Discussion on Neural Tangent Kernels (NTK) and Graph Neural Tangent Kernels (GNTK)

Prepared by Jiayao Zhang
jiayaozhang@acm.org

July 13, 2019

Contents

1. Neural Tangent Kernels (NTK)
   - Contruction of the NTK due to [ADH+19]
   - Convergence to the NTK at Initialization [ADH+19]
   - Equivalence between NN and NTK
2. Graph Neural Nets (GNN)
3. Graph Neural Tangent Kernel (GNTK)
   - Overview
   - Formulae
   - Theoretical Analyses
     - Proof of Theorem 4.2 [DHP+19]
     - Proof of Theorem 4.3 [DHP+19]

Contents

1. Neural Tangent Kernels (NTK)
   - Contruction of the NTK due to [ADH+19]
   - Convergence to the NTK at Initialization [ADH+19]
   - Equivalence between NN and NTK
2. Graph Neural Nets (GNN)
3. Graph Neural Tangent Kernel (GNTK)
   - Overview
   - Formulae
   - Theoretical Analyses
     - Proof of Theorem 4.2 [DHP+19]
     - Proof of Theorem 4.3 [DHP+19]

Overview

- The Neural Tangent Kernel (NTK) can be given by the kernel function

$$k\left(\boldsymbol{x}, \boldsymbol{x}'\right) = \mathbb{E}_{\boldsymbol{\theta}} \left\langle \frac{\partial f(\boldsymbol{\theta}, \boldsymbol{x})}{\partial \boldsymbol{\theta}}, \frac{\partial f(\boldsymbol{\theta}, \boldsymbol{x}')}{\partial \boldsymbol{\theta}} \right\rangle, \tag{1}$$

- where the gradient $\frac{\partial f(\boldsymbol{\theta}, \boldsymbol{x})}{\partial \boldsymbol{\theta}}$ appears from the gradient descent.

f = beta_0 + beta_1 \cdot x
k(x, x') = \Exp_{\beta = [\beta_0, \beta_1]} [x^T x' + 1]

Infinite Width Limit of the MLP (1/7)

- Write $f(\boldsymbol{\theta}, \boldsymbol{x}) \in \mathbb{R}$ the output of a NN, $\boldsymbol{\theta} \in \mathbb{R}^N$ all the parameters, and $\boldsymbol{x} \in \mathbb{R}^d$ the input. Given a training dataset $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n \subset \mathbb{R}^d \times \mathbb{R}$, consider training the neural network by minimizing the squared loss over training data:

$$\ell(\boldsymbol{\theta}) = \frac{1}{2} \sum_{i=1}^{n} \left(f(\boldsymbol{\theta}, \boldsymbol{x}_i) - y_i \right)^2 . \tag{2}$$

Infinite Width Limit of the MLP (2/7)

- For an Multi-Layer Perceptron (MLP, feedforward nets with fully-connected layers) with L layers, $d_1, \ldots, d_L$ neurons per layer, and $d_0 = d$, where $\boldsymbol{x} \in \mathbb{R}^d$. Let $\boldsymbol{x} \in \mathbb{R}^d$ be the input, write $\boldsymbol{g}^{(0)}(\boldsymbol{x}) = \boldsymbol{x}$ and $d_0 = d$, and

$$\begin{aligned} \boldsymbol{f}^{(h)}(\boldsymbol{x}) &= \boldsymbol{W}^{(h)} \boldsymbol{g}^{(h-1)}(\boldsymbol{x}) \in \mathbb{R}^{d_h}, \\ \boldsymbol{g}^{(h)}(\boldsymbol{x}) &= \sqrt{\frac{c_\sigma}{d_h}} \sigma\left(\boldsymbol{f}^{(h)}(\boldsymbol{x})\right) \in \mathbb{R}^{d_h}, \qquad h = 1, 2, \ldots, L. \end{aligned} \tag{3}$$

LeCun

where $\boldsymbol{W}^{(h)} \in \mathbb{R}^{d_h \times d_{h-1}}$ is the weight matrix in the h-th layer ($h \in [L]$), $\sigma : \mathbb{R} \to \mathbb{R}$ is a coordinate-wise activation function, and $c_\sigma = \left(\mathbb{E}_{z \sim \mathcal{N}(0,1)}\left[\sigma(z)^2\right]\right)^{-1}$. for ReLU, c_{\sigma} = 2.

expected length / variance of all neurons.
(Poole et al., 2016, exponential expressive)

Infinite Width Limit of the MLP (3/7)

- The last layer of the neural network is

$$\begin{aligned} f(\boldsymbol{\theta}, \boldsymbol{x}) =& f^{(L+1)}(\boldsymbol{x}) = \boldsymbol{W}^{(L+1)} \cdot \boldsymbol{g}^{(L)}(\boldsymbol{x}) \\ =& \boldsymbol{W}^{(L+1)} \cdot \sqrt{\frac{c_\sigma}{d_L}} \sigma \left(\boldsymbol{W}^{(L)} \cdot \sqrt{\frac{c_\sigma}{d_{L-1}}} \right. \\ & \left. \times \sigma \left(\boldsymbol{W}^{(L-1)} \cdots \sqrt{\frac{c_\sigma}{d_1}} \sigma \left(\boldsymbol{W}^{(1)} \boldsymbol{x} \right) \right) \right), \end{aligned} \tag{4}$$

where $\boldsymbol{W}^{(L+1)} \in \mathbb{R}^{1 \times d_L}$ is the weights in the final layer, and $\boldsymbol{\theta} = \left(\boldsymbol{W}^{(1)}, \ldots, \boldsymbol{W}^{(L+1)} \right)$ represents all the parameters in the network.

Infinite Width Limit of the MLP (4/7)

- All the weights are initialized as i.i.d. $\mathcal{N}(0,1)$. In the limit of $d_1, d_2, \ldots, d_L \to \infty$, the scaling factor $\sqrt{c_\sigma / d_h}$ in Equation (4) ensures that the norm of $\boldsymbol{g}^{(h)}(\boldsymbol{x})$ for each $h \in [L]$ is approximately preserved at initialization (see[#] `du2018global`).
- In particular, for ReLU activation, we have $\mathbb{E}\left[\|\boldsymbol{g}^{(h)}(\boldsymbol{x})\|^2\right] = \|\boldsymbol{x}\|^2$ $(\forall h \in [L])$.

Du et al. Gradient descent finds global minima of deep neural networks. https://arxiv.org/abs/1811.03804

Infinite Width Limit of the MLP (5/7)

- From [LBN+17], one has the preactivations of each layer $h \in [L]$ have their each coordinates tending to Gaussian process of covariance $\Sigma^{(h-1)} : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$:

$$\begin{aligned}
\Sigma^{(0)}(\boldsymbol{x}, \boldsymbol{x}') &= \boldsymbol{x}^\top \boldsymbol{x}', \\
\boldsymbol{\Lambda}^{(h)}(\boldsymbol{x}, \boldsymbol{x}') &= \begin{pmatrix} \Sigma^{(h-1)}(\boldsymbol{x}, \boldsymbol{x}) & \Sigma^{(h-1)}(\boldsymbol{x}, \boldsymbol{x}') \\ \Sigma^{(h-1)}(\boldsymbol{x}', \boldsymbol{x}) & \Sigma^{(h-1)}(\boldsymbol{x}', \boldsymbol{x}') \end{pmatrix} \in \mathbb{R}^{2\times 2}, \qquad (5) \\
\Sigma^{(h)}(\boldsymbol{x}, \boldsymbol{x}') &= c_\sigma \mathbb{E}_{(u,v)\sim\mathcal{N}\left(\mathbf{0}, \boldsymbol{\Lambda}^{(h)}\right)} \left[\sigma(u)\, \sigma(v)\right],
\end{aligned}$$

Infinite Width Limit of the MLP (6/7)

- The intuition is that

$$\left[\boldsymbol{f}^{(h+1)}(\boldsymbol{x})\right]_i = \sum_{j=1}^{d_h} \left[\boldsymbol{W}^{(h+1)}\right]_{i,j} \left[\boldsymbol{g}^{(h)}(\boldsymbol{x})\right]_j \tag{6}$$

  is a centered Gaussian process conditioned on $\boldsymbol{f}^{(h)}$ $(\forall i \in [d_{h+1}])$, with covariance

$$\mathbb{E}\left[\left[\boldsymbol{f}^{(h+1)}(\boldsymbol{x})\right]_i \cdot \left[\boldsymbol{f}^{(h+1)}(\boldsymbol{x}')\right]_i \middle| \boldsymbol{f}^{(h)}\right]. \tag{7}$$

Infinite Width Limit of the MLP (7/7)

- We further have

if i = k,
Exp[z ^ 2] = 1;
otherwise
Exp[z_1 z_2] = 0;
z_1, z_2, z ~iid N(0,1)

$$
\begin{aligned}
&\mathbb{E}\left[\left[\boldsymbol{f}^{(h+1)}(\boldsymbol{x})\right]_i \cdot \left[\boldsymbol{f}^{(h+1)}(\boldsymbol{x}')\right]_i \middle| \boldsymbol{f}^{(h)}\right] \\
=&\mathbb{E}\left[\sum_{j,k\in[d_h]} \boldsymbol{W}_{ij}^{(h+1)}\boldsymbol{W}_{ik}^{(h+1)}\boldsymbol{g}^{(h)}(\boldsymbol{x})_j\boldsymbol{g}^{(h)}(\boldsymbol{x}')_k \middle| \boldsymbol{f}^{(h)}\right] \\
=&\sum_{j,k\in[d_h]} \delta_{jk}\boldsymbol{g}^{(h)}(\boldsymbol{x})_j\boldsymbol{g}^{(h)}(\boldsymbol{x}')_k = \langle \boldsymbol{g}^{(h)}(\boldsymbol{x}), \boldsymbol{g}^{(h)}(\boldsymbol{x}')\rangle \\
=&\frac{c_\sigma}{d_h}\langle \sigma\left(f^{(h)}(\boldsymbol{x})\right), \sigma\left(f^{(h)}(\boldsymbol{x}')\right)\rangle \\
=&\frac{c_\sigma}{d_h}\sum_{j\in[d_h]} \sigma\left(f^{(h)}(\boldsymbol{x})\right)_j \cdot \sigma\left(f^{(h)}(\boldsymbol{x}')\right)_j \\
\xrightarrow[\text{a.s.}]{\text{LLN}}& c_\sigma \mathbb{E}_{(u,v)\sim\mathcal{N}(\boldsymbol{0},\boldsymbol{\Lambda}^{(h)})}\left[\sigma(u)\sigma(v)\right] \equiv \Sigma^{(h)}(\boldsymbol{x},\boldsymbol{x}'),
\end{aligned}
\tag{8}
$$

as $d_h \to \infty$ given that each $\boldsymbol{f}_j^{(h)}$ is a centered Gaussian process with covariance $\Sigma^{(h-1)}$.

Derivation of the NTK (1/7)

- To obtain the NTK, one computes the value that

$$\left\langle \frac{\partial f(\boldsymbol{\theta}, \boldsymbol{x})}{\partial \boldsymbol{\theta}}, \frac{\partial f(\boldsymbol{\theta}, \boldsymbol{x}')}{\partial \boldsymbol{\theta}} \right\rangle \tag{9}$$

  converges to at random initialization in the infinite width limit.

Derivation of the NTK (2/7)

- We can write the partial derivative with respect to a particular weight matrix $\boldsymbol{W}^{(h)}$ in a compact form:

$$\frac{\partial f(\boldsymbol{\theta}, \boldsymbol{x})}{\partial \boldsymbol{W}^{(h)}} = \mathbf{b}^{(h)}(\boldsymbol{x}) \cdot \left(\boldsymbol{g}^{(h-1)}(\boldsymbol{x})\right)^{\top}, \qquad h = 1, 2, \ldots, L+1, \quad (10)$$

(backprop)

where

$$\mathbf{b}^{(h)}(\boldsymbol{x}) = \begin{cases} 1 \in \mathbb{R}, & \text{for } h = L+1, \\ \sqrt{\frac{c_\sigma}{d_h}} \boldsymbol{D}^{(h)}(\boldsymbol{x}) \left(\boldsymbol{W}^{(h+1)}\right)^{\top} \mathbf{b}^{(h+1)}(\boldsymbol{x}) \in \mathbb{R}^{d_h}, & \text{for } h \in [L]. \end{cases} \quad (11)$$

$$\boldsymbol{D}^{(h)}(\boldsymbol{x}) = \operatorname{diag}\left(\dot{\sigma}\left(\boldsymbol{f}^{(h)}(\boldsymbol{x})\right)\right) \in \mathbb{R}^{d_h \times d_h}, \qquad h \in [L]. \quad (12)$$

Derivation of the NTK (3/7)

- Then, for any $h \in [L+1]$,

$$\begin{aligned}\left\langle \frac{\partial f(\boldsymbol{\theta}, \boldsymbol{x})}{\partial \boldsymbol{W}^{(h)}}, \frac{\partial f(\boldsymbol{\theta}, \boldsymbol{x}')}{\partial \boldsymbol{W}^{(h)}} \right\rangle &= \left\langle \mathbf{b}^{(h)}(\boldsymbol{x}) \cdot \left(\boldsymbol{g}^{(h-1)}(\boldsymbol{x})\right)^{\top}, \right. \\ &\qquad \left. \mathbf{b}^{(h)}(\boldsymbol{x}') \cdot \left(\boldsymbol{g}^{(h-1)}(\boldsymbol{x}')\right)^{\top} \right\rangle \\ &= \left\langle \boldsymbol{g}^{(h-1)}(\boldsymbol{x}), \boldsymbol{g}^{(h-1)}(\boldsymbol{x}') \right\rangle \\ &\qquad \times \left\langle \mathbf{b}^{(h)}(\boldsymbol{x}), \mathbf{b}^{(h)}(\boldsymbol{x}') \right\rangle .\end{aligned} \tag{13}$$

- Note that we have established in Equation (7) that

$$\left\langle \boldsymbol{g}^{(h-1)}(\boldsymbol{x}), \boldsymbol{g}^{(h-1)}(\boldsymbol{x}') \right\rangle \to \Sigma^{(h-1)}\left(\boldsymbol{x}, \boldsymbol{x}'\right). \tag{14}$$

Derivation of the NTK (4/7)

- For the other factor $\left\langle \mathbf{b}^{(h)}(\boldsymbol{x}), \mathbf{b}^{(h)}(\boldsymbol{x}') \right\rangle$, by definition (11),

$$\begin{aligned} &\left\langle \mathbf{b}^{(h)}(\boldsymbol{x}), \mathbf{b}^{(h)}(\boldsymbol{x}') \right\rangle \\ &\qquad = \left\langle \sqrt{\frac{c_\sigma}{d_h}} \boldsymbol{D}^{(h)}(\boldsymbol{x}) \left(\boldsymbol{W}^{(h+1)}\right)^\top \mathbf{b}^{h+1}(\boldsymbol{x}), \right. \\ &\qquad\qquad \left. \sqrt{\frac{c_\sigma}{d_h}} \boldsymbol{D}^{(h)}(\boldsymbol{x}') \left(\boldsymbol{W}^{(h+1)}\right)^\top \mathbf{b}^{h+1}(\boldsymbol{x}') \right\rangle . \end{aligned} \tag{15}$$

- In Eq. (11) of [ADH+19], there is another factor $\frac{c_\sigma^{L-h}}{d_{h+1} \cdots d_L}$ preceeding the RHS of Eq. (15), which is likely to be a typo.

Derivation of the NTK (5/7)

- Although $\boldsymbol{W}^{(h+1)}$ and $\mathbf{b}^{h+1}(\boldsymbol{x})$ are dependent, the Gaussian initialization of $\boldsymbol{W}^{(h+1)}$ allows us to replace $\boldsymbol{W}^{(h+1)}$ with a fresh new sample $\widetilde{\boldsymbol{W}}^{(h+1)}$ without changing its limit.

$$
\begin{aligned}
&\left\langle \sqrt{\frac{c_\sigma}{d_h}} \boldsymbol{D}^{(h)}(\boldsymbol{x}) \left(\boldsymbol{W}^{(h+1)}\right)^\top \mathbf{b}^{h+1}(\boldsymbol{x}), \sqrt{\frac{c_\sigma}{d_h}} \boldsymbol{D}^{(h)}(\boldsymbol{x}') \left(\boldsymbol{W}^{(h+1)}\right)^\top \mathbf{b}_{h+1}(\boldsymbol{x}') \right\rangle \\
\approx &\left\langle \sqrt{\frac{c_\sigma}{d_h}} \boldsymbol{D}^{(h)}(\boldsymbol{x}) \left(\widetilde{\boldsymbol{W}}^{(h+1)}\right)^\top \mathbf{b}_{h+1}(\boldsymbol{x}), \sqrt{\frac{c_\sigma}{d_h}} \boldsymbol{D}^{(h)}(\boldsymbol{x}') \left(\widetilde{\boldsymbol{W}}^{(h+1)}\right)^\top \mathbf{b}^{h+1}(\boldsymbol{x}') \right\rangle \\
\to &\frac{c_\sigma}{d_h} \operatorname{tr}\left(\boldsymbol{D}^{(h)}(\boldsymbol{x}) \boldsymbol{D}^{(h)}(\boldsymbol{x}')\right) \left\langle \mathbf{b}^{(h+1)}(\boldsymbol{x}), \mathbf{b}^{(h+1)}(\boldsymbol{x}') \right\rangle \\
\to &\dot{\Sigma}^{(h)}\left(\boldsymbol{x}, \boldsymbol{x}'\right) \left\langle \mathbf{b}^{(h+1)}(\boldsymbol{x}), \mathbf{b}^{(h+1)}(\boldsymbol{x}') \right\rangle,
\end{aligned}
\tag{16}
$$

where

$$
\dot{\Sigma}^{(h)}(\boldsymbol{x}, \boldsymbol{x}') = c_\sigma \mathbb{E}_{(u,v)\sim\mathcal{N}\left(\mathbf{0}, \boldsymbol{\Lambda}^{(h)}\right)} \left[\dot{\sigma}(u)\dot{\sigma}(v)\right]. \tag{17}
$$

Derivation of the NTK (6/7)

- Applying this approximation inductively in Equation (15), we get

$$\left\langle \mathbf{b}^{(h)}(\boldsymbol{x}), \mathbf{b}^{(h)}(\boldsymbol{x}') \right\rangle \to \prod_{h'=h}^{L} \dot{\Sigma}^{(h')}(\boldsymbol{x}, \boldsymbol{x}'). \tag{18}$$

- Finally, since

$$\left\langle \frac{\partial f(\boldsymbol{\theta}, \boldsymbol{x})}{\partial \boldsymbol{\theta}}, \frac{\partial f(\boldsymbol{\theta}, \boldsymbol{x}')}{\partial \boldsymbol{\theta}} \right\rangle = \sum_{h=1}^{L+1} \left\langle \frac{\partial f(\boldsymbol{\theta}, \boldsymbol{x})}{\partial \boldsymbol{W}^{(h)}}, \frac{\partial f(\boldsymbol{\theta}, \boldsymbol{x}')}{\partial \boldsymbol{W}^{(h)}} \right\rangle, \tag{19}$$

Derivation of the NTK (7/7)

one has

$$\Theta^{(L)}(\boldsymbol{x}, \boldsymbol{x}') = \sum_{h=1}^{L+1} \left(\Sigma^{(h-1)}(\boldsymbol{x}, \boldsymbol{x}') \cdot \prod_{h'=h}^{L+1} \dot{\Sigma}^{(h')}(\boldsymbol{x}, \boldsymbol{x}') \right), \tag{20}$$

NTK

where one writes $\dot{\Sigma}^{(L+1)}(\boldsymbol{x}, \boldsymbol{x}') = 1$.

