Why Do Deep Residual Networks Generalize Better than Deep Feedforward Networks? — A Neural Tangent Kernel Perspective

Kaixuan Huang*
Peking University
hackyhuang@pku.edu.cn

Yuqing Wang*
Georgia Institute of Technology
ywang3398@gatech.edu

Molei Tao
Georgia Institute of Technology
mtao@gatech.edu

Tuo Zhao
Georgia Institute of Technology
tourzhao@gatech.edu

Abstract

Deep residual networks (ResNets) have demonstrated better generalization performance than deep feedforward networks (FFNets). However, the theory behind such a phenomenon is still largely unknown. This paper studies this fundamental problem in deep learning from a so-called "neural tangent kernel" perspective. Specifically, we first show that under proper conditions, as the width goes to infinity, training deep ResNets can be viewed as learning reproducing kernel functions with some kernel function. We then compare the kernel of deep ResNets with that of deep FFNets and discover that the class of functions induced by the kernel of FFNets is asymptotically not learnable, as the depth goes to infinity. In contrast, the class of functions induced by the kernel of ResNets does not exhibit such degeneracy. Our discovery partially justifies the advantages of deep ResNets over deep FFNets in generalization abilities. Numerical results are provided to support our claim.

1 Introduction

Deep Neural Networks (DNNs) have made significant progress in a variety of real-world applications, such as computer vision [1, 2, 3], speech recognition, natural language processing [4, 5, 6], recommendation systems, etc. Among various network architectures, Residual Networks (ResNets, [7]) are undoubtedly a breakthrough. Residual Networks are equipped with residual connections, which skip layers in the forward step. Similar ideas based on gating mechanisms are also adopted in Highway Networks [8], and further inspire many follow-up works such as Densely Connected Networks [9].

Compared with conventional Feedforward Networks (FFNets), residual networks demonstrate surprising generalization abilities. Existing literature rarely considers deep feedforward networks with more than 30 layers. This is because many experimental results have suggested that very deep feedforward networks yield worse generalization performance than their shallow counterparts [7]. In contrast, we can train residual networks with hundreds of layers, and achieve better generalization performance than that of feedforward networks. For example, ResNet-152 [7], achieving a 19.38% top-1 error on the ImageNet data set, consists of 152 layers; ResNet-1001 [10], achieving a 4.92% error on the CIFAR-10 data set, consists of 1000 layers.

Despite the great success and popularity of the residual networks, the reason why they generalize so well is still largely unknown. There have been several lines of research attempting to demystify this phenomenon. One line of research focuses on empirical studies of residual networks, and provides

*Equal contribution.

34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada.

intriguing observations. For example, [11] show that residual networks behave like an ensemble of weakly dependent networks of much smaller sizes, and meanwhile, they also show that the gradient vanishing issue is also significantly mitigated due to these smaller networks. [12] further provide a more refined elaboration on the gradient vanishing issue. They demonstrate that the gradient magnitude in residual networks only shows sublinear decay (with respect to the layer), which is much slower than the exponential decay of gradient magnitude in feedforward neural networks. [13] propose a visualization approach for analyzing the landscape of neural networks, and further demonstrate that residual networks have smoother optimization landscape due to the skip-layer connections.

Another line of research focuses on theoretical investigations of residual networks under *simplified network architectures*. A commonly adopted structure, which is a reformulation of FFNets, is

$$x_\ell = \phi(x_{\ell-1} + \alpha W_\ell x_{\ell-1}), \tag{1}$$

where ℓ is the number of layers and the skip-connection only bypasses the weight matrix W_ℓ at each layer [14, 15, 16, 17, 18]. Specifically, [16] study the optimization landscape with linear activation; [17] study using Stochastic Gradient Descent (SGD) to train a two-layer ResNet. [18] study using Gradient Descent (GD) to train a two-layer non-overlapping residual network. [14, 15] both take the perturbation analysis approach to show convergence of such ResNets. A more realistic structure is

$$x_\ell = x_{\ell-1} + \phi(\alpha W_\ell x_{\ell-1}), \tag{2}$$

where the skip-connection bypasses the activation function [19, 20]. [20] only consider separable setting and take the perturbation analysis to show the convergence and generalization property of such ResNet. These results, however, are only loosely related to the generalization abilities of residual networks, and often considered to be overoptimistic, due to the oversimplified assumptions.

Some more recent works provide a new theoretical framework for analyzing *overparameterized* neural networks [21, 22, 23, 24, 14, 25, 26, 27]. They focus on connecting two- or three-layer overparameterized (sufficiently wide) neural networks to *reproducing kernel Hilbert spaces*. Specifically, they show that under proper conditions, the weight matrices of a well trained overparameterized neural network (achieving any given small training error) are actually very close to their initialization. Accordingly, the training process can be described as searching within some class of reproducing kernel functions, where the associated kernel is called the "*neural tangent kernel*" (*NTK*, [21]) and only depends on the initialization of the weights. Accordingly, the generalization properties of the overparameterized neural network are equivalent to those of the associated NTK function class. Based on such a framework, [19] derived the NTK of the ResNet (2) when only the last layer is trained, and proved the convergence of such ResNet. However, they did not provide an explicit formula for the NTK when all layers are trained, which is required for characterizing the generalization property of ResNets.

To better understand the generalization abilities of deep feedforward and residual networks, we propose to investigate the NTKs associated with these networks when all but the last layers are trained, and consider the case when both widths and depths go to infinity[2]. For the structure of ResNets, we adopt (2) only with a slight modification, since it captures the essence of the skip-connection; see Section 2

$$x_\ell = x_{\ell-1} + \alpha\sqrt{\frac{1}{m}} V_\ell \sigma_0\Big(\sqrt{\frac{2}{m}} W_\ell x_{\ell-1}\Big). \tag{3}$$

Specifically, we prove that similar to what has been shown for feedforward networks [21], as the width of deep residual networks increases to infinity, training residual networks can also be viewed as learning reproducing kernel functions with some NTK. However, such an NTK associated with the residual networks exhibits a very different behavior from that of feedforward networks.

To demonstrate such a difference, we further consider the regime, where the depths of both feedforward and residual networks are allowed to increase to infinity. Accordingly, both NTKs associated with deep feedforward and residual networks converge to their limiting forms sublinearly (in terms of the depth). For notational simplicity, we refer to the limiting form of the NTKs as the limiting NTK. Besides asymptotic analysis, we also provide nonasymptotic bounds, which demonstrate equivalence between limiting NTKs and neural networks with sufficient depth and width.

When comparing their limiting NTKs, we find that the class of functions induced by the limiting NTKs associated with deep feedforward networks is essentially not learnable. Such a class of

[2]More precisely, our analysis considers the regime, where the widths go to infinity first, and then the depths go to infinity. See more details in Section 4.

functions is sufficient to overfit training data. Given any finite sample size, however, the learned function cannot generalize. In contrast, the class of functions induced by the limiting NTKs associated with deep residual networks does not exhibit such degeneracy. Our discovery partially justifies the advantages of deep residual networks over deep feedforward networks in terms of generalization abilities. Numerical results are provided to support our claim.

Our work is closely related to [28]. They also investigate the so-called "Gaussian Process" kernel induced by feedforward networks under the regime where the depth is allowed to increase to infinity. However, their studied neural networks are essentially some specific implementations of the reproducing kernels using random features, since the training process only updates the last layer of the neural networks, and keeps other layers unchanged. In contrast, we assume the training process updates all layers except for the last layer.

Notations: We use $\sigma_0(z) = \max(0, z)$ to denote the ReLU activation function in neural networks. We use $\sigma(z)$ to denote the normalized ReLU function $\sigma(z) = \sqrt{2}\max(0, z)$. The derivative [3] of ReLU function (step function) is $\sigma_0'(z) = \mathbb{I}_{\{z\geq 0\}}$. Then $\sigma'(z) = \sqrt{2}\mathbb{I}_{\{z\geq 0\}}$ is the normalized step function. We use D to denote the input dimension and $\mathbb{S}^{D-1}$ to denote the unit sphere in $\mathbb{R}^D$. We use m to denote the network width (the number of neurons at each layer) and L to denote the depth. Let $\mathcal{M}_+^2$ be the set of all 2×2 positive semi-definite matrices. We use $\mathcal{F}$ to denote the set of all symmetric and positive semi-definite functions from $\mathbb{R}^D \times \mathbb{R}^D$ to $\mathbb{R}$. We use $\|\cdot\|_{\max}$ to denote the entry-wise ℓ_∞ norm for matrices and use $\|\cdot\|$ to denote the ℓ_2 norm for vectors and the spectral norm for matrices. We use $\mathrm{diag}(\cdot)$ to denote the diagonal matrix. We use I_n to denote the $n \times n$ identity matrix. We use x and $\tilde{x}$ to denote a pair of inputs. We use x_ℓ and $\tilde{x}_\ell$ to denote the output of the ℓ-th layer of a network for the input x and $\tilde{x}$, respectively. We use f and $\tilde{f}$ to denote the final output of the network for x and $\tilde{x}$, respectively. We use $\nabla_\theta f = \nabla_\theta f_\theta(x)$ to denote the derivative of parametrized model f_θ w.r.t. θ at the input x, and $\nabla_\theta \tilde{f}$ to denote the counterpart at the input $\tilde{x}$.

2 Background

For self-containedness, we first briefly review feedforward networks, residual networks and dual kernels associated with neural networks.

Feedforward Networks. We define an L-layer feedforward network (FFNet) $f(x)$ with ReLU activation in a recursive manner,

$$x_0 = x;\ x_\ell = \sqrt{\frac{2}{m}}\sigma_0(W_\ell x_{\ell-1}),\ \ell = 1, \cdots, L;\ f(x) = v^\top x_L, \tag{4}$$

where $W_1 \in \mathbb{R}^{m\times D}$ and $W_2, \cdots, W_L \in \mathbb{R}^{m\times m}$ are weight matrices, and $v \in \mathbb{R}^m$ is the output weight vector. For simplicity, we only consider feedforward networks with scalar outputs.

Residual Networks. We define an L-layer residual network (ResNet) $f(x)$ in a recursive manner,

$$x_0 = \sqrt{\frac{1}{m}}Ax;\ x_\ell = x_{\ell-1} + \alpha\sqrt{\frac{1}{m}}V_\ell\sigma_0\Big(\sqrt{\frac{2}{m}}W_\ell x_{\ell-1}\Big),\ \ell = 1, \cdots, L;\ f(x) = v^\top x_L, \tag{5}$$

where $W_\ell, V_\ell \in \mathbb{R}^{m\times m}$ for $\ell = 1, \cdots, L$, $A \in \mathbb{R}^{m\times D}$, $v \in \mathbb{R}^m$, and $\alpha = L^{-\gamma}$ is the scaling factor of the bottleneck layers. The scaling factor α is necessary for controlling the norm of x_l.

The network architecture in (5) is similar to the "pre-activation" shortcuts in [10], except that each bottleneck layer only contains one activation - between W_ℓ and V_ℓ. We remove the activation of the input due to some technical issues (See more details in Section 3).

Dual and Normalized Kernels. The dual kernel technique was first proposed in [29] and motivated several follow-up works such as [28, 30]. Here we adopt the description in [28]. We use K to denote a kernel function on the input space $\mathbb{R}^D$, i.e., $K : \mathbb{R}^D \times \mathbb{R}^D \to \mathbb{R}$. We denote

$$\Sigma(x, \tilde{x}) = \begin{pmatrix} K(x, x) & K(x, \tilde{x}) \\ K(\tilde{x}, x) & K(\tilde{x}, \tilde{x}) \end{pmatrix} \text{ and } N_\rho = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix},$$

where $K \in \mathcal{F}$, $\rho \in \mathbb{R}$. Given an activation function $\phi : \mathbb{R} \to \mathbb{R}$, its dual activation function $\hat{\phi} : [-1, 1] \to [-1, 1]$ is defined to be $\hat{\phi}(\rho) = \mathbb{E}_{(X,\tilde{X})\sim\mathcal{N}(0,N_\rho)}\phi(X)\phi(\tilde{X})$.

We then define the dual kernel as follows.

Definition 1. *We say that $\Gamma_\phi(K) : \mathbb{R}^D \times \mathbb{R}^D \to \mathbb{R}$ is the dual kernel of K with respect to the activation ϕ, if we have $\Gamma_\phi(K)(x, \tilde{x}) = \mathbb{E}_{(X,\tilde{X})\sim\mathcal{N}(0,\Sigma(x,\tilde{x}))}\phi(X)\phi(\tilde{X})$.*

[3] Although the ReLU function σ_0 is not differentiable at 0, we call σ_0' derivative for notational convenience.

Note that $\Gamma_\phi(K)$ is also positive semi-definite. We also define the normalized kernel.

Definition 2. *We say that a kernel $K \in \mathcal{F}$ is normalized, if $K(x,x) = 1$ for all $x \in \mathbb{R}^D$. For a general kernel $K \in \mathcal{F}$, we define its normalized kernel by $\overline{K}$ where $\overline{K}(x,\tilde{x}) = \frac{K(x,\tilde{x})}{\sqrt{K(x,x)K(\tilde{x},\tilde{x})}}$.*

For *normalized ReLU function* $\sigma(z) = \sqrt{2}\max(0,z)$, [28] show $\hat{\sigma}(\rho) = \frac{\sqrt{1-\rho^2}+(\pi-\cos^{-1}(\rho))\rho}{\pi}$. Since $\sigma(z)$ is positive homogeneous, we have $\Gamma_\sigma(K)(x,\tilde{x}) = \sqrt{K(x,x)K(\tilde{x},\tilde{x})}\,\hat{\sigma}(\overline{K}(x,\tilde{x}))$. For *derivative of normalized ReLU function* $\sigma'(z) = \sqrt{2}\mathbb{I}_{\{z\geq 0\}}$, [28] show that $\widehat{\sigma'}(\rho) = \frac{\pi-\cos^{-1}(\rho)}{\pi}$. Since $\sigma'(z)$ is zeroth-order positive homogeneous, we have $\Gamma_{\sigma'}(K)(x,\tilde{x}) = \widehat{\sigma'}(\overline{K}(x,\tilde{x}))$. For more technical details of the dual kernel, we refer the readers to [28].

3 Neural Tangent Kernels of Deep Networks

There are two approaches to connecting neural networks to kernels: one is *Gaussian Process Kernel* (GP Kernel); the other is *Neural Tangent Kernel* (NTK). GP Kernel corresponds to the regime where the first L layers are fixed after random initialization, and only the last layer is trained. Therefore, the first L layers are essentially random feature mapping [31]. This is inconsistent with the practice, as the first L layers should also be trained. In contrast, NTK corresponds to the regime where the first L layers are also trained. For both GP Kernel and NTK, we consider the case when the width of the neural network goes to infinity. Due to space limit, we only provide some proof sketches for our theory, and all technical details are deferred to the appendix.

3.1 Feedforward Networks

We consider the Feedforward Network (FFNet) defined in (4), where $W_1 \in \mathbb{R}^{m\times D}$, $W_2, \cdots, W_L \in \mathbb{R}^{m\times m}$ and $v \in \mathbb{R}^m$ are all initialized as i.i.d. $\mathcal{N}(0,1)$ variables.[4] Given such random initialization, the outputs converge to a Gaussian process, as the width goes to infinity [32, 21]. Accordingly, the GP kernel is defined as follows.

Proposition 1 ([28, 21])**.** *The GP kernel of the L-layer FFNet defined in* (4) *is*

$$K_0(x,\tilde{x}) = x^\top \tilde{x};\ K_\ell(x,\tilde{x}) = \Gamma_\sigma(K_{\ell-1})(x,\tilde{x}),\ \ell = 1,\cdots,L. \tag{6}$$

Theorem 1 ([28])**.** *For the FFNet defined in* (4)*, there exists an absolute constant C, given the width $m \geq C\epsilon^{-2}L^2\log(8L/\delta)$, with probability at least $1-\delta$ over the randomness of the initialization, for input $x, \tilde{x}$ on the unit sphere, the inner product of the outputs of the ℓ-th layer can be approximated by $K_\ell(x,\tilde{x})$, i.e.,*

$$|\langle x_\ell, \tilde{x}_\ell\rangle - K_\ell(x,\tilde{x})| \leq \epsilon, \text{ for all } \ell = 1,\cdots,L.$$

The next proposition shows the NTK of this FFNet. Unlike the GP kernel, the NTK corresponds to the case when $\theta = (W_1,\cdots,W_L)$ are trained.

Proposition 2 ([21])**.** *The NTK of the FFNet can be derived in terms of the GP kernels as*

$$\Omega_L(x,\tilde{x}) = \sum_{\ell=1}^{L}\Big[K_{\ell-1}(x,\tilde{x})\prod_{i=\ell}^{L}\Gamma_{\sigma'}(K_{i-1})(x,\tilde{x})\Big]. \tag{7}$$

Besides the asymptotic result, [22] further provide a nonasymptotic bound as follows.

Theorem 2 ([22])**.** *For the FFNet defined in* (4)*, when the width $m \geq CL^6\epsilon^{-4}\log(L/\delta)$, where C is a constant, with probability at least $1-\delta$ over the initialization, for input $x, \tilde{x}$ on the unit sphere, the Neural Tangent Kernel can be approximated by $\Omega_L(x,\tilde{x})$, i.e.,*

$$\big|\langle\nabla_\theta f, \nabla_\theta \tilde{f}\rangle - \Omega_L(x,\tilde{x})\big| \leq L\epsilon.$$

[22] then showed that a sufficiently wide FFNet trained by gradient flow is close to the kernel regression predictor via its NTK.

Remark 1. *For self-containedness, we directly adopt the results from existing literature in this subsection. For more technical details on gradient flow and kernel ridge regression, we refer the readers to [28, 21, 22].*

3.2 Residual Networks

We consider the Residual Network (ResNet) in (5), where all parameters $(A, v, W_1,\cdots,W_L, V_1,\cdots,V_L)$ are independently initialized from the standard Gaussian

[4] In general, the weight matrices do not need to be square matrices, nor do they need to be of the same size.

distribution. For simplicity, we only train $\theta = (W_1, \cdots, W_L, V_1, \cdots, V_L)$, but not A or v, and the NTK of the ResNet is computed accordingly. Note that our theory can be naturally generalized to the setting where all parameters including A and v are trained, but the analysis will be more involved. Our next proposition derives the GP kernel of the ResNet.

Proposition 3. *The GP kernel of the ResNet is*

$$K_0(x, \tilde{x}) = x^\top \tilde{x};\ K_\ell(x, \tilde{x}) = K_{\ell-1}(x, \tilde{x}) + \alpha^2 \Gamma_\sigma(K_{\ell-1})(x, \tilde{x}),$$

where $\ell = 1, \cdots, L$, and $\alpha = L^{-\gamma}$ for $0.5 \leq \gamma \leq 1$.

