STAT 946 Fall 2025 Scribes: Disen Liao
Nov 3, 2025

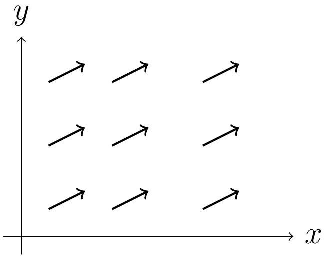
Lecture 14: Depth Scaling

1 Vector Field and ODE

An Ordinary Differential Equation (ODE) is an equation that relates an unknown function of a single independent variable (like time, t) to its own derivatives.The system is defined by a vector field b and an initial condition x_0:

$$\begin{cases} \dot{X}_t = b(X_t) \\ X_0 = x_0 \end{cases}$$

Vector field illustration:



2 Euler Discretization

To solve this numerically, we can approximate the continuous solution X_t with a discrete sequence Y_k, where $Y_k \approx X_{k\eta}$. This is done using the Euler method, which advances the solution in small time steps of size η.The update rule is defined as:

$$\begin{cases} Y_{k+1} = Y_k + \eta\, b(Y_k) \\ Y_0 = x_0 \end{cases} \qquad \text{Step size } \eta > 0$$

This rule is derived from the definition of the derivative. By rearranging the update rule, we can see it's a finite difference approximation of the original ODE's dynamics:

$$b(Y_k) = \frac{Y_{k+1} - Y_k}{\eta} \approx \frac{X_{(k+1)\eta} - X_{k\eta}}{\eta} \approx \dot{X}_{k\eta}$$

Here, the discrete difference $(Y_{k+1} - Y_k)/\eta$ approximates the true continuous derivative $\dot{X}$ at time $t = k\eta$.

2.1 Error Analysis

We try to analyze the local truncation error, the error we make in just one step. Here we make two Assumptions:

- $|b(x)| \leq C_0$ (Boundedness).
- $|b(x) - b(y)| \leq C_1|x - y|$ (Lipschitz Condition).

Then we will have:

$$X_\eta - X_0 = \int_0^\eta b(X_s)\, ds = \eta\, b(X_0) + \int_0^\eta [b(X_s) - b(X_0)]\, ds,$$

the second term is the *error term*.

$$\begin{aligned} \text{Error} &\leq C_1 \int_0^\eta |X_s - X_0|\, ds \\ &\leq C_1 \int_0^\eta \max_{s' \leq s} |X_{s'} - X_0|\, ds \\ &\leq C_0 C_1 \eta^2 = O(\eta^2) \end{aligned}$$

2.2 Stepwise Error Accumulation

This section analyzes how the small errors from each step add up over the entire simulation interval $[0, T]$. This is known as the global error. We start with the local error for a single step, which we previously found to be proportional to the square of the step size:

$$\text{Error}_{\text{step}} \approx C\eta^2$$

This is an $O(\eta^2)$ error.

Now, we must sum these local errors over all the steps required to cover the total time T. The number of steps, k, is determined by the total time T and the step size η:

$$k = T/\eta \quad (\text{or } k\eta = T)$$

A simplified view of the total error is to multiply the number of steps by the error per step:

$$\text{Total error} \approx k \times (\text{Local error}) \approx O(k\eta^2)$$

By substituting $k = T/\eta$, we can see how the global error depends on the step size η:

$$\text{Total error} \approx O((T/\eta) \cdot \eta^2) = O(T\eta)$$

Since T is a fixed constant, the total error is directly proportional to η. This is the key result: as the step size η goes to zero, the total accumulated error also goes to zero.

$$\eta \to 0 \quad \implies \quad \text{Total error} \to 0$$

This demonstrates that the Euler method converges. To achieve this, we must let the number of steps k go to infinity while the step size η goes to zero, such that their product remains the constant simulation time T.

