Deep Architecture Connectivity Matters for Its Convergence: A Fine-Grained Analysis

Wuyang Chen*
University of Texas at Austin

Wei Huang*
RIKEN AIP

Xinyu Gong
University of Texas at Austin

Boris Hanin
Princeton University

Zhangyang Wang
University of Texas at Austin

Abstract

Advanced deep neural networks (DNNs), designed by either human or AutoML algorithms, are growing increasingly complex. Diverse operations are connected by complicated connectivity patterns, e.g., various types of skip connections. Those topological compositions are empirically effective and observed to smooth the loss landscape and facilitate the gradient flow in general. However, it remains elusive to derive any principled understanding of their effects on the DNN capacity or trainability, and to understand why or in which aspect one specific connectivity pattern is better than another. In this work, we theoretically characterize the impact of connectivity patterns on the convergence of DNNs under gradient descent training in fine granularity. By analyzing a wide network's Neural Network Gaussian Process (NNGP), we are able to depict how the spectrum of an NNGP kernel propagates through a particular connectivity pattern, and how that affects the bound of convergence rates. As one practical implication of our results, we show that by a simple filtration on "unpromising" connectivity patterns, we can trim down the number of models to evaluate, and significantly accelerate the large-scale neural architecture search without any overhead. Code is available at: `https://github.com/VITA-Group/architecture_convergence`.

1 Introduction

Recent years have witnessed substantial progress in designing better deep neural network architectures. The common objective is to build networks that are easy to optimize, of superior trade-offs between efficiency and accuracy, and are generalizable to diverse tasks and datasets. When developing deep networks, how operations (linear transformations and non-linear activations) are connected and stacked together is vital, which is studied in network's convergence [18, 71, 75], complexity [47, 50, 21], generalization [14, 8, 61], loss landscapes [36, 20, 51], etc.

Although it has been widely observed that the performance of deep networks keeps being improved by advanced design options, our understanding remains limited on how a network's properties are influenced by its architectures. For example, in computer vision, design trends have shifted from vanilla chain-like stacked layers [33, 32, 53] to manually elaborated connectivity patterns (ResNet [25], DenseNet [28], etc.). While people observed smoothed loss landscapes [36], mitigated gradient vanishing problem [4], and better generalization [29], these findings only explain the effectiveness of adding skip connections in general, but barely lead to further "finer-grained" insight on more sophisticated composition of skip connections beyond ResNet. Recently, the AutoML community tries to relieve human efforts and propose to automatically discover novel networks from

*Equal Contribution.

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

gigantic architecture spaces [72, 46]. Despite the strong performance [57, 27, 62, 58], the searched architectures are often composed of highly complex (or even randomly wired) connections, leaving it challenging to analyze theoretically.

Understanding the principles of deep architecture connectivity is of significant importance. Scientifically, this helps answer why composing complicated skip connection patterns has been such an effective "trick" in improving deep networks' empirical performance. Practically, this has direct guidance in designing more efficient and expressive architectures. To close the gap between our theoretical understandings and practical architectures, in this work we target two concrete questions:

Q1: Can we understand the precise roles of different connectivity patterns in deep networks?

Q2: Can we summarize principles on how connectivity should be designed in deep networks?

One standard way to study how operations and connections affect the network is to analyze the model's convergence under gradient descent [19, 2, 18]. Here, we systematically study the relationship between the connectivity pattern of a neural network and the bound of its convergence rate. A deep architecture can be viewed as a directed acyclic computational graph (DAG), where feature maps are represented as nodes and operations in different layers are directed edges linking features. Under this formulation, by analyzing the spectrum of the Neural Network Gaussian Process (NNGP) kernel, we show that the bound of the convergence rate of the DAG is jointly determined by the number of unique paths in a DAG and the number of parameterized operations on each path. Note that, although several prior arts [44, 1, 10, 11] explored deep learning theories to predict the promise of architectures, their indicators were developed for the general deep network *functions*, not for fine-grained characterization for specific deep architecture *topology patterns*. Therefore, their correlations for selecting better architecture topologies are only *empirically observed*, but not *theoretically justified*. In contrast, our fine-grained conclusion is theory-ground and also empirically verified.

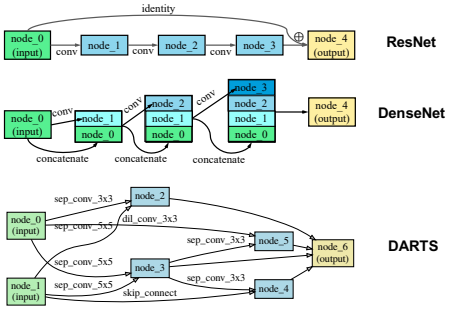

Figure 1: Connectivity patterns of deep networks are growingly more complex, including ResNet [25], DenseNet [28], and architectures sampled from the DARTS search space [40] commonly used in neural architecture search (NAS).

Based on this conclusion, we present two intuitive and practical principles for designing deep architecture DAGs: the "effective depth" $\bar{d}$ and the "effective width" $\bar{m}$. Experiments on diverse architecture benchmarks and datasets demonstrate that $\bar{d}$ and $\bar{m}$ can jointly distinguish promising connectivity patterns from unpromising ones. As a practical implication, our work also suggests a cheap "plug-and-play" method to accelerate the neural architecture search by filtering out potentially unpromising architectures at almost zero cost, before any gradient-based architecture evaluations. Our contributions are summarized below:

- We first theoretically analyze the convergence of gradient descent of diverse neural network architectures, and find the connectivity patterns largely impact their bound of convergence rate.
- From the theoretical analysis, we abstract two practical principles on designing the network's connectivity pattern: "effective depth" $\bar{d}$ and "effective width" $\bar{m}$.
- Both our convergence analysis and principles on effective depth/width are verified by experiments on diverse architectures and datasets. Our method can further significantly accelerate the neural architecture search without introducing any extra cost.

2 Related works

2.1 Global convergence of deep networks

Many works analyzed the convergence of networks trained with gradient descent [31, 68, 22, 9]. Convergence rates were originally explored for wide two-layer neural networks [60, 54, 55, 38, 19, 43, 3]. More recent works extended the analysis to deeper neural networks [2, 18, 42, 74, 71, 75, 30] and showed that over-parameterized networks can converge to the global minimum with random initialization if the width (number of neurons) is polynomially large as the number of training samples. In comparison, we expand such analysis to bridge the gap between theoretical understandings of general neural networks and practical selections of better "fine-grained" architectures.

2.2 Residual structures in deep networks

Architectures for computer vision have evolved from plain chain-like stacked layers [33, 32, 53] to elaborated connectivity patterns (ResNet [25], DenseNet [28], etc.). With the development of AutoML algorithms, novel networks of complicated operations/connections were discovered. Despite their strong performance [57, 27, 62, 58], these architectures are often composed of highly complex connections and are hard to analyze. [52] defined the depth and width for complex connectivity patterns, and studied the impacts of network topologies in different cases. To better understand these residual patterns, people tried to unify the architectures with different formulations. One seminal way is to represent network structures into graphs, then randomly sample different architectures from the graph distribution and empirically correlate their generalization with graph-related metrics [62, 67]. Other possible ways include geometric topology analysis [6], network science [5], etc. For example, [5] proposed NN-Mass by analyzing the network's layer-wise dynamic isometry, and then empirically linked to the network's convergence. However, none of those works directly connect the convergence rate analysis to different residual structures.

2.3 Theory-guided design of neural architectures

Particularly related to our work is an emerging research direction, that tries to connect recent deep learning theories to guide the design of novel network architectures. The main idea is to find theoretical indicators that have strong correlations with network's training or testing performance. [44] pioneered a training-free NAS method, which empirically leveraged sample-wise activation patterns to rank architectures. [45] leveraged the network's NNGP features to approximate its predictions. Different training-free indicators were evaluated in [1], and the "synflow" measure [59] was leveraged as the main ranking metric. [10] incorporated two theory-inspired metrics with supernet pruning as the search method. However, these works mainly adapted theoretical properties of the general deep neural network *function*: their correlations with the concrete network architecture *topology* are only *empirically observed*, but not *theoretically justified*.

3 Topology matters: convergence analysis with connectivity patterns

In this section, we study the convergence of gradient descent for deep networks, whose connectivity patterns can be formulated as small but representative direct acyclic graphs (DAGs). Our goal is to compare the convergence rate bounds of different DAGs, and further establish links to their connectivity patterns, leading to abstracting design principles.

3.1 Problem setup and architectures notations

We consider a network's DAG as illustrated in Figure 2. $\boldsymbol{X}^{(h)}$ $(h \in [0, H])$ is the feature map (node) and $\boldsymbol{W}$ is the operation (edge). $\boldsymbol{X}^{(0)}$ is the input node, $\boldsymbol{X}^{(H)}$ is the output node ($H = 3$ in this case), and $\boldsymbol{X}^{(1)}, \cdots, \boldsymbol{X}^{(H-1)}$ are intermediate nodes. The DAG is allowed to be fully connected: any two nodes could be connected by an operation. Feature maps from multiple edges coming to one vertex will be directly summed up. This DAG has many practical instances: for example, it is used in NAS-Bench-201 [17], a popular NAS benchmark. The forward process of the network in Figure 2 can be formulated as:

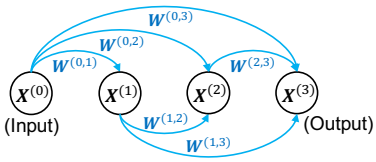

Figure 2: A network represented as a direct acyclic graph (DAG). $\boldsymbol{X}^{(h)}$ is the feature map (node). $\boldsymbol{W}$ is the operation (edge). $h \in [0, H]$, here $H = 3$. If we remove some edges or set some $\boldsymbol{W}$ as skip-connections (i.e., identity mappings), how would the convergence of this DAG be affected?

$$
\begin{aligned}
\boldsymbol{X}^{(1)} &= \rho(\boldsymbol{W}^{(0,1)}\boldsymbol{X}^{(0)}) \\
\boldsymbol{X}^{(2)} &= \rho(\boldsymbol{W}^{(0,2)}\boldsymbol{X}^{(0)}) + \rho(\boldsymbol{W}^{(1,2)}\boldsymbol{X}^{(1)}) \\
\boldsymbol{X}^{(3)} &= \rho(\boldsymbol{W}^{(0,3)}\boldsymbol{X}^{(0)}) + \rho(\boldsymbol{W}^{(1,3)}\boldsymbol{X}^{(1)}) + \rho(\boldsymbol{W}^{(2,3)}\boldsymbol{X}^{(2)}) \\
\boldsymbol{u} &= \boldsymbol{a}^{\top}\boldsymbol{X}^{(3)}.
\end{aligned}
\tag{1}
$$

Feature $\boldsymbol{X}^{(s)} \in \mathbb{R}^{m \times 1}$, where m is the absolute width of an edge (i.e. number of neurons), and $s \in \{1, 2, 3\}$. $\boldsymbol{a}$ is the final layer and $\boldsymbol{u}$ is the network's output. We consider three candidate operations for each edge: a linear transformation followed by a non-linear activation, or a skip-

connection (identity mapping), or a broken edge (zero mapping):

$$\boldsymbol{W}^{(s,t)} \begin{cases} = \mathbf{0} & \text{zero} \\ = \boldsymbol{I}^{m\times m} & \text{skip-connection} \\ \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}^{m\times m}) & \text{linear transformation} \end{cases}, \quad \rho(x) = \begin{cases} 0 & \text{zero} \\ 1 & \text{skip-connection} \\ \sqrt{\frac{c_\sigma}{m}}\sigma(x), & \text{linear transformation} \end{cases}. \tag{2}$$

$s, t \in \{1, 2, 3\}$, $\mathcal{N}$ is the Gaussian distribution, and σ is the activation function. $c_\sigma = \left(\mathbb{E}_{x\sim N(0,1)}\left[\sigma(x)^2\right]\right)^{-1}$ is a scaling factor to normalize the input in the initialization phase.

Consider the Neural Network Gaussian Process (NNGP) in the infinite-width limit [34], we define our NNGP variance as $\boldsymbol{K}_{ij}^{(s)} = \langle \boldsymbol{X}_i^{(s)}, \boldsymbol{X}_j^{(s)} \rangle$ and NNGP mean as $\boldsymbol{b}_i^{(s)} = \mathbb{E}[\boldsymbol{X}_i^{(s)}]$, $i, j \in 1, \cdots, N$ for N training samples in total. Both $\boldsymbol{K}_{ij}$ and $\boldsymbol{b}_i$ are taken the expectation over the weight distributions.

3.2 Preliminary: bounding the network's linear convergence rate via NNGP spectrum

Before we analyze different connectivity patterns (Section 3.3), we first give the linear convergence of our DAG networks (Theorem 3.1) and also show the guarantee of the full-rankness of $\lambda_{\min}(\boldsymbol{K}^{(H)})$ (Lemma 3.1). For a sufficiently wide neural network, its bound of convergence rate to the global minimum can be governed by the NNGP kernel. The linear convergence rate for a deep neural network of a DAG-like connectivity pattern is shown as follows:

Theorem 3.1 (Linear Convergence of DAG)**.** *Consider a DAG of H nodes and P_H end-to-end paths. At k-th gradient descent step on N training samples, with MSE loss $\mathcal{L}(k) = \frac{1}{2}\left\|\boldsymbol{y} - \boldsymbol{u}(k)\right\|_2^2$, suppose the learning rate $\eta = O\left(\frac{\lambda_{\min}(\boldsymbol{K}^{(H)})}{(NP_H)^2} 2^{O(H)}\right)$ and the number of neurons per layer $m = \Omega\left(\max\left\{\frac{(NP_H)^4}{\lambda_{\min}^4(\boldsymbol{K}^{(H)})}, \frac{NP_H H}{\delta}, \frac{(NP_H)^2 \log\left(\frac{HN}{\delta}\right) 2^{O(H)}}{\lambda_{\min}^2(\boldsymbol{K}^{(H)})}\right\}\right)$, we have*

$$\left\|\boldsymbol{y} - \boldsymbol{u}(k)\right\|_2^2 \leq \left(1 - \frac{\eta\lambda_{\min}(\boldsymbol{K}^{(H)})}{2}\right)^k \left\|\boldsymbol{y} - \boldsymbol{u}(0)\right\|_2^2, \tag{3}$$

where P_H is number of end-to-end paths from $\boldsymbol{X}^{(0)}$ to $\boldsymbol{X}^{(H)}$.

Remark 3.2. In our work, we will use a fixed small learning rate η across different network architectures, to focus our analysis on the impact of $\lambda_{\min}(\boldsymbol{K}^{(H)})$. This is motivated by the widely adopted setting in popular architecture benchmarks [48, 17] that we will follow in our experiments.

The above theorem shows that the convergence rate is bounded by the smallest eigenvalue of $\boldsymbol{K}^{(H)}$, and also indicates that networks of larger least eigenvalues will more likely to converge faster. This requires that the $\boldsymbol{K}^{(H)}$ has full rankness, which is demonstrated by the following lemma:

Lemma 3.1 (Full Rankness of $\boldsymbol{K}^{(H)}$)**.** *Suppose $\sigma(\cdot)$ is analytic and not a polynomial function. If no parallel data points, then $\lambda_{\min}\left(\boldsymbol{K}^{(H)}\right) > 0$.*

3.3 How does NNGP propagate through DAG?

Now we are ready to link the DAG's connectivity pattern to the bound of its convergence rate. Although different DAGs are all of the linear convergence under gradient descent, they are very likely to have different bounds of convergence rates. Finding the exact mapping from $\lambda_{\min}(\boldsymbol{K}^{(0)})$ to $\lambda_{\min}(\boldsymbol{K}^{(H)})$ will lead us to a fine-grained comparison between different connectivity patterns.

First, for fully-connected operations, we can obtain the propagation of the NNGP variance and mean between two consecutive layers:

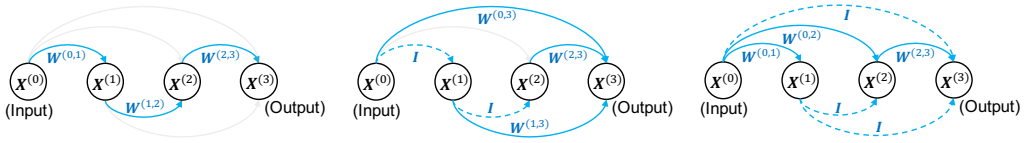

Figure 3: Three example DAGs. Solid blue arrows are parameterized operations (e.g. fully-connected layer). Dashed arrows are non-parameterized operations (e.g. skip-connections). Removed edges are in light grey. See Section 3.3 for their convergence analysis. Left: sequential connectivity (DAG#1). Middle: parallel connectivity (DAG#2). Right: mixed of sequential and parallel connectivity (DAG#3).