Proposition 3 demonstrates that each layer of the ResNet recursively "contributes" to the kernel in an incremental manner, which is quite different from that of the FFNet (shown in Proposition 1). Proposition 3 essentially provides a rigorous justification for the intuition discussed by [33]. Besides the above asymptotic result, we also derive a nonasymptotic bound as follows.

Theorem 3. *For the ResNet defined in* (5)*, given two inputs on the unit sphere $x, \tilde{x} \in \mathbb{S}^{D-1}$, $\epsilon < 0.5$, and*

$$m \geq C\epsilon^{-2} L^{2-2\gamma} \log(36(L+1)/\delta),$$

where C is a constant and $0.5 \leq \gamma \leq 1$, with probability at least $1 - \delta$ over the randomness of the initialization, for all layers $\ell = 0, \cdots, L$ and $(x^{(1)}, x^{(2)}) \in \{(x, x), (x, \tilde{x}), (\tilde{x}, \tilde{x})\}$, we have

$$|\langle x_\ell^{(1)}, x_\ell^{(2)} \rangle - K_\ell(x^{(1)}, x^{(2)})| \leq \epsilon,$$

where K_ℓ is recursively defined in Proposition 3.

Theorem 3 implies that sufficiently wide residual networks are mimicking the GP kernel under proper conditions. The proof can be found in Appendix A. Next we present the NTK of the ResNet defined in (5) in the following proposition.

Proposition 4. *The NTK of the ResNet is $\Omega_L(x, \tilde{x}) = \alpha^2 \sum_{\ell=1}^{L} \big[B_{\ell+1}(x, \tilde{x}) \Gamma_\sigma(K_{\ell-1})(x, \tilde{x}) + K_{\ell-1}(x, \tilde{x}) B_{\ell+1}(x, \tilde{x}) \Gamma_{\sigma'}(K_{\ell-1})(x, \tilde{x})\big]$, where K_ℓ's are defined in Proposition 3; $B_{L+1}(x, \tilde{x}) = 1$, and for $\ell = 1, \cdots, L$, B_ℓ's are defined as*

$$B_{\ell+1}(x, \tilde{x}) = B_{\ell+2}(x, \tilde{x}) + \alpha^2 B_{\ell+2}(x, \tilde{x}) \Gamma_{\sigma'}(K_\ell)(x, \tilde{x}).$$

Proposition 4 implies that similar to what has been proved for the FFNet, the ResNet trained by gradient flow is also equivalent to the kernel regression predictor with some NTK. Note that Proposition 4 is an asymptotic result. We defer the proof, as it can be straightforwardly derived from the nonasymptotic bound as follows.

Theorem 4. *For the ResNet defined in* (5)*, given two inputs on the unit sphere $x, \tilde{x} \in \mathbb{S}^{D-1}$, $\epsilon < 0.5$, and*

$$m \geq C\epsilon^{-4} L^{2-2\gamma} \big(\log(320(L^2+1)/\delta) + 1\big),$$

where C is a constant, with probability at least $1 - \delta$ over the randomness of the initialization, we have

$$\big|\big\langle \nabla_\theta f, \nabla_\theta \tilde{f} \big\rangle - \Omega_L(x, \tilde{x})\big| \leq 2L\alpha^2\epsilon,$$

where $\alpha = L^{-\gamma}$ with $\gamma \in [0.5, 1]$, $\Omega_L(x, \tilde{x})$ is defined in Proposition 4.

Proof Sketch of Proposition 4 and Theorem 4. For simplicity, we use $\phi_W : \mathbb{R}^m \to \mathbb{R}^m$ to denote $\phi_W(z) = \sqrt{\frac{2}{m}}\sigma_0(Wz)$. Then its derivative w.r.t. z is as follows, $\phi'_W(z) = \sqrt{\frac{2}{m}} D(Wz) W$, where $D(Wz)$ is an operator defined as $D(Wz) \equiv \mathrm{diag}(\sigma_0'(Wz)) = \mathrm{diag}([\mathbb{I}_{\{W_{1,\cdot} z \geq 0\}}, \cdots, \mathbb{I}_{\{W_{m,\cdot} z \geq 0\}}]^\top)$.

For simplicity, we denote $D_\ell = D(W_\ell x_{\ell-1})$, where $\ell = 1, 2, \cdots, L$. Note that D_ℓ is essentially the activation pattern of the ℓ-th bottleneck layer on the input x. We denote $\widetilde{D}_\ell$ for $\tilde{x}$ in a similar fashion. Then we have $\frac{\partial x_\ell}{\partial x_{\ell-1}} = I_m + \alpha\sqrt{\frac{1}{m}} V_\ell \sqrt{\frac{2}{m}} D_\ell W_\ell$. For $\ell = 1, \cdots, L$, we denote $b_{\ell+1} = \nabla_{x_\ell} f$. Then we have $b_{\ell+1} = \big(v^\top \frac{\partial x_L}{\partial x_{L-1}} \frac{\partial x_{L-1}}{\partial x_{L-2}} \cdots \frac{\partial x_{\ell+1}}{\partial x_\ell}\big)^\top$.

Combining all above derivations, we have $\nabla_{V_\ell} f = \frac{\alpha}{\sqrt{m}} b_{\ell+1} \cdot (\phi_{W_\ell}(x_{\ell-1}))^\top$, and $\nabla_{W_\ell} f = \frac{\alpha}{\sqrt{m}} \sqrt{\frac{2}{m}} D_\ell V_\ell^\top b_{\ell+1} \cdot x_{\ell-1}^\top$. Then we can derive the kernel $\sum_{\ell=1}^{L} \langle \nabla_{W_\ell} f, \nabla_{W_\ell} \tilde{f} \rangle + \sum_{\ell=1}^{L} \langle \nabla_{V_\ell} f, \nabla_{V_\ell} \tilde{f} \rangle$, where $\langle \nabla_{V_\ell} f, \nabla_{V_\ell} \tilde{f} \rangle = \alpha^2 \underbrace{\frac{1}{m} \langle b_{\ell+1}, \tilde{b}_{\ell+1} \rangle}_{T_{\ell,1}} \underbrace{\langle \phi_{W_\ell}(x_{\ell-1}), \phi_{W_\ell}(\tilde{x}_{\ell-1}) \rangle}_{T_{\ell,2}}$,

$\langle \nabla_{W_\ell} f, \nabla_{W_\ell} \tilde{f} \rangle = \alpha^2 \underbrace{\langle x_{\ell-1}, \tilde{x}_{\ell-1} \rangle}_{T_{\ell,3}} \underbrace{\frac{2}{m^2} \tilde{b}_{\ell+1}^\top V_\ell \widetilde{D}_\ell D_\ell V_\ell^\top b_{\ell+1}}_{T_{\ell,4}}$. Note that the concentration of $T_{\ell,3}$ can be shown by Theorem 3. We then show the concentration of $T_{\ell,1}$, $T_{\ell,2}$ and $T_{\ell,4}$, respectively.

For simplicity, we define two matrices for each layer,

$$\widehat{\Sigma}_\ell(x, \tilde{x}) = \begin{bmatrix} \langle x_\ell, x_\ell \rangle & \langle x_\ell, \tilde{x}_\ell \rangle \\ \langle \tilde{x}_\ell, x_\ell \rangle & \langle \tilde{x}_\ell, \tilde{x}_\ell \rangle \end{bmatrix}, \ \Sigma_\ell(x, \tilde{x}) = \begin{bmatrix} K_\ell(x, x) & K_\ell(x, \tilde{x}) \\ K_\ell(\tilde{x}, x) & K_\ell(\tilde{x}, \tilde{x}) \end{bmatrix}.$$

We define $\psi_\sigma : \mathcal{M}_+^2 \to \mathbb{R}$ as $\psi_\sigma(\Sigma) = \mathbb{E}_{(X,\tilde{X}) \sim \mathcal{N}(0,\Sigma)} \sigma(X)\sigma(\tilde{X})$ and $\psi_{\sigma'} : \mathcal{M}_+^2 \to \mathbb{R}$ as $\psi_{\sigma'}(\Sigma) = \mathbb{E}_{(X,\tilde{X}) \sim \mathcal{N}(0,\Sigma)} \sigma'(X)\sigma'(\tilde{X})$. Note $\Gamma_\sigma(K_{\ell-1}) = \psi_\sigma(\Sigma_{\ell-1})$ and $\Gamma_{\sigma'}(K_{\ell-1}) = \psi_{\sigma'}(\Sigma_{\ell-1})$.

The following lemmas are technical results and very involved. Please see Appendix B for details.

Lemma 1. *Suppose that for $\ell = 1, \cdots, L$,*

$$\|\widehat{\Sigma}_{\ell-1}(x, \tilde{x}) - \Sigma_{\ell-1}(x, \tilde{x})\|_{\max} \leq c\epsilon^2, \ \ m \geq C_1 \epsilon^{-2} L^{2-2\gamma} \big(\log(80L^2/\delta) + 1 \big), \tag{8}$$

with probability at least $1 - 3\delta$, we have $|T_{\ell,1} - B_{\ell+1}(x, \tilde{x})| \leq c_1\epsilon$, for $\ell = 1, \cdots, L$, where C_1, c_1, and c are constants.

Lemma 2. *Suppose* (8) *holds for $\ell = 1, \cdots, L$. With probability at least $1 - \delta$, we have $|T_{\ell,2} - \Gamma_\sigma(K_{\ell-1})(x, \tilde{x})| \leq c_2\epsilon$, for $\ell = 1, \cdots, L$, where C_2 and c_2 are constants.*

Lemma 3. *Suppose that* (8) *holds for $\ell = 1, \cdots, L$. With probability at least $1 - 3\delta$, we have $|T_{\ell,4} - B_{\ell+1}(x, \tilde{x}) \Gamma_{\sigma'}(K_{\ell-1})(x, \tilde{x})| \leq c_3\epsilon$, for $\ell = 1, \cdots, L$, where c_3 is a constant.*

We remark: (1) Lemma 1 is proved by reverse induction; (2) Lemma 2 exploits the concentration properties of W_ℓ and local Lipschitz properties of ψ_σ; (3) We prove Lemma 3 and Lemma 1 simultaneously with the Hölder continuity of $\psi_{\sigma'}$. Combining all results above, we complete Theorem 4. Moreover, taking $m \to \infty$, we have Proposition 4. □

4 Deep Feedforward v.s. Residual Networks

To compare the NTKs associated with deep FFNets and ResNets, we consider proper normalization, which avoids the kernel function blowing up or vanishing as the depth L goes to infinity.

4.1 The Limiting NTK of the Feedforward Networks

Recall that the NTK of the L-layer FFNet defined in (4) is $\Omega_L(x, \tilde{x}) = \sum_{\ell=1}^L \big[K_{\ell-1}(x, \tilde{x}) \cdot \prod_{i=\ell}^L \Gamma_{\sigma'}(K_{i-1})(x, \tilde{x}) \big]$. One can check that $\Omega_L(x, x) = L$ for all $x \in \mathbb{S}^{D-1}$. To avoid $\Omega_L(x, x) \to \infty$, as $L \to \infty$. We consider a normalized version as

$$\overline{\Omega}_L(x, \tilde{x}) = \frac{1}{L} \Omega_L(x, \tilde{x}).$$

We characterize the impact of the depth L on the NTK in the following theorem.

Theorem 5. *For the NTK of the FFNet, as $L \to \infty$, given $x, \tilde{x} \in \mathbb{S}^{D-1}$ and $|1 - x^\top \tilde{x}| \geq \delta > 0$, where δ is a constant and does not scale with L, we have*

$$\Big| \overline{\Omega}_L(x, \tilde{x}) - 1/4 \Big| = \mathcal{O}\Big(\frac{\mathrm{polylog}(L)}{L} \Big),$$

When $x = \tilde{x}$, we have $\overline{\Omega}_L(x, \tilde{x}) = 1, \forall L$.

Proof Sketch of Theorem 5 . The main challenge comes from the sophisticated recursion of the kernel. To handle the recursion, we employ the following bound.

Lemma 4. *When L is large enough, we have*

$$\cos\left(\pi \left(1 - \left(\frac{n}{n+1} \right)^{3 + \frac{\log(L)^2}{L}} \right) \right) \leq K_n(x, \tilde{x}) \leq \cos\left(\pi \left(1 - \left(\frac{n + \log(L)^p}{n + \log(L)^p + 1} \right)^{3 - \frac{\log(L)^2}{L}} \right) \right),$$

where p is a positive constant depending on δ.

By Lemma 4, we can further bound $\prod_{i=\ell}^{L}\Gamma_{\sigma'}(K_{i-1}(x,\tilde{x}))$ by

$$\Big(\frac{\ell-1}{L}\Big)^{3+\frac{\log(L)^2}{L}} \leq \prod_{i=\ell}^{L}\Gamma_{\sigma'}(K_{i-1}(x,\tilde{x})) \leq \Big(\frac{\ell+\log(L)^p-1}{L+\log(L)^p}\Big)^{3-\frac{\log(L)^2}{L}} \tag{9}$$

Hence we can measure the rate of convergence. The detailed proof is the following. □

As can be seen from Theorem 5, the NTK of the FFNet converges to a limiting form, i.e.,

$$\overline{\Omega}_\infty(x,\tilde{x}) = \lim_{L\to\infty}\overline{\Omega}_L(x,\tilde{x}) = \begin{cases} 1/4, & x \neq \tilde{x} \\ 1, & x = \tilde{x} \end{cases}.$$

For simplicity, we refer to $\overline{\Omega}_\infty$ as the limiting NTK of the FFNets.

The limiting NTK of the FFNets is actually a non-informative kernel. For example, we consider a kernel regression problem with n independent observations $\{(x_i,y_i)\}_{i=1}^n$, where $x_i \in \mathbb{R}^D$ is the feature vector, and $y_i \in \mathbb{R}$ is the response. Without loss of generality, we assume that the training samples have been properly processed such that $x_i \neq x_j$ for $i \neq j$, and $\sum_{i=1}^n y_i = 0$. By the Representer theorem [34], we know that the kernel regression function can be represented by $f(\cdot) = \sum_{i=1}^n \beta_i \overline{\Omega}_\infty(x_i,\cdot)$. We then minimize the regularized empirical risk as follows.

$$\hat{\beta} = \min_{\beta} \|y - \widetilde{\Omega}\beta\|^2 + \lambda\beta^\top\widetilde{\Omega}\beta, \tag{10}$$

where $\beta = (\beta_1,...\beta_n)^\top \in \mathbb{R}^n$, $y = (y_1,...,y_n)^\top \in \mathbb{R}^n$, $\widetilde{\Omega} \in \mathbb{R}^{n\times n}$ with $\widetilde{\Omega}_{ij} = \overline{\Omega}_\infty(x_i,x_j)$, and λ is the regularization parameter and usually very small for large n. One can check that (10) admits a closed form solution $\hat{\beta} = (\widetilde{\Omega}+\lambda I_n)^{-1}y$. Note that we have $\tilde{\Omega}+\lambda I_n = 1/4 J_n + (\lambda+3/4)I_n$, which is the sum of a diagonal matrix and a rank-one matrix and J_n is $n\times n$ all-ones matrix. By Sherman – Morrison formula

$$(A+uv^\top)^{-1} = A^{-1} - \frac{A^{-1}uv^\top A^{-1}}{1+v^\top A^{-1}u}, \text{ we have } \hat{\beta} = \frac{1}{\lambda+3/4}\Big(I_n - \frac{1}{n+4\lambda+3}J_n\Big)y.$$

Then we further have $f(x_j) = \sum_{i=1}^n \hat{\beta}_i\overline{\Omega}_\infty(x_i,x_j) = \frac{3}{4\lambda+3}y_j$.

As can be seen, for sufficiently large n and sufficiently small λ, we have $f(x_j) \approx y_j$, which means that we can fit the training data well. However, for an unseen data point x^*, where $x^* \neq x_1,...,x_n$, the regression function f always gives an output 0, i.e.,

$$f(x^*) = \sum_{i=1}^n \hat{\beta}_i\overline{\Omega}_\infty(x_i,x^*) = \frac{1}{4}\sum_{i=1}^n\hat{\beta}_i = 0.$$

This indicates that the function class induced by the limiting NTK of the FFNets $\overline{\Omega}_\infty$ is not learnable.

4.2 The Limiting NTK of the Residual Networks

Recall that the infinite-width NTK of the L-layer ResNet is

$$\Omega_L(x,\tilde{x}) = \alpha^2\sum_{\ell=1}^{L}\Big[B_{\ell+1}(x,\tilde{x})\Gamma_\sigma(K_{\ell-1})(x,\tilde{x}) + K_{\ell-1}(x,\tilde{x})B_{\ell+1}(x,\tilde{x})\Gamma_{\sigma'}(K_{\ell-1})(x,\tilde{x})\Big],$$

where $B_{L+1}(x,\tilde{x}) = 1$ and for $\ell = 1,..,L-1$, $B_{\ell+1}(x,\tilde{x}) = \prod_{i=\ell}^{L-1}(1+\alpha^2\Gamma_{\sigma'}(K_i)(x,\tilde{x}))$. One can check that for $x \in \mathbb{S}^{D-1}$, $\Omega_L(x,x) = 2L\alpha^2(1+\alpha^2)^{L-1}$.

Different from the NTK of the FFNet, $\Omega_L(x,x) \to 0$ as $L\to\infty$. Therefore, we also consider the normalized NTK for the ResNet to prevent the kernel from vanishing. Specifically, the normalized NTK of the ResNet on $\mathbb{S}^{D-1}\times\mathbb{S}^{D-1}$, $\overline{\Omega}_L(x,\tilde{x})$, is defined as follows,

$$\frac{1/(2L)}{(1+\alpha^2)^{L-1}}\sum_{\ell=1}^{L}\Big[B_{\ell+1}(x,\tilde{x})\Gamma_\sigma(K_{\ell-1})(x,\tilde{x}) + K_{\ell-1}(x,\tilde{x})B_{\ell+1}(x,\tilde{x})\Gamma_{\sigma'}(K_{\ell-1})(x,\tilde{x})\Big]. \tag{11}$$

We then analyze the limiting NTK of the ResNets. Recall that $\alpha = L^{-\gamma}$. Our next theorem only considers $\gamma = 1$, i.e., $\alpha = 1/L$.

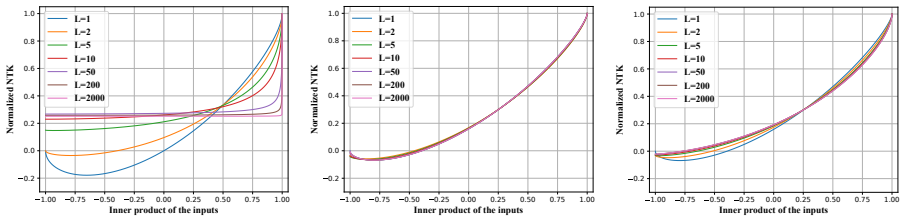
Theorem 6. *For the NTK of the ResNet, as $L\to\infty$, given $\alpha = \frac{1}{L}$ and $x,\tilde{x} \in \mathbb{S}^{D-1}$ such that $|1-x^\top\tilde{x}| \geq \delta > 0$, where δ is a constant and does not scale with L, we have*

$$\big|\overline{\Omega}_L(x,\tilde{x}) - \overline{\Omega}_1(x,\tilde{x})\big| = \mathcal{O}\left(1/L\right),$$

where $\overline{\Omega}_1(x,\tilde{x}) = \frac{1}{2}\Big(\hat{\sigma}(x^\top\tilde{x}) + x^\top\tilde{x}\cdot\hat{\sigma'}(x^\top\tilde{x})\Big)$.

