})\mathrm{Tr}(\mathbf{M}_i^{(h+1)}\mathbf{D}\mathbf{M}_i^{(h+1)}) - \prod_{h'=h}^{L}\dot{\Sigma}^{(h')}(\mathbf{x}^{(1)},\mathbf{x}^{(2)})\right| \\
&\leq \left|\sum_i \mathbf{b}_i^{(h+1)}(\mathbf{x}^{(1)})\mathbf{b}_i^{(h+1)}(\mathbf{x}^{(2)})\left(\frac{2}{d_h}\mathrm{Tr}(\mathbf{M}_i^{(h+1)}\mathbf{D}\mathbf{M}_i^{(h+1)}) - \dot{\Sigma}^{(h)}(\mathbf{x}^{(1)},\mathbf{x}^{(2)})\right)\right| \\
&+ \left|\dot{\Sigma}^{(h)}(\mathbf{x}^{(1)}\mathbf{x}^{(2)})\right|\left|\left\langle \mathbf{b}^{(h+1)}(\mathbf{x}^{(1)})\mathbf{b}^{(h+1)}(\mathbf{x}^{(2)})\right\rangle - \prod_{h'=h+1}^{L}\dot{\Sigma}^{(h')}(\mathbf{x}^{(1)},\mathbf{x}^{(2)})\right| \\
&\leq \left\|\mathbf{b}^{(h+1)}(\mathbf{x}^{(1)})\right\|_2\left\|\mathbf{b}^{(h+1)}(\mathbf{x}^{(2)})\right\|_2\epsilon_4 + \epsilon_2 \\
&= 2\epsilon_4 + \epsilon_2.
\end{aligned}$$

□

Before we prove Proposition 9.20 and Proposition 9.27, we first prove a convenient result.

Proposition 9.19. $\mathbf{M}_i^{(h+1)}$ *commutes with* $\Pi_{\mathbf{G}_i}$ *(and thus* $\Pi_{\mathbf{G}_i}^{\perp}$*).*

Proof. We can decompose $\Pi_{\mathbf{G}_i} = \Pi_{\mathbf{M}_i\mathbf{g}_1} + \Pi_{\mathbf{G}_i/\mathbf{M}_i\mathbf{g}_1}$. Observe that $\Pi_{\mathbf{G}_i/\mathbf{M}_i\mathbf{g}_1}$ is projecting a vector into the space spanned by $\mathbf{M}_i\mathbf{g}_2 - \langle\mathbf{M}_i\mathbf{g}_1, \mathbf{M}_i\mathbf{g}_2\rangle\,\mathbf{M}_i\mathbf{g}_1 = \mathbf{M}_i(\mathbf{g}_2 - \langle\mathbf{M}_i\mathbf{g}_1, \mathbf{M}_i\mathbf{g}_2\rangle\,\mathbf{g}_1)$. Thus, we can first prove $\mathbf{M}_i$ commutes with $\Pi_{\mathbf{M}_i\mathbf{g}_1}$ and the same result follows for $\Pi_{\mathbf{G}_i/\mathbf{M}_i\mathbf{g}_1}$. Notice that $\mathbf{M}_i\Pi_{\mathbf{M}_i\mathbf{g}_1} = \mathbf{M}_i\frac{\mathbf{M}_i\mathbf{g}_1(\mathbf{M}_i\mathbf{g}_1)^\top}{\|\mathbf{M}_i\mathbf{g}_1\|_2^2} = \frac{1}{\sqrt{\alpha}}\frac{\mathbf{M}_i\mathbf{g}_1(\mathbf{M}_i\mathbf{g}_1)^\top}{\|\mathbf{M}_i\mathbf{g}_1\|_2^2} = \frac{\mathbf{M}_i\mathbf{g}_1(\mathbf{M}_i\mathbf{g}_1)^\top}{\|\mathbf{M}_i\mathbf{g}_1\|_2^2}\mathbf{M}_i = \Pi_{\mathbf{M}_i\mathbf{g}_1}\mathbf{M}_i$. □

9.6.1 Bounding the Independent Part

Proposition 9.20 (Formal Version of Proposition 5.2)**.** *Conditioned on the event in Lemma 9.9 occurs. With probability at least* $1-\delta_2$*, if* $\overline{\mathcal{A}}^L(\epsilon_1^2/2)\bigcap\overline{\mathcal{B}}^{h+1}(\epsilon_2)\bigcap\overline{\mathcal{C}}(\epsilon_3)\bigcap\overline{\mathcal{D}}^h(\epsilon_4)$*, then for any* $(\mathbf{x}^{(1)},\mathbf{x}^{(2)}) \in \{(\mathbf{x},\mathbf{x}),(\mathbf{x},\mathbf{x}'),(\mathbf{x}',\mathbf{x}')\}$*, we have*

$$\begin{aligned}
&\left|\frac{c_\sigma}{d_h}\sum_{i,j}\mathbf{b}_i^{(h+1)}(\mathbf{x}^{(1)})\mathbf{b}_j^{(h+1)}(\mathbf{x}^{(2)})\left(\widetilde{\mathbf{w}}_i^{(h+1)}\right)^\top\Pi_{\mathbf{G}_i}^{\perp}\mathbf{M}_i^{(h+1)}\mathbf{D}\mathbf{M}_j^{(h+1)}\Pi_{\mathbf{G}_j}^{\perp}\widetilde{\mathbf{w}}_j^{(h+1)}\right. \\
&\quad\left. - \frac{2}{d_h}\sum_i \mathbf{b}_i^{(h+1)}(\mathbf{x}^{(1)})\mathbf{b}_i^{(h+1)}(\mathbf{x}^{(2)})\mathrm{Tr}(\mathbf{M}_i^{(h+1)}\mathbf{D}\mathbf{M}_i^{(h+1)})\right| \leq 3\sqrt{\frac{8\log\frac{6}{\delta_2}}{\alpha d_h}},
\end{aligned}$$

which implies for any $\mathbf{x}^{(1)} \in \{\mathbf{x},\mathbf{x}'\}$*,*

$$\left\|\sqrt{\frac{c_\sigma}{d_h}}\sum_i \mathbf{b}_i^{(h+1)}(\mathbf{x}^{(1)})\left(\widetilde{\mathbf{w}}_i^{(h+1)}\right)^\top\Pi_{\mathbf{G}_i}^{\perp}\mathbf{M}_i^{(h+1)}\mathbf{D}\right\|_2 \leq \sqrt{4+3\sqrt{\frac{8\log\frac{6}{\delta_2}}{\alpha d_h}}} \leq 6,$$

if $d_h \geq \frac{8}{\alpha}\log\frac{6}{\delta_2}$.

Proof. First, we compute the difference between the projected version of the inner product and normal inner product in expectation: First we have

$$\begin{aligned}
&\mathop{\mathbb{E}}_{\widetilde{\mathbf{W}}^{(h+1)}}\left(\frac{c_\sigma}{d_h}\sum_{i,j}\mathbf{b}_i^{(h+1)}(\mathbf{x}^{(1)})\mathbf{b}_j^{(h+1)}(\mathbf{x}^{(2)})\left(\widetilde{\mathbf{w}}_i^{(h+1)}\right)^\top\mathbf{M}_i^{(h+1)}\mathbf{D}\mathbf{M}_j^{(h+1)}\widetilde{\mathbf{w}}_j^{(h+1)}\right) \\
&= \frac{c_\sigma}{d_h}\sum_i \mathbf{b}_i^{(h+1)}(\mathbf{x}^{(1)})\mathbf{b}_i^{(h+1)}(\mathbf{x}^{(2)})\mathrm{Tr}(\mathbf{M}_i^{(h+1)}\mathbf{D}\mathbf{M}_i^{(h+1)}).
\end{aligned}$$

Then,

$$\begin{aligned}
&\mathop{\mathbb{E}}_{\widetilde{\mathbf{W}}^{(h+1)}}\left(\frac{c_\sigma}{d_h}\sum_{i,j}\mathbf{b}_i^{(h+1)}(\mathbf{x}^{(1)})\mathbf{b}_j^{(h+1)}(\mathbf{x}^{(2)})\left(\widetilde{\mathbf{w}}_i^{(h+1)}\right)^\top\Pi_{\widetilde{\mathbf{G}}_i}^{\perp}\mathbf{M}_i^{(h+1)}\mathbf{D}\mathbf{M}_j^{(h+1)}\Pi_{\widetilde{\mathbf{G}}_j}^{\perp}\widetilde{\mathbf{w}}_j^{(h+1)}\right.\\
&\quad\left.-\frac{c_\sigma}{d_h}\sum_{i,j}\mathbf{b}_i^{(h+1)}(\mathbf{x}^{(1)})\mathbf{b}_j^{(h+1)}(\mathbf{x}^{(2)})\left(\widetilde{\mathbf{w}}_i^{(h+1)}\right)^\top\mathbf{M}_i^{(h+1)}\mathbf{D}\mathbf{M}_j^{(h+1)}\widetilde{\mathbf{w}}_j^{(h+1)}\right)\\
&=\frac{c_\sigma}{d_h}\sum_{i}\mathbf{b}_i^{(h+1)}(\mathbf{x}^{(1)})\mathbf{b}_i^{(h+1)}(\mathbf{x}^{(2)})\mathrm{Tr}(\Pi_{\widetilde{\mathbf{G}}_i}^{\perp}\mathbf{M}_i^{(h+1)}\mathbf{D}\mathbf{M}_i^{(h+1)}\Pi_{\widetilde{\mathbf{G}}_i}^{\perp}-\mathbf{M}_i^{(h+1)}\mathbf{D}\mathbf{M}_i^{(h+1)})\\
&=\frac{c_\sigma}{d_h}\sum_{i}\mathbf{b}_i^{(h+1)}(\mathbf{x}^{(1)})\mathbf{b}_i^{(h+1)}(\mathbf{x}^{(2)})\mathrm{Tr}((\Pi_{\widetilde{\mathbf{G}}_i}^{\perp}-I)\mathbf{M}_i^{(h+1)}\mathbf{D}\mathbf{M}_i^{(h+1)})\\
&=\frac{c_\sigma}{d_h}\sum_{i}\mathbf{b}_i^{(h+1)}(\mathbf{x}^{(1)})\mathbf{b}_i^{(h+1)}(\mathbf{x}^{(2)})\mathrm{Tr}(\Pi_{\mathbf{G}_i}\mathbf{M}_i^{(h+1)}\mathbf{D}\mathbf{M}_i^{(h+1)}),
\end{aligned}$$

where the third last equality is true because we can interchange between $\mathbf{M}_i^{(h+1)}$ and $\Pi_{\widetilde{\mathbf{G}}_i}^{\perp}$. And the second last equality is because $\mathrm{Tr}(\Pi_{\widetilde{\mathbf{G}}_i}^{\perp}\mathbf{M}_i^{(h+1)}\mathbf{D}\mathbf{M}_i^{(h+1)}\Pi_{\widetilde{\mathbf{G}}_i}^{\perp}-\mathbf{M}_i^{(h+1)}\mathbf{D}\mathbf{M}_i^{(h+1)})=\mathrm{Tr}(\Pi_{\widetilde{\mathbf{G}}_i}^{\perp}\mathbf{M}_i^{(h+1)}\mathbf{D}\mathbf{M}_i^{(h+1)}\Pi_{\widetilde{\mathbf{G}}_i}^{\perp})-\mathrm{Tr}(\mathbf{M}_i^{(h+1)}\mathbf{D}\mathbf{M}_i^{(h+1)})=\mathrm{Tr}(\Pi_{\widetilde{\mathbf{G}}_i}^{\perp}\mathbf{M}_i^{(h+1)}\mathbf{D}\mathbf{M}_i^{(h+1)})-\mathrm{Tr}(\mathbf{M}_i^{(h+1)}\mathbf{D}\mathbf{M}_i^{(h+1)})=\mathrm{Tr}((\Pi_{\widetilde{\mathbf{G}}_i}^{\perp}-I)\mathbf{M}_i^{(h+1)}\mathbf{D}\mathbf{M}_i^{(h+1)})=\mathrm{Tr}(\Pi_{\mathbf{G}_i}\mathbf{M}_i^{(h+1)}\mathbf{D}\mathbf{M}_i^{(h+1)})$. Since $\mathrm{rank}(\Pi_{\mathbf{G}_i})\le 2$ and $\left\|\mathbf{M}_i^{(h+1)}\mathbf{D}\mathbf{M}_i^{(h+1)}\right\|_2\le\frac{1}{\alpha}$, we have

$$0\le\mathrm{Tr}(\Pi_{\mathbf{G}_i}\mathbf{M}_i^{(h+1)}\mathbf{D}\mathbf{M}_i^{(h+1)})\le\frac{2}{\alpha}.$$