Contents

1. Neural Tangent Kernels (NTK)
   - Contruction of the NTK due to [ADH+19]
   - Convergence to the NTK at Initialization [ADH+19]
   - Equivalence between NN and NTK
2. Graph Neural Nets (GNN)
3. Graph Neural Tangent Kernel (GNTK)
   - Overview
   - Formulae
   - Theoretical Analyses
     - Proof of Theorem 4.2 [DHP+19]
     - Proof of Theorem 4.3 [DHP+19]

Convergence of NTK of MLP (1/2)

Theorem 3.1 [ADH+19]

Theorem 3.1 (Convergence to the NTK at initialization)

Fix $\epsilon > 0$ *and* $\delta \in (0,1)$. *Suppose* $\sigma(z) = \max(0,z)$ $(z \in \mathbb{R})$, $\min_{h\in[L]} d_h \geq \text{poly}(L, 1/\epsilon) \cdot \log(L/\delta)$, *and* $[\boldsymbol{W}^{(h)}]_{i,j} \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$ $\forall h \in [L+1], i \in [d_h], j \in [d_{h-1}]$. *Then for any inputs* $\boldsymbol{x}, \boldsymbol{x}' \in \mathbb{R}^{d_0}$ *such that* $\|\boldsymbol{x}\| \leq 1, \|\boldsymbol{x}'\| \leq 1$, *with probability at least* $1-\delta$ *we have:*

$$\left| \left\langle \frac{\partial f(\boldsymbol{\theta}, \boldsymbol{x})}{\partial \boldsymbol{\theta}}, \frac{\partial f(\boldsymbol{\theta}, \boldsymbol{x}')}{\partial \boldsymbol{\theta}} \right\rangle - \Theta^{(L)}(\boldsymbol{x}, \boldsymbol{x}') \right| \leq \epsilon. \tag{21}$$

one realization of finite width

theoretical limit as $d_h \to \infty$
- dependent on data
- indep of params

Convergence of NTK of MLP (2/2)

Theorem 3.1 [ADH+19]

- Compared with [JGH18] and [Yan19], Theorem 3.1 of [ADH+19] is non-asymptotic.
- [JGH18] requires $d_0, \ldots, d_L \to \infty$ sequentially; [Yan19] requires $d_0, \ldots, d_L \to \infty$ at the same rate. But [ADH+19] requires $\min_h d_h \to \infty$.

Notations (1/8)

Definition B.1 (k-homogeneous function)

A function $f : \mathbb{R} \to \mathbb{R}$ is said to be k-homogeneous, if $f(\lambda x) = \lambda^k f(x)$ for all $x \in \mathbb{R}, \lambda > 0$.

Definition B.2

Let $\mathcal{S}^+$ be the set of positive semi-definite kernels over $\mathbb{R}^d$, that is

$$\mathcal{S}^+ = \left\{ K : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R} \middle| \forall N \in \mathbb{N}, \boldsymbol{x}_1, \dots \boldsymbol{x}_N \in \mathbb{R}^d, c_1, \dots, c_N \in \mathbb{R}, \sum_{i=1}^N \sum_{j=1}^N c_i c_j K(x_i, x_j) \geq 0. \right\} \tag{22}$$

Notations (2/8)

Let $\sigma : \mathbb{R} \to \mathbb{R}$ be the activation function, and $\mathcal{T}_\sigma : \mathcal{S}^+ \to \mathcal{S}^+$ be the operator induced by σ, i.e.,

$$\forall \boldsymbol{x}, \boldsymbol{x}' \in \mathbb{R}^d, \quad \mathcal{T}_\sigma(K)(\boldsymbol{x}, \boldsymbol{x}') = c_\sigma \mathbb{E}_{(u,v)\sim\mathcal{N}\left(\boldsymbol{0}, \boldsymbol{K}|_{\boldsymbol{x},\boldsymbol{x}'}\right)} \left[\sigma(u)\,\sigma(v)\right], \tag{23}$$

where $\boldsymbol{K}|_{\boldsymbol{x},\boldsymbol{x}'} \in \mathbb{R}^{2\times 2}$, $\boldsymbol{K}|_{\boldsymbol{x},\boldsymbol{x}'} = \begin{bmatrix} K(\boldsymbol{x},\boldsymbol{x}) & K(\boldsymbol{x},\boldsymbol{x}') \\ K(\boldsymbol{x}',\boldsymbol{x}) & K(\boldsymbol{x}',\boldsymbol{x}') \end{bmatrix}$.

Define

$$t_\sigma(\boldsymbol{\Sigma}) = c_\sigma \mathbb{E}_{(u,v)\sim\mathcal{N}(\boldsymbol{0},\boldsymbol{\Sigma})} \left[\sigma(u)\,\sigma(v)\right], \tag{24}$$

$$\hat{t}_\sigma(\rho)\hat{t}_\sigma(\rho) = c_\sigma \mathbb{E}_{(u,v)\sim\boldsymbol{\Sigma}'} \left[\sigma(u)\,\sigma(v)\right], \text{ with } \boldsymbol{\Sigma}' = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix} \tag{25}$$

Notations (3/8)

When σ is k-homogeneous function, we have

$$t_\sigma(\mathbf{\Sigma}) = c_\sigma \left(\Sigma_{11}\Sigma_{22}\right)^{\frac{k}{2}} \mathbb{E}_{(u,v)\sim\mathcal{N}(\mathbf{0},\mathbf{\Sigma}')} \left[\sigma\left(u\right)\sigma\left(v\right)\right], \tag{26}$$

with

$$\mathbf{\Sigma}' = \begin{bmatrix} 1 & \frac{\Sigma_{12}}{\sqrt{\Sigma_{11}\Sigma_{22}}} \\ \frac{\Sigma_{12}}{\sqrt{\Sigma_{11}\Sigma_{22}}} & 1 \end{bmatrix}. \tag{27}$$

Thus $t_\sigma(\mathbf{\Sigma})$ can be written as $c_\sigma \left(\Sigma_{11}\Sigma_{22}\right)^{\frac{k}{2}} \hat{t}(\frac{\Sigma_{12}}{\sqrt{\Sigma_{11}\Sigma_{22}}})$,

Notations (4/8)

Fact B.1 (Some facts about $\sigma(z) = \max(0, z)$ and $\mathcal{T}_\sigma$)

1. *For all activation function* σ, $t_\sigma\left(\begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}\right) = 1$.
2. *For all* ~~1~~ k*-homogeneous activation* σ, $\hat{t}_\sigma(1) = 1$ *and* $t_\sigma\left(\begin{bmatrix} a & a \\ a & a \end{bmatrix}\right) = a^k$.
3. *For* $\sigma(z) = \max(0, z)$, $\hat{t}_\sigma(\rho) = \frac{\sqrt{1-\rho^2} + \rho \arcsin \rho}{\pi} + \frac{\cancel{\pi}}{2}$,1 $\hat{t}_{\dot{\sigma}}(\rho) = \frac{1}{2} + \frac{\arcsin \rho}{\pi}$ *and* $c_\sigma = c_{\dot{\sigma}} = 2$. (cf. 1905)

Notations (5/8)

Recall the definition in Equation (5) and (17), we have

$$
\begin{aligned}
&\Sigma^{(0)}(\boldsymbol{x}, \boldsymbol{x}') = \boldsymbol{x}^\top \boldsymbol{x}', \\
&\boldsymbol{\Lambda}^{(h)}(\boldsymbol{x}, \boldsymbol{x}') = \boldsymbol{\Sigma}^{(h-1)}\Big|_{\boldsymbol{x},\boldsymbol{x}'} = \begin{pmatrix} \Sigma^{(h-1)}(\boldsymbol{x}, \boldsymbol{x}) & \Sigma^{(h-1)}(\boldsymbol{x}, \boldsymbol{x}') \\ \Sigma^{(h-1)}(\boldsymbol{x}', \boldsymbol{x}) & \Sigma^{(h-1)}(\boldsymbol{x}', \boldsymbol{x}') \end{pmatrix} \in \mathbb{R}^{2\times 2}, \\
&\Sigma^{(h)}(\boldsymbol{x}, \boldsymbol{x}') = c_\sigma \mathbb{E}_{(u,v)\sim\mathcal{N}\left(\mathbf{0}, \boldsymbol{\Lambda}^{(h)}\right)} \left[\sigma(u)\,\sigma(v)\right], \\
&\dot{\Sigma}^{(h)}(\boldsymbol{x}, \boldsymbol{x}') = c_\sigma \mathbb{E}_{(u,v)\sim\mathcal{N}\left(\mathbf{0}, \boldsymbol{\Lambda}^{(h)}\right)} \left[\dot{\sigma}(u)\dot{\sigma}(v)\right]
\end{aligned}
\tag{28}
$$

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

For $\sigma(z) = \max(z, 0)$, we have

$$
\Sigma^{(h)}(\boldsymbol{x}, \boldsymbol{x}) = \|\boldsymbol{x}\|^2, \quad \forall 0 \le h \le L. \tag{29}
$$

Notations (6/8)

Let $\boldsymbol{D} = \boldsymbol{D}(\boldsymbol{x}, \boldsymbol{x}') = \boldsymbol{D}^{(h)}(\boldsymbol{x})\boldsymbol{D}^{(h)}(\boldsymbol{x}')$ is a 0-1 diagonal matrix. We define the following events:

- **Control of the post activations.**
  - $\mathcal{A}^h(\boldsymbol{x}, \boldsymbol{x}', \epsilon_1) := \left\{ \left| \boldsymbol{g}^{(h)}(x^{(0)})^\top \boldsymbol{g}^{(h)}(\boldsymbol{x}) - \boldsymbol{\Sigma}^{(h)}(\boldsymbol{x}^{(0)}, \boldsymbol{x}) \right| \le \epsilon_1 \right\}$, $\forall 0 \le h \le L$
  - $\overline{\mathcal{A}}^h(\boldsymbol{x}, \boldsymbol{x}', \epsilon_1) = \mathcal{A}^h(\boldsymbol{x}, \boldsymbol{x}, \epsilon_1) \cap \mathcal{A}^h(\boldsymbol{x}, \boldsymbol{x}', \epsilon_1) \cap \mathcal{A}^h(\boldsymbol{x}', \boldsymbol{x}', \epsilon_1)$;
  - $\overline{\mathcal{A}}(\boldsymbol{x}, \boldsymbol{x}', \epsilon_1) = \bigcup_{h=0}^{L} \overline{\mathcal{A}}^h(\epsilon_1)$.
- **Control of the backprop.**
  - $\mathcal{B}^h(\boldsymbol{x}, \boldsymbol{x}', \epsilon_2) = \left\{ \left| \left\langle \mathbf{b}^{(h)}(\boldsymbol{x}), \mathbf{b}^{(h)}(\boldsymbol{x}') \right\rangle - \prod_{h=h}^{L} \dot{\Sigma}^{(h)}(\boldsymbol{x}, \boldsymbol{x}') \right| < \epsilon_2 \right\}$;
  - $\overline{\mathcal{B}}^h(\boldsymbol{x}, \boldsymbol{x}', \epsilon_2) = \mathcal{B}^h(\boldsymbol{x}, \boldsymbol{x}, \epsilon_2) \cap \mathcal{B}^h(\boldsymbol{x}, \boldsymbol{x}', \epsilon_2) \cap \mathcal{B}^h(\boldsymbol{x}', \boldsymbol{x}', \epsilon_2)$;
  - $\overline{\mathcal{B}}(\boldsymbol{x}, \boldsymbol{x}', \epsilon_2) = \bigcup_{h=1}^{L+1} \overline{\mathcal{B}}^h(\epsilon_2)$;

Notations (7/8)

- **Control of the output.**

$$\overline{\mathcal{C}}\left(\boldsymbol{x}, \boldsymbol{x}', \epsilon_3\right) = \left\{|f(\boldsymbol{\theta}, \boldsymbol{x})| \leq \epsilon_3, |f(\boldsymbol{\theta}, \boldsymbol{x}')| \leq \epsilon_3\right\} \tag{30}$$

- **Control of the derivative.**
  - $\mathcal{D}^h\left(\boldsymbol{x}, \boldsymbol{x}', \epsilon_4\right) = \left\{|2\frac{\operatorname{tr}\left(\boldsymbol{D}(\boldsymbol{x},\boldsymbol{x}')\right)}{d_h} - \dot{\Sigma}^{(h)}(\boldsymbol{x}, \boldsymbol{x}')| < \epsilon_4\right\}$;
  - $\overline{\mathcal{D}}^h\left(\boldsymbol{x}, \boldsymbol{x}', \epsilon_4\right) = \mathcal{D}^h\left(\boldsymbol{x}, \boldsymbol{x}, \epsilon_4\right) \cap \mathcal{D}^h\left(\boldsymbol{x}, \boldsymbol{x}', \epsilon_4\right) \cap \mathcal{D}^h\left(\boldsymbol{x}', \boldsymbol{x}', \epsilon_1\right)$;
  - $\overline{\mathcal{D}}\left(\boldsymbol{x}, \boldsymbol{x}', \epsilon_4\right) = \bigcup_{h=1}^{L+1} \overline{\mathcal{D}}^h\left(\epsilon_4\right)$.

For simplicity, we will omit $\boldsymbol{x}, \boldsymbol{x}'$ when there's no ambiguity. For events $\mathcal{A}, \mathcal{B}$, we define the event $\mathcal{A} \Rightarrow \mathcal{B}$ as $\neg\mathcal{A} \wedge \mathcal{B}$.

Notations (8/8)

Recall

$$\Theta^{(L)}(\boldsymbol{x}, \boldsymbol{x}') = \sum_{h=1}^{L+1} \left(\Sigma^{(h-1)}(\boldsymbol{x}, \boldsymbol{x}') \cdot \prod_{h'=h}^{L+1} \dot{\Sigma}^{(h')}(\boldsymbol{x}, \boldsymbol{x}') \right), \tag{31}$$

where one writes $\dot{\Sigma}^{(L+1)}(\boldsymbol{x}, \boldsymbol{x}') = 1$; and

Theorem 3.1 (Convergence to the NTK at initialization)

Fix $\epsilon > 0$ and $\delta \in (0, 1)$. Suppose $\sigma(z) = \max(0, z)$ $(z \in \mathbb{R})$, $\min_{h \in [L]} d_h \geq \text{poly}(L, 1/\epsilon) \cdot \log(L/\delta)$, and $[\boldsymbol{W}^{(h)}]_{i,j} \overset{\text{iid}}{\sim} \mathcal{N}(0, 1)$ $\forall h \in [L+1], i \in [d_h], j \in [d_{h-1}]$. Then for any inputs $\boldsymbol{x}, \boldsymbol{x}' \in \mathbb{R}^{d_0}$ such that $\|\boldsymbol{x}\| \leq 1, \|\boldsymbol{x}'\| \leq 1$, with probability at least $1 - \delta$ we have:

$$\left| \left\langle \frac{\partial f(\boldsymbol{\theta}, \boldsymbol{x})}{\partial \boldsymbol{\theta}}, \frac{\partial f(\boldsymbol{\theta}, \boldsymbol{x}')}{\partial \boldsymbol{\theta}} \right\rangle - \Theta^{(L)}(\boldsymbol{x}, \boldsymbol{x}') \right| \leq \epsilon. \tag{32}$$

Proof Outline of Theorem 3.1 [ADH+19] (1/5)

Theorem B.1 (Corollary 16 in [DFS16])

Let $\sigma(z) = \max(0, z), z \in \mathbb{R}$ *and* $[\boldsymbol{W}^{(h)}]_{ij} \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$, $\forall h \in [L], i \in [d^{h+1}], j \in [d^h]$, *there exist constants* c_1,c_2, *such that if* $c_1 \frac{L^2 \log(\frac{8L}{\delta})}{\epsilon^2} \leq \min_{1 \leq h \leq L} d_h$ *and* $\epsilon \leq \min(c_2, \frac{1}{L})$, *then for any fixed* $\boldsymbol{x}, \boldsymbol{x}' \in \mathbb{R}^{d_0}$, $\|\boldsymbol{x}\|, \|\boldsymbol{x}'\| \leq 1$, *we have w.p.* $\geq 1 - \delta$, $\forall 0 \leq h \leq L$, $\forall (\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}) \in \{(\boldsymbol{x}, \boldsymbol{x}), (\boldsymbol{x}, \boldsymbol{x}'), (\boldsymbol{x}', \boldsymbol{x}')\}$,

$$\left| \boldsymbol{g}^{(h)}(\boldsymbol{x}^{(2)})^\top \boldsymbol{g}^{(h)}(\boldsymbol{x}^{(1)}) - \boldsymbol{\Sigma}^{(h)}(\boldsymbol{x}^{(2)}, \boldsymbol{x}^{(1)}) \right| \leq \epsilon. \tag{33}$$

In other words, if $\min_{h \in [L]} d_h \geq c_1 \frac{L^2 \log(\frac{L}{\delta_1})}{\epsilon_1^2}$, $\epsilon_1 \leq \min(c_2, \frac{1}{L})$, *then for fixed* $\boldsymbol{x}, \boldsymbol{x}'$,

$$\mathbb{P}\left[\textit{there is at least one layer with the above ineq held}\right] \geq 1 - \delta_1.$$

Proof Outline of Theorem 3.1 [ADH+19] (2/5)

Theorem B.2

Let $\sigma(z) = \max(0, z), z \in \mathbb{R}$, *if* $[\boldsymbol{W}^{(h)}]_{ij} \overset{i.i.d.}{\sim} N(0,1)$,
$\forall h \in [L+1], i \in [d^{h+1}], j \in [d^h]$, *there exist constants* c_1,c_2, *such that if*
$\min_{h\in[L]} d_h \geq c_1 \frac{L^2 \log(\frac{L}{\delta})}{\epsilon^4}, \epsilon \leq \frac{c_2}{L}$, *then for any fixed* $\boldsymbol{x}, \boldsymbol{x}' \in \mathbb{R}^{d_0}$,
$\|\boldsymbol{x}\|, \|\boldsymbol{x}'\| \leq 1$,*we have w.p.* $1-\delta$, $\forall 0 \leq h \leq L$,
$\forall (\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}) \in \{(\boldsymbol{x}, \boldsymbol{x}), (\boldsymbol{x}, \boldsymbol{x}'), (\boldsymbol{x}', \boldsymbol{x}')\}$,

$$\left|\boldsymbol{g}^{(h)}(\boldsymbol{x}^{(2)})^\top \boldsymbol{g}^{(h)}(\boldsymbol{x}^{(1)}) - \boldsymbol{\Sigma}^{(h)}(\boldsymbol{x}^{(2)}, \boldsymbol{x}^{(1)})\right| \leq \frac{\epsilon^2}{2}, \tag{34}$$

and

$$\left|\left\langle \mathbf{b}^{(h)}(\boldsymbol{x}^{(1)}), \mathbf{b}^{(h)}(\boldsymbol{x}^{(2)})\right\rangle - \prod_{h'=h}^{L} \dot{\Sigma}^{(h')}(\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)})\right| < 3L\epsilon. \tag{35}$$

Proof Outline of Theorem 3.1 [ADH+19] (3/5)

Theorem B.2

In other words, if $\min_{h\in[L]} d_h \geq c_1 \frac{L^2 \log(\frac{L}{\delta_1})}{\epsilon_1^2}$, $\epsilon_1 \leq \frac{c_2}{L}$, *then for fixed* $\boldsymbol{x}, \boldsymbol{x}'$,

$$\mathbb{P}\left[\overline{\mathcal{A}}\left(\frac{\epsilon^2}{8}\right) \bigwedge \overline{\mathcal{B}}\left(3L\epsilon\right)\right] \geq 1-\delta \tag{36}$$

Proof Outline of Theorem 3.1 [ADH+19] (4/5)

Proof of Theorem 3.1.

Recall that $\Theta^{(L)}(\boldsymbol{x}, \boldsymbol{x}') = \sum_{h=1}^{L+1} \left(\Sigma^{(h-1)}(\boldsymbol{x}, \boldsymbol{x}') \cdot \prod_{h'=h}^{L+1} \dot{\Sigma}^{(h')}(\boldsymbol{x}, \boldsymbol{x}') \right)$, thus it suffices to show that w.p. $1-\epsilon$, for every $0 \le h \le L$, it holds that

$$\left| \left\langle \frac{\partial f(\boldsymbol{\theta}, \boldsymbol{x})}{\partial \boldsymbol{W}^{(h)}}, \frac{\partial f(\boldsymbol{\theta}, \boldsymbol{x}')}{\partial \boldsymbol{W}^{(h)}} \right\rangle - \Sigma^{(h-1)}(\boldsymbol{x}, \boldsymbol{x}') \cdot \prod_{h'=h}^{L+1} \dot{\Sigma}^{(h')}(\boldsymbol{x}, \boldsymbol{x}') \right| \le \frac{\epsilon}{L+1}, \tag{37}$$

which is a direct consequence of Theorem B.2 □

Proof Outline of Theorem 3.1 [ADH+19] (5/5)

Roadmap.