Proof Sketch of Theorem 6. The main technical challenge here is also handling the recursion. Specifically, we denote $K_{\ell,L}$ to be the ℓ-th layer of the GP kernel when the depth is L, which is originally denoted by $K_\ell(x, \tilde{x})$. Let $S_0 = K_0(x, \tilde{x})$ and $S_{\ell,L} = \frac{K_{\ell,L}}{(1+\alpha^2)^\ell} = \frac{K_{\ell,L}}{(1+1/L^2)^\ell}$. We have $\Gamma_\sigma(K_{\ell,L}) = (1+\alpha^2)^\ell \hat{\sigma}(S_{\ell,L})$ and $\Gamma_{\sigma'}(K_{\ell,L}) = \widehat{\sigma'}(S_{\ell,L})$. We rewrite the recursion of $K_{\ell,L}$ as $S_{\ell,L} = \frac{S_{\ell-1,L}+\alpha^2\hat{\sigma}(S_{\ell-1,L})}{(1+\alpha^2)} \geq S_{\ell-1,L}$, which eases the technical difficulty. However, the proof is still highly involved, and more details can be found in Appendix E. □

Note that we do not consider $\gamma = 0.5$ for technical concerns, as $\overline{\Omega}_L(x, \tilde{x})$ in (11) becomes very complicated to compute, as $L \to \infty$. Also we find that considering $\gamma = 1$ is sufficient to provide us new th



(a) *FFNets* (b) *ResNets with* $\gamma = 1$ (c) *ResNets with* $\gamma = 0.5$

Figure 1: *Normalized Neural Tangent Kernels Associated with Different Deep Networks.*

Different from FFNets, the class of functions induced by the NTKs of the ResNets does not significantly change, as the depth L increases. Surprisingly, we actually have $\overline{\Omega}_\infty = \overline{\Omega}_1$ for $\alpha = 1/L$, i.e., infinitely deep and 1-layer ResNets induce the same NTK. To further visualize such a difference, we plot the NTKs of the ResNets in Fig. 1(b) and 1(c) for $\alpha = 1/L$ and $\alpha = 1/\sqrt{L}$, respectively. As can be seen, the increase of the depth yields very small changes to the NTKs. This partially explains why increasing the depth of the ResNet does not significantly deteriorate the generalization.

Moreover, as long as $x \neq \widetilde{x}$, i.e., $\langle x, \widetilde{x}\rangle \neq 1$, the limiting NTK of the FFNets always yields $1/4$ regardless how different x is from $\widetilde{x}$. In contrast, the residual networks do not suffer from this drawback. The limiting NTK of the ResNets can greatly distinguish the difference between x and $\widetilde{x}$, e.g., $\langle x, \widetilde{x}\rangle = -0.5$, 0, and 0.5 yield different values. Therefore, for an unseen data point, the corresponding regression model does not always output 0, which is in sharp contrast to that of the limiting NTK of the FFNets.

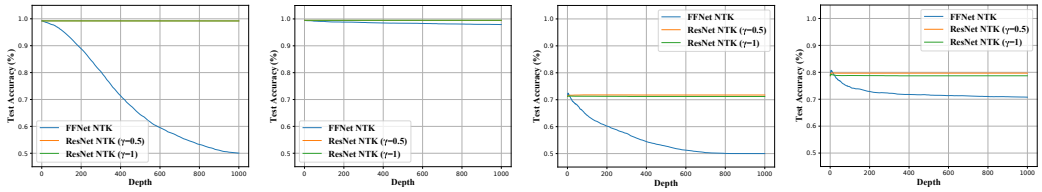
5 Experiments

We demonstrate the generalization properties of the kernel regression based on the NTKs of the FFNets and the ResNets with varying depths. Our experiments follow similar settings to [22, 23]. We adopt two widely used data sets – MNIST [35] and CIFAR10 [36], which are popular in existing literature. Note that both MNIST and CIFAR10 contains 10 classes of images. For simplicity, we select 2 classes out of 10 (digits "0" and "8" for MNIST, categories "airplane" and "ship" for CIFAR10), respectively, which results in two binary classification problems, denoted by MNIST2 and CIFAR2.

Similar to [22, 23], we use the kernel regression model for classification. Specifically, given the training data $(x_1, y_1), \cdots, (x_n, y_n)$, where $x_i \in \mathbb{R}^D$ and $y_i \in \{-1, +1\}$ for $i = 1, ..., n$, we compute the kernel matrix $\tilde{K} = [\tilde{K}_{ij}]_{i,j=1}^n$ using the NTKs associated with the FFNets and the ResNets, where $\tilde{K}_{ij} = \overline{\Omega}_L(x_i, x_j)$. Then we compute the kernel regression function $f(x) = \sum_{i=1}^n \alpha_i \overline{\Omega}_L(x, x_i)$, where $[\alpha_1, ..., \alpha_n]^\top = (\tilde{K} + \lambda I)^{-1} y$, $y = [y_1, ..., y_n]^\top$ and $\lambda = 0.1/n$ is a very small constant. We predict the label of x to be $\mathrm{sign}(f(x))$.

Our experiments adopt the NTKs associated with three network architectures: (1) FFNets, (2) ResNets ($\gamma = 0.5$) and (3) ResNets ($\gamma = 1$). We set $n = 200$ and $n = 2000$. For each data set, we randomly select n training data points ($n/2$ for each class) and 2000 testing data points (1000 for each class). When training the kernel regression models, we normalize all training data points to have zero mean and unit norm. We repeat the procedure for 20 simulations. We find that the training errors of all simulations (L varies from 1 to 2000) are 0.0, which means that all NTK-based models are sufficient to overfit the training data, regardless $n = 200$ or $n = 2000$. The test accuracies of the kernel regression models with different kernels and depths are shown in Figure 2.

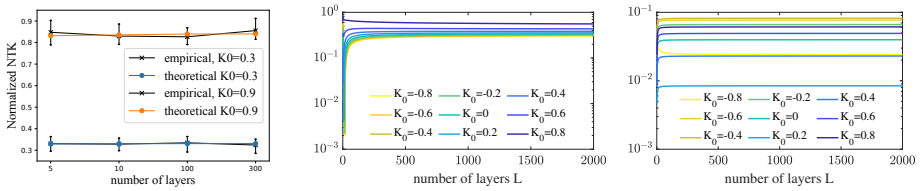
As can be seen, the test accuracies of the kernel regression models of ResNets (both $\gamma = 0.5$ and $\gamma = 1$) are not sensitive to the depth. In contrast, the test accuracies of the kernel regression models of the FFNets significantly decrease, as the depth L increases. Especially when the sample size is small ($n = 200$), the kernel regression models behave like random guess for both MNIST2 and CIFAR2 when $L > 1000$. This is consistent with our analysis.



(a) *MNIST2* ($n = 200$) (b) *MNIST2* ($n = 2000$) (c) *CIFAR2* ($n = 200$) (d) *CIFAR2* ($n = 2000$)

Figure 2: *Test accuracies of the kernel regression models evaluated on MNIST2 and CIFAR2.*

Next we provide numerical verifications for our theorems. For Theorem 4, we randomly initialize the ResNet with width=500, scaling factor $\gamma = 1$ and depth $L = 5, 10, 100, 300$, and then calculate the inner product of the Jacobians of the ResNet for two different inputs as in the definition of NTK. We repeat the procedure for 500 times and plot the mean value (black cross) and the 1/4, 3/4 quantiles ("I"-shape line) of the sampled random NTKs and the theoretical NTK value in Fig. 3(a), which shows the two results match very well. For Theorem 5 and Theorem 6, Fig. 3(b) and Fig. 3(c) show that $\lim_{L\to\infty} |\overline{\Omega}_L(x, \tilde{x}) - 1/4| \cdot L/\log(\mathrm{L}) \approx$ constant and $\lim_{L\to\infty} |\overline{\Omega}_L(x, \tilde{x}) - \overline{\Omega}_1(x, \tilde{x})| \cdot L \approx$ constant with $x^\top \tilde{x} = K_0$ chosen at 9 points.



(a) *Theorem 4* (b) *Theorem 5* (c) *Theorem 6*

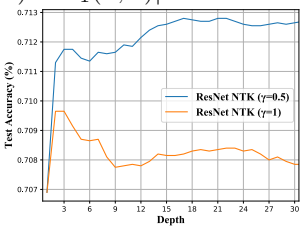
Figure 3: *Verification of main theorems. (a) Theorem 4,* $m = 500$ *and scaling* $\gamma = 1$*; (b) Theorem 5, y-axis is* $|\overline{\Omega}_L(x, \tilde{x}) - 1/4| \cdot L/\log(\mathrm{L})$*; (c) Theorem 6, y-axis is* $|\overline{\Omega}_L(x, \tilde{x}) - \overline{\Omega}_1(x, \tilde{x})| \cdot L$

6 Discussion

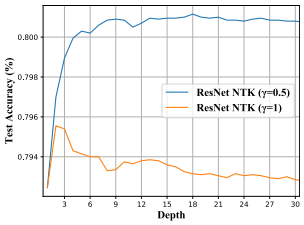
We discuss the NTK of the ResNet in more details. We remark unless specified, the NTK mentioned below indicates the normalized NTK.

Our theory shows the function class induced by the NTK of the deep ResNet asymptotically converges to that by the NTK of the 1-layer ResNet, as the depth increases. This indicates that the complexity of such a function class is not significantly different from that by the NTK of the 1-layer ResNet, for large enough L. Thus, the generalization gap does not significantly increase, as L increases.

On the other hand, our experiments suggest that, as illustrated in Figure 4, the NTK of the ResNet with $\gamma = 1$ actually achieves the best testing accuracy for CIFAR2 when $L = 2$. The accuracy slightly decreases as L increases, and becomes stable when $L \geq 9$. For the NTK of the ResNet with $\gamma = 0.5$, the accuracy achieves the best when $L \approx 15$, and becomes stable for $L \geq 15$. Such evidence suggests that the function class induced by the NTKs of the ResNets with large L and large γ are possibly not as flexible as those by the NTKs of the deep ResNets with small L and small γ.



(a) *CIFAR2* (n=200)



(b) *CIFAR2* (n=2000)

Figure 4: *Test accuracies of the kernel regression models evaluated on CIFAR2.*

Existing literature connects overparameterized neural networks to NTKs only under some very specific regime. Practical neural networks, however, are trained under more complicated regimes. Therefore, there still exists a significant theoretical gap between NTKs and practical neural networks. For example, Theorem 6 shows that the NTK of the infinitely deep ResNet is identical to that of the 1-layer ResNet, while practical ResNets often show better generalization performance, as the depth increases. Also, we do not consider batch norm in our networks but refer to [37] if necessary. We will leave these challenges for future investigation.

Broader Impact

This paper makes a significant contribution to extending the frontier of deep learning theory, and increases the intellectual rigor. To the best of our knowledge, our results are the first one for analyzing the effect of depth on the generalization of neural tangent kernels (NTKs). Moreover, our results are also the first one establishing the non-asymptotic bounds for NTKs of ResNets when all but the last layers are trained, which enables us to successfully analyze the generalization properties of ResNets through the perspective of NTK. This is in sharp contrast to the existing impractical theoretical results for NTKs of ResNets, which either only apply to an over-simplified structure of ResNets or only deal with the case when the last layer is trained.

Acknowledgement

Molei Tao was partially supported by NSF DMS-1847802 and ECCS-1936776 and Yuqing Wang was partially supported by NSF DMS-1847802.

References

[1] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, pages 1097–1105, 2012.

[2] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, pages 2672–2680, 2014.

[3] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

[4] Alex Graves, Abdel-rahman Mohamed, and Geoffrey Hinton. Speech recognition with deep recurrent neural networks. In *2013 IEEE international conference on acoustics, speech and signal processing*, pages 6645–6649. IEEE, 2013.

[5] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

[6] Tom Young, Devamanyu Hazarika, Soujanya Poria, and Erik Cambria. Recent trends in deep learning based natural language processing. *ieee Computational intelligenCe magazine*, 13(3):55–75, 2018.

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[8] Rupesh K Srivastava, Klaus Greff, and Jürgen Schmidhuber. Training very deep networks. In *Advances in Neural Information Processing Systems*, pages 2377–2385, 2015.

[9] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In *European conference on computer vision*, pages 630–645. Springer, 2016.

[11] Andreas Veit, Michael J Wilber, and Serge Belongie. Residual networks behave like ensembles of relatively shallow networks. In *Advances in Neural Information Processing Systems*, pages 550–558, 2016.

[12] David Balduzzi, Marcus Frean, Lennox Leary, JP Lewis, Kurt Wan-Duo Ma, and Brian McWilliams. The shattered gradients problem: If resnets are the answer, then what is the question? *arXiv preprint arXiv:1702.08591*, 2017.

[13] Hao Li, Zheng Xu, Gavin Taylor, Christoph Studer, and Tom Goldstein. Visualizing the loss landscape of neural nets. In *Advances in Neural Information Processing Systems*, pages 6389–6399, 2018.

[14] Zeyuan Allen-Zhu, Yuanzhi Li, and Zhao Song. A convergence theory for deep learning via over-parameterization. *arXiv preprint arXiv:1811.03962*, 2018.

[15] Huishuai Zhang, Da Yu, Mingyang Yi, Wei Chen, and Tie-Yan Liu. Convergence theory of learning over-parameterized resnet: A full characterization, 2019.

[16] Moritz Hardt and Tengyu Ma. Identity matters in deep learning. *arXiv preprint arXiv:1611.04231*, 2016.

[17] Yuanzhi Li and Yang Yuan. Convergence analysis of two-layer neural networks with relu activation. In *Advances in neural information processing systems*, pages 597–607, 2017.

[18] Tianyi Liu, Minshuo Chen, Mo Zhou, Simon S Du, Enlu Zhou, and Tuo Zhao. Towards understanding the importance of shortcut connections in residual networks. In *Advances in Neural Information Processing Systems*, pages 7890–7900, 2019.

[19] Simon S Du, Jason D Lee, Haochuan Li, Liwei Wang, and Xiyu Zhai. Gradient descent finds global minima of deep neural networks. *arXiv preprint arXiv:1811.03804*, 2018.

[20] Spencer Frei, Yuan Cao, and Quanquan Gu. Algorithm-dependent generalization bounds for overparameterized deep residual networks. In *Advances in Neural Information Processing Systems*, pages 14769–14779, 2019.

[21] Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in Neural Information Processing Systems*, pages 8571–8580, 2018.

[22] Sanjeev Arora, Simon S Du, Wei Hu, Zhiyuan Li, Ruslan Salakhutdinov, and Ruosong Wang. On exact computation with an infinitely wide neural net. *arXiv preprint arXiv:1904.11955*, 2019.

[23] Sanjeev Arora, Simon S Du, Wei Hu, Zhiyuan Li, and Ruosong Wang. Fine-grained analysis of optimization and generalization for overparameterized two-layer neural networks. *arXiv preprint arXiv:1901.08584*, 2019.

[24] Zeyuan Allen-Zhu and Yuanzhi Li. Can sgd learn recurrent neural networks with provable generalization? *arXiv preprint arXiv:1902.01028*, 2019.

[25] Zeyuan Allen-Zhu, Yuanzhi Li, and Yingyu Liang. Learning and generalization in overparameterized neural networks, going beyond two layers. *arXiv preprint arXiv:1811.04918*, 2018.

[26] Yuanzhi Li and Yingyu Liang. Learning overparameterized neural networks via stochastic gradient descent on structured data. In *Advances in Neural Information Processing Systems*, pages 8157–8166, 2018.

[27] Difan Zou, Yuan Cao, Dongruo Zhou, and Quanquan Gu. Stochastic gradient descent optimizes over-parameterized deep relu networks. arxiv e-prints, art. *arXiv preprint arXiv:1811.08888*, 2018.

[28] Amit Daniely, Roy Frostig, and Yoram Singer. Toward deeper understanding of neural networks: The power of initialization and a dual view on expressivity. In *Advances In Neural Information Processing Systems*, pages 2253–2261, 2016.

[29] Youngmin Cho and Lawrence K. Saul. Kernel methods for deep learning. In Y. Bengio, D. Schuurmans, J. D. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 342–350. Curran Associates, Inc., 2009.

[30] Julien Mairal, Piotr Koniusz, Zaid Harchaoui, and Cordelia Schmid. Convolutional kernel networks. In *Advances in neural information processing systems*, pages 2627–2635, 2014.

[31] Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. In *Advances in neural information processing systems*, pages 1177–1184, 2008.

[32] Jaehoon Lee, Yasaman Bahri, Roman Novak, Samuel S Schoenholz, Jeffrey Pennington, and Jascha Sohl-Dickstein. Deep neural networks as gaussian processes. *arXiv preprint arXiv:1711.00165*, 2017.

[33] Adrià Garriga-Alonso, Carl Edward Rasmussen, and Laurence Aitchison. Deep convolutional networks as shallow gaussian processes. *arXiv preprint arXiv:1808.05587*, 2018.

[34] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. *The elements of statistical learning*, volume 1. Springer series in statistics New York, 2001.

[35] Yann LeCun. The mnist database of handwritten digits. *http://yann.lecun.com/exdb/mnist/*, 1998.

[36] Alex Krizhevsky, Vinod Nair, and Geoffrey Hinton. Cifar-10 and cifar-100 datasets. *URl: https://www. cs. toronto. edu/kriz/cifar. html (vi sited on Mar. 1, 2016)*, 2009.

[37] Arthur Jacot, Franck Gabriel, Franccois Gaston Ged, and Clément Hongler. Order and chaos: Ntk views on dnn normalization, checkerboard and boundary artifacts. 2019.

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

[39] Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of machine learning*. MIT press, 2018.

[40] Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford University Press, 2013.

A Proof of GP Kernels of ResNets

A.1 Notation and Main Idea

For a fixed pair of inputs x and $\tilde{x}$, we introduce two matrices for each layer

$$\hat{\Sigma}_\ell(x,\tilde{x}) = \begin{bmatrix} \langle x_\ell, x_\ell\rangle & \langle x_\ell, \tilde{x}_\ell\rangle \\ \langle \tilde{x}_\ell, x_\ell\rangle & \langle \tilde{x}_\ell, \tilde{x}_\ell\rangle \end{bmatrix},$$

and

$$\Sigma_\ell(x,\tilde{x}) = \begin{bmatrix} K_\ell(x,x) & K_\ell(x,\tilde{x}) \\ K_\ell(\tilde{x},x) & K_\ell(\tilde{x},\tilde{x}) \end{bmatrix}.$$

$\hat{\Sigma}_\ell(x,\tilde{x})$ is the empirical Gram matrix of the outputs of the ℓ-th layer, while $\Sigma_\ell(x,\tilde{x})$ is the infinite-width version. Theorem 3 says that with high probability, for each layer ℓ, the difference of these two matrices measured by the entry-wise L_∞ norm (denoted by $\|\cdot\|_{\max}$) is small.