2.3 Continuous-Time Interpolation

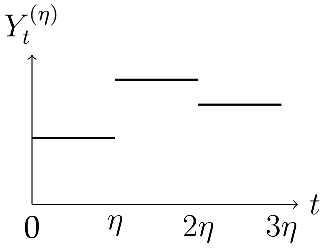
The analysis above shows convergence at the final time T. We now introduce a way to formalize the convergence over the entire time interval $[0, T]$.We define a continuous-time interpolation $Y_t^{(\eta)}$ from our discrete solution points $(Y_0, Y_\eta, Y_{2\eta}, \dots)$. This function "fills in the gaps" between our discrete points.

$$Y_t^{(\eta)} := Y_{\lfloor t/\eta \rfloor \cdot \eta} \quad \text{for } t \in [0, T].$$

Where $\lfloor \cdot \rfloor$ is the floor operation. This creates a piecewise-constant (step function) that holds the value Y_k for the entire time interval $[k\eta, (k+1)\eta)$.With this definition, we can state the formal convergence theorem. We will have uniform convergence of the approximation to the true solution:

$$\sup_{t \in [0,T]} |X_t - Y_t^{(\eta)}| \to 0 \quad \text{as } \eta \to 0$$

This notation means that the maximum possible error (the sup, or supremum) between the true solution X_t and our blocky approximation $Y_t^{(\eta)}$ over the entire interval $[0, T]$ shrinks to zero as the step size η shrinks to zero.



3 Residual Networks [HZRS16]

Consider the standard MLP:

$$h_{\ell+1} = \frac{1}{\sqrt{n}} W_\ell \, \phi(h_\ell).$$

When training very deep very deep MLPs, the gradient must be back propagated through all the layers. This involves a long chain of matrix multiplications, which will result in vanishing gradients/exploding gradients. ResNet addresses this problem by adding a "skip connection" that passes the input h_ℓ directly to the next layer, by passing the residual block:

$$h_{\ell+1} = \underbrace{h_\ell}_{\text{skip}} + \underbrace{\frac{1}{\sqrt{n}} W_\ell \, \phi(h_\ell)}_{\text{residual}} .$$

3.1 The Continuous Limit

We can view a very deep ResNet as a discrete system evolving over "layer time". This leads to a connection with ODE. Here we assume recurrent weight sharing: $W_1 = W_2 = \cdots = W_d = W$. And we can introduce a small step size ϵ, so we can rewrite the shared-weight ResNet update as:

$$h_{\ell+1} = h_\ell + \underbrace{\varepsilon}_{\text{step}} \underbrace{\frac{1}{\sqrt{n}} W \, \phi(h_\ell)}_{\text{vec. field}} .$$

Now, we keep n fixed, and $d \to \infty$, $\epsilon \to 0$. We keep the total "layer time" $\bar{\tau} = \epsilon d > 0$ fixed. Just as in the Euler convergence, we can define a continuous function $h_\tau^{(\epsilon)}$:

$$h_\tau^{(\epsilon)} = h_{\lfloor \tau/\epsilon \rfloor}$$

As we take the limit, this step-function approximation converges to a smooth, continuous path h_τ:

$$h_\tau^{(\epsilon)} \to h_\tau, \qquad \tau \in [0, \bar{\tau}].$$

This limiting continuous path h_τ is the solution to the following ODE:

$$\dot{h}_\tau = \frac{1}{\sqrt{n}} W \phi(h_\tau)$$

Here, τ represents the continuous "layer time" (where $\tau \sim \epsilon \ell$). This results is a Neural ODE [CRBD18].

$$h_\tau^{(\varepsilon)} = h_{\lfloor \tau/\varepsilon \rfloor} \;\longrightarrow\; h_\tau, \qquad \tau \in [0, \bar{\tau}].$$

$$\dot{h}_\tau = \frac{1}{\sqrt{n}} W\, \phi(h_\tau), \qquad \tau \sim \varepsilon \ell.$$

3.2 Variance (non-recurrent case)

We now consider a different scaling for the ResNet update, specifically for the non-recurrent case (where each layer has independent weights W_ℓ). This scaling is often used when analyzing the variance of activations as the network depth increases.This scaling convention is different from the one used to derive the Neural ODE.Recall the standard ResNet block from eq:resnet:$h_{\ell+1} = h_\ell + \frac{1}{\sqrt{n}} W_\ell\, \phi(h_\ell)$. Now, we introduce the ε parameter (related to depth, where $\varepsilon \to 0$ as $d \to \infty$) but with a different scaling:

$$h_{\ell+1} = h_\ell + \underbrace{\frac{1}{d}}_{\varepsilon} \frac{1}{\sqrt{n}} W_\ell\, \phi(h_\ell)$$

We can also have CLT scaling:

$$h_{\ell+1} = h_\ell + \frac{1}{\sqrt{d}} \frac{1}{\sqrt{n}} W_\ell\, \phi(h_\ell).$$

(Why this works?)