- Theorem 3.1 $\Leftarrow$ Theorem B.2;
- Theorem B.2 $\Leftarrow$ Theorem B.1, Lemmas B.4, B.6, B.7 (induction);
- Lemma B.7 $\Leftarrow$ Claims B.1, B.2, B.3;
- Claim B.2 $\Leftarrow$ Claim B.1, Lemmas B.3, B.10;
- Claim B.3 $\Leftarrow$ Lemma B.3;
- Lemma B.6 $\Leftarrow$ Lemma B.5 $\Leftarrow$ Lemmas B.8, B.9

Some Lemmas (1/5)

Lemma B.1 (Uniform Continuity of arcsin z)

1. *For any* $-\frac{\pi}{2} \leq y' \leq y \leq \frac{\pi}{2}$, $\sin y - \sin y' \geq 2\sin^2 \frac{y-y'}{2}$.
2. $\sin y \geq \frac{2y}{\pi}, \forall y \in [0, \frac{\pi}{2}]$.
3. arcsin *is uniform continuous: for every* $\epsilon \in \mathbb{R}^+$, $|z - z'| < \frac{2\epsilon^2}{\pi^2} \Rightarrow |\arcsin z - \arcsin z'| < \epsilon$.
4. *For* $\sigma(z) = \max(0, z)$, $\hat{t}_{\dot{\sigma}}$ *is uniform continuous: for every* $\epsilon \in \mathbb{R}^+$, $|z - z'| < 2\epsilon^2 \Rightarrow |\hat{t}_{\dot{\sigma}}(z) - \hat{t}_{\dot{\sigma}}(z')| < \epsilon$.

Some Lemmas (2/5)

Proof of Lemma B.1.

(1). From $-\frac{\pi}{2} \le y' \le y' \le \frac{\pi}{2}$ we know $\frac{-\pi}{2} + \frac{y-y'}{2} \le \frac{y+y'}{2} \le \frac{\pi}{2} - \frac{y-y'}{2}$, which implies that $\cos(\frac{y+y'}{2}) \ge \sin(\frac{y-y'}{2})$. Thus,

$$\sin y \sin y' = 2\cos\frac{y+y'}{2}\sin\frac{y-y'}{2} \ge 2\sin^2\frac{y-y'}{2}.$$

(2). Note that $\left(\frac{\sin y}{y}\right)' = \frac{y\cos y - \sin y}{y^2} = \frac{\cos y}{y^2}(y - \tan y) < 0$, $\frac{\sin y}{y}$ is decreasing on $[0, \frac{\pi}{2}]$. Thus $\frac{\sin y}{y} \ge \frac{1}{\frac{\pi}{2}} = \frac{2}{\pi}, \forall y \in [0, \frac{\pi}{2}]$.

(3). Let $y, y' \in [-\frac{\pi}{2}, \frac{\pi}{2}]$, such that $\sin y = z, \sin y' = z'$. W.l.o.g., we assume $y' < y$, $z' < z$. Combing (1) and (2), we have

$z - z' = \sin y - \sin y' \ge 2\sin^2\frac{y-y'}{2} \ge \frac{2(y-y')^2}{\pi^2}$. Thus

$z - z' \le \frac{2\epsilon^2}{\pi^2} \Longrightarrow \arcsin z - \arcsin z' = y - y' \le \epsilon.$ □

Some Lemmas (3/5)

Lemma B.2

$$\mathbb{P}\left[\mathcal{A} \Rightarrow \mathcal{B}\right] \geq \mathbb{P}\left[\mathcal{B} \mid \mathcal{A}\right].$$

Proof of Lemma B.2.

$$\mathbb{P}\left[\mathcal{A} \Rightarrow \mathcal{B}\right] = \mathbb{P}\left[\neg\mathcal{A} \wedge \mathcal{B}\right] = 1 - \mathbb{P}\left[\mathcal{A} \vee \neg\mathcal{B}\right] = 1 - \mathbb{P}\left[\neg\mathcal{B} \mid \mathcal{A}\right]\mathbb{P}\left[\mathcal{A}\right] \geq 1 - \mathbb{P}\left[\neg\mathcal{B} \mid \mathcal{A}\right] = \mathbb{P}\left[\mathcal{B} \mid \mathcal{A}\right].$$

□

For matrix $\boldsymbol{A}$, define the projection matrix for the column space of $\boldsymbol{A}$, $\boldsymbol{\Pi}_{\boldsymbol{A}} := \boldsymbol{A}\boldsymbol{A}^{\dagger}$ and the orthogonal projection matrix $\boldsymbol{\Pi}_{\boldsymbol{A}}^{\perp} = I - \boldsymbol{A}\boldsymbol{A}^{\dagger}$. For two random variables X and Y, $X \overset{\mathrm{d}}{=\!=}_{\mathcal{A}} Y$ means X is equal to Y in distribution conditioned on the σ-algebra generated by $\mathcal{A}$.

Some Lemmas (4/5)

Lemma B.3

Let $\boldsymbol{w} \sim \mathcal{N}(0, \boldsymbol{I}_d)$, $\boldsymbol{G} \in \mathbb{R}^{d \times k}$ *be some fixed matrix, and random vector* $\boldsymbol{F} = \boldsymbol{w}^\top \boldsymbol{G}$, *then conditioned on the value of* $\boldsymbol{G}$, $\boldsymbol{w}$ *remains gaussian in the null space of the row space of* $\boldsymbol{G}$. *Mathematically, it means*

$$\Pi_{\boldsymbol{G}}^{\perp} \boldsymbol{w} \overset{d}{=\!=}_{\boldsymbol{F}=\boldsymbol{w}^\top \boldsymbol{G}} \Pi_{\boldsymbol{G}}^{\perp} \widetilde{\boldsymbol{w}},$$

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

Some Lemmas (5/5)

Proof of Lemma B.3.

This lemma is straightforward when $\Pi^{\perp}_{\boldsymbol{G}}$ is a diagonal matrix.
In general, let $\boldsymbol{G} = \boldsymbol{U}\boldsymbol{G}'$, where $\boldsymbol{U} \in \mathbb{R}^{d\times d}$ is orthogonal and $\Pi^{\perp}_{G'}$ is diagonal. Now we have

$$\Pi^{\perp}_{\boldsymbol{G}}\boldsymbol{w} = \boldsymbol{U}\Pi^{\perp}_{\boldsymbol{G}'}\boldsymbol{U}^{\top}\boldsymbol{w} \overset{\mathrm{d}}{=\!=}_{\boldsymbol{F}=(\boldsymbol{U}^{\top}\boldsymbol{w})^{\top}\boldsymbol{G}'} \boldsymbol{U}\Pi^{\perp}_{\boldsymbol{G}'}\boldsymbol{U}^{\top}\widetilde{\boldsymbol{w}}, = \Pi^{\perp}_{\boldsymbol{G}}\widetilde{\boldsymbol{w}}$$

where we used the fact that if $\boldsymbol{w} \sim \mathcal{N}(0, \boldsymbol{I}_d)$, then for any orthogonal $\boldsymbol{U}$, $\boldsymbol{U}\boldsymbol{w} \sim \mathcal{N}(0, \boldsymbol{I}_d)$ twice. □

Proof of Theorem B.2 (1/47)

Theorem B.2

Let $\sigma(z) = \max(0, z), z \in \mathbb{R}$, *if* $[\boldsymbol{W}^{(h)}]_{ij} \overset{i.i.d.}{\sim} N(0,1)$,
$\forall h \in [L+1], i \in [d^{h+1}], j \in [d^h]$, *there exist constants* c_1,c_2, *such that if*
$\min_{h\in[L]} d_h \geq c_1 \frac{L^2 \log(\frac{L}{\delta})}{\epsilon^4}, \epsilon \leq \frac{c_2}{L}$, *then for any fixed* $\boldsymbol{x}, \boldsymbol{x}' \in \mathbb{R}^{d_0}$,
$\|\boldsymbol{x}\|, \|\boldsymbol{x}'\| \leq 1$,*we have w.p.* $1-\delta$, $\forall 0 \leq h \leq L$,
$\forall (\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}) \in \{(\boldsymbol{x}, \boldsymbol{x}), (\boldsymbol{x}, \boldsymbol{x}'), (\boldsymbol{x}', \boldsymbol{x}')\}$,

$$\left|\boldsymbol{g}^{(h)}(\boldsymbol{x}^{(2)})^\top \boldsymbol{g}^{(h)}(\boldsymbol{x}^{(1)}) - \boldsymbol{\Sigma}^{(h)}(\boldsymbol{x}^{(2)}, \boldsymbol{x}^{(1)})\right| \leq \frac{\epsilon^2}{2}, \tag{38}$$

and

$$\left|\left\langle \mathbf{b}^{(h)}(\boldsymbol{x}^{(1)}), \mathbf{b}^{(h)}(\boldsymbol{x}^{(2)})\right\rangle - \prod_{h'=h}^{L} \dot{\Sigma}^{(h')}(\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)})\right| < 3L\epsilon. \tag{39}$$

Proof of Theorem B.2 (2/47)

Theorem B.2

In other words, if $\min_{h\in[L]} d_h \geq c_1 \frac{L^2 \log(\frac{L}{\delta_1})}{\epsilon_1^2}$, $\epsilon_1 \leq \frac{c_2}{L}$, *then for fixed* $\boldsymbol{x}, \boldsymbol{x}'$,

$$\mathbb{P}\left[\overline{\mathcal{A}}\left(\frac{\epsilon^2}{8}\right) \bigwedge \overline{\mathcal{B}}\left(3L\epsilon\right)\right] \geq 1-\delta \tag{40}$$

Proof of Theorem B.2 (3/47)

Note that for $c_\sigma = 2$ for $\sigma(z) = \max(0, z)$, by definition of $\mathbf{b}^{(h)}$, we have

$$\begin{aligned}&\left\langle \mathbf{b}^{(h)}(\boldsymbol{x}), \mathbf{b}^{(h)}(\boldsymbol{x}') \right\rangle \\ =&\frac{2}{d_h} \mathbf{b}^{(h+1)}(\boldsymbol{x})^\top \boldsymbol{W}^{(h+1)} \boldsymbol{D}^{(h)}(\boldsymbol{x}) \boldsymbol{D}^{(h)}(\boldsymbol{x}') \left(\boldsymbol{W}^{(h+1)}\right)^\top \mathbf{b}^{(h+1)}(\boldsymbol{x}').\end{aligned} \tag{41}$$

Intuitively, when d_h is large, we can replace $\boldsymbol{W}^{(h+1)}$ by a fresh i.i.d copy $\widetilde{\boldsymbol{W}}$ with a small difference by $\tilde{O}(\frac{1}{\sqrt{d_h}})$ as below. Similar techniques are used in [Yan19].

Proof of Theorem B.2 (4/47)

$$
\begin{aligned}
\left\langle \mathbf{b}^{(h)}(\boldsymbol{x}), \mathbf{b}^{(h)}(\boldsymbol{x}') \right\rangle =& \frac{2}{d_h} \mathbf{b}^{(h+1)}(\boldsymbol{x})^\top \boldsymbol{W}^{(h+1)} \boldsymbol{D}^{(h)}(\boldsymbol{x}) \\
& \cdot \boldsymbol{D}^{(h)}(\boldsymbol{x}') \left(\boldsymbol{W}^{(h+1)}\right)^\top \mathbf{b}^{(h+1)}(\boldsymbol{x}') \\
\approx & \frac{2}{d_h} \mathbf{b}^{(h+1)}(\boldsymbol{x})^\top \widetilde{\boldsymbol{W}} \boldsymbol{D}^{(h)}(\boldsymbol{x}) \boldsymbol{D}^{(h)}(\boldsymbol{x}') \widetilde{\boldsymbol{W}}^\top \mathbf{b}^{(h+1)}(\boldsymbol{x}') \\
\approx & \operatorname{tr}\left(\frac{2}{d_h} \boldsymbol{D}^{(h)}(\boldsymbol{x}) \boldsymbol{D}^{(h)}(\boldsymbol{x}') \right) \mathbf{b}^{(h+1)}(\boldsymbol{x})^\top \mathbf{b}^{(h+1)}(\boldsymbol{x}') \\
\approx & \dot{\Sigma}^{(h)}(\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}) \prod_{h'=h+1}^{L} \dot{\Sigma}^{(h')}(\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)})
\end{aligned}
\tag{42}
$$

The proof is based on a careful control of the following events.

Proof of Theorem B.2 (5/47)

Lemma B.4

$$\mathbb{P}\left[\overline{\mathcal{A}}^{L}\left(\epsilon_1^2/2\right) \Longrightarrow \overline{\mathcal{C}}\left(2\sqrt{\log\frac{4}{\delta_3}}\right)\right] \geq 1-\delta_3, \quad \forall \epsilon_1 \in [0,1], \delta_3 \in (0,1). \tag{43}$$

Lemma B.5

$$\mathbb{P}\left[\overline{\mathcal{A}}^{h+1}\left(\epsilon_1^2/2\right) \Longrightarrow \overline{\mathcal{D}}^{h}\left(\epsilon_1 + \sqrt{\frac{2\log\frac{6}{\delta_4}}{d_h}}\right)\right] \geq 1-\delta_4, \tag{44}$$
$$\forall \epsilon_1 \in [0,1], \delta_4 \in (0,1).$$

Proof of Theorem B.2 (6/47)

Lemma B.6

$$\mathbb{P}\left[\overline{\mathcal{A}}\left(\epsilon_1^2/2\right) \Longrightarrow \overline{\mathcal{D}}\left(\epsilon_1 + \sqrt{\frac{2\log\frac{6L}{\delta_4}}{\min_h d_h}}\right)\right] \geq 1-\delta_4, \tag{45}$$

$$\forall \epsilon_1 \in [0,1], \delta_4 \in (0,1).$$

Proof of Lemma B.6.

Apply union bound on Lemma B.5. □

Proof of Theorem B.2 (7/47)

Lemma B.7

There exists constant $C, C' \in \mathbb{R}$, for any $\epsilon_2, \epsilon_3, \epsilon_4 \in [0,1]$, we have

$$\begin{aligned}
&\mathbb{P}\left[\overline{\mathcal{A}}^L\left(\epsilon_1^2/2\right) \bigwedge \overline{\mathcal{B}}^{h+1}\left(\epsilon_2\right) \bigwedge \overline{\mathcal{C}}\left(\epsilon_3\right) \bigwedge \overline{\mathcal{D}}^h\left(\epsilon_4\right)\right. \\
&\left.\quad \Longrightarrow \overline{\mathcal{B}}^h\left(\epsilon_2 + \frac{C'\epsilon_3}{\sqrt{d_h}} + 2\epsilon_4 + C\sqrt{\frac{\log\frac{1}{\delta_2}}{d_h}}\right)\right] \qquad (46) \\
&\geq 1 - \delta_2.
\end{aligned}$$

Proof of Theorem B.2 (8/47)

Proof of Theorem B.2.

We will use induction on Lemma B.7 to prove Theorem B.2. In the statement of Theorem B.1, we set $\delta_1 = \frac{\delta}{4}$, $\epsilon_1 = \frac{\epsilon^2}{8}$, for some c_1, c_2, we have

$$\mathbb{P}\left[\overline{\mathcal{A}}^L\left(\epsilon^2/8\right)\right] \geq 1 - \delta/4 \tag{47}$$

Proof of Theorem B.2 (9/47)

Proof of Theorem B.2.

In the statement of Lemma B.6, we set $\delta_4 = \frac{\delta_2}{4}$, and $\epsilon_1 = \frac{\epsilon}{2}$. Note that for c_1 large enough $\sqrt{\frac{2\log\frac{24L}{\delta}}{\min_h d_h}} \leq \frac{\epsilon}{2}$ and thus we have

$$\begin{aligned}\mathbb{P}\left[\overline{\mathcal{A}}\left(\epsilon^2/8\right) \Rightarrow \overline{\mathcal{D}}\left(\epsilon\right)\right] &\geq \mathbb{P}\left[\overline{\mathcal{A}}\left(\epsilon^2/8\right) \Rightarrow \overline{\mathcal{D}}\left(\epsilon/2 + \sqrt{\frac{2\log\frac{24L}{\delta}}{\min_h d_h}}\right)\right] \\ &\geq 1 - \delta/4\end{aligned} \tag{48}$$

In the statement of Lemma B.4, we set $\delta_3 = \frac{\delta}{4}$, and $\epsilon_1 = \frac{\epsilon^2}{8}$, we have

$$\mathbb{P}\left[\overline{\mathcal{A}}^L\left(\epsilon^2/8\right) \Rightarrow \overline{\mathcal{C}}\left(2\sqrt{\log\frac{16}{\delta}}\right)\right] \geq 1 - \delta/4 \tag{49}$$

Proof of Theorem B.2 (10/47)

Proof of Theorem B.2.

Using union bound on Equation (47),(48),(49), we have

$$\mathbb{P}\left[\overline{\mathcal{A}}^{L}\left(\epsilon^{2}/8\right)\bigwedge\overline{\mathcal{C}}\left(2\sqrt{\log\frac{16}{\delta}}\right)\bigwedge\overline{\mathcal{D}}\left(\epsilon\right)\right]\geq 1-\frac{3\delta}{4} \tag{50}$$

Now we will begin the induction argument. First of all, note that $\mathbb{P}\left[\overline{\mathcal{B}}^{L+1}\left(0\right)\right]=1$ by definition.

Proof of Theorem B.2 (11/47)

Proof of Theorem B.2.

For $1 \leq h \leq L$ in the statement of Lemma B.7, we set $\epsilon_2 = 3(L+1-h)\epsilon$, $\epsilon_3 = 3\sqrt{\log \frac{16}{\delta}}$, $\epsilon_2 = \epsilon$, $\delta_4 = \frac{\delta}{4L}$. Note that for c_1 large enough, $C\sqrt{\frac{\log \frac{1}{\delta_2}}{d_h}} + C'\sqrt{\frac{\log \frac{L}{\delta_2}}{d_h}} < \epsilon$. Thus we have

Proof of Theorem B.2 (12/47)

Proof of Theorem B.2.

$$
\begin{aligned}
&\mathbb{P}\left[\overline{\mathcal{B}}^{h+1}\left((3L-3h)\epsilon\right)\bigwedge\overline{\mathcal{C}}\left(3\sqrt{\log\frac{16}{\delta_2}}\right)\bigwedge\overline{\mathcal{D}}^{h}\left(\epsilon\right)\right.\\
&\quad\left.\Rightarrow\overline{\mathcal{B}}^{h}\left((3L+2-3h)\epsilon+C\sqrt{\frac{\log\frac{1}{\delta}}{d_h}}+3C'\sqrt{\frac{\log\frac{16}{\delta}}{d_h}}\right)\right]\\
\geq&\mathbb{P}\left[\overline{\mathcal{B}}^{h+1}\left((3L-3h)\epsilon\right)\bigwedge\overline{\mathcal{C}}\left(3\sqrt{\log\frac{16}{\delta_2}}\right)\bigwedge\overline{\mathcal{D}}^{h}\left(\epsilon\right)\right.\\
&\quad\left.\Rightarrow\overline{\mathcal{B}}^{h}\left((3L+3-3h)\epsilon\right)\right]\\
\geq&1-\frac{\delta}{4L}
\end{aligned}
\tag{51}
$$

Proof of Theorem B.2 (13/47)

Proof of Theorem B.2.