Now notice that

$$\sum_i\mathbf{b}_i^{(h+1)}(\mathbf{x}^{(1)})\mathbf{b}_i^{(h+1)}(\mathbf{x}^{(2)})\mathrm{Tr}(\Pi_{\mathbf{G}_i}\mathbf{M}_i^{(h+1)}\mathbf{D}\mathbf{M}_i^{(h+1)})=\mathbf{b}^{(h+1)}(\mathbf{x}^{(1)})^\top\mathbf{T}\mathbf{b}^{(h+1)}(\mathbf{x}^{(2)}),$$

where

$$\mathbf{T}=\begin{bmatrix}\mathrm{Tr}(\Pi_{\mathbf{G}_1}\mathbf{M}_1^{(h+1)}\mathbf{D}\mathbf{M}_1^{(h+1)}) & 0 & \dots & 0\\ 0 & \mathrm{Tr}(\Pi_{\mathbf{G}_2}\mathbf{M}_2^{(h+1)}\mathbf{D}\mathbf{M}_2^{(h+1)}) & \dots & 0\\ \vdots & \vdots & & \vdots\\ 0 & 0 & \dots & \mathrm{Tr}(\Pi_{\mathbf{G}_{d_{h+1}}}\mathbf{M}_{d_{h+1}}^{(h+1)}\mathbf{D}\mathbf{M}_{d_{h+1}}^{(h+1)})\end{bmatrix}.$$

Notice that $\|\mathbf{T}\|_2\le\frac{2}{\alpha}$ and thus, $|\mathbf{b}^{(h+1)}(\mathbf{x}^{(1)})^\top\mathbf{T}\mathbf{b}^{(h+1)}(\mathbf{x}^{(2)})|\le\frac{2}{\alpha}\left\|\mathbf{b}^{(h+1)}(\mathbf{x}^{(1)})\right\|_2\left\|\mathbf{b}^{(h+1)}(\mathbf{x}^{(2)})\right\|_2$. Therefore, we have

$$\begin{aligned}
&\mathop{\mathbb{E}}_{\widetilde{\mathbf{W}}^{(h+1)}}\left(\frac{c_\sigma}{d_h}\sum_{i,j}\mathbf{b}_i^{(h+1)}(\mathbf{x}^{(1)})\mathbf{b}_j^{(h+1)}(\mathbf{x}^{(2)})\left(\widetilde{\mathbf{w}}_i^{(h+1)}\right)^\top\Pi_{\widetilde{\mathbf{G}}_i}^{\perp}\mathbf{M}_i^{(h+1)}\mathbf{D}\mathbf{M}_j^{(h+1)}\Pi_{\widetilde{\mathbf{G}}_j}^{\perp}\widetilde{\mathbf{w}}_j^{(h+1)}\right.\\
&\quad\left.-\frac{c_\sigma}{d_h}\sum_{i,j}\mathbf{b}_i^{(h+1)}(\mathbf{x}^{(1)})\mathbf{b}_j^{(h+1)}(\mathbf{x}^{(2)})\left(\widetilde{\mathbf{w}}_i^{(h+1)}\right)^\top\mathbf{M}_i^{(h+1)}\mathbf{D}\mathbf{M}_j^{(h+1)}\widetilde{\mathbf{w}}_j^{(h+1)}\right)\\
&\le\frac{c_\sigma}{d_h}\frac{2}{\alpha}\left\|\mathbf{b}^{(h+1)}(\mathbf{x}^{(1)})\right\|_2\left\|\mathbf{b}^{(h+1)}(\mathbf{x}^{(2)})\right\|_2\le\frac{c_\sigma}{d_h}\frac{8}{\alpha}.
\end{aligned}\tag{19}$$

Next, we analyze concentration of

$$\frac{c_\sigma}{d_h}\sum_{i,j}\mathbf{b}_i^{(h+1)}(\mathbf{x}^{(1)})\mathbf{b}_j^{(h+1)}(\mathbf{x}^{(2)})\left(\widetilde{\mathbf{w}}_i^{(h+1)}\right)^\top\Pi_{\widetilde{\mathbf{G}}_i}^{\perp}\mathbf{M}_i^{(h+1)}\mathbf{D}\mathbf{M}_j^{(h+1)}\Pi_{\widetilde{\mathbf{G}}_j}^{\perp}\widetilde{\mathbf{w}}_j^{(h+1)}.$$

Since the following new random vector has multivariate Gaussian distribution, we can write

$$\left[\sum_{i=1}^{d_{h+1}}\mathbf{b}_i^{(h+1)}(\mathbf{x}^{(1)})((\widetilde{\mathbf{w}}_i^{(h+1)})^\top\Pi_{\widetilde{\mathbf{G}}_i}^{\perp})\odot\mathbf{m}_i^{(h+1)}\quad\sum_{i=1}^{d_{h+1}}\mathbf{b}_i^{(h+1)}(\mathbf{x}^{(2)})((\widetilde{\mathbf{w}}_i^{(h+1)})^\top\Pi_{\widetilde{\mathbf{G}}_i}^{\perp})\odot\mathbf{m}_i^{(h+1)}\right]^\top\overset{D}{=}\mathbf{M}\boldsymbol{\xi},$$

where $\boldsymbol{\xi} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_{2d_h})$, and $\mathbf{M} \in \mathbb{R}^{2d_h \times 2d_h}$ and its covariance matrix is given by a blocked symmetric matrix

$$\mathbf{C} = \begin{bmatrix} \mathbf{C}(\mathbf{x}^{(1)}, \mathbf{x}^{(1)}) & \mathbf{C}(\mathbf{x}^{(1)}, \mathbf{x}^{(2)}) \\ \mathbf{C}(\mathbf{x}^{(1)}, \mathbf{x}^{(2)}) & \mathbf{C}(\mathbf{x}^{(2)}, \mathbf{x}^{(2)}) \end{bmatrix} = \mathbf{M}\mathbf{M}^\top,$$

where each block is given by

$$\begin{aligned}
&\mathbf{C}(\mathbf{x}^{(p)}, \mathbf{x}^{(q)}) \\
&= \mathop{\mathbb{E}}_{\widetilde{\mathbf{W}}^{(h+1)}} \left(\sum_{i=1}^{d_{h+1}} \mathbf{b}_i^{(h+1)}(\mathbf{x}^{(p)}) ((\widetilde{\mathbf{w}}_i^{(h+1)})^\top \Pi_{\mathbf{G}_i}^\perp) \odot \mathbf{m}_i^{(h+1)} \right)^\top \left(\sum_{j=1}^{d_{h+1}} \mathbf{b}_j^{(h+1)}(\mathbf{x}^{(q)}) ((\widetilde{\mathbf{w}}_j^{(h+1)})^\top \Pi_{\mathbf{G}_j}^\perp) \odot \mathbf{m}_j^{(h+1)} \right) \\
&= \mathop{\mathbb{E}}_{\widetilde{\mathbf{W}}^{(h+1)}} \left(\sum_{i=1}^{d_{h+1}} \mathbf{b}_i^{(h+1)}(\mathbf{x}^{(p)}) (\Pi_{\mathbf{G}_i}^\perp \widetilde{\mathbf{w}}_i^{(h+1)}) \odot \mathbf{m}_i^{(h+1)} \right) \left(\sum_{j=1}^{d_{h+1}} \mathbf{b}_j^{(h+1)}(\mathbf{x}^{(q)}) ((\widetilde{\mathbf{w}}_j^{(h+1)})^\top \Pi_{\mathbf{G}_j}^\perp) \odot \mathbf{m}_j^{(h+1)} \right) \\
&= \mathop{\mathbb{E}}_{\widetilde{\mathbf{W}}^{(h+1)}} \left(\sum_{i=1}^{d_{h+1}} \sum_{j=1}^{d_{h+1}} \mathbf{b}_i^{(h+1)}(\mathbf{x}^{(p)}) \mathbf{b}_j^{(h+1)}(\mathbf{x}^{(q)}) \Pi_{\mathbf{G}_i}^\perp (\widetilde{\mathbf{w}}_i^{(h+1)} \odot \mathbf{m}_i^{(h+1)}) (\widetilde{\mathbf{w}}_j^{(h+1)} \odot \mathbf{m}_j^{(h+1)})^\top \Pi_{\mathbf{G}_j}^\perp \right) \\
&= \sum_{i=1}^{d_{h+1}} \mathbf{b}_i^{(h+1)}(\mathbf{x}^{(p)}) \mathbf{b}_i^{(h+1)}(\mathbf{x}^{(q)}) \Pi_{\mathbf{G}_i}^\perp \left(\mathop{\mathbb{E}}_{\widetilde{\mathbf{W}}^{(h+1)}} (\widetilde{\mathbf{w}}_i^{(h+1)} \odot \mathbf{m}_i^{(h+1)}) (\widetilde{\mathbf{w}}_i^{(h+1)} \odot \mathbf{m}_i^{(h+1)})^\top \right) \Pi_{\mathbf{G}_i}^\perp \\
&= \sum_{i=1}^{d_{h+1}} \mathbf{b}_i^{(h+1)}(\mathbf{x}^{(p)}) \mathbf{b}_i^{(h+1)}(\mathbf{x}^{(q)}) \Pi_{\mathbf{G}_i}^\perp \mathrm{diag} \left(\left(\mathbf{m}_i^{(h+1)} \right)^2 \right) \Pi_{\mathbf{G}_i}^\perp,
\end{aligned}$$

where the third equality is from Proposition 9.19 and the square on a vector in the last equality is applied element-wise. Therefore, we can write

$$\begin{aligned}
\mathbf{C} &= \begin{bmatrix} \mathbf{C}(\mathbf{x}^{(1)}, \mathbf{x}^{(1)}) & \mathbf{C}(\mathbf{x}^{(1)}, \mathbf{x}^{(2)}) \\ \mathbf{C}(\mathbf{x}^{(1)}, \mathbf{x}^{(2)}) & \mathbf{C}(\mathbf{x}^{(2)}, \mathbf{x}^{(2)}) \end{bmatrix} \\
&= \sum_{i=1}^{d_{h+1}} \begin{bmatrix} \mathbf{b}_i^{(h+1)}(\mathbf{x}^{(1)}) \mathbf{b}_i^{(h+1)}(\mathbf{x}^{(1)}) & \mathbf{b}_i^{(h+1)}(\mathbf{x}^{(1)}) \mathbf{b}_i^{(h+1)}(\mathbf{x}^{(2)}) \\ \mathbf{b}_i^{(h+1)}(\mathbf{x}^{(1)}) \mathbf{b}_i^{(h+1)}(\mathbf{x}^{(2)}) & \mathbf{b}_i^{(h+1)}(\mathbf{x}^{(2)}) \mathbf{b}_i^{(h+1)}(\mathbf{x}^{(2)}) \end{bmatrix} \otimes \Pi_{\mathbf{G}_i}^\perp \mathrm{diag} \left(\left(\mathbf{m}_i^{(h+1)} \right)^2 \right) \Pi_{\mathbf{G}_i}^\perp.
\end{aligned}$$

Bounding the Operator Norm of the Covariance Matrix $\mathbf{C}$

Next, we want to show that

$$\sum_{i=1}^{d_{h+1}} \begin{bmatrix} \mathbf{b}_i^{(h+1)}(\mathbf{x}^{(1)}) \mathbf{b}_i^{(h+1)}(\mathbf{x}^{(1)}) & \mathbf{b}_i^{(h+1)}(\mathbf{x}^{(1)}) \mathbf{b}_i^{(h+1)}(\mathbf{x}^{(2)}) \\ \mathbf{b}_i^{(h+1)}(\mathbf{x}^{(1)}) \mathbf{b}_i^{(h+1)}(\mathbf{x}^{(2)}) & \mathbf{b}_i^{(h+1)}(\mathbf{x}^{(2)}) \mathbf{b}_i^{(h+1)}(\mathbf{x}^{(2)}) \end{bmatrix} \otimes \left(\frac{1}{\alpha} \mathbf{I} - \Pi_{\mathbf{G}_i}^\perp \mathrm{diag} \left(\left(\mathbf{m}_i^{(h+1)} \right)^2 \right) \Pi_{\mathbf{G}_i}^\perp \right) \succeq \mathbf{0}. \tag{20}$$

