Lecture 3: The Principle of Sparsity

[§4, §8, §11.2, and §∞.3 of
"The Principles of Deep Learning Theory (PDLT),"
arXiv:2106.10165]

Problems 1, 2, & 3

Fully-trained network output, Taylor-expanded around initialization:

$$z^\star = z + (\theta^\star - \theta)\frac{dz}{d\theta} + \frac{1}{2}(\theta^\star - \theta)^2\frac{d^2 z}{d\theta^2} + \ldots$$

Problems 1, 2, & 3

Fully-trained network output, Taylor-expanded around initialization:

$$z^\star = z + (\theta^\star - \theta)\frac{dz}{d\theta} + \frac{1}{2}(\theta^\star - \theta)^2\frac{d^2 z}{d\theta^2} + \dots$$

- Problem 1: too many terms in general

Problems 1, 2, & 3

Fully-trained network output, Taylor-expanded around initialization:

$$z^\star = z + (\theta^\star - \theta)\frac{dz}{d\theta} + \frac{1}{2}(\theta^\star - \theta)^2 \frac{d^2 z}{d\theta^2} + \ldots$$

- Problem 1: too many terms in general
- Problem 2: complicated mapping

$$p(\theta) \to p\left(\theta, z, \frac{dz}{d\theta}, \frac{d^2 z}{d\theta^2}, \ldots\right)$$

initial distributions over model parameters

statistics at *initialization*

Problems 1, 2, & 3

Fully-trained network output, Taylor-expanded around initialization:

$$z^\star = z + (\theta^\star - \theta)\frac{dz}{d\theta} + \frac{1}{2}(\theta^\star - \theta)^2\frac{d^2 z}{d\theta^2} + \ldots$$

- Problem 1: too many terms in general
- Problem 2: complicated mapping
- Problem 3: complicated dynamics $\theta^\star = [\theta^\star]\left(\theta, z, \frac{dz}{d\theta}, \frac{d^2 z}{d\theta^2}, \ldots; \text{algorithm}; \text{data}\right)$

$$p(\theta) \to p\left(\theta, z, \frac{dz}{d\theta}, \frac{d^2 z}{d\theta^2}, \ldots\right)$$

initial distributions over model parameters

statistics at *initialization*

Dan has covered Dynamics

Fully-trained network output, Taylor-expanded around initialization:

$$z^\star = z + (\theta^\star - \theta)\frac{dz}{d\theta} + \frac{1}{2}(\theta^\star - \theta)^2\frac{d^2 z}{d\theta^2} + \dots$$

- ~~Problem 1: too many terms in general~~
- Problem 2: complicated mapping
- Problem 3: complicated dynamics $z^\star = [z^\star]\left(\theta, z, \frac{dz}{d\theta}, \frac{d^2 z}{d\theta^2}, \dots; \text{algorithm}; \text{data}\right)$

$$p(\theta) \to p\left(\theta, z, \frac{dz}{d\theta}, \frac{d^2 z}{d\theta^2}, \dots\right)$$

initial distributions
over
model parameters

statistics at *initialization*

Dan has covered Dynamics

Fully-trained network output, Taylor-expanded around initialization:

$$z^\star = z + (\theta^\star - \theta)\frac{dz}{d\theta} + \frac{1}{2}(\theta^\star - \theta)^2\frac{d^2 z}{d\theta^2} + \dots$$

- ~~Problem 1: too many terms in general~~
- Problem 2: complicated mapping
- Problem 3: complicated dynamics $z^\star = [z^\star]\left(\theta, z, \frac{dz}{d\theta}, \frac{d^2 z}{d\theta^2}, \dots; \text{algorithm}; \text{data}\right)$

$$p(\theta) \to p\left(\theta, z, \frac{dz}{d\theta}, \frac{d^2 z}{d\theta^2}, \dots\right) \to p(z^\star)$$

initial distributions over model parameters

statistics at *initialization*

statistics *after training*

Sho will cover Statistics

Fully-trained network output, Taylor-expanded around initialization:

$$z^\star = z + (\theta^\star - \theta)\frac{dz}{d\theta} + \frac{1}{2}(\theta^\star - \theta)^2\frac{d^2 z}{d\theta^2} + \dots$$

- Problem 1: too many terms in general
- Problem 2: complicated mapping
- Problem 3: complicated dynamics

$$p(\theta) \rightarrow p\left(\theta, z, \frac{dz}{d\theta}, \frac{d^2 z}{d\theta^2}, \dots\right) \rightarrow p(z^\star)$$

initial distributions over model parameters

statistics at *initialization*

statistics *after training*

Sho will cover Statistics

$$p(\theta) \rightarrow p\left(\theta, z, \frac{dz}{d\theta}, \frac{d^2 z}{d\theta^2}, \ldots\right)$$

initial distributions
over
model parameters

statistics at *initialization*

for WIDE & DEEP neural networks

Sho will cover Statistics

Lecture 3: **The Principle of Sparsity**, deriving recursions

Lecture 4: **The Principle of Criticality**, solving recursions

$$p(\theta) \rightarrow p\left(\theta, z, \frac{dz}{d\theta}, \frac{d^2 z}{d\theta^2}, \ldots\right)$$

initial distributions
over
model parameters

statistics at *initialization*

for WIDE & DEEP neural networks

Outline

1. Neural Networks 101
2. One-Layer Neural Networks
3. Two-Layer Neural Networks
4. Deep Neural Networks

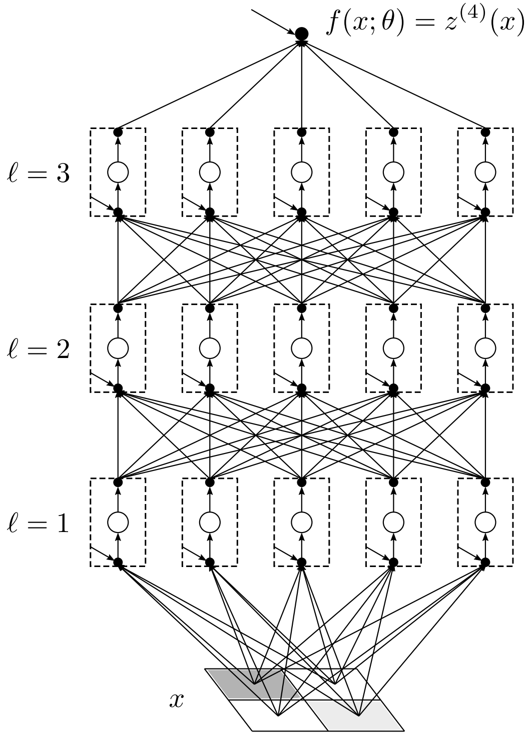
1. Neural Networks 101

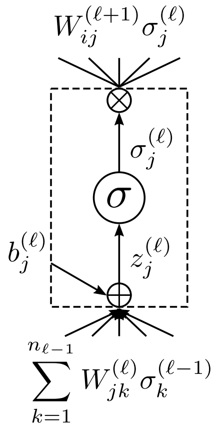
Neural Networks

$$\widehat{z}_i^{(1)}(x) \equiv b_i^{(1)} + \sum_{j=1}^{n_0} W_{ij}^{(1)} x_j \quad \text{for} \quad i = 1, \ldots, n_1 \,,$$

$$\widehat{z}_i^{(\ell+1)}(x) \equiv b_i^{(\ell+1)} + \sum_{j=1}^{n_\ell} W_{ij}^{(\ell+1)} \sigma\left(\widehat{z}_j^{(\ell)}(x)\right) \quad \text{for} \quad i = 1, \ldots, n_{\ell+1} \,;\; \ell = 1, \ldots, L-1$$

$$\widehat{z}_{i;\delta} = \widehat{z}_i^{(L)}(x_\delta)$$





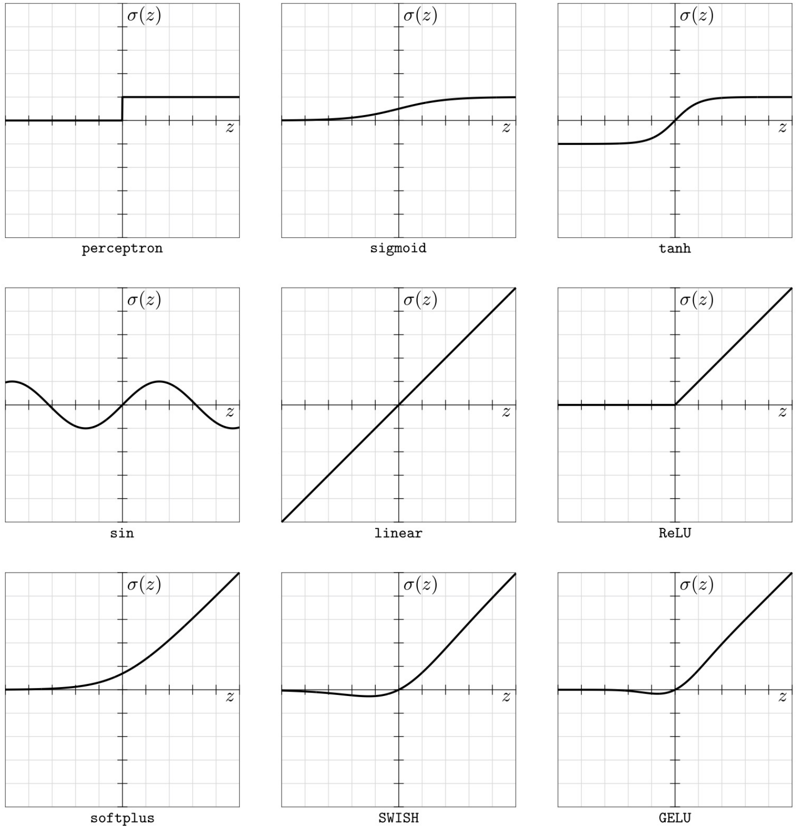
Neural Networks

$$\widehat{z}_i^{(1)}(x) \equiv b_i^{(1)} + \sum_{j=1}^{n_0} W_{ij}^{(1)} x_j \quad \text{for} \quad i = 1, \ldots, n_1 \,,$$