4 Random Walk

The discrete random walk considers a discrete-time process Y_k:

$$Y_{k+1} = Y_k + \sqrt{\eta}\, \sigma(Y_k)\, \xi_k,$$

where η is a small step size (can be set to $1/n$), ξ_k are i.i.d. random variables with $\mathbb{E}[\xi_k] = 0$ and $\mathbb{E}[\xi_k^2] = 1$. The $\sqrt{\eta}$ is the CLT scaling.

We define a continuous-time interpolation $Y_t^{(\eta)} = Y_{\lfloor t/\eta \rfloor}$. As we take the limit $\eta \to 0$, this discrete process converges to a continuous process X_t:

$$Y_t^{(\eta)} = Y_{\lfloor t/\eta \rfloor} \;\longrightarrow\; X_t$$

The limiting process X_t is the solution to a stochastic different equation (SDE):

$$dX_t = \sigma(X_t)\, dB_t$$

The solution to the SDE can be written as an integral equation:

$$X_t = X_0 + \int_0^t \sigma(X_s)\, dB_s$$

This is not a standard Riemann integral. It is an Itô stochastic integral, which is defined as a specific limit of a sum as the time partition Δt shrinks to 0:

$$\int_0^t \sigma(X_s)\, dB_s = \lim_{\Delta t \to 0} \sum_{i=1}^{n} \sigma(X_{t_i})\big(B_{t_{i+1}} - B_{t_i}\big)$$

This function $\sigma(X_{t_i})$ is evaluated at t_i, which is the left end point of the time interval $[t_i, t_{i+1}]$.

In general, a process with both a deterministic step and a random step:

$$Y_{k+1} = Y_k + \underbrace{\eta\, b(Y_k)}_{\text{Drift (ODE part)}} + \underbrace{\sqrt{\eta}\, \sigma(Y_k)\, \xi_k}_{\text{Diffusion (SDE part)}}$$

In the limit $\eta \to 0$, converges to a full SDE:

$$dX_t = b(X_t)\, dt + \sigma(X_t)\, dB_t$$

where the $O(\eta)$ term (Riemann integral) becomes the drift term dt, and the $O(\sqrt{\eta})$ term (CLT scaling) becomes the diffusion term dB_t.

4.1 Application to Scaled ResNet

We now apply this to scaled ResNet, recall the "CLT scaled" ResNet:

$$h_{\ell+1} = h_\ell + \frac{1}{\sqrt{d}} \frac{1}{\sqrt{n}} W_\ell \phi(h_\ell)$$

We can match this to the general SDE by setting:

- Step size η = 1/d
- Drift $b(h_\ell)$ = 0 (assiming W_ℓ has zero mean, $\mathbf{E}[W_\ell] = 0$)
- Stochastic term: $\sqrt{\eta}\sigma(h_\ell)\xi_k = \frac{1}{\sqrt{d}}(\frac{1}{\sqrt{n}} W_\ell \phi(h_\ell))$

To find the diffusion σ, we can write the ResNet as:

$$h_{\ell+1} = h_\ell + \frac{1}{\sqrt{d}} \tilde{\xi}_\ell$$

Where $\tilde{\xi}_\ell$ has i.i.d. entry and each entry can be computed as:

$$\tilde{\xi}_{\ell,i} = \langle W_{\ell,i}, \phi(h_\ell) \rangle = \sum_{j=1}^{n} W_{\ell,ij}\phi_j(h_\ell)$$

Assuming $W_{\ell,ij}$ are i.i.d. with $\mathbb{E}[W_{\ell,ij}] = 0$ and $\mathbb{E}[W_{\ell,ij}^2] = 1$ (standard initialization):

$$\mathbb{E}[\tilde{\xi}_{\ell,i}^2] = \mathbb{E}\left[\left(\sum_{j=1}^{n} W_{\ell,ij}\phi_j(h_\ell)\right)^2\right]$$

Because the weights $W_{\ell,ij}$ are independent, the cross-terms $\mathbb{E}[W_{\ell,ij}W_{\ell,ik}]$ are 0 for $j \neq k$.