Given this, since Kronecker product preserves two norm we have that

$$\begin{aligned}
\|\mathbf{C}\|_2 &\leq \frac{1}{\alpha} \left\| \sum_{i=1}^{d_{h+1}} \begin{bmatrix} \mathbf{b}_i^{(h+1)}(\mathbf{x}^{(1)}) \mathbf{b}_i^{(h+1)}(\mathbf{x}^{(1)}) & \mathbf{b}_i^{(h+1)}(\mathbf{x}^{(1)}) \mathbf{b}_i^{(h+1)}(\mathbf{x}^{(2)}) \\ \mathbf{b}_i^{(h+1)}(\mathbf{x}^{(1)}) \mathbf{b}_i^{(h+1)}(\mathbf{x}^{(2)}) & \mathbf{b}_i^{(h+1)}(\mathbf{x}^{(2)}) \mathbf{b}_i^{(h+1)}(\mathbf{x}^{(2)}) \end{bmatrix} \right\|_2 \\
&= \frac{1}{\alpha} \left\| \begin{bmatrix} \left\langle \mathbf{b}^{(h+1)}(\mathbf{x}^{(1)}), \mathbf{b}^{(h+1)}(\mathbf{x}^{(1)}) \right\rangle & \left\langle \mathbf{b}^{(h+1)}(\mathbf{x}^{(1)}), \mathbf{b}^{(h+1)}(\mathbf{x}^{(2)}) \right\rangle \\ \left\langle \mathbf{b}^{(h+1)}(\mathbf{x}^{(1)}), \mathbf{b}^{(h+1)}(\mathbf{x}^{(2)}) \right\rangle & \left\langle \mathbf{b}^{(h+1)}(\mathbf{x}^{(2)}), \mathbf{b}^{(h+1)}(\mathbf{x}^{(2)}) \right\rangle \end{bmatrix} \right\|_2 \\
&\leq \frac{1}{\alpha} \sqrt{2} \left(\left\langle \mathbf{b}^{(h+1)}(\mathbf{x}^{(1)}), \mathbf{b}^{(h+1)}(\mathbf{x}^{(1)}) \right\rangle + \left\langle \mathbf{b}^{(h+1)}(\mathbf{x}^{(1)}), \mathbf{b}^{(h+1)}(\mathbf{x}^{(2)}) \right\rangle \right),
\end{aligned}$$

where the last inequality is by applying $\|\mathbf{A}\|_2 \leq \sqrt{m} \|\mathbf{A}\|_\infty$.

We prove the matrix in Equation (20) is positive semi-definite by constructing a multivariate Gaussian distribution such that its covariance matrix is exactly the matrix and exploring the fact that the covariance matrix of two independent Gaussian distribution is the sum of the two covariance matrix. First, notice that

$$\frac{1}{\alpha} \mathbf{I} - \Pi_{\mathbf{G}_i}^\perp \mathrm{diag} \left(\left(\mathbf{m}_i^{(h+1)} \right)^2 \right) \Pi_{\mathbf{G}_i}^\perp = \frac{1}{\alpha} \left(\Pi_{\mathbf{G}_i}^\perp + \Pi_{\mathbf{G}_i} \right) - \Pi_{\mathbf{G}_i}^\perp \mathrm{diag} \left(\left(\mathbf{m}_i^{(h+1)} \right)^2 \right) \Pi_{\mathbf{G}_i}^\perp$$

$$
\begin{aligned}
&= \Pi_{\mathbf{G}_i}^{\perp}\left(\frac{1}{\alpha}\mathbf{I} - \operatorname{diag}\left(\left(\mathbf{m}_i^{(h+1)}\right)^2\right)\Pi_{\mathbf{G}_i}^{\perp}\right) + \frac{1}{\alpha}\Pi_{\mathbf{G}_i} \\
&= \Pi_{\mathbf{G}_i}^{\perp}\left(\frac{1}{\alpha}\left(\Pi_{\mathbf{G}_i}^{\perp} + \Pi_{\mathbf{G}_i}\right) - \operatorname{diag}\left(\left(\mathbf{m}_i^{(h+1)}\right)^2\right)\Pi_{\mathbf{G}_i}^{\perp}\right) + \frac{1}{\alpha}\Pi_{\mathbf{G}_i} \\
&= \Pi_{\mathbf{G}_i}^{\perp}\left(\frac{1}{\alpha}\Pi_{\mathbf{G}_i} + \left(\frac{1}{\alpha}\mathbf{I} - \operatorname{diag}\left(\left(\mathbf{m}_i^{(h+1)}\right)^2\right)\right)\Pi_{\mathbf{G}_i}^{\perp}\right) + \frac{1}{\alpha}\Pi_{\mathbf{G}_i} \\
&= \Pi_{\mathbf{G}_i}^{\perp}\left(\frac{1}{\alpha}\mathbf{I} - \operatorname{diag}\left(\left(\mathbf{m}_i^{(h+1)}\right)^2\right)\right)\Pi_{\mathbf{G}_i}^{\perp} + \frac{1}{\alpha}\Pi_{\mathbf{G}_i}.
\end{aligned}
$$

The final Gaussian is constructed by the sum of the following two groups of Gaussian: let $\mathbf{W}_1, \mathbf{W}_2$ be two independent standard Gaussian matrices,

$$
\left[\sum_{i=1}^{d_{h+1}} \mathbf{b}_i(\mathbf{x}^{(1)})(\mathbf{w}_{1,i}^{(h+1)})^\top\left(\tfrac{1}{\sqrt{\alpha}}\mathbf{I} - \operatorname{diag}\left(\mathbf{m}_i^{(h+1)}\right)\right)\Pi_{\mathbf{G}_i}^{\perp} \quad \sum_{i=1}^{d_{h+1}} \mathbf{b}_i(\mathbf{x}^{(2)})(\mathbf{w}_{1,i}^{(h+1)})^\top\left(\tfrac{1}{\sqrt{\alpha}}\mathbf{I} - \operatorname{diag}\left(\mathbf{m}_i^{(h+1)}\right)\right)\Pi_{\mathbf{G}_i}^{\perp}\right],
$$
$$
\left[\sum_{i=1}^{d_{h+1}} \mathbf{b}_i(\mathbf{x}^{(1)})(\mathbf{w}_{2,i}^{(h+1)})^\top\tfrac{1}{\sqrt{\alpha}}\Pi_{\mathbf{G}_i} \quad \sum_{i=1}^{d_{h+1}} \mathbf{b}_i(\mathbf{x}^{(2)})(\mathbf{w}_{2,i}^{(h+1)})^\top\tfrac{1}{\sqrt{\alpha}}\Pi_{\mathbf{G}_i}\right],
$$

where $\mathbf{w}_{i,j}$ denote the j-th row of $\mathbf{W}_i$.

Now conditioned on $\{\mathbf{b}^{(h+1)}(\mathbf{x}^{(1)}), \mathbf{b}^{(h+1)}(\mathbf{x}^{(2)}), \mathbf{g}^{(h)}(\mathbf{x}^{(1)}), \mathbf{g}^{(h)}(\mathbf{x}^{(2)})\}$, we have

$$
\begin{aligned}
&\left(\sum_{i=1}^{d_{h+1}} \mathbf{b}_i^{(h+1)}(\mathbf{x}^{(1)})(\mathbf{w}_i^{(h+1)} \odot \mathbf{m}_i^{(h+1)})^\top\Pi_{\mathbf{G}_i}^{\perp}\right)\mathbf{D}\left(\sum_{i=1}^{d_{h+1}} \mathbf{b}_i^{(h+1)}(\mathbf{x}^{(2)})(\mathbf{w}_i^{(h+1)} \odot \mathbf{m}_i^{(h+1)})^\top\Pi_{\mathbf{G}_i}^{\perp}\right) \\
&\overset{D}{=} \left(\sum_{i=1}^{d_{h+1}} \mathbf{b}_i^{(h+1)}(\mathbf{x}^{(1)})(\widetilde{\mathbf{w}}_i^{(h+1)} \odot \mathbf{m}_i^{(h+1)})^\top\Pi_{\mathbf{G}_i}^{\perp}\right)\mathbf{D}\left(\sum_{i=1}^{d_{h+1}} \mathbf{b}_i^{(h+1)}(\mathbf{x}^{(2)})(\widetilde{\mathbf{w}}_i^{(h+1)} \odot \mathbf{m}_i^{(h+1)})^\top\Pi_{\mathbf{G}_i}^{\perp}\right) \\
&\overset{D}{=} ([\mathbf{I}_{d_h} \quad \mathbf{0}]\mathbf{M}\boldsymbol{\xi})^\top\mathbf{D}([\mathbf{0} \quad \mathbf{I}_{d_h}]\mathbf{M}\boldsymbol{\xi}) \\
&\overset{D}{=} \frac{1}{2}\boldsymbol{\xi}^\top\mathbf{M}^\top\begin{bmatrix}\mathbf{0} & \mathbf{D} \\ \mathbf{D} & \mathbf{0}\end{bmatrix}\mathbf{M}\boldsymbol{\xi}.
\end{aligned}
$$

Now, let

$$
\mathbf{A} = \frac{1}{2}\mathbf{M}^\top\begin{bmatrix}\mathbf{0} & \mathbf{D} \\ \mathbf{D} & \mathbf{0}\end{bmatrix}\mathbf{M},
$$

and we have

$$
\begin{aligned}
\|\mathbf{A}\|_2 &\le \frac{1}{2}\|\mathbf{M}\|_2^2\|\mathbf{D}\|_2 \\
&= \frac{1}{2}\left\|\mathbf{M}\mathbf{M}^\top\right\|_2\|\mathbf{D}\|_2 \\
&= \frac{1}{2}\|\mathbf{C}\|_2 \\
&\le \frac{1}{2\alpha}\sqrt{2}\left(\left\langle\mathbf{b}^{(h+1)}(\mathbf{x}^{(1)}), \mathbf{b}^{(h+1)}(\mathbf{x}^{(1)})\right\rangle + \left\langle\mathbf{b}^{(h+1)}(\mathbf{x}^{(1)}), \mathbf{b}^{(h+1)}(\mathbf{x}^{(2)})\right\rangle\right) \\
&\le \frac{2\sqrt{2}}{\alpha}.
\end{aligned}
$$

Bounding the Trace of the Covariance Matrix C

Naively apply 2-norm-Frobenius-norm bound for matrices will give us

$$
\|\mathbf{A}\|_F \le \sqrt{2d_h}\|\mathbf{A}\|_2 \le \frac{4\sqrt{d_h}}{\alpha}.
$$

We prove a better bound. Observe that

$$
\frac{1}{d_h}\|\mathbf{A}\|_F = \frac{1}{d_h}\left\|\frac{1}{2}\mathbf{M}^\top\begin{bmatrix}\mathbf{0} & \mathbf{D} \\ \mathbf{D} & \mathbf{0}\end{bmatrix}\mathbf{M}\right\|_F \le \frac{1}{2d_h}\|\mathbf{M}\|_2\|\mathbf{M}\|_F\|\mathbf{D}\|_2 = \frac{1}{2d_h\sqrt{\alpha}}\|\mathbf{M}\|_F = \frac{1}{2d_h\sqrt{\alpha}}\sqrt{\operatorname{Tr}(\mathbf{M}\mathbf{M}^\top)}.
$$