$$\widehat{z}_i^{(\ell+1)}(x) \equiv b_i^{(\ell+1)} + \sum_{j=1}^{n_\ell} W_{ij}^{(\ell+1)} \sigma\left(\widehat{z}_j^{(\ell)}(x)\right) \quad \text{for} \quad i = 1, \ldots, n_{\ell+1}\,;\ \ell = 1, \ldots, L-1$$

$$\widehat{z}_{i;\delta} = \widehat{z}_i^{(L)}(x_\delta)$$

activation function $\sigma(z)$



Neural Networks

preactivations

$$\widehat{z}_i^{(1)}(x) \equiv b_i^{(1)} + \sum_{j=1}^{n_0} W_{ij}^{(1)} x_j \quad \text{for} \quad i = 1, \dots, n_1 \,,$$

$$\widehat{z}_i^{(\ell+1)}(x) \equiv b_i^{(\ell+1)} + \sum_{j=1}^{n_\ell} W_{ij}^{(\ell+1)} \sigma\left(\widehat{z}_j^{(\ell)}(x)\right) \quad \text{for} \quad i = 1, \dots, n_{\ell+1} \,; \; \ell = 1, \dots, L-1$$

network output

$$\widehat{z}_{i;\delta} = \widehat{z}_i^{(L)}(x_\delta)$$

Neural Networks

$$\widehat{z}_i^{(1)}(x) \equiv b_i^{(1)} + \sum_{j=1}^{n_0} W_{ij}^{(1)} x_j \quad \text{for} \quad i = 1, \ldots, n_1\,,$$

$$\widehat{z}_i^{(\ell+1)}(x) \equiv b_i^{(\ell+1)} + \sum_{j=1}^{n_\ell} W_{ij}^{(\ell+1)} \sigma\left(\widehat{z}_j^{(\ell)}(x)\right) \quad \text{for} \quad i = 1, \ldots, n_{\ell+1}\,;\ \ell = 1, \ldots, L-1$$

$$\widehat{z}_{i;\delta} = \widehat{z}_i^{(L)}(x_\delta)$$

hat @ initialization

Neural Networks

$$\widehat{z}_i^{(1)}(x) \equiv b_i^{(1)} + \sum_{j=1}^{n_0} W_{ij}^{(1)} x_j \quad \text{for} \quad i = 1, \ldots, n_1 \, ,$$

$$\widehat{z}_i^{(\ell+1)}(x) \equiv b_i^{(\ell+1)} + \sum_{j=1}^{n_\ell} W_{ij}^{(\ell+1)} \sigma\left(\widehat{z}_j^{(\ell)}(x)\right) \quad \text{for} \quad i = 1, \ldots, n_{\ell+1} \, ; \; \ell = 1, \ldots, L-1$$

$$\widehat{z}_{i;\delta} = \widehat{z}_i^{(L)}(x_\delta)$$

Biases and weights (model parameters) are independently (& symmetrically) distributed with variances

$$\mathbb{E}\left[b_{i_1}^{(\ell)} b_{i_2}^{(\ell)}\right] = \delta_{i_1 i_2} C_b^{(\ell)} \, , \quad \mathbb{E}\left[W_{i_1 j_1}^{(\ell)} W_{i_2 j_2}^{(\ell)}\right] = \delta_{i_1 i_2} \delta_{j_1 j_2} \frac{C_W^{(\ell)}}{n_{\ell-1}}$$

Neural Networks

$$\widehat{z}_i^{(1)}(x) \equiv b_i^{(1)} + \sum_{j=1}^{n_0} W_{ij}^{(1)} x_j \quad \text{for} \quad i = 1, \ldots, n_1\,,$$

$$\widehat{z}_i^{(\ell+1)}(x) \equiv b_i^{(\ell+1)} + \sum_{j=1}^{n_\ell} W_{ij}^{(\ell+1)} \sigma\!\left(\widehat{z}_j^{(\ell)}(x)\right) \quad \text{for} \quad i = 1, \ldots, n_{\ell+1}\,;\ \ell = 1, \ldots, L-1$$

$$\widehat{z}_{i;\delta} = \widehat{z}_i^{(L)}(x_\delta)$$

Biases and weights (model parameters) are independently (& symmetrically) distributed with variances

$$\mathbb{E}\left[b_{i_1}^{(\ell)} b_{i_2}^{(\ell)}\right] = \delta_{i_1 i_2} C_b^{(\ell)}\,, \qquad \mathbb{E}\left[W_{i_1 j_1}^{(\ell)} W_{i_2 j_2}^{(\ell)}\right] = \delta_{i_1 i_2} \delta_{j_1 j_2} \frac{C_W^{(\ell)}}{n_{\ell-1}}$$

initialization hyperparameters

Neural Networks

$$\widehat{z}_i^{(1)}(x) \equiv b_i^{(1)} + \sum_{j=1}^{n_0} W_{ij}^{(1)} x_j \quad \text{for} \quad i = 1, \ldots, n_1 \, ,$$

$$\widehat{z}_i^{(\ell+1)}(x) \equiv b_i^{(\ell+1)} + \sum_{j=1}^{n_\ell} W_{ij}^{(\ell+1)} \sigma\left(\widehat{z}_j^{(\ell)}(x)\right) \quad \text{for} \quad i = 1, \ldots, n_{\ell+1} \, ; \; \ell = 1, \ldots, L-1$$

$$\widehat{z}_{i;\delta} = \widehat{z}_i^{(L)}(x_\delta)$$

Biases and weights (model parameters) are independently (& symmetrically) distributed with variances

$$\mathbb{E}\left[b_{i_1}^{(\ell)} b_{i_2}^{(\ell)}\right] = \delta_{i_1 i_2} C_b^{(\ell)} \, , \quad \mathbb{E}\left[W_{i_1 j_1}^{(\ell)} W_{i_2 j_2}^{(\ell)}\right] = \delta_{i_1 i_2} \delta_{j_1 j_2} \frac{C_W^{(\ell)}}{n_{\ell-1}}$$

good wide limit

One Aside on Gradient Descent

$$\theta_\mu(t+1) = \theta_\mu(t) - \eta \sum_\nu \lambda_{\mu\nu} \left(\sum \frac{\partial \mathcal{L}}{\partial z_{j;\tilde{\alpha}}} \frac{dz_{j;\tilde{\alpha}}}{d\theta_\nu} \right)$$

[Cf. Andrea's "S" matrix]

One Aside on Gradient Descent

$$\theta_\mu(t+1) = \theta_\mu(t) - \eta \sum_\nu \lambda_{\mu\nu} \left(\sum \frac{\partial \mathcal{L}}{\partial z_{j;\tilde{\alpha}}} \frac{dz_{j;\tilde{\alpha}}}{d\theta_\nu} \right)$$

Taylor expansion:

$$\begin{aligned} z_{i;\delta}(t+1) = & z_{i;\delta}(t) \\ & - \eta \sum_{j,\tilde{\alpha}} \left(\sum_{\mu,\nu} \lambda_{\mu\nu} \frac{dz_{i;\delta}}{d\theta_\mu} \frac{dz_{j;\tilde{\alpha}}}{d\theta_\nu} \right) \frac{\partial \mathcal{L}}{\partial z_{j;\tilde{\alpha}}} \\ & + \ldots \end{aligned}$$

One Aside on Gradient Descent

$$\theta_\mu(t+1) = \theta_\mu(t) - \eta \sum_\nu \lambda_{\mu\nu} \left(\sum \frac{\partial \mathcal{L}}{\partial z_{j;\tilde{\alpha}}} \frac{dz_{j;\tilde{\alpha}}}{d\theta_\nu} \right)$$

Taylor expansion:

Neural Tangent Kernel (NTK) $H(t)$ [Cf. Dan's $k_{\delta\tilde{\alpha}}$]

$$\begin{aligned} z_{i;\delta}(t+1) = & z_{i;\delta}(t) \\ & - \eta \sum_{j,\tilde{\alpha}} \left(\sum_{\mu,\nu} \lambda_{\mu\nu} \frac{dz_{i;\delta}}{d\theta_\mu} \frac{dz_{j;\tilde{\alpha}}}{d\theta_\nu} \right) \frac{\partial \mathcal{L}}{\partial z_{j;\tilde{\alpha}}} \\ & + \dots \end{aligned}$$

One Aside on Gradient Descent

$$\theta_\mu(t+1) = \theta_\mu(t) - \eta \sum_\nu \lambda_{\mu\nu} \left(\sum \frac{\partial \mathcal{L}}{\partial z_{j;\tilde{\alpha}}} \frac{dz_{j;\tilde{\alpha}}}{d\theta_\nu} \right)$$

Taylor expansion:

$$z_{i;\delta}(t+1) = z_{i;\delta}(t)$$

Neural Tangent Kernel (NTK) $H(t)$ [Cf. Dan's $k^{\mathrm{E}}_{ij;\delta\tilde{\alpha}}(\theta)$]

$$- \eta \sum_{j,\tilde{\alpha}} \left(\sum_{\mu,\nu} \lambda_{\mu\nu} \frac{dz_{i;\delta}}{d\theta_\mu} \frac{dz_{j;\tilde{\alpha}}}{d\theta_\nu} \right) \frac{\partial \mathcal{L}}{\partial z_{j;\tilde{\alpha}}}$$

differential of NTK (dNTK) $\mathrm{d}H(t)$ [Cf. Dan's $\mu_{\delta\tilde{\alpha}_1\tilde{\alpha}_2}$]

$$+ \frac{\eta^2}{2} \sum_{j_1, j_2, \tilde{\alpha}_1 \tilde{\alpha}_2} \left(\sum_{\mu_1,\nu_1,\mu_2,\nu_2} \lambda_{\mu_1\nu_1} \lambda_{\mu_2\nu_2} \frac{d^2 z_{i;\delta}}{d\theta_{\mu_1} d\theta_{\mu_2}} \frac{dz_{j_1;\tilde{\alpha}_1}}{d\theta_{\nu_1}} \frac{dz_{j_2;\tilde{\alpha}_2}}{d\theta_{\nu_2}} \right) \frac{\partial \mathcal{L}}{\partial z_{j_1;\tilde{\alpha}_1}} \frac{\partial \mathcal{L}}{\partial z_{j_2;\tilde{\alpha}_2}}$$

$$+ \ldots$$

One Aside on Gradient Descent

$$\theta_\mu(t+1) = \theta_\mu(t) - \eta \sum_\nu \lambda_{\mu\nu} \left(\sum \frac{\partial \mathcal{L}}{\partial z_{j;\tilde{\alpha}}} \frac{dz_{j;\tilde{\alpha}}}{d\theta_\nu} \right)$$