Lemma 3.2 (Propagation of $\boldsymbol{K}$)**.** *Define the propagation as* $\boldsymbol{K}^{(l)} = f(\boldsymbol{K}^{(l-1)})$ *and* $\boldsymbol{b}^{(l)} = g(\boldsymbol{b}^{(l-1)})$*. When the edge operation is a linear transformation, we have:*

$$\begin{aligned}
\boldsymbol{K}_{ii}^{(l)} &= f(\boldsymbol{K}_{ii}^{(l-1)}) = \int \mathcal{D}_z c_\sigma \sigma^2\left(\sqrt{\boldsymbol{K}_{ii}^{(l-1)}} z\right) \\
\boldsymbol{K}_{ij}^{(l)} &= f(\boldsymbol{K}_{ij}^{(l-1)}) = \int \mathcal{D}_{z_1}\mathcal{D}_{z_2} c_\sigma \sigma\left(\sqrt{\boldsymbol{K}_{ii}^{(l-1)}} z_1\right)\sigma\left(\sqrt{\boldsymbol{K}_{jj}^{(l-1)}}(\boldsymbol{C}_{ij}^{(l-1)} z_1 + \sqrt{1-(\boldsymbol{C}_{ij}^{(l-1)})^2} z_2)\right) \\
\boldsymbol{C}_{ij}^{(l)} &= \boldsymbol{K}_{ij}^{(l)} / \sqrt{\boldsymbol{K}_{ii}^{(l)} \boldsymbol{K}_{jj}^{(l)}} \\
\boldsymbol{b}_i^{(l)} &= g(\boldsymbol{b}_i^{(l-1)}) = \int \mathcal{D}_z \sqrt{c_\sigma} \sigma(\sqrt{\boldsymbol{K}_{ii}^{(l)}} z)
\end{aligned} \tag{4}$$

where z*,* z_1 *and* z_2 *are independent standard Gaussian random variables. Besides,* $\int \mathcal{D}_z = \frac{1}{\sqrt{2\pi}} \int dz e^{-\frac{1}{2}z^2}$ *is the measure for a normal distribution.*

Remark 3.3. Since $\boldsymbol{b}_i^{(l)}$ is a constant, it will not affect our analysis and we will omit it below.

Remark 3.4. With centered normalized inputs, $\boldsymbol{K}_{ii}^{(0)} = 1$.

Our last lemma to prepare states that we can bound the smallest eigenvalue of the NNGP kernel of $N \times N$ by its 2×2 principal submatrix case.

Lemma 3.3. *For a positive definite symmetric matrix* $\boldsymbol{K} \in \mathbb{R}^{N \times N}$*, the smallest eigenvalue is bounded by the smallest eigenvalue of its* 2×2 *principal sub-matrix.*

$$\lambda_{\min}(\boldsymbol{K}) \leq \min_{i \neq j} \lambda_{\min} \begin{bmatrix} \boldsymbol{K}_{ii} & \boldsymbol{K}_{ij} \\ \boldsymbol{K}_{ji} & \boldsymbol{K}_{jj} \end{bmatrix} \tag{5}$$

Now we can analyze the smallest eigenvalue $\lambda_{\min}(\boldsymbol{K}^{(H)})$ for different DAGs. Note that here we adopt ReLU activation for three reasons: 1) Analyzing ReLU with the initialization at the edge-of-chaos [23] exhibits promising results; 2) Previous works also imply that the convergence rate of ReLU-based networks depends on λ_0 [19]; 3) ReLU is the most commonly used activation in practice. Specifically, we set $c_\sigma = 2$ for ReLU [18, 23]. In Figure 3, we show three representative DAGs ($H = 3$). For a fair comparison, each DAG has three linear transformation layers, i.e., they have the same number of parameters and mainly differ in how nodes are connected. After these three case studies, we will show a more general rule for the propagation of $\boldsymbol{K}^{(0)}$.

DAG#1 (sequential connections). In this case, $\boldsymbol{W}^{(0,1)}, \boldsymbol{W}^{(1,2)}, \boldsymbol{W}^{(2,3)}$ = linear transformation, and $\boldsymbol{W}^{(0,2)}, \boldsymbol{W}^{(0,3)}, \boldsymbol{W}^{(1,2)}$ = zero (broken edge). There is only <u>one unique path</u> connecting from $\boldsymbol{X}^{(0)}$ to $\boldsymbol{X}^{(3)}$, with <u>three parameterized operations</u> on it.

$$\boldsymbol{K}^{(1)} = f(\boldsymbol{K}^{(0)}) \qquad \boldsymbol{K}^{(2)} = f(\boldsymbol{K}^{(1)}) \qquad \boldsymbol{K}^{(3)} = f(\boldsymbol{K}^{(2)}) \tag{6}$$

By Lemma 3.3, we calculate the upper bound of the least eigenvalue of $\boldsymbol{K}^{(3)}$, denoted as λ_{dag1}:

$$\begin{aligned}
&\lambda_{\min} \begin{bmatrix} \boldsymbol{K}_{ii}^{(3)} & \boldsymbol{K}_{ij}^{(3)} \\ \boldsymbol{K}_{ji}^{(3)} & \boldsymbol{K}_{jj}^{(3)} \end{bmatrix} = \lambda_{\min} \begin{bmatrix} \boldsymbol{K}_{ii}^{(2)} & f(\boldsymbol{K}_{ij}^{(2)}) \\ f(\boldsymbol{K}_{ji}^{(2)}) & \boldsymbol{K}_{jj}^{(2)} \end{bmatrix} = \lambda_{\min} \begin{bmatrix} \boldsymbol{K}_{ii}^{(1)} & f(f(\boldsymbol{K}_{ij}^{(1)})) \\ f(f(\boldsymbol{K}_{ji}^{(1)})) & \boldsymbol{K}_{jj}^{(1)} \end{bmatrix} \\
&= \lambda_{\min} \begin{bmatrix} \boldsymbol{K}_{ii}^{(0)} & f(f(f(\boldsymbol{K}_{ij}^{(0)}))) \\ f(f(f(\boldsymbol{K}_{j,i}^{(0)}))) & \boldsymbol{K}_{jj}^{(0)} \end{bmatrix} = 1 - f^3(\boldsymbol{K}_{ij}^{(0)}) \equiv \lambda_{\text{dag1}},
\end{aligned} \tag{7}$$

where we denote the function composition $f^3(\boldsymbol{K}_{ij}^{(0)}) = f(f(f(\boldsymbol{K}_{ij}^{(0)})))$.

DAG#2 (parallel connections). In this case, $\boldsymbol{W}^{(0,3)}, \boldsymbol{W}^{(1,3)}, \boldsymbol{W}^{(2,3)}$ = linear transformation, $\boldsymbol{W}^{(0,1)}, \boldsymbol{W}^{(1,2)}$ = skip-connection, and $\boldsymbol{W}^{(0,2)}$ = zero (broken edge). There are three unique paths connecting from $\boldsymbol{X}^{(0)}$ to $\boldsymbol{X}^{(3)}$, with one parameterized operation on each path.

$$\begin{aligned} \boldsymbol{K}^{(2)} &= \boldsymbol{K}^{(1)} = \boldsymbol{K}^{(0)} \\ \boldsymbol{K}^{(3)} &= f(\boldsymbol{K}^{(0)}) + f(\boldsymbol{K}^{(1)}) + f(\boldsymbol{K}^{(2)}) = 3f(\boldsymbol{K}^{(0)}) \end{aligned} \tag{8}$$

where $\boldsymbol{C}$ is a constant matrix with all entries being the same. Therefore:

$$\begin{aligned} \lambda_{\min} \begin{bmatrix} \boldsymbol{K}_{ii}^{(3)} & \boldsymbol{K}_{ij}^{(3)} \\ \boldsymbol{K}_{ji}^{(3)} & \boldsymbol{K}_{jj}^{(3)} \end{bmatrix} &= \lambda_{\min} \begin{bmatrix} 3\boldsymbol{K}_{ii}^{(0)} & 3f(\boldsymbol{K}_{ij}^{(0)}) \\ 3f(\boldsymbol{K}_{ji}^{(0)}) & 3\boldsymbol{K}_{jj}^{(0)} \end{bmatrix} \\ = 3\lambda_{\min} \begin{bmatrix} \boldsymbol{K}_{ii}^{(0)} & f(\boldsymbol{K}_{ij}^{(0)}) \\ f(\boldsymbol{K}_{ji}^{(0)}) & \boldsymbol{K}_{jj}^{(0)} \end{bmatrix} &= 3(1 - f(\boldsymbol{K}_{ij}^{(0)})) \equiv \lambda_{\text{dag2}}. \end{aligned} \tag{9}$$

DAG#3 (mixture of sequential and parallel connections). In this case, $\boldsymbol{W}^{(0,1)}, \boldsymbol{W}^{(0,2)}, \boldsymbol{W}^{(2,3)}$ = linear transformation, $\boldsymbol{W}^{(0,3)}, \boldsymbol{W}^{(1,2)}, \boldsymbol{W}^{(1,3)}$ = skip connection. There are four unique paths connecting from $\boldsymbol{X}^{(0)}$ to $\boldsymbol{X}^{(3)}$, with 0/1/2 parameterized operations on each path.

$$\begin{aligned} &\boldsymbol{K}^{(1)} = f(\boldsymbol{K}^{(0)}) \qquad \boldsymbol{K}^{(2)} = \boldsymbol{K}^{(1)} + f(\boldsymbol{K}^{(0)}) \\ &\boldsymbol{K}^{(3)} = f(\boldsymbol{K}^{(2)}) + \boldsymbol{K}^{(1)} + \boldsymbol{K}^{(0)} = \boldsymbol{K}^{(0)} + f(\boldsymbol{K}^{(0)}) + f(2f(\boldsymbol{K}^{(0)})) \end{aligned} \tag{10}$$

Then we have:

$$\lambda_{\min} \begin{bmatrix} \boldsymbol{K}_{ii}^{(3)} & \boldsymbol{K}_{ij}^{(3)} \\ \boldsymbol{K}_{j,i}^{(3)} & \boldsymbol{K}_{jj}^{(3)} \end{bmatrix} = \lambda_{\min} \begin{bmatrix} 4\boldsymbol{K}_{ii}^{(0)} & \tilde{f}(\boldsymbol{K}_{ij}^{(0)}) \\ \tilde{f}(\boldsymbol{K}_{j,i}^{(0)}) & 4\boldsymbol{K}_{jj}^{(0)} \end{bmatrix} = 4 - \tilde{f}(\boldsymbol{K}_{j,i}^{(0)}) \equiv \lambda_{\text{dag3}}, \tag{11}$$

where $\tilde{f}(\boldsymbol{K}_{ij}^{(0)}) = \boldsymbol{K}_{ij}^{(0)} + f(\boldsymbol{K}_{ij}^{(0)}) + f(2f(\boldsymbol{K}_{ij}^{(0)}))$

Conclusion: by comparing Eq. 7, 9, and 11, we can show the rank of three upper bounds of least eigenvalues as: $\lambda_{\text{dag1}} < \lambda_{\text{dag2}} < \lambda_{\text{dag3}}$. Therefore, the bound of the convergence rate of DAG#3 is better than DAG#2, and DAG#1 is the worst. See our Appendix C in supplement for detailed analysis.

General rules of propagation from $\boldsymbol{K}^{(0)}$ to $\boldsymbol{K}^{(H)}$.

- Diagonal elements of $\boldsymbol{K}^{(H)}$ is determined by the number of unique paths from $\boldsymbol{X}^{(0)}$ to $\boldsymbol{X}^{(H)}$.
- Off-diagonal elements of $\boldsymbol{K}^{(H)}$ is determined by the number of incoming edges of $\boldsymbol{X}^{(H)}$ and the number of parameterized operations on each path.

Thus, we could simplify our rules as:

$$\lambda_{\min}(\boldsymbol{K}^{(H)}) \leq \min_{i \neq j} \left(P - \sum_{p=1}^{P} f^{d_p}\left(\boldsymbol{K}_{ij}^{(0)}\right) \right), \tag{12}$$

where P is number of end-to-end paths from $\boldsymbol{X}^{(0)}$ to $\boldsymbol{X}^{(H)}$, d_p is the number of linear transformation operations on the p-th path, and f^{d_p} indicates a d_p-power composition of f. This summary is based on the propagation of the NNGP variance. With this rule, we can quickly link the relationship between the smallest eigenvalue of $\boldsymbol{K}^{(H)}$ and $\boldsymbol{K}^{(0)}$.

4 Practical principle of connectivity: effective depth and effective width

Eq. 12 precisely characterizes the impact of a network's topology on its convergence. Inspired by above results, we observe two factors that control the $\lambda_{\min}(\boldsymbol{K}^{(H)})$: (1) P, the "width" of a DAG; (2) $d_p (p \in [1, P])$, the "depth" of a DAG. However, directly comparing convergences via Eq. 12 is non-trivial (see our Appendix C), because: (1) the complicated form of the NNGP propagation "f" (Eq. 3.2); (2) "P" and "d_p"s are not truly free variables, as they are still coupled together in the discrete topology space of a fully connected DAG.

Motivated by these two challenges, we propose a practical version of our principle. We simplify the analysis of Eq. 12 by reducing "P" and "d_p"s to only two intuitive variables highly related to the network's connectivity patterns: the effective depth and effective width.

Definition 4.1 (Effective Depth and Width)**.** Consider the directed acyclic graph (DAG) of a network. Suppose there are P unique paths connecting the starting and the ending vertex. Denote the number of parameterized operations on the p-th path as $d_p, p \in [1, P]$. We define:

$$\text{Effective Depth} \quad \bar{d} = \frac{\sum_{p=1}^{P} d_p}{P}, \qquad \text{Effective Width} \quad \bar{m} = \frac{\sum_{p=1}^{P} \mathbb{1}(d_p > 0)}{\bar{d}}, \tag{13}$$

where $\mathbb{1}(d_p > 0) = 1$ if $d_p > 0$ otherwise is 0.

Remark 4.2. In our experiments below, we consider fully-connected or convolutional layers as parameterized operations, ignoring their detailed configurations (kernel sizes, dilations, groups, etc.). We consider skip-connections and pooling layers as non-parameterized operations.

The effective depth considers the averaging effects of multiple parallel paths, and the effective width considers the amount of unique information flowing from the starting vertex to the end, normalized by the overall depth. We will demonstrate later in Section 5.2 that performances of networks from diverse architecture spaces show positive correlations with these two aspects. While $\bar{d}$ and $\bar{m}$ may be loose to predict the best architecture, in Section 5.3 we will show that they are very stable in distinguishing bad ones. Therefore, using these two principles, we can quickly determine if an architecture is potentially unpromising at almost zero cost, by only analyzing its connectivity without any forward or backward calculations.

5 Experiments

5.1 Empirical convergence confirms our analysis

We first experimentally verify our convergence analysis in Section 3.3. In all cases we use ReLU nonlinearities with Kaiming normal initialization [24]. We build the same three computational graphs of fully-connected layers in Figure 3. Three networks have hidden layers of a constant width of 1024. We train the network using SGD with a mini-batch of size 128. The learning rate is fixed at 1×10^{-5}. No augmentation, weight decay, learning rate decay, or momentum is adopted.

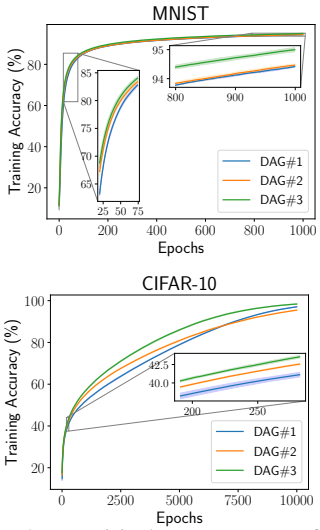

Figure 4: Empirical convergence of three DAG cases (Figure 3) can verify our theoretical analysis in Section 3. Small standard deviations of three runs are shown in shadows.

Based on our analysis, on both MNIST and CIFAR-10, the convergence rate of DAG#1 (Figure 3 left) is worse than DAG#2 (Figure 3 middle), and is further worse than DAG#3 (Figure 3 right). The experimental observation is consistent with this analysis: the training accuracy of DAG#3 increases the fastest, and DAG#2 is faster than DAG#1. Although on CIFAR-10 DAG#1 slightly outperforms DAG#2 in the end, DAG#2 still benefits from faster convergence in most early epochs. This result confirms that by our convergence analysis, we can effectively compare the convergence of different connectivity patterns.

5.2 Extreme $\bar{d}$ or $\bar{m}$ leads to bad performance on architecture benchmarks

In this experiment, our goal is to verify the two principles we proposed in Section 4. Specifically, we will leverage a large number of network architectures of random connectivities, calculate their $\bar{d}$ and $\bar{m}$, and compare their performance. We consider popular architecture benchmarks that are widely studied in the community of neural architecture search (NAS). Licenses are publicly available.