Using the similar idea from bounding the 2-norm of $\mathbf{C} = \mathbf{M}\mathbf{M}^\top$, we want to show that

$$\begin{aligned}
&\sum_{i=1}^{d_{h+1}} \begin{bmatrix} \mathbf{b}_i^{(h+1)}(\mathbf{x}^{(1)})\mathbf{b}_i^{(h+1)}(\mathbf{x}^{(1)}) & \mathbf{b}_i^{(h+1)}(\mathbf{x}^{(1)})\mathbf{b}_i^{(h+1)}(\mathbf{x}^{(2)}) \\ \mathbf{b}_i^{(h+1)}(\mathbf{x}^{(1)})\mathbf{b}_i^{(h+1)}(\mathbf{x}^{(2)}) & \mathbf{b}_i^{(h+1)}(\mathbf{x}^{(2)})\mathbf{b}_i^{(h+1)}(\mathbf{x}^{(2)}) \end{bmatrix} \otimes \left(\left(\mathbf{M}_i^{(h+1)}\right)^2 - \Pi_{\mathbf{G}_i}^{\perp} \left(\mathbf{M}_i^{(h+1)}\right)^2 \Pi_{\mathbf{G}_i}^{\perp} \right) \\
&= \sum_{i=1}^{d_{h+1}} \begin{bmatrix} \mathbf{b}_i^{(h+1)}(\mathbf{x}^{(1)})\mathbf{b}_i^{(h+1)}(\mathbf{x}^{(1)}) & \mathbf{b}_i^{(h+1)}(\mathbf{x}^{(1)})\mathbf{b}_i^{(h+1)}(\mathbf{x}^{(2)}) \\ \mathbf{b}_i^{(h+1)}(\mathbf{x}^{(1)})\mathbf{b}_i^{(h+1)}(\mathbf{x}^{(2)}) & \mathbf{b}_i^{(h+1)}(\mathbf{x}^{(2)})\mathbf{b}_i^{(h+1)}(\mathbf{x}^{(2)}) \end{bmatrix} \otimes \Pi_{\mathbf{G}_i} \left(\mathbf{M}_i^{(h+1)}\right)^2 \Pi_{\mathbf{G}_i} \succeq \mathbf{0}. \qquad (21)
\end{aligned}$$

If this equation is true, then we have

$$\begin{aligned}
\frac{1}{d_h}\mathrm{Tr}(\mathbf{M}\mathbf{M}^\top) &\leq \frac{1}{d_h}\mathrm{Tr}\left(\sum_{i=1}^{d_{h+1}} \begin{bmatrix} \mathbf{b}_i^{(h+1)}(\mathbf{x}^{(1)})\mathbf{b}_i^{(h+1)}(\mathbf{x}^{(1)}) & \mathbf{b}_i^{(h+1)}(\mathbf{x}^{(1)})\mathbf{b}_i^{(h+1)}(\mathbf{x}^{(2)}) \\ \mathbf{b}_i^{(h+1)}(\mathbf{x}^{(1)})\mathbf{b}_i^{(h+1)}(\mathbf{x}^{(2)}) & \mathbf{b}_i^{(h+1)}(\mathbf{x}^{(2)})\mathbf{b}_i^{(h+1)}(\mathbf{x}^{(2)}) \end{bmatrix} \otimes \mathbf{M}_i^2 \right) \\
&= \frac{1}{d_h} \sum_{i=1}^{d_{h+1}} \left[\left(\mathbf{b}_i^{(h+1)}(\mathbf{x}^{(1)})\right)^2 + \left(\mathbf{b}_i^{(h+1)}(\mathbf{x}^{(2)})\right)^2 \right] \mathrm{Tr}\left(\left(\mathbf{M}_i^{(h+1)}\right)^2 \right) \\
&\leq \frac{1}{d_h} \sum_{i=1}^{d_{h+1}} \left[\left(\mathbf{b}_i^{(h+1)}(\mathbf{x}^{(1)})\right)^2 + \left(\mathbf{b}_i^{(h+1)}(\mathbf{x}^{(2)})\right)^2 \right] \max_i \mathrm{Tr}\left(\left(\mathbf{M}_i^{(h+1)}\right)^2 \right) \\
&= \max_i \frac{1}{d_h}\mathrm{Tr}\left(\left(\mathbf{M}_i^{(h+1)}\right)^2 \right) \left(\left\|\mathbf{b}^{(h+1)}(\mathbf{x}^{(1)})\right\|_2^2 + \left\|\mathbf{b}^{(h+1)}(\mathbf{x}^{(2)})\right\|_2^2 \right) \\
&\leq 4 \max_i \frac{1}{d_h}\mathrm{Tr}\left(\left(\mathbf{M}_i^{(h+1)}\right)^2 \right).
\end{aligned}$$

By Lemma 9.9, with probability $\geq 1-\delta$, $\max_i \frac{1}{d_h}\mathrm{Tr}\left(\left(\mathbf{M}_i^{(h+1)}\right)^2 \right) \leq 1+1 = 2$. Thus, we have

$$\frac{1}{d_h}\left\|\mathbf{A}\right\|_F \leq \sqrt{\frac{2}{d_h \alpha}}.$$

To prove Equation (21), since $\mathbf{M}_i$ commutes with $\Pi_{\mathbf{G}_i}^{\perp}$, we have

$$\mathbf{M}_i^2 - \Pi_{\mathbf{G}_i}^{\perp}\mathbf{M}_i^2\Pi_{\mathbf{G}_i}^{\perp} = \mathbf{M}_i^2 - \mathbf{M}_i^2\Pi_{\mathbf{G}_i}^{\perp} = \mathbf{M}_i^2\Pi_{\mathbf{G}_i} = \Pi_{\mathbf{G}_i}\mathbf{M}_i^2\Pi_{\mathbf{G}_i}.$$

The Gaussian vector given by

$$\left[\sum_{i=1}^{d_{h+1}} \mathbf{b}_i(\mathbf{x}^{(1)})(\mathbf{w}_{2,i}^{(h+1)})^\top \mathbf{M}_i \Pi_{\mathbf{G}_i} \quad \sum_{i=1}^{d_{h+1}} \mathbf{b}_i(\mathbf{x}^{(2)})(\mathbf{w}_{2,i}^{(h+1)})^\top \mathbf{M}_i \Pi_{\mathbf{G}_i} \right]$$

has the covariance matrix.

Now apply Gaussian chaos concentration bound (Lemma 9.4), we have with probability $1 - \frac{\delta_2}{6}$,

$$\begin{aligned}
\frac{1}{d_h}|\boldsymbol{\xi}^\top\mathbf{A}\boldsymbol{\xi} - \mathbb{E}[\boldsymbol{\xi}^\top\mathbf{A}\boldsymbol{\xi}]| &\leq \frac{1}{d_h}\left(2\left\|\mathbf{A}\right\|_F \sqrt{\log\frac{6}{\delta_2}} + 2\left\|\mathbf{A}\right\|_2 \log\frac{6}{\delta_2} \right) \\
&\leq \sqrt{\frac{8\log\frac{6}{\delta_2}}{\alpha d_h}} + 4\sqrt{2}\frac{\log\frac{6}{\delta_2}}{\alpha d_h}. \qquad (22)
\end{aligned}$$

Finally, combining Equation 19 and Equation 22, we have

$$\begin{aligned}
&\left| \frac{c_\sigma}{d_h} \sum_{i,j} \mathbf{b}_i^{(h+1)}(\mathbf{x}^{(1)})\mathbf{b}_j^{(h+1)}(\mathbf{x}^{(2)}) \left(\widetilde{\mathbf{w}}_i^{(h+1)}\right)^\top \mathbf{M}_i^{(h+1)} \Pi_{\mathbf{G}_i}^{\perp} \mathbf{D} \Pi_{\mathbf{G}_j}^{\perp} \mathbf{M}_j^{(h+1)} \widetilde{\mathbf{w}}_j^{(h+1)} \right. \\
&\quad \left. - \frac{2}{d_h} \sum_i \mathbf{b}_i^{(h+1)}(\mathbf{x}^{(1)})\mathbf{b}_i^{(h+1)}(\mathbf{x}^{(2)}) \mathrm{Tr}(\mathbf{M}_i^{(h+1)}\mathbf{D}\mathbf{M}_i^{(h+1)}) \right| \\
&\leq \frac{2}{d_h}|\boldsymbol{\xi}^\top\mathbf{A}\boldsymbol{\xi} - \mathbb{E}[\boldsymbol{\xi}^\top\mathbf{A}\boldsymbol{\xi}]| + \left| \frac{2}{d_h}\mathbb{E}\left[\boldsymbol{\xi}^\top\mathbf{A}\boldsymbol{\xi}\right] - \frac{2}{d_h}\sum_i \mathbf{b}_i^{(h+1)}(\mathbf{x}^{(1)})\mathbf{b}_i^{(h+1)}(\mathbf{x}^{(2)}) \mathrm{Tr}(\mathbf{M}_i^{(h+1)}\mathbf{D}\mathbf{M}_i^{(h+1)}) \right|
\end{aligned}$$

$$\leq \frac{c_\sigma}{d_h}\frac{8}{\alpha} + \sqrt{\frac{8\log\frac{6}{\delta_2}}{\alpha d_h}} + 4\sqrt{2}\frac{\log\frac{6}{\delta_2}}{\alpha d_h} \leq 3\sqrt{\frac{8\log\frac{6}{\delta_2}}{\alpha d_h}}.$$

where we choose $d_h \geq \frac{8}{\alpha}\log\frac{6}{\delta_2}$. Then take a union bound over $(\mathbf{x},\mathbf{x}),(\mathbf{x},\mathbf{x}'),(\mathbf{x}',\mathbf{x}')$. Finally, taking $\mathbf{x}^{(1)} = \mathbf{x}^{(2)}$, we have

$$\begin{aligned}
&\left\|\sqrt{\frac{c_\sigma}{d_h}}\sum_i \mathbf{b}_i^{(h+1)}(\mathbf{x}^{(1)})\left(\widetilde{\mathbf{w}}_i^{(h+1)}\right)^\top \mathbf{M}_i^{(h+1)}\Pi_{\mathbf{G}_i}^{\perp}\mathbf{D}\right\|_2 \\
&\leq \sqrt{\left|\frac{c_\sigma}{d_h}\sum_{i,j}\mathbf{b}_i^{(h+1)}(\mathbf{x}^{(1)})\left(\widetilde{\mathbf{w}}_i^{(h+1)}\right)^\top \mathbf{M}_i^{(h+1)}\Pi_{\mathbf{G}_i}^{\perp}\mathbf{D}\Pi_{\mathbf{G}_j}^{\perp}\mathbf{M}_j^{(h+1)}\widetilde{\mathbf{w}}_j^{(h+1)}\mathbf{b}_j^{(h+1)}(\mathbf{x}^{(1)})\right|} \\
&\leq \sqrt{\frac{2}{d_h}\sum_i \mathbf{b}_i^{(h+1)}(\mathbf{x}^{(1)})\mathbf{b}_i^{(h+1)}(\mathbf{x}^{(2)})\mathrm{Tr}(\mathbf{M}_i^{(h+1)}\mathbf{D}\mathbf{M}_i^{(h+1)}) + 3\sqrt{\frac{8\log\frac{6}{\delta_2}}{\alpha d_h}}} \\
&\leq \sqrt{4 + 3\sqrt{\frac{8\log\frac{6}{\delta_2}}{\alpha d_h}}} \leq 6.
\end{aligned}$$

□

9.6.2 Proof of Lemma 9.13: Bounding Pseudo Networks' Output

This is the most involving part of the proof. To facilitate the proof, we first introduce a special property of the standard Gaussian vector.

Proposition 9.21. *For any given nonzero vectors* $\mathbf{x},\mathbf{y}$*, the distribution of* $(\mathbf{w}^\top\mathbf{x})^2\mathbb{I}(\mathbf{w}^\top\mathbf{y} > 0)$ *is the same as* $(\mathbf{w}^\top\mathbf{x})^2\mathbb{I}(\mathbf{w}^\top\mathbf{x} > 0)$ *where* $\mathbf{w}\sim\mathcal{N}(\mathbf{0},\mathbf{I})$.