$$\begin{aligned}
\mathbb{E}\left[\left(\sum_{j=1}^{n} W_{\ell,ij}\phi_j(h_\ell)\right)^2\right] &= \mathbb{E}\left[\sum_{j,j'=1}^{n} W_{\ell,ij}W_{\ell,ij'}\phi_j\phi_{j'}\right] \\
&= \sum_{j=1}^{n} \mathbb{E}[W_{\ell,ij}^2]\phi_j^2 \\
&= \sum_{j=1}^{n} (1) \cdot \phi_j^2 \\
&= |\phi(h_\ell)|^2
\end{aligned}$$

This defines the variance of each block, and this matches the SDE form $\sqrt{\eta}\sigma(h_\ell)$ where $\eta = 1/d$ and $\sigma(h_\ell) = \frac{1}{\sqrt{n}}|\phi(h_\ell)|$.Thus, as $d \to \infty$ (so $\eta \to 0$), the ResNet dynamics converge to [HY23]:

$$dh_\tau = \frac{1}{\sqrt{n}}\,|\phi(h_\tau)|\,dB_\tau$$

4.2 Joint Limits

1. Proportional Scaling (Standard MLP, no skip connection)

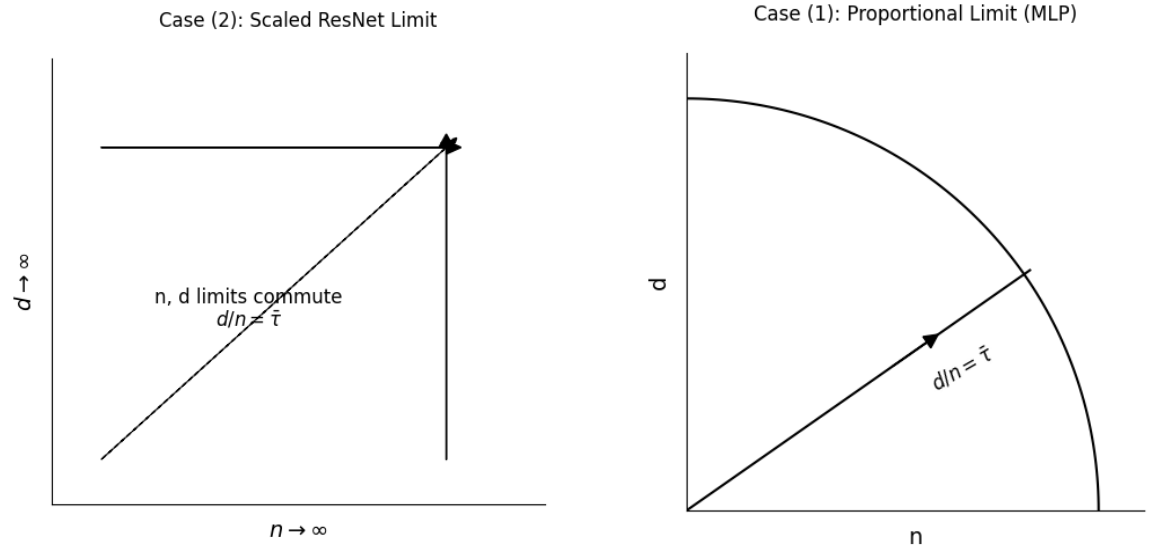
$$h_{\ell+1} = \frac{1}{\sqrt{n}}\,W_\ell\,\phi(h_\ell)$$

   Here, the limits must be taken jointly, with their ratio held constant: $d/n \to \bar{\tau}$. This is typical in Random Matrix Theory analysis.

2. Scaled ResNet

$$h_{\ell+1} = h_\ell + \frac{1}{\sqrt{d}} \cdot [\ldots] \quad (\text{“some block”})$$

For this model, the n, d limits commute. You can let width $n \to \infty$ first (mean-field limit) and then let depth $d \to \infty$ (SDE limit), or vice versa, or take them jointly, and you will arrive at the same limiting process. This is a very robust property.