- The *NAS-Bench-201* [17] provides 15,625 architectures that are stacked by repeated DAGs of four nodes (exactly the same DAG we considered in Section 3 and Figure 2). It contains architecture's performance on three datasets (CIFAR-10, CIFAR-100, ImageNet-16-120 [15]) evaluated under a unified protocol (i.e. same learning rate, batch size, etc., for all architectures). The operation space contains *zero*, *skip-connection*, *conv*1×1, *conv*3×3 *convolution*, and *average pooling* 3×3.
- Large-scale architecture spaces: NASNet [73], Amoeba [49], PNAS [39], ENAS [46], DARTS [40]. These spaces have more operation types and more complicated rules on allowed DAG connectivity

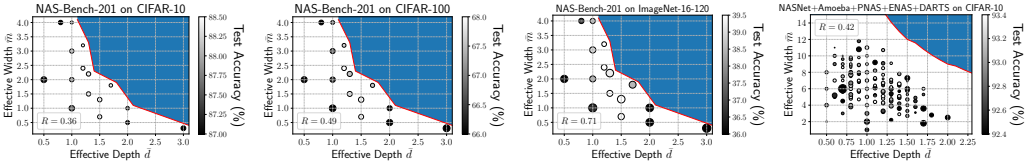

Figure 5: In a complete architecture space, being extremely large or small on either the effective depth ($\bar{d}$) or the effective width ($\bar{m}$) leads a connectivity pattern to bad performance (black) and large variance (large circle size). Each dot represents a subset of architectures of the same $\bar{d}$ and $\bar{m}$. The left three plots are results on CIFAR-10, CIFAR-100, and ImageNet-16-120 from the NAS-Bench-201 [17]. The rightmost plot is on CIFAR-10 from a union of multiple large-scale spaces: NASNet [73], Amoeba [49], PNAS [39], ENAS [46], DARTS [40]. Blue areas indicate invalid DAG regions: a DAG cannot be both deep and wide at the same time.

patterns, see Figure 1 bottom as an example. We refer to their papers for details. A benchmark is further proposed in [48]: about 5000 architectures are randomly selected from each space above, and in total we have 24827 architectures pretrained on CIFAR-10. We will use this prepared data[2].

We show our results in Figure 5. From all four plots[3] (across different architecture spaces and datasets), it can be observed that, architectures of extreme depths or widths suffer from bad performance. We also calculate the multi-correlation[4] (R) between the accuracy and the joint of both $\bar{d}$ and $\bar{m}$, and show as legends in Figure 5. All four plots show positive correlations. It is worth noting that here each dot represents a subset of architectures that share the same $\bar{d}$ and $\bar{m}$, with possible different fine-grained topology or layer types, indicating the generalizability of our proposed principles. The variance of performance in each subset is represented by the radius of the dot.

Remark 5.1. Our notion of "extreme $\bar{d}$ or $\bar{m}$" targets a complete space of DAGs: given the total number of nodes, any two nodes in the graph can be connected, and there are no topological restrictions or disabled edges (i.e., one cannot introduce any prior to the distribution of graph topologies).

5.3 A "Plug-and-Play" method for accelerating NAS: bypassing extreme $\bar{d}$ and $\bar{m}$

Inspired by the bad performance from extreme $\bar{d}$ or $\bar{m}$ discussed in Section 5.2, we are motivated to further contribute a "plug-and-play" tool to accelerate practical NAS applications.

Method. In NAS, the bottleneck of search efficiency is the evaluation cost during the search. Our core idea is to use $\bar{d}$ and $\bar{m}$ to pre-select architectures that are very likely to be unpromising before we spend extra cost to evaluate them. Training-based NAS is slow: one has to perform gradient descent on each sampled architecture and use the loss or accuracy to rank different networks. We instead choose to improve two recent training-free NAS methods: NAS-WOT [44] and TE-NAS [10]. This is because our goal here is to accelerate NAS, and these two are state-of-the-art NAS works of extremely low search time cost. We plan to show that our method can even further accelerate these two training-free NAS works. The rationale is that, our $\bar{d}$ and $\bar{m}$ are even much cheaper: we only do a simple calculation on a network's DAG structure, but NAS-WOT and TE-NAS still have to perform forward or backward on a mini-batch of real data. Although our $\bar{d}$ and $\bar{m}$ are only inspired from the optimization perspective, not the complexity or generalization, our method mainly filters out bad architectures at a coarse level, but does not promote elites in a fine-grained way.

NAS-WOT proposed to rank architectures by their local linear maps. Given a mini-batch of data, each point can be identified by its activation pattern through the network, and a network that can assign more unique activation patterns will be considered promising. NAS-WOT uses random sampling to search the architectures and rank them based on this training-free score. **Our version**: We will skip potentially unpromising architecture before leaving them for NAS-WOT to calculate their scores.

TE-NAS proposed to rank architectures by combining both the trainability (condition number of NTK) and expressivity (number of linear regions). TE-NAS progressively pruned a super network to a single-path network, by removing unpromising edges of low scores.

[2]Data is publicly available at `https://github.com/facebookresearch/nds`, only test accuracy available.
[3]To fairly compare different connectivity patterns, we fix the total number of convolutional layers per model as 3 (out of 6 edges) on NAS-Bench-201, and 5 (out of 10 edges) on the union of large-scale architecture spaces.
[4]See Appendix D for its definition.

Table 1: NAS Search Performance in DARTS space on ImageNet. Our standard deviations over three random runs are included in parentheses.

Architecture	Test Error(%)		Params. (M)	Search Cost (GPU days)	Search Method
	top-1	top-5			
NASNet-A [73]	26.0	8.4	5.3	2000	RL
AmoebaNet-C [49]	24.3	7.6	6.4	3150	evolution
PNAS [39]	25.8	8.1	5.1	225	SMBO
MnasNet-92 [56]†	25.2	8.0	4.4	288 (TPU)	RL
DARTS (2nd) [40]	26.7	8.7	4.7	4.0‡	gradient
SNAS (mild) [63]	27.3	9.2	4.3	1.5	gradient
GDAS [16]	26.0	8.5	5.3	0.21	gradient
BayesNAS [70]	26.5	8.9	3.9	0.2	gradient
P-DARTS [13]	24.4	7.4	4.9	0.3	gradient
PC-DARTS [64]	25.1	7.8	5.3	0.1‡	gradient
PC-DARTS (ImageNet) [64]†	24.2	7.3	5.3	3.8	gradient
ProxylessNAS (GPU) [7]†	24.9	7.5	7.1	8.3	gradient
SGAS (Cri 1. avg) [35]	24.42 (0.16)	7.29 (0.09)	5.3	0.25	gradient
DrNAS [12]†	23.7	7.1	5.7	4.6	gradient
NAS-WOT [44]†⋆	26.2	8.2	4.4	0.0036‡	training-free
NAS-WOT **+ DAG (ours)**†	25.9 (0.4)	8.2	4.4	0.003‡	training-free
TE-NAS [10]†	24.5	7.5	5.4	0.17‡	training-free
TE-NAS **+ DAG (ours)**†	24.2 (0.3)	7.4	6.1	0.1‡	training-free

† Architectures searched on ImageNet. Other methods searched on CIFAR-10 and then transfered to the ImageNet.
⋆ Our reproduced result, the original work did not provide results on the ImageNet.
‡ Recorded on a single GTX 1080Ti GPU.

Table 2: Search Performance from NAS-Bench-201. "optimal" indicates the best test accuracy achievable in the space. Our standard deviations over three random runs are included in parentheses.

Architecture	CIFAR-10	CIFAR-100	ImageNet-16-120	Search Cost (GPU sec.)	Search Method
ResNet [25]	93.97	70.86	43.63	-	-
RSPS [37]	87.66(1.69)	58.33(4.34)	31.14(3.88)	8007.13	random
ENAS [46]	54.30(0.00)	15.61(0.00)	16.32(0.00)	13314.51	RL
DARTS (1st) [40]	54.30(0.00)	15.61(0.00)	16.32(0.00)	10889.87	gradient
DARTS (2nd) [40]	54.30(0.00)	15.61(0.00)	16.32(0.00)	29901.67	gradient
GDAS [16]	93.61(0.09)	70.70(0.30)	41.84(0.90)	28925.91	gradient
WOT [44]	92.81 (0.99)	69.48 (1.70)	43.10 (3.16)	30.01	training-free
WOT + DAG (ours)	92.98 (0.78)	69.86 (1.24)	43.44 (2.64)	17.95 (-40.2%)	training-free
TE-NAS [10]	93.9 (0.47)	71.24 (0.56)	42.38 (0.46)	1558	training-free
TE-NAS + DAG (ours)	93.6 (0.37)	71.26 (0.77)	44.38 (0.76)	1077 (-30.9%)	training-free
FP-NAS [65]	77.4 (16.6)	64.7 (5.3)	26.7 (10.4)	4837	gradient
FP-NAS + DAG (ours)	93.3 (0.3)	70.8 (0.4)	44.5 (1.4)	2612 (-46.0%)	gradient
Optimal	94.37	73.51	47.31	-	-

Our version: We will skip potentially unpromising architecture before leaving them for TE-NAS to prune.

We provide a pseudocode algorithm in Appendix A to demonstrate the usage of our method.

How to choose $\bar{d}$ and $\bar{m}$? Although $\bar{d}$ and $\bar{m}$ are hyperparameters, it is worth noting that they can be determined in a highly principled way. In practice, given a search space, we only calculate the center $(\bar{d}^*, \bar{m}^*) = (\frac{\bar{d}_{\max}+\bar{d}_{\min}}{2}, \frac{\bar{m}_{\max}+\bar{m}_{\min}}{2})$ and the radius $(r_{\bar{d}}, r_{\bar{m}}) = (\frac{\bar{d}_{\max}-\bar{d}_{\min}}{2}, \frac{\bar{m}_{\max}-\bar{m}_{\min}}{2})$. We by default only keep architectures within half of the radius from the center for evaluation: $|\bar{d}-\bar{d}^*| \leq \frac{1}{2}r_{\bar{d}}$ and $|\bar{m}-\bar{m}^*| \leq \frac{1}{2}r_{\bar{m}}$. This principle leads to $(\bar{d}^*, \bar{m}^*) = (1.5, 10.5)$ with $\frac{1}{2}(r_{\bar{d}}, r_{\bar{m}}) = (0.7, 4.8)$ on DARTS, and $(\bar{d}^*, \bar{m}^*) = (1.6, 2.2)$ with $\frac{1}{2}(r_{\bar{d}}, r_{\bar{m}}) = (0.7, 0.9)$ on NAS-Bench-201.

Implementation settings. We train searched architectures for 250 epochs using SGD, with a learning rate as 0.5, a cosine scheduler, momentum as 0.9, weight decay as 3×10^{-5}, and a batch size as 768. This setting follows previous works [1, 44, 69, 41, 66, 26, 12, 10].

DARTS space on ImageNet. As shown in Table 1, for both two training-free NAS methods, by adopting our pre-filtration strategy, we can further reduce the search time cost and achieve better search results. For NAS-WOT, we can save 16.7% search time cost and reduce 0.3% top-1 error. For TE-NAS, we can significantly save 41.2% search time cost, and still improve the top-1 error by 0.4%. FLOPs of our search architectures are 0.68G (TE-NAS + DAG) and 0.56G (WOT + DAG).

NAS-Bench-201 Space. As shown in Table 2, for both NAS-WOT and TE-NAS, we reduce over 30% search time cost with strong accuracy. We also include a training-based method FP-NAS [65], where we even achieve 46% seach cost reduction with better performance. Moreover, we show that our method is robust to the choices of $\bar{d}$ and $\bar{m}$. In the ablation study in Table 3, by changing different ranges of $\bar{d}$ and $\bar{m}$, our method remains strong over the WOT baseline.

Table 3: Our method is robust to choices of $\bar{d}$ and $\bar{m}$ (WOT [44] on NAS-Bench-201).

Ranges of accepted $\bar{d}$ and $\bar{m}$	CIFAR-100
$\lvert\bar{d}-\bar{d}^*\rvert \leq r_{\bar{d}}, \lvert\bar{m}-\bar{m}^*\rvert \leq r_{\bar{m}}$ (baseline)	69.48 (1.70)
$\lvert\bar{d}-\bar{d}^*\rvert \leq \frac{3}{4}r_{\bar{d}}, \lvert\bar{m}-\bar{m}^*\rvert \leq \frac{3}{4}r_{\bar{m}}$	69.51 (1.70)
$\lvert\bar{d}-\bar{d}^*\rvert \leq \frac{1}{2}r_{\bar{d}}, \lvert\bar{m}-\bar{m}^*\rvert \leq \frac{1}{2}r_{\bar{m}}$	69.86 (1.24)
$\lvert\bar{d}-\bar{d}^*\rvert \leq \frac{1}{4}r_{\bar{d}}, \lvert\bar{m}-\bar{m}^*\rvert \leq \frac{1}{4}r_{\bar{m}}$	69.97 (1.19)

6 Conclusion

In this work, we show that it is possible to conduct fine-grained convergence analysis on networks of complex connectivity patterns. By analyzing how an NNGP kernel propagates through the networks, we can fairly compare different networks' bounds of convergence rates. This theoretical analysis and comparison are empirically verified on MNIST and CIFAR-10. To make our convergence analysis more practical and general, we propose two intuitive principles on how to design a network's connectivity patterns: the effective depth and the effective width. Experiments on diverse architecture benchmarks and datasets demonstrate that networks with an extreme depth or width show bad performance, indicating that both the depth and width are important. Finally, we apply our principles to the large-scale neural architecture search application, and our method can largely accelerate the search cost of two training-free efficient NAS works with faster and better search performance. Our work bridge the gap between the Deep Learning theory and the application part, making the theoretical analysis more practical in architecture designs.

Acknowledgement

B. Hanin and Z. Wang are supported by NSF Scale-MoDL (award numbers: 2133806, 2133861).

References

[1] Mohamed Abdelfattah, Abhinav Mehrotra, Lukasz Dudziak, and Donald Nicholas Lane. Zero-cost proxies for lightweight nas. In *International Conference on Learning Representations*, 2021.

[2] Zeyuan Allen-Zhu, Yuanzhi Li, and Zhao Song. A convergence theory for deep learning via over-parameterization. In *International Conference on Machine Learning*, pages 242–252. PMLR, 2019.

[3] Sanjeev Arora, Simon Du, Wei Hu, Zhiyuan Li, and Ruosong Wang. Fine-grained analysis of optimization and generalization for overparameterized two-layer neural networks. In *International Conference on Machine Learning*, pages 322–332. PMLR, 2019.

[4] David Balduzzi, Marcus Frean, Lennox Leary, JP Lewis, Kurt Wan-Duo Ma, and Brian McWilliams. The shattered gradients problem: If resnets are the answer, then what is the question? In *International Conference on Machine Learning*, pages 342–350. PMLR, 2017.

[5] Kartikeya Bhardwaj, Guihong Li, and Radu Marculescu. How does topology influence gradient propagation and model performance of deep networks with densenet-type skip connections? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13498–13507, 2021.

[6] Kaifeng Bu, Yaobo Zhang, and Qingxian Luo. Depth-width trade-offs for neural networks via topological entropy. *arXiv preprint arXiv:2010.07587*, 2020.

[7] Han Cai, Ligeng Zhu, and Song Han. Proxylessnas: Direct neural architecture search on target task and hardware. *arXiv preprint arXiv:1812.00332*, 2018.

[8] Yuan Cao and Quanquan Gu. Generalization bounds of stochastic gradient descent for wide and deep neural networks. *Advances in Neural Information Processing Systems*, 32:10836–10846, 2019.

[9] Xiangyu Chang, Yingcong Li, Samet Oymak, and Christos Thrampoulidis. Provable benefits of overparameterization in model compression: From double descent to pruning neural networks. *arXiv preprint arXiv:2012.08749*, 2020.

[10] Wuyang Chen, Xinyu Gong, and Zhangyang Wang. Neural architecture search on imagenet in four gpu hours: A theoretically inspired perspective. In *International Conference on Learning Representations*, 2021.

[11] Wuyang Chen, Wei Huang, Xianzhi Du, Xiaodan Song, Zhangyang Wang, and Denny Zhou. Auto-scaling vision transformers without training. In *International Conference on Learning Representations*, 2022.

[12] Xiangning Chen, Ruochen Wang, Minhao Cheng, Xiaocheng Tang, and Cho-Jui Hsieh. Drnas: Dirichlet neural architecture search. *arXiv preprint arXiv:2006.10355*, 2020.

[13] Xin Chen, Lingxi Xie, Jun Wu, and Qi Tian. Progressive differentiable architecture search: Bridging the depth gap between search and evaluation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1294–1303, 2019.