Proof. Define random variables $z_1 = \left(\mathbf{w}^\top\mathbf{x}\right)^2\mathbb{I}(\mathbf{w}^\top\mathbf{y} > 0)$ and $z_2 = \left(\mathbf{w}^\top\mathbf{x}\right)^2\mathbb{I}(\mathbf{w}^\top\mathbf{x} > 0)$. Let F_1, F_2 be the cumulative distribution function of z_1, z_2. It is easy to see that both z_1 and z_2 has probability $1/2$ of being zero and thus we consider the probability that z_1 and z_2 are not identically zero. Then for $z > 0$,

$$\begin{aligned}
\mathbb{P}[0 < z_1 \leq z] &= \int_{\{\mathbf{w}:\mathbf{w}^\top\mathbf{y}>0,|\mathbf{w}^\top\mathbf{x}|\leq\sqrt{z}\}} \frac{1}{(2\pi)^{k/2}}e^{-\frac{1}{2}\|\mathbf{w}\|_2^2}\,d\mathbf{w} \\
&= \int_{\{\mathbf{w}:\mathbf{w}^\top\mathbf{x}>0,\mathbf{w}^\top\mathbf{y}>0,|\mathbf{w}^\top\mathbf{x}|\leq\sqrt{z}\}\cup\{\mathbf{w}:\mathbf{w}^\top\mathbf{x}\leq 0,\mathbf{w}^\top\mathbf{y}>0,|\mathbf{w}^\top\mathbf{x}|\leq\sqrt{z}\}} \frac{1}{(2\pi)^{k/2}}e^{-\frac{1}{2}\|\mathbf{w}\|_2^2}\,d\mathbf{w} \\
&= \int_{\{\mathbf{w}:\mathbf{w}^\top\mathbf{x}>0,\mathbf{w}^\top\mathbf{y}>0,|\mathbf{w}^\top\mathbf{x}|\leq\sqrt{z}\}} \frac{1}{(2\pi)^{k/2}}e^{-\frac{1}{2}\|\mathbf{w}\|_2^2}\,d\mathbf{w} \\
&\quad + \int_{\{\mathbf{w}:\mathbf{w}^\top\mathbf{x}\leq 0,\mathbf{w}^\top\mathbf{y}>0,|\mathbf{w}^\top\mathbf{x}|\leq\sqrt{z}\}} \frac{1}{(2\pi)^{k/2}}e^{-\frac{1}{2}\|\mathbf{w}\|_2^2}\,d\mathbf{w} \\
&= \int_{\{\mathbf{w}:\mathbf{w}^\top\mathbf{x}>0,\mathbf{w}^\top\mathbf{y}>0,|\mathbf{w}^\top\mathbf{x}|\leq\sqrt{z}\}} \frac{1}{(2\pi)^{k/2}}e^{-\frac{1}{2}\|\mathbf{w}\|_2^2}\,d\mathbf{w} \\
&\quad + \int_{\{\mathbf{w}:\mathbf{w}^\top\mathbf{x}>0,\mathbf{w}^\top\mathbf{y}\leq 0,|\mathbf{w}^\top\mathbf{x}|\leq\sqrt{z}\}} \frac{1}{(2\pi)^{k/2}}e^{-\frac{1}{2}\|\mathbf{w}\|_2^2}\,d\mathbf{w} \\
&= \int_{\{\mathbf{w}:\mathbf{w}^\top\mathbf{x}>0,|\mathbf{w}^\top\mathbf{x}|\leq\sqrt{z}\}} \frac{1}{(2\pi)^{k/2}}e^{-\frac{1}{2}\|\mathbf{w}\|_2^2}\,d\mathbf{w} \\
&= \mathbb{P}[0 < z_2 \leq z],
\end{aligned}$$

where the third last equality is by spherical symmetry of Gaussian and take $\mathbf{w} := -\mathbf{w}$ over the region. □

Mask-Induced Pseudo-Network

It turns out that the term in Equation (24) is closely related to a network structure which we defined as follows.

Definition 9.22 (Pseudo-network induced by mask)**.** *Define the pseudo-network induced by the h-th layer j-th column of sparse masks $\mathbf{m}^{(h)}$ denoted by $\mathbf{m}_{\cdot j}^{(h)}$ for all $h \in \{2,\dots,L\}$, $j \in [d_{h-1}]$ and $h' \in \{h+1, h+2, \dots, L\}$ to be*

$$\begin{aligned}
\mathbf{g}^{(h,j,h)}(\mathbf{x}) &= \sqrt{\frac{c_\sigma}{d_h}}\mathbf{D}^{(h)}(\mathbf{x})\mathrm{diag}_i\left(\frac{\mathbf{m}_{ij}^{(h)}\sqrt{\alpha}}{\left\|\mathbf{g}^{(h-1)}\odot\mathbf{m}_i^{(h)}\right\|_2^2}\right)\mathbf{f}^{(h)}(\mathbf{x}),\\
\mathbf{f}^{(h,j,h')}(\mathbf{x}) &= \left(\mathbf{W}^{(h')}\odot\mathbf{m}^{(h')}\right)\mathbf{g}^{(h,j,h'-1)}(\mathbf{x}),\\
\mathbf{g}^{(h,j,h')}(\mathbf{x}) &= \sqrt{\frac{c_\sigma}{d_{h'}}}\mathbf{D}^{(h')}(\mathbf{x})\mathbf{f}^{(h,j,h')}(\mathbf{x}).
\end{aligned}$$

where $f^{(h,j,L+1)}(\mathbf{x})$ is the output of the pseudo-network.

We would like to bound $|f^{(h+1,j,L+1)}(\mathbf{x})|$ for all $h \in \{2,\dots,L\}$, $j \in [d_{h-1}]$. Observe that without the diagonal matrix in $\mathbf{g}^{(h,j,h)}(\mathbf{x})$ we have

$$\mathbf{g}^{(h+1)}(\mathbf{x}) = \sqrt{\frac{c_\sigma}{d_{h+1}}}\mathbf{D}^{(h+1)}(\mathbf{x})\mathbf{f}^{(h+1)}(\mathbf{x}).$$

Conditioned on $\mathbf{g}^{(h+1,j,L)}(\mathbf{x})$, $f^{(h+1,j,L+1)}(\mathbf{x})$ has distribution $\mathcal{N}(0, \left\|\mathbf{g}^{(h+1,j,L)}(\mathbf{x})\odot\mathbf{m}^{(L+1)}\right\|_2^2)$. Therefore, the magnitude of $|f^{(h+1,j,L+1)}(\mathbf{x})|$ would depend on $\left\|\mathbf{g}^{(h+1,j,L)}(\mathbf{x})\odot\mathbf{m}^{(L+1)}\right\|_2$.

Definition 9.23. *Define the event*

$$\begin{aligned}
&\mathcal{C}_1(\epsilon) = \left\{\left|\left\|\mathbf{g}^{(h,j,h')}\right\|_2^2 - \mathbb{E}\left\|\mathbf{g}^{(h,j,h')}\right\|_2^2\right| < \epsilon, \quad \forall h\in\{2,\dots,L\}, j\in[d_{h-1}], h'\in\{h+1,h+2,\dots,L\}\right\},\\
&\mathcal{C}_2\left(\mathbf{x}, 2\sqrt{\log\frac{4\sum_{h'=1}^{L-1}d_{h'}}{\delta}}\right)\\
&\quad= \left\{|f^{(h,j,L+1)}(\mathbf{x})| < 2\sqrt{\log\frac{4\sum_{h'=1}^{L-1}d_{h'}}{\delta}},\ \forall h\in\{2,\dots,L\}, j\in[d_{h-1}], h'\in\{h+1,\dots,L\}\right\},\\
&\overline{\mathcal{C}}\left(2\sqrt{\log\frac{4\sum_{h'=1}^{L-1}d_{h'}}{\delta}}\right) = \mathcal{C}_1(\epsilon)\cap\mathcal{C}_2\left(\mathbf{x}, 2\sqrt{\log\frac{4\sum_{h'=1}^{L-1}d_{h'}}{\delta}}\right)\cap\mathcal{C}_2\left(\mathbf{x}', 2\sqrt{\log\frac{4\sum_{h'=1}^{L-1}d_{h'}}{\delta}}\right).
\end{aligned}$$

We are going to show that the event $\overline{\mathcal{C}}$ holds with probability $1-\delta$.

First, we show that

Lemma 9.24. *Assume $\overline{\mathcal{A}}(\epsilon_1)$ holds for $\epsilon_1 < 1/2$. For all $h \in \{2,\dots,L\}$, $j \in [d_{h-1}]$, it holds that for all $h' \in \{h+1, h+2, h+3, \dots, L\}$,*

$$\mathop{\mathbb{E}}_{\mathbf{W}^{(h+1)},\mathbf{m}^{(h+1)}\dots,\mathbf{W}^{(h')},\mathbf{m}^{(h')}}\left[\left\|\mathbf{g}^{(h,j,h')}(\mathbf{x})\right\|_2^2\middle|\mathbf{g}^{(h,j,h)}\right] \le 2\mathop{\mathbb{E}}_{\mathbf{W}^{(h+1)},\mathbf{m}^{(h+1)},\dots,\mathbf{W}^{(h')},\mathbf{m}^{(h')}}\left[\left\|\mathbf{g}^{(h')}(\mathbf{x})\right\|_2^2\middle|\mathbf{g}^{(h)}(\mathbf{x})\right].$$

Proof. By Proposition 9.21, for two non-zero vectors $\mathbf{x}, \mathbf{y}$ we have

$$\mathop{\mathbb{E}}_{\mathbf{w}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}\left[\left(\mathbf{w}^\top\mathbf{x}\right)^2\mathbb{I}(\mathbf{w}^\top\mathbf{y}>0)\right] = \mathop{\mathbb{E}}_{\mathbf{w}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}\left[\left(\mathbf{w}^\top\mathbf{x}\right)^2\mathbb{I}(\mathbf{w}^\top\mathbf{x}>0)\right].$$

This equation tells us that the direction of $\mathbf{y}$ doesn't matter which implies

$$\mathop{\mathbb{E}}_{\mathbf{w}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}\left[\left(\mathbf{w}^\top\mathbf{x}\right)^2\frac{c_\sigma}{d_{h+1}}\dot\sigma\left(\mathbf{w}^\top\mathbf{y}\right)\right] = \mathop{\mathbb{E}}_{\mathbf{w}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}\left[\left(\mathbf{w}^\top\mathbf{x}\right)^2\frac{c_\sigma}{d_{h+1}}\dot\sigma\left(\mathbf{w}^\top\mathbf{x}\right)\right] = \frac{\|\mathbf{x}\|_2^2}{d_{h+1}}.$$

Now, this implies that conditioned on $\mathbf{m}$,

$$\begin{aligned}&\mathop{\mathbb{E}}_{\mathbf{w}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}\left[\left((\mathbf{w}\odot\mathbf{m})^\top\mathbf{x}\right)^2\frac{c_\sigma}{d_{h+1}}\dot\sigma\left((\mathbf{w}\odot\mathbf{m})^\top\mathbf{y}\right)\right]\\&=\mathop{\mathbb{E}}_{\mathbf{w}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}\left[\left((\mathbf{w}\odot\mathbf{m})^\top\mathbf{x}\right)^2\frac{c_\sigma}{d_{h+1}}\dot\sigma\left((\mathbf{w}\odot\mathbf{m})^\top\mathbf{x}\right)\right]=\frac{\|\mathbf{x}\odot\mathbf{m}\|_2^2}{d_{h+1}}.\end{aligned}\tag{23}$$

Now, we fix h and j and prove the inequality holds for all h'. By Equation (23),

$$\begin{aligned}&\mathop{\mathbb{E}}_{\mathbf{m}^{(h+1)},\mathbf{W}^{(h+1)}}\left[\left\|\mathbf{g}^{(h,j,h+1)}\right\|_2^2\right]\\&=\mathop{\mathbb{E}}_{\mathbf{m}^{(h+1)}}\left[\sum_{i=1}^{d_{h+1}}\mathop{\mathbb{E}}_{\mathbf{w}_i^{(h+1)}}\left[\left(\left(\mathbf{w}_i^{(h+1)}\odot\mathbf{m}_i^{(h+1)}\right)^\top\mathbf{g}^{(h,j,h)}\right)^2\frac{c_\sigma}{d_{h+1}}\dot\sigma\left(\left(\mathbf{w}_i^{(h+1)}\odot\mathbf{m}_i^{(h+1)}\right)^\top\mathbf{g}^{(h)}\right)\middle|\mathbf{m}^{(h+1)}\right]\right]\\&=\left\|\mathbf{g}^{(h,j,h)}\right\|_2^2.\end{aligned}$$

Hence, by iterated expectation, we have for all $h'\in\{h+1,h+2,h+3,\dots,L\}$,

$$\begin{aligned}\mathop{\mathbb{E}}_{\mathbf{W}^{(h+1)},\mathbf{m}^{(h+1)},\dots,\mathbf{W}^{(h')},\mathbf{m}^{(h')}}\left[\left\|\mathbf{g}^{(h,j,h')}(\mathbf{x})\right\|_2^2\middle|\mathbf{g}^{(h,j,h)}\right]&=\left\|\mathbf{g}^{(h,j,h)}\right\|_2^2,\\\mathop{\mathbb{E}}_{\mathbf{W}^{(h+1)},\mathbf{m}^{(h+1)},\dots,\mathbf{W}^{(h')},\mathbf{m}^{(h')}}\left[\left\|\mathbf{g}^{(h')}(\mathbf{x})\right\|_2^2\middle|\mathbf{g}^{(h)}(\mathbf{x})\right]&=\left\|\mathbf{g}^{(h)}(\mathbf{x})\right\|_2^2.\end{aligned}$$