Case (2): Scaled ResNet Limit
d → ∞
n, d limits commute
d/n = τ̄
n → ∞
Case (1): Proportional Limit (MLP)
d
d/n = τ̄
n

5 Linear Network at Init

We consider the case the input is a single data point $(m = 1)$. The network are given by:

$$h_{\ell+1} = \frac{1}{\sqrt{n}} W_\ell h_\ell$$

where $h_\ell \in \mathbb{R}^n$ is the activation at layer ℓ, and $W_\ell \in \mathbb{R}^{n \times n}$ is a random weight matrix. We assume the entries of W_ℓ are i.i.d. $N(0, 1)$. So we have $h_\ell \mid h_{\ell-1} \sim \mathcal{N}\big(0, \frac{1}{n}|h_\ell|^2 \otimes I_n\big)$. Then, we trach the evolution of the normalized squared norm, $\frac{1}{n}|h_\ell|^2$.

$$\frac{1}{n}|h_{\ell+1}|^2 = \frac{1}{n}\left|\frac{1}{\sqrt{n}} W_\ell h_\ell\right|^2 = \frac{1}{n^2}|W_\ell h_\ell|^2$$

We can separate the magnitude and direction of h_ℓ (where $v_\ell = h_\ell/|h_\ell|$ is a unit vector):

$$\frac{1}{n}|h_{\ell+1}|^2 = \frac{1}{n^2}|W_\ell(v_\ell|h_\ell|)|^2 = \left(\frac{1}{n}|h_\ell|^2\right) \cdot \left(\frac{1}{n}|W_\ell v_\ell|^2\right)$$

Let's define a new random variable $\frac{1}{n}|\xi_\ell|^2 = \frac{1}{n}|W_\ell v_\ell|^2$. Since v_ℓ is a unit vector and W_ℓ is a standard Gaussian matrix, the vector $\xi_\ell = W_\ell v_\ell$ is also standard Gaussian, $\xi_\ell \sim N(0, I_n)$, and is independent of h_ℓ.

This gives us the recursive relation:

$$\frac{1}{n}|h_{\ell+1}|^2 = \left(\frac{1}{n}|h_\ell|^2\right) \cdot \left(\frac{1}{n}|\xi_\ell|^2\right)$$

Unrolling this recursion back to the input $h_0 = x$ (where n_0 is the input dimension), we get:

$$\frac{1}{n}|h_{\ell+1}|^2 = \left(\frac{1}{n_0}|x|^2\right) \cdot \prod_{k=1}^{\ell} \left(\frac{1}{n}|\xi_k|^2\right)$$

To make the problem manageable, taking log on both sides:

$$\log\left(\frac{1}{n}|h_d|^2\right) = \log\left(\frac{1}{n_0}|x|^2\right) + \sum_{\ell=0}^{d-1} \log\left(\frac{1}{n}|\xi_\ell|^2\right)$$

We now analyze the properties of the random term $\frac{1}{n}|\xi_\ell|^2 = \frac{1}{n}\sum_{i=1}^n \xi_{\ell,i}^2$, where $\xi_{\ell,i} \sim N(0,1)$ are i.i.d.

1. Expectation: By the Law of Large Numbers, this term converges to its mean.

$$\mathbb{E}\left[\frac{1}{n}|\xi_\ell|^2\right] = \frac{1}{n}\sum_{i=1}^n \mathbb{E}[\xi_{\ell,i}^2] = \frac{1}{n}\sum_{i=1}^n (1) = 1$$

2. Variance:We first compute the second moment. Recall for a standard Gaussian, $\mathbb{E}[\xi^2] = 1$ and $\mathbb{E}[\xi^4] = 3$.

$$\mathbb{E}\left[\left(\frac{1}{n}|\xi_\ell|^2\right)^2\right] = \frac{1}{n^2}\mathbb{E}\left[\left(\sum_{i=1}^n \xi_{\ell,i}^2\right)^2\right] = \frac{1}{n^2}\mathbb{E}\left[\sum_{i,j=1}^n \xi_{\ell,i}^2\xi_{\ell,j}^2\right]$$

We split the sum into diagonal ($i = j$) and off-diagonal ($i \neq j$) terms:

$$= \frac{1}{n^2}\left(\sum_{i=1}^n \mathbb{E}[\xi_{\ell,i}^4] + \sum_{i \neq j} \mathbb{E}[\xi_{\ell,i}^2]\mathbb{E}[\xi_{\ell,j}^2]\right)$$

$$= \frac{1}{n^2}\big(n \cdot 3 + n(n-1) \cdot 1 \cdot 1\big) = \frac{3n + n^2 - n}{n^2} = \frac{n^2 + 2n}{n^2} = 1 + \frac{2}{n}$$

The variance is therefore:

$$\text{Var}\left(\frac{1}{n}|\xi_\ell|^2\right) = \mathbb{E}[(\cdot)^2] - (\mathbb{E}[\cdot])^2 = \left(1 + \frac{2}{n}\right) - (1)^2 = \frac{2}{n}$$

Since Var $\to 0$ as $n \to \infty$, the term concentrates around its mean of 1.