Taylor expansion:

$$\begin{aligned} z_{i;\delta}(t+1) =& z_{i;\delta}(t) \\ &- \eta \sum_{j,\tilde{\alpha}} \left(\sum_{\mu,\nu} \lambda_{\mu\nu} \frac{dz_{i;\delta}}{d\theta_\mu} \frac{dz_{j;\tilde{\alpha}}}{d\theta_\nu} \right) \frac{\partial \mathcal{L}}{\partial z_{j;\tilde{\alpha}}} \\ &+ \frac{\eta^2}{2} \sum_{j_1,j_2,\tilde{\alpha}_1\tilde{\alpha}_2} \left(\sum_{\mu_1,\nu_1,\mu_2,\nu_2} \lambda_{\mu_1\nu_1}\lambda_{\mu_2\nu_2} \frac{d^2 z_{i;\delta}}{d\theta_{\mu_1} d\theta_{\mu_2}} \frac{dz_{j_1;\tilde{\alpha}_1}}{d\theta_{\nu_1}} \frac{dz_{j_2;\tilde{\alpha}_2}}{d\theta_{\nu_2}} \right) \frac{\partial \mathcal{L}}{\partial z_{j_1;\tilde{\alpha}_1}} \frac{\partial \mathcal{L}}{\partial z_{j_2;\tilde{\alpha}_2}} \\ &- \frac{\eta^3}{6} \sum \left(\sum \lambda_{\mu_1\nu_1}\lambda_{\mu_2\nu_2}\lambda_{\mu_3\nu_3} \frac{d^3 z_{i;\delta}}{d\theta_{\mu_1} d\theta_{\mu_2} d\theta_{\mu_3}} \frac{dz_{j_1;\tilde{\alpha}_1}}{d\theta_{\nu_1}} \frac{dz_{j_2;\tilde{\alpha}_2}}{d\theta_{\nu_2}} \frac{dz_{j_3;\tilde{\alpha}_3}}{d\theta_{\nu_3}} \right) \frac{\partial \mathcal{L}}{\partial z_{j_1;\tilde{\alpha}_1}} \frac{\partial \mathcal{L}}{\partial z_{j_2;\tilde{\alpha}_2}} \frac{\partial \mathcal{L}}{\partial z_{j_3;\tilde{\alpha}_3}} \\ &+ \ldots \end{aligned}$$

Neural Tangent Kernel (NTK) $H(t)$

differential of NTK (dNTK) $\mathrm{d}H(t)$

ddNTK $\mathrm{dd}H(t)$

One Aside on Gradient Descent

$$\theta_\mu(t+1) = \theta_\mu(t) - \eta \sum_\nu \lambda_{\mu\nu} \left(\sum \frac{\partial \mathcal{L}}{\partial z_{j;\tilde{\alpha}}} \frac{dz_{j;\tilde{\alpha}}}{d\theta_\nu} \right)$$

Taylor expansion:

$z_{i;\delta}(t+1) = z_{i;\delta}(t)$ Neural Tangent Kernel (NTK) $H(t)$

$$- \eta \sum_{j,\tilde{\alpha}} \left(\sum_{\mu,\nu} \lambda_{\mu\nu} \frac{dz_{i;\delta}}{d\theta_\mu} \frac{dz_{j;\tilde{\alpha}}}{d\theta_\nu} \right) \frac{\partial \mathcal{L}}{\partial z_{j;\tilde{\alpha}}}$$

differential of NTK (dNTK) $\mathrm{d}H(t)$

$$O(1/n) \left(\begin{aligned} &+ \frac{\eta^2}{2} \sum_{j_1, j_2, \tilde{\alpha}_1 \tilde{\alpha}_2} \left(\sum_{\mu_1, \nu_1, \mu_2, \nu_2} \lambda_{\mu_1\nu_1} \lambda_{\mu_2\nu_2} \frac{d^2 z_{i;\delta}}{d\theta_{\mu_1} d\theta_{\mu_2}} \frac{dz_{j_1;\tilde{\alpha}_1}}{d\theta_{\nu_1}} \frac{dz_{j_2;\tilde{\alpha}_2}}{d\theta_{\nu_2}} \right) \frac{\partial \mathcal{L}}{\partial z_{j_1;\tilde{\alpha}_1}} \frac{\partial \mathcal{L}}{\partial z_{j_2;\tilde{\alpha}_2}} \\ &- \frac{\eta^3}{6} \sum \left(\sum \lambda_{\mu_1\nu_1} \lambda_{\mu_2\nu_2} \lambda_{\mu_3\nu_3} \frac{d^3 z_{i;\delta}}{d\theta_{\mu_1} d\theta_{\mu_2} d\theta_{\mu_3}} \frac{dz_{j_1;\tilde{\alpha}_1}}{d\theta_{\nu_1}} \frac{dz_{j_2;\tilde{\alpha}_2}}{d\theta_{\nu_2}} \frac{dz_{j_3;\tilde{\alpha}_3}}{d\theta_{\nu_3}} \right) \frac{\partial \mathcal{L}}{\partial z_{j_1;\tilde{\alpha}_1}} \frac{\partial \mathcal{L}}{\partial z_{j_2;\tilde{\alpha}_2}} \frac{\partial \mathcal{L}}{\partial z_{j_3;\tilde{\alpha}_3}} \end{aligned} \right.$$

ddNTK $\mathrm{dd}H(t)$

$$O(1/n^2) \left(+ \ldots \right.$$

Neural Tangent Kernel (NTK)

$$\widehat{H}^{(\ell)}_{i_1 i_2;\delta_1\delta_2} \equiv \sum_{\mu,\nu} \lambda_{\mu\nu} \frac{d\widehat{z}^{(\ell)}_{i_1;\delta_1}}{d\theta_\mu} \frac{d\widehat{z}^{(\ell)}_{i_2;\delta_2}}{d\theta_\nu}$$

$$\{\theta_\mu\} = \{b^{(\ell)}_i, W^{(\ell)}_{ij}\}$$

$$\widehat{H}_{i_1 i_2;\delta_1\delta_2} = \widehat{H}^{(L)}_{i_1 i_2;\delta_1\delta_2}$$

Neural Tangent Kernel (NTK)

$$\widehat{H}^{(\ell)}_{i_1 i_2;\delta_1\delta_2} \equiv \sum_{\mu,\nu} \lambda_{\mu\nu} \frac{d\widehat{z}^{(\ell)}_{i_1;\delta_1}}{d\theta_\mu} \frac{d\widehat{z}^{(\ell)}_{i_2;\delta_2}}{d\theta_\nu}$$

$$\{\theta_\mu\} = \{b^{(\ell)}_i, W^{(\ell)}_{ij}\}$$

$$\widehat{H}_{i_1 i_2;\delta_1\delta_2} = \widehat{H}^{(L)}_{i_1 i_2;\delta_1\delta_2}$$

Diagonal, group-by-group, learning rate:

$$\lambda_{b^{(\ell)}_{i_1} b^{(\ell)}_{i_2}} = \delta_{i_1 i_2} \lambda^{(\ell)}_b\,, \qquad \lambda_{W^{(\ell)}_{i_1 j_1} W^{(\ell)}_{i_2 j_2}} = \delta_{i_1 i_2} \delta_{j_1 j_2} \frac{\lambda^{(\ell)}_W}{n_{\ell-1}}$$

[Cf. Andrea's "S" matrix]

Neural Tangent Kernel (NTK)

$$\widehat{H}^{(\ell)}_{i_1 i_2;\delta_1\delta_2} \equiv \sum_{\mu,\nu} \lambda_{\mu\nu} \frac{d\widehat{z}^{(\ell)}_{i_1;\delta_1}}{d\theta_\mu} \frac{d\widehat{z}^{(\ell)}_{i_2;\delta_2}}{d\theta_\nu}$$

$$\{\theta_\mu\} = \{b^{(\ell)}_i, W^{(\ell)}_{ij}\}$$

$$\widehat{H}_{i_1 i_2;\delta_1\delta_2} = \widehat{H}^{(L)}_{i_1 i_2;\delta_1\delta_2}$$