The idea is to bound how much the ℓ-th layer magnifies the input error to the output. Specifically, if the outputs of $(\ell-1)$-th layer satisfy

$$\left\|\hat{\Sigma}_{\ell-1}(x,\tilde{x}) - \Sigma_{\ell-1}(x,\tilde{x})\right\|_{\max} \le \tau,$$

we hope to prove that with high probability over the randomness of W_ℓ and V_ℓ, we have

$$\left\|\hat{\Sigma}_{\ell}(x,\tilde{x}) - \Sigma_{\ell}(x,\tilde{x})\right\|_{\max} \le \left(1+\mathcal{O}\left(\frac{1}{L}\right)\right)\tau.$$

Then the theorem is proved by first showing that w.h.p. $\left\|\hat{\Sigma}_0(x,\tilde{x}) - \Sigma_0(x,\tilde{x})\right\|_{\max} \le (1+\mathcal{O}(1/L))^{-L}\epsilon$ and then applying the result above for each layer.

A.2 Lemmas

We introduce the following lemmas. The first lemma shows the boundedness of $K_\ell(x,\tilde{x})$.

Lemma 5. *For the ResNet defined in Eqn.* (5)*, $K_\ell(x,x) = (1+\alpha^2)^\ell$ for all $x \in \mathbb{S}^{D-1}$, $\ell = 0,1,\cdots,L$. Also $K_\ell(x,x)$ is bounded uniformly when $0.5 \le \gamma \le 1$.*

Recall that $\phi_{W_\ell}(z) = \sqrt{\frac{2}{m}}\sigma_0(W_\ell z)$. Since W_ℓ is Gaussian, we know that $\phi_{W_\ell}(x_{\ell-1})$ and $\phi_{W_\ell}(\tilde{x}_{\ell-1})$ are both sub-Gaussian random vectors over the randomness of W_ℓ. Then their inner product enjoys sub-exponential property.

Lemma 6 (Sub-exponential concentration)**.** *With probability at least $1-\delta'$ over the randomness of $W_\ell \sim \mathcal{N}(0,I)$, when $m \ge c'\log(6/\delta')$, the following hold simultaneously*

$$\left|\langle \phi_{W_\ell}(x_{\ell-1}), \phi_{W_\ell}(\tilde{x}_{\ell-1})\rangle - \psi_\sigma(\hat{\Sigma}_{\ell-1}(x,\tilde{x}))\right| \le \sqrt{\frac{c'\log(6/\delta')}{m}}\|x_{\ell-1}\|\|\tilde{x}_{\ell-1}\|, \tag{12}$$

$$\left|\|\phi_{W_\ell}(x_{\ell-1})\|^2 - \|x_{\ell-1}\|^2\right| \le \sqrt{\frac{c'\log(6/\delta')}{m}}\|x_{\ell-1}\|^2, \tag{13}$$

$$\left|\|\phi_{W_\ell}(\tilde{x}_{\ell-1})\|^2 - \|\tilde{x}_{\ell-1}\|^2\right| \le \sqrt{\frac{c'\log(6/\delta')}{m}}\|\tilde{x}_{\ell-1}\|^2. \tag{14}$$

Lemma 7 (Locally Lipschitzness, based on [28])**.** *ψ_σ is $(1+\frac{1}{\pi}(\frac{r}{\mu})^2)$-Lipschitz w.r.t. $\max$ norm in $\mathcal{M}_{\mu,r} = \left\{\begin{bmatrix} a & b \\ b & c\end{bmatrix} | a,c \in [\mu-r,\mu+r]; ac-b^2>0\right\}$ for all $\mu>0$, $0<r\le\mu/2$. That means, if (i). $\|\hat{\Sigma}_{\ell-1}(x,\tilde{x}) - \Sigma_{\ell-1}(x,\tilde{x})\|_{\max} \le \tau$ and (ii). $K_{\ell-1}(x,x) = K_{\ell-1}(\tilde{x},\tilde{x}) = \mu$, for $\tau \le \mu/2$, we have*

$$\left|\psi_\sigma(\hat{\Sigma}_{\ell-1}(x,\tilde{x})) - \psi_\sigma(\Sigma_{\ell-1}(x,\tilde{x}))\right| \le \left(1+\frac{1}{\pi}\left(\frac{\tau}{\mu}\right)^2\right)\tau.$$

A.3 Proof of Theorem 3

Proof. In this proof, we also show the following hold with the same probability.

1. For $\ell = 0,1,\cdots,L$, $\|x_\ell\|$ and $\|\tilde{x}_\ell\|$ are bounded by an absolute constant C_1 ($C_1 = 4$).

2. For $\ell = 1, \cdots, L$, $\|\phi_{W_\ell}(x_{\ell-1})\|$ and $\|\phi_{W_\ell}(\tilde{x}_{\ell-1})\|$ are bounded by an absolute constant C_2 ($C_2 = 8$).

3. $\left|\langle \phi_{W_\ell}(x_{\ell-1}^{(1)}), \phi_{W_\ell}(x_{\ell-1}^{(2)})\rangle - \Gamma_\sigma(K_{\ell-1})(x^{(1)}, x^{(2)})\right| \leq 2\epsilon$ for all $\ell = 1, \cdots, L$ and $(x^{(1)}, x^{(2)}) \in \{(x, x), (x, \tilde{x}), (\tilde{x}, \tilde{x})\}$.

We focus on the ℓ-th layer. Let $\tau = \left\|\hat{\Sigma}_{\ell-1}(x, \tilde{x}) - \Sigma_{\ell-1}(x, \tilde{x})\right\|_{\max}$. Recall that $\Gamma_\sigma(K_{\ell-1})(x, \tilde{x}) = \psi_\sigma(\Sigma_{\ell-1}(x, \tilde{x})) = \mathbb{E}_{(X,\tilde{X})\sim\mathcal{N}(0,\Sigma_{\ell-1}(x,\tilde{x}))}\sigma(X)\sigma(\tilde{X})$. Then

$$K_\ell(x, \tilde{x}) = K_{\ell-1}(x, \tilde{x}) + \alpha^2\psi_\sigma(\Sigma_{\ell-1}(x, \tilde{x})).$$

Since $x_\ell = x_{\ell-1} + \frac{\alpha}{\sqrt{m}} V_\ell \phi_{W_\ell}(x_{\ell-1})$, we have

$$\begin{aligned}
\langle x_\ell, \tilde{x}_\ell\rangle &= \langle x_{\ell-1}, \tilde{x}_{\ell-1}\rangle + \frac{\alpha^2}{m}\langle V_\ell\phi_{W_\ell}(x_{\ell-1}), V_\ell\phi_{W_\ell}(\tilde{x}_{\ell-1})\rangle \\
&\quad + \alpha\frac{1}{\sqrt{m}}\big(\langle V_\ell\phi_{W_\ell}(x_{\ell-1}), \tilde{x}_{\ell-1}\rangle + \langle V_\ell\phi_{W_\ell}(\tilde{x}_{\ell-1}), x_{\ell-1}\rangle\big) \\
&= \langle x_{\ell-1}, \tilde{x}_{\ell-1}\rangle + \alpha^2 P + \alpha(Q + R),
\end{aligned}$$

where

$$\begin{aligned}
P &\equiv \frac{1}{m}\langle V_\ell\phi_{W_\ell}(x_{\ell-1}), V_\ell\phi_{W_\ell}(\tilde{x}_{\ell-1})\rangle, \\
Q &\equiv \frac{1}{\sqrt{m}}\big(\langle V_\ell\phi_{W_\ell}(x_{\ell-1}), \tilde{x}_{\ell-1}\rangle\big), \\
R &\equiv \frac{1}{\sqrt{m}}\big(\langle V_\ell\phi_{W_\ell}(\tilde{x}_{\ell-1}), x_{\ell-1}\rangle\big).
\end{aligned}$$

Under the randomness of V_ℓ, P is sub-exponential, and Q and R are Gaussian random variables. Therefore, for a given δ_0, if $m \geq c_0 \log(2/\delta_0)$, with probability at least $1 - \delta_0$ over the randomness of V_ℓ, we have

$$\left|P - \langle \phi_{W_\ell}(x_{\ell-1}), \phi_{W_\ell}(\tilde{x}_{\ell-1})\rangle\right| \leq \|\phi_{W_\ell}(x_{\ell-1})\|\|\phi_{W_\ell}(\tilde{x}_{\ell-1})\|\sqrt{\frac{c_0\log(2/\delta_0)}{m}}; \tag{15}$$

for a given $\tilde{\delta}$, with probability at least $1 - 2\tilde{\delta}$ over the randomness of V_ℓ, we have

$$|Q| \leq \|\phi_{W_\ell}(x_{\ell-1})\|\|\tilde{x}_{\ell-1}\|\sqrt{\frac{\tilde{c}\log(2/\tilde{\delta})}{m}}, \tag{16}$$

and

$$|R| \leq \|\phi_{W_\ell}(\tilde{x}_{\ell-1})\|\|x_{\ell-1}\|\sqrt{\frac{\tilde{c}\log(2/\tilde{\delta})}{m}}, \tag{17}$$

where $c_0, \tilde{c} > 0$ are absolute constants.

Using the above result and Lemma 6 and setting $\delta_0 = \tilde{\delta} = \frac{\delta}{18(L+1)}$, $\delta' = \frac{\delta}{6(L+1)}$, when $m \geq C\log(36(L+1)/\delta)$, we have (15), (16), (17), (12), (13), and (14) hold with probability at least $1 - \frac{\delta}{3(L+1)}$.

Recall that $\tau = \left\|\hat{\Sigma}_{\ell-1}(x, \tilde{x}) - \Sigma_{\ell-1}(x, \tilde{x})\right\|_{\max}$. Conditioned on $\tau < 0.5$, we have

$$\|x_{\ell-1}\|^2 \leq K_{\ell-1}(x, x) + \tau \leq (1 + \alpha^2)^L + \tau \leq e + \tau.$$

Similarly we can show $\|\tilde{x}_{\ell-1}\|^2$ is bounded by $e + \tau$. By (13) and (14) we have $\|\phi_{W_\ell}(x_{\ell-1})\|^2 \leq 2\|x_{\ell-1}\|^2$ and $\|\phi_{W_\ell}(\tilde{x}_{\ell-1})\|^2 \leq 2\|\tilde{x}_{\ell-1}\|^2$, which are both bounded.

Then

$$\begin{aligned}
&\Big|\langle x_\ell, \tilde{x}_\ell\rangle - \big(\alpha^2\psi_\sigma(\Sigma_{\ell-1}(x,\tilde{x})) + K_{\ell-1}(x,\tilde{x})\big)\Big| \\
&\quad\le \tau + \alpha^2\big(P - \psi_\sigma(\Sigma_{\ell-1}(x,\tilde{x}))\big) + \alpha(|Q|+|R|) \\
&\quad\le \tau + \alpha^2\Big|P - \langle\phi_{W_\ell}(x_{\ell-1}), \phi_{W_\ell}(\tilde{x}_{\ell-1})\rangle\Big| + \alpha\sqrt{\frac{\tilde{c}\log(2/\tilde{\delta})}{m}}\big(\|\phi_{W_\ell}(\tilde{x}_{\ell-1})\|\|x_{\ell-1}\| + \|\phi_{W_\ell}(x_{\ell-1})\|\|\tilde{x}_{\ell-1}\|\big) \\
&\qquad + \alpha^2\Big|\psi_\sigma(\hat{\Sigma}_{\ell-1}(x,\tilde{x})) - \psi_\sigma(\Sigma_{\ell-1}(x,\tilde{x}))\Big| + \alpha^2\Big|\langle\phi_{W_\ell}(x_{\ell-1}), \phi_{W_\ell}(\tilde{x}_{\ell-1})\rangle - \psi_\sigma(\hat{\Sigma}_{\ell-1}(x,\tilde{x}))\Big| \\
&\quad\le \tau + (\alpha^2+\alpha)\sqrt{\frac{C_3\log(36(L+1)/\delta)}{m}} + \alpha^2\tau\bigg(1 + \frac{1}{\pi}\bigg(\frac{\tau}{K_{\ell-1}(x,x)}\bigg)^2\bigg) \\
&\quad\le \tau + (\alpha^2+\alpha)\sqrt{\frac{C_3\log(36(L+1)/\delta)}{m}} + \alpha^2\tau\bigg(1 + \frac{1}{4\pi}\bigg).
\end{aligned}$$

When $\alpha = \frac{1}{L^\gamma}$, $\gamma \in [0.5, 1]$, we have $\alpha^2 \le 1/L$. Then when

$$m \ge \frac{C_3 L^{2(1-\gamma)}\log(36(L+1)/\delta)}{\tau^2},$$

we have

$$\Big|\langle x_\ell, \tilde{x}_\ell\rangle - K_\ell(x,\tilde{x})\Big| \le \tau + \frac{4}{L}\tau.$$

As a byproduct, we have

$$\begin{aligned}
&\Big|\langle\phi_{W_\ell}(x_{\ell-1}), \phi_{W_\ell}(\tilde{x}_{\ell-1})\rangle - \psi_\sigma(\Sigma_{\ell-1}(x,\tilde{x}))\Big| \\
&\le \sqrt{\frac{C_4\log(36(L+1)/\delta)}{m}} + \Big(1 + \frac{1}{\pi}\Big(\frac{\tau}{\mu}\Big)^2\Big)\tau \le 2\tau.
\end{aligned}$$

Repeat the above for $(x_{\ell-1}, x_{\ell-1})$ and $(\tilde{x}_{\ell-1}, \tilde{x}_{\ell-1})$, we have with probability at least $1 - \delta/(L+1)$ over the randomness of V_ℓ and W_ℓ,

$$\begin{aligned}
&\Big\|\hat{\Sigma}_{\ell-1}(x,\tilde{x}) - \Sigma_{\ell-1}(x,\tilde{x})\Big\|_{\max} \le \tau \Rightarrow \\
&\Big\|\hat{\Sigma}_{\ell}(x,\tilde{x}) - \Sigma_{\ell}(x,\tilde{x})\Big\|_{\max} \le (1+4/L)\tau.
\end{aligned} \tag{18}$$

Finally, when $m \ge \frac{C_5\log(6(L+1)/\delta)}{(\epsilon/e^4)^2}$, with probability at least $1 - \delta/(L+1)$ over the randomness of A, we have

$$\Big\|\hat{\Sigma}_0(x,\tilde{x}) - \Sigma_0(x,\tilde{x})\Big\|_{\max} \le \epsilon/e^4.$$

Then the result follows by successively using (18). □

A.4 proof of lemma 7

Proof. [28] showed that

$$\Big\|\nabla\psi_\sigma\begin{bmatrix} a & b \\ b & c\end{bmatrix}\Big\|_1 = \frac{1}{2}\frac{a+c}{\sqrt{ac}}\left|\hat{\sigma}\left(\frac{b}{\sqrt{ac}}\right) - \frac{b}{\sqrt{ac}}\hat{\sigma}'\left(\frac{b}{\sqrt{ac}}\right)\right| + \hat{\sigma}'\left(\frac{b}{\sqrt{ac}}\right).$$

When $a, c \in [\mu - r, \mu + r]$, we have

$$\frac{1}{2}\frac{a+c}{\sqrt{ac}} = \frac{1}{2}\bigg(\sqrt{\frac{a}{c}} + \sqrt{\frac{c}{a}}\bigg) \le \frac{1}{2}\bigg(\sqrt{\frac{\mu+r}{\mu-r}} + \sqrt{\frac{\mu-r}{\mu+r}}\bigg) = \bigg(1 - \Big(\frac{r}{\mu}\Big)^2\bigg)^{-1/2} \le 1 + \Big(\frac{r}{\mu}\Big)^2.$$

The last inequality holds when $r < \frac{\mu}{2}$.

Define $\rho = \frac{b}{\sqrt{ac}}$, we have $\rho \in [-1, 1]$. Then

$$\begin{aligned}
\|\nabla\phi_\sigma\|_1 &\le \bigg(1 + \Big(\frac{r}{\mu}\Big)^2\bigg)\Big|\hat{\sigma}(\rho) - \rho\hat{\sigma}'(\rho)\Big| + \hat{\sigma}'(\rho) \\
&= \bigg(1 + \Big(\frac{r}{\mu}\Big)^2\bigg)\left|\frac{\sqrt{1-\rho^2}}{\pi}\right| + 1 - \frac{\cos^{-1}\rho}{\pi} \\
&\le \frac{\sqrt{1-\rho^2}}{\pi} + 1 - \frac{\cos^{-1}\rho}{\pi} + \frac{1}{\pi}\Big(\frac{r}{\mu}\Big)^2 \\
&\le 1 + \frac{1}{\pi}\Big(\frac{r}{\mu}\Big)^2.
\end{aligned}$$

□

B Proof of Theorem 4

B.1 Notation and Main Idea

We already know that when the network width m is large enough, $\langle x_{\ell-1}, \tilde{x}_{\ell-1}\rangle \approx K_{\ell-1}(x,\tilde{x})$, and $\langle \phi_{W_\ell}(x_{\ell-1}), \phi_{W_\ell}(\tilde{x}_{\ell-1})\rangle \approx \Gamma_\sigma(K_{\ell-1})(x,\tilde{x})$.

Next we need to show the concentration of the inner product of $\frac{b_\ell}{\sqrt{m}}$ and $\frac{\tilde{b}_\ell}{\sqrt{m}}$. We define two matrices for each layer

$$\hat{\Theta}_\ell(x,\tilde{x}) = \frac{1}{m}\begin{bmatrix}\langle b_\ell, b_\ell\rangle & \langle b_\ell, \tilde{b}_\ell\rangle \\ \langle \tilde{b}_\ell, b_\ell\rangle & \langle \tilde{b}_\ell, \tilde{b}_\ell\rangle\end{bmatrix},$$

and

$$\Theta_\ell(x,\tilde{x}) = \begin{bmatrix}B_\ell(x,x) & B_\ell(x,\tilde{x}) \\ B_\ell(\tilde{x},x) & B_\ell(\tilde{x},\tilde{x})\end{bmatrix}.$$

Recall that

$$b_\ell = \alpha\sqrt{\frac{1}{m}}\sqrt{\frac{2}{m}}W_\ell^\top D_\ell V_\ell^\top b_{\ell+1} + b_{\ell+1}.$$

We aim to show that when $\|\hat{\Theta}_{\ell+1}(x,\tilde{x}) - \Theta_{\ell+1}(x,\tilde{x})\|_{\max} \le \tau$, with high probability over the randomness of W_ℓ and V_ℓ, we have $\|\hat{\Theta}_\ell(x,\tilde{x}) - \Theta_\ell(x,\tilde{x})\|_{\max} \le (1+\mathcal{O}(1/L))\tau$. Notice that $b_{\ell+1}$ and $\tilde{b}_{\ell+1}$ contain the information of W_ℓ and V_ℓ; they are not independent. Nevertheless we can decompose the randomness of W_ℓ and V_ℓ to show the concentration. This technique is also used in [22].