To analyze the sum of logs, we define a normalized random variable ζ_ℓ with mean 0 and variance 1:

$$\frac{1}{n}|\xi_\ell|^2 = 1 + \sqrt{\frac{2}{n}}\zeta_\ell$$

Now, we use the Taylor expansion $\log(1+x) \approx x - \frac{x^2}{2} + O(x^3)$:

$$\begin{aligned}\log\left(\frac{1}{n}|\xi_\ell|^2\right) &= \log\left(1 + \sqrt{\frac{2}{n}}\zeta_\ell\right) \\ &\approx \left(\sqrt{\frac{2}{n}}\zeta_\ell\right) - \frac{1}{2}\left(\sqrt{\frac{2}{n}}\zeta_\ell\right)^2 + O(n^{-3/2}) \\ &\approx \sqrt{\frac{2}{n}}\zeta_\ell - \frac{1}{n}\zeta_\ell^2 + O(n^{-3/2})\end{aligned}$$

Substituting this back into our sum (and writing $\log(\frac{1}{n_0}|x_0|^2)$ for the initial term):

$$\log\left(\frac{1}{n}|h_d|^2\right) \approx \log\left(\frac{1}{n_0}|x_0|^2\right) + \sum_{\ell=0}^{d-1}\left(\sqrt{\frac{2}{n}}\zeta_\ell - \frac{1}{n}\zeta_\ell^2\right) + O(d \cdot n^{-3/2})$$

We analyze the behavior in the joint limit where $d \to \infty$ and $n \to \infty$ such that their ratio is constant: $d/n \to \bar{\tau}$.

$$\log\left(\frac{1}{n}|h_d|^2\right) \approx \log\left(\frac{1}{n_0}|x_0|^2\right) + \underbrace{\sqrt{\frac{2}{n}}\sum_{\ell=0}^{d-1}\zeta_\ell}_{\text{(1) CLT Term}} - \underbrace{\frac{1}{n}\sum_{\ell=0}^{d-1}\zeta_\ell^2}_{\text{(2) LLN Term}} + \underbrace{O(d \cdot n^{-3/2})}_{\text{(3) Remainder}}$$

1. CLT Term: $\sqrt{\frac{2}{n}}\sum\zeta_\ell = \sqrt{2}\sqrt{\frac{d}{n}}\left(\frac{1}{\sqrt{d}}\sum\zeta_\ell\right)$. By the Central Limit Theorem (CLT), $\frac{1}{\sqrt{d}}\sum\zeta_\ell \to N(0,1)$. Thus, the term converges to $\sqrt{2}\sqrt{\bar{\tau}} \cdot N(0,1) \sim N(0, 2\bar{\tau})$.

2. LLN Term: This term converges to its expected value. By the Law of Large Numbers (LLN), $\frac{1}{d}\sum\zeta_\ell^2 \to \mathbb{E}[\zeta_\ell^2] = 1$. The term is $-\frac{d}{n}(\frac{1}{d}\sum\zeta_\ell^2)$, hence it converges to $-\bar{\tau}$.

3. Remainder: $O(d \cdot n^{-3/2}) = O(\frac{d}{n} \cdot n^{-1/2}) = O(\bar{\tau} n^{-1/2}) \to 0$.

Theorem 1 ([HN19])**.** *Combining these limits, the log-norm of the output activation converges to a Gaussian distribution:*

$$\log\left(\frac{1}{n}|h_d|^2\right) \xrightarrow{d/n \to \bar{\tau}} N\left(\log\left(\frac{1}{n_0}|x_0|^2\right) - \bar{\tau}, \quad 2\bar{\tau}\right)$$

References

[CRBD18] Ricky TQ Chen, Yulia Rubanova, Jesse Bettencourt, and David K Duvenaud. Neural ordinary differential equations. *Advances in neural information processing systems*, 31, 2018.

[HN19] Boris Hanin and Mihai Nica. Products of many large random matrices and gradients in deep neural networks. *Communications in Mathematical Physics*, 376(1):287–322, December 2019.

[HY23] Soufiane Hayou and Greg Yang. Width and depth limits commute in residual networks, 2023.

[HZRS16] 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.