Using union bound again on Equation (50) and Equation (51) for every h in $\{1, 2, \ldots, L\}$, we have

$$
\begin{aligned}
&\mathbb{P}\left[\overline{\mathcal{A}}^L\left(\epsilon^2/8\right) \bigwedge \overline{\mathcal{C}}\left(\epsilon\right) \bigwedge \overline{\mathcal{D}}\left(\epsilon\right) \bigwedge \overline{\mathcal{B}}\left(3L\epsilon\right)\right] \\
\geq &\mathbb{P}\left[\overline{\mathcal{A}}^L\left(\epsilon^2/8\right) \bigwedge \overline{\mathcal{C}}\left(\epsilon\right) \bigwedge \overline{\mathcal{D}}\left(\epsilon\right) \bigwedge_{h=1}^{L} \overline{\mathcal{B}}\left(3(L+1-h)\epsilon\right)\right] \\
\geq &\left(1 - \mathbb{P}\left[\overline{\mathcal{A}}^L\left(\epsilon^2/8\right) \bigwedge \overline{\mathcal{C}}\left(\epsilon\right) \bigwedge \overline{\mathcal{D}}\left(\epsilon\right)\right]\right) \\
&+ \sum_{h=1}^{L}\left(1 - \mathbb{P}\left[\overline{\mathcal{B}}^{h+1}\left((3L-3h)\epsilon\right) \bigwedge \overline{\mathcal{C}}\left(\epsilon\right) \bigwedge \overline{\mathcal{D}}^h\left(\epsilon\right) \Rightarrow \overline{\mathcal{B}}^h\left((3L+3-3h)\epsilon\right)\right]\right) \\
\geq &1 - \delta
\end{aligned}
\tag{52}
$$

□

Proof of Theorem B.2 (14/47)

Lemma B.4

$$\mathbb{P}\left[\overline{\mathcal{A}}^{L}\left(\epsilon_1^2/2\right) \Longrightarrow \overline{\mathcal{C}}\left(2\sqrt{\log\frac{4}{\delta_3}}\right)\right] \geq 1-\delta_3, \quad \forall \epsilon_1 \in [0,1], \delta_3 \in (0,1). \tag{53}$$

Proof of Theorem B.2 (15/47)

Proof of Lemma B.4.

For fixed $\boldsymbol{g}^{(L)}(\boldsymbol{x})$, $f(\boldsymbol{\theta}, \boldsymbol{x}) = \boldsymbol{W}^{(L+1)}\boldsymbol{g}^{(L)}(\boldsymbol{x}) \overset{\mathrm{d}}{=\!=} N(0, \|\boldsymbol{g}^{(L)}(\boldsymbol{x})\|^2$. Thus by subgaussian concentration[cite], we know w.p. $\geq 1-\delta$ over the randomness of $\boldsymbol{W}^{(L+1)}$, $|f(\boldsymbol{\theta}, \boldsymbol{x})| \leq \sqrt{2\log\frac{2}{\delta}}\|\boldsymbol{g}^{(L)}(\boldsymbol{x})\|$.

For $\epsilon_1 \leq 1$, we have $\epsilon_1^2/2 < 1$, which implies $\|\boldsymbol{g}^{(L)}(\boldsymbol{x})\|^2 \leq 1 + \frac{\epsilon_1^2}{2} \leq 2$, and thus taking union bound over $\boldsymbol{x}, \boldsymbol{x}'$, we have w.p. $\geq 1-\delta$, $|f(\boldsymbol{\theta}, \boldsymbol{x})| \leq 2\sqrt{\log\frac{2}{\delta}}$,$|f(\boldsymbol{\theta}, \boldsymbol{x}')| \leq 2\sqrt{\log\frac{2}{\delta}}$.

Proof of Theorem B.2 (16/47)

Proof of Lemma B.4.

$$\mathbb{P}\left[\overline{\mathcal{A}}^{L}\left(\epsilon_1^2/2\right) \Rightarrow \overline{\mathcal{C}}\left(2\sqrt{\log\frac{4}{\delta_3}}\right)\right] \geq \mathbb{P}\left[\overline{\mathcal{C}}\left(2\sqrt{\log\frac{4}{\delta_3}}\right) \mid \overline{\mathcal{A}}^{L}\left(\epsilon_1^2/2\right)\right] \geq 1-\delta \tag{54}$$

□

Lemma B.5

$$\mathbb{P}\left[\overline{\mathcal{A}}^{h+1}\left(\epsilon_1^2/2\right) \Longrightarrow \overline{\mathcal{D}}^{h}\left(\epsilon_1 + \sqrt{\frac{2\log\frac{6}{\delta_4}}{d_h}}\right)\right] \geq 1-\delta_4, \quad \forall \epsilon_1 \in [0,1], \delta_4 \in (0,1). \tag{55}$$

Proof of Theorem B.2 (17/47)

Lemma B.8

Define $\boldsymbol{G}^{(h)}(\boldsymbol{x},\boldsymbol{x}') = \begin{bmatrix} \boldsymbol{g}^{(h)}(\boldsymbol{x})^\top \boldsymbol{g}^{(h)}(\boldsymbol{x}) & \boldsymbol{g}^{(h)}(\boldsymbol{x})^\top \boldsymbol{g}^{(h)}(\boldsymbol{x}') \\ \boldsymbol{g}^{(h)}(\boldsymbol{x}')^\top \boldsymbol{g}^{(h)}(\boldsymbol{x}) & \boldsymbol{g}^{(h)}(\boldsymbol{x}')^\top \boldsymbol{g}^{(h)}(\boldsymbol{x}') \end{bmatrix}$, *we have for every* $1 \le h \le L$,

$$\begin{aligned} &\|\boldsymbol{G}^{(h)}(\boldsymbol{x},\boldsymbol{x}') - \boldsymbol{\Lambda}^{(h)}(\boldsymbol{x},\boldsymbol{x}')\|_\infty \le \frac{\epsilon^2}{2} \\ \Rightarrow &\left| t_{\dot\sigma}\left(\boldsymbol{G}^{(h)}(\boldsymbol{x},\boldsymbol{x}')\right) - t_{\dot\sigma}\left(\boldsymbol{\Lambda}^{(h)}(\boldsymbol{x},\boldsymbol{x}')\right)\right| \le \epsilon, \forall 0 \le \epsilon \le 1. \end{aligned} \tag{56}$$

Proof of Theorem B.2 (18/47)

Proof of Lemma B.8.

For simplicity, we denote $\boldsymbol{G}^{(h)}(\boldsymbol{x}, \boldsymbol{x}')$, $\boldsymbol{\Lambda}^{(h)}(\boldsymbol{x}, \boldsymbol{x}')$ by $\boldsymbol{G}$,$\boldsymbol{\Lambda}$ respectively. Since $\dot{\sigma}(z) = \mathbf{1}[z \geq 0]$ is 0-homogeneous, we have

$$\begin{aligned} t_{\dot{\sigma}}(\boldsymbol{G}) =& \hat{t}_{\dot{\sigma}}\left(\frac{G_{12}}{\sqrt{G_{11}G_{22}}}\right) = \frac{1}{2} + \arcsin\frac{G_{12}}{\sqrt{G_{11}G_{22}}} \\ t_{\dot{\sigma}}(\boldsymbol{\Lambda}) =& \hat{t}_{\dot{\sigma}}\left(\frac{\Lambda_{12}}{\sqrt{\Lambda_{11}\Lambda_{22}}}\right) = \frac{1}{2} + \arcsin\frac{\Lambda_{12}}{\sqrt{\Lambda_{11}\Lambda_{22}}} = \frac{1}{2} + \arcsin\Lambda_{12} \end{aligned} \tag{57}$$

It is easy to verify that $|\sqrt{G_{11}G_{22}} - 1| \leq \epsilon^2/2$, and thus

$$\begin{aligned} \left|\frac{G_{12}}{\sqrt{G_{11}G_{22}}} - \Lambda_{12}\right| &\leq \left|\frac{G_{12}}{\sqrt{G_{11}G_{22}}} - \frac{\Lambda_{12}}{\sqrt{G_{11}G_{22}}}\right| + |\Lambda_{12}|\left|1 - \frac{1}{\sqrt{G_{11}G_{22}}}\right| \\ &\leq \frac{\epsilon^2/2}{1-\epsilon^2/2} + \frac{\epsilon^2/2}{1-\epsilon^2/2} \leq 2\epsilon^2. \end{aligned} \tag{58}$$

Proof of Theorem B.2 (19/47)

Proof of Lemma B.8.

Thus, by Lemma B.1

$$|t_{\dot{\sigma}}(\boldsymbol{G}) - t_{\dot{\sigma}}(\boldsymbol{\Lambda})| \leq \left|\frac{1}{2} + \arcsin\frac{G_{12}}{\sqrt{G_{11}G_{22}}} - \frac{1}{2} + \arcsin\Lambda_{12}\right| \leq \epsilon. \quad (59)$$

□

Proof of Theorem B.2 (20/47)

Lemma B.9

For any $0 \leq h \leq L-1$, *any fixed* $\{\boldsymbol{W}^{(i)}\}_{i=1}^{h}$, *w.p.* $1-\delta$ *over the randomness of* $\boldsymbol{W}^{(h+1)} \in \mathbb{R}^{d^{h+1} \times d^h}$, *we have*

$$\left|2\frac{\operatorname{tr}(\boldsymbol{D})}{d_h} - \hat{t}_{\dot{\sigma}}\left(\boldsymbol{G}^{(h)}(\boldsymbol{x}, \boldsymbol{x}')\right)\right| < \sqrt{\frac{2\log\frac{2}{\delta}}{d_h}}. \tag{60}$$

Proof of Lemma B.9.

Notice that $\mathbb{E}2\frac{\operatorname{tr}(\boldsymbol{D})}{d_h} = \hat{t}_{\dot{\sigma}}\left(\boldsymbol{G}^{(h)}(\boldsymbol{x}, \boldsymbol{x}')\right)$, the proof is completed by Chernoff Bound. □

Proof of Theorem B.2 (21/47)

Proof of Lemma B.5.

Note that $\dot{\Sigma}^{(h)}(\boldsymbol{x}, \boldsymbol{x}') = t_{\sigma'}\left(\boldsymbol{\Sigma}^{(h)}\Big|_{\boldsymbol{x},\boldsymbol{x}'}\right) = \hat{t}_{\sigma'}\left(\boldsymbol{\Lambda}^{(h)}(\boldsymbol{x}, \boldsymbol{x}')\right)$.

Combining Lemma B.8 and Lemma B.9, we have for any $(\boldsymbol{x}, \boldsymbol{x}')$,

$$\mathbb{P}\left[\mathcal{D}^h\left(\boldsymbol{x}, \boldsymbol{x}', \epsilon_1 + \sqrt{\frac{2\log\frac{6}{\delta}}{d_h}}\right) \mid \mathcal{A}^{h+1}\left(\boldsymbol{x}, \boldsymbol{x}', \epsilon_1^2/2\right)\right] \geq 1 - \frac{\delta}{3}.$$

Proof of Theorem B.2 (22/47)

Proof of Lemma B.5.

Taking union bound over $(\boldsymbol{x}, \boldsymbol{x}), (\boldsymbol{x}, \boldsymbol{x}'), (\boldsymbol{x}', \boldsymbol{x}')$ for the choice of $(\boldsymbol{x}, \boldsymbol{x}')$, we have

$$\begin{aligned} &\mathbb{P}\left[\overline{\mathcal{A}}^{h+1}\left(\epsilon_1^2/2\right) \Rightarrow \overline{\mathcal{D}}^{h}\left(\epsilon_1 + \sqrt{\frac{2\log\frac{6}{\delta}}{d_h}}\right)\right] \\ \geq &\mathbb{P}\left[\overline{\mathcal{D}}^{h}\left(\epsilon_1 + \sqrt{\frac{2\log\frac{6}{\delta}}{d_h}}\right) \mid \overline{\mathcal{A}}^{h+1}\left(\epsilon_1^2/2\right)\right] \geq 1-\delta \end{aligned} \tag{61}$$

Proof of Theorem B.2 (23/47)

Lemma B.7

There exists constant $C, C' \in \mathbb{R}$, for any $\epsilon_2, \epsilon_3, \epsilon_4 \in [0,1]$, we have

$$\begin{aligned}
&\mathbb{P}\left[\overline{\mathcal{A}}^L\left(\epsilon_1^2/2\right) \bigwedge \overline{\mathcal{B}}^{h+1}\left(\epsilon_2\right) \bigwedge \overline{\mathcal{C}}\left(\epsilon_3\right) \bigwedge \overline{\mathcal{D}}^h\left(\epsilon_4\right)\right. \\
&\left.\quad \Longrightarrow \overline{\mathcal{B}}^h\left(\epsilon_2 + \frac{C'\epsilon_3}{\sqrt{d_h}} + 2\epsilon_4 + C\sqrt{\frac{\log\frac{1}{\delta_2}}{d_h}}\right)\right] \qquad (62) \\
&\geq 1 - \delta_2.
\end{aligned}$$

Proof of Theorem B.2 (24/47)

The proof of Lemma B.7 is based on the following 3 claims, Claim B.1, B.2 and B.3.

Claim B.1

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

$$\left|\frac{2\operatorname{tr}(\boldsymbol{D})}{d_h}\left\langle \mathbf{b}^{(h)}(\boldsymbol{x}^{(2)}), \mathbf{b}^{(h)}(\boldsymbol{x}^{(1)})\right\rangle - \prod_{h'=h}^{L} \dot{\Sigma}^{(h')}(\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)})\right| \leq \epsilon_2 + 2\epsilon_4. \tag{63}$$

Proof of Theorem B.2 (25/47)

Proof of Claim B.1.

$$
\begin{aligned}
&\left|\frac{2\operatorname{tr}(\boldsymbol{D})}{d_h}\left\langle \mathbf{b}^{(h)}(\boldsymbol{x}^{(2)}), \mathbf{b}^{(h)}(\boldsymbol{x}^{(1)})\right\rangle - \prod_{h'=h}^{L}\dot{\Sigma}^{(h')}(\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)})\right| \\
\leq &\left|\frac{2\operatorname{tr}(\boldsymbol{D})}{d_h} - \dot{\Sigma}^{(h)}(\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)})\right| \cdot \left|\left\langle \mathbf{b}^{(h)}(\boldsymbol{x}^{(2)}), \mathbf{b}^{(h)}(\boldsymbol{x}^{(1)})\right\rangle\right| \\
+ &\left|\dot{\Sigma}^{(h)}(\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)})\right| \cdot \left|\left\langle \mathbf{b}^{(h)}(\boldsymbol{x}^{(2)}), \mathbf{b}^{(h)}(\boldsymbol{x}^{(1)})\right\rangle - \prod_{h'=h+1}^{L}\dot{\Sigma}^{(h')}(\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)})\right| \\
\leq &2\epsilon_4 + \epsilon_2
\end{aligned}
\tag{64}
$$

□

Proof of Theorem B.2 (26/47)

For any fixed h, let $\boldsymbol{G} = [\boldsymbol{g}^{(h)}(\boldsymbol{x})\boldsymbol{g}^{(h)}(\boldsymbol{x}')]$,

Claim B.2

w.p. $\geq 1 - \frac{\delta_2}{2}$, *if* $\overline{\mathcal{A}}^L(\epsilon_1^2/2) \bigwedge \overline{\mathcal{B}}^{h+1}(\epsilon_2) \bigwedge \overline{\mathcal{C}}(\epsilon_3) \bigwedge \overline{\mathcal{D}}^h(\epsilon_4)$, *then we have for any* $(\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}) \in \{(\boldsymbol{x}, \boldsymbol{x}), (\boldsymbol{x}, \boldsymbol{x}'), (\boldsymbol{x}', \boldsymbol{x}')\}$,

$$\begin{aligned}\Bigg|\frac{2}{d_h}\mathbf{b}^{(h+1)}(\boldsymbol{x}^{(1)})^\top \boldsymbol{W}^{(h+1)}\Pi_{\boldsymbol{G}}^{\perp}\boldsymbol{D}\Pi_{\boldsymbol{G}}^{\perp}\left(\boldsymbol{W}^{(h+1)}\right)^\top \mathbf{b}^{(h+1)}(\boldsymbol{x}^{(2)}) \\ -\frac{2\operatorname{tr}(\boldsymbol{D})}{d_h}\left\langle \mathbf{b}^{(h)}(\boldsymbol{x}^{(2)}), \mathbf{b}^{(h)}(\boldsymbol{x}^{(1)})\right\rangle\Bigg| \leq 16\sqrt{\frac{\log\frac{6}{\delta_2}}{d_h}}.\end{aligned} \tag{65}$$

As a by-product, for any $\boldsymbol{x}^{(1)} \in \{\boldsymbol{x}, \boldsymbol{x}'\}$, *we have*

$$\sqrt{\frac{2}{d_h}}\|\mathbf{b}^{(h+1)}(\boldsymbol{x}^{(1)})^\top \boldsymbol{W}^{(h+1)}\Pi_{\boldsymbol{G}}^{\perp}\boldsymbol{D}\| \leq 4\sqrt{\frac{\log\frac{6}{\delta_2}}{d_h}}. \tag{66}$$

Proof of Theorem B.2 (27/47)

Lemma B.10 (Gaussian chaos of order 2,[BLM13])

Let $\boldsymbol{\xi} \sim N(0, \boldsymbol{I}_n)$ *be an* n*-dimensional unit gaussian random vector,* $\boldsymbol{A} \in \mathbb{R}^{n \times n}$ *be a symmetric matrix, then for any* $t > 0$,

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

Or,

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

Proof of Theorem B.2 (28/47)

Proof of Claim B.2.

It suffices to prove this claim conditioned on every possible realization of

$$\{\mathbf{b}^{(h+1)}(\boldsymbol{x}^{(1)}), \mathbf{b}^{(h+1)}(\boldsymbol{x}^{(2)}), \boldsymbol{f}^{(h)}(\boldsymbol{x}^{(1)}), \boldsymbol{f}^{(h)}(\boldsymbol{x}^{(2)})\}.$$

Recall that $\boldsymbol{G} = \left[\boldsymbol{g}^{(h)}(\boldsymbol{x}^{(1)})\boldsymbol{g}^{(h)}(\boldsymbol{x}^{(2)})\right]$, we further define $\boldsymbol{F} = \left[\boldsymbol{f}^{(h)}(\boldsymbol{x}^{(1)})\boldsymbol{f}^{(h)}(\boldsymbol{x}^{(2)})\right]$. Applying Lemma B.3 on each row of $\boldsymbol{W}^{h+1}$, we have

$$\boldsymbol{W}^{(h+1)}\Pi_{\boldsymbol{G}}^{\perp} \overset{\mathrm{d}}{=\!=}_{\boldsymbol{F}=\boldsymbol{W}^{(h+1)}\boldsymbol{G}} \widetilde{\boldsymbol{W}}\Pi_{\boldsymbol{G}}^{\perp}, \tag{67}$$

where $\widetilde{\boldsymbol{W}}$ is an iid copy of $\boldsymbol{W}^{(h+1)}$.

Proof of Theorem B.2 (29/47)

Proof of Claim B.2.

Note that $[\mathbf{b}^{(h+1)}(\boldsymbol{x})^\top \widetilde{\boldsymbol{W}} \quad \mathbf{b}^{(h+1)}(\boldsymbol{x}^{(1)})^\top \widetilde{\boldsymbol{W}}]^\top \in \mathbb{R}^{2d_h}$ follows a joint zero-mean gaussian distribution with covariance matrix $\boldsymbol{\Sigma} = \begin{bmatrix} b_{11}\boldsymbol{I}_n & b_{12}\boldsymbol{I}_n \\ b_{21}\boldsymbol{I}_n & b_{22}\boldsymbol{I}_n \end{bmatrix}$, where $b_{ij} = \mathbf{b}^{(h)}(\boldsymbol{x}^{(i)})^\top \mathbf{b}^{(h)}(\boldsymbol{x}^{(j)})$, for $i, j = 1, 2$. In other words, there exists $\boldsymbol{M} \in \mathbb{R}^{2d_h \times 2d_h}$, s.t. $\boldsymbol{M}\boldsymbol{M}^\top = \boldsymbol{\Sigma}$, and

$$[\mathbf{b}^{(h+1)}(\boldsymbol{x})^\top \widetilde{\boldsymbol{W}} \quad \mathbf{b}^{(h+1)}(\boldsymbol{x}^{(1)})^\top \widetilde{\boldsymbol{W}}]^\top \overset{\mathrm{d}}{=\!=} M\boldsymbol{\xi},$$

where $\boldsymbol{\xi} \sim N(\mathbf{0}, \boldsymbol{I}_{2d_h})$.

Proof of Theorem B.2 (30/47)

Proof of Claim B.2.