[14] Zixiang Chen, Yuan Cao, Difan Zou, and Quanquan Gu. How much over-parameterization is sufficient to learn deep relu networks? *arXiv preprint arXiv:1911.12360*, 2019.

[15] Patryk Chrabaszcz, Ilya Loshchilov, and Frank Hutter. A downsampled variant of imagenet as an alternative to the cifar datasets, 2017.

[16] Xuanyi Dong and Yi Yang. Searching for a robust neural architecture in four gpu hours. In *Proceedings of the IEEE Conference on computer vision and pattern recognition*, pages 1761–1770, 2019.

[17] Xuanyi Dong and Yi Yang. Nas-bench-102: Extending the scope of reproducible neural architecture search. *arXiv preprint arXiv:2001.00326*, 2020.

[18] Simon Du, Jason Lee, Haochuan Li, Liwei Wang, and Xiyu Zhai. Gradient descent finds global minima of deep neural networks. In *International Conference on Machine Learning*, pages 1675–1685. PMLR, 2019.

[19] Simon S Du, Xiyu Zhai, Barnabas Poczos, and Aarti Singh. Gradient descent provably optimizes over-parameterized neural networks. *arXiv preprint arXiv:1810.02054*, 2018.

[20] Stanislav Fort and Stanislaw Jastrzebski. Large scale structure of neural network loss landscapes. *Advances in Neural Information Processing Systems*, 32:6709–6717, 2019.

[21] Boris Hanin, Ryan Jeong, and David Rolnick. Deep relu networks preserve expected length. *arXiv preprint arXiv:2102.10492*, 2021.

[22] Boris Hanin and Mihai Nica. Finite depth and width corrections to the neural tangent kernel. *arXiv preprint arXiv:1909.05989*, 2019.

[23] Soufiane Hayou, Arnaud Doucet, and Judith Rousseau. On the impact of the activation function on deep neural networks training. *arXiv preprint arXiv:1902.06853*, 2019.

[24] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pages 1026–1034, 2015.

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

[26] Pengfei Hou, Ying Jin, and Yukang Chen. Single-darts: Towards stable architecture search. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 373–382, 2021.

[27] Andrew Howard, Mark Sandler, Grace Chu, Liang-Chieh Chen, Bo Chen, Mingxing Tan, Weijun Wang, Yukun Zhu, Ruoming Pang, Vijay Vasudevan, et al. Searching for mobilenetv3. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1314–1324, 2019.

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

[29] Kaixuan Huang, Yuqing Wang, Molei Tao, and Tuo Zhao. Why do deep residual networks generalize better than deep feedforward networks?–a neural tangent kernel perspective. *arXiv preprint arXiv:2002.06262*, 2020.

[30] Wei Huang, Weitao Du, and Richard Yi Da Xu. On the neural tangent kernel of deep networks with orthogonal initialization. *arXiv preprint arXiv:2004.05867*, 2020.

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

[32] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[33] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[34] 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.

[35] Guohao Li, Guocheng Qian, Itzel C Delgadillo, Matthias Muller, Ali Thabet, and Bernard Ghanem. Sgas: Sequential greedy architecture search. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1620–1630, 2020.

[36] Hao Li, Zheng Xu, Gavin Taylor, Christoph Studer, and Tom Goldstein. Visualizing the loss landscape of neural nets. *arXiv preprint arXiv:1712.09913*, 2017.

[37] Liam Li and Ameet Talwalkar. Random search and reproducibility for neural architecture search. In *Uncertainty in Artificial Intelligence*, pages 367–377. PMLR, 2020.

[38] Yuanzhi Li and Yingyu Liang. Learning overparameterized neural networks via stochastic gradient descent on structured data. *arXiv preprint arXiv:1808.01204*, 2018.

[39] Chenxi Liu, Barret Zoph, Maxim Neumann, Jonathon Shlens, Wei Hua, Li-Jia Li, Li Fei-Fei, Alan Yuille, Jonathan Huang, and Kevin Murphy. Progressive neural architecture search. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 19–34, 2018.

[40] Hanxiao Liu, Karen Simonyan, and Yiming Yang. Darts: Differentiable architecture search. *arXiv preprint arXiv:1806.09055*, 2018.

[41] Vasco Lopes, Miguel Santos, Bruno Degardin, and Luís A Alexandre. Guided evolution for neural architecture search. *arXiv preprint arXiv:2110.15232*, 2021.

[42] Yiping Lu, Chao Ma, Yulong Lu, Jianfeng Lu, and Lexing Ying. A mean field analysis of deep resnet and beyond: Towards provably optimization via overparameterization from depth. In *International Conference on Machine Learning*, pages 6426–6436. PMLR, 2020.

[43] Song Mei, Theodor Misiakiewicz, and Andrea Montanari. Mean-field theory of two-layers neural networks: dimension-free bounds and kernel limit. In *Conference on Learning Theory*, pages 2388–2464. PMLR, 2019.

[44] Joe Mellor, Jack Turner, Amos Storkey, and Elliot J Crowley. Neural architecture search without training. In *International Conference on Machine Learning*, pages 7588–7598. PMLR, 2021.

[45] Daniel S Park, Jaehoon Lee, Daiyi Peng, Yuan Cao, and Jascha Sohl-Dickstein. Towards nngp-guided neural architecture search. *arXiv preprint arXiv:2011.06006*, 2020.

[46] Hieu Pham, Melody Y Guan, Barret Zoph, Quoc V Le, and Jeff Dean. Efficient neural architecture search via parameter sharing. *arXiv preprint arXiv:1802.03268*, 2018.

[47] Ben Poole, Subhaneil Lahiri, Maithra Raghu, Jascha Sohl-Dickstein, and Surya Ganguli. Exponential expressivity in deep neural networks through transient chaos. In *Advances in neural information processing systems*, pages 3360–3368, 2016.

[48] Ilija Radosavovic, Justin Johnson, Saining Xie, Wan-Yen Lo, and Piotr Dollár. On network design spaces for visual recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1882–1890, 2019.

[49] Esteban Real, Alok Aggarwal, Yanping Huang, and Quoc V Le. Regularized evolution for image classifier architecture search. In *Proceedings of the aaai conference on artificial intelligence*, volume 33, pages 4780–4789, 2019.

[50] Bastian Rieck, Matteo Togninalli, Christian Bock, Michael Moor, Max Horn, Thomas Gumbsch, and Karsten Borgwardt. Neural persistence: A complexity measure for deep neural networks using algebraic topology. *arXiv preprint arXiv:1812.09764*, 2018.

[51] Alexander Shevchenko and Marco Mondelli. Landscape connectivity and dropout stability of sgd solutions for over-parameterized neural networks. In *International Conference on Machine Learning*, pages 8773–8784. PMLR, 2020.

[52] Yao Shu, Wei Wang, and Shaofeng Cai. Understanding architectures learnt by cell-based neural architecture search. In *International Conference on Learning Representations*, 2019.

[53] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[54] Mahdi Soltanolkotabi, Adel Javanmard, and Jason D Lee. Theoretical insights into the optimization landscape of over-parameterized shallow neural networks. *IEEE Transactions on Information Theory*, 65(2):742–769, 2018.

[55] Mei Song, Andrea Montanari, and P Nguyen. A mean field view of the landscape of two-layers neural networks. *Proceedings of the National Academy of Sciences*, 115:E7665–E7671, 2018.

[56] Mingxing Tan, Bo Chen, Ruoming Pang, Vijay Vasudevan, Mark Sandler, Andrew Howard, and Quoc V Le. Mnasnet: Platform-aware neural architecture search for mobile. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2820–2828, 2019.

[57] Mingxing Tan and Quoc V Le. Efficientnet: Rethinking model scaling for convolutional neural networks. *arXiv preprint arXiv:1905.11946*, 2019.

[58] Mingxing Tan and Quoc V Le. Efficientnetv2: Smaller models and faster training. *arXiv preprint arXiv:2104.00298*, 2021.

[59] Hidenori Tanaka, Daniel Kunin, Daniel LK Yamins, and Surya Ganguli. Pruning neural networks without any data by iteratively conserving synaptic flow. *arXiv preprint arXiv:2006.05467*, 2020.

[60] Luca Venturi, Afonso S Bandeira, and Joan Bruna. Spurious valleys in two-layer neural network optimization landscapes. *arXiv preprint arXiv:1802.06384*, 2018.

[61] Lechao Xiao, Jeffrey Pennington, and Samuel S Schoenholz. Disentangling trainability and generalization in deep learning. *arXiv preprint arXiv:1912.13053*, 2019.

[62] Saining Xie, Alexander Kirillov, Ross Girshick, and Kaiming He. Exploring randomly wired neural networks for image recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1284–1293, 2019.

[63] Sirui Xie, Hehui Zheng, Chunxiao Liu, and Liang Lin. Snas: stochastic neural architecture search. *arXiv preprint arXiv:1812.09926*, 2018.

[64] Yuhui Xu, Lingxi Xie, Xiaopeng Zhang, Xin Chen, Guo-Jun Qi, Qi Tian, and Hongkai Xiong. Pc-darts: Partial channel connections for memory-efficient architecture search. In *International Conference on Learning Representations*, 2019.

[65] Zhicheng Yan, Xiaoliang Dai, Peizhao Zhang, Yuandong Tian, Bichen Wu, and Matt Feiszli. Fp-nas: Fast probabilistic neural architecture search. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15139–15148, 2021.

[66] Peng Ye, Baopu Li, Yikang Li, Tao Chen, Jiayuan Fan, and Wanli Ouyang. β-darts: Beta-decay regularization for differentiable architecture search. *arXiv preprint arXiv:2203.01665*, 2022.

[67] Jiaxuan You, Jure Leskovec, Kaiming He, and Saining Xie. Graph structure of neural networks. *arXiv preprint arXiv:2007.06559*, 2020.

[68] Xiao Zhang, Yaodong Yu, Lingxiao Wang, and Quanquan Gu. Learning one-hidden-layer relu networks via gradient descent. In *The 22nd international conference on artificial intelligence and statistics*, pages 1524–1534. PMLR, 2019.

[69] Xuanyang Zhang, Pengfei Hou, Xiangyu Zhang, and Jian Sun. Neural architecture search with random labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10907–10916, 2021.

[70] Hongpeng Zhou, Minghao Yang, Jun Wang, and Wei Pan. Bayesnas: A bayesian approach for neural architecture search. *arXiv preprint arXiv:1905.04919*, 2019.

[71] Pan Zhou, Caiming Xiong, Richard Socher, and Steven CH Hoi. Theory-inspired path-regularized differential network architecture search. *arXiv preprint arXiv:2006.16537*, 2020.

[72] Barret Zoph and Quoc V Le. Neural architecture search with reinforcement learning. *arXiv preprint arXiv:1611.01578*, 2016.

[73] Barret Zoph, Vijay Vasudevan, Jonathon Shlens, and Quoc V Le. Learning transferable architectures for scalable image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8697–8710, 2018.

[74] Difan Zou, Yuan Cao, Dongruo Zhou, and Quanquan Gu. Gradient descent optimizes over-parameterized deep relu networks. *Machine Learning*, 109(3):467–492, 2020.

[75] Difan Zou, Philip M Long, and Quanquan Gu. On the global convergence of training deep linear resnets. *arXiv preprint arXiv:2003.01094*, 2020.

Checklist

The checklist follows the references. Please read the checklist guidelines carefully for information on how to answer these questions. For each question, change the default **[TODO]** to [Yes] , [No] , or [N/A] . You are strongly encouraged to include a **justification to your answer**, either by referencing the appropriate section of your paper or providing a brief inline description. For example:

- Did you include the license to the code and datasets? [Yes] See Section **??**.
- Did you include the license to the code and datasets? [No] The code and the data are proprietary.
- Did you include the license to the code and datasets? [N/A]

Please do not modify the questions and only use the provided macros for your answers. Note that the Checklist section does not count towards the page limit. In your paper, please delete this instructions block and only keep the Checklist section heading above along with the questions/answers below.

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] In Remark 5.1, our claim "Extreme $\bar{d}$ or $\bar{m}$ leads to bad performance" requires architectures from a complete space of DAGs, without introducing any prior on the DAG topology.
   (c) Did you discuss any potential negative societal impacts of your work? [No] Our work does not have negative societal impacts
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes] We include our proofs in supplement.
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] We include our code and Readme of instructions in supplement. Data is publicly available online.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] See Section 5.3.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] We run experiments for three random seeds.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] See Table 1.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]

(b) Did you mention the license of the assets? [Yes] Licenses are publicly available

(c) Did you include any new assets either in the supplemental material or as a URL? [N/A]

(d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [No]

(e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]

5. If you used crowdsourcing or conducted research with human subjects...

   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]

   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]

   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

A Pseudocode of our "Plug-and-Play" method

Here we provide a pseudocode of using our "Plug-and-Play" method in the simplified WOT algorithm [44]. Black lines are from original WOT. Red lines are our method, which is extremely easy to use and cheap to calculate.

Algorithm 1: Pseudocode of "WOT + ours" in a PyTorch-like style.

```
# depth_center, r_depth: center and radius of depths defined in Section 5.3
# width_center, r_width: center and radius of widths defined in Section 5.3
# search_space: a given space of architectures, like DARTS or NAS-Bench-201
# N: number of sampled architectures to evaluate
# WOT: the Jacobian-based scoring method in [44].

archs = random_sample(search_space, N) # randomly sample N architectures to evaluate
scores = []
for arch in archs:
    depth, width = get_depth_width(arch) # get DAG's effective depth and width
    if not (abs(depth - depth_center) <= 0.5 * r_depth and abs(width - width_center) <= 0.5 * r_width):
        # this DAG has large depth or large width, bypass it
        scores.append(-inf)
        continue
    score = WOT(arch)
    scores.append(score)
return archs[argmax(scores)]
```

B Proofs for Section 3.3

Lemma 3.2 (Propagation of $\boldsymbol{K}$)**.** *Define the propagation as* $\boldsymbol{K}^{(l)} = f(\boldsymbol{K}^{(l-1)})$ *and* $\boldsymbol{b}^{(l)} = g(\boldsymbol{b}^{(l-1)})$*. When the edge operation is a linear transformation, we have,*

$$\begin{aligned}
\boldsymbol{K}_{ii}^{(l)} &= f(\boldsymbol{K}_{ii}^{(l-1)}) = \int \mathcal{D}_z c_\sigma \sigma^2\left(\sqrt{\boldsymbol{K}_{ii}^{(l-1)}} z\right) \\
\boldsymbol{K}_{ij}^{(l)} &= f(\boldsymbol{K}_{ij}^{(l-1)}) = \int \mathcal{D}_{z_1}\mathcal{D}_{z_2} c_\sigma \sigma\left(\sqrt{\boldsymbol{K}_{ii}^{(l-1)}} z_1\right)\sigma\left(\sqrt{\boldsymbol{K}_{jj}^{(l-1)}}(\boldsymbol{C}_{ij}^{(l-1)} z_1 + \sqrt{1-(\boldsymbol{C}_{ij}^{(l-1)})^2} z_2)\right) \\
\boldsymbol{C}_{ij}^{(l)} &= \boldsymbol{K}_{ij}^{(l)} / \sqrt{\boldsymbol{K}_{ii}^{(l)} \boldsymbol{K}_{jj}^{(l)}} \\
\boldsymbol{b}_i^{(l)} &= g(\boldsymbol{b}_i^{(l-1)}) = \int \mathcal{D}_z \sqrt{c_\sigma} \sigma(\sqrt{\boldsymbol{K}_{ii}^{(l)}} z)
\end{aligned} \tag{14}$$

where z*,* z_1 *and* z_2 *are independent standard Gaussian random variables. Besides,* $\int \mathcal{D}_z = \frac{1}{\sqrt{2\pi}} \int dz e^{-\frac{1}{2}z^2}$ *is the measure for a normal distribution.*

Proof of Lemma 3.2. According to the property of Gaussian process for neural networks [34], for each layer l, we have

$$\boldsymbol{K}_{ii}^{(l)} = \int \mathcal{D}_z c_\sigma \sigma^2\left(\sqrt{\boldsymbol{K}_{ii}^{(l-1)}} z\right)$$

$$\boldsymbol{K}_{ij}^{(l)} = \int \mathcal{D}_{z_1}\mathcal{D}_{z_2} c_\sigma \sigma(u)\sigma(v)$$

where $u = \sqrt{\boldsymbol{K}_{ii}^{(l-1)}} z_1$ and $v = \sqrt{\boldsymbol{K}_{jj}^{(l-1)}}\left(\boldsymbol{C}_{ij}^{(l-1)} z_1 + \sqrt{1-(\boldsymbol{C}_{ij}^{(l-1)})^2} z_2\right)$, with $\boldsymbol{C}_{ij}^{(l)} = \boldsymbol{K}_{ij}^{(l)} / \sqrt{\boldsymbol{K}_{ii}^{(l)} \boldsymbol{K}_{jj}^{(l)}}$. Besides, $\int \mathcal{D}_z = \frac{1}{\sqrt{2\pi}} \int dz e^{-\frac{1}{2}z^2}$ is the measure for a normal distribution.