By our assumption $\left\|\mathbf{g}^{(h-1)}\odot\mathbf{m}^{(h)}\right\|_2^2\geq 1-\epsilon_1^2\geq 1/2$, we have $\left\|\mathbf{g}^{(h,j,h)}(\mathbf{x})\right\|_2^2\leq 2\left\|\mathbf{g}^{(h)}(\mathbf{x})\right\|_2^2$. This proves the lemma. □

Corollary 9.25. *Assume $\overline{\mathcal{A}}(\epsilon_1)$ holds for $\epsilon_1<1/2$. For all $h\in\{2,\dots,L\}$, $j\in[d_{h-1}]$, $h'\in\{h+1,h+2,h+3,\dots,L\}$ and $i\in[d_{h'+1}]$,*

$$\begin{aligned}&\mathop{\mathbb{E}}_{\mathbf{W}^{(h+1)},\mathbf{m}^{(h+1)}\dots,\mathbf{W}^{(h')},\mathbf{m}^{(h')},\mathbf{m}^{(h'+1)}}\left[\left\|\mathbf{g}^{(h,j,h')}(\mathbf{x})\odot\mathbf{m}_i^{(h'+1)}\right\|_2^2\middle|\mathbf{g}^{(h,j,h)}\right]\\&\leq 2\mathop{\mathbb{E}}_{\mathbf{W}^{(h+1)},\mathbf{m}^{(h+1)},\dots,\mathbf{W}^{(h')},\mathbf{m}^{(h')},\mathbf{m}^{(h'+1)}}\left[\left\|\mathbf{g}^{(h')}(\mathbf{x})\odot\mathbf{m}_i^{(h'+1)}\right\|_2^2\middle|\mathbf{g}^{(h)}(\mathbf{x})\right].\end{aligned}$$

Proof. Use the fact that the mask $\mathbf{m}_i^{(h'+1)}$ is independent and preserve the 2-norm in expectation. □

Lemma 9.26. *Assume $\overline{\mathcal{A}}(\epsilon_1)$ holds for $\epsilon_1<1/2$. Let $\epsilon\in(0,1)$. If for all $h\in L$, it satisfies that $d_h\geq\Omega(\frac{1}{\alpha}\frac{L^2}{\epsilon^2}\log\frac{2Ld_{h+1}\sum_{h'=1}^{h-1}d'_h}{\delta})=\widetilde{\Omega}(\frac{1}{\alpha}\frac{L^2}{\epsilon^2})$, then with probability at least $1-\delta$ over the randomness in the initialization of weights and masks, we have for all $h\in\{2,\dots,L\}$, $j\in[d_{h-1}]$,*

$$|f^{(h,j,L+1)}(\mathbf{x})|,\ |f^{(h,j,L+1)}(\mathbf{x}')|\leq 2\sqrt{\log\frac{4\sum_{h'=1}^{L-1}d_{h'}}{\delta}}.$$

In other words, if $d_h\geq\Omega(\frac{1}{\alpha}\frac{L^2}{\epsilon^2}\log\frac{2Ld_{h+1}\sum_{h'=1}^{h-1}d'_h}{\delta_3})=\widetilde{\Omega}(\frac{1}{\alpha}\frac{L^2}{\epsilon^2})$, then

$$\mathbb{P}\left[\overline{\mathcal{A}}(\epsilon_1)\Rightarrow\overline{\mathcal{C}}\left(2\sqrt{\log\frac{4\sum_{h'=1}^{L-1}d_{h'}}{\delta_3}}\right)\right]\geq 1-\delta_3$$

Proof. Proposition 9.21 proved that conditioned on $\mathbf{g},\widetilde{\mathbf{g}},\mathbf{m}$, the random variable $((\mathbf{w}\odot\mathbf{m})^\top\widetilde{\mathbf{g}}\sqrt{\frac{c_\sigma}{d_h}})^2\dot\sigma((\mathbf{w}\odot\mathbf{m})^\top\mathbf{g})$ has the same distribution as $((\mathbf{w}\odot\mathbf{m})^\top\widetilde{\mathbf{g}}\sqrt{\frac{c_\sigma}{d_h}})^2\dot\sigma((\mathbf{w}\odot\mathbf{m})^\top\widetilde{\mathbf{g}})$, which implies their concentration properties are the same. At a given layer h', we want this concentration to holds for all $\left\|\mathbf{g}^{(h,j,h')}(\mathbf{x})\odot\mathbf{m}^{(h'+1)}\right\|_2^2$ where $2\leq h\leq h'$ and

$h \in [d_{h-1}]$. Thus there is in total $\sum_{h=1}^{h'-1} d_h$ events. Therefore, by Theorem 9.11, if $d_{h'} \geq \Omega(\frac{1}{\alpha}\frac{L^2}{\epsilon^2}\log\frac{8d_{h'+1}L\sum_{h=1}^{h'-1}d_h}{\delta})$, with probability $1-\delta/2$, for all layer h', for all $h \in \{2,\ldots,L\}$, $j \in [d_{h-1}]$ and $h' \in \{h+1, h+2, \ldots, L\}$, and for both $\mathbf{x}, \mathbf{x}'$

$$\left\|\mathbf{g}^{(h,j,h')}(\mathbf{x}) \odot \mathbf{m}^{(h'+1)}\right\|_2^2 \leq 2\,\mathbb{E}\left[\left\|\mathbf{g}^{(h')}(\mathbf{x}) \odot \mathbf{m}^{(h'+1)}\right\|_2^2\right] + \epsilon \leq 3.$$

By Lemma 9.16, this implies with probability $1-\delta/2$, for all $j \in [d_h]$,

$$|f^{(h,j,L+1)}(\mathbf{x})|,\ |f^{(h,j,L+1)}(\mathbf{x}')| \leq 2\sqrt{\log\frac{4\sum_{h'=1}^{L-1} d_h}{\delta}}.$$

□

9.6.3 Bounding the Dependent Part

Proposition 9.27 (Formal Version of Proposition 5.3)**.** *If $d_{h'} \geq \Omega(\frac{1}{\alpha}\frac{L^2}{\epsilon^2}\log\frac{8d_{h'+1}L\sum_{h\leq h'}d_h}{\delta})$, with probability $1-\delta_3/2$, the event $\overline{\mathcal{C}}(\sqrt{\log\frac{\sum d_h}{\delta_3}})$ (which we define in the proof) holds and at layer h', for all $j \in [d_h]$,*

$$\left\|\sum_i \mathbf{b}_i^{(h+1)}(\mathbf{x}^{(1)})\left(\mathbf{w}_i^{(h+1)}\right)^\top \Pi_{\mathbf{G}_i}\mathbf{M}_i^{(h+1)}\right\|_2 \leq 2 + 2\sqrt{\frac{1}{\alpha}\log\frac{8}{\delta_2}} + \frac{4}{\alpha}\sqrt{\log\frac{4\sum d_h}{\delta_3}}.$$

Proof. By triangle inequality, combining the result from Lemma 9.28 and Lemma 9.29, we have

$$\begin{aligned}\left\|\sum_i \mathbf{b}_i^{(h+1)}(\mathbf{x}^{(1)})\left(\mathbf{w}_i^{(h+1)}\right)^\top \Pi_{\mathbf{G}_i}\mathbf{M}_i^{(h+1)}\right\|_2 &\leq \left\|\sum_i \mathbf{b}_i^{(h+1)}(\mathbf{x}^{(1)})\left(\mathbf{w}_i^{(h+1)}\right)^\top \Pi_{(\mathbf{g}^{(h)}(\mathbf{x})\odot\mathbf{m}_i^{(h+1)})}\mathbf{M}_i^{(h+1)}\right\|_2 \\ &\quad + \left\|\sum_i \mathbf{b}_i^{(h+1)}(\mathbf{x}^{(1)})\left(\mathbf{w}_i^{(h+1)}\right)^\top \Pi_{\mathbf{G}_i/(\mathbf{g}^{(h)}(\mathbf{x})\odot\mathbf{m}_i^{(h+1)})}\mathbf{M}_i^{(h+1)}\right\|_2 \qquad (24) \\ &\leq 2 + 2\sqrt{\frac{1}{\alpha}\log\frac{8}{\delta_2}} + \frac{4}{\alpha}\sqrt{\log\frac{\sum_h d_h}{\delta_3}}.\end{aligned}$$

□

Thus, we need to upper bound the two terms in Equation (24). We first bound the second term which is easier.

Lemma 9.28. *With probability $1-\delta_2$,*

$$\left\|\sum_i \mathbf{b}_i^{(h+1)}(\mathbf{x}^{(1)})\left(\mathbf{w}_i^{(h+1)}\right)^\top \Pi_{\mathbf{G}_i/(\mathbf{g}^{(h)}(\mathbf{x})\odot\mathbf{m}_i^{(h+1)})}\mathbf{M}_i^{(h+1)}\right\|_2 \leq 2\left(1+\sqrt{\frac{1}{\alpha}\log\frac{8}{\delta_2}}\right).$$

Proof. We omit the superscript denoting layers in this proof when there is no confusion. Notice that $\mathbf{G}_i/(\mathbf{g}^{(h)}(\mathbf{x}) \odot \mathbf{m}_i^{(h+1)})$ is spanned by the vector

$$\mathbf{u}^{(i)} := \mathbf{g}^{(h)}(\mathbf{x}') \odot \mathbf{m}_i^{(h+1)} - \left\langle \mathbf{g}^{(h)}(\mathbf{x}) \odot \mathbf{m}_i^{(h+1)}, \mathbf{g}^{(h)}(\mathbf{x}') \odot \mathbf{m}_i^{(h+1)} \right\rangle \mathbf{g}^{(h)}(\mathbf{x}) \odot \mathbf{m}_i^{(h+1)}.$$

Now conditioned on $\mathbf{g}^{(h)}(\mathbf{x}), \mathbf{b}^{(h+1)}(\mathbf{x})$, observe that $\sum_i \mathbf{b}_i(\widetilde{\mathbf{w}}_i^\top \mathbf{u}^{(i)})\mathbf{u}^{(i)} = \sum_i \mathbf{b}_i w_i \mathbf{u}^{(i)}$ where $w_i \sim \mathcal{N}(0,1)$ is Gaussian (independent of $\mathbf{g}^{(h)}(\mathbf{x}), \mathbf{b}^{(h+1)}(\mathbf{x})$). Let $\mathbf{w} := [w_1, w_2, \ldots, w_{d_{h+1}}]$. Its covariance matrix is given by

$$\mathop{\mathbb{E}}_{\widetilde{\mathbf{w}}}\left(\sum_i \mathbf{b}_i w_i \mathbf{u}^{(i)}\right)\left(\sum_j \mathbf{b}_j w_j \mathbf{u}^{(j)}\right)^\top = \mathop{\mathbb{E}}_{\mathbf{w}}\sum_{i,j}\mathbf{b}_i\mathbf{b}_j w_i w_j \mathbf{u}^{(i)}\left(\mathbf{u}^{(j)}\right)^\top = \sum_i \mathbf{b}_i^2 \mathbf{u}^{(i)}\left(\mathbf{u}^{(i)}\right)^\top.$$

Let the eigenvalue decomposition of this matrix be $\mathbf{U}\mathbf{D}\mathbf{U}^\top$, then the vector $\sum_i \mathbf{b}_i w_i \mathbf{u}^{(i)}$ has the same distribution as $\mathbf{U}\mathbf{D}^{1/2}\widetilde{\mathbf{w}}$ where $\widetilde{\mathbf{w}} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. Thus,

$$\mathop{\mathbb{E}}_{\mathbf{w}}\left[\left\|\sum_i \mathbf{b}_i w_i \mathbf{u}^{(i)}\right\|_2^2\right] = \mathop{\mathbb{E}}_{\widetilde{\mathbf{w}}}\left[\widetilde{\mathbf{w}}^\top \mathbf{D}^{1/2}\mathbf{U}^\top \mathbf{U}\mathbf{D}^{1/2}\widetilde{\mathbf{w}}\right] = \mathrm{Tr}(\mathbf{D}).$$