B.2 Lemmas

In this part we introduce some useful lemmas. The first one shows the property of the step activation function.

Lemma 8 (Property of σ')**.** *[22]*

(1). Sub-Gaussian concentration. With probability at least $1-\delta$ over the randomness of W_ℓ, we have

$$\left|\frac{2}{m}\operatorname{Tr}(D_\ell\widetilde{D}_\ell) - \psi_{\sigma'}(\hat{\Sigma}_{\ell-1}(x,\tilde{x}))\right| \le \sqrt{\frac{c\log(2/\delta)}{m}}.$$

(2). Holder continuity. Fix $\mu > 0, 0 < r \le \mu$. For all $A, B \in \mathcal{M}_{\mu,r} = \left\{\begin{bmatrix}a & b\\ b & c\end{bmatrix}\middle| a,c \in [\mu-r,\mu+r]; ac-b^2>0\right\}$, if $\|A-B\|_{\max} \le (\mu-r)\epsilon^2$, then

$$|\psi_{\sigma'}(A) - \psi_{\sigma'}(B)| \le \epsilon.$$

The following lemma shows that regardless the fact that $b_{\ell+1}$ and $\tilde{b}_{\ell+1}$ depend on V_ℓ, we can treat V_ℓ as a Gaussian matrix independent of $b_{\ell+1}$ and $\tilde{b}_{\ell+1}$ when the network width is large enough.

Lemma 9. *Assume the following inequality hold simultaneously for all $\ell = 1, 2, \cdots, L$*

$$\left\|\frac{1}{\sqrt{m}}W_\ell\right\| \le C, \quad \left\|\frac{1}{\sqrt{m}}V_\ell\right\| \le C.$$

Fix an ℓ. Further assume that

$$\|\hat{\Theta}_{\ell+1}(x,\tilde{x}) - \Theta_{\ell+1}(x,\tilde{x})\|_{\max} \le 1.$$

When $m \ge \max\{\frac{C}{\epsilon^2}(1+\log\frac{6}{\delta}), \frac{C}{\epsilon^2}\log\frac{8L}{\delta'}, cL^{2-2\gamma}\log\frac{8L}{\delta'}\}$, the following holds for all $(x^{(1)}, x^{(2)}) \in \{(x,x),(x,\tilde{x}),(\tilde{x},\tilde{x})\}$ with probability at least $1-\delta-\delta'$

$$\left|\frac{2}{m}\frac{b_{\ell+1}^{(1)}}{\sqrt{m}}^\top V_\ell D_\ell^{(1)} D_\ell^{(2)} V_\ell^\top \frac{b_{\ell+1}^{(2)}}{\sqrt{m}} - \langle\frac{b_{\ell+1}^{(1)}}{\sqrt{m}}, \frac{b_{\ell+1}^{(2)}}{\sqrt{m}}\rangle\frac{2}{m}\operatorname{Tr}(D_\ell^{(1)}D_\ell^{(2)})\right| \le \epsilon.$$

The following lemma shows the same thing for W_ℓ as V_ℓ in Lemma 9.

Lemma 10. *Assume the conditions and the results of Lemma 9 hold.*

(1). When $m \ge \max\{\frac{C}{\epsilon^2}(1+\log\frac{6}{\delta}), \frac{C}{\epsilon^2}\log\frac{8L}{\delta'}, cL^{2-2\gamma}\log\frac{8L}{\delta'}\}$,the following holds for all $(x^{(1)}, x^{(2)}) \in \{(x,x),(x,\tilde{x}),(\tilde{x},\tilde{x})\}$ with probability at least $1-\delta-\delta'$

$$\left|\frac{1}{m}\frac{2}{m}\langle W_\ell^\top D_\ell^{(1)} V_\ell^\top \frac{b_{\ell+1}^{(1)}}{\sqrt{m}}, W_\ell^\top D_\ell^{(2)} V_\ell^\top \frac{b_{\ell+1}^{(2)}}{\sqrt{m}}\rangle - \frac{2}{m}\langle D_\ell^{(1)} V_\ell^\top \frac{b_{\ell+1}^{(1)}}{\sqrt{m}}, D_\ell^{(2)} V_\ell^\top \frac{b_{\ell+1}^{(2)}}{\sqrt{m}}\rangle\right| \le \epsilon.$$

(2). When $m \geq \max\{\frac{C}{\tilde{\epsilon}^2}\log\frac{16L}{\tilde{\delta}}, cL^{2-2\gamma}\log\frac{16L}{\tilde{\delta}}\}$, *for all* $(x^{(1)}, x^{(2)}) \in \{(x,x),(x,\tilde{x}),(\tilde{x},x),(\tilde{x},\tilde{x})\}$, *the following holds with probability at least* $1-\tilde{\delta}$

$$\left|\frac{1}{m}\sqrt{\frac{1}{m}}\sqrt{\frac{2}{m}}\langle W_\ell^\top D_\ell^{(1)} V_\ell^\top b_{\ell+1}^{(1)}, b_{\ell+1}^{(2)}\rangle\right| \leq \tilde{\epsilon}.$$

B.3 Proof of Theorem 4

Proof. In this proof we are going to prove that when m satisfies the assumption, with probability at least $1-\delta_0$, the following hold for $\ell = 1, \cdots, L$.

$$\left|\frac{1}{\alpha^2}\langle \nabla_{V_\ell} f, \nabla_{V_\ell}\tilde{f}\rangle - B_{\ell+1}(x,\tilde{x})\Gamma_\sigma(K_{\ell-1})(x,\tilde{x})\right| \leq \epsilon_0,$$
$$\left|\frac{1}{\alpha^2}\langle \nabla_{W_\ell} f, \nabla_{W_\ell}\tilde{f}\rangle - K_{\ell-1}(x,\tilde{x})B_{\ell+1}(x,\tilde{x})\Gamma_{\sigma'}(K_{\ell-1})(x,\tilde{x})\right| \leq \epsilon_0.$$

We break the proof into several steps. Each step is based on the result of the previous steps. Note that the absolute constants c and C may vary throughout the proof.

Step 1. Norm Control of the Gaussian Matrices

With probability at least $1-\delta_1$, when $m > c\log\frac{4L}{\delta_1}$, one can show that the following hold simultaneously for all $\ell = 1, 2, \cdots, L$ [38]

$$\left\|\frac{1}{\sqrt{m}}W_\ell\right\| \leq C, \quad \left\|\frac{1}{\sqrt{m}}V_\ell\right\| \leq C.$$

Step 2. Concentration of the GP kernels

By Theorem 3, with probability at least $1-\delta_2$, when

$$m \geq \frac{C}{\epsilon_2^4}L^{2-2\gamma}\log\frac{36(L+1)}{\delta_2},$$

we have

1. For $\ell = 0, \cdots, L$, $\left\|\Sigma_\ell(x,\tilde{x}) - \hat{\Sigma}_\ell(x,\tilde{x})\right\|_{\max} \leq c\epsilon_2^2$;
2. For $\ell = 0, 1, \cdots, L$, $\|x_\ell\|$ and $\|\tilde{x}_\ell\|$ are bounded by an absolute constant C_1 ($C_1 = 4$);
3. For $\ell = 1, \cdots, L$, $\|\phi_{W_\ell}(x_{\ell-1})\|$ and $\|\phi_{W_\ell}(\tilde{x}_{\ell-1})\|$ are bounded by an absolute constant C_2 ($C_2 = 8$);
4. $\left|\langle\phi_{W_\ell}(x_{\ell-1}^{(1)}), \phi_{W_\ell}(x_{\ell-1}^{(2)})\rangle - \Gamma_\sigma(K_{\ell-1})(x^{(1)}, x^{(2)})\right| \leq 2c\epsilon_2^2$ for all $\ell = 1, \cdots, L$ and $(x^{(1)}, x^{(2)}) \in \{(x,x),(x,\tilde{x}),(\tilde{x},\tilde{x})\}$.

Step 3. Concentration of σ'

By Lemma 8, when $m \geq \frac{C}{\epsilon_2^2}\log\frac{6L}{\delta_3}$, with probability at least $1-\delta_3$, for all $\ell = 1, 2, \cdots, L$ and $(x^{(1)}, x^{(2)}) \in \{(x,x),(x,\tilde{x}),(\tilde{x},\tilde{x})\}$, we have

$$\left|\frac{2}{m}\operatorname{Tr}(D_\ell^{(1)}D_\ell^{(2)}) - \Gamma_{\sigma'}(K_{\ell-1})(x^{(1)}, x^{(2)})\right| \leq \sqrt{\frac{c\log(6L/\delta_3)}{m}} + \sqrt{2\left\|\hat{\Sigma}_{\ell-1}(x,\tilde{x}) - \Sigma_{\ell-1}(x,\tilde{x})\right\|_{\max}} \leq \epsilon_2.$$

Step 4. Concentration of B_ℓ

Recall that

$$b_{\ell+1} = \left(v^\top \frac{\partial x_L}{\partial x_{L-1}}\frac{\partial x_{L-1}}{\partial x_{L-2}}\cdots\frac{\partial x_{\ell+1}}{\partial x_\ell}\right)^\top.$$

We have

$$b_{L+1} = v,$$

and for $\ell = 1, 2, \cdots, L-1$,

$$b_{\ell+1} = \frac{\partial x_{\ell+1}}{\partial x_\ell}^\top b_{\ell+2} = \alpha\sqrt{\frac{1}{m}}\sqrt{\frac{2}{m}}W_{\ell+1}^\top D_{\ell+1}V_{\ell+1}^\top b_{\ell+2} + b_{\ell+2}.$$

Following the same idea in Thm 3, we prove by induction. First of all, for b_{L+1}, we have $\Theta_{L+1}(x,\tilde{x}) = \begin{bmatrix}1 & 1\\ 1 & 1\end{bmatrix}, \hat{\Theta}_{L+1}(x,\tilde{x}) = \frac{\|v\|^2}{m}\begin{bmatrix}1 & 1\\ 1 & 1\end{bmatrix}$. Then by Bernstein inequality [39], with probability at least $1-\frac{\delta_4}{L}$, when $m \geq \frac{C}{\epsilon_4^2}\log\frac{2L}{\delta_4}$, we have

$$\left|\frac{\|v\|^2}{m} - 1\right| \leq \epsilon_4.$$

Fix $\ell \in \{2,3,\cdots,L\}$. Assume that

$$\left\|\hat{\Theta}_{\ell+1}(x,\tilde{x}) - \Theta_{\ell+1}(x,\tilde{x})\right\|_{\max} \leq \tau \leq 1,$$

we hope to prove with high probability,

$$\left\|\hat{\Theta}_{\ell}(x,\tilde{x}) - \Theta_{\ell}(x,\tilde{x})\right\|_{\max} \leq (1+\mathcal{O}(1/L))\tau.$$

First write

$$\frac{1}{m}\langle b_\ell^{(1)}, b_\ell^{(2)}\rangle = \frac{1}{m}\langle b_{\ell+1}^{(1)}, b_{\ell+1}^{(2)}\rangle + \alpha^2 P + \alpha(Q+R),$$

where

$$\begin{aligned}
P &= \frac{1}{m}\frac{2}{m}\langle W_\ell^\top D_\ell^{(1)} V_\ell^\top \frac{b_{\ell+1}^{(1)}}{\sqrt{m}}, W_\ell^\top D_\ell^{(2)} V_\ell^\top \frac{b_{\ell+1}^{(2)}}{\sqrt{m}}\rangle,\\
Q &= \frac{1}{m}\sqrt{\frac{1}{m}}\sqrt{\frac{2}{m}}\langle W_\ell^\top D_\ell^{(1)} V_\ell^\top b_{\ell+1}^{(1)}, b_{\ell+1}^{(2)}\rangle,\\
R &= \frac{1}{m}\sqrt{\frac{1}{m}}\sqrt{\frac{2}{m}}\langle W_\ell^\top D_\ell^{(2)} V_\ell^\top b_{\ell+1}^{(2)}, b_{\ell+1}^{(1)}\rangle.
\end{aligned}$$

Then

$$\begin{aligned}
&\left|\frac{1}{m}\langle b_\ell^{(1)}, b_\ell^{(2)}\rangle - (B_{\ell+1}(x^{(1)},x^{(2)}) + \alpha^2 B_{\ell+1}(x^{(1)},x^{(2)})\Gamma_{\sigma'}(K_{\ell-1})(x^{(1)},x^{(2)})\right|\\
&\leq \left|\frac{1}{m}\langle b_{\ell+1}^{(1)}, b_{\ell+1}^{(2)}\rangle - B_{\ell+1}(x^{(1)},x^{(2)})\right| + \alpha^2\left|P - B_{\ell+1}(x^{(1)},x^{(2)})\Gamma_{\sigma'}(K_{\ell-1})(x^{(1)},x^{(2)})\right| + \alpha|Q| + \alpha|R|\\
&\leq \tau + \alpha^2\left|P - \frac{2}{m}\langle D_\ell^{(1)} V_\ell^\top \frac{b_{\ell+1}^{(1)}}{\sqrt{m}}, D_\ell^{(2)} V_\ell^\top \frac{b_{\ell+1}^{(2)}}{\sqrt{m}}\rangle\right|\\
&\quad + \alpha^2\left|\frac{2}{m}\langle D_\ell^{(1)} V_\ell^\top \frac{b_{\ell+1}^{(1)}}{\sqrt{m}}, D_\ell^{(2)} V_\ell^\top \frac{b_{\ell+1}^{(2)}}{\sqrt{m}}\rangle - \langle \frac{b_{\ell+1}^{(1)}}{\sqrt{m}}, \frac{b_{\ell+1}^{(2)}}{\sqrt{m}}\rangle \frac{2}{m}\operatorname{Tr}(D_\ell^{(1)} D_\ell^{(2)})\right|\\
&\quad + \alpha^2\left|\langle \frac{b_{\ell+1}^{(1)}}{\sqrt{m}}, \frac{b_{\ell+1}^{(2)}}{\sqrt{m}}\rangle - B_{\ell+1}(x^{(1)},x^{(2)})\right|\left|\frac{2}{m}\operatorname{Tr}(D_\ell^{(1)} D_\ell^{(2)})\right|\\
&\quad + \alpha^2\left|B_{\ell+1}(x^{(1)},x^{(2)})\right|\left|\frac{2}{m}\operatorname{Tr}(D_\ell^{(1)} D_\ell^{(2)}) - \Gamma_{\sigma'}(K_{\ell-1})(x^{(1)},x^{(2)})\right|\\
&\quad + \alpha|Q| + \alpha|R|.
\end{aligned}$$

In Lemma 9 and Lemma 10, set $\tilde{\epsilon} = cL^{\gamma-1}\tau$, $\epsilon = c\tau$, $\delta = \tilde{\delta} = \delta' = \delta_4/5L$. When $m \geq \max\{\frac{C}{\tau^2}(1+\log\frac{30L}{\delta_4}), \frac{C}{\tau^2}\log\frac{40L^2}{\delta_4}, \frac{C}{\tau^2}L^{2-2\gamma}\log\frac{80L^2}{\delta_4}, cL^{2-2\gamma}\log\frac{80L^2}{\delta_4}\}$, with probability at least $1-\frac{\delta_4}{L}$, the results of Lemma 9 and Lemma 10 hold. Then for all $(x^{(1)},x^{(2)}) \in \{(x,x),(x,\tilde{x}),(\tilde{x},\tilde{x})\}$,

$$\begin{aligned}
\left|\frac{1}{m}\langle b_\ell^{(1)}, b_\ell^{(2)}\rangle - B_\ell(x^{(1)},x^{(2)})\right| &\leq \tau + \alpha^2 c\tau + \alpha^2 c\tau + \alpha^2 2\tau + \alpha^2 e\epsilon_2 + 2\alpha c L^{a-1}\tau\\
&\leq \tau(1+\mathcal{O}(1/L)). \quad (\text{Set } \epsilon_2 \leq c\tau.)
\end{aligned}$$

By taking union bound, with probability at least $1-\delta_4$, we have for all $\ell = 1,2,\cdots,L$,

$$\|\hat{\Theta}_{\ell+1}(x,\tilde{x}) - \Theta_{\ell+1}(x,\tilde{x})\|_{\max} \leq (1+\mathcal{O}(1/L))^L\epsilon_4 \leq C\epsilon_4.$$

Meanwhile, we have for all $(x^{(1)},x^{(2)}) \in \{(x,x),(x,\tilde{x}),(\tilde{x},\tilde{x})\}$ and $\ell = 1,\cdots,L$,

$$\left|\frac{2}{m}\langle D_\ell^{(1)} V_\ell^\top \frac{b_{\ell+1}^{(1)}}{\sqrt{m}}, D_\ell^{(2)} V_\ell^\top \frac{b_{\ell+1}^{(2)}}{\sqrt{m}}\rangle - B_{\ell+1}(x^{(1)},x^{(2)})\Gamma_{\sigma'}(K_{\ell-1})(x^{(1)},x^{(2)})\right| \leq (2+c)\tau + e\epsilon_2 \leq C\epsilon_4.$$

Step 5. Summary

Using previous results, for all ℓ, we have

$$\begin{aligned}
&\left|\frac{1}{\alpha^2}\langle\nabla_{V_\ell}f,\nabla_{V_\ell}\tilde{f}\rangle - B_{\ell+1}\Gamma_\sigma(K_{\ell-1})\right| \\
&\le \left|\frac{1}{m}\langle b_{\ell+1},\tilde{b}_{\ell+1}\rangle - B_{\ell+1}\right|\cdot|\langle\phi_{W_\ell}(x_{\ell-1}),\phi_{W_\ell}(\tilde{x}_{\ell-1})\rangle| + |B_{\ell+1}|\cdot|\langle\phi_{W_\ell}(x_{\ell-1}),\phi_{W_\ell}(\tilde{x}_{\ell-1})\rangle - \Gamma_\sigma(K_{\ell-1})| \\
&\le C\epsilon_4 + C\epsilon_2^2,
\end{aligned}$$

and

$$\begin{aligned}
&\left|\frac{1}{\alpha^2}\langle\nabla_{W_\ell}f,\nabla_{W_\ell}\tilde{f}\rangle - K_{\ell-1}B_{\ell+1}\Gamma_{\sigma'}(K_{\ell-1})\right| \\
&\le \left|\frac{1}{m}\langle x_{\ell-1},\tilde{x}_{\ell-1}\rangle - K_{\ell-1}\right|\cdot\left|\frac{2}{m}\tilde{b}_{\ell+1}^\top V_\ell\widetilde{D}_\ell D_\ell V_\ell^\top b_{\ell+1}\right| + |K_{\ell-1}|\cdot\left|\frac{2}{m}\tilde{b}_{\ell+1}^\top V_\ell\widetilde{D}_\ell D_\ell V_\ell^\top b_{\ell+1} - B_{\ell+1}\Gamma_{\sigma'}(K_{\ell-1})\right| \\
&\le C\epsilon_2^2 + C\epsilon_4.
\end{aligned}$$

To sum up, by choosing $\epsilon_4 = c\epsilon_0$, $\epsilon_2 = c\epsilon_4$, and $\delta_1 = \delta_2 = \delta_3 = \delta_4 = \delta_0/4$, then with probability at least $1-\delta_0$, when

$$\begin{aligned}
m &\ge \frac{C}{\epsilon_0^4}L^{2-2\gamma}\left(\log\frac{320(L^2+1)}{\delta_0}+1\right) \\
&\ge \max\Bigg\{c\log\frac{16L}{\delta_0}, \frac{C}{\epsilon_0^4}L^{2-2\gamma}\log\frac{144(L+1)}{\delta_0}, \frac{C}{\epsilon_0^2}\log\frac{24L}{\delta_0}, \\
&\quad \frac{C}{\epsilon_0^2}\log\frac{8L}{\delta_0}, \frac{C}{\epsilon_0^2}(1+\log\frac{120L}{\delta_0}), \frac{C}{\epsilon_0^2}\log\frac{160L^2}{\delta_0}, \frac{C}{\epsilon_0^2}L^{2-2\gamma}\log\frac{320L^2}{\delta_0}, cL^{2-2\gamma}\log\frac{320L^2}{\delta_0^4}\Bigg\},
\end{aligned}$$

the desired results hold.