Note that the diagonal entry $\boldsymbol{K}_{ii}^{(l)}$ of NNGP is a self-correlation term thus governed by one random variance. On the other hand, $\boldsymbol{K}_{ij}^{(l)}$ is the correlation between $\boldsymbol{X}_i^{(l)}$ and $\boldsymbol{X}_j^{(l)}$, therefore it contains two independent random variables. Finally, we calculate the bias term:

$$\boldsymbol{b}_i^{(l)} = g(\boldsymbol{b}_i^{(l-1)}) = \sqrt{c_\sigma} \int \mathcal{D}_z \sigma(\sqrt{\boldsymbol{K}_{ii}^{(l)}} z)$$

□

Lemma 3.3. *For a positive definite symmetric matrix $\boldsymbol{K} \in \mathbb{R}^{N\times N}$, the smallest eigenvalue is bounded by the smallest eigenvalue of its 2×2 principal sub-matrix.*

$$\lambda_{\min}(\boldsymbol{K}) \leq \min_{i\neq j} \lambda_{\min} \begin{bmatrix} \boldsymbol{K}_{ii} & \boldsymbol{K}_{ij} \\ \boldsymbol{K}_{ji} & \boldsymbol{K}_{jj} \end{bmatrix} \tag{15}$$

Proof of Lemma 3.3. We first have the **Cauchy Interlace Theorem**: Suppose $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ is symmetric. Let $\boldsymbol{A}^{(n-1)} \in \mathbb{R}^{(n-1)\times(n-1)}$ be a principal submatrix of $\boldsymbol{A}$ (obtained by removing both i-th row and i-th column for any $i \in [1, n]$). Suppose $\boldsymbol{A}$ has eigenvalues $\lambda_1^{(n)} \geq \lambda_2^{(n)} \geq \cdots \geq \lambda_n^{(n)}$ and $\boldsymbol{A}^{(n-1)}$ has eigenvalues $\lambda_1^{(n-1)} \geq \lambda_2^{(n-1)} \geq \cdots \geq \lambda_{n-1}^{(n-1)}$. Then:

$$\lambda_1^{(n)} \geq \lambda_1^{(n-1)} \geq \lambda_2^{(n)} \geq \lambda_2^{(n-1)} \geq \cdots \geq \lambda_{n-1}^{(n-1)} \geq \lambda_n^{(n)}.$$

By repeating this theorem for $n-2$ times, we can have:

$$\lambda_2^{(2)} \geq \lambda_3^{(3)} \geq \cdots \geq \lambda_{n-1}^{(n-1)} \geq \lambda_n^{(n)}.$$

This means we can upper bound $\lambda_n^{(n)}$ by the minimal smallest eigenvalue of any 2×2 principal submatrix of $\boldsymbol{A}$.

□

C Detailed Analysis on three DAGs in Figure 3 and Section 3.3.

We take the condition of ReLU that $\sigma(x) = \max\{0, x\}$ and $c_\sigma = 2$ into Lemma 3.2 and obtain:

$$\begin{aligned}
\boldsymbol{K}_{ii}^{(l)} &= \int c_\sigma \mathcal{D}_z \sigma^2\left(\sqrt{\boldsymbol{K}_{ii}^{(l-1)}} z\right) = \boldsymbol{K}_{ii}^{(l-1)} \\
\boldsymbol{K}_{ij}^{(l)} &= \frac{2\boldsymbol{C}_{ij}^{(l-1)} \arcsin \boldsymbol{C}_{ij}^{(l-1)} + 2\sqrt{1-(\boldsymbol{C}_{ij}^{(l-1)})^2} + \pi \boldsymbol{C}_{ij}^{(l-1)}}{2\pi} \cdot \sqrt{\boldsymbol{K}_{ii}^{(l)} \boldsymbol{K}_{jj}^{(l)}} \\
\boldsymbol{b}_i^l &= \int \sqrt{c_\sigma \mathcal{D}_z} \sigma(\sqrt{\boldsymbol{K}_{ii}^{(l-1)}} z) = \frac{\sqrt{2}}{2}.
\end{aligned}$$

We provide more properties regarding the propagation of normalized non-diagonal elements in NNGP by studying the function

$$\boldsymbol{C}_{ij}^{(l)} = h(\boldsymbol{C}_{ij}^{(l-1)}) = \frac{2\boldsymbol{C}_{ij}^{(l-1)} \arcsin \boldsymbol{C}_{ij}^{(l-1)} + 2\sqrt{1-(\boldsymbol{C}_{ij}^{(l-1)})^2} + \pi \boldsymbol{C}_{ij}^{(l-1)}}{2\pi}.$$

It is known that f is differentiable and satisfies:

$$h'(\boldsymbol{C}_{ij}^{(l-1)}) = \frac{1}{\pi} \arcsin(\boldsymbol{C}_{ij}^{(l-1)}) + \frac{1}{2};\ h''(\boldsymbol{C}_{ij}^{(l-1)}) = \frac{1}{\pi\sqrt{1-(\boldsymbol{C}_{ij}^{(l-1)})^2}}.$$

Therefor we have $h(\boldsymbol{C}_{ii}^{(l-1)}) > \boldsymbol{C}_{ii}^{(l-1)}$, for $\boldsymbol{C}_{ii}^{(l-1)} \in (0, 1)$. Finally, we show the Taylor expansion for $h(\boldsymbol{C}_{ii}^{(l-1)})$ when $\boldsymbol{C}_{ii}^{(l-1)}$ is close to 1:

$$h(\boldsymbol{C}_{ij}^{(l)}) \overset{\boldsymbol{C}_{ij}^{(l)} \to 1-}{=} \boldsymbol{C}_{ij}^{(l)} + \frac{2\sqrt{2}}{3\pi}(1-\boldsymbol{C}_{ij}^{(l)})^{3/2} + O((1-\boldsymbol{C}_{ij}^{(l)})^{5/2}).$$

This can be obtained by expanding each term in $h(\boldsymbol{C}_{ij}^{(l-1)})$.

Before our analysis, we re-state some assumptions and facts:

- $h(\cdot)$ is a monotonically increasing function in $[0, 1)$, and $\lim_{\boldsymbol{C}_{ij}^{(l-1)} \to 1-} h(\boldsymbol{C}_{ij}^{(l-1)}) = 1$. Please refer to the blue curve in Figure 2(b) in [23] for the behavior of $h(\cdot)$.

- We assume no parallel data points in the training data, thus $\boldsymbol{C}_{ij}^{(0)} \in [0,1)$ and $\boldsymbol{K}_{ij}^{0} \in [0,1)$. This further indicates $h(\boldsymbol{C}_{ij}^{(0)}), h(\boldsymbol{K}_{ij}^{(0)}) \in [1/\pi, 1)$.
- $f(\boldsymbol{K}_{ij}^{(l-1)}) > \boldsymbol{K}_{ij}^{(l-1)}$.

We first show that $\lambda_{\text{dag1}} < \lambda_{\text{dag2}}$, namely, the upper bounds of the least eigenvalues of DAG#1 is smaller than that of DAG#2 in Figure 3:

$$\lambda_{\text{dag2}} = 3(1 - f(\boldsymbol{K}_{ij}^{(0)})) > 1 - f(\boldsymbol{K}_{ij}^{(0)}) > 1 - f^3(\boldsymbol{K}_{ij}^{(0)}) = \lambda_{\text{dag1}}.$$

Before we compare λ_{dag2} and λ_{dag3}, we first try to simplify λ_{dag3}. Specifically, we have:

$$f(\boldsymbol{K}_{ij}^{(0)}) = h(\boldsymbol{C}_{ij}^{(0)})\sqrt{\boldsymbol{K}_{ii}^{(1)}\boldsymbol{K}_{jj}^{(1)}} = h(\boldsymbol{C}_{ij}^{(0)})$$

and

$$\begin{aligned} f(2f(\boldsymbol{K}_{ij}^{(0)})) &= h(\frac{2f(\boldsymbol{K}_{ij}^{(0)})}{\sqrt{2f(\boldsymbol{K}_{ii}^{(0)}) \cdot 2f(\boldsymbol{K}_{jj}^{(0)})}})\sqrt{2\boldsymbol{K}_{ii}^{(1)} \cdot 2\boldsymbol{K}_{jj}^{(1)}} \\ &= h(f(\boldsymbol{K}_{ij}^{(0)})) \cdot 2 = 2h(h(\boldsymbol{C}_{ij}^{(0)})) \end{aligned}$$

Now we compare λ_{dag2} and λ_{dag3}:

$$\begin{aligned} \lambda_{\text{dag3}} - \lambda_{\text{dag2}} &= 4 - \left(\boldsymbol{K}_{ij}^{(0)} + f(\boldsymbol{K}_{ij}^{(0)}) + f(2f(\boldsymbol{K}_{ij}^{(0)}))\right) - 3(1 - f(\boldsymbol{K}_{ij}^{(0)})) \\ &= 1 - \boldsymbol{K}_{ij}^{(0)} + 2f(\boldsymbol{K}_{ij}^{(0)}) - f(2f(\boldsymbol{K}_{ij}^{(0)})) \\ &= 1 - \boldsymbol{K}_{ij}^{(0)} + 2\left(h(\boldsymbol{C}_{ij}^{(0)}) - h(h(\boldsymbol{C}_{ij}^{(0)}))\right). \end{aligned}$$

As $\boldsymbol{K}_{ij}^{(0)} \to 1-$, we have $\boldsymbol{C}_{ij}^{(0)} \to 1-$ and $h(\boldsymbol{C}_{ij}^{(l-1)}) \to 1-$. Thus $\lim_{\boldsymbol{K}_{ij}^{(0)} \to 1-} \lambda_{\text{dag3}} - \lambda_{\text{dag2}} = 0$. Moreover, take the derivative of $\lambda_{\text{dag3}} - \lambda_{\text{dag2}}$ w.r.t. $\boldsymbol{K}_{ij}^{(0)}$, we have:

$$\begin{aligned} \frac{\partial(\lambda_{\text{dag3}} - \lambda_{\text{dag2}})}{\partial \boldsymbol{K}_{ij}^{(0)}} &= -1 + 2h'(\boldsymbol{K}_{ij}^{(0)})(1 - h'(h(\boldsymbol{K}_{ij}^{(0)}))) \\ &< -1 + 2 \cdot 1 \cdot (1 - \frac{1}{2}) = 0. \end{aligned}$$

The inequality considers the fact that $h' \in [\frac{1}{2}, 1)$. Therefore, $\lambda_{\text{dag3}} - \lambda_{\text{dag2}}$ is a monotonically decreasing function w.r.t. $\boldsymbol{K}_{ij}^{(0)}$ on $[0,1)$. Thus, we have $\lambda_{\text{dag3}} - \lambda_{\text{dag2}} > 0$ on $[0,1)$.

In conclusion, we have $\lambda_{\text{dag3}} > \lambda_{\text{dag2}} > \lambda_{\text{dag1}}$. Therefore, the convergence rates are ranked as: DAG#3 > DAG#2 > DAG#1.

D Coefficient of Multiple Correlation R

Our correlation coefficient R between the accuracy and the joint of both $\bar{d}$ and $\bar{m}$ is calculated as blow:

$$\begin{aligned} R^2 &= \boldsymbol{c}^\top \boldsymbol{R}_{\bar{d},\bar{m}}^{-1} \boldsymbol{c} \\ \boldsymbol{c} &= [r_{\bar{d},y}, r_{\bar{m},y}]^T \\ \boldsymbol{R}_{\bar{d},\bar{m}} &= \begin{bmatrix} r_{x_{\bar{d}}x_{\bar{d}}} & r_{x_{\bar{d}}x_{\bar{m}}} \\ r_{x_{\bar{m}}x_{\bar{d}}} & r_{x_{\bar{m}}x_{\bar{m}}} \end{bmatrix} \end{aligned} \tag{16}$$

E Proof of Theorem 3.1

Theorem 3.1 (Linear Convergence of DAG)**.** *At k-th iteration of gradient descent on N training data samples, we have a network parameterized by $\boldsymbol{\theta}(k)$ with its output as $\boldsymbol{u}(k)$. The network's*

connectivity pattern can be formulated as a DAG of H nodes and P end-to-end paths. With MSE loss $\mathcal{L}(\boldsymbol{\theta}(k)) = \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{u}(k)\|_2^2$, suppose the learning rate $\eta = O\left(\frac{\lambda_{\min}\left(\boldsymbol{K}^{(H)}\right)}{(NP)^2} 2^{O(H)}\right)$ and the number of neurons (width) per layer $m = \Omega\left(\max\left\{\frac{(NP_H)^4}{\lambda_{\min}^4\left(\boldsymbol{K}^{(H)}\right)}, \frac{NP_H H}{\delta}, \frac{(NP_H)^2 \log\left(\frac{HN}{\delta}\right) 2^{O(H)}}{\lambda_{\min}^2\left(\boldsymbol{K}^{(H)}\right)}\right\}\right)$, we have

$$\|\mathbf{y} - \mathbf{u}(k)\|_2^2 \leq \left(1 - \frac{\eta \lambda_{\min}(\boldsymbol{K}^{(H)})}{2}\right)^k \|\mathbf{y} - \mathbf{u}(0)\|_2^2 \tag{17}$$

Proof Sketch of Theorem 3.1. The main purpose of this theorem is to extend the convergence of MLP and ResNet studied in [18] to general DAG networks. One can first decompose the induction hypothesis:

Condition E.1.1. *At the k-th iteration, we have*

$$\|\boldsymbol{y} - \boldsymbol{u}(k)\|_2^2 \leq (1 - \frac{\eta \lambda_0}{2})^k \|\boldsymbol{y} - \boldsymbol{u}(0)\|_2^2 .$$

into:

$$\begin{aligned}
&\|\boldsymbol{y} - \boldsymbol{u}(k+1)\|_2^2 \\
=& \|\boldsymbol{y} - \boldsymbol{u}(k) - (\boldsymbol{u}(k+1) - \boldsymbol{u}(k))\|_2^2 \\
=& \|\boldsymbol{y} - \boldsymbol{u}(k)\|_2^2 - 2\left(\boldsymbol{y} - \boldsymbol{u}(k)\right)^\top (\boldsymbol{u}(k+1) - \boldsymbol{u}(k)) + \|\boldsymbol{u}(k+1) - \boldsymbol{u}(k)\|_2^2 \\
=& \|\boldsymbol{y} - \boldsymbol{u}(k)\|_2^2 - 2\left(\boldsymbol{y} - \boldsymbol{u}(k)\right)^\top \boldsymbol{I}_1(k) - 2\left(\boldsymbol{y} - \boldsymbol{u}(k)\right)^\top \boldsymbol{I}_2(k) + \|\boldsymbol{u}(k+1) - \boldsymbol{u}(k)\|_2^2 \\
\leq& \left(1 - \eta \lambda_{\min}\left(\boldsymbol{G}^{(H)}(k)\right)\right) \|\boldsymbol{y} - \boldsymbol{u}(k)\|_2^2 - 2\left(\boldsymbol{y} - \boldsymbol{u}(k)\right)^\top \boldsymbol{I}_2(k) + \|\boldsymbol{u}(k+1) - \boldsymbol{u}(k)\|_2^2 ,
\end{aligned} \tag{18}$$

where $\boldsymbol{G}^{(H)}(k) = \sum_{s=0}^{H-1} \boldsymbol{G}^{(s,H)}(k)$ as the NTK of incoming edges into the last node, composed by the NTK of each edge.

This induction implies that the convergence rate of gradient descent dynamics of DAG networks with an MSE loss is linear. To prove the induction in Theorem 3.1, three core steps are:

Step 1. Show at initialization, $\lambda_{\min}\left(\boldsymbol{G}^{(H)}(0)\right) \geq \frac{3\lambda_{\min}(\boldsymbol{K}^{(H)})}{4}$ when m is sufficiently large.

Step 2. Show that during gradient descent, $\left\|\boldsymbol{G}^{(H)}(k) - \boldsymbol{G}^{(H)}(0)\right\|_2 \leq \frac{1}{4}\lambda_{\min}(\boldsymbol{K}^{(H)})$ when m is sufficiently large.

Step 3. Show the last two terms, both $-2\left(\boldsymbol{y} - \boldsymbol{u}(k)\right)^\top \boldsymbol{I}_2(k)$ and $\|\boldsymbol{u}(k+1) - \boldsymbol{u}(k)\|_2^2$ are proportional to $\eta^2 \|\boldsymbol{y} - \boldsymbol{u}(k)\|_2^2$, and if we set η sufficiently small, they are smaller than $\eta \lambda_{\min}\left(\boldsymbol{G}^{(H)}(k)\right) \|\boldsymbol{y} - \boldsymbol{u}(k)\|_2^2$, leading the MSE loss linearly decrease.