Diagonal, group-by-group, learning rate:

$$\lambda_{b^{(\ell)}_{i_1} b^{(\ell)}_{i_2}} = \delta_{i_1 i_2} \lambda^{(\ell)}_b \,, \qquad \lambda_{W^{(\ell)}_{i_1 j_1} W^{(\ell)}_{i_2 j_2}} = \delta_{i_1 i_2} \delta_{j_1 j_2} \frac{\lambda^{(\ell)}_W}{n_{\ell-1}}$$

good wide limit

Two Pedagogical Simplifications

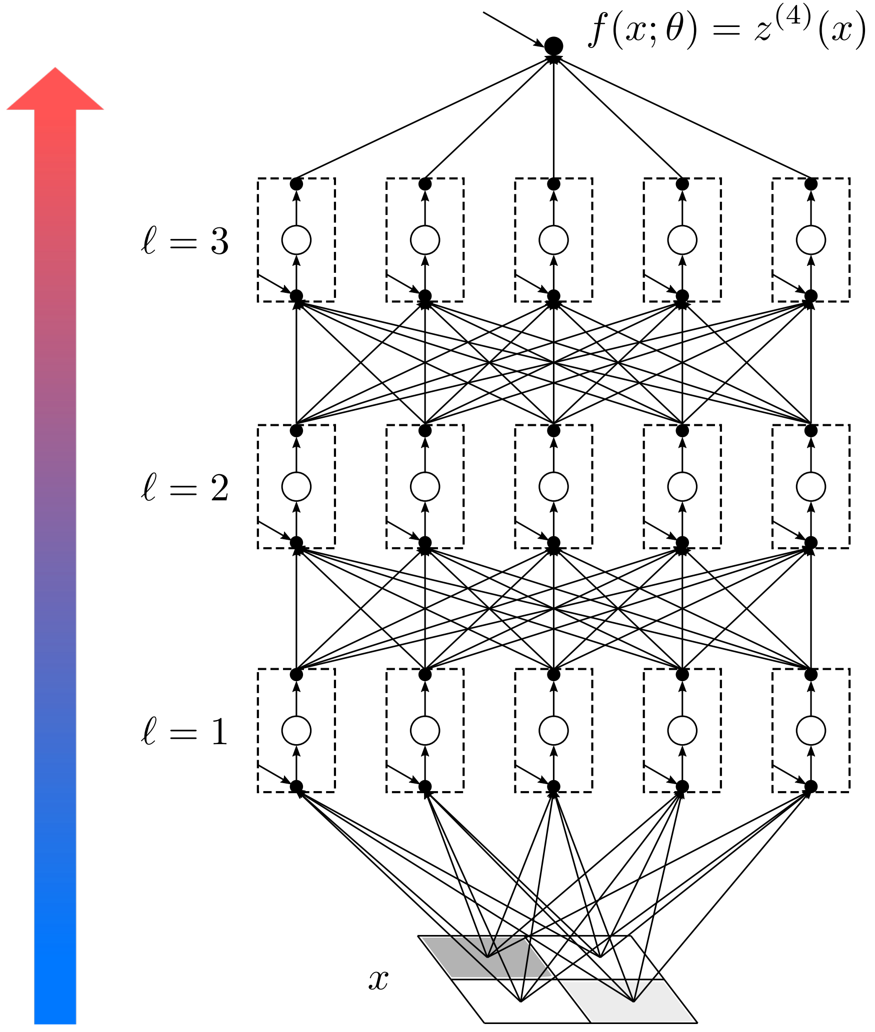
[See "PDLT" (arXiv:2106.10165) for more general cases.]

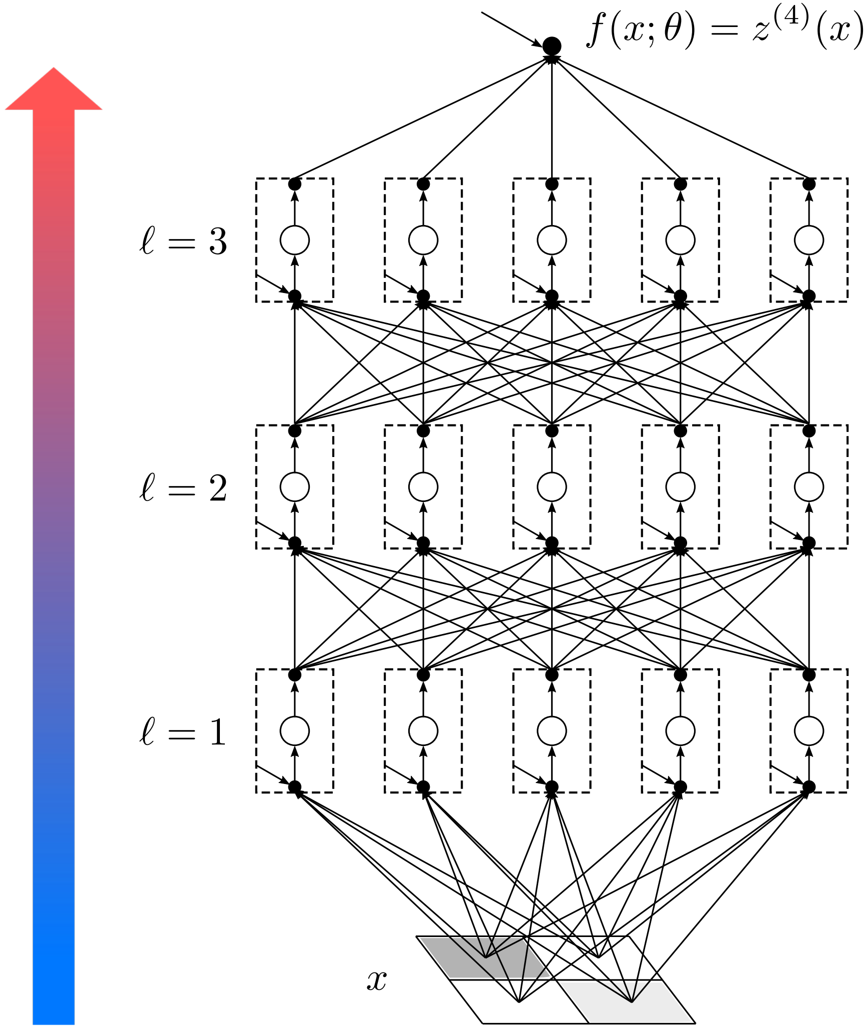
1. Single input; drop sample indices

$$x_{j;\delta} \to x_j\,, \quad \widehat{z}^{(\ell)}_{j;\delta} \to \widehat{z}^{(\ell)}_{j}\,, \quad \widehat{H}^{(\ell)}_{i_1 i_2;\delta_1 \delta_2} \to \widehat{H}^{(\ell)}_{i_1 i_2}$$

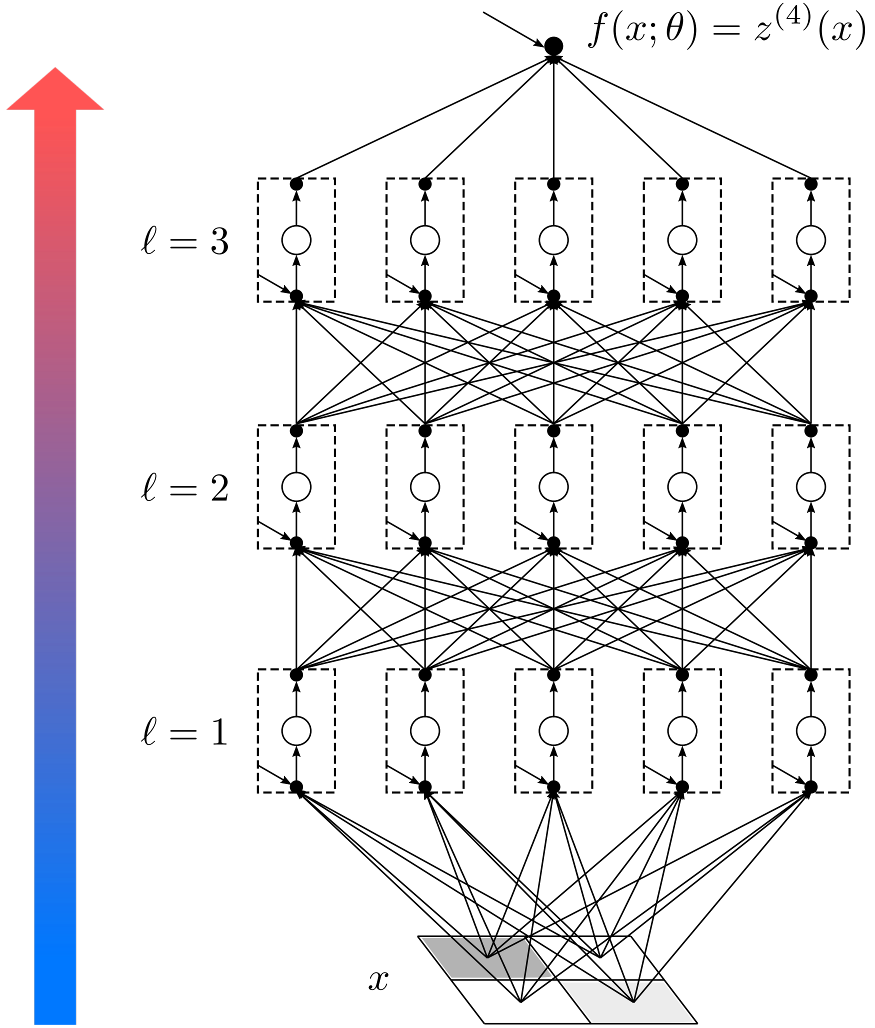
2. Layer-independent hyperparameters; drop layer indices from them

$$C_b^{(\ell)} = C_b, \quad C_W^{(\ell)} = C_W\,, \quad \lambda_b^{(\ell)} = \lambda_b, \quad \lambda_W^{(\ell)} = \lambda_W$$