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

$$
\begin{aligned}
& \mathbf{b}^{(h+1)}(\boldsymbol{x}^{(1)})^\top \boldsymbol{W}^{(h+1)} \Pi_{\boldsymbol{G}}^{\perp} \boldsymbol{D} \Pi_{\boldsymbol{G}}^{\perp} \left(\boldsymbol{W}^{(h+1)}\right)^\top \mathbf{b}^{(h+1)}(\boldsymbol{x}^{(2)}) \\
\overset{\mathrm{d}}{=\!=} & \mathbf{b}^{(h+1)}(\boldsymbol{x}^{(1)})^\top \widetilde{\boldsymbol{W}} \Pi_{\boldsymbol{G}}^{\perp} \boldsymbol{D} \Pi_{\boldsymbol{G}}^{\perp} \left(\widetilde{\boldsymbol{W}}\right)^\top \mathbf{b}^{(h+1)}(\boldsymbol{x}^{(2)}) \\
\overset{\mathrm{d}}{=\!=} & \left(\begin{bmatrix} \boldsymbol{I}_{d_h} & \mathbf{0} \end{bmatrix} \boldsymbol{M} \boldsymbol{\xi}\right)^\top \Pi_{\boldsymbol{G}}^{\perp} \boldsymbol{D} \Pi_{\boldsymbol{G}}^{\perp} \left(\begin{bmatrix} \mathbf{0} & \boldsymbol{I}_{d_h} \end{bmatrix} \boldsymbol{M} \boldsymbol{\xi}\right) \\
\overset{\mathrm{d}}{=\!=} & \frac{1}{2} \boldsymbol{\xi}^\top \boldsymbol{M}^\top \begin{bmatrix} \mathbf{0} & \Pi_{\boldsymbol{G}}^{\perp} \boldsymbol{D} \Pi_{\boldsymbol{G}}^{\perp} \\ \Pi_{\boldsymbol{G}}^{\perp} \boldsymbol{D} \Pi_{\boldsymbol{G}}^{\perp} & \mathbf{0} \end{bmatrix} \boldsymbol{M} \boldsymbol{\xi}.
\end{aligned}
$$

Proof of Theorem B.2 (31/47)

Proof of Claim B.2.

Now we are ready to prove Claim B.1 by applying Lemma B.10. Let $\boldsymbol{A} = \frac{1}{2}\boldsymbol{M}^\top \begin{bmatrix} \mathbf{0} & \Pi_{\boldsymbol{G}}^\perp \boldsymbol{D} \Pi_{\boldsymbol{G}}^\perp \\ \Pi_{\boldsymbol{G}}^\perp \boldsymbol{D} \Pi_{\boldsymbol{G}}^\perp & \mathbf{0} \end{bmatrix} \boldsymbol{M}$, we have

$$\begin{aligned} \mathbb{E}\boldsymbol{\xi}^\top \boldsymbol{A}\boldsymbol{\xi} = \operatorname{tr}(\boldsymbol{A}) &= \frac{1}{2}\operatorname{tr}\left(\begin{bmatrix} \mathbf{0} & \Pi_{\boldsymbol{G}}^\perp \boldsymbol{D} \Pi_{\boldsymbol{G}}^\perp \\ \Pi_{\boldsymbol{G}}^\perp \boldsymbol{D} \Pi_{\boldsymbol{G}}^\perp & \mathbf{0} \end{bmatrix} \boldsymbol{\Sigma}\right) \\ &= b_{12}\operatorname{tr}\left(\Pi_{\boldsymbol{G}}^\perp \boldsymbol{D} \Pi_{\boldsymbol{G}}^\perp \boldsymbol{I}_n\right) = b_{12}\operatorname{tr}\left(\boldsymbol{D}\Pi_{\boldsymbol{G}}^\perp\right). \end{aligned} \tag{68}$$

Note that by definition $\Pi_{\boldsymbol{G}}^\perp = \boldsymbol{I}_{d_h} - \Pi_{\boldsymbol{G}}$, and $\operatorname{rank}(\Pi_{\boldsymbol{G}}) \le 2$, we have

$$\begin{aligned} \operatorname{tr}\left(\boldsymbol{D}\Pi_{\boldsymbol{G}}^\perp\right) &= \operatorname{tr}\left(\boldsymbol{D}(\boldsymbol{I} - \Pi_{\boldsymbol{G}})\right) = \operatorname{tr}(D) - \operatorname{tr}\left(\boldsymbol{D}\Pi_{\boldsymbol{G}}\right) \\ &= \operatorname{tr}(D) - \operatorname{tr}\left(\Pi_{\boldsymbol{G}}\boldsymbol{D}\Pi_{\boldsymbol{G}}\right). \end{aligned} \tag{69}$$

Proof of Theorem B.2 (32/47)

Proof of Claim B.2.

Since $\mathbf{0} \preceq \boldsymbol{D} \preceq \boldsymbol{I}_{d_h}$, we have $0 \leq \operatorname{tr}(\Pi_{\boldsymbol{G}} \boldsymbol{D} \Pi_{\boldsymbol{G}}) \leq 2$, and thus $b_{12}(\operatorname{tr}(\boldsymbol{D}) - 2) \leq \mathbb{E}\boldsymbol{\xi}^\top \boldsymbol{A}\boldsymbol{\xi} \leq b_{12} \operatorname{tr}(\boldsymbol{D})$. For the upper bound of spectrum, note that $\|\boldsymbol{M}\|_2^2 = \|\boldsymbol{\Sigma}\|_2 = \left\| \begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{bmatrix} \right\|_2 \leq b_{11} + b_{12}$, and $\mathbf{0} \preceq \Pi_{\boldsymbol{G}}^{\perp}, \boldsymbol{D} \preceq \boldsymbol{I}_{d_h}$, we have

$$\begin{aligned} \|\boldsymbol{A}\|_2 &\leq \frac{1}{2}\|\boldsymbol{M}\|_2^2 \|\Pi_{\boldsymbol{G}}^{\perp} \boldsymbol{D} \Pi_{\boldsymbol{G}}^{\perp}\|_2 \leq \frac{1}{2}\|\boldsymbol{M}\|_2^2 \|\Pi_{\boldsymbol{G}}^{\perp}\|_2 \|\boldsymbol{D}\|_2 \|\Pi_{\boldsymbol{G}}^{\perp}\|_2 \\ &\leq \frac{b_{11} + b_{12}}{2} \leq \sqrt{2}, \end{aligned} \tag{70}$$

and

$$\|\boldsymbol{A}\|_F \leq \sqrt{2d_h}\|\boldsymbol{A}\|_2 = \frac{\sqrt{2d_h}(b_{11} + b_{22})}{2} \leq 2\sqrt{d_h}.$$

Proof of Theorem B.2 (33/47)

Proof of Claim B.2.

Thus by Lemma B.10 with $t = \log \frac{6}{\delta_2}$ we have w.p. $1 - \frac{\delta_2}{6}$,

$$\begin{aligned} \frac{1}{d_h} \left| \boldsymbol{\xi}^\top \boldsymbol{A} \boldsymbol{\xi} - \mathbb{E} \boldsymbol{\xi}^\top \boldsymbol{A} \boldsymbol{\xi} \right| &\leq \frac{1}{d_h} \left(2 \|\boldsymbol{A}\|_F \sqrt{t} + 2 \|\boldsymbol{A}\|_2 t \right) \\ &= 4 \sqrt{\frac{\log \frac{6}{\delta_2}}{d_h}} + 2\sqrt{2} \frac{\log \frac{6}{\delta_2}}{d_h} .. \end{aligned} \tag{71}$$

Proof of Theorem B.2 (34/47)

Proof of Claim B.2.

Thus we have

$$
\begin{aligned}
&\left|\frac{2}{d_h}\mathbf{b}^{(h+1)}(\boldsymbol{x}^{(1)})^\top \boldsymbol{W}^{(h+1)}\Pi_{\boldsymbol{G}}^{\perp}\boldsymbol{D}\Pi_{\boldsymbol{G}}^{\perp}\left(\boldsymbol{W}^{(h+1)}\right)^\top \mathbf{b}^{(h+1)}(\boldsymbol{x}^{(2)})\right. \\
&\left.-\frac{2\operatorname{tr}(\boldsymbol{D})}{d_h}\left\langle \mathbf{b}^{(h)}(\boldsymbol{x}^{(2)}), \mathbf{b}^{(h)}(\boldsymbol{x}^{(1)})\right\rangle\right| \\
\leq &\frac{2}{d_h}\left|\boldsymbol{\xi}^\top \boldsymbol{A}\boldsymbol{\xi} - \mathbb{E}\boldsymbol{\xi}^\top \boldsymbol{A}\boldsymbol{\xi}\right| + \left|2\mathbb{E}\boldsymbol{\xi}^\top \boldsymbol{A}\boldsymbol{\xi} - \frac{2\operatorname{tr}(\boldsymbol{D})}{d_h}\left\langle \mathbf{b}^{(h)}(\boldsymbol{x}^{(2)}), \mathbf{b}^{(h)}(\boldsymbol{x}^{(1)})\right\rangle\right| \qquad (72) \\
\leq &8\sqrt{\frac{\log\frac{6}{\delta_2}}{d_h}} + 4\sqrt{2}\frac{\log\frac{6}{\delta_2}}{d_h} + \frac{4b_{12}}{d_h}
\end{aligned}
$$

Proof of Theorem B.2 (35/47)

Proof of Claim B.2.

$$\begin{aligned}
\leq & 14\sqrt{\frac{\log\frac{6}{\delta_2}}{d_h}} + \frac{4(1+\epsilon_2)}{d_h} \quad (2\sqrt{2} \leq 3 \wedge \log\frac{6}{\delta_2} \leq d_h) \\
\leq & 16\sqrt{\frac{\log\frac{6}{\delta_2}}{d_h}} \quad (\epsilon_2 \leq 1 \wedge \sqrt{d_h \log 6} \geq 4).
\end{aligned} \tag{73}$$

Proof of Theorem B.2 (36/47)

Proof of Claim B.2.

The main part of the claim is completed by taking union bound over $(\boldsymbol{x}, \boldsymbol{x}), (\boldsymbol{x}, \boldsymbol{x}'), (\boldsymbol{x}', \boldsymbol{x}')$. For the by-product, let $\boldsymbol{x}^{(2)} = \boldsymbol{x}^{(1)}$, and we have

$$
\begin{aligned}
&\sqrt{\frac{2}{d_h}} \|\mathbf{b}^{(h+1)}(\boldsymbol{x}^{(1)})^\top \boldsymbol{W}^{(h+1)} \Pi_{\boldsymbol{G}}^{\perp} \boldsymbol{D}\| \\
\leq &\sqrt{\left| \frac{2}{d_h} \mathbf{b}^{(h+1)}(\boldsymbol{x}^{(1)})^\top \boldsymbol{W}^{(h+1)} \Pi_{\boldsymbol{G}}^{\perp} \boldsymbol{D} \Pi_{\boldsymbol{G}}^{\perp} \left(\boldsymbol{W}^{(h+1)}\right)^\top \mathbf{b}^{(h+1)}(\boldsymbol{x}^{(2)}) \right|} \\
\leq &\sqrt{\frac{2\,\mathrm{tr}\,(\boldsymbol{D})}{d_h} \langle \mathbf{b}^{(h)}(\boldsymbol{x}^{(2)}), \mathbf{b}^{(h)}(\boldsymbol{x}^{(1)}) \rangle + \left(16 \sqrt{\frac{\log \frac{6}{\delta_2}}{d_h}} \right)^2}
\end{aligned}
\tag{74}
$$

Proof of Theorem B.2 (37/47)

Proof of Claim B.2.

$$\leq \sqrt{4 + \left(16\sqrt{\frac{\log\frac{6}{\delta_2}}{d_h}}\right)^2} \tag{75}$$

$$\leq 2 + 4\sqrt{\frac{\log\frac{6}{\delta_2}}{d_h}} \leq 6 \ (\log\frac{6}{\delta_2} \leq d_h)$$

□

Proof of Theorem B.2 (38/47)

Claim B.3

w.p. $\geq 1 - \frac{\delta_2}{2}$, *if* $\overline{\mathcal{A}}^L\left(\epsilon_1^2/2\right) \bigwedge \overline{\mathcal{B}}^{h+1}\left(\epsilon_2\right) \bigwedge \overline{\mathcal{C}}\left(\epsilon_3\right) \bigwedge \overline{\mathcal{D}}^h\left(\epsilon_4\right)$, *then*

$$\begin{aligned} \left\|\boldsymbol{\Pi}_{\boldsymbol{G}}\left(\boldsymbol{W}^{(h+1)}\right)^{\top} \mathbf{b}^{(h+1)}(\boldsymbol{x})\right\| &\leq 2\sqrt{\log\frac{8}{\delta_2}} + \sqrt{2}\epsilon_3, \\ \left\|\boldsymbol{\Pi}_{\boldsymbol{G}}\left(\boldsymbol{W}^{(h+1)}\right)^{\top} \mathbf{b}^{(h+1)}(\boldsymbol{x}')\right\| &\leq 2\sqrt{\log\frac{8}{\delta_2}} + \sqrt{2}\epsilon_3. \end{aligned} \tag{76}$$

Proof of Theorem B.2 (39/47)

Proof of Claim B.3.

It suffices to prove the claim for $\boldsymbol{x}$. We will denote $\boldsymbol{x}$ by $\boldsymbol{x}$, $\boldsymbol{g}^{(h)}(\boldsymbol{x})$ by $\boldsymbol{g}^{(h)}$ and $\mathbf{b}^{(h+1)}(\boldsymbol{x})$ by $\mathbf{b}^{(h+1)}$. We also define $\Pi_{\boldsymbol{g}}$ as $\boldsymbol{g}\boldsymbol{g}^\top$, and $\Pi_{\boldsymbol{G}/\boldsymbol{g}} = \Pi_{\boldsymbol{G}} - \Pi_{\boldsymbol{g}}$. Clearly, $\Pi_{\boldsymbol{G}/\boldsymbol{g}}$ is still a projection matrix of rank 0 or 1.

Since $\|\boldsymbol{\Pi}_{\boldsymbol{G}} \left(\boldsymbol{W}^{(h+1)}\right)^\top \mathbf{b}^{(h+1)}(\boldsymbol{x})\| \leq$ $\|\boldsymbol{\Pi}_{\boldsymbol{g}} \left(\boldsymbol{W}^{(h+1)}\right)^\top \mathbf{b}^{(h+1)}\| + \|\boldsymbol{\Pi}_{\boldsymbol{G}/\boldsymbol{g}} \left(\boldsymbol{W}^{(h+1)}\right)^\top \mathbf{b}^{(h+1)}\|$, it suffices to bound these two terms separately.

Proof of Theorem B.2 (40/47)

Proof of Claim B.3.

Recall $\mathbf{b}^{(h+1)}$ is defined as the gradient of $f(\boldsymbol{\theta}, \boldsymbol{x})$ with respect to the pre-activation of layer $h+1$, $\boldsymbol{f}^{h+1}$, thus if we view g as a function $\boldsymbol{g}^{(h)}, \boldsymbol{W}^{(h+1)}, \ldots, \boldsymbol{W}^{(L+1)}$, by the rule of back propagation, we have

$$\frac{\partial g(\boldsymbol{g}^{(h)}, \boldsymbol{W}^{(h+1)}, \ldots, \boldsymbol{W}^{(L+1)})}{\partial \boldsymbol{g}^{(h)}} = (\mathbf{b}^{(h+1)})^{\top} \boldsymbol{W}^{(h+1)}.$$

Note that relu is 1-homogeneous, namely $\forall \lambda \in \mathbb{R}, \sigma(\lambda z) = \lambda \sigma(z)$, the whole network is also 1-homogeneous in $\boldsymbol{g}^{(h)}$. In other words, we have

Proof of Theorem B.2 (41/47)

Proof of Claim B.3.

$$
\begin{aligned}
& f(\boldsymbol{g}^{(h)}, \boldsymbol{W}^{(h+1)}, \ldots, \boldsymbol{W}^{(L+1)}) \\
= & \left. \frac{\partial f(\lambda \boldsymbol{g}^{(h)}, \boldsymbol{W}^{(h+1)}, \ldots, \boldsymbol{W}^{(L+1)})}{\partial \lambda} \right|_{\lambda=1} \\
= & \left\langle \left. \frac{\partial f(\lambda \boldsymbol{g}^{(h)}, \boldsymbol{W}^{(h+1)}, \ldots, \boldsymbol{W}^{(L+1)})}{\partial \lambda \boldsymbol{g}^{(h)}} \right|_{\lambda=1}, \left. \frac{\partial \lambda \boldsymbol{g}^{(h)}}{\partial \lambda} \right|_{\lambda=1} \right\rangle \qquad (77) \\
= & \left\langle \frac{\partial g(\boldsymbol{g}^{(h)}, \boldsymbol{W}^{(h+1)}, \ldots, \boldsymbol{W}^{(L+1)})}{\partial \boldsymbol{g}^{(h)}}, \boldsymbol{g}^{(h)} \right\rangle \\
= & (\boldsymbol{g}^{(h)})^{\top} \left(\boldsymbol{W}^{(h+1)} \right)^{\top} \mathbf{b}^{(h+1)}
\end{aligned}
$$

Proof of Theorem B.2 (42/47)

Proof of Claim B.3.

By definition of $\Pi_{\boldsymbol{g}}$, we have

$$\begin{aligned}\left\|\boldsymbol{\Pi}_{\boldsymbol{g}}\left(\boldsymbol{W}^{(h+1)}\right)^{\top}\mathbf{b}\right\| &= \left\|\frac{\boldsymbol{g}^{(h)}(\boldsymbol{g}^{(h)})^{\top}}{\|\boldsymbol{g}^{(h)}\|^{2}}\left(\boldsymbol{W}^{(h+1)}\right)^{\top}\mathbf{b}^{(h+1))}\right\| \\ &= \left|\frac{(\boldsymbol{g}^{(h)})^{\top}}{\|\boldsymbol{g}^{(h)}\|}\left(\boldsymbol{W}^{(h+1)}\right)^{\top}\mathbf{b}^{(h+1))}\right| \\ &= \frac{|f(\boldsymbol{\theta},\boldsymbol{x})|}{\|\boldsymbol{g}^{(h)}\|}.\end{aligned} \tag{78}$$

Proof of Theorem B.2 (43/47)

Proof of Claim B.3.

Note that $\boldsymbol{g}^{(h)}(x^{(0)})^\top \boldsymbol{g}^{(h)}(\boldsymbol{x}) \geq 1 - \epsilon_1^2/2 \geq \frac{1}{2}$, we have

$$\|\boldsymbol{\Pi}_{\boldsymbol{g}} \left(\boldsymbol{W}^{(h+1)}\right)^\top \mathbf{b}\| = \frac{|f(\boldsymbol{\theta}, \boldsymbol{x})|}{\|\boldsymbol{g}^{(h)}\|} \leq \sqrt{2}\epsilon_3.$$

For the second term $\boldsymbol{\Pi}_{\boldsymbol{G}/\boldsymbol{g}} \left(\boldsymbol{W}^{(h+1)}\right)^\top \boldsymbol{b}^{(h+1)}$, note that conditioned on $\boldsymbol{g}^{(h)}, \boldsymbol{f}^h = \frac{1}{\sqrt{d_{h+1}}} \boldsymbol{W}^{(h+1)} \boldsymbol{g}^{(h)}$ and all $\{\boldsymbol{W}^{(h)}\})_{h'}^{L+1}$ (thus $\mathbf{b}^{(h+1)}$), by Lemma B.3, $\Pi_{\boldsymbol{G}/\boldsymbol{g}}(\boldsymbol{W}^{(h+1)}) \overset{\mathrm{d}}{=\!=} \Pi_{\boldsymbol{G}/\boldsymbol{g}} \widetilde{\boldsymbol{W}}$, where $\widetilde{\boldsymbol{W}}$ is an iid copy of $\boldsymbol{W}^{(h+1)}$.

Proof of Theorem B.2 (44/47)

Proof of Claim B.3.

Thus if $\operatorname{rank}(\boldsymbol{\Pi}_{\boldsymbol{G}/\boldsymbol{g}}) = 1$, suppose $\boldsymbol{\Pi}_{\boldsymbol{G}/\boldsymbol{g}} = \boldsymbol{u}\boldsymbol{u}^\top$ for some unit vector $\boldsymbol{u}$, we have

$$\begin{aligned} \|\boldsymbol{\Pi}_{\boldsymbol{G}/\boldsymbol{g}} \left(\boldsymbol{W}^{(h+1)}\right)^\top \boldsymbol{b}^{(h+1)}\| &= \left|\boldsymbol{u}^\top \left(\boldsymbol{W}^{(h+1)}\right)^\top \boldsymbol{b}^{(h+1)}\right| \\ &\overset{\mathrm{d}}{=} \left|\boldsymbol{u}^\top \left(\widetilde{\boldsymbol{W}}\right)^\top \boldsymbol{b}^{(h+1)}\right| \\ &\overset{\mathrm{d}}{=} |t|, \end{aligned} \tag{79}$$

where $t \sim N(0, \|\mathbf{b}^{(h+1)}\|)$. Hence w.p. $\geq 1 - \delta_2/4$ over the randomness of $\boldsymbol{W}^{(L)}$, $\|\boldsymbol{\Pi}_{\boldsymbol{G}/\boldsymbol{g}} \left(\boldsymbol{W}^{(h+1)}\right)^\top \boldsymbol{b}^{(h+1)}\| \leq \sqrt{2\log\frac{8}{\delta_2}}\|\mathbf{b}^{(h+1)}\| \leq \sqrt{2\log\frac{8}{\delta_2}} \leq 2\sqrt{\log\frac{8}{\delta_2}}$ $(\epsilon_2 < 1)$.