□

C Proofs of the Lemmas

C.1 Supporting lemmas

Lemma 11. *Define $G = [\phi_{W_\ell}(x_{\ell-1}), \phi_{W_\ell}(\tilde{x}_{\ell-1})]$, and $\Pi_G^\perp$ as the orthogonal projection onto the orthogonal complement of the column space of G. when $m \ge 1 + \log\frac{6}{\delta}$, the following holds with probability at least $1-\delta$ for all $(x^{(1)}, x^{(2)}) \in \{(x,x),(x,\tilde{x}),(\tilde{x},\tilde{x})\}$,*

$$\left|\frac{2}{m}\frac{b_{\ell+1}^{(1)}}{\sqrt{m}}^\top V_\ell\Pi_G^\perp D_\ell^{(1)}D_\ell^{(2)}\Pi_G^\perp V_\ell^\top\frac{b_{\ell+1}^{(2)}}{\sqrt{m}} - \langle\frac{b_{\ell+1}^{(1)}}{\sqrt{m}},\frac{b_{\ell+1}^{(1)}}{\sqrt{m}}\rangle\frac{2}{m}\operatorname{Tr}(D_\ell^{(1)}D_\ell^{(2)})\right| \le (4+4\sqrt{2})M\sqrt{\frac{1+\log\frac{6}{\delta}}{m}},$$

where

$$M = \max\left\{\frac{\|b_{\ell+1}\|^2}{m}, \frac{\|\tilde{b}_{\ell+1}\|^2}{m}\right\}.$$

proof of Lemma 11. We prove the lemma on any realization of $(A, W_1, V_1, \cdots, W_{\ell-1}, V_{\ell-1}, W_\ell, W_{\ell+1}, V_{\ell+1}, \cdots, W_L, V_L, v)$, $V_\ell\phi_{W_\ell}(x_{\ell-1})$ and $V_\ell\phi_{W_\ell}(\tilde{x}_{\ell-1})$, and consider the remaining randomness of V_ℓ. In this case, D_ℓ, $\widetilde{D}_\ell$, $b_{\ell+1}$ and $\tilde{b}_{\ell+1}$ are fixed.

One can show that conditioned on the realization of $V_\ell G$ (whose "degree of freedom" is $2m$), $V_\ell\Pi_G^\perp$ is identically distributed as $\widetilde{V}_\ell\Pi_G^\perp$, where $\widetilde{V}_\ell$ is an i.i.d. copy of V_ℓ. The remaining m^2-2m "degree of freedom" is enough for a good concentration. For the proof of this result, we refer the readers to Lemma E.3 in [22].

Denote $T = \Pi_G^\perp D_\ell^{(1)}D_\ell^{(2)}\Pi_G^\perp$,

$$S = \begin{bmatrix} \widetilde{V}_\ell^\top\frac{b_{\ell+1}^{(1)}}{\sqrt{m}} \\ \widetilde{V}_\ell^\top\frac{b_{\ell+1}^{(2)}}{\sqrt{m}} \end{bmatrix}.$$

We know that S is a $2m$-dimensional Gaussian random vector, and

$$S \sim \mathcal{N}\left(0, \begin{bmatrix} \langle\frac{b_{\ell+1}^{(1)}}{\sqrt{m}},\frac{b_{\ell+1}^{(1)}}{\sqrt{m}}\rangle I_m & \langle\frac{b_{\ell+1}^{(1)}}{\sqrt{m}},\frac{b_{\ell+1}^{(2)}}{\sqrt{m}}\rangle I_m \\ \langle\frac{b_{\ell+1}^{(2)}}{\sqrt{m}},\frac{b_{\ell+1}^{(1)}}{\sqrt{m}}\rangle I_m & \langle\frac{b_{\ell+1}^{(2)}}{\sqrt{m}},\frac{b_{\ell+1}^{(2)}}{\sqrt{m}}\rangle I_m \end{bmatrix}\right).$$

Then there exists a matrix $P \in \mathbb{R}^{2m\times 2m}$, such that

$$PP^\top = \begin{bmatrix} \langle \frac{b_{\ell+1}^{(1)}}{\sqrt{m}}, \frac{b_{\ell+1}^{(1)}}{\sqrt{m}} \rangle I_m & \langle \frac{b_{\ell+1}^{(1)}}{\sqrt{m}}, \frac{b_{\ell+1}^{(2)}}{\sqrt{m}} \rangle I_m \\ \langle \frac{b_{\ell+1}^{(2)}}{\sqrt{m}}, \frac{b_{\ell+1}^{(1)}}{\sqrt{m}} \rangle I_m & \langle \frac{b_{\ell+1}^{(2)}}{\sqrt{m}}, \frac{b_{\ell+1}^{(2)}}{\sqrt{m}} \rangle I_m \end{bmatrix},$$

and $S \overset{d}{=} P\xi, \xi \sim \mathcal{N}(0, I_{2m})$.

Thus

$$\frac{b_{\ell+1}^{(1)}}{\sqrt{m}}^\top \widetilde{V}_\ell \Pi_G^\perp D_\ell^{(1)} D_\ell^{(2)} \Pi_G^\perp \widetilde{V}_\ell^\top \frac{b_{\ell+1}^{(2)}}{\sqrt{m}} \overset{d}{=} \xi^\top P^\top \begin{bmatrix} I_m \\ 0 \end{bmatrix}^\top T \begin{bmatrix} 0 \\ I_m \end{bmatrix} P\xi = \frac{1}{2}\xi^\top P^\top \begin{bmatrix} 0 & T \\ T & 0 \end{bmatrix} P\xi.$$

We have

$$\begin{aligned} \left\| \frac{1}{2} P^\top \begin{bmatrix} 0 & T \\ T & 0 \end{bmatrix} P \right\| &\le \frac{1}{2} \left\|P^\top\right\| \cdot \|P\| \cdot \left\| \begin{bmatrix} 0 & T \\ T & 0 \end{bmatrix} \right\| \\ &= \frac{1}{2} \left\|PP^\top\right\| \cdot \|T\| \\ &\le \frac{1}{2} \left\| \begin{bmatrix} \langle \frac{b_{\ell+1}^{(1)}}{\sqrt{m}}, \frac{b_{\ell+1}^{(1)}}{\sqrt{m}} \rangle I_m & \langle \frac{b_{\ell+1}^{(1)}}{\sqrt{m}}, \frac{b_{\ell+1}^{(2)}}{\sqrt{m}} \rangle I_m \\ \langle \frac{b_{\ell+1}^{(2)}}{\sqrt{m}}, \frac{b_{\ell+1}^{(1)}}{\sqrt{m}} \rangle I_m & \langle \frac{b_{\ell+1}^{(2)}}{\sqrt{m}}, \frac{b_{\ell+1}^{(2)}}{\sqrt{m}} \rangle I_m \end{bmatrix} \right\| \left\|\Pi_G^\perp\right\| \left\|D_\ell^{(1)}\right\| \left\|D_\ell^{(2)}\right\| \left\|\Pi_G^\perp\right\| \\ &\le \frac{\langle \frac{b_{\ell+1}^{(1)}}{\sqrt{m}}, \frac{b_{\ell+1}^{(1)}}{\sqrt{m}} \rangle + \langle \frac{b_{\ell+1}^{(2)}}{\sqrt{m}}, \frac{b_{\ell+1}^{(2)}}{\sqrt{m}} \rangle}{2} \le M. \end{aligned}$$

And $\left\| \frac{1}{2} P^\top \begin{bmatrix} 0 & T \\ T & 0 \end{bmatrix} P \right\|_F \le \sqrt{2m} M.$

Then by the Hanson-Wright Inequality for Gaussian chaos [40], we have with probability at least $1 - \delta/3$,

$$\begin{aligned} &\frac{2}{m} \left| \frac{b_{\ell+1}^{(1)}}{\sqrt{m}}^\top \widetilde{V}_\ell \Pi_G^\perp D_\ell^{(1)} D_\ell^{(2)} \Pi_G^\perp \widetilde{V}_\ell^\top \frac{b_{\ell+1}^{(2)}}{\sqrt{m}} - \mathbb{E}_{\widetilde{V}_\ell} \left[\frac{b_{\ell+1}^{(1)}}{\sqrt{m}}^\top \widetilde{V}_\ell \Pi_G^\perp D_\ell^{(1)} D_\ell^{(2)} \Pi_G^\perp \widetilde{V}_\ell^\top \frac{b_{\ell+1}^{(2)}}{\sqrt{m}} \right] \right| \\ \le\ &\frac{4}{m} \left(\sqrt{2m} M \sqrt{\log \frac{6}{\delta}} + M \log \frac{6}{\delta} \right), \end{aligned}$$

Furthermore, we have

$$\mathbb{E}_{\widetilde{V}_\ell} \left[\frac{b_{\ell+1}^{(1)}}{\sqrt{m}}^\top \widetilde{V}_\ell \Pi_G^\perp D_\ell^{(1)} D_\ell^{(2)} \Pi_G^\perp \widetilde{V}_\ell^\top \frac{b_{\ell+1}^{(2)}}{\sqrt{m}} \right] = \langle \frac{b_{\ell+1}^{(1)}}{\sqrt{m}}, \frac{b_{\ell+1}^{(1)}}{\sqrt{m}} \rangle \operatorname{Tr}(\Pi_G^\perp D_\ell^{(1)} D_\ell^{(2)}).$$

Thus

$$\begin{aligned} &\left| \frac{2}{m} \mathbb{E}_{\widetilde{V}_\ell} \left[\frac{b_{\ell+1}^{(1)}}{\sqrt{m}}^\top \widetilde{V}_\ell \Pi_G^\perp D_\ell^{(1)} D_\ell^{(2)} \Pi_G^\perp \widetilde{V}_\ell^\top \frac{b_{\ell+1}^{(2)}}{\sqrt{m}} \right] - \langle \frac{b_{\ell+1}^{(1)}}{\sqrt{m}}, \frac{b_{\ell+1}^{(1)}}{\sqrt{m}} \rangle \frac{2}{m} \operatorname{Tr}(D_\ell^{(1)} D_\ell^{(2)}) \right| \\ =\ &\frac{2}{m} \left| \langle \frac{b_{\ell+1}^{(1)}}{\sqrt{m}}, \frac{b_{\ell+1}^{(1)}}{\sqrt{m}} \rangle \operatorname{Tr}(\Pi_G D_\ell^{(1)} D_\ell^{(2)}) \right| \\ \le\ &\frac{2}{m} M \operatorname{Tr}(\Pi_G D_\ell^{(1)} D_\ell^{(2)} \Pi_G) \\ \le\ &\frac{4}{m} M. \end{aligned}$$

By taking union bound, we have with probability at least $1 - \delta$, for all $(x^{(1)}, x^{(2)}) \in \{(x,x), (x,\tilde{x}), (\tilde{x},\tilde{x})\}$,

$$\begin{aligned} &\left| \frac{2}{m} \frac{b_{\ell+1}^{(1)}}{\sqrt{m}}^\top V_\ell \Pi_G^\perp D_\ell^{(1)} D_\ell^{(2)} \Pi_G^\perp V_\ell^\top \frac{b_{\ell+1}^{(2)}}{\sqrt{m}} - \langle \frac{b_{\ell+1}^{(1)}}{\sqrt{m}}, \frac{b_{\ell+1}^{(1)}}{\sqrt{m}} \rangle \frac{2}{m} \operatorname{Tr}(D_\ell^{(1)} D_\ell^{(2)}) \right| \\ \le\ &\frac{4}{m} \left(\sqrt{2m} M \sqrt{\log \frac{6}{\delta}} + M \log \frac{6}{\delta} \right) + \frac{4}{m} M \\ \le\ &(4 + 4\sqrt{2}) M \sqrt{\frac{1 + \log \frac{6}{\delta}}{m}}, \end{aligned}$$

where the last inequality holds when $m \geq 1 + \log \frac{6}{\delta}$. □

Lemma 12 (Norm controls of $b_{\ell+1}$)**.** *Assume the following inequalities hold simultaneously for all* $\ell = 1, 2, \cdots, L$

$$\left\| \frac{1}{\sqrt{m}} W_\ell \right\| \leq C, \quad \left\| \frac{1}{\sqrt{m}} V_\ell \right\| \leq C.$$

Then for any fixed input x, $1 \leq \ell \leq L$ *and* $u \in \mathbb{R}^m$*, when*

$$m \geq cL^{2-2\gamma} \log \frac{2L}{\delta'},$$

with probability at least $1 - \delta'$ *over the randomness of* $W_{\ell+1}, V_{\ell+1}, \cdots, W_L, V_L, v$*, we have*

$$|\langle u, b_{\ell+1} \rangle| \leq C' \|u\| \sqrt{\log \frac{2L}{\delta'}}.$$

proof of Lemma 12. Denote $u_\ell = u$, and

$$u_{i+1} = \alpha \sqrt{\frac{1}{m}} \sqrt{\frac{2}{m}} V_{i+1} D_{i+1} W_{i+1} u_i + u_i, \quad i = \ell, \ell+1, \cdots, L-1.$$

One can show that $\langle u, b_{\ell+1} \rangle = \langle v, u_L \rangle$. Next we show that $\|u_{i+1}\| = (1 + \mathcal{O}(\frac{1}{L}))\|u_i\|$ with high probability. First write

$$\|u_{i+1}\|^2 = \|u_i\|^2 + \alpha^2 \left\| \sqrt{\frac{1}{m}} \sqrt{\frac{2}{m}} V_{i+1} D_{i+1} W_{i+1} u_i \right\|^2 + 2\alpha \left\langle u_i, \sqrt{\frac{1}{m}} \sqrt{\frac{2}{m}} V_{i+1} D_{i+1} W_{i+1} u_i \right\rangle.$$

By the assumption we have

$$\left\| \sqrt{\frac{2}{m}} D_{i+1} W_{i+1} u_i \right\| \leq \sqrt{2} C \|u_i\|,$$

$$\left\| \sqrt{\frac{1}{m}} \sqrt{\frac{2}{m}} V_{i+1} D_{i+1} W_{i+1} u_i \right\| \leq \sqrt{2} C^2 \|u_i\|.$$

With probability at least $1 - \delta'/L$ over the randomness of V_{i+1}, we have

$$\left\| \langle u_i, \sqrt{\frac{1}{m}} \sqrt{\frac{2}{m}} V_{i+1} D_{i+1} W_{i+1} u_i \rangle \right\| \leq \|u_i\| \cdot \left\| \sqrt{\frac{2}{m}} D_{i+1} W_{i+1} u_i \right\| \sqrt{\frac{c \log \frac{2L}{\delta'}}{m}}.$$

Then when

$$m \geq cL^{2-2\gamma} \log \frac{2L}{\delta'},$$

we have

$$\begin{aligned}
\|u_{i+1}\|^2 &= \|u_i\|^2 + \alpha^2 \left\| \sqrt{\frac{1}{m}} \sqrt{\frac{2}{m}} V_{i+1} D_{i+1} W_{i+1} u_i \right\|^2 + 2\alpha \langle u_i, \sqrt{\frac{1}{m}} \sqrt{\frac{2}{m}} V_{i+1} D_{i+1} W_{i+1} u_i \rangle \\
&\leq (1 + 2C^4/L)\|u_i\|^2 + 2\alpha\sqrt{2} C \|u_i\|^2 \sqrt{\frac{c \log \frac{2L}{\delta'}}{m}} \\
&\leq (1 + 2C^4/L + 2\sqrt{2}C/L)\|u_i\|^2 = (1 + \mathcal{O}(1/L))\|u_i\|^2.
\end{aligned}$$

Then with probability at least $1 - \delta'(L-1)/L$ we have $\|u_L\| \leq C\|u\|$. Finally the result holds from the standard concentration bound for Gaussian random variables [39]. □

C.2 Proofs of Lemma 9

proof of Lemma 9. By the assumption, we have

$$\frac{1}{m} \|b_{\ell+1}\|^2 \leq B_{\ell+1}(x, x) + 1 \leq 4.$$

Similarly, $\frac{1}{m}\|\tilde{b}_{\ell+1}\|^2 \leq 4$. Then by Lemma 11, when $m \geq \frac{C}{\epsilon^2}(1 + \log \frac{6}{\delta})$, we have for all $(x^{(1)}, x^{(2)}) \in \{(x, x), (x, \tilde{x}), (\tilde{x}, \tilde{x})\}$,

$$\left| \frac{2}{m} \frac{b_{\ell+1}^{(1)}}{\sqrt{m}}^\top V_\ell \Pi_G^\perp D_\ell^{(1)} D_\ell^{(2)} \Pi_G^\perp V_\ell^\top \frac{b_{\ell+1}^{(2)}}{\sqrt{m}} - \langle \frac{b_{\ell+1}^{(1)}}{\sqrt{m}}, \frac{b_{\ell+1}^{(1)}}{\sqrt{m}} \rangle \frac{2}{m} \operatorname{Tr}(D_\ell^{(1)} D_\ell^{(2)}) \right| \leq c\epsilon.$$