$f(x;\theta) = z^{(4)}(x)$
$\ell = 3$
$\ell = 2$
$\ell = 1$
x

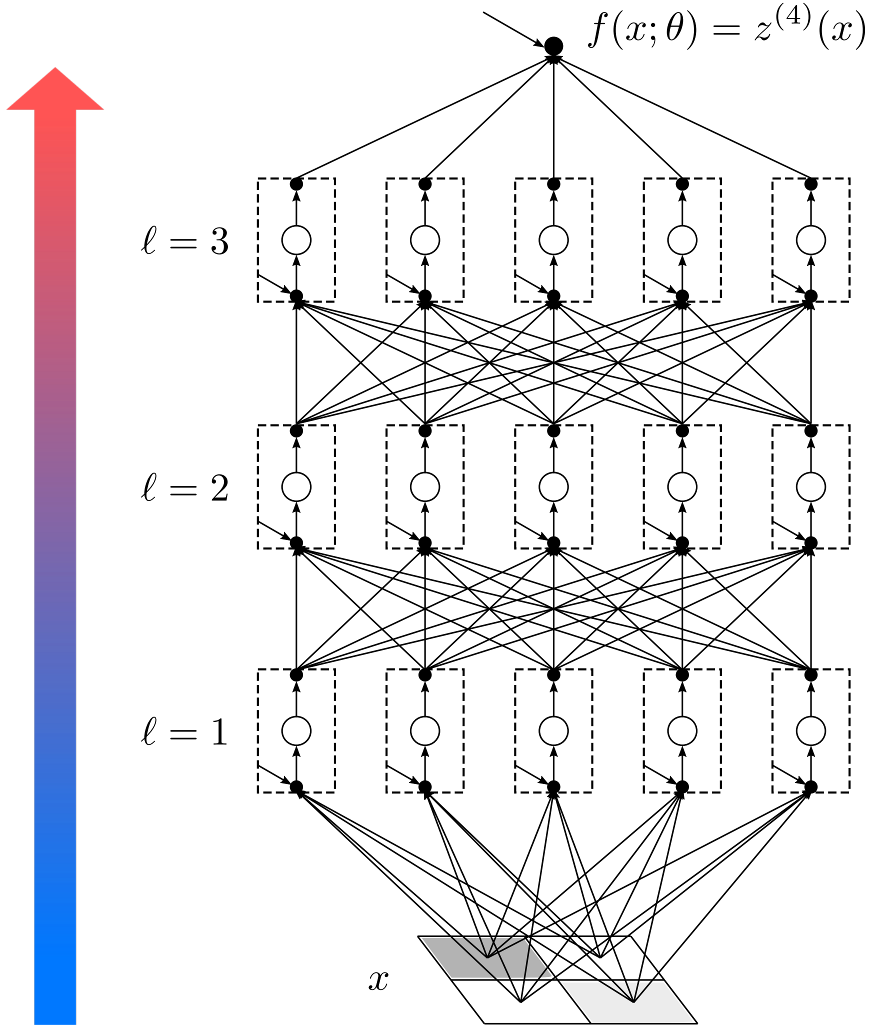


$$p\left(\widehat{z}^{(1)}, \widehat{H}^{(1)}, \widehat{\mathrm{d}H}^{(1)}, \ldots\right)$$



$$p\left(\widehat{z}^{(2)}, \widehat{H}^{(2)}, \widehat{\mathrm{d}H}^{(2)}, \ldots\right)$$

$$p\left(\widehat{z}^{(1)}, \widehat{H}^{(1)}, \widehat{\mathrm{d}H}^{(1)}, \ldots\right)$$



$$p\left(\widehat{z}^{(4)}, \widehat{H}^{(4)}, \widehat{\mathrm{d}H}^{(4)}, \ldots\right)$$

$$p\left(\widehat{z}^{(3)}, \widehat{H}^{(3)}, \widehat{\mathrm{d}H}^{(3)}, \ldots\right)$$

$$p\left(\widehat{z}^{(2)}, \widehat{H}^{(2)}, \widehat{\mathrm{d}H}^{(2)}, \ldots\right)$$

$$p\left(\widehat{z}^{(1)}, \widehat{H}^{(1)}, \widehat{\mathrm{d}H}^{(1)}, \ldots\right)$$

2. One-Layer Neural Networks

$$p\left(\widehat{z}^{(1)}, \widehat{H}^{(1)}, \widehat{\mathrm{d}H}^{(1)}, \ldots\right)$$

Statistics of $\widehat{z}_i^{(1)} = b_i^{(1)} + \sum_{j=1}^{n_0} W_{ij}^{(1)} x_j$

$$p\left(\widehat{z}^{(1)}\right)$$

$$\mathbb{E}[\widehat{z}_i^{(1)}],\ \mathbb{E}[\widehat{z}_{i_1}^{(1)}\widehat{z}_{i_2}^{(1)}],\ \mathbb{E}[\widehat{z}_{i_1}^{(1)}\widehat{z}_{i_2}^{(1)}\widehat{z}_{i_3}^{(1)}],\ \mathbb{E}[\widehat{z}_{i_1}^{(1)}\widehat{z}_{i_2}^{(1)}\widehat{z}_{i_3}^{(1)}\widehat{z}_{i_4}^{(1)}],\ \ldots$$

Statistics of $\widehat{z}_i^{(1)} = b_i^{(1)} + \sum_{j=1}^{n_0} W_{ij}^{(1)} x_j$

$$p\left(\widehat{z}^{(1)}\right)$$

$$\cancel{\mathbb{E}[\widehat{z}_i^{(1)}]},\ \mathbb{E}[\widehat{z}_{i_1}^{(1)}\widehat{z}_{i_2}^{(1)}],\ \cancel{\mathbb{E}[\widehat{z}_{i_1}^{(1)}\widehat{z}_{i_2}^{(1)}\widehat{z}_{i_3}^{(1)}]},\ \mathbb{E}[\widehat{z}_{i_1}^{(1)}\widehat{z}_{i_2}^{(1)}\widehat{z}_{i_3}^{(1)}\widehat{z}_{i_4}^{(1)}],\ \ldots$$

Statistics of $\widehat{z}_i^{(1)} = b_i^{(1)} + \sum_{j=1}^{n_0} W_{ij}^{(1)} x_j$

$$\mathbb{E}\left[\widehat{z}_{i_1}^{(1)}\widehat{z}_{i_2}^{(1)}\right] = \mathbb{E}\left[\left(b_{i_1}^{(1)} + \sum_{j_1=1}^{n_0} W_{i_1 j_1}^{(1)} x_{j_1}\right)\left(b_{i_2}^{(1)} + \sum_{j_2=1}^{n_0} W_{i_2 j_2}^{(1)} x_{j_2}\right)\right]$$

$$\mathbb{E}\left[b_{i_1}^{(1)} b_{i_2}^{(1)}\right] = \delta_{i_1 i_2} C_b\,, \quad \mathbb{E}\left[W_{i_1 j_1}^{(1)} W_{i_2 j_2}^{(1)}\right] = \delta_{i_1 i_2}\delta_{j_1 j_2}\frac{C_W}{n_0}$$

Statistics of $\widehat{z}_i^{(1)} = b_i^{(1)} + \sum_{j=1}^{n_0} W_{ij}^{(1)} x_j$

"Wick contraction"

$$\mathbb{E}\left[\widehat{z}_{i_1}^{(1)}\widehat{z}_{i_2}^{(1)}\right] = \mathbb{E}\left[\left(b_{i_1}^{(1)} + \sum_{j_1=1}^{n_0} W_{i_1 j_1}^{(1)} x_{j_1}\right)\left(b_{i_2}^{(1)} + \sum_{j_2=1}^{n_0} W_{i_2 j_2}^{(1)} x_{j_2}\right)\right]$$

$$= C_b \delta_{i_1 i_2} + \sum_{j_1, j_2=1}^{n_0} \frac{C_W}{n_0} \delta_{i_1 i_2} \delta_{j_1 j_2} x_{j_1} x_{j_2}$$

$$\mathbb{E}\left[b_{i_1}^{(1)} b_{i_2}^{(1)}\right] = \delta_{i_1 i_2} C_b \,, \quad \mathbb{E}\left[W_{i_1 j_1}^{(1)} W_{i_2 j_2}^{(1)}\right] = \delta_{i_1 i_2} \delta_{j_1 j_2} \frac{C_W}{n_0}$$

Statistics of $\widehat{z}_i^{(1)} = b_i^{(1)} + \sum_{j=1}^{n_0} W_{ij}^{(1)} x_j$

$$\mathbb{E}\left[\widehat{z}_{i_1}^{(1)}\widehat{z}_{i_2}^{(1)}\right] = \mathbb{E}\left[\left(b_{i_1}^{(1)} + \sum_{j_1=1}^{n_0} W_{i_1 j_1}^{(1)} x_{j_1}\right)\left(b_{i_2}^{(1)} + \sum_{j_2=1}^{n_0} W_{i_2 j_2}^{(1)} x_{j_2}\right)\right]$$

$$= C_b \delta_{i_1 i_2} + \sum_{j_1, j_2=1}^{n_0} \frac{C_W}{n_0} \delta_{i_1 i_2} \delta_{j_1 j_2} x_{j_1} x_{j_2}$$

$$\mathbb{E}\left[b_{i_1}^{(1)} b_{i_2}^{(1)}\right] = \delta_{i_1 i_2} C_b\,, \quad \mathbb{E}\left[W_{i_1 j_1}^{(1)} W_{i_2 j_2}^{(1)}\right] = \delta_{i_1 i_2}\delta_{j_1 j_2}\frac{C_W}{n_0}$$

Statistics of $\widehat{z}_i^{(1)} = b_i^{(1)} + \sum_{j=1}^{n_0} W_{ij}^{(1)} x_j$

$$
\begin{aligned}
\mathbb{E}\left[\widehat{z}_{i_1}^{(1)}\widehat{z}_{i_2}^{(1)}\right] =&\mathbb{E}\left[\left(b_{i_1}^{(1)} + \sum_{j_1=1}^{n_0} W_{i_1 j_1}^{(1)} x_{j_1}\right)\left(b_{i_2}^{(1)} + \sum_{j_2=1}^{n_0} W_{i_2 j_2}^{(1)} x_{j_2}\right)\right] \\
=&C_b\delta_{i_1 i_2} + \sum_{j_1,j_2=1}^{n_0} \frac{C_W}{n_0}\delta_{i_1 i_2}\delta_{j_1 j_2} x_{j_1} x_{j_2} \\
=&\delta_{i_1 i_2}\left[C_b + C_W\left(\frac{1}{n_0}\sum_{j=1}^{n_0} x_j^2\right)\right] \equiv \delta_{i_1 i_2} G^{(1)}
\end{aligned}
$$