In our proof, we mainly focus on deriving the condition on m (the required number neurons per layer) and η (learning rate) in our DAG setting, by analyzing $\lambda_{\min}\left(\boldsymbol{G}^{(H)}(0)\right)$. Our proof relies on lemmas below, which are based on first analyzing the gradient on each path and then considering the DAG as a union bound over all P_H number of end-to-end paths.

□

We want to point out another benefit of bounding both $\lambda_{\min}\left(\boldsymbol{G}^{(H)}(k)\right)$ and the linear convergence rate via $\lambda_{\min}(\boldsymbol{K}^{(H)})$. Since we try to find the DAG-specific propagation from $\lambda_{\min}(\boldsymbol{K}^{(0)})$ to $\lambda_{\min}(\boldsymbol{K}^{(H)})$, and all DAGs share the same $\lambda_{\min}(\boldsymbol{K}^{(0)})$, we can more fairly compare the bound of convergence rate of different DAGs.

E.1 Lemmas

Definition E.2. The geometric series function:

$$g_\alpha(n) = \sum_{i=0}^{n-1} \alpha^i$$

Lemma E.1 (Lemma on Initialization Norms)**.** *If $\sigma(\cdot)$ is a $L-$Lipschitz activation and $m = \Omega\left(\frac{NP_H^2 H g_C(H)^2}{\delta}\right)$, where $C \triangleq c_\sigma L\left(2\left|\sigma(0)\right|\sqrt{\frac{2}{\pi}} + 2L\right)$ and P_H is the total number of end-to-end paths, then with probability at least $1-\delta$ over random initialization, for every $h \in [H]$ and $i \in [N]$, we have $0 \le \left\|\boldsymbol{X}_i^{(h)}(0)\right\|_2 \le c_{x,0}$, where $c_{x,0} = 2P_H \max(1, c_{w,0}L)$ and $c_{w,0}$ is a universal constant.*

We follow the proof sketch described in Section E. We first analyze the spectral property of $\boldsymbol{G}^{(H)}(0)$ at the initialization phase. The following lemma lower bounds its least eigenvalue.

Lemma E.2 (Least Eigenvalue at the Initialization)**.** *If $m = \Omega\left(\frac{(NP_H)^2 \log(HN/\delta) 2^{O(H)}}{\lambda_{\min}(\boldsymbol{K}^{(0)})^2}\right)$, we have*

$$\lambda_{\min}(\boldsymbol{G}^{(H)}(0)) \ge \frac{3}{4}\lambda_{\min}(\boldsymbol{K}^{(0)}).$$

Next, we characterize how the perturbation on the weight matrices affects the input of each layer.

Lemma E.3. *Suppose $\sigma(\cdot)$ is L-Lipschitz and for $h \in [H]$, $\left\|\boldsymbol{W}^{(s,t)}(0)\right\|_2 \le c_{w,0}\sqrt{m}$, $\left\|\boldsymbol{X}^{(t)}(0)\right\|_2 \le c_{x,0}$ and $\left\|\boldsymbol{W}^{(s,t)}(k) - \boldsymbol{W}^{(s,t)}(0)\right\|_F \le \sqrt{m}R$ for some constant $c_{w,0}, c_{x,0} > 0$ and $R \le c_{w,0}$. Then we have*

$$\left\|\boldsymbol{X}^{(t)}(k) - \boldsymbol{X}^{(t)}(0)\right\|_2 \le 2\prod_{s=1}^{t} P_s\sqrt{c_\sigma}c_{w,0}L + \sum_{s=1}^{t} P_s\sqrt{c_\sigma}Lc_{x,0}R.$$

Next, we characterize how the perturbation on the weight matrices affect $\boldsymbol{G}^{(H)}$.

Lemma E.4. *Suppose $\sigma(\cdot)$ is differentiable, $L-$Lipschitz and $\beta-$smooth. Suppose for $s,t \in [H]$, $\left\|\boldsymbol{W}^{(s,t)}(0)\right\|_2 \le c_{w,0}\sqrt{m}$, $\|\boldsymbol{a}(0)\|_2 \le a_{2,0}\sqrt{m}$, $\|\boldsymbol{a}(0)\|_4 \le a_{4,0}m^{1/4}$, $0 \le \left\|\boldsymbol{x}^{(t)}(0)\right\|_2 \le c_{x,0}$, if $\left\|\boldsymbol{W}^{(s,t)}(k) - \boldsymbol{W}^{(s,t)}(0)\right\|_F, \|\boldsymbol{a}(k) - \boldsymbol{a}(0)\|_2 \le \sqrt{m}R$ where $R \le c\lambda_0(P_H N)^{-1}$ and $R \le c$ for some small constant c, we have*

$$\left\|\boldsymbol{G}^{(H)}(k) - \boldsymbol{G}^{(H)}(0)\right\|_2 \le \frac{\lambda_0}{2}.$$

Lemma E.5. *If Condition E.1.1 holds for $1,\ldots,k$, we have for any $k' \in [k+1]$ and any $s,t \in [H]$*

$$\left\|\boldsymbol{W}^{(s,t)}(k') - \boldsymbol{W}^{(s,t)}(0)\right\|_F, \|\boldsymbol{a}(k') - \boldsymbol{a}(0)\|_2 \le R\sqrt{m},$$
$$\left\|\boldsymbol{W}^{(s,t)}(k') - \boldsymbol{W}^{(s,t)}(k'-1)\right\|_F, \|\boldsymbol{a}(k') - \boldsymbol{a}(k'-1)\|_2 \le \eta Q(k'-1),$$

where $R = P_H \frac{16\sqrt{c_\sigma}c_{x,0}a_{2,0}Le^{2\sqrt{c_\sigma}c_{w,0}L}\sqrt{N}\|\boldsymbol{y}-\boldsymbol{u}(0)\|_2}{\lambda_0\sqrt{m}} < c$ for some small constant c and $Q(k') = 4P_H\sqrt{c_\sigma}c_{x,0}a_{2,0}Le^{2\sqrt{c_\sigma}c_{w,0}L}\sqrt{N}\,\|\boldsymbol{y} - \boldsymbol{u}(k')\|_2$.

The next lemma bounds the $\boldsymbol{I}_2$ term.

Lemma E.6. *If Condition E.1.1 holds for $1,\ldots,k$ and $\eta \le c\lambda_0(NP_H)^{-2}$ for some small constant c, we have*

$$\|\boldsymbol{I}_2(k)\|_2 \le \frac{1}{8}\eta\lambda_0\,\|\boldsymbol{y} - \boldsymbol{u}(k)\|_2\,.$$

Next we bound the quadratic term.

Lemma E.7. *If Condition E.1.1 holds for $1,\ldots,k$ and $\eta \le c\lambda_0(NP_H)^{-2}$ for some small constant c, we have $\|\boldsymbol{u}(k+1) - \boldsymbol{u}(k)\|_2^2 \le \frac{1}{8}\eta\lambda_0\,\|\boldsymbol{y} - \boldsymbol{u}(k)\|_2^2$.*

E.2 Proofs of Lemmas

Proof of Lemma E.1. We will bound $\left\|\boldsymbol{X}_i^{(t)}(0)\right\|_2$ by induction on layers. For the first layer, when $\boldsymbol{W}^{(0,1)}$ is a parameterized linear operator, we can calculate:

$$\begin{aligned}
\mathbb{E}\left[\left\|\boldsymbol{X}_i^{(1)}(0)\right\|_2^2\right] =& c_\sigma \mathbb{E}\left[\sigma\left(\boldsymbol{W}^{(0,1)}(0)^\top \boldsymbol{X}_i\right)^2\right] \\
=& c_\sigma \mathbb{E}_{X\sim N(0,1)} \sigma(X)^2 \\
=& 1. \\
\mathrm{Var}\left[\left\|\boldsymbol{X}_i^{(1)}(0)\right\|_2^2\right] =& \frac{c_\sigma^2}{m} \mathrm{Var}\left[\sigma\left(\boldsymbol{W}^{(0,1)}(0)^\top \boldsymbol{X}_i(0)\right)^2\right] \\
\leq& \frac{c_\sigma^2}{m} \mathbb{E}_{X\sim N(0,1)} \sigma(X)^4 \\
\leq& \frac{c_\sigma^2}{m} \mathbb{E}\left[\left(|\sigma(0)| + L\left|\boldsymbol{W}^{(0,1)}(0)^\top \boldsymbol{X}_i\right|\right)^4\right] \\
\leq& \frac{C_2}{m},
\end{aligned}$$

where $C_2 \triangleq \sigma(0)^4 + 4|\sigma(0)|^3 L\sqrt{2/\pi} + 6\sigma(0)^2 L^2 + 8|\sigma(0)| L^3 \sqrt{2/\pi} + 32L^4$. Considering $\boldsymbol{W}^{(0,1)}$ could be a zero or an identity mapping, We have with probability at least $1 - \frac{\delta}{n}$,

$$0 \leq \left\|\boldsymbol{X}_i^{(1)}(0)\right\|_2 \leq 2.$$

By definition we have for $2 \leq t \leq H$,

$$0 \leq \left\|\boldsymbol{X}_i^{(t)}(0)\right\|_2 \leq \sum_{s=0}^{P_t} \left\|\sqrt{\frac{c_\sigma}{m}} \sigma\left(\boldsymbol{W}^{(s,t)}(0) \boldsymbol{X}_i^{(s)}(0)\right)\right\|_2.$$

Thus we have:

$$\begin{aligned}
\mathrm{Var}\left[\left\|\boldsymbol{X}_i^{(t)}(0)\right\|_2^2\right] &= \mathrm{Var}\left[\left\|\sum_{s=0}^{P_t} \sqrt{\frac{c_\sigma}{m}} \sigma\left(\boldsymbol{W}^{(s,t)}(0) \boldsymbol{X}_i^{(s)}(0)\right)\right\|_2^2\right] \\
&\leq \sum_{s_1=0}^{P_t} \sum_{s_2=0}^{P_t} \frac{C_2}{m} = P_t^2 \frac{C_2}{m}.
\end{aligned}$$

Applying Chebyshev's inequality and plugging in our assumption on m, we have with probability $1 - \frac{\delta}{NH}$ over $\boldsymbol{W}^{(s,t)}$,

$$\left|\left\|\boldsymbol{X}_i^{(t)}(0)\right\|_2^2 - \mathbb{E}\left\|\boldsymbol{X}_i^{(t)}(0)\right\|_2^2\right| \leq \frac{1}{2g_C(H)}.$$

Since $\boldsymbol{W}^{(s,t)}$ could be linear, identity, or zero mapping, we have:

$$\left\|\sqrt{\frac{c_\sigma}{m}} \sigma\left(\boldsymbol{W}^{(s,t)}(0) \boldsymbol{X}_i^{(s)}(0)\right)\right\|_2 \leq \left\|\boldsymbol{X}_i^{(s)}(0)\right\|_2 \max(1, c_{w,0} L).$$

Thus

$$0 \leq \left\|\boldsymbol{X}_i^{(t)}(0)\right\|_2 \leq \max(1, c_{w,0} L) \sum_{s=0}^{P_t} \left\|\boldsymbol{X}_i^{(s)}(0)\right\|_2,$$

which implies

$$0 \leq \left\|\boldsymbol{X}^{(t)}(0)\right\|_2 \leq 2P_H \max(1, c_{w,0} L).$$

Choosing $c_{x,0} = P_H c_{w,0} L$ and using union bounds over $[n]$, we prove the lemma.

□

Proof of Lemma E.2. The purpose of this proof is to study how large m is needed to ensure the randomly generated Gram matrices ($\boldsymbol{X}^\top \boldsymbol{X}$) is close to the population Gram matrices ($\boldsymbol{K}$) in networks of general DAG connectivity patterns, via extending the general framework in Section E in [18]. For complete definitions, we refer readers to Section E.1 in [18].

Theorem E.3. *With probability* $1-\delta$ *over the* $\left\{\boldsymbol{W}^{(s,t)}\right\}_{s,t}$, *suppose* $m = \Omega\left(\frac{(NP_H)^2 \log\left(\frac{HN}{\delta}\right) 2^{O(H)}}{\lambda_{\min}^2\left(\boldsymbol{K}^{(H)}\right)}\right)$, *for any* $1 \le t \le H-1, 1 \le i,j \le N$,

$$\left\| \frac{1}{m} \sum_{\alpha=1}^{m} (\boldsymbol{X}_i^{(t),(\alpha)})^\top \boldsymbol{X}_j^{(t),(\alpha)} - \boldsymbol{K}_{ij}^{(t)} \right\|_\infty \le \mathcal{E} \sqrt{\frac{\log(HN/\delta)}{m}} \tag{19}$$

and any $t \in [H-1], \forall 1 \le i \le N$,

$$\left\| \frac{1}{m} \sum_{\alpha=1}^{m} \boldsymbol{X}_i^{(t),(\alpha)} - \boldsymbol{b}_i^{(t)} \right\|_\infty \le \mathcal{E} \sqrt{\frac{\log(HN/\delta)}{m}} \tag{20}$$

The error constant $\mathcal{E}$ *satisfies there exists an absolute constant* $C > 0$ *such that*

$$\mathcal{E} \le C \left(\prod_{t=1}^{H-1} \left(|P_0^{(t)}| + \sum_{h' \in P_1^{(t)}} \left(\Lambda \mathfrak{W}_{(t)} + 2CB\mathfrak{W}_{(t)} + 2C\sqrt{\mathfrak{W}_{(t)} M} \right) \right) \right) \times \max\{\mathfrak{W}_{(t)} \sqrt{(1+C^2)M^2}, \sqrt{C^2 M}\}$$

where $P_0^{(t)}$ *and* $P_1^{(t)}$ *are sets of indices of nodes with skip-connection and parameterized layers into* $\boldsymbol{X}^{(s)}$, *respectively. We have* $M = O(1), B = O(1), C = O(1), \Lambda = O(1), \mathfrak{W}_{(t)} = O(1)$, *and they are defined by:*

- $\mathcal{W}^{(s,t)}\left(\boldsymbol{K}^{(s)}\right) = \mathbb{E}_{\boldsymbol{W}}\left[\boldsymbol{W}^{(s,t)} \boldsymbol{K}^{(s)} (\boldsymbol{W}^{(s,t)})^\top\right]$.
- $M = 1 + 100 \max_{i,j,s,t} |\mathcal{W}^{(s,t)}(\boldsymbol{K}_{ij}^{(s)})|$,
- $B = 1 + 100 \max_{i,t} |\boldsymbol{b}_i^{(t)}|$,
- $C = |\rho(0)| + \sup_{x \in \mathbb{R}} |\rho'(x)|$,
- Λ *is a constant that only depends on* ρ,
- $\mathfrak{W}_{(t)} = 1 + \max_s \|\mathcal{W}^{(s,t)}\|_{L^\infty \to L^\infty}$.