Specifically, we have

$$\left\| \sqrt{\frac{2}{m}} \frac{b_{\ell+1}}{\sqrt{m}}^\top V_\ell \Pi_G^\perp D_\ell \right\| \leq \sqrt{c\epsilon + \frac{2}{m} \operatorname{Tr}(D_\ell) \frac{1}{m} \|b_{\ell+1}\|^2} \leq \mathcal{O}(1),$$

and similarly

$$\left\| \sqrt{\frac{2}{m}} \frac{\tilde{b}_{\ell+1}}{\sqrt{m}}^\top V_\ell \Pi_G^\perp \widetilde{D}_\ell \right\| \le \mathcal{O}(1).$$

Next we bound

$$\left\| \frac{b_{\ell+1}}{\sqrt{m}}^\top V_\ell \Pi_G \right\|.$$

Notice that Π_G is a orthogonal projection onto the column space of G, which is at most 2-dimension. One can write $\Pi_G = u_1 u_1^\top + u_2 u_2^\top$, where $\|u_i\| = 1$ or 0. By Lemma 12, fixing u_1, u_2 and V_ℓ, w.p greater than $1 - \delta'$ over the randomness of $W_{\ell+1}, V_{\ell+1}, \cdots, W_L, V_L, v$, we have

$$\left| b_{\ell+1}^\top \frac{1}{\sqrt{m}} V_\ell u_i \right| \le C'' \sqrt{\log \frac{8L}{\delta'}},$$

and

$$\left| \tilde{b}_{\ell+1}^\top \frac{1}{\sqrt{m}} V_\ell u_i \right| \le C'' \sqrt{\log \frac{8L}{\delta'}},$$

for both $i = 1, 2$ when

$$m \ge c L^{2-2\gamma} \log \frac{8L}{\delta'}.$$

Therefore

$$\left\| \frac{b_{\ell+1}}{\sqrt{m}}^\top V_\ell \Pi_G \right\|, \left\| \frac{\tilde{b}_{\ell+1}}{\sqrt{m}}^\top V_\ell \Pi_G \right\| \le \mathcal{O}\left(\sqrt{\log \frac{8L}{\delta'}} \right).$$

Finally, using $I_m = \Pi_G + \Pi_G^\perp$, we have

$$\begin{aligned}
&\left| \frac{2}{m} \frac{b_{\ell+1}^{(1)}}{\sqrt{m}}^\top V_\ell D_\ell^{(1)} D_\ell^{(2)} V_\ell^\top \frac{b_{\ell+1}^{(2)}}{\sqrt{m}} - \langle \frac{b_{\ell+1}^{(1)}}{\sqrt{m}}, \frac{b_{\ell+1}^{(2)}}{\sqrt{m}} \rangle \frac{2}{m} \operatorname{Tr}(D_\ell^{(1)} D_\ell^{(2)}) \right| \\
\le &\left| \frac{2}{m} \frac{b_{\ell+1}^{(1)}}{\sqrt{m}}^\top V_\ell \Pi_G^\perp D_\ell^{(1)} D_\ell^{(2)} \Pi_G^\perp V_\ell^\top \frac{b_{\ell+1}^{(2)}}{\sqrt{m}} - \langle \frac{b_{\ell+1}^{(1)}}{\sqrt{m}}, \frac{b_{\ell+1}^{(1)}}{\sqrt{m}} \rangle \frac{2}{m} \operatorname{Tr}(D_\ell^{(1)} D_\ell^{(2)}) \right| \\
&+ \sqrt{\frac{2}{m}} \left| \frac{b_{\ell+1}^{(1)}}{\sqrt{m}}^\top V_\ell \Pi_G D_\ell^{(1)} D_\ell^{(2)} \Pi_G^\perp V_\ell^\top \frac{b_{\ell+1}^{(2)}}{\sqrt{m}} \sqrt{\frac{2}{m}} \right| \\
&+ \sqrt{\frac{2}{m}} \left| \sqrt{\frac{2}{m}} \frac{b_{\ell+1}^{(1)}}{\sqrt{m}}^\top V_\ell \Pi_G^\perp D_\ell^{(1)} D_\ell^{(2)} \Pi_G V_\ell^\top \frac{b_{\ell+1}^{(2)}}{\sqrt{m}} \right| \\
&+ \frac{2}{m} \left| \frac{b_{\ell+1}^{(1)}}{\sqrt{m}}^\top V_\ell \Pi_G D_\ell^{(1)} D_\ell^{(2)} \Pi_G V_\ell^\top \frac{b_{\ell+1}^{(2)}}{\sqrt{m}} \right| \\
\le &\, c\epsilon + \sqrt{\frac{2}{m}} \mathcal{O}\left(\sqrt{\log \frac{8L}{\delta'}} \right) + \frac{2}{m} \mathcal{O}\left(\log \frac{8L}{\delta'} \right) \le \epsilon.
\end{aligned}$$

The last inequality holds when $m \ge \frac{C}{\epsilon^2} \log \frac{8L}{\delta'}$. □

C.3 Proof of Lemma 10

proof of Lemma 10. The first part of the proof is essentially the same as Lemma 9. Define

$$d_{\ell+1} = D_\ell \frac{1}{\sqrt{m}} V_\ell^\top \frac{b_{\ell+1}}{\sqrt{m}}, \quad \tilde{d}_{\ell+1} = \widetilde{D}_\ell \frac{1}{\sqrt{m}} V_\ell^\top \frac{\tilde{b}_{\ell+1}}{\sqrt{m}}.$$

We know that $d_{\ell+1}$ and $\tilde{d}_{\ell+1}$ depend on W_ℓ only through $W_\ell x_{\ell-1}$ and $W_\ell \tilde{x}_{\ell-1}$. Let $H = [x_{\ell-1}, \tilde{x}_{\ell-1}]$. Then

$$\begin{aligned}
&\left| \frac{2}{m} \langle W_\ell^\top d_{\ell+1}^{(1)}, W_\ell^\top d_{\ell+1}^{(2)} \rangle - 2 \langle d_{\ell+1}^{(1)}, d_{\ell+1}^{(2)} \rangle \right| \\
\le &\left| \frac{2}{m} \langle \Pi_H^\perp W_\ell^\top d_{\ell+1}^{(1)}, \Pi_H^\perp W_\ell^\top d_{\ell+1}^{(2)} \rangle - 2 \langle d_{\ell+1}^{(1)}, d_{\ell+1}^{(2)} \rangle \right| + \left| \frac{2}{m} \langle \Pi_H W_\ell^\top d_{\ell+1}^{(1)}, \Pi_H^\perp W_\ell^\top d_{\ell+1}^{(2)} \rangle \right| \\
&+ \left| \frac{2}{m} \langle \Pi_H^\perp W_\ell^\top d_{\ell+1}^{(1)}, \Pi_H W_\ell^\top d_{\ell+1}^{(2)} \rangle \right| + \left| \frac{2}{m} \langle \Pi_H W_\ell^\top d_{\ell+1}^{(1)}, \Pi_H W_\ell^\top d_{\ell+1}^{(2)} \rangle \right|.
\end{aligned}$$

Since $\|d_{\ell+1}\|, \|\tilde{d}_{\ell+1}\| = \mathcal{O}(1)$, similar to Lemma 11, when $m \ge 1 + \log \frac{6}{\delta}$, w.p at least $1 - \delta$ we have

$$\left| \frac{2}{m} \langle \Pi_H^\perp W_\ell^\top d_{\ell+1}^{(1)}, \Pi_H^\perp W_\ell^\top d_{\ell+1}^{(2)} \rangle - 2 \langle d_{\ell+1}^{(1)}, d_{\ell+1}^{(2)} \rangle \right| \le \mathcal{O}\left(\sqrt{\frac{1 + \log \frac{6}{\delta}}{m}} \right),$$

and

$$\left\|\sqrt{\frac{2}{m}}\Pi_H^{\perp} W_\ell^\top d_{\ell+1}^{(i)}\right\| = \mathcal{O}(1), \quad i = 1, 2,$$

Using the same argument as in the proof of Lemma 9, we decompose Π_H into two vectors w_1 and w_2, whose randomness comes from $W_1, V_1, \cdots, W_{\ell-1}, V_{\ell-1}$. By writing

$$w_i^\top W_\ell^\top d_{\ell+1}^{(i)} = \langle b_{\ell+1}^{(i)}, \frac{1}{\sqrt{m}} V_\ell D_\ell^{(i)} \frac{1}{\sqrt{m}} W_\ell w_i \rangle,$$

we can also apply Lemma 12. Then we conclude that w.p. greater than $1 - \delta'$ over the randomness of v, we have

$$\|\Pi_H W_\ell^\top d_{\ell+1}\|, \|\Pi_H W_\ell^\top \tilde{d}_{\ell+1}\| = \mathcal{O}\left(\sqrt{\log \frac{8L}{\delta'}}\right),$$

when

$$m \geq cL^{2-2\gamma} \log \frac{8L}{\delta'}.$$

Then exactly the same result of Lemma 9 holds.

For the second part, notice that

$$\begin{aligned}
\frac{1}{m}\sqrt{\frac{1}{m}}\sqrt{\frac{2}{m}}\langle W_\ell^\top D_\ell^{(1)} V_\ell^\top b_{\ell+1}^{(1)}, b_{\ell+1}^{(2)}\rangle &= \sqrt{\frac{2}{m}}\langle W_\ell^\top D_\ell^{(1)} \sqrt{\frac{1}{m}} V_\ell^\top \frac{b_{\ell+1}^{(1)}}{\sqrt{m}}, \frac{b_{\ell+1}^{(2)}}{\sqrt{m}}\rangle \\
&= \sqrt{\frac{2}{m}}\langle W_\ell^\top d_{\ell+1}^{(1)}, \frac{b_{\ell+1}^{(2)}}{\sqrt{m}}\rangle \\
&= \sqrt{\frac{2}{m}}\langle \Pi_H^{\perp} W_\ell^\top d_{\ell+1}^{(1)}, \frac{b_{\ell+1}^{(2)}}{\sqrt{m}}\rangle + \sqrt{\frac{2}{m}}\langle \Pi_H \frac{1}{\sqrt{m}} W_\ell^\top d_{\ell+1}^{(1)}, b_{\ell+1}^{(2)}\rangle.
\end{aligned}$$

Conditioned on $x_{\ell-1}, \tilde{x}_{\ell-1}, W_\ell x_{\ell-1}$, and $W_\ell \tilde{x}_{\ell-1}$, W_ℓ is independent of $b_{\ell+1}, \tilde{b}_{\ell+1}, d_{\ell+1}$, and $\tilde{d}_{\ell+1}$. Furthermore, we have $\Pi_H^{\perp} W_\ell^\top =_d \Pi_H^{\perp} \widehat{W}_\ell^\top$, where $\widehat{W}_\ell$ is an i.i.d. copy of W_ℓ. Then for the first term, with probability at least $1 - \tilde{\delta}/2$, we have for all $(x^{(1)}, x^{(2)}) \in \{(x, x), (x, \tilde{x}), (\tilde{x}, x), (\tilde{x}, \tilde{x})\}$,

$$\left|\sqrt{\frac{2}{m}}\langle \Pi_H^{\perp} W_\ell^\top d_{\ell+1}^{(1)}, \frac{b_{\ell+1}^{(2)}}{\sqrt{m}}\rangle\right| \leq \left\|\Pi_H^{\perp} \frac{b_{\ell+1}^{(2)}}{\sqrt{m}}\right\| \|d_{\ell+1}^{(1)}\| \sqrt{\frac{2c \log \frac{16}{\tilde{\delta}}}{m}} \leq \mathcal{O}\left(\sqrt{\frac{\log \frac{16}{\tilde{\delta}}}{m}}\right).$$

For the second term, write $\Pi_H = w_1 w_1^\top + w_2 w_2^\top$, where $\|w_i\| = 1$ or 0. Then by Lemma 12, with probability at least $1 - \tilde{\delta}/2$, for all $(x^{(1)}, x^{(2)}) \in \{(x, x), (x, \tilde{x}), (\tilde{x}, x), (\tilde{x}, \tilde{x})\}$, when $m \geq cL^{2-2\gamma} \log \frac{16L}{\tilde{\delta}}$, we have

$$\begin{aligned}
\left|\sqrt{\frac{2}{m}}\langle w_i w_i^\top \frac{1}{\sqrt{m}} W_\ell^\top d_{\ell+1}^{(1)}, b_{\ell+1}^{(2)}\rangle\right| &= \left|\sqrt{\frac{2}{m}} w_i^\top \frac{1}{\sqrt{m}} W_\ell^\top d_{\ell+1}^{(1)} \langle w_i, b_{\ell+1}^{(2)}\rangle\right| \\
&\leq \sqrt{\frac{2}{m}}\|w_i\| \left\|\frac{1}{\sqrt{m}} W_\ell^\top\right\| \|d_{\ell+1}^{(1)}\| \left|\langle w_i, b_{\ell+1}^{(2)}\rangle\right| \\
&\leq \mathcal{O}\left(\sqrt{\frac{\log \frac{16L}{\tilde{\delta}}}{m}}\right).
\end{aligned}$$

□

D Proof of Theorem 5

Proof. For $x, \tilde{x} \in \mathbb{S}^{D-1}$, we have $K_\ell(x, x) = K_\ell(\tilde{x}, \tilde{x}) = 1$ for all ℓ. Hence we only need to study when $x \neq \tilde{x}$. Note we have

$$K_\ell(x, \tilde{x}) = \Gamma_\sigma(K_{\ell-1})(x, \tilde{x}) = \hat{\sigma}(K_{\ell-1}(x, \tilde{x})), \text{ and } \Gamma_{\sigma'}(K_\ell)(x, \tilde{x}) = \widehat{\sigma'}(K_\ell(x, \tilde{x})).$$

For simplicity, we use K_ℓ to denote $K_\ell(x, \tilde{x})$, where $x \neq \tilde{x}$ and $x, \tilde{x} \in \mathbb{S}^{D-1}$.

Recall that

$$\hat{\sigma}(\rho) = \frac{\sqrt{1 - \rho^2} + \left(\pi - \cos^{-1}(\rho)\right)\rho}{\pi}, \text{ and } \widehat{\sigma'}(\rho) = \frac{\pi - \cos^{-1}(\rho)}{\pi}.$$

Hence we have $\hat{\sigma}(1) = 1$, $K_{\ell-1} \leq \hat{\sigma}(K_{\ell-1}) = K_\ell$, $(\hat{\sigma})'(\rho) = \widehat{\sigma'}(\rho) \in [0, 1]$, and $(\widehat{\sigma'})'(\rho) \geq 0$. Then $\hat{\sigma}$ is a convex function.

Since $\{K_\ell\}$ is an increasing sequence and $|K_\ell| \le 1$, we have K_ℓ converges as $\ell \to \infty$. Taking the limit of both sides of $\hat{\sigma}(K_{\ell-1}) = K_\ell$, we have $K_\ell \to 1$ as $\ell \to \infty$.
For K_ℓ, we also have

$$K_\ell = \hat{\sigma}(K_{\ell-1}) = \frac{\sqrt{1-K_{\ell-1}^2} + (\pi - \cos^{-1}(K_{\ell-1}))K_{\ell-1}}{\pi} = K_{\ell-1} + \frac{\sqrt{1-K_{\ell-1}^2} - \cos^{-1}(K_{\ell-1})K_{\ell-1}}{\pi}.$$

Let $e_\ell = 1 - K_\ell$, we can easily check that

$$e_{\ell-1} - \frac{e_{\ell-1}^{3/2}}{\pi} \le e_\ell \le e_{\ell-1} - \frac{2\sqrt{2}e_{\ell-1}^{3/2}}{3\pi}. \tag{19}$$

Hence as $e_\ell \to 0$, we have $\frac{e_\ell}{e_{\ell-1}} \to 1$, which implies $\{K_\ell\}$ converges sublinearly.

Assume $e_\ell = \frac{C}{\ell^p} + \mathcal{O}(\ell^{-(p+1)})$. By taking the assumption into (19) and comparing the highest order of both sides, we have $p = 2$.

Thus $\exists C$, s.t. $|1 - K_\ell| \le \frac{C}{\ell^2}$, i.e. the convergence rate of K_ℓ is $\mathcal{O}\left(\frac{1}{\ell^2}\right)$.

Lemma 13. *For each $K_0 < 1$, there exists $p > 0$ and $n_0 = n_0(\delta) > 0$, such that $K_n \le 1 - \frac{9\pi^2}{2(n+n_0)^{2+\frac{\log(L)^p}{L}}}$, $\forall n = 0, \dots, L$, when L is large.*

Proof. First, solve $K_0 \le 1 - \frac{9\pi^2}{2n^{2+\frac{\log(L)^p}{L}}}$. Then we can choose $n_0 \ge \sqrt{\frac{9\pi^2}{2\delta}} \ge \sqrt{\frac{9\pi^2}{2(1-K_0)}}$, which is independent of L and n. For the rest of the proof, without loss of generality, we just use n instead of $n+n_0$. Also for small δ(when δ is not small enough we can pick a small $\delta_0 < \delta$ and let $n_0 \ge \sqrt{\frac{9\pi^2}{2\delta_0}}$), we have $\frac{9\pi^2}{2(n+n_0)^{2+\frac{\log(L)^p}{L}}} \le \delta(\text{or } \delta_0)$ which is also small.

Let $K_n = 1 - \epsilon$. Then, when ϵ is small, we have

$$K_{n+1} - K_n = \hat{\sigma}(K_n) - K_n = \mathcal{O}(\epsilon^{3/2}).$$

Also, we have

$$\begin{aligned}\left(1 - \frac{9\pi^2}{2(n+1)^{2+\frac{\log(L)^p}{L}}}\right) - \left(1 - \frac{9\pi^2}{2n^{2+\frac{\log(L)^p}{L}}}\right) &= \mathcal{O}\left(\frac{1}{n^{3+\frac{\log(L)^p}{L}}}\right) \\ &\ge \mathcal{O}\left(\left(\frac{1}{n^{2+\frac{\log(L)^p}{L}}}\right)^{3/2}\right) = \mathcal{O}\left(\frac{1}{n^{3+\frac{3\log(L)^p}{2L}}}\right).\end{aligned}$$

Overall, we want an upper bound for K_n and from the above we only know that K_n is of order $1 - \mathcal{O}(n^{-2})$ but this order may hide some terms of logarithmic order. Hence we use the order $1 - \mathcal{O}(n^{-(2+\epsilon)})$ to provide an upper bound of K_n. Here $\frac{\log(L)^p}{L}$ is constructed for the convenience of the rest of the proof. □

Let $N_0 = N_0(L)$ be the solution of

$$\cos\left(\pi\left(1 - \left(\frac{n+1}{n+2}\right)^{3-\frac{\log(L)^2}{L}}\right)\right) = \hat{\sigma}\left(\cos\left(\pi\left(1 - \left(\frac{n}{n+1}\right)^{3-\frac{\log(L)^2}{L}}\right)\right)\right),$$

where for $N_0 < n < N_L$ with some N_L, we have

$$\cos\left(\pi\left(1 - \left(\frac{n+1}{n+2}\right)^{3-\frac{\log(L)^2}{L}}\right)\right) \ge \hat{\sigma}\left(\cos\left(\pi\left(1 - \left(\frac{n}{n+1}\right)^{3-\frac{\log(L)^2}{L}}\right)\right)\right).$$

One can check by series expansion that $N_0 = N_0(L) \le 5\frac{L}{\log(L)^2}$.