Proof of Theorem B.2 (45/47)

Proof of Claim B.3.

If $\operatorname{rank}(\boldsymbol{\Pi}_{\boldsymbol{G}/\boldsymbol{g}}) = 0$, then $\|\boldsymbol{\Pi}_{\boldsymbol{G}/\boldsymbol{g}} \left(\boldsymbol{W}^{(h+1)}\right)^\top \boldsymbol{b}^{(h+1)}\| = 0 < 2\sqrt{\log \frac{8}{\delta_2}}$.
Thus w.p. $\geq 1 - \frac{\delta_2}{4}$,

$$
\begin{aligned}
&\|\boldsymbol{\Pi}_{\boldsymbol{G}} \left(\boldsymbol{W}^{(h+1)}\right)^\top \mathbf{b}^{(h+1)}(\boldsymbol{x})\| \\
\leq &\|\boldsymbol{\Pi}_{\boldsymbol{g}} \left(\boldsymbol{W}^{(h+1)}\right)^\top \mathbf{b}^{(h+1)}\| + \|\boldsymbol{\Pi}_{\boldsymbol{G}/\boldsymbol{g}} \left(\boldsymbol{W}^{(h+1)}\right)^\top \mathbf{b}^{(h+1)}\| \qquad (80) \\
\leq &2\sqrt{\log \frac{8}{\delta_2}} + \sqrt{2}\epsilon_3.
\end{aligned}
$$

Thus by assumption $\log \frac{8}{\delta_2} \leq d_h$, we have
$2\sqrt{\log \frac{8}{\delta_2}} + \sqrt{2}\epsilon_3 \leq 2\sqrt{d_h} + \sqrt{2} \leq 3\sqrt{2d_h}$. □

Proof of Theorem B.2 (46/47)

Wrapping things up, by combining Claim B.2 and Claim B.3, we have w.p. $\geq 1-\delta_2$, for any pair of $(\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}) \in \{(\boldsymbol{x},\boldsymbol{x}), (\boldsymbol{x},\boldsymbol{x}'), (\boldsymbol{x}',\boldsymbol{x}'))\}$,

$$
\begin{aligned}
&\Big| \frac{2}{d_h} \mathbf{b}^{(h+1)}(\boldsymbol{x}^{(1)})^\top \left(\boldsymbol{W}^{(h+1)}\right) \boldsymbol{D}^{(h)}(\boldsymbol{x}^{(1)}) \boldsymbol{D}^{(h)}(\boldsymbol{x}^{(2)}) \left(\boldsymbol{W}^{(h+1)}\right)^\top \mathbf{b}^{(h+1)}(\boldsymbol{x}^{(2)}) - \\
&\frac{2}{d_h} \mathbf{b}^{(h+1)}(\boldsymbol{x}^{(1)})^\top \boldsymbol{W}^{(h+1)} \Pi_{\boldsymbol{G}}^{\perp} \boldsymbol{D} \Pi_{\boldsymbol{G}}^{\perp} \left(\boldsymbol{W}^{(h+1)}\right)^\top \mathbf{b}^{(h+1)}(\boldsymbol{x}^{(2)}) \Big| \\
\leq &\| \frac{2}{d_h} \mathbf{b}^{(h+1)}(\boldsymbol{x}^{(1)})^\top \boldsymbol{W}^{(h+1)} \Pi_{\boldsymbol{G}} \boldsymbol{D} \| \cdot \| \boldsymbol{D} \Pi_{\boldsymbol{G}}^{\perp} \left(\boldsymbol{W}^{(h+1)}\right)^\top \mathbf{b}^{(h+1)}(\boldsymbol{x}^{(2)}) \| \qquad (81) \\
+ &\| \frac{2}{d_h} \mathbf{b}^{(h+1)}(\boldsymbol{x}^{(1)})^\top \boldsymbol{W}^{(h+1)} \Pi_{\boldsymbol{G}}^{\perp} \boldsymbol{D} \| \cdot \| \boldsymbol{D} \Pi_{\boldsymbol{G}} \left(\boldsymbol{W}^{(h+1)}\right)^\top \mathbf{b}^{(h+1)}(\boldsymbol{x}^{(2)}) \| \\
+ &\| \frac{2}{d_h} \mathbf{b}^{(h+1)}(\boldsymbol{x}^{(1)})^\top \boldsymbol{W}^{(h+1)} \Pi_{\boldsymbol{G}} \| \cdot \| \Pi_{\boldsymbol{G}} \left(\boldsymbol{W}^{(h+1)}\right)^\top \mathbf{b}^{(h+1)}(\boldsymbol{x}^{(2)}) \|
\end{aligned}
$$

Proof of Theorem B.2 (47/47)

$$
\begin{aligned}
&\leq \left(12\sqrt{2}\sqrt{\frac{\ln\frac{8}{\delta_2}}{d_h}}+12\epsilon_3\right)+\left(12\sqrt{2}\sqrt{\frac{\ln\frac{8}{\delta_2}}{d_h}}+12\epsilon_3\right)+\left(12\sqrt{2}\sqrt{\frac{\ln\frac{8}{\delta_2}}{d_h}}+12\epsilon_3\right)\\
&=36\sqrt{\frac{2\ln\frac{8}{\delta_2}}{d_h}}+36\epsilon_3.
\end{aligned}
\tag{82}
$$

Using Equation (81) together with Claim B.1 and Claim B.2, we've finished the proof for Lemma B.7.

Contents

1. Neural Tangent Kernels (NTK)
   - Contruction of the NTK due to [ADH+19]
   - Convergence to the NTK at Initialization [ADH+19]
   - Equivalence between NN and NTK
2. Graph Neural Nets (GNN)
3. Graph Neural Tangent Kernel (GNTK)
   - Overview
   - Formulae
   - Theoretical Analyses
     - Proof of Theorem 4.2 [DHP+19]
     - Proof of Theorem 4.3 [DHP+19]

Lemma 3.1 [ADH+19] (1/4)

Lemma 3.1 (Evolution of outputs.)

- Recall in minimizing

$$\ell(\boldsymbol{\theta}) = \frac{1}{2}\sum_{i=1}^{n}\left(f(\boldsymbol{\theta}, \boldsymbol{x}_i) - y_i\right)^2, \tag{83}$$

with infinitesimally small learning rate, one has

$$\frac{\mathrm{d}\,\boldsymbol{\theta}(t)}{\mathrm{d}\,t} = -\nabla\ell(\boldsymbol{\theta}(t)). \tag{84}$$

Lemma 3.1 [ADH+19] (2/4)

Lemma 3.1 (Evolution of outputs.)

- Let $\boldsymbol{u}(t) = (f(\boldsymbol{\theta}(t), \boldsymbol{x}_i))_{i\in[n]} \in \mathbb{R}^n$ be the network outputs on all $\boldsymbol{x}_i$'s at time t, and $\boldsymbol{y} = (y_i)_{i\in[n]}$ be the desired outputs.

  Then $\boldsymbol{u}(t)$ follows

$$\frac{\mathrm{d}\,\boldsymbol{u}(t)}{\mathrm{d}\,t} = -\boldsymbol{H}(t)\cdot(\boldsymbol{u}(t)-\boldsymbol{y}), \tag{85}$$

  where $\boldsymbol{H}(t)$ is an $n\times n$ positive semidefinite matrix whose (i,j)-th entry is $\left\langle \frac{\partial f(\boldsymbol{\theta}(t),\boldsymbol{x}_i)}{\partial\boldsymbol{\theta}}, \frac{\partial f(\boldsymbol{\theta}(t),\boldsymbol{x}_j)}{\partial\boldsymbol{\theta}} \right\rangle$:

Lemma 3.1 [ADH+19] (3/4)

Lemma 3.1 (Evolution of outputs.)

- For *over-parameterized* deep neural networks including (i) fully-connected neural networks with wide hidden layers and (ii) convolutional neural networks with many channels in hidden layers, with suitable parameterization and initialization schemes, it can be shown that the kernel $\boldsymbol{H}(t)$ remains approximately unchanged during training, i.e.

$$\boldsymbol{H}(t) \approx \boldsymbol{H}(0). \tag{86}$$

  Furthermore, as the hidden widths go to infinity, the (stochastic) kernel $\boldsymbol{H}(0)$ at initialization converges to a deterministic limit $\boldsymbol{H}^*$. This gives rise to the *neural tangent kernel (NTK)*, as described in Equation (1), with $[\boldsymbol{H}^*]_{i,j} = k(\boldsymbol{x}_i, \boldsymbol{x}_j)$. In Section **??**, we describe the specific parameterizations for fully-connected and convolutional neural networks that give rise to the NTK phenomenon.

Lemma 3.1 [ADH+19] (4/4)

Lemma 3.1 (Evolution of outputs.)

Under approximations $\boldsymbol{H}(t) \approx \boldsymbol{H}(0) \approx \boldsymbol{H}^*$, if we work under the infinite width limit, Equation (90) can be replaced with the following *linear* dynamical system:

$$\frac{\boldsymbol{du}(t)}{\boldsymbol{d}t} = -\boldsymbol{H}^* \cdot (\boldsymbol{u}(t) - \boldsymbol{y}).$$

This is the same as the dynamics of *kernel regression* under gradient flow, for which at time $t \to \infty$ the final prediction function is (assuming $\boldsymbol{u}(0) = \boldsymbol{0}$)

$$f^*(\boldsymbol{x}) = (k(\boldsymbol{x}, \boldsymbol{x}_1), \ldots, k(\boldsymbol{x}, \boldsymbol{x}_n)) \cdot (\boldsymbol{H}^*)^{-1}\boldsymbol{y}. \tag{87}$$

Therefore, we can evaluate the performance of infinitely wide deep neural networks by looking at the kernel regression solution (87) using NTK. The problem boils down to the computation of NTK, which is the subject of Section **??**.

Proof of Lemma 3.1 [ADH+19] (1/2)

Proof

The parameters $\boldsymbol{\theta}$ evolve according to

$$\frac{\mathrm{d}\,\boldsymbol{\theta}(t)}{\mathrm{d}\,t} = -\nabla \ell(\boldsymbol{\theta}(t)) = -\sum_{i=1}^{n} \left(f(\boldsymbol{\theta}(t), \boldsymbol{x}_i) - y_i\right) \frac{\partial f(\boldsymbol{\theta}(t), \boldsymbol{x}_i)}{\partial \boldsymbol{\theta}}, \tag{88}$$

where $t \geq 0$ is a continuous time index. Under Equation (88), the evolution of the network output $f(\boldsymbol{\theta}(t), \boldsymbol{x}_i)$ can be written as

$$\begin{aligned} &\frac{\mathrm{d}\, f(\boldsymbol{\theta}(t), \boldsymbol{x}_i)}{\mathrm{d}\,t} \\ &= -\sum_{j=1}^{n} (f(\boldsymbol{\theta}(t), \boldsymbol{x}_j) - y_j) \left\langle \frac{\partial f(\boldsymbol{\theta}(t), \boldsymbol{x}_i)}{\partial \boldsymbol{\theta}}, \frac{\partial f(\boldsymbol{\theta}(t), \boldsymbol{x}_j)}{\partial \boldsymbol{\theta}} \right\rangle, \qquad \forall i \in [n]. \end{aligned} \tag{89}$$

Proof of Lemma 3.1 [ADH+19] (2/2)

Proof

Since $\boldsymbol{u}(t) = (f(\boldsymbol{\theta}(t), \boldsymbol{x}_i))_{i\in[n]} \in \mathbb{R}^n$ is the network outputs on all $\boldsymbol{x}_i$'s at time t, and $\boldsymbol{y} = (y_i)_{i\in[n]}$ is the desired outputs, Equation (89) can be written more compactly as

$$\frac{\mathrm{d}\,\boldsymbol{u}(t)}{\mathrm{d}\,t} = -\boldsymbol{H}(t)\cdot(\boldsymbol{u}(t) - \boldsymbol{y}), \tag{90}$$

where $\boldsymbol{H}(t) \in \mathbb{R}^{n\times n}$ is a kernel matrix defined as

$$\boldsymbol{H}(t)_{i,j} = \left\langle \frac{\partial f(\boldsymbol{\theta}(t), \boldsymbol{x}_i)}{\partial \boldsymbol{\theta}}, \frac{\partial f(\boldsymbol{\theta}(t), \boldsymbol{x}_j)}{\partial \boldsymbol{\theta}} \right\rangle, \tag{91}$$

$\forall i, j \in [n]$.

Equivalence Between NTK and NN (1/1)

Theorem 3.2 [ADH+19]

Theorem 3.2 (Main theorem)

Suppose $\sigma(z) = \max(0, z)$ $(z \in \mathbb{R})$, $1/\kappa = \mathrm{poly}(1/\epsilon, \log(n/\delta))$ *and* $d_1 = d_2 = \cdots = d_L = m$ *with* $m \geq \mathrm{poly}(1/\kappa, L, 1/\lambda_0, n, \log(1/\delta))$. *Then for any* $\boldsymbol{x}_{te} \in \mathbb{R}^d$ *with* $\|\boldsymbol{x}_{te}\| = 1$, *with probability at least* $1 - \delta$ *over the random initialization, we have*

$$|f_{nn}(\boldsymbol{x}_{te}) - f_{ntk}(\boldsymbol{x}_{te})| \leq \epsilon. \tag{92}$$

Conclusions

- GD on DNN equivalent to Kernel GD wrt the NTK.
- NTK for MLPs ([JGH18, ADH+19]); NTK for convolutional nets (CNTK, [ADH+19]); NTK for graph nets (GNTK, [DHP+19])

Contents

1 Neural Tangent Kernels (NTK)
- Contruction of the NTK due to [ADH+19]
- Convergence to the NTK at Initialization [ADH+19]
- Equivalence between NN and NTK

2 Graph Neural Nets (GNN)

3 Graph Neural Tangent Kernel (GNTK)
- Overview
- Formulae
- Theoretical Analyses
  - Proof of Theorem 4.2 [DHP+19]
  - Proof of Theorem 4.3 [DHP+19]

Notations and Setup

- Let $G = (V, E)$ be a graph with node features $\boldsymbol{h}_v \in \mathbb{R}^d$ for each $v \in V$.
- Denote the neighborhood of node v by $\mathcal{N}(v)$.
- Consider the graph classification task: given a set of graphs $\{G_1, ..., G_n\} \subseteq \mathcal{G}$ and their labels $\{y_1, ..., y_n\} \subseteq \mathcal{Y}$, to learn to predict labels of unseen graphs.

BLOCK Operation (1/2)

- A BLOCK operation aggregates features over a neighborhood $\mathcal{N}(u) \cup \{u\}$ via, e.g., summation, and transforms the aggregated features with non-linearity, e.g., MLP or a fully-connected layer followed by ReLU.
- Denote the number of fully-connected layers in each BLOCK operation, i.e., the number of hidden layers of an MLP, by R.
- When $R = 1$, the BLOCK operation can be formulated as

$$\text{BLOCK}^{(\ell)}(u) = \sqrt{\frac{c_\sigma}{m}} \cdot \sigma\left(\boldsymbol{W}_\ell \cdot c_u \sum_{v \in \mathcal{N}(u) \cup \{u\}} \boldsymbol{h}_v^{(\ell-1)}\right). \tag{93}$$

Here, $\boldsymbol{W}_\ell$ are learnable weights, initialized as Gaussian random variables. σ is an activation function like ReLU. m is the output dimension of $\boldsymbol{W}_\ell$.

BLOCK Operation (2/2)

- When the number of fully-connected layers $R = 2$, the BLOCK operation can be written as

$$\begin{aligned}&\text{BLOCK}^{(\ell)}(u)\\ =&\sqrt{\frac{c_\sigma}{m}}\sigma\left(\boldsymbol{W}_{\ell,2}\sqrt{\frac{c_\sigma}{m}}\cdot\sigma\left(\boldsymbol{W}_{\ell,1}\cdot c_u\sum_{v\in\mathcal{N}(u)\cup\{u\}}\boldsymbol{h}_v^{(\ell-1)}\right)\right),\end{aligned}\tag{94}$$

where $\boldsymbol{W}_{\ell,1}$ and $\boldsymbol{W}_{\ell,2}$ are learnable weights. BLOCK operations can be defined similarly for $R > 2$.

READOUT Operation

- To get the representation of an entire graph $\boldsymbol{h}_G$ after L steps of aggregation, we take the summation over all node features, i.e.,

$$\boldsymbol{h}_G = \text{READOUT}\left(\left\{\boldsymbol{h}_u^{(L)}, u \in V\right\}\right) = \sum_{u \in V} \boldsymbol{h}_u^{(L)}. \tag{95}$$

- There are more sophisticated READOUT operations than a simple summation. Jumping Knowledge Network (JK-Net) considers graph structures of different granularity, and aggregates graph features across all layers as

$$\boldsymbol{h}_G = \text{READOUT}^{\text{JK}}\left(\left\{\boldsymbol{h}_u^{(\ell)}, u \in V, \ell \in [L]\right\}\right) = \sum_{u \in V} \left[\boldsymbol{h}_u^{(0)}; \ldots; \boldsymbol{h}_u^{(L)}\right]. \tag{96}$$

Building GNNs using BLOCK and READOUT

- Most modern GNNs are constructed using the BLOCK operation and the READOUT operation. We denote the number of BLOCK operations (aggregation steps) in a GNN by L.
- For each $\ell \in [L]$ and $u \in V$, we define $\boldsymbol{h}_u^{(\ell)} = \mathrm{BLOCK}^{(\ell)}(u)$. The graph-level feature is then

$$\boldsymbol{h}_G = \mathrm{READOUT}\left(\left\{\boldsymbol{h}_u^{(L)}, u \in V\right\}\right) \tag{97}$$

or

$$\boldsymbol{h}_G = \mathrm{READOUT}^{\mathrm{JK}}\left(\left\{\boldsymbol{h}_u^{(\ell)}, u \in V, \ell \in [L]\right\}\right), \tag{98}$$

depending on whether jumping knowledge is applied or not.

Contents

1 Neural Tangent Kernels (NTK)
- Contruction of the NTK due to [ADH+19]
- Convergence to the NTK at Initialization [ADH+19]
- Equivalence between NN and NTK

2 Graph Neural Nets (GNN)

3 Graph Neural Tangent Kernel (GNTK)
- Overview
- Formulae
- Theoretical Analyses
  - Proof of Theorem 4.2 [DHP+19]
  - Proof of Theorem 4.3 [DHP+19]

Contents

1. Neural Tangent Kernels (NTK)
   - Contruction of the NTK due to [ADH+19]
   - Convergence to the NTK at Initialization [ADH+19]
   - Equivalence between NN and NTK
2. Graph Neural Nets (GNN)
3. Graph Neural Tangent Kernel (GNTK)
   - Overview
   - Formulae
   - Theoretical Analyses
     - Proof of Theorem 4.2 [DHP+19]
     - Proof of Theorem 4.3 [DHP+19]

GNTK: Problem Formulation

- Let $f(\theta, G) \in \mathbb{R}$ be the output of the corresponding GNN under parameters θ and input graph G, for two given graphs G and G', to calculate the corresponding GNTK value, we need to calculate the expected value of

$$\left\langle \frac{\partial f(\theta, G)}{\partial \theta}, \frac{\partial f(\theta, G')}{\partial \theta} \right\rangle \tag{99}$$

  in the limit that $m \to \infty$ and θ are all Gaussian random variables, which can be viewed as a Gaussian process.

- For each layer in the GNN, denote by $\mathbf{\Sigma}$ the covariance matrix of outputs of that layer and $\dot{\mathbf{\Sigma}}$ the covariance matrix corresponds to the derivative of that layer.