Proof of Theorem E.3. The proof is by induction. Suppose that Equation (19) and (20) hold for $1 \le s \le t$ with probability at least $1 - \frac{t}{H}\delta$, now we want to show the equations holds for $t+1$ with probability at least $1 - \delta/H$ conditioned on previous layers satisfying Equation (19) and (20). Let $s = t+1$. recall

$$\boldsymbol{X}_i^{(s),(\alpha)} = \sum_{h' \in P_0^{(s)}} \boldsymbol{X}^{(h')} + \sum_{h' \in P_1^{(s)}} \rho^{(h')} \left(\frac{\sum_\beta \boldsymbol{W}_{(\beta)}^{(h',h),(\alpha)} \boldsymbol{X}_i^{(h'),(\beta)}}{\sqrt{m}} \right),$$

Now we get the following formula for the expectation:

$$\begin{aligned}
\mathbb{E}^{(s)}[\hat{\boldsymbol{K}}_{ij}^{(s)}] &= \sum_{h' \in P_0^{(s)}} \hat{\boldsymbol{K}}_{ij}^{(h')} + \sum_{h' \in P_1^{(s)}} \mathbb{E}_{(\boldsymbol{U},\boldsymbol{V})} \left(\rho(\boldsymbol{U})^\top (\hat{\boldsymbol{b}}_j^{(h')}) + (\hat{\boldsymbol{b}}_i^{(h')})^\top \rho(\boldsymbol{V}) + \rho(\boldsymbol{U})^\top \rho(\boldsymbol{V})) \right) \\
\mathbb{E}^{(s)} \hat{\boldsymbol{b}}_i^{(s)} &= \sum_{h' \in P_0^{(s)}} \hat{\boldsymbol{b}}_i^{(h')} + \sum_{h' \in P_1^{(s)}} \mathbb{E}_{\boldsymbol{U}} \rho(\boldsymbol{U})
\end{aligned}$$

with

$$(\boldsymbol{U}, \boldsymbol{V}) \sim N\left(\boldsymbol{0}, \begin{pmatrix} \mathcal{W}^{(h',s)}(\hat{\boldsymbol{K}}_{ii}^{(h')}) & \mathcal{W}^{(h',s)}(\hat{\boldsymbol{K}}_{ij}^{(h')}) \\ \mathcal{W}^{(h',s)}(\hat{\boldsymbol{K}}_{ji}^{(h')}) & \mathcal{W}^{(h',s)}(\hat{\boldsymbol{K}}_{jj}^{(h')}) \end{pmatrix}\right)$$

To bound the differences that determine how the error propagates through layers:

$$\max_{ij} \left\| \mathbb{E}^{(s)} \hat{\boldsymbol{K}}_{ij}^{(s)} - \boldsymbol{K}_{ij}^{(s)} \right\|_\infty \text{ and } \max_{i} \left\| \mathbb{E}^{(s)} \hat{\boldsymbol{b}}_{ij}^{(s)} - \boldsymbol{b}_{ij}^{(s)} \right\|_\infty ,$$

we analyze the error:

$$\begin{aligned}
&\left\| \mathbb{E}^{(s)} \hat{\boldsymbol{K}}_{ij}^{(s)} - \boldsymbol{K}_{ij}^{(s)} \right\|_\infty \\
\leq &\sum_{h' \in P_0^{(s)}} \left\| \hat{\boldsymbol{K}}_{ij}^{(h')} - \boldsymbol{K}_{ij}^{(h')} \right\|_\infty \\
&+ \sum_{h' \in P_1^{(s)}} (\left\| \mathbb{E}_{(\boldsymbol{U},\boldsymbol{V})\sim\hat{\boldsymbol{A}}} \rho(\boldsymbol{U})^\top (\hat{\boldsymbol{b}}_j^{(h')}) - \mathbb{E}_{(\boldsymbol{U},\boldsymbol{V})\sim\hat{\boldsymbol{A}}} \rho(\boldsymbol{U})^\top (\boldsymbol{b}_j^{(h')}) \right\|_\infty \\
&+ \left\| \mathbb{E}_{(\boldsymbol{U},\boldsymbol{V})\sim\hat{\boldsymbol{A}}} \rho(\boldsymbol{U})^\top (\boldsymbol{b}_j^{(h')}) - \mathbb{E}_{(\boldsymbol{U},\boldsymbol{V})\sim\boldsymbol{A}} \rho(\boldsymbol{U})^\top (\boldsymbol{b}_j^{(h')}) \right\|_\infty \\
&+ \left\| \mathbb{E}_{(\boldsymbol{U},\boldsymbol{V})\sim\hat{\boldsymbol{A}}} (\hat{\boldsymbol{b}}_i^{(h')})^\top \rho(\boldsymbol{V}) - \mathbb{E}_{(\boldsymbol{U},\boldsymbol{V})\sim\hat{\boldsymbol{A}}} ((\boldsymbol{b}_i^{(h')})^\top \rho(\boldsymbol{V}) \right\|_\infty \\
&+ \left\| \mathbb{E}_{(\boldsymbol{U},\boldsymbol{V})\sim\hat{\boldsymbol{A}}} (\boldsymbol{b}_i^{(h')})^\top \rho(\boldsymbol{V}) - \mathbb{E}_{(\boldsymbol{U},\boldsymbol{V})\sim\boldsymbol{A}} (\boldsymbol{b}_i^{(h')})^\top \rho(\boldsymbol{V}) \right\|_\infty \\
&+ \left\| \mathbb{E}_{(\boldsymbol{U},\boldsymbol{V})\sim\hat{\boldsymbol{A}}} \rho(\boldsymbol{U})^\top \rho(\boldsymbol{V})) - \mathbb{E}_{(\boldsymbol{U},\boldsymbol{V})\sim\boldsymbol{A}} \rho(\boldsymbol{U})^\top \rho(\boldsymbol{V})) \right\|_\infty)
\end{aligned}$$

where we define

$$\hat{\boldsymbol{A}} = \begin{pmatrix} \mathcal{W}^{(h',s)}(\hat{\boldsymbol{K}}_{ii}^{(h')}) & \mathcal{W}^{(h',s)}(\hat{\boldsymbol{K}}_{ij}^{(h')}) \\ \mathcal{W}^{(h',s)}(\hat{\boldsymbol{K}}_{ji}^{(h')}) & \mathcal{W}^{(h',s)}(\hat{\boldsymbol{K}}_{jj}^{(h')}) \end{pmatrix} \text{ and } \boldsymbol{A} = \begin{pmatrix} \mathcal{W}^{(h',s)}(\boldsymbol{K}_{ii}^{(h')}) & \mathcal{W}^{(h',s)}(\boldsymbol{K}_{ij}^{(h')}) \\ \mathcal{W}^{(h',s)}(\boldsymbol{K}_{ji}^{(h')}) & \mathcal{W}^{(h',s)}(\boldsymbol{K}_{jj}^{(h')}) \end{pmatrix}$$

By definition, we have

$$\|\boldsymbol{A} - \hat{\boldsymbol{A}}\|_\infty \leq \mathfrak{W}_{(s)} \max_{ij} \|\hat{\boldsymbol{K}}_{ij}^{(h')} - \boldsymbol{K}_{ij}^{(h')}\|_\infty$$

We can also estimate other terms

$$\begin{aligned}
&\left\| \mathbb{E}_{(\boldsymbol{U},\boldsymbol{V})\sim\hat{\boldsymbol{A}}} \rho(\boldsymbol{U})^\top (\hat{\boldsymbol{b}}_j^{(h')}) - \mathbb{E}_{(\boldsymbol{U},\boldsymbol{V})\sim\hat{\boldsymbol{A}}} \rho(\boldsymbol{U})^\top (\boldsymbol{b}_j^{(h')}) \right\|_\infty \\
\leq &\left\| \mathbb{E}_{(\boldsymbol{U},\boldsymbol{V})\sim\hat{\boldsymbol{A}}} \rho(\boldsymbol{U})^\top \left(\hat{\boldsymbol{b}}_j^{(h')} - \boldsymbol{b}_j^{(h')} \right) \right\|_\infty \\
\leq &C \sqrt{\mathfrak{W}_{(s)} \max_{ij} \|\hat{\boldsymbol{K}}_{ij}^{(s)}\|_\infty} \max_i \left\| \hat{\boldsymbol{b}}_{ij}^{(h')} - \boldsymbol{b}_{ij}^{(h')} \right\|_\infty \\
\leq &C \sqrt{\mathfrak{W}_{(s)} M} \max_i \left\| \hat{\boldsymbol{b}}_{ij}^{(h')} - \boldsymbol{b}_{ij}^{(h')} \right\|_\infty ,
\end{aligned}$$

$$\begin{aligned}
&\left\| \mathbb{E}_{(\boldsymbol{U},\boldsymbol{V})\sim\hat{\boldsymbol{A}}} \rho(\boldsymbol{U})^\top (\boldsymbol{b}_j^{(h')}) - \mathbb{E}_{(\boldsymbol{U},\boldsymbol{V})\sim\boldsymbol{A}} \rho(\boldsymbol{U})^\top (\boldsymbol{b}_j^{(h')}) \right\|_\infty \\
\leq &BC \|\boldsymbol{A} - \hat{\boldsymbol{A}}\|_\infty \\
\leq &BC \mathfrak{W}_{(s)} \max_{ij} \|\hat{\boldsymbol{K}}_{ij}^{(h')} - \boldsymbol{K}_{ij}^{(h')}\|_\infty ,
\end{aligned}$$

and

$$\begin{aligned}
&\left\| \mathbb{E}_{(\boldsymbol{U},\boldsymbol{V})\sim\hat{\boldsymbol{A}}} \rho(\boldsymbol{U})^\top \rho(\boldsymbol{V}) - \mathbb{E}_{(\boldsymbol{U},\boldsymbol{V})\sim\boldsymbol{A}} \rho(\boldsymbol{U})^\top \rho(\boldsymbol{V}) \right\|_\infty \\
\leq &\Lambda \|\boldsymbol{A} - \hat{\boldsymbol{A}}\|_\infty \\
\leq &\Lambda \mathfrak{W}_{(s)} \max_{ij} \|\hat{\boldsymbol{K}}_{ij}^{(h')} - \boldsymbol{K}_{ij}^{(h')}\|_\infty .
\end{aligned}$$

Putting these estimates together, we have

$$
\begin{aligned}
&\max_{ij} \|\mathbb{E}^{(s)} \hat{\boldsymbol{K}}_{ij}^{(s)} - \boldsymbol{K}_{ij}^{(s)}\|_\infty \\
&\leq \sum_{h' \in P_0^{(s)}} \left(\max_{ij} \|\hat{\boldsymbol{K}}_{ij}^{(h')} - \boldsymbol{K}_{ij}^{(h')}\|_\infty \vee \max_i \|\hat{\boldsymbol{b}}_{ij}^{(h')} - \boldsymbol{b}_{ij}^{(h')}\|_\infty \right) \\
&+ \sum_{h' \in P_1^{(s)}} \left(\Lambda \mathfrak{W}_{(s)} + 2CB\mathfrak{W}_{(s)} + 2C\sqrt{\mathfrak{W}_{(s)} M} \right) \left(\max_{ij} \|\hat{\boldsymbol{K}}_{ij}^{(h')} - \boldsymbol{K}_{ij}^{(h')}\|_\infty \vee \max_i \|\hat{\boldsymbol{b}}_{ij}^{(h')} - \boldsymbol{b}_{ij}^{(h')}\|_\infty \right)
\end{aligned}
$$

and

$$
\begin{aligned}
\max_i \left\| \mathbb{E}^{(s)} \hat{\boldsymbol{b}}_{ij}^{(s)} - \boldsymbol{b}_{ij}^{(s)} \right\|_\infty &\leq \sum_{h' \in P_0^{(s)}} \left(\max_{ij} \|\hat{\boldsymbol{K}}_{ij}^{(h')} - \boldsymbol{K}_{ij}^{(h')}\|_\infty \vee \max_i \|\hat{\boldsymbol{b}}_{ij}^{(h')} - \boldsymbol{b}_{ij}^{(h')}\|_\infty \right) \\
&+ \sum_{h' \in P_1^{(s)}} \left(\Lambda \mathfrak{W}_{(s)} \right) \left(\max_{ij} \|\hat{\boldsymbol{K}}_{ij}^{(h')} - \boldsymbol{K}_{ij}^{(h')}\|_\infty \vee \max_i \|\hat{\boldsymbol{b}}_{ij}^{(h')} - \boldsymbol{b}_{ij}^{(h')}\|_\infty \right).
\end{aligned}
$$

These two bounds imply the theorem.

□

Remark E.4. Combing Theorem E.3 and Lemma E.1 in [18] and standard matrix perturbation bound directly have Lemma E.2.

□

Proof of Lemma E.3. We prove this lemma by induction. Our induction hypothesis is

$$
\left\| \boldsymbol{X}^{(t)}(k) - \boldsymbol{X}^{(t)}(0) \right\|_2 \leq g(t),
$$

where

$$
g(t) = \sqrt{c_\sigma} P_t \left(2g(t-1) c_{w,0} L + L c_{x,0} R \right).
$$

For $h = 1$, we have

$$
\begin{aligned}
\left\| \boldsymbol{X}^{(1)}(k) - \boldsymbol{X}^{(1)}(0) \right\|_2 &\leq \sqrt{\frac{c_\sigma}{m}} \left\| \sigma\left(\boldsymbol{W}^{(0,1)}(k) \boldsymbol{X} \right) - \sigma\left(\boldsymbol{W}^{(0,1)}(0) \boldsymbol{X} \right) \right\|_2 \\
&\leq \sqrt{\frac{c_\sigma}{m}} \left\| \boldsymbol{W}^{(0,1)}(k) - \boldsymbol{W}^{(0,1)}(0) \right\|_F \leq \sqrt{c_\sigma} L R,
\end{aligned}
$$

which implies $g(1) = \sqrt{c_\sigma} L R$. For $2 \leq t \leq H$, following the proof of Lemma C.3. in [18], we have

$$
\begin{aligned}
\left\| \boldsymbol{X}^{(t)}(k) - \boldsymbol{X}^{(t)}(0) \right\|_2 &= \sqrt{\frac{c_\sigma}{m}} \left\| \sum_{s=0}^{P_t} \sigma\left(\boldsymbol{W}^{(s,t)}(k) \boldsymbol{X}^{(s)}(k) \right) - \sigma\left(\boldsymbol{W}^{(s,t)}(0) \boldsymbol{x}^{(s)}(0) \right) \right\|_2 \\
&\leq \sqrt{\frac{c_\sigma}{m}} \sum_{s=0}^{P_t} \left\| \sigma\left(\boldsymbol{W}^{(s,t)}(k) \boldsymbol{X}^{(s)}(k) \right) - \sigma\left(\boldsymbol{W}^{(s,t)}(0) \boldsymbol{x}^{(s)}(0) \right) \right\|_2 \\
&\leq P_t \left(2\sqrt{c_\sigma} c_{w,0} L g(t-1) + \sqrt{c_\sigma} L c_{x,0} R \right).
\end{aligned}
$$

Lastly, simple calculations show $g(t) \leq 2 \prod_{s=1}^t P_s \sqrt{c_\sigma} c_{w,0} L + \sum_{s=1}^t P_s \sqrt{c_\sigma} L c_{x,0} R$. □

Proof of Lemma E.4. Following the proof of Lemma C.4. in [18], we can obtain

$$
\left| \boldsymbol{G}_{i,j}^{(H)}(k) - \boldsymbol{G}_{i,j}^{(H)}(0) \right| \leq P_H \left(I_1^{i,j} + I_2^{i,j} + I_3^{i,j} \right).
$$

Therefore we can bound the perturbation

$$
\begin{aligned}
\left\| \boldsymbol{G}^{(H)}(k) - \boldsymbol{G}^{(H)}(0) \right\|_F &= \sqrt{\sum_{(i,j)}^{N,N} \left| \boldsymbol{G}_{i,j}^{(H)}(k) - \boldsymbol{G}_{i,j}^{(H)}(0) \right|^2} \\
&\leq P_H L^2 N R \left[3c_{x,0} c_x a_{2,0}^2 + 2\beta c_{x,0}^2 \left(2c_x c_{w,0} + c_{x,0} \right) a_{4,0}^2 + 12 c_{x,0}^2 a_{2,0} \right],
\end{aligned}
$$

where $c_x \triangleq 2\prod_{s=1}^{t} P_s\sqrt{c_\sigma}c_{w,0}\frac{L}{R} + \sum_{s=1}^{t} P_s\sqrt{c_\sigma}Lc_{x,0}$. Plugging in the bound on R, we have the desired result.