Now, we use the fact that the sum of the eigenvalues of a SPD matrix is its trace and we have

$$\mathrm{Tr}(\mathbf{D}) = \mathrm{Tr}\left(\sum_i \mathbf{b}_i^2 \mathbf{u}^{(i)}\left(\mathbf{u}^{(i)}\right)^\top\right) = \sum_j \sum_i \mathbf{b}_i^2 \left(\mathbf{u}_j^{(i)}\right)^2 = \sum_i \mathbf{b}_i^2 = \|\mathbf{b}\|_2^2.$$

By Jensen's inequality, we have

$$\mathop{\mathbb{E}}_{\mathbf{w}}\left[\left\|\sum_i \mathbf{b}_i w_i \mathbf{u}^{(i)}\right\|_2\right] \le \sqrt{\mathop{\mathbb{E}}_{\mathbf{w}}\left[\left\|\sum_i \mathbf{b}_i w_i \mathbf{u}^{(i)}\right\|_2^2\right]} = \|\mathbf{b}\|_2.$$

Further, use the definition of two norm we can write

$$\left\|\sum_i \mathbf{b}_i w_i \mathbf{u}^{(i)}\right\|_2 = \sup_{\|\mathbf{x}\|_2=1}\left\langle \mathbf{x}, \sum_i \mathbf{b}_i w_i \mathbf{u}^{(i)}\right\rangle \overset{\mathcal{D}}{=} \sup_{\|\mathbf{x}\|_2=1}\left\langle \mathbf{x}, \mathbf{U}\mathbf{D}^{1/2}\widetilde{\mathbf{w}}\right\rangle = \sup_{\|\mathbf{x}\|_2=1}\left\langle \mathbf{x}\mathbf{D}^{1/2}, \widetilde{\mathbf{w}}\right\rangle.$$

The last quantity is in form of a Gaussian complexity and, by Lemma 9.5, has sub-Gaussian concentration with variance proxy $\sigma^2 = \max_i \mathbf{D}_{ii} \le \mathrm{Tr}(\mathbf{D}) = \|\mathbf{b}\|_2^2$. Thus, with probability $1 - \delta_2/4$,

$$\left\|\sum_i \mathbf{b}_i^{(h+1)}(\mathbf{x}^{(1)})\left(\mathbf{w}_i^{(h+1)}\right)^\top \Pi_{\mathbf{G}_i/(\mathbf{g}^{(h)}(\mathbf{x})\odot\mathbf{m}_i^{(h+1)})}\mathbf{M}_i^{(h+1)}\right\|_2 \le \left(1+\sqrt{\frac{2}{\alpha}\log\frac{8}{\delta_2}}\right)\left\|\mathbf{b}^{(h+1)}\right\|_2 \le 2\left(1+\sqrt{\frac{1}{\alpha}\log\frac{8}{\delta_2}}\right).$$

□

Now we bound the first term in Equation (24).

Lemma 9.29. *With probability* $1-\delta$,

$$\left\|\sum_i \mathbf{b}_i^{(h+1)}(\mathbf{x}^{(1)})\left(\mathbf{w}_i^{(h+1)}\right)^\top \Pi_{(\mathbf{g}^{(h)}\odot\mathbf{m}_i^{(h+1)})}\mathbf{M}_i^{(h+1)}\right\|_2 \le \frac{4}{\alpha}\sqrt{\log\frac{4\sum_h d_h}{\delta}}.$$

Proof. Since $\Pi_{(\mathbf{g}^{(h)}\odot\mathbf{m}_i^{(h+1)})} = \frac{(\mathbf{g}^{(h)}\odot\mathbf{m}_i^{(h+1)})(\mathbf{g}^{(h)}\odot\mathbf{m}_i^{(h+1)})^\top}{\left\|(\mathbf{g}^{(h)}\odot\mathbf{m}_i^{(h+1)})\right\|_2^2}$ we have

$$\begin{aligned}
&\left\|\sum_i \mathbf{b}_i^{(h+1)}(\mathbf{x}^{(1)})\left(\mathbf{w}_i^{(h+1)}\right)^\top \Pi_{(\mathbf{g}^{(h)}\odot\mathbf{m}_i^{(h+1)})}\mathbf{M}_i^{(h+1)}\right\|_2 \\
&= \left\|\sum_i \mathbf{b}_i^{(h+1)}(\mathbf{x}^{(1)})\left(\mathbf{w}_i^{(h+1)}\right)^\top \frac{(\mathbf{g}^{(h)}\odot\mathbf{m}_i^{(h+1)})(\mathbf{g}^{(h)}\odot\mathbf{m}_i^{(h+1)})^\top}{\left\|(\mathbf{g}^{(h)}\odot\mathbf{m}_i^{(h+1)})\right\|_2^2}\mathbf{M}_i^{(h+1)}\right\|_2 \\
&= \left\|\frac{1}{\sqrt{\alpha}}\sum_i \mathbf{b}_i^{(h+1)}(\mathbf{x}^{(1)})\left(\mathbf{w}_i^{(h+1)}\right)^\top \frac{(\mathbf{g}^{(h)}\odot\mathbf{m}_i^{(h+1)})(\mathbf{g}^{(h)}\odot\mathbf{m}_i^{(h+1)})^\top}{\left\|(\mathbf{g}^{(h)}\odot\mathbf{m}_i^{(h+1)})\right\|_2^2}\right\|_2.
\end{aligned}$$

Now let's look at the j-th coordinate of this vector:

$$\left(\frac{1}{\sqrt{\alpha}}\sum_i \mathbf{b}_i^{(h+1)}(\mathbf{x}^{(1)})\left(\mathbf{w}_i^{(h+1)}\right)^\top \frac{(\mathbf{g}^{(h)}\odot\mathbf{m}_i^{(h+1)})(\mathbf{g}^{(h)}\odot\mathbf{m}_i^{(h+1)})^\top}{\left\|(\mathbf{g}^{(h)}\odot\mathbf{m}_i^{(h+1)})\right\|_2^2}\right)_j$$

$$
\begin{aligned}
&= \frac{1}{\sqrt{\alpha}} \sum_i \mathbf{b}_i^{(h+1)}(\mathbf{x}^{(1)}) \left(\mathbf{w}_i^{(h+1)}\right)^\top \frac{(\mathbf{g}^{(h)} \odot \mathbf{m}_i^{(h+1)}) \mathbf{m}_{ij}^{(h+1)} \mathbf{g}_j^{(h)}}{\left\|(\mathbf{g}^{(h)} \odot \mathbf{m}_i^{(h+1)})\right\|_2^2} \\
&= \frac{1}{\alpha} \mathbf{g}_j^{(h)} \left(\mathbf{b}^{(h+1)}(\mathbf{x}^{(1)})\right)^\top \mathrm{diag}_i \left(\frac{\mathbf{m}_{ij}^{(h+1)} \sqrt{\alpha}}{\left\|\mathbf{g}^{(h)} \odot \mathbf{m}_i^{(h+1)}\right\|_2^2} \right) \begin{bmatrix} \left(\mathbf{w}_1^{(h+1)} \odot \mathbf{m}_1^{(h+1)}\right)^\top (\mathbf{g}^{(h)} \odot \mathbf{m}_1^{(h+1)}) \sqrt{\alpha} \\ \left(\mathbf{w}_2^{(h+1)} \odot \mathbf{m}_2^{(h+1)}\right)^\top (\mathbf{g}^{(h)} \odot \mathbf{m}_2^{(h+1)}) \sqrt{\alpha} \\ \vdots \\ \left(\mathbf{w}_{d_{h+1}}^{(h+1)} \odot \mathbf{m}_{d_{h+1}}^{(h+1)}\right)^\top (\mathbf{g}^{(h)} \odot \mathbf{m}_{d_{h+1}}^{(h+1)}) \sqrt{\alpha} \end{bmatrix} \\
&= \frac{1}{\alpha} \mathbf{g}_j^{(h)} \left(\mathbf{b}^{(h+1)}(\mathbf{x}^{(1)})\right)^\top \mathrm{diag}_i \left(\frac{\mathbf{m}_{ij}^{(h+1)} \sqrt{\alpha}}{\left\|\mathbf{g}^{(h)} \odot \mathbf{m}_i^{(h+1)}\right\|_2^2} \right) \mathbf{f}^{(h+1)}(\mathbf{x}^{(1)}) \\
&= \frac{1}{\alpha} \mathbf{g}_j^{(h)} \left(\mathbf{w}^{(L+1)} \odot \mathbf{m}^{(L+1)}\right)^\top \sqrt{\frac{c_\sigma}{d_L}} \mathbf{D}^{(L)}(\mathbf{x}^{(1)}) \left(\mathbf{W}^{(L)} \odot \mathbf{m}^{(L)}\right) \\
&\quad \cdots \sqrt{\frac{c_\sigma}{d_{h+1}}} \mathbf{D}^{(h+1)}(\mathbf{x}^{(1)}) \mathrm{diag}_i \left(\frac{\mathbf{m}_{ij}^{(h+1)} \sqrt{\alpha}}{\left\|\mathbf{g}^{(h)} \odot \mathbf{m}_i^{(h+1)}\right\|_2^2} \right) \mathbf{f}^{(h+1)}(\mathbf{x}^{(1)}) \\
&= \frac{1}{\alpha} \mathbf{g}_j^{(h)} f^{(h+1,j,L+1)}(\mathbf{x}^{(1)})
\end{aligned}
$$

By Lemma 9.26, we have

$$
|f^{(h,j,L+1)}| \leq 2 \sqrt{\log \frac{4 \sum_{h'=1}^{L-1} d_{h'}}{\delta}}.
$$

Finally, by Theorem 9.11, we have $\left\|\mathbf{g}^{(h)}\right\|_2 \leq 2$. This implies

$$
\left\| \sum_i \mathbf{b}_i^{(h+1)}(\mathbf{x}^{(1)}) \left(\mathbf{w}_i^{(h+1)}\right)^\top \mathbf{M}_i^{(h+1)} \Pi_{(\mathbf{g}^{(h)} \odot \mathbf{m}_i^{(h+1)})} \right\|_2 \leq \frac{4}{\alpha} \sqrt{\log \frac{4 \sum_{h'=1}^{L-1} d_h}{\delta_3}}
$$

□

Continuing Proof of Lemma 9.14. Wrapping things up, from Equation (18), by Proposition 9.20 and Proposition 9.27,

$$
\begin{aligned}
&\left| \frac{c_\sigma}{d_h} \sum_{i,j} \mathbf{b}_i^{(h+1)}(\mathbf{x}^{(1)}) \mathbf{b}_j^{(h+1)}(\mathbf{x}^{(2)}) \left(\mathbf{w}_i^{(h+1)}\right) \mathbf{M}_i^{(h+1)} \mathbf{D}^{(h)}(\mathbf{x}^{(1)}) \mathbf{D}^{(h)}(\mathbf{x}^{(2)}) \mathbf{M}_j^{(h+1)} \mathbf{w}_j^{(h+1)} \right. \\
&\quad \left. - \frac{c_\sigma}{d_h} \sum_{i,j} \mathbf{b}_i^{(h+1)}(\mathbf{x}^{(1)}) \mathbf{b}_j^{(h+1)}(\mathbf{x}^{(2)}) \left(\mathbf{w}_i^{(h+1)}\right)^\top \mathbf{M}_i^{(h+1)} \Pi_{\mathbf{G}_i}^\perp \mathbf{D}^{(h)}(\mathbf{x}^{(1)}) \mathbf{D}^{(h)}(\mathbf{x}^{(2)}) \Pi_{\mathbf{G}_j}^\perp \mathbf{M}_j^{(h+1)} \mathbf{w}_j^{(h+1)} \right| \\
&\leq \left\| \sqrt{\frac{c_\sigma}{d_h}} \sum_i \mathbf{b}_i^{(h+1)}(\mathbf{x}^{(1)}) \left(\mathbf{w}_i^{(h+1)}\right)^\top \mathbf{M}_i^{(h+1)} \Pi_{\mathbf{G}_i}^\perp \mathbf{D}^{(h)}(\mathbf{x}^{(1)}) \right\| \left\| \sqrt{\frac{c_\sigma}{d_h}} \sum_j \mathbf{b}_j^{(h+1)}(\mathbf{x}^{(2)}) \mathbf{D}^{(h)}(\mathbf{x}^{(2)}) \Pi_{\mathbf{G}_j} \mathbf{M}_j^{(h+1)} \mathbf{w}_j^{(h+1)} \right\| \\
&+ \left\| \sqrt{\frac{c_\sigma}{d_h}} \sum_i \mathbf{b}_i^{(h+1)}(\mathbf{x}^{(1)}) \left(\mathbf{w}_i^{(h+1)}\right)^\top \mathbf{M}_i^{(h+1)} \Pi_{\mathbf{G}_i} \mathbf{D}^{(h)}(\mathbf{x}^{(1)}) \right\| \left\| \sqrt{\frac{c_\sigma}{d_h}} \sum_j \mathbf{b}_j^{(h+1)}(\mathbf{x}^{(2)}) \mathbf{D}^{(h)}(\mathbf{x}^{(2)}) \Pi_{\mathbf{G}_j}^\perp \mathbf{M}_j^{(h+1)} \mathbf{w}_j^{(h+1)} \right\| \\
&+ \left\| \sqrt{\frac{c_\sigma}{d_h}} \sum_i \mathbf{b}_i^{(h+1)}(\mathbf{x}^{(1)}) \left(\mathbf{w}_i^{(h+1)}\right)^\top \mathbf{M}_i^{(h+1)} \Pi_{\mathbf{G}_i} \mathbf{D}^{(h)}(\mathbf{x}^{(1)}) \right\| \left\| \sqrt{\frac{c_\sigma}{d_h}} \sum_j \mathbf{b}_j^{(h+1)}(\mathbf{x}^{(2)}) \mathbf{D}^{(h)}(\mathbf{x}^{(2)}) \Pi_{\mathbf{G}_j} \mathbf{M}_j^{(h+1)} \mathbf{w}_j^{(h+1)} \right\| \\
&\leq 2 \left(\frac{12\sqrt{2}}{\sqrt{d_h}} + 12 \sqrt{\frac{2}{\alpha} \frac{\log \frac{8}{\delta_2}}{d_h}} + \frac{24}{\alpha} \frac{\sqrt{2 \log \frac{4 \sum d_h}{\delta_3}}}{\sqrt{d_h}} \right) + \frac{2}{d_h} \left(2 + 2 \sqrt{\frac{1}{\alpha} \log \frac{8}{\delta_2}} + \frac{4}{\alpha} \sqrt{\log \frac{4 \sum d_h}{\delta_3}} \right)^2
\end{aligned}
$$