Statistics of $\widehat{z}_i^{(1)} = b_i^{(1)} + \sum_{j=1}^{n_0} W_{ij}^{(1)} x_j$

$$
\begin{aligned}
&\mathbb{E}\left[\widehat{z}_{i_1}^{(1)}\widehat{z}_{i_2}^{(1)}\widehat{z}_{i_3}^{(1)}\widehat{z}_{i_4}^{(1)}\right] \\
=&\mathbb{E}\left[\left(b_{i_1}^{(1)} + \sum_{j_1=1}^{n_0} W_{i_1 j_1}^{(1)} x_{j_1}\right)\left(b_{i_2}^{(1)} + \sum_{j_2=1}^{n_0} W_{i_2 j_2}^{(1)} x_{j_2}\right)\left(b_{i_3}^{(1)} + \sum_{j_3=1}^{n_0} W_{i_3 j_3}^{(1)} x_{j_3}\right)\left(b_{i_4}^{(1)} + \sum_{j_4=1}^{n_0} W_{i_4 j_4}^{(1)} x_{j_4}\right)\right] \\
=&\left(\delta_{i_1 i_2}\delta_{i_3 i_4} + \delta_{i_1 i_3}\delta_{i_2 i_4} + \delta_{i_1 i_4}\delta_{i_2 i_3}\right) \\
&\times\left(C_b^2 + 2C_b C_W \frac{1}{n_0}\sum_{j=1}^{n_0} x_j^2 + C_W^2 \frac{1}{n_0^2}\sum_{j_1,j_2=0}^{n_0} x_{j_1}^2 x_{j_2}^2\right) \\
=&\left(G^{(1)}\right)^2\left(\delta_{i_1 i_2}\delta_{i_3 i_4} + \delta_{i_1 i_3}\delta_{i_2 i_4} + \delta_{i_1 i_4}\delta_{i_2 i_3}\right)
\end{aligned}
$$

Statistics of $\widehat{z}_i^{(1)} = b_i^{(1)} + \sum_{j=1}^{n_0} W_{ij}^{(1)} x_j$

$$
\begin{aligned}
&\mathbb{E}\left[\widehat{z}_{i_1}^{(1)}\widehat{z}_{i_2}^{(1)}\widehat{z}_{i_3}^{(1)}\widehat{z}_{i_4}^{(1)}\right] \\
=&\mathbb{E}\left[\left(b_{i_1}^{(1)} + \sum_{j_1=1}^{n_0} W_{i_1 j_1}^{(1)} x_{j_1}\right)\left(b_{i_2}^{(1)} + \sum_{j_2=1}^{n_0} W_{i_2 j_2}^{(1)} x_{j_2}\right)\left(b_{i_3}^{(1)} + \sum_{j_3=1}^{n_0} W_{i_3 j_3}^{(1)} x_{j_3}\right)\left(b_{i_4}^{(1)} + \sum_{j_4=1}^{n_0} W_{i_4 j_4}^{(1)} x_{j_4}\right)\right] \\
=&\left(\delta_{i_1 i_2}\delta_{i_3 i_4} + \delta_{i_1 i_3}\delta_{i_2 i_4} + \delta_{i_1 i_4}\delta_{i_2 i_3}\right) \\
&\times\left(\underline{C_b^2} + 2C_b C_W \frac{1}{n_0}\sum_{j=1}^{n_0} x_j^2 + C_W^2 \frac{1}{n_0^2}\sum_{j_1,j_2=0}^{n_0} x_{j_1}^2 x_{j_2}^2\right) \\
=&\left(G^{(1)}\right)^2\left(\delta_{i_1 i_2}\delta_{i_3 i_4} + \delta_{i_1 i_3}\delta_{i_2 i_4} + \delta_{i_1 i_4}\delta_{i_2 i_3}\right)
\end{aligned}
$$

$$
\mathbb{E}\left[b_{i_1}^{(1)} b_{i_2}^{(1)}\right] = \delta_{i_1 i_2} C_b\,, \quad \mathbb{E}\left[W_{i_1 j_1}^{(1)} W_{i_2 j_2}^{(1)}\right] = \delta_{i_1 i_2}\delta_{j_1 j_2}\frac{C_W}{n_0}
$$

Statistics of $\widehat{z}_i^{(1)} = b_i^{(1)} + \sum_{j=1}^{n_0} W_{ij}^{(1)} x_j$

$$
\begin{aligned}
&\mathbb{E}\left[\widehat{z}_{i_1}^{(1)}\widehat{z}_{i_2}^{(1)}\widehat{z}_{i_3}^{(1)}\widehat{z}_{i_4}^{(1)}\right] \\
=&\mathbb{E}\left[\left(b_{i_1}^{(1)} + \sum_{j_1=1}^{n_0} W_{i_1 j_1}^{(1)} x_{j_1}\right)\left(b_{i_2}^{(1)} + \sum_{j_2=1}^{n_0} W_{i_2 j_2}^{(1)} x_{j_2}\right)\left(b_{i_3}^{(1)} + \sum_{j_3=1}^{n_0} W_{i_3 j_3}^{(1)} x_{j_3}\right)\left(b_{i_4}^{(1)} + \sum_{j_4=1}^{n_0} W_{i_4 j_4}^{(1)} x_{j_4}\right)\right] \\
=&\left(\delta_{i_1 i_2}\delta_{i_3 i_4} + \delta_{i_1 i_3}\delta_{i_2 i_4} + \delta_{i_1 i_4}\delta_{i_2 i_3}\right) \\
&\times\left(\underline{C_b^2} + 2 C_b C_W \frac{1}{n_0}\sum_{j=1}^{n_0} x_j^2 + C_W^2 \frac{1}{n_0^2}\sum_{j_1, j_2 = 0}^{n_0} x_{j_1}^2 x_{j_2}^2\right) \\
=&\left(G^{(1)}\right)^2\left(\delta_{i_1 i_2}\delta_{i_3 i_4} + \delta_{i_1 i_3}\delta_{i_2 i_4} + \delta_{i_1 i_4}\delta_{i_2 i_3}\right)
\end{aligned}
$$

$$
\mathbb{E}\left[b_{i_1}^{(1)} b_{i_2}^{(1)}\right] = \delta_{i_1 i_2} C_b\,, \quad \mathbb{E}\left[W_{i_1 j_1}^{(1)} W_{i_2 j_2}^{(1)}\right] = \delta_{i_1 i_2}\delta_{j_1 j_2}\frac{C_W}{n_0}
$$

Statistics of $\widehat{z}_i^{(1)} = b_i^{(1)} + \sum_{j=1}^{n_0} W_{ij}^{(1)} x_j$

$$
\begin{aligned}
&\mathbb{E}\left[\widehat{z}_{i_1}^{(1)}\widehat{z}_{i_2}^{(1)}\widehat{z}_{i_3}^{(1)}\widehat{z}_{i_4}^{(1)}\right] \\
=&\mathbb{E}\left[\left(b_{i_1}^{(1)} + \sum_{j_1=1}^{n_0} W_{i_1 j_1}^{(1)} x_{j_1}\right)\left(b_{i_2}^{(1)} + \sum_{j_2=1}^{n_0} W_{i_2 j_2}^{(1)} x_{j_2}\right)\left(b_{i_3}^{(1)} + \sum_{j_3=1}^{n_0} W_{i_3 j_3}^{(1)} x_{j_3}\right)\left(b_{i_4}^{(1)} + \sum_{j_4=1}^{n_0} W_{i_4 j_4}^{(1)} x_{j_4}\right)\right] \\
=& \left(\delta_{i_1 i_2}\delta_{i_3 i_4} + \delta_{i_1 i_3}\delta_{i_2 i_4} + \delta_{i_1 i_4}\delta_{i_2 i_3}\right) \\
&\times \left(C_b^2 + 2C_b C_W \frac{1}{n_0}\sum_{j=1}^{n_0} x_j^2 + C_W^2 \frac{1}{n_0^2}\sum_{j_1,j_2=0}^{n_0} x_{j_1}^2 x_{j_2}^2\right) \\
=& \left(G^{(1)}\right)^2 \left(\delta_{i_1 i_2}\delta_{i_3 i_4} + \delta_{i_1 i_3}\delta_{i_2 i_4} + \delta_{i_1 i_4}\delta_{i_2 i_3}\right)
\end{aligned}
$$

$$
\mathbb{E}\left[b_{i_1}^{(1)} b_{i_2}^{(1)}\right] = \delta_{i_1 i_2} C_b\,, \quad \mathbb{E}\left[W_{i_1 j_1}^{(1)} W_{i_2 j_2}^{(1)}\right] = \delta_{i_1 i_2}\delta_{j_1 j_2}\frac{C_W}{n_0}
$$

Statistics of $\widehat{z}_i^{(1)} = b_i^{(1)} + \sum_{j=1}^{n_0} W_{ij}^{(1)} x_j$

$$
\begin{aligned}
&\mathbb{E}\left[\widehat{z}_{i_1}^{(1)}\widehat{z}_{i_2}^{(1)}\widehat{z}_{i_3}^{(1)}\widehat{z}_{i_4}^{(1)}\right] \\
=&\mathbb{E}\left[\left(b_{i_1}^{(1)} + \sum_{j_1=1}^{n_0} W_{i_1 j_1}^{(1)} x_{j_1}\right)\left(b_{i_2}^{(1)} + \sum_{j_2=1}^{n_0} W_{i_2 j_2}^{(1)} x_{j_2}\right)\left(b_{i_3}^{(1)} + \sum_{j_3=1}^{n_0} W_{i_3 j_3}^{(1)} x_{j_3}\right)\left(b_{i_4}^{(1)} + \sum_{j_4=1}^{n_0} W_{i_4 j_4}^{(1)} x_{j_4}\right)\right] \\
=&\left(\delta_{i_1 i_2}\delta_{i_3 i_4} + \delta_{i_1 i_3}\delta_{i_2 i_4} + \delta_{i_1 i_4}\delta_{i_2 i_3}\right) \\
&\times\left(C_b^2 + \underset{bbWW}{2C_b C_W \frac{1}{n_0}\sum_{j=1}^{n_0} x_j^2} + \underset{WWWW}{C_W^2 \frac{1}{n_0^2}\sum_{j_1,j_2=0}^{n_0} x_{j_1}^2 x_{j_2}^2}\right) \\
=&\left(G^{(1)}\right)^2\left(\delta_{i_1 i_2}\delta_{i_3 i_4} + \delta_{i_1 i_3}\delta_{i_2 i_4} + \delta_{i_1 i_4}\delta_{i_2 i_3}\right)
\end{aligned}
$$

$$
\mathbb{E}\left[b_{i_1}^{(1)} b_{i_2}^{(1)}\right] = \delta_{i_1 i_2} C_b\,, \quad \mathbb{E}\left[W_{i_1 j_1}^{(1)} W_{i_2 j_2}^{(1)}\right] = \delta_{i_1 i_2}\delta_{j_1 j_2}\frac{C_W}{n_0}
$$