□

Proof of Lemma E.5. We will prove this corollary by induction. The induction hypothesis is

$$\left\|\boldsymbol{W}^{(s,t)}(k') - \boldsymbol{W}^{(s,t)}(0)\right\|_F \leq \sum_{k''=0}^{k'-1}(1-\frac{\eta\lambda_0}{2})^{k''/2}\frac{1}{4}\eta\lambda_0 R\sqrt{m} \leq R\sqrt{m}, k' \in [k+1],$$

$$\left\|\boldsymbol{a}(k') - \boldsymbol{a}(0)\right\|_2 \leq \sum_{k''=0}^{k'-1}(1-\frac{\eta\lambda_0}{2})^{k''/2}\frac{1}{4}\eta\lambda_0 R\sqrt{m} \leq R\sqrt{m}, k' \in [k+1].$$

First it is easy to see it holds for $k''=0$. Now suppose it holds for $k''=0,\ldots,k'$, we consider $k''=k'+1$. Applying union bound over P_H number of end-to-end paths, we have

$$\begin{aligned}
&\left\|\boldsymbol{W}^{(s,t)}(k'+1) - \boldsymbol{W}^{(s,t)}(k')\right\|_F \\
\leq&\eta P_t L\sqrt{\frac{c_\sigma}{m}}\left\|\boldsymbol{a}\right\|_2\sum_{i=1}^{N}|y_i - u_i(k')|\left\|\boldsymbol{X}_i^{(s-1)}(k')\right\|_2\prod_{s'=t+1}^{H}\left\|\boldsymbol{I}+\lambda_0^{3/2}\sqrt{\frac{c_\sigma}{m}}\boldsymbol{J}_i^{(s',s)}(k')\boldsymbol{W}^{(s,s')}(k')\right\|_2 \\
\leq&2\eta P_t\sqrt{c_\sigma}c_{x,0}La_{2,0}e^{2\sqrt{c_\sigma}c_{w,0}L}\sqrt{N}\left\|\boldsymbol{y}-\boldsymbol{u}(k')\right\|_2 \\
=&\eta P_t Q(k') \\
\leq&P_H(1-\frac{\eta\lambda_0}{2})^{k'/2}\frac{1}{4}\eta\lambda_0 R\sqrt{m},
\end{aligned}$$

where

$$\boldsymbol{J}_i^{(s',s)} \triangleq \boldsymbol{diag}\left(\rho'\left((\boldsymbol{W}_1^{(s',s)})^\top\boldsymbol{X}_i^{(s')}\right),\ldots,\rho'\left((\boldsymbol{W}_m^{(s',s)})^\top\boldsymbol{X}_i^{(s')}\right)\right) \in \mathbb{R}^{m\times m}$$

are the derivative matrices induced by the activation function. Similarly, we have

$$\begin{aligned}
\left\|\boldsymbol{a}(k'+1) - \boldsymbol{a}(k')\right\|_2 \leq&2\eta c_{x,0}\sum_{i=1}^{N}|y_i - u(k')| \\
\leq&\eta Q(k') \\
\leq&(1-\frac{\eta\lambda_0}{2})^{k'/2}\frac{1}{4}\eta\lambda_0 R\sqrt{m}.
\end{aligned}$$

Thus

$$\begin{aligned}
&\left\|\boldsymbol{W}^{(s,t)}(k'+1) - \boldsymbol{W}^{(s,t)}(0)\right\|_F \\
\leq&\left\|\boldsymbol{W}^{(s,t)}(k'+1) - \boldsymbol{W}^{(s,t)}(k')\right\|_F + \left\|\boldsymbol{W}^{(s,t)}(k') - \boldsymbol{W}^{(s,t)}(0)\right\|_F \\
\leq&\sum_{k''=0}^{k'}\eta(1-\frac{\eta\lambda_0}{2})^{k''/2}\frac{1}{4}\eta\lambda_0 R\sqrt{m}.
\end{aligned}$$

Similarly,

$$\begin{aligned}
&\left\|\boldsymbol{a}(k'+1) - \boldsymbol{a}(0)\right\|_2 \\
\leq&\sum_{k''=0}^{k'}\eta(1-\frac{\eta\lambda_0}{2})^{k''/2}\frac{1}{4}\eta\lambda_0 R\sqrt{m}.
\end{aligned}$$

□

Proof of Lemma E.6. We apply union bound over P_H number of end-to-end paths.

$$
\begin{aligned}
&\left\|\mathcal{L}'^{(s,t)}(k)\right\|_F \triangleq \left\|\frac{\partial \mathcal{L}}{\partial \boldsymbol{W}^{(s,t)}(k)}\right\|_F \\
=&P_t \left\|\sqrt{\frac{c_\sigma}{m}} \sum_{i=1}^{N} (y_i - u_i(k)) \boldsymbol{X}_i^{(s)}(k) \cdot \left[\boldsymbol{a}(k)^\top \prod_{l=t+1}^{H} \left(\boldsymbol{I} + \sqrt{\frac{c_\sigma}{m}} \boldsymbol{J}_i^{(t,l)}(k) \boldsymbol{W}^{(t,l)}(k)\right) \boldsymbol{J}_i^{(t,l)}(k)\right]\right\|_F \\
\leq& P_t L \sqrt{\frac{c_\sigma}{m}} \left\|\boldsymbol{a}(k)\right\|_2 \sum_{i=1}^{N} |y_i - u_i(k)| \left\|\boldsymbol{X}^{(s)}(k)\right\|_2 \prod_{l=t+1}^{H} \left\|\boldsymbol{I} + \sqrt{\frac{c_\sigma}{m}} \boldsymbol{J}_i^{(t,l)}(k) \boldsymbol{W}^{(t,l)}(k)\right\|_2 .
\end{aligned}
$$

We have bounded the RHS in the proof for Lemma E.5, thus

$$
\left\|\mathcal{L}'^{(s,t)}(k)\right\|_F \leq \lambda_0 Q(k).
$$

Denote $\boldsymbol{\theta}(k, k') = \boldsymbol{\theta}(k) - k' \mathcal{L}'(\boldsymbol{\theta}(k))$, we have

$$
\begin{aligned}
&\left\|u_i'^{(s,t)}\left(\boldsymbol{\theta}(k)\right) - u_i'^{(s,t)}\left(\boldsymbol{\theta}(k,k')\right)\right\|_F = \\
&P_t \sqrt{\frac{c_\sigma}{m}} \left\|\boldsymbol{X}_i^{(s)}(k) \boldsymbol{a}(k)^\top \prod_{l=t+1}^{H} \left(\boldsymbol{I} + \sqrt{\frac{c_\sigma}{m}} \boldsymbol{J}_i^{(s,l)}(k) \boldsymbol{W}^{(s,l)}(k)\right) \boldsymbol{J}_i^{(s,t)}(k)\right. \\
&\left. - \boldsymbol{X}_i^{(s)}(k,k') \boldsymbol{a}(k,k')^\top \prod_{l=t+1}^{H} \left(\boldsymbol{I} + \sqrt{\frac{c_\sigma}{m}} \boldsymbol{J}_i^{(s,l)}(k,k') \boldsymbol{W}^{(s,l)}(k,k')\right) \boldsymbol{J}_i^{(s,t)}(k,k')\right\|_F .
\end{aligned}
$$

Through standard calculations, we have

$$
\begin{aligned}
\left\|\boldsymbol{W}^{(s,l)}(k) - \boldsymbol{W}^{(s,l)}(k,s)\right\|_F \leq& \eta Q(k), \\
\left\|\boldsymbol{a}(k) - \boldsymbol{a}(k,s)\right\|_F \leq& \eta Q(k), \\
\left\|\boldsymbol{X}_i^{(s)}(k) - \boldsymbol{X}_i^{(s)}(k,s)\right\|_F \leq& \eta c_x \frac{Q(k)}{\sqrt{m}}, \\
\left\|\boldsymbol{J}^{(s,l)}(k) - \boldsymbol{J}^{(s,l)}(k,s)\right\|_F \leq& 2\left(c_{x,0} + c_{w,0} c_x\right) \eta \beta Q(k).
\end{aligned}
$$

Thus we have

$$
\begin{aligned}
&\left\|u_i'^{(t)}\left(\boldsymbol{\theta}(k)\right) - u_i'^{(t)}\left(\boldsymbol{\theta}(k,k')\right)\right\|_F \\
\leq& 4 c_{x,0} L a_{2,0} e^{2L c_{w,0}} \eta Q(k) \sqrt{\frac{c_\sigma}{m}} \left(\frac{c_x}{c_{x,0}} + \frac{2}{L}\left(c_{x,0} + c_{w,0} c_x\right) \beta \sqrt{m} + 4 c_{w,0}\left(c_{x,0} + c_{w,0} c_x\right) \beta + L + 1\right) \\
\leq& 32 c_\sigma c_{x,0} a_{2,0} e^{2L c_{w,0}} \left(c_{x,0} + c_{w,0} c_x\right) \beta \eta Q(k).
\end{aligned}
$$

Thus:

$$
\left|I_2^i\right| \leq 32 c_\sigma c_{x,0} a_{2,0} e^{2L c_{w,0}} \left(c_{x,0} + c_{w,0} c_x\right) \beta \eta^2 Q(k)^2 \leq \frac{1}{8} \eta \lambda_0 \left\|\boldsymbol{y} - \boldsymbol{u}(k)\right\|_2 ,
$$

where we used the bound of η and that $\left\|\boldsymbol{y} - \boldsymbol{u}(0)\right\|_2 = O(\sqrt{N})$,. □

Proof of Lemma E.7.

$$
\begin{aligned}
&\|\boldsymbol{u}(k+1)-\boldsymbol{u}(k)\|_2^2 = \sum_{i=1}^{N}\left(\boldsymbol{a}(k+1)^\top \boldsymbol{X}_i^{(H)}(k+1) - \boldsymbol{a}(k)^\top \boldsymbol{X}_i^{(H)}(k)\right)^2 \\
&= \sum_{i=1}^{N}\left(\left[\boldsymbol{a}(k+1)-\boldsymbol{a}(k)\right]^\top \boldsymbol{X}_i^{(H)}(k+1) + \boldsymbol{a}(k)^\top\left[\boldsymbol{X}_i^{(H)}(k+1)-\boldsymbol{X}_i^{(H)}(k)\right]\right)^2 \\
&\le 2\|\boldsymbol{a}(k+1)-\boldsymbol{a}(k)\|_2^2 \sum_{i=1}^{N}\left\|\boldsymbol{X}_i^{(H)}(k+1)\right\|_2^2 + 2\|\boldsymbol{a}(k)\|_2^2 \sum_{i=1}^{N}\left\|\boldsymbol{X}_i^{(H)}(k+1)-\boldsymbol{X}_i^{(H)}(k)\right\|_2^2 \\
&\le 8N\eta^2 c_{x,0}^2 Q(k)^2 + 4N\left(\eta a_{2,0} c_x Q(k)\right)^2 \\
&\le \frac{1}{8}\eta\lambda_0\|\boldsymbol{y}-\boldsymbol{u}(k)\|_2^2 .
\end{aligned}
$$

□

F Proof of Lemma 3.1

Lemma 3.1 (Full Rankness of $\mathbf{K}^{(H)}$)**.** *Suppose $\sigma(\cdot)$ is analytic and not a polynomial function. If no parallel data points, then $\lambda_{\min}\left(\mathbf{K}^{(H)}\right) > 0$.*

Proof of Lemma 3.1. First, we know that $\boldsymbol{K}^{(0)}$ is positive definite by our assumption on the data point. The next step is to prove propagation of either skip-connection or linear operation will keep the positive definite property. For a linear operation, the NNGP variance matrix can be expressed as,

$$\boldsymbol{K}_{ij} = \mathbb{E}_{\boldsymbol{W}\sim\mathcal{N}(0,\boldsymbol{I})}[\sigma(\boldsymbol{W}^\top\boldsymbol{X}_i)\sigma(\boldsymbol{W}^\top\boldsymbol{X}_j)]$$

Define the feature map induced by NNGP kernel as $\phi_{\boldsymbol{X}_i}(\boldsymbol{W}) = \sigma(\boldsymbol{W}^\top\boldsymbol{X}_i)\boldsymbol{X}_i$. Then we will show that $\phi_{\boldsymbol{X}_1}(\boldsymbol{W}),\ldots\phi_{\boldsymbol{X}_N}(\boldsymbol{W})$ are linearly independent. Assume that there are $a_i \in \mathbb{R}$ such that,

$$\sum_{i=1}^{N} a_i\phi_{\boldsymbol{X}_i} = \sum_{i=1}^{N} a_i\sigma(\boldsymbol{W}^\top\boldsymbol{X}_i)\boldsymbol{X}_i = 0$$

Differentiating the above equation $(N-2)$ times with respect to $\boldsymbol{W}$, we then have,

$$\sum_{i=1}^{N}(a_i\sigma^{(N)}(\boldsymbol{W}^\top\boldsymbol{X}_i))\boldsymbol{X}_i^{\otimes(N-1)} = 0$$

The linearly independents of $\boldsymbol{X}_i^{\otimes(N-1)}$ can be proven by induction. For $N=2$, $\boldsymbol{X}_1\boldsymbol{X}_1^\top$ and $\boldsymbol{X}_2\boldsymbol{X}_2^\top$ are linearly independent under the non-parallel assumption. Suppose $\{\mathrm{vec}(\boldsymbol{X}_1^{\otimes(N-1)}),\ldots,\mathrm{vec}(\boldsymbol{X}_{N-1}^{\otimes(N-1)})\}$ are linearly independent. Then we prove the linearly independent for case of $\boldsymbol{X}_i^{\otimes N}$ by contradiction. Suppose there exists $\alpha_1,\ldots,\alpha_N \in \mathbb{R}$ not identically 0, such that

$$\sum_{i=1}^{N}\alpha_i\mathrm{vec}(\boldsymbol{X}_i^{\otimes N}) = 0$$

which implies for $j = 1,\ldots,d$

$$\sum_{i=1}^{N}\alpha_i\boldsymbol{X}_{ij}\mathrm{vec}(\boldsymbol{X}_i^{\otimes(N-1)}) = 0$$

Because we have supposed the linearly independence of $\boldsymbol{X}_i^{\otimes(N-1)}$, we know there must exist $j \in [d]$ that $\boldsymbol{X}_{ij} \neq 0$ for all $i \in [N]$. However, this is contradict with the assumption that $\alpha_1,\ldots,\alpha_N \neq 0$. Thus we can conclude that $\{\boldsymbol{X}_i^{\otimes(N-1)}\}_{i=1}^N$ are linearly independent. Finally, we have

$$\lambda_{\min}(\boldsymbol{K}) > 0$$

Then we consider the NNGP for node t:

$$\boldsymbol{K}^{(t)} = \sum_{s \in \mathcal{S}} \boldsymbol{K}^{(s,t)} + \sum_{s_1, s_2 \in \mathcal{S}} \boldsymbol{b}^{(s_1,t)} \boldsymbol{b}^{(s_2,t)}$$

where $\mathcal{S}$ is the set of nodes in the DAG except t. When $\boldsymbol{W}^{(s,t)}$ is the skip-connection operator, then we have $\boldsymbol{K}^{(s,t)} = \boldsymbol{K}^{(s)}$. On the other hand, when $\boldsymbol{W}^{(s,t)}$ corresponds to linear transformation, we have $\boldsymbol{K}_{ij}^{(s,t)} = \mathbb{E}_{\boldsymbol{W} \sim \mathcal{N}(0,\boldsymbol{I})}[\sigma(\boldsymbol{W}^\top \boldsymbol{z}_i)\sigma(\boldsymbol{W}^\top \boldsymbol{z}_j)]$, with $\boldsymbol{z} \in \boldsymbol{Z}$ and $\boldsymbol{Z} = \boldsymbol{D}^{1/2}\boldsymbol{U}^\top$, where $\boldsymbol{U}\boldsymbol{D}\boldsymbol{U}^\top = \boldsymbol{K}^{(m)}$. Thus for each $\boldsymbol{K}_{ij}^{(s,t)}$ we can apply the result above and conclude each matrix is positive definite.

In addition, in each matrix $\boldsymbol{b}^{(s_1,t)}\boldsymbol{b}^{(s_2,t)}$, all entries are the same. Because

$$\lambda_{\min} \begin{bmatrix} \boldsymbol{K}_{ii} & \boldsymbol{K}_{ij} \\ \boldsymbol{K}_{ji} & \boldsymbol{K}_{jj} \end{bmatrix} = \lambda_{\min} \left(\begin{bmatrix} \boldsymbol{K}_{ii} & \boldsymbol{K}_{ij} \\ \boldsymbol{K}_{ji} & \boldsymbol{K}_{jj} \end{bmatrix} + \begin{bmatrix} \boldsymbol{b}_i(\boldsymbol{b}_i)^\top & \boldsymbol{b}_i(\boldsymbol{b}_j)^\top \\ \boldsymbol{b}_j(\boldsymbol{b}_i)^\top & \boldsymbol{b}_j(\boldsymbol{b}_j)^\top \end{bmatrix} \right)$$

Thus we know that $\boldsymbol{K}^{(t)}$ is positive definite.

Repeat the argument above, we can conclude that $\boldsymbol{K}^{(H)}$ is full rankness.

□

G Convergence vs. DAG Topology

In Figure 6, we sample 408 networks from our DAG space ($H = 3$). We empirically measure how many epochs each network requires to reach 80% training accuracy as its convergence rate, and plot these epochs with networks' topologies. We observe that wide and shallow networks show faster convergence. This observation is aligned with Eq. 12, which states that networks with larger widths (P_H, the number of end-to-end paths) and smaller depths (d_p, the number of "Linear + ReLU" layers on each path) will have large λ_0. However, as convergence only partially explains the final test accuracy (with missing factors like generalization and model complexity), Figure 6 does not conflict with Figure 5.

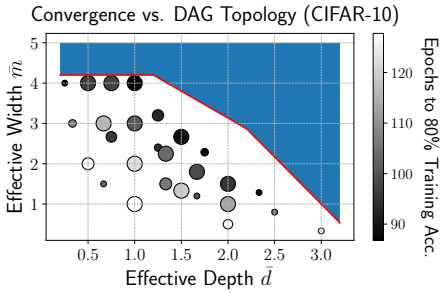

Figure 6: In a complete architecture space, being extremely large or small on either the effective depth ($\bar{d}$) or the effective width ($\bar{m}$) leads a connectivity pattern to have slow convergence (white circles). Each dot represents a subset of architectures of the same $\bar{d}$ and $\bar{m}$. Blue areas indicate invalid DAG regions: a DAG cannot be both deep and wide at the same time.