- These covariance matrices can be calculated via dynamic programming.

Contents

1 Neural Tangent Kernels (NTK)
- Contruction of the NTK due to [ADH+19]
- Convergence to the NTK at Initialization [ADH+19]
- Equivalence between NN and NTK

2 Graph Neural Nets (GNN)

3 Graph Neural Tangent Kernel (GNTK)
- Overview
- Formulae
- Theoretical Analyses
  - Proof of Theorem 4.2 [DHP+19]
  - Proof of Theorem 4.3 [DHP+19]

GNTK: Formulae

Setup

- Given two graphs $G = (V, E), G' = (V', E')$ with $|V| = n$, $|V'| = n'$ and a GNN with L BLOCK operations and R fully-connected layers with ReLU activation in each BLOCK operation.
- First define the covariance matrix between input features of two input graphs G, G', $\boldsymbol{\Sigma}^{(0)}(G, G') \in \mathbb{R}^{n \times n'}$.
- For two nodes $u \in V$ and $u' \in V'$, $\left[\boldsymbol{\Sigma}^{(0)}(G, G')\right]_{uu'}$ is defined to be $\boldsymbol{h}_u^\top \boldsymbol{h}_{u'}$, where $\boldsymbol{h}_u$ and $\boldsymbol{h}_{u'}$ are the input features of $u \in V$ and $u' \in V'$.

GNTK: Formulae for BLOCK Operation (1/3)

- A BLOCK operation in GNTK calculates a covariance matrix $\mathbf{\Sigma}^{(\ell)}_{(R)}(G,G') \in \mathbb{R}^{n\times n'}$ using $\mathbf{\Sigma}^{(\ell-1)}_{(R)}(G,G') \in \mathbb{R}^{n\times n'}$, and calculates intermediate kernel values $\mathbf{\Theta}^{(\ell)}_{(r)}(G,G') \in \mathbb{R}^{n\times n'}$.
- First perform a neighborhood aggregation operation

$$\begin{aligned}\left[\mathbf{\Sigma}^{(\ell)}_{(0)}(G,G')\right]_{uu'} =& c_u c_{u'} \sum_{v\in\mathcal{N}(u)\cup\{u\}} \sum_{v'\in\mathcal{N}(u')\cup\{u'\}} \left[\mathbf{\Sigma}^{(\ell-1)}_{(R)}(G,G')\right]_{vv'},\\ \left[\mathbf{\Theta}^{(\ell)}_{(0)}(G,G')\right]_{uu'} =& c_u c_{u'} \sum_{v\in\mathcal{N}(u)\cup\{u\}} \sum_{v'\in\mathcal{N}(u')\cup\{u'\}} \left[\mathbf{\Theta}^{(\ell-1)}_{(R)}(G,G')\right]_{vv'}.\end{aligned} \tag{100}$$

  Here define $\mathbf{\Sigma}^{(0)}_{(R)}(G,G')$ and $\mathbf{\Theta}^{(0)}_{(R)}(G,G')$ as $\mathbf{\Sigma}^{(0)}(G,G')$.

GNTK: Formulae for BLOCK Operation (2/3)

- Next perform R transformations that correspond to the R fully-connected layers with ReLU activation. $\sigma(z) = \max\{0, z\}$, and $\dot{\sigma}(z) = \mathbf{1}[z \geq 0]$.
- For each $r \in [R]$, define
  - For $u \in V, u' \in V'$,

$$\begin{aligned}&\left[\boldsymbol{A}_{(r)}^{(\ell)}(G, G')\right]_{uu'} \\ &= \begin{pmatrix} \left[\boldsymbol{\Sigma}_{(r-1)}^{(\ell)}(G, G)\right]_{u,u} & \left[\boldsymbol{\Sigma}_{(r-1)}^{(\ell)}(G, G')\right]_{uu'} \\ \left[\boldsymbol{\Sigma}_{(r-1)}^{(\ell)}(G', G)\right]_{uu'} & \left[\boldsymbol{\Sigma}_{(r-1)}^{(\ell)}(G', G')\right]_{u'u'} \end{pmatrix} \in \mathbb{R}^{2\times 2}. \end{aligned} \tag{101}$$

GNTK: Formulae for BLOCK Operation (3/3)

- For $u \in V, u' \in V'$,

$$\begin{aligned}\left[\boldsymbol{\Sigma}_{(r)}^{(\ell)}(G, G')\right]_{uu'} =& c_\sigma \mathbb{E}_{(a,b)\sim\mathcal{N}\left(\mathbf{0}, \left[\boldsymbol{A}_{(r)}^{(\ell)}(G,G')\right]_{uu'}\right)} \left[\sigma(a)\,\sigma(b)\right], \\ \left[\dot{\boldsymbol{\Sigma}}_{(r)}^{(\ell)}(G, G')\right]_{uu'} =& c_\sigma \mathbb{E}(a,b) \sim \mathcal{N}\left(\mathbf{0}, \left[\boldsymbol{A}_{(r)}^{(\ell)}(G, G')\right]_{uu'}\right) \left[\dot{\sigma}(a)\dot{\sigma}(b)\right].\end{aligned} \tag{102}$$

- For $u \in V, u' \in V'$,

$$\begin{aligned}\left[\boldsymbol{\Theta}_{(r)}^{(\ell)}(G, G')\right]_{uu'} =& \left[\boldsymbol{\Theta}_{(r-1)}^{(\ell)}(G, G')\right]_{uu'} \left[\dot{\boldsymbol{\Sigma}}_{(r)}^{(\ell)}(G, G')\right]_{uu'} \\ &+ \left[\boldsymbol{\Sigma}_{(r)}^{(\ell)}(G, G')\right]_{uu'}.\end{aligned} \tag{103}$$

GNTK: Formulae for READOUT Operation

- Given these intermediate outputs, the final output of GNTK is

$$\Theta(G, G') = \begin{cases} \sum_{u \in V, u' \in V'} \left[\mathbf{\Theta}_{(R)}^{(\ell)} (G, G') \right]_{uu'} & \text{w/o jumping,} \\ \sum_{u \in V, u' \in V'} \left[\sum_{\ell=0}^{L} \mathbf{\Theta}_{(R)}^{(\ell)} (G, G') \right]_{uu'} & \text{w/ jumping.} \end{cases} . \tag{104}$$

GNTK: An Example

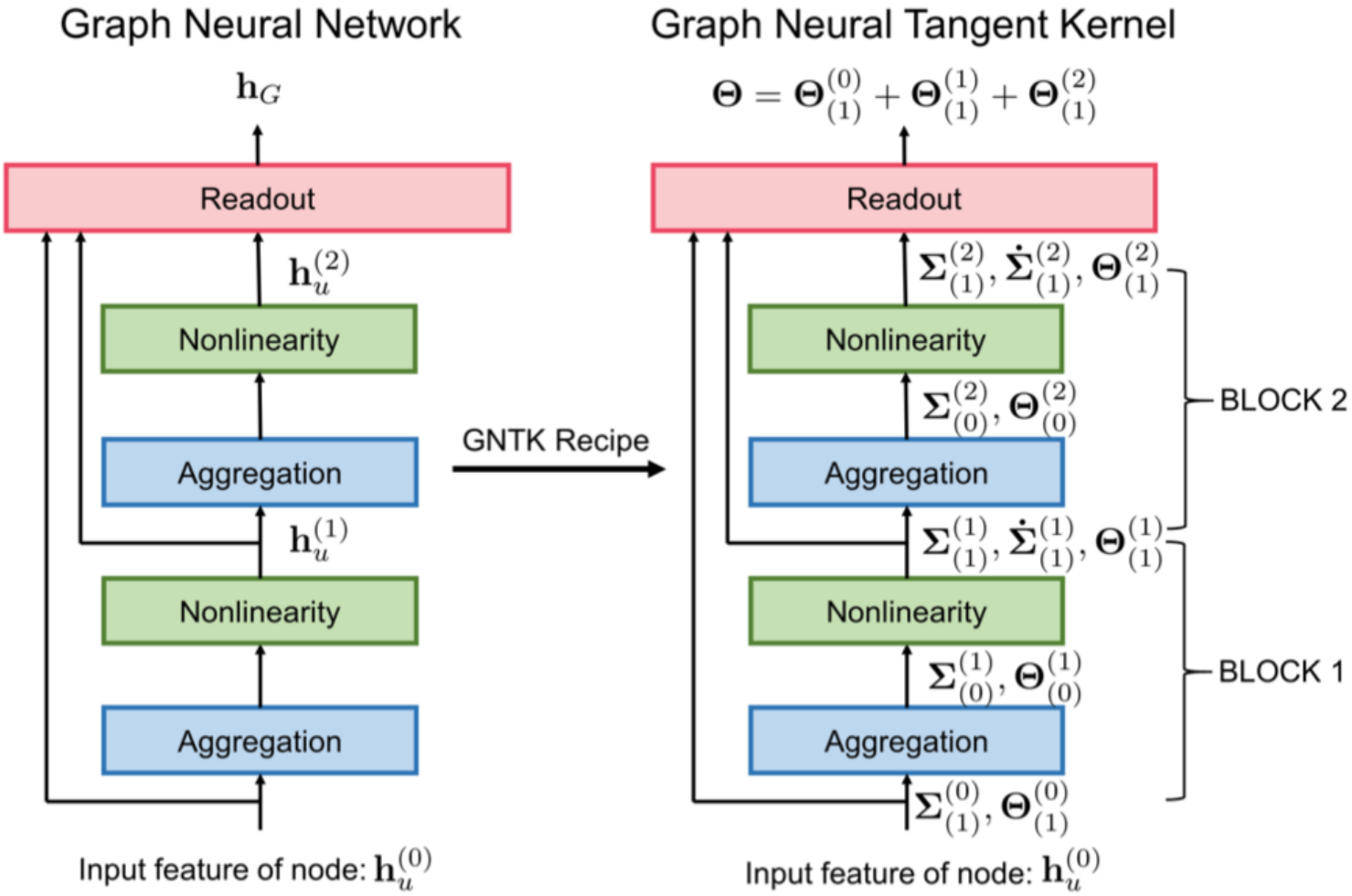


Figure: Exapmle of GNTK from a GNN.

GNTK: An Example

- Consider a GNN with $L = 2$ BLOCKs, $R = 1$ fully-connected layer in each BLOCK, and jumping knowledge.
- For two graphs G and G', first calculate

$$\left[\boldsymbol{\Theta}^{(0)}_{(1)}(G, G')\right]_{uu'} = \left[\boldsymbol{\Sigma}^{(0)}_{(1)}(G, G')\right]_{uu'} = \left[\boldsymbol{\Sigma}^{(0)}(G, G')\right]_{uu'} = \boldsymbol{h}_u^\top \boldsymbol{h}_{u'}. \tag{105}$$

- Then follow the kernel formulae to calculate $\boldsymbol{\Sigma}^{(\ell)}_{(0)}, \boldsymbol{\Theta}^{(\ell)}_{(0)}$ using $\boldsymbol{\Sigma}^{(\ell-1)}_{(R)}, \boldsymbol{\Theta}^{(\ell-1)}_{(R)}$ (Aggregation) and calculate $\boldsymbol{\Sigma}^{(\ell)}_{(r)}, \dot{\boldsymbol{\Sigma}}^{(\ell)}_{(r)}, \boldsymbol{\Theta}^{(\ell)}_{(r)}$ using $\boldsymbol{\Sigma}^{(\ell)}_{(r-1)}, \boldsymbol{\Theta}^{(\ell)}_{(r-1)}$ (Nonlinearity).
- The final output is

$$\Theta(G, G') = \sum_{u \in V, u' \in V'} \left[\sum_{\ell=0}^{L} \boldsymbol{\Theta}^{(\ell)}_{(R)}(G, G')\right]_{uu'}. \tag{106}$$

Contents

1. Neural Tangent Kernels (NTK)
   - Contruction of the NTK due to [ADH+19]
   - Convergence to the NTK at Initialization [ADH+19]
   - Equivalence between NN and NTK
2. Graph Neural Nets (GNN)
3. Graph Neural Tangent Kernel (GNTK)
   - Overview
   - Formulae
   - Theoretical Analyses
     - Proof of Theorem 4.2 [DHP+19]
     - Proof of Theorem 4.3 [DHP+19]

GNTK: Generalization (1/2)

- Given n training data $\{(G_i, y_i)\}_{i=1}^n$ drawn i.i.d. from the underlying distribution $\mathcal{D}$, where G_i is the i-th input graph and y_i is its label.
- Consider a GNN with a single BLOCK operation, followed by the READOUT operation (without jumping knowledge).
- Set $c_u = \left(\left\|\sum_{v \in \mathcal{N}(u) \cup \{u\}} \boldsymbol{h}_v\right\|_2\right)^{-1}$, and $\boldsymbol{\Theta} \in \mathbb{R}^{n \times n}$ to be the kernel matrix, where $\boldsymbol{\Theta}_{ij} = \Theta(G_i, G_j)$.

GNTK: Generalization (2/2)

- Assume the kernel matrix $\mathbf{\Theta} \in \mathbb{R}^{n \times n}$ is invertible.
- For a testing point G_{te}, the prediction of kernel regression using GNTK on this testing point is

$$f_{ker}(G_{te}) = [\Theta(G_{te}, G_1), \Theta(G_{te}, G_1), \ldots, \Theta(G_{te}, G_n)]^\top \mathbf{\Theta}^{-1} \boldsymbol{y}. \tag{107}$$

GNTK Theorem 4.1 [DHP+19] (1/1)

Theorem 4.1 (Generalization bound of kernel regression [BM02].)

Given n training data $\{(G_i, y_i)\}_{i=1}^n$ drawn i.i.d. from the underlying distribution $\mathcal{D}$. Consider any loss function $\ell : \mathbb{R} \times \mathbb{R} \to [0, 1]$ that is 1-Lipschitz in the first argument such that $\ell(y, y) = 0$. With probability at least $1 - \delta$, the population loss of the GNTK predictor can be upper bounded by

$$\begin{aligned} L_{\mathcal{D}}\left(f_{ker}\right) &= \boldsymbol{E}_{(G,y)\sim\mathcal{D}}\left[\ell(f_{ker}(G), y)\right] \\ &= \leq O\left(\frac{\sqrt{\boldsymbol{y}^\top \boldsymbol{\Theta}^{-1}\boldsymbol{y} \cdot \operatorname{tr}\left(\boldsymbol{\Theta}\right)}}{n} + \sqrt{\frac{\log(1/\delta)}{n}}\right). \end{aligned} \tag{108}$$

- Using the above theorem, if one can bound $\boldsymbol{y}^\top \boldsymbol{\Theta}^{-1}\boldsymbol{y}$ and $\operatorname{tr}\left(\boldsymbol{\Theta}\right)$, then one can obtain a sample complexity bound.

GNTK Theorem 4.2 [DHP+19] (1/1)

Theorem 4.2 (An upper bound of $\boldsymbol{y}^\top \boldsymbol{\Theta}^{-1}\boldsymbol{y}$)

For each $i \in [n]$, if the labels $\{y_i\}_{i=1}^n$ satisfy

$$y_i = \alpha_1 \sum_{u\in V} \left(\overline{\boldsymbol{h}}_u^\top \boldsymbol{\beta}_1\right) + \sum_{l=1}^{\infty} \alpha_{2l} \sum_{u\in V} \left(\overline{\boldsymbol{h}}_u^\top \boldsymbol{\beta}_{2l}\right)^{2l}, \tag{109}$$

where $\overline{\boldsymbol{h}}_u = c_u \sum_{v\in\mathcal{N}(u)\cup\{u\}} \boldsymbol{h}_v$, $\alpha_1, \alpha_2, \alpha_4, \ldots \in \mathbb{R}$, $\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{\beta}_4, \ldots \in \mathbb{R}^d$, and $G_i = (V, E)$, then we have

$$\sqrt{\boldsymbol{y}^\top \boldsymbol{\Theta}^{-1}\boldsymbol{y}} \le 2|\alpha_1| \cdot \|\boldsymbol{\beta}_1\|_2 + \sum_{l=1}^{\infty} \sqrt{2\pi}(2l-1)|\alpha_{2l}| \cdot \|\boldsymbol{\beta}_{2l}\|_2^{2l}. \tag{110}$$

Proof of Theorem 4.2 [DHP+19] (1/10)

For two graph G and G', the GNTK kernel function that corresponds to the simple GNN can be described as

$$\begin{aligned}&\Theta(G,G')\\&=\sum_{u\in V,u'\in V'}\Big(\underbrace{\left[\boldsymbol{\Sigma}^{(1)}_{(0)}(G,G')\right]_{uu'}\left[\dot{\boldsymbol{\Sigma}}^{(1)}_{(1)}(G,G')\right]_{uu'}}_{\text{Theta\_1}}+\left[\boldsymbol{\Sigma}^{(1)}_{(1)}(G,G')\right]_{uu'}\Big).\end{aligned}\tag{111}$$

Here, we have

$$\left[\boldsymbol{\Sigma}^{(1)}_{(0)}(G,G')\right]_{uu'}=c_uc_{u'}\left(\sum_{v\in\mathcal{N}(u)\cup\{u\}}\boldsymbol{h}_v\right)^\top\left(\sum_{v'\in\mathcal{N}(u')\cup\{u'\}}\boldsymbol{h}_{v'}\right)=\overline{\boldsymbol{h}}_u^\top\overline{\boldsymbol{h}}_{u'}.\tag{112}$$

Proof of Theorem 4.2 [DHP+19] (2/10)

Recall that

$$\begin{aligned}\left[\boldsymbol{\Sigma}_{(r)}^{(\ell)}(G,G')\right]_{uu'} =& c_\sigma \mathbb{E}_{(a,b)\sim\mathcal{N}\left(\mathbf{0},\left[\boldsymbol{A}_{(r)}^{(\ell)}(G,G')\right]_{uu'}\right)}\left[\sigma(a)\,\sigma(b)\right],\\ \left[\dot{\boldsymbol{\Sigma}}_{(r)}^{(\ell)}(G,G')\right]_{uu'} =& c_\sigma \mathbb{E}_{(a,b)\sim\mathcal{N}\left(\mathbf{0},\left[\boldsymbol{A}_{(r)}^{(\ell)}(G,G')\right]_{uu'}\right)}\left[\dot\sigma(a)\dot\sigma(b)\right]\end{aligned} \tag{113}$$

and

$$\left[\boldsymbol{A}_{(r)}^{(\ell)}(G,G')\right]_{uu'} = \begin{pmatrix}\left[\boldsymbol{\Sigma}_{(r-1)}^{(\ell)}(G,G)\right]_{u,u} & \left[\boldsymbol{\Sigma}_{(r-1)}^{(\ell)}(G,G')\right]_{uu'}\\ \left[\boldsymbol{\Sigma}_{(r-1)}^{(\ell)}(G',G)\right]_{uu'} & \left[\boldsymbol{\Sigma}_{(r-1)}^{(\ell)}(G',G')\right]_{u'u'}\end{pmatrix} \in \mathbb{R}^{2\times 2}. \tag{114}$$

Proof of Theorem 4.2 [DHP+19] (3/10)

Since $\sigma(z) = \max\{0, z\}$ is the ReLU activation function, and $\dot{\sigma}(z) = \mathbb{1}[z \geq 0]$ is the derivative of the ReLU activation function, and $\|\overline{\boldsymbol{h}}_u\|_2 = 1$ for all nodes u, by calculation, we have

$$\left[\dot{\boldsymbol{\Sigma}}^{(1)}_{(1)}(G, G')\right]_{uu'} = \frac{\pi - \arccos\left(\left[\boldsymbol{\Sigma}^{(1)}_{(0)}(G, G')\right]_{uu'}\right)}{2\pi},$$

$$\left[\boldsymbol{\Sigma}^{(1)}_{(1)}(G, G')\right]_{uu'}$$

$$= \frac{\pi - \arccos\left(\left[\boldsymbol{\Sigma}^{(1)}_{(0)}(G, G')\right]_{uu'}\right) + \sqrt{1 - \left[\boldsymbol{\Sigma}^{(1)}_{(0)}(G, G')\right]^2_{uu'}}}{2\pi}. \tag{115}$$

Proof of Theorem 4.2 [DHP+19] (4/10)