Statistics of $\widehat{z}_i^{(1)} = b_i^{(1)} + \sum_{j=1}^{n_0} W_{ij}^{(1)} x_j$

$$
\begin{aligned}
&\mathbb{E}\left[\widehat{z}_{i_1}^{(1)}\widehat{z}_{i_2}^{(1)}\widehat{z}_{i_3}^{(1)}\widehat{z}_{i_4}^{(1)}\right] \\
=&\mathbb{E}\left[\left(b_{i_1}^{(1)} + \sum_{j_1=1}^{n_0} W_{i_1 j_1}^{(1)} x_{j_1}\right)\left(b_{i_2}^{(1)} + \sum_{j_2=1}^{n_0} W_{i_2 j_2}^{(1)} x_{j_2}\right)\left(b_{i_3}^{(1)} + \sum_{j_3=1}^{n_0} W_{i_3 j_3}^{(1)} x_{j_3}\right)\left(b_{i_4}^{(1)} + \sum_{j_4=1}^{n_0} W_{i_4 j_4}^{(1)} x_{j_4}\right)\right] \\
=&\left(\delta_{i_1 i_2}\delta_{i_3 i_4} + \delta_{i_1 i_3}\delta_{i_2 i_4} + \delta_{i_1 i_4}\delta_{i_2 i_3}\right) \\
&\times\left(C_b^2 + 2C_b C_W \frac{1}{n_0}\sum_{j=1}^{n_0} x_j^2 + C_W^2 \frac{1}{n_0^2}\sum_{j_1, j_2=0}^{n_0} x_{j_1}^2 x_{j_2}^2\right) \\
=&\left(G^{(1)}\right)^2 \left(\delta_{i_1 i_2}\delta_{i_3 i_4} + \delta_{i_1 i_3}\delta_{i_2 i_4} + \delta_{i_1 i_4}\delta_{i_2 i_3}\right)
\end{aligned}
$$

$$\mathbb{E}\left[\widehat{z}_{i_1}^{(1)}\widehat{z}_{i_2}^{(1)}\right] = G^{(1)}\delta_{i_1 i_2} = \left[C_b + C_W\left(\frac{1}{n_0}\sum_j x_j^2\right)\right]\delta_{i_1 i_2}$$

Statistics of $\widehat{z}_i^{(1)} = b_i^{(1)} + \sum_{j=1}^{n_0} W_{ij}^{(1)} x_j$

$$\mathbb{E}\left[\widehat{z}_{i_1}^{(1)}\widehat{z}_{i_2}^{(1)}\right] = G^{(1)}\delta_{i_1 i_2}$$

$$\mathbb{E}\left[\widehat{z}_{i_1}^{(1)}\widehat{z}_{i_2}^{(1)}\widehat{z}_{i_3}^{(1)}\widehat{z}_{i_4}^{(1)}\right] = \left(G^{(1)}\right)^2 \left(\delta_{i_1 i_2}\delta_{i_3 i_4} + \delta_{i_1 i_3}\delta_{i_2 i_4} + \delta_{i_1 i_4}\delta_{i_2 i_3}\right)$$

$$\cdots$$

Statistics of $\widehat{z}_i^{(1)} = b_i^{(1)} + \sum_{j=1}^{n_0} W_{ij}^{(1)} x_j$

$$\mathbb{E}\left[\widehat{z}_{i_1}^{(1)}\widehat{z}_{i_2}^{(1)}\right] = G^{(1)}\delta_{i_1 i_2}$$

$$\mathbb{E}\left[\widehat{z}_{i_1}^{(1)}\widehat{z}_{i_2}^{(1)}\widehat{z}_{i_3}^{(1)}\widehat{z}_{i_4}^{(1)}\right] = \left(G^{(1)}\right)^2 (\delta_{i_1 i_2}\delta_{i_3 i_4} + \delta_{i_1 i_3}\delta_{i_2 i_4} + \delta_{i_1 i_4}\delta_{i_2 i_3})$$

$$\cdots$$

$$\boxed{p\left(\widehat{z}^{(1)}\right) \propto \exp\left[-\frac{1}{2G^{(1)}}\sum_{i=1}^{n_1}\left(\widehat{z}_i^{(1)}\right)^2\right] = \prod_{i=1}^{n_1}\left\{\exp\left[-\frac{1}{2G^{(1)}}\left(\widehat{z}_i^{(1)}\right)^2\right]\right\}}$$

Statistics of $\widehat{z}_i^{(1)} = b_i^{(1)} + \sum_{j=1}^{n_0} W_{ij}^{(1)} x_j$

$$\boxed{p\left(\widehat{z}^{(1)}\right) \propto \exp\left[-\frac{1}{2G^{(1)}} \sum_{i=1}^{n_1} \left(\widehat{z}_i^{(1)}\right)^2\right] = \prod_{i=1}^{n_1} \left\{\exp\left[-\frac{1}{2G^{(1)}} \left(\widehat{z}_i^{(1)}\right)^2\right]\right\}}$$

- Neurons don’t talk to each other; they are statistically independent.
- We marginalized over/integrated out $b_i^{(1)}$ and $W_{ij}^{(1)}$.
- Two interpretations:
  (i) outputs of one-layer networks; or
  (ii) preactivations in the first layer of deeper networks.

Statistics of $\widehat{H}^{(1)}_{i_1 i_2}$

$$\widehat{H}^{(1)}_{i_1 i_2} \equiv \sum_{\mu,\nu} \lambda_{\mu\nu} \frac{d\widehat{z}^{(1)}_{i_1}}{d\theta_\mu} \frac{d\widehat{z}^{(1)}_{i_2}}{d\theta_\nu}$$

Statistics of $\widehat{H}^{(1)}_{i_1 i_2}$

$$\widehat{H}^{(1)}_{i_1 i_2} \equiv \sum_{\mu,\nu} \lambda_{\mu\nu} \frac{d\widehat{z}^{(1)}_{i_1}}{d\theta_\mu} \frac{d\widehat{z}^{(1)}_{i_2}}{d\theta_\nu}$$

$$= \lambda_b \sum_{j=1}^{n_1} \frac{d\widehat{z}^{(1)}_{i_1}}{db^{(1)}_j} \frac{d\widehat{z}^{(1)}_{i_2}}{db^{(1)}_j} + \frac{\lambda_W}{n_0} \sum_{j=1}^{n_1} \sum_{k=1}^{n_0} \frac{d\widehat{z}^{(1)}_{i_1}}{dW^{(1)}_{jk}} \frac{d\widehat{z}^{(1)}_{i_2}}{dW^{(1)}_{jk}}$$

$$\lambda_{b^{(1)}_{i_1} b^{(1)}_{i_2}} = \delta_{i_1 i_2} \lambda_b \,, \qquad \lambda_{W^{(1)}_{i_1 j_1} W^{(1)}_{i_2 j_2}} = \delta_{i_1 i_2} \delta_{j_1 j_2} \frac{\lambda_W}{n_0}$$

Statistics of $\widehat{H}^{(1)}_{i_1 i_2}$

$$\widehat{H}^{(1)}_{i_1 i_2} \equiv \sum_{\mu,\nu} \lambda_{\mu\nu} \frac{d\widehat{z}^{(1)}_{i_1}}{d\theta_\mu} \frac{d\widehat{z}^{(1)}_{i_2}}{d\theta_\nu}$$

$$= \lambda_b \sum_{j=1}^{n_1} \frac{d\widehat{z}^{(1)}_{i_1}}{db^{(1)}_j} \frac{d\widehat{z}^{(1)}_{i_2}}{db^{(1)}_j} + \frac{\lambda_W}{n_0} \sum_{j=1}^{n_1} \sum_{k=1}^{n_0} \frac{d\widehat{z}^{(1)}_{i_1}}{dW^{(1)}_{jk}} \frac{d\widehat{z}^{(1)}_{i_2}}{dW^{(1)}_{jk}}$$

$$= \lambda_b \sum_{j=1}^{n_1} \delta_{i_1 j} \delta_{i_2 j} + \frac{\lambda_W}{n_0} \sum_{j=1}^{n_1} \sum_{k=1}^{n_0} \delta_{i_1 j} x_k \delta_{i_2 j} x_k$$

$$\widehat{z}^{(1)}_i = b^{(1)}_i + \sum_{j=1}^{n_0} W^{(1)}_{ij} x_j$$

Statistics of $\widehat{H}^{(1)}_{i_1 i_2}$

$$
\begin{aligned}
\widehat{H}^{(1)}_{i_1 i_2} \equiv & \sum_{\mu,\nu} \lambda_{\mu\nu} \frac{d\widehat{z}^{(1)}_{i_1}}{d\theta_\mu} \frac{d\widehat{z}^{(1)}_{i_2}}{d\theta_\nu} \\
= & \lambda_b \sum_{j=1}^{n_1} \frac{d\widehat{z}^{(1)}_{i_1}}{db^{(1)}_j} \frac{d\widehat{z}^{(1)}_{i_2}}{db^{(1)}_j} + \frac{\lambda_W}{n_0} \sum_{j=1}^{n_1} \sum_{k=1}^{n_0} \frac{d\widehat{z}^{(1)}_{i_1}}{dW^{(1)}_{jk}} \frac{d\widehat{z}^{(1)}_{i_2}}{dW^{(1)}_{jk}} \\
= & \lambda_b \sum_{j=1}^{n_1} \delta_{i_1 j} \delta_{i_2 j} + \frac{\lambda_W}{n_0} \sum_{j=1}^{n_1} \sum_{k=1}^{n_0} \delta_{i_1 j} x_k \delta_{i_2 j} x_k \\
= & \lambda_b \delta_{i_1 i_2} + \frac{\lambda_W}{n_0} \delta_{i_1 i_2} \sum_{k=1}^{n_0} x_k x_k \\
= & \delta_{i_1 i_2} \left[\lambda_b + \lambda_W \left(\frac{1}{n_0} \sum_{j=1}^{n_0} x_j^2 \right) \right] \equiv \delta_{i_1 i_2} H^{(1)}
\end{aligned}
$$

Statistics of $\widehat{H}^{(1)}_{i_1 i_2}$

$$\boxed{\widehat{H}^{(1)}_{i_1 i_2} = \delta_{i_1 i_2} H^{(1)} = \delta_{i_1 i_2} \left[\lambda_b + \lambda_W \left(\frac{1}{n_0} \sum_{j=1}^{n_0} x_j^2 \right) \right]}$$

- “Deterministic”: it doesn't depend on any particular initialization; you always get the same number.
- “Frozen”: it cannot evolve during training; no representation learning.