Next we would like to find n such that

$$K_n = \cos\left(\pi\left(1 - \left(\frac{5\frac{L}{\log(L)^2}}{5\frac{L}{\log(L)^2} + 1}\right)^{3-\frac{\log(L)^2}{L}}\right)\right).$$

By series expansion, we know

$$\cos\left(\pi\left(1-\left(\frac{5\frac{L}{\log(L)^2}}{5\frac{L}{\log(L)^2}+1}\right)^{3-\frac{\log(L)^2}{L}}\right)\right)\geq 1-\frac{9\pi^2}{2\left(\frac{5L}{\log(L)^2}\right)^2}.$$

Then it suffices to solve

$$1-\frac{9\pi^2}{2(\frac{5L}{\log(L)^2})^2}\geq 1-\frac{9\pi^2}{2n^{2+\frac{\log(L)^p}{L}}}\geq K_n,\ i.e.,\ n^{2+\frac{\log(L)^p}{L}}\leq\left(\frac{5L}{\log(L)^2}\right)^2. \tag{20}$$

Lemma 14. *When $q>p-1$, we have $n\lesssim\frac{5L}{\log(L)^2}-\log(L)^q$ satisfies (20).*

Proof. If the condition above holds, we have

$$n^{2+\frac{\log(L)^p}{L}}\leq\left(\frac{5L}{\log(L)^2}-\log(L)^q\right)^{2+\frac{\log(L)^p}{L}},$$

which is

$$\begin{aligned}n^{1+\frac{\log(L)^p}{2L}}&\leq\left(\frac{5L}{\log(L)^2}-\log(L)^q\right)\left(\frac{5L}{\log(L)^2}-\log(L)^q\right)^{\frac{\log(L)^p}{2L}}\\&\leq\left(\frac{5L}{\log(L)^2}-\log(L)^q\right)\left(1+\frac{\log(L)^p\log(\frac{5L}{\log(L)^2})}{2L}\right)\\&=\frac{5L}{\log(L)^2}-\log(L)^q+\frac{5}{2}\log(L)^{p-2}\log\left(\frac{5L}{\log(L)^2}\right)-\frac{1}{2L}\log(L)^{p+q}\log\left(\frac{5L}{\log(L)^2}\right),\end{aligned}$$

where $\left(\frac{5L}{\log(L)^2}-\log(L)^q\right)^{\frac{\log(L)^p}{2L}}\to 1$ as $L\to\infty$.

Thus we have $q>p-1$. □

Just pick $q=p$. Then we have $n^{1+\frac{\log(L)^p}{2L}}\lesssim\frac{5L}{\log(L)^2}$ and $n\lesssim\frac{5L}{\log(L)^2}-\log(L)^p$.

Lemma 15. *When L is large enough, we have*

$$\cos\left(\pi\left(1-\left(\frac{n}{n+1}\right)^{3+\frac{\log(L)^2}{L}}\right)\right)\leq K_n\leq\cos\left(\pi\left(1-\left(\frac{n+\log(L)^p}{n+\log(L)^p+1}\right)^{3-\frac{\log(L)^2}{L}}\right)\right).$$

Proof. Let $F(n)=\cos\left(\pi\left(1-\left(\frac{n+\log(L)^p}{n+\log(L)^p+1}\right)^{3-\frac{\log(L)^p}{L}}\right)\right)$.

For the right hand side, when $n\gtrsim\frac{5L}{\log(L)^2}-\log(L)^p$, we have, by series expansion, $F(n+1)\geq\hat{\sigma}\left(F(n)\right)$. Also, when $n\sim aL$, where $0<a\leq 1$, we have

$$F(n+1)-\hat{\sigma}(F(n))=\mathcal{O}\left(\frac{3\left(2\pi^2\mathrm{a}\log^{10}(L)+\pi^2\log^8(L)\right)}{2L^4\left(\mathrm{a}\log^2(L)+5\right)^4}\right)>0.$$

Then for $\frac{5L}{\log(L)^2}-\log(L)^p\lesssim n\lesssim L$, we have $F(n+1)\geq\hat{\sigma}\left(F(n)\right)$ and thus $K_n\leq F(n)$.

When $n\lesssim\frac{5L}{\log(L)^2}-\log(L)^p$, we have $F(n+1)\leq\hat{\sigma}\left(F(n)\right)$. Hence $K_n\leq F(n)$.

For the left hand side,

$$\begin{aligned}&\cos\left(\pi\left(1-\left(\frac{n+1}{n+2}\right)^{3+\frac{\log(L)^2}{L}}\right)\right)-\hat{\sigma}\left(\cos\left(\pi\left(1-\left(\frac{n}{n+1}\right)^{3+\frac{\log(L)^2}{L}}\right)\right)\right)\\&\qquad\sim-\frac{27\pi^2}{2n^4}-\frac{3\pi^2\log(L)^2}{n^3L},\ \forall n=1,...,L.\end{aligned}$$

Hence we have the left hand side. □

From Lemma 15, by series expansion, we have

$$|1 - K_n| \leq \frac{\left(3\pi + \frac{\pi \log(L)^2}{L}\right)^2}{2n^2} \sim \frac{9\pi^2}{2n^2},$$

when L is large.

Moreover, we can get

$$\left(\frac{n}{n+1}\right)^{3+\frac{\log(L)^2}{L}} \leq \Gamma_{\sigma'}(K_n) \leq \left(\frac{n+\log(L)^p}{n+\log(L)^p+1}\right)^{3-\frac{\log(L)^2}{L}}.$$

Then

$$\left(\frac{\ell-1}{L}\right)^{3+\frac{\log(L)^2}{L}} \leq \prod_{i=\ell}^{L} \Gamma_{\sigma'}(K_{i-1}) \leq \left(\frac{\ell+\log(L)^p-1}{L+\log(L)^p}\right)^{3-\frac{\log(L)^2}{L}}.$$

Let $N = \log(L)^p$. For the right hand side, if we sum over ℓ, we have

$$\begin{aligned}\frac{1}{L}\sum_{\ell=1}^{L}\left(\frac{\ell+N-1}{L+N}\right)^{3-\frac{\log(L)^2}{L}} &\leq \frac{1}{L}\int_1^{L+1}\left(\frac{x+N-1}{L+N}\right)^{3-\frac{\log(L)^2}{L}} dx \\ &= \frac{\left((L+N)^{4-\frac{\log(L)^2}{L}} - (N)^{4-\frac{\log(L)^2}{L}}\right)}{L(L+N)^{3-\frac{\log(L)^2}{L}}\left(4-\frac{\log(L)^2}{L}\right)}.\end{aligned}$$

Taking the limit of both sides, we have

$$\lim_{L\to\infty}\frac{1}{L}\sum_{\ell=1}^{L}\left(\frac{\ell+N-1}{L+N}\right)^{3-\frac{\log(L)^2}{L}} \leq \frac{1}{4}.$$

Similarly, by

$$\frac{1}{L}\sum_{i=1}^{L}\left(\frac{\ell-1}{L}\right)^{3+\frac{\log(L)^2}{L}} \geq \frac{1}{L}\int_1^{L}\left(\frac{x-1}{L}\right)^{3+\frac{\log(L)^2}{L}} dx = \frac{(L-1)^{4+\frac{\log(L)^2}{L}}}{\left(4+\frac{\log(L)^2}{L}\right)L^{4+\frac{\log(L)^2}{L}}},$$

we have

$$\lim_{L\to\infty}\frac{1}{L}\sum_{i=1}^{L}\left(\frac{\ell-1}{L}\right)^{3+\frac{\log(L)^2}{L}} \geq \frac{1}{4}.$$

Hence,

$$\begin{aligned}\lim_{L\to\infty}\frac{1}{L}\sum_{\ell=1}^{L}\left(\frac{\ell+N-1}{L+N}\right)^{3-\frac{\log(L)^2}{L}} &= \lim_{L\to\infty}\frac{1}{L}\sum_{\ell=1}^{L}\left(\frac{\ell-1}{L}\right)^{3+\frac{\log(L)^2}{L}} \\ &= \lim_{L\to\infty}\frac{1}{L}\sum_{\ell=1}^{L}\prod_{i=\ell}^{L}\Gamma_{\sigma'}(K_{i-1}) = \frac{1}{4}.\end{aligned}$$

Recall from previous discussion, $K_\ell = 1 - \mathcal{O}(\frac{1}{\ell^2})$. Therefore,

$$\lim_{L\to\infty}\frac{1}{L}\sum_{\ell=1}^{L}K_{\ell-1}\prod_{i=\ell}^{L}\Gamma_{\sigma'}(K_{i-1}) = \frac{1}{4}.$$

Also, when L is large, we have

$$\frac{\left((L+N)^{4-\frac{\log(L)^2}{L}} - (N)^{4-\frac{\log(L)^2}{L}}\right)}{L(L+N)^{3-\frac{\log(L)^2}{L}}\left(4-\frac{\log(L)^2}{L}\right)} > \frac{1}{4} > \frac{(L-1)^{4+\frac{\log(L)^2}{L}}}{\left(4+\frac{\log(L)^2}{L}\right)L^{4+\frac{\log(L)^2}{L}}}.$$

Hence we can estimate the convergence rate of the normalized kernel

$$\left|\frac{1}{L}\sum_{\ell=1}^{L}K_{\ell-1}\prod_{i=\ell}^{L}\Gamma_{\sigma'}(K_{i-1}) - \frac{1}{4}\right| = \left|\frac{1}{L}\sum_{\ell=1}^{L}\left(K_{\ell-1}\left(\prod_{i=\ell}^{L}\Gamma_{\sigma'}(K_{i-1}) - \frac{1}{4}\right) + \frac{1}{4}(K_{\ell-1}-1)\right)\right|$$

$$
\begin{aligned}
&\leq \left|\frac{1}{L}\sum_{\ell=1}^{L}\prod_{i=\ell}^{L}\Gamma_{\sigma'}(K_{i-1})-\frac{1}{4}\right|+\frac{1}{4}\left|\frac{1}{L}\sum_{\ell=1}^{L}(K_{\ell-1}-1)\right| \\
&\leq \left|\frac{\left((L+N)^{4-\frac{\log(L)^2}{L}}-(N)^{4-\frac{\log(L)^2}{L}}\right)}{L(L+N)^{3-\frac{\log(L)^2}{L}}\left(4-\frac{\log(L)^2}{L}\right)}-\frac{(L-1)^{4+\frac{\log(L)^2}{L}}}{\left(4+\frac{\log(L)^2}{L}\right)L^{4+\frac{\log(L)^2}{L}}}\right| \\
&\qquad\qquad +\frac{1}{4}\left|\frac{1}{L}\sum_{i=1}^{L}(K_{\ell-1}-1)\right| \\
&\lesssim \frac{4\log(L)^p+\log(L)^2}{16L}=\mathcal{O}\left(\frac{\text{poly}\log(L))}{L}\right)
\end{aligned}
$$

□

E Proof of Theorem 6

Proof. We denote $K_{\ell,L}$ to be the ℓ-th layer of K when the depth is L, which is originally denoted by K_ℓ.

Let $S_{\ell,L}=\frac{K_{\ell,L}}{(1+\alpha^2)^\ell}=\frac{K_{\ell,L}}{(1+1/L^2)^\ell}$ and $S_0=K_0$, then $\Gamma_\sigma(K_{\ell,L})=(1+\alpha^2)^\ell\hat{\sigma}(S_{\ell,L})$ and $\Gamma_{\sigma'}(K_{\ell,L})=\widehat{\sigma'}(S_{\ell,L})$. Hence we can rewrite the recursion to be

$$
S_{\ell,L}=\frac{S_{\ell-1,L}+\alpha^2\hat{\sigma}(S_{\ell-1,L})}{(1+\alpha^2)}\geq S_{\ell-1,L}. \tag{21}
$$

Moreover, since $S_{\ell,L}-S_{\ell-1,L}=\frac{\alpha^2}{1+\alpha^2}(\hat{\sigma}(S_{\ell-1,L})-S_{\ell-1,L})$ and $(\hat{\sigma}(S_{\ell-1,L})-S_{\ell-1,L})$ is decreasing, we can have

$$
S_{\ell,L}\leq S_0+\frac{(\hat{\sigma}(S_0)-S_0)\ell}{L^2}.
$$

Denote $P_{\ell+1,L}=B_{\ell+1,L}(1+\alpha^2)^{-(L-\ell)}=\prod_{i=\ell}^{L-1}\frac{1+\alpha^2\widehat{\sigma'}(S_{i,L})}{1+\alpha^2}$. Since

$$
1-\frac{1+\alpha^2\widehat{\sigma'}(S_{i,L})}{1+\alpha^2}=\frac{\alpha^2(1-\widehat{\sigma'}(S_{i,L}))}{1+\alpha^2}=\frac{1-\widehat{\sigma'}(S_{i,L})}{L^2+1},
$$

we have

$$
1-P_{\ell+1,L}=1-\prod_{i=\ell}^{L-1}\left(1-\frac{1-\widehat{\sigma'}(S_{i,L})}{L^2+1}\right)\leq\sum_{i=\ell}^{L-1}\frac{1-\widehat{\sigma'}(S_{i,L})}{L^2+1}=\frac{L-\ell-\sum_{i=\ell}^{L-1}\widehat{\sigma'}(S_{i,L})}{L^2+1},
$$

where $\ell=1,\ldots,L-1$. For $P_{L+1,L}$, we have $1-P_{L+1,L}=0$.

Then we can rewrite the normalized kernel to be

$$
\overline{\Omega}_L=\frac{1}{2L}\sum_{\ell=1}^{L}P_{\ell+1,L}(\hat{\sigma}(S_{\ell-1,L})+S_{\ell-1,L}\widehat{\sigma'}(S_{\ell-1,L})).
$$

Hence we have the bound for each layer

$$
\begin{aligned}
&\left|P_{\ell+1,L}(\hat{\sigma}(S_{\ell-1,L})+S_{\ell-1,L}\widehat{\sigma'}(S_{\ell-1,L}))-(\hat{\sigma}(S_0)+S_0\widehat{\sigma'}(S_0))\right| \\
&\quad\leq\left|P_{\ell+1,L}\right|\cdot\left|(\hat{\sigma}(S_{\ell-1,L})+S_{\ell-1,L}\widehat{\sigma'}(S_{\ell-1,L}))-(\hat{\sigma}(S_0)+S_0\widehat{\sigma'}(S_0))\right|+\left|\hat{\sigma}(S_0)+S_0\widehat{\sigma'}(S_0)\right|\cdot\left|1-P_{\ell+1,L}\right| \\
&\quad\leq\left|\widehat{\sigma'}(S_{\ell-1,L})(S_{\ell-1,L}-S_0)\right|+\left|\widehat{\sigma'}(S_{\ell-1,L})S_{\ell-1,L}-\widehat{\sigma'}(S_0)S_0\right|+\left|\hat{\sigma}(S_0)+S_0\widehat{\sigma'}(S_0)\right|\cdot\left|1-P_{\ell+1,L}\right| \\
&\quad=2\left|\widehat{\sigma'}(S_{\ell-1,L})(S_{\ell-1,L}-S_0)\right|+\left|S_0(\widehat{\sigma'}(S_{\ell-1,L})-\widehat{\sigma'}(S_0))\right|+\left|\hat{\sigma}(S_0)+S_0\widehat{\sigma'}(S_0)\right|\cdot\left|1-P_{\ell+1,L}\right| \\
&\quad\leq\frac{2\widehat{\sigma'}(S_{\ell-1,L})(\hat{\sigma}(S_0)-S_0)\ell}{L^2}+\frac{|S_0|(\hat{\sigma}(S_0)-S_0)(\ell-1)}{\pi L^2\sqrt{1-S_{\ell-1,L}^2}}+\left|\hat{\sigma}(S_0)+S_0\widehat{\sigma'}(S_0)\right|\frac{L-\ell-\sum_{i=\ell}^{L-1}\widehat{\sigma'}(S_{i,L})}{L^2+1} \\
&\quad\leq\frac{2\widehat{\sigma'}(S_{\ell-1,L})(\hat{\sigma}(S_0)-S_0)\ell}{L^2}+\frac{|S_0|(\hat{\sigma}(S_0)-S_0)(\ell-1)}{\pi L^2\sqrt{1-S_{\ell-1,L}^2}}+\left|\hat{\sigma}(S_0)+S_0\widehat{\sigma'}(S_0)\right|\frac{L-\ell-(L-\ell)\widehat{\sigma'}(S_0)}{L^2+1}.
\end{aligned}
$$

Therefore we have the bound for the normalized kernel

$$
\begin{aligned}
&\left|\overline{\Omega}_L - \frac{1}{2}(\hat{\sigma}(S_0) + S_0\widehat{\sigma'}(S_0))\right| \\
&\quad = \left|\frac{1}{2L}\sum_{\ell=1}^{L}\left(P_{\ell+1,L}(\hat{\sigma}(S_{\ell-1,L}) + S_{\ell-1,L}\widehat{\sigma'}(S_{\ell-1,L}))\right) - \frac{1}{2}(\hat{\sigma}(S_0) + S_0\widehat{\sigma'}(S_0))\right| \\
&\quad \le \frac{1}{2L}\sum_{\ell=1}^{L}\left(\frac{2\widehat{\sigma'}(S_{\ell-1,L})(\hat{\sigma}(S_0) - S_0)\ell}{L^2} + \frac{|S_0|(\hat{\sigma}(S_0) - S_0)(\ell-1)}{\pi L^2\sqrt{1 - S_{\ell-1,L}^2}}\right) \\
&\qquad\qquad + \frac{1}{2L}\sum_{\ell=1}^{L-1}\left(\left|\hat{\sigma}(S_0) + S_0\widehat{\sigma'}(S_0)\right|\frac{L - \ell - (L-\ell)\widehat{\sigma'}(S_0)}{L^2+1}\right) \\
&\quad \le \frac{1}{2L}\left(\frac{L+1}{L}(\hat{\sigma}(S_0) - S_0) + \frac{|S_0|(\hat{\sigma}(S_0) - S_0)L(L-1)}{2\pi L^2 C} + \left|\hat{\sigma}(S_0) + S_0\widehat{\sigma'}(S_0)\right|\frac{\frac{L(L-1)}{2}(1 - \widehat{\sigma'}(S_0))}{L^2+1}\right) \\
&\quad \sim \left(\frac{(\hat{\sigma}(S_0) - S_0)}{2}\left(1 + \frac{|S_0|}{2\pi C}\right) + \frac{1}{2}\left|\hat{\sigma}(S_0) + S_0\widehat{\sigma'}(S_0)\right|(1 - \widehat{\sigma'}(S_0))\right)\frac{1}{L}
\end{aligned}
$$

where $C = C(\delta) = \sqrt{1 - (1-\delta)^2}$ and $S_0 = K_0$. □