$$\leq \frac{48\sqrt{2}}{\sqrt{d_h}} + 48\sqrt{\frac{2}{\alpha}\frac{\log\frac{8}{\delta_2}}{d_h}} + \frac{96}{\alpha}\frac{\sqrt{2\log\frac{4\sum d_h}{\delta_3}}}{\sqrt{d_h}}$$

□

10 ADDITIONAL EXPERIMENT RESULTS

10.1 Experimental Setup

All of our models are trained with SGD and the detailed settings are summarized below.

Table 1: Summary of architectures, dataset and training hyperparameters

MODEL	DATA	EPOCH	BATCH SIZE	LR	MOMENTUM	LR DECAY, EPOCH	WEIGHT DECAY
LENET	MNIST	40	128	0.1	0	0	0
VGG	CIFAR-10	160	128	0.1	0.9	$0.1 \times [80, 120]$	0.0001
RESNETS	CIFAR-10	160	128	0.1	0.9	$0.1 \times [80, 120]$	0.0001

10.2 Further Experiment Results

10.2.1 MNIST

For MNIST dataset, we train a fully-connected neural network with 2-hidden layers of width 2048. The performance is shown in Figure 4.

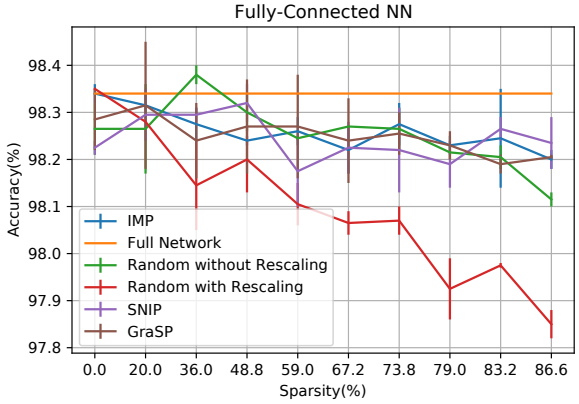


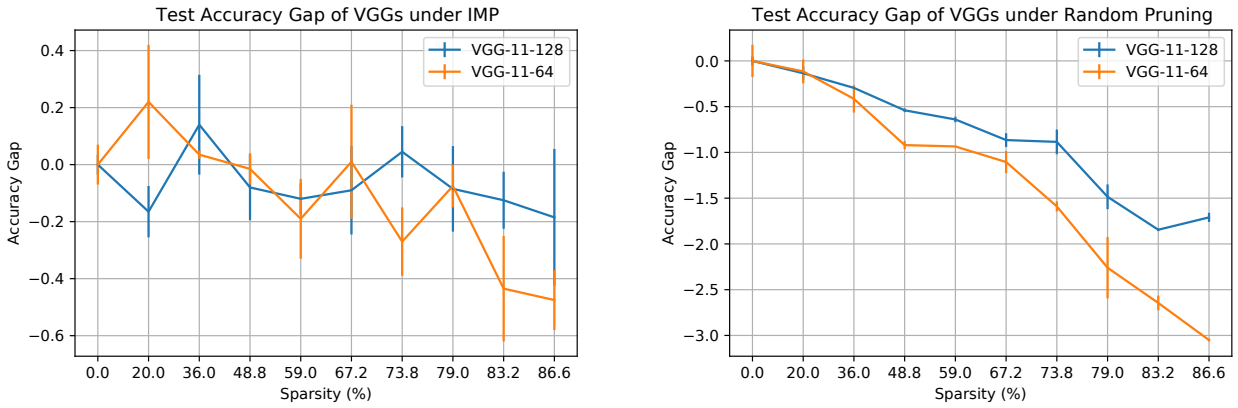
Figure 4: Comparing the performance of random pruning with/without rescaling with IMP, SNIP and GraSP using a fully-connected neural network with 2 hidden layers of width 2048 on MNIST dataset.

10.2.2 CIFAR-10

VGG. We train standard VGG-11 (i.e., VGG-11-64) and VGG-11-128 on CIFAR-10 dataset. The results are shown in Figure 5a and Figure 5b.

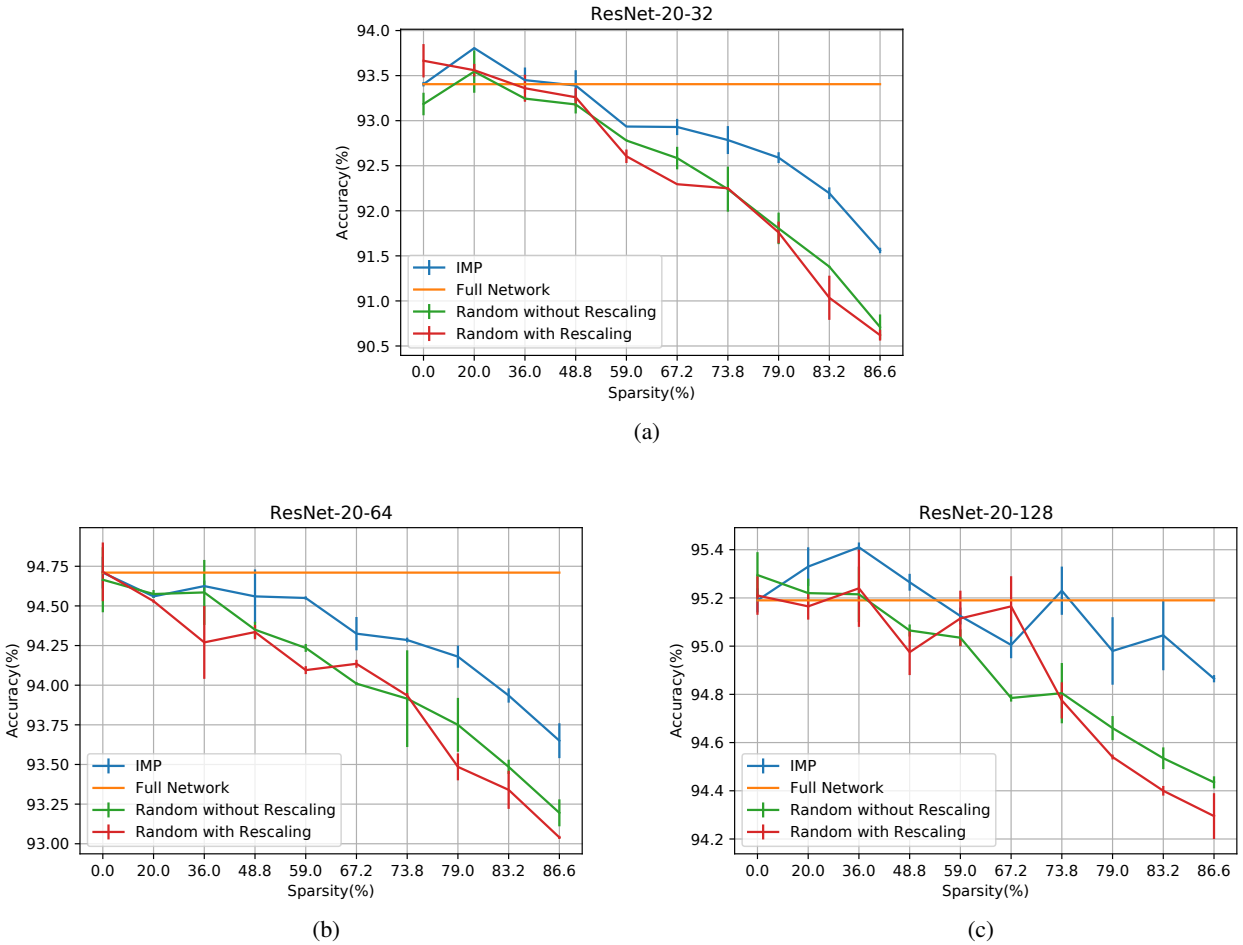
ResNet. We further train ResNet-20 of width 32, 64 and 128 and compare the performance of random pruning with and without rescaling against IMP. The results are shown in Figure 6a, Figure 6b and Figure 6c.

We further plot random pruning with rescaling across different width in Figure 7 and pruning by IMP in Figure 8. The result further shows under the same pruning rate, increasing width can make the pruned model perform on par with the full model.



(a) (b)

Figure 5: The performance of random pruning and IMP using VGG-11 of different width on CIFAR-10 dataset.



(a)

(b) (c)

Figure 6: The performance of random pruning with/without rescaling and IMP using ResNet-20 of different width on CIFAR-10 dataset.

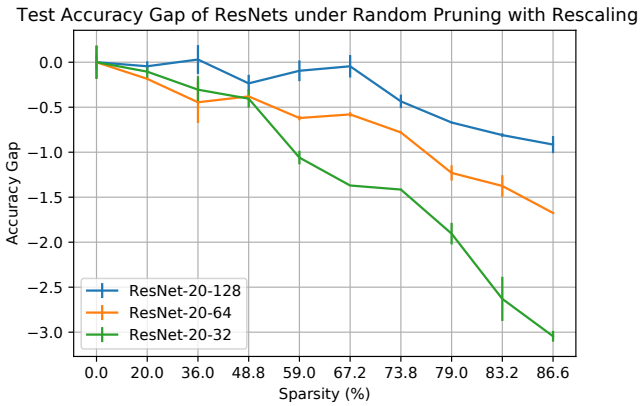


Figure 7: The test accuracy gap of random pruning with rescaling using ResNet-20 of different width on CIFAR-10 dataset.

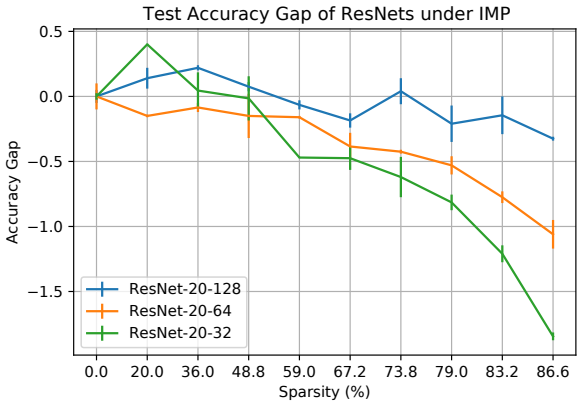


Figure 8: The test accuracy gap of IMP using ResNet-20 of different width on CIFAR-10 dataset.