Since

$$\arcsin(x) = \sum_{l=0}^{\infty} \frac{(2l-1)!!}{(2l)!!} \cdot \frac{x^{2l+1}}{2l+1}, \tag{116}$$

we have

$$\begin{aligned}
&\left[\boldsymbol{\Sigma}_{(0)}^{(1)}(G,G')\right]_{uu'} \left[\dot{\boldsymbol{\Sigma}}_{(1)}^{(1)}(G,G')\right]_{uu'} \\
=&\frac{1}{4}\left[\boldsymbol{\Sigma}_{(0)}^{(1)}(G,G')\right]_{uu'} + \frac{1}{2\pi}\left[\boldsymbol{\Sigma}_{(0)}^{(1)}(G,G')\right]_{uu'} \arcsin\left(\left[\boldsymbol{\Sigma}_{(0)}^{(1)}(G,G')\right]_{uu'}\right) \\
=&\frac{1}{4}\left[\boldsymbol{\Sigma}_{(0)}^{(1)}(G,G')\right]_{uu'} + \frac{1}{2\pi}\sum_{l=1}^{\infty} \frac{(2l-3)!!}{(2l-2)!!\cdot(2l-1)} \cdot \left[\boldsymbol{\Sigma}_{(0)}^{(1)}(G,G')\right]_{uu'}^{2l} \\
=&\frac{1}{4}\overline{\boldsymbol{h}}_u^\top \overline{\boldsymbol{h}}_{u'} + \frac{1}{2\pi}\sum_{l=1}^{\infty} \frac{(2l-3)!!}{(2l-2)!!\cdot(2l-1)} \cdot \left(\overline{\boldsymbol{h}}_u^\top \overline{\boldsymbol{h}}_{u'}\right)^{2l}.
\end{aligned} \tag{117}$$

Proof of Theorem 4.2 [DHP+19] (5/10)

Let $\boldsymbol{\Phi}^{(2l)}(\cdot)$ be the feature map of the polynomial kernel of degree $2l$, i.e.,

$$k^{(2l)}(\boldsymbol{x}, \boldsymbol{y}) = \left(\boldsymbol{x}^\top \boldsymbol{y}\right)^{2l} = \boldsymbol{\Phi}^{(2l)}(\boldsymbol{x})^\top \boldsymbol{\Phi}^{(2l)}(\boldsymbol{y}). \tag{118}$$

We have

$$\begin{aligned} &\left[\boldsymbol{\Sigma}^{(1)}_{(0)}(G, G')\right]_{uu'} \left[\dot{\boldsymbol{\Sigma}}^{(1)}_{(1)}(G, G')\right]_{uu'} \\ =&\frac{1}{4}\overline{\boldsymbol{h}}_u^\top \overline{\boldsymbol{h}}_{u'} + \frac{1}{2\pi}\sum_{l=1}^{\infty} \frac{(2l-3)!!}{(2l-2)!! \cdot (2l-1)} \cdot \left(\boldsymbol{\Phi}^{(2l)}(\overline{\boldsymbol{h}}_u)\right)^\top \boldsymbol{\Phi}^{(2l)}(\overline{\boldsymbol{h}}_{u'}). \end{aligned} \tag{119}$$

Let

Proof of Theorem 4.2 [DHP+19] (6/10)

$$\Theta_1(G,G') = \sum_{u\in V, u'\in V'} \left[\boldsymbol{\Sigma}^{(1)}_{(0)}(G,G')\right]_{uu'} \left[\dot{\boldsymbol{\Sigma}}^{(1)}_{(1)}(G,G')\right]_{uu'}, \tag{120}$$

we have

$$\begin{aligned}
&\Theta_1(G,G') \\
=&\frac{1}{4}\left(\sum_{u\in V}\overline{\boldsymbol{h}}_u\right)^\top\left(\sum_{u'\in V'}\overline{\boldsymbol{h}}_{u'}\right) + \frac{1}{2\pi}\sum_{l=1}^{\infty}\frac{(2l-3)!!}{(2l-2)!!\cdot(2l-1)}\cdot\left(\sum_{u\in V}\boldsymbol{\Phi}^{(2l)}(\overline{\boldsymbol{h}}_u)\right)
\end{aligned} \tag{121}$$

Proof of Theorem 4.2 [DHP+19] (7/10)

Since $\mathbf{\Theta} = \mathbf{\Theta}_1 + \mathbf{\Theta}_2$ where $\mathbf{\Theta}_2$ is a kernel matrix (and thus positive semi-definite), for any $y \in \mathbb{R}^n$, we have

$$\boldsymbol{y}^\top \mathbf{\Theta}^{-1} \boldsymbol{y} \leq \boldsymbol{y}^\top \mathbf{\Theta}_1^{-1} \boldsymbol{y}. \tag{122}$$

JY: I was not able to prove this yet.

Proof of Theorem 4.2 [DHP+19] (8/10)

Recall that

$$y_i = \alpha_1 \sum_{u \in V} \left(\overline{\boldsymbol{h}}_u^\top \boldsymbol{\beta}_1 \right) + \sum_{l=1}^{\infty} \alpha_{2l} \sum_{u \in V} \left(\overline{\boldsymbol{h}}_u^\top \boldsymbol{\beta}_{2l} \right)^{2l}.$$

We rewrite

$$y_i = y_i^{(0)} + \sum_{l=1}^{\infty} y_i^{(2l)}, \tag{123}$$

where

$$y_i^{(0)} = \alpha_1 \left(\sum_{u \in V} \overline{\boldsymbol{h}}_u \right)^\top \boldsymbol{\beta}_1, \tag{124}$$

and for each $l \geq 1$,

Proof of Theorem 4.2 [DHP+19] (9/10)

$$
\begin{aligned}
y_i^{(2l)} &= \alpha_{2l} \sum_{u \in V} \left(\overline{\boldsymbol{h}}_u^\top \boldsymbol{\beta}_{2l} \right)^{2l} = \alpha_{2l} \sum_{u \in V} \left(\boldsymbol{\Phi}^{2l} \left(\overline{\boldsymbol{h}}_u \right) \right)^\top \boldsymbol{\Phi}^{2l} \left(\boldsymbol{\beta}_{2l} \right) \\
&= \alpha_{2l} \left(\sum_{u \in V} \boldsymbol{\Phi}^{2l} \left(\overline{\boldsymbol{h}}_u \right) \right)^\top \boldsymbol{\Phi}^{2l} \left(\boldsymbol{\beta}_{2l} \right).
\end{aligned} \tag{125}
$$

We have

$$
\boldsymbol{y} = \boldsymbol{y}^{(0)} + \sum_{l=1}^{\infty} \boldsymbol{y}^{(2l)}. \tag{126}
$$

Thus,

$$
\sqrt{\boldsymbol{y}^\top \boldsymbol{\Theta}^{-1} \boldsymbol{y}} \leq \sqrt{\boldsymbol{y}^\top \boldsymbol{\Theta}_1^{-1} \boldsymbol{y}} \leq \sqrt{\left(\boldsymbol{y}^{(0)} \right)^\top \boldsymbol{\Theta}_1^{-1} \boldsymbol{y}^{(0)}} + \sum_{l=1}^{\infty} \sqrt{\left(\boldsymbol{y}^{(2l)} \right)^\top \boldsymbol{\Theta}_1^{-1} \boldsymbol{y}^{(2l)}}. \tag{127}
$$

Proof of Theorem 4.2 [DHP+19] (10/10)

When $l = 0$, we have

$$\sqrt{(\boldsymbol{y}^0)^\top \boldsymbol{\Theta}_1^{-1} \boldsymbol{y}^0} \leq 2|\alpha_1| \|\boldsymbol{\beta}_1\|_2. \tag{128}$$

When $l \geq 1$, we have

$$\sqrt{(\boldsymbol{y}^{2l})^\top \boldsymbol{\Theta}_1^{-1} \boldsymbol{y}^{2l}} \leq \sqrt{2\pi}(2l-1)|\alpha_{2l}| \left\| \boldsymbol{\Phi}^{2l}(\boldsymbol{\beta}_{2l}) \right\|_2. \tag{129}$$

Notice that

$$\left\| \boldsymbol{\Phi}^{2l}(\boldsymbol{\beta}_{2l}) \right\|_2^2 = \left(\boldsymbol{\Phi}^{2l}(\boldsymbol{\beta}_{2l}) \right)^\top \boldsymbol{\Phi}^{2l}(\boldsymbol{\beta}_{2l}) = \|\boldsymbol{\beta}_{2l}\|_2^{4l}. \tag{130}$$

Thus,

$$\sqrt{\boldsymbol{y}^\top \boldsymbol{\Theta}^{-1} \boldsymbol{y}} \leq 2|\alpha_1| \|\boldsymbol{\beta}_1\|_2 + \sum_{l=1}^{\infty} \sqrt{2\pi}(2l-1)|\alpha_{2l}| \|\boldsymbol{\beta}_{2l}\|_2^{2l}. \tag{131}$$

GNTK Theorem 4.3 [DHP+19]

Theorem 4.3 (An upper bound of $\mathrm{tr}(\mathbf{\Theta})$)

If for all graphs $G_i = (V_i, E_i)$ in the training set, $|V_i|$ is upper bounded by $\overline{V}$, then $\mathrm{tr}(\mathbf{\Theta}) \leq O(n\overline{V}^2)$. Here n is the number of training samples.

$=$

Proof of Theorem 4.3 [DHP+19] (1/4)

Recall that

$$\begin{aligned}&\Theta(G,G')\\ =&\sum_{u\in V,u'\in V'}\left(\left[\boldsymbol{\Sigma}^{(1)}_{(0)}(G,G')\right]_{uu'}\left[\dot{\boldsymbol{\Sigma}}^{(1)}_{(1)}(G,G')\right]_{uu'}+\left[\boldsymbol{\Sigma}^{(1)}_{(1)}(G,G')\right]_{uu'}\right),\end{aligned}\tag{132}$$

where

$$\left[\boldsymbol{\Sigma}^{(1)}_{(0)}(G,G')\right]_{uu'}=c_uc_{u'}\left(\sum_{v\in\mathcal{N}(u)\cup\{u\}}\boldsymbol{h}_v\right)^\top\left(\sum_{v'\in\mathcal{N}(u')\cup\{u'\}}\boldsymbol{h}_{v'}\right)=\overline{\boldsymbol{h}}_u^\top\overline{\boldsymbol{h}}_{u'}\tag{133}$$

Proof of Theorem 4.3 [DHP+19] (2/4)

and

$$\begin{aligned}
\left[\dot{\mathbf{\Sigma}}^{(1)}_{(1)}(G,G')\right]_{uu'} &= \frac{\pi - \arccos\left(\left[\mathbf{\Sigma}^{(1)}_{(0)}(G,G')\right]_{uu'}\right)}{2\pi}, \\
\left[\mathbf{\Sigma}^{(1)}_{(1)}(G,G')\right]_{uu'} & \\
= \frac{\pi - \arccos\left(\left[\mathbf{\Sigma}^{(1)}_{(0)}(G,G')\right]_{uu'}\right) + \sqrt{1 - \left[\mathbf{\Sigma}^{(1)}_{(0)}(G,G')\right]^2_{uu'}}}{2\pi}&.
\end{aligned} \tag{134}$$

Proof of Theorem 4.3 [DHP+19] (3/4)

Since for each node u, $\overline{\boldsymbol{h}}_u = c_u \sum_{v \in \mathcal{N}(u) \cup \{u\}} \boldsymbol{h}_v$, and $c_u = \left(\left\| \sum_{v \in \mathcal{N}(u) \cup \{u\}} \boldsymbol{h}_v \right\|_2 \right)^{-1}$, we have $\|\overline{\boldsymbol{h}}_u\|_2 = 1$. Moreover,

$$\left[\dot{\boldsymbol{\Sigma}}^{(1)}_{(1)}(G, G') \right]_{uu'} = \frac{\pi - \arccos\left(\left[\boldsymbol{\Sigma}^{(1)}_{(0)}(G, G') \right]_{uu'} \right)}{2\pi} \leq 1/2 \tag{135}$$

and

$$\begin{aligned} &\left[\boldsymbol{\Sigma}^{(1)}_{(1)}(G, G') \right]_{uu'} \\ =&\frac{\pi - \arccos\left(\left[\boldsymbol{\Sigma}^{(1)}_{(0)}(G, G') \right]_{uu'} \right) + \sqrt{1 - \left[\boldsymbol{\Sigma}^{(1)}_{(0)}(G, G') \right]^2_{uu'}}}{2\pi} \leq \frac{1 + \pi}{2\pi} \leq 1, \end{aligned} \tag{136}$$

Proof of Theorem 4.3 [DHP+19] (4/4)

we have

$$\begin{aligned}
&\Theta(G, G') \\
&\quad = \sum_{u \in V, u' \in V'} \left(\left[\boldsymbol{\Sigma}^{(1)}_{(0)}(G, G') \right]_{uu'} \left[\dot{\boldsymbol{\Sigma}}^{(1)}_{(1)}(G, G') \right]_{uu'} + \left[\boldsymbol{\Sigma}^{(1)}_{(1)}(G, G') \right]_{uu'} \right) \\
&\quad \leq \left(\frac{1}{2} + 1 \right) |V||V'| \\
&\quad \leq 2|V||V'|.
\end{aligned} \tag{137}$$

Thus,

$$\mathrm{tr}(\boldsymbol{\Theta}) \leq 2n\overline{V}^2. \tag{138}$$

GNTK: Remarks

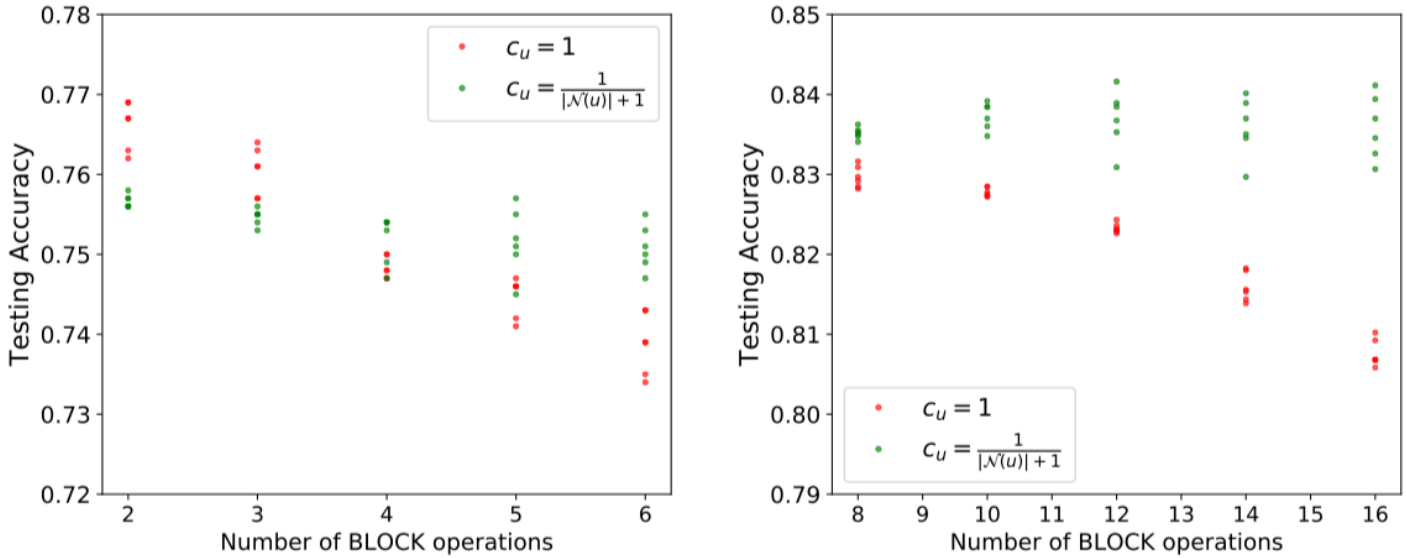
- Combining these three results, we know if

$$2|\alpha_1| \cdot \|\boldsymbol{\beta}_1\|_2 + \sum_{l=1}^{\infty} \sqrt{2\pi}(2l-1)|\alpha_{2l}| \cdot \|\boldsymbol{\beta}_{2l}\|_2^{2l} \tag{139}$$

  is bounded, and $|V_i|$ is bounded for all graphs $G_i = (V_i, E_i)$ in the training set, then the GNTK that corresponds to the simple GNN described above can learn functions of forms in (109), with polynomial number of samples.

GNTK: Experiments I

- Test accuracy is correlated with the dataset and architecture.

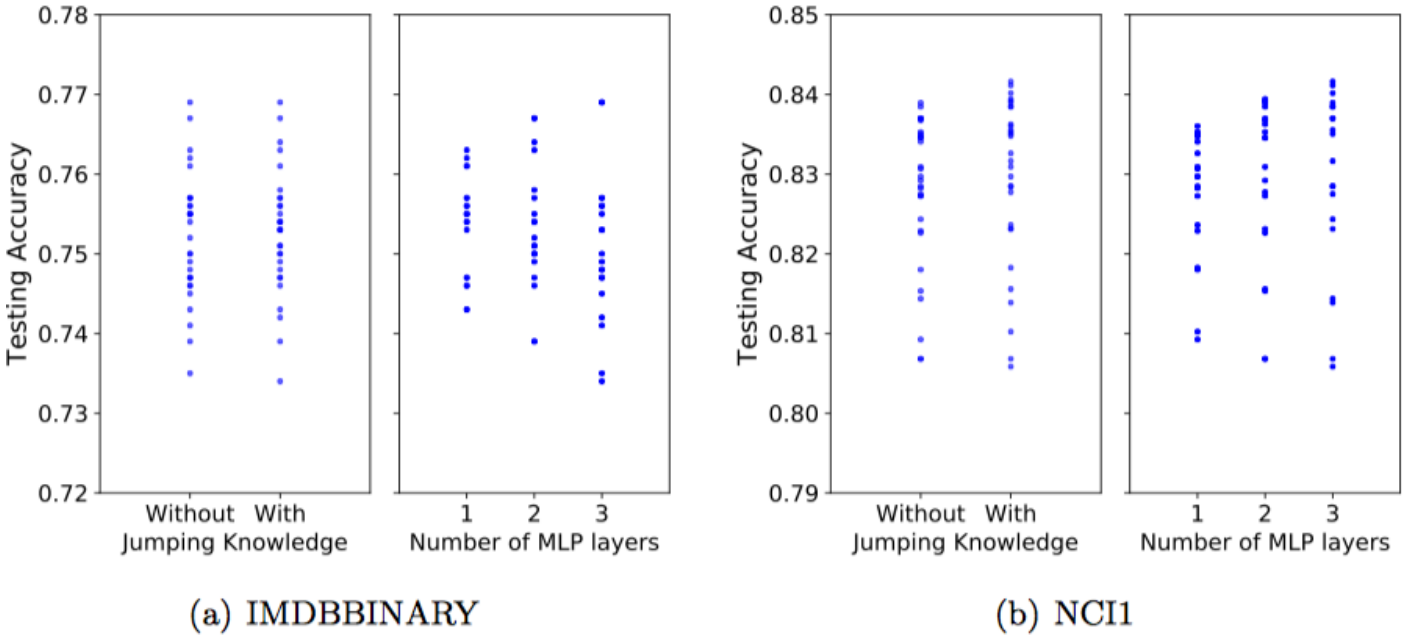


(a) IMDBBINARY (b) NCI1

Figure: Effect of number of BLOCK operations.

GNTK: Experiments II

- Jump knowledge is expected to improve performance.
- The authors of [DHP+19] observed a 0.8% improvement in accuracy.



(a) IMDBBINARY (b) NCI1

Figure: Effect of jump knowledge.

References (1/2)

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).

Stéphane Boucheron, Gábor Lugosi, and Pascal Massart, *Concentration inequalities: A nonasymptotic theory of independence*, Oxford university press, 2013.

Peter L Bartlett and Shahar Mendelson, *Rademacher and gaussian complexities: Risk bounds and structural results*, Journal of Machine Learning Research **3** (2002), no. Nov, 463–482.

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

Simon S. Du, Kangcheng Hou, Barnabás Póczos, Ruslan Salakhutdinov, Ruosong Wang, and Keyulu Xu, *Graph neural tangent kernel: Fusing graph neural networks with graph kernels*, CoRR **abs/1905.13192** (2019).

References (2/2)


Arthur Jacot, Franck Gabriel, and Clément Hongler, *Neural tangent kernel: Convergence and generalization in neural networks*, Advances in neural information processing systems, 2018, pp. 8571–8580.


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).


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