Statistics of $\widehat{\mathrm{d}H}^{(1)}_{i_0 i_1 i_2}$

$$\widehat{\mathrm{d}H}^{(1)}_{i_0 i_1 i_2} \equiv \sum_{\mu_1,\nu_1,\mu_2,\nu_2} \lambda_{\mu_1\nu_1}\lambda_{\mu_2\nu_2} \frac{d^2\widehat{z}^{(1)}_{i_0}}{d\theta_{\mu_1} d\theta_{\mu_2}} \frac{d\widehat{z}^{(1)}_{i_1}}{d\theta_{\nu_1}} \frac{d\widehat{z}^{(1)}_{i_2}}{d\theta_{\nu_2}}$$

Statistics of $\widehat{\mathrm{d}H}^{(1)}_{i_0 i_1 i_2}$

$$\widehat{\mathrm{d}H}^{(1)}_{i_0 i_1 i_2} \equiv \sum_{\mu_1,\nu_1,\mu_2,\nu_2} \lambda_{\mu_1\nu_1}\lambda_{\mu_2\nu_2} \cancel{\frac{d^2\widehat{z}^{(1)}_{i_0}}{d\theta_{\mu_1}d\theta_{\mu_2}}} \frac{d\widehat{z}^{(1)}_{i_1}}{d\theta_{\nu_1}} \frac{d\widehat{z}^{(1)}_{i_2}}{d\theta_{\nu_2}} = 0$$

$$\widehat{z}^{(1)}_i = b^{(1)}_i + \sum_{j=1}^{n_0} W^{(1)}_{ij} x_j$$

Statistics of $\widehat{\mathrm{d}H}^{(1)}_{i_0 i_1 i_2}$

$$\boxed{\widehat{\mathrm{d}H}^{(1)}, \widehat{\mathrm{dd}H}^{(1)}, \ldots = 0}$$

- No representation learning.
- No algorithm dependence.

Statistics of One-Layer Neural Networks

$$p\left(\widehat{z}^{(1)}\right) \propto \exp\left[-\frac{1}{2G^{(1)}}\sum_{i=1}^{n_1}\left(\widehat{z}_i^{(1)}\right)^2\right] = \prod_{i=1}^{n_1}\left\{\exp\left[-\frac{1}{2G^{(1)}}\left(\widehat{z}_i^{(1)}\right)^2\right]\right\}$$

$$\widehat{H}^{(1)}_{i_1 i_2} = \delta_{i_1 i_2} H^{(1)} = \delta_{i_1 i_2}\left[\lambda_b + \lambda_W\left(\frac{1}{n_0}\sum_{j=1}^{n_0} x_j^2\right)\right]$$

$$\widehat{\mathrm{d}H}^{(1)}, \widehat{\mathrm{dd}H}^{(1)}, \ldots = 0$$

Statistics of One-Layer Neural Networks

$$\boxed{p\left(\widehat{z}^{(1)}\right) \propto \exp\left[-\frac{1}{2G^{(1)}}\sum_{i=1}^{n_1}\left(\widehat{z}_i^{(1)}\right)^2\right] = \prod_{i=1}^{n_1}\left\{\exp\left[-\frac{1}{2G^{(1)}}\left(\widehat{z}_i^{(1)}\right)^2\right]\right\}}$$

$$\boxed{\widehat{H}^{(1)}_{i_1 i_2} = \delta_{i_1 i_2} H^{(1)} = \delta_{i_1 i_2}\left[\lambda_b + \lambda_W\left(\frac{1}{n_0}\sum_{j=1}^{n_0} x_j^2\right)\right]}$$

$$\boxed{\widehat{\mathrm{d}H}^{(1)}, \widehat{\mathrm{dd}H}^{(1)}, \ldots = 0}$$

Linear dynamics: $$z_{i;\delta}(t+1) = z_{i;\delta}(t) - \eta \sum H_{\delta\tilde{\alpha}}\left[z_{i;\tilde{\alpha}}(t) - y_{i;\tilde{\alpha}}\right]$$

Simple solution: $$z^{\star}_{i;\delta} = \widehat{z}_{i;\delta} - \sum H^{(1)}_{\delta\tilde{\alpha}_1}\left(\left(\widetilde{H}^{(1)}\right)^{-1}\right)^{\tilde{\alpha}_1\tilde{\alpha}_2}\left[\widehat{z}_{i;\tilde{\alpha}_2} - y_{i;\tilde{\alpha}_2}\right]$$

Statistics of One-Layer Neural Networks

$$\boxed{p\left(\widehat{z}^{(1)}\right) \propto \exp\left[-\frac{1}{2G^{(1)}}\sum_{i=1}^{n_1}\left(\widehat{z}_i^{(1)}\right)^2\right] = \prod_{i=1}^{n_1}\left\{\exp\left[-\frac{1}{2G^{(1)}}\left(\widehat{z}_i^{(1)}\right)^2\right]\right\}}$$

$$\boxed{\widehat{H}^{(1)}_{i_1 i_2} = \delta_{i_1 i_2} H^{(1)} = \delta_{i_1 i_2}\left[\lambda_b + \lambda_W\left(\frac{1}{n_0}\sum_{j=1}^{n_0} x_j^2\right)\right]}$$

$$\boxed{\widehat{\mathrm{d}H}^{(1)}, \widehat{\mathrm{dd}H}^{(1)}, \ldots = 0}$$

$$z^\star_{i;\delta} = \widehat{z}_{i;\delta} - \sum H^{(1)}_{\delta\tilde{\alpha}_1}\left(\left(\widetilde{H}^{(1)}\right)^{-1}\right)^{\tilde{\alpha}_1\tilde{\alpha}_2}\left[\widehat{z}_{i;\tilde{\alpha}_2} - y_{i;\tilde{\alpha}_2}\right]$$

$$p\left(\widehat{z}, H\right) \to p\left(z^\star\right)$$

statistics at *initialization* statistics *after training*

$$\mathbb{E}\left[z^\star_{i;\delta}\right] = \sum_{\tilde{\alpha}_1,\tilde{\alpha}_2\in\mathcal{A}} H^{(1)}_{\delta\tilde{\alpha}_1}\left(\left(H^{(1)}\right)^{-1}\right)^{\tilde{\alpha}_1\tilde{\alpha}_2} y_{i;\tilde{\alpha}_2}$$

Statistics of One-Layer Neural Networks

$$\boxed{p\left(\widehat{z}^{(1)}\right) \propto \exp\left[-\frac{1}{2G^{(1)}}\sum_{i=1}^{n_1}\left(\widehat{z}_i^{(1)}\right)^2\right] = \prod_{i=1}^{n_1}\left\{\exp\left[-\frac{1}{2G^{(1)}}\left(\widehat{z}_i^{(1)}\right)^2\right]\right\}}$$

$$\boxed{\widehat{H}^{(1)}_{i_1 i_2} = \delta_{i_1 i_2} H^{(1)} = \delta_{i_1 i_2}\left[\lambda_b + \lambda_W\left(\frac{1}{n_0}\sum_{j=1}^{n_0} x_j^2\right)\right]}$$

$$\boxed{\widehat{\mathrm{d}H}^{(1)}, \widehat{\mathrm{dd}H}^{(1)}, \ldots = 0}$$

- Same trivial statistics for infinite-width neural networks of *any* fixed depth.
- No representation learning, no algorithm dependence; not a good model of deep learning.

We must study *deeper* networks of *finite width*!

3. Two-Layer Neural Networks

$$p\left(\widehat{z}^{(2)}, \widehat{H}^{(2)}, \widehat{\mathrm{d}H}^{(2)}, \ldots\right)$$

Statistics of $\widehat{z}_i^{(2)} = b_i^{(2)} + \sum_{j=1}^{n_1} W_{ij}^{(2)} \sigma\left(\widehat{z}_j^{(1)}\right)$

$$\mathbb{E}\left[\widehat{z}_{i_1}^{(2)}\widehat{z}_{i_2}^{(2)}\right] = \mathbb{E}\left[\left(b_{i_1}^{(2)} + \sum_{j_1=1}^{n_1} W_{i_1 j_1}^{(2)} \sigma\left(\widehat{z}_{j_1}^{(1)}\right)\right)\left(b_{i_2}^{(2)} + \sum_{j_2=1}^{n_1} W_{i_2 j_2}^{(2)} \sigma\left(\widehat{z}_{j_2}^{(1)}\right)\right)\right]$$

$$\mathbb{E}\left[b_{i_1}^{(2)} b_{i_2}^{(2)}\right] = \delta_{i_1 i_2} C_b\,, \quad \mathbb{E}\left[W_{i_1 j_1}^{(2)} W_{i_2 j_2}^{(2)}\right] = \delta_{i_1 i_2}\delta_{j_1 j_2}\frac{C_W}{n_1}$$

Statistics of $\widehat{z}_i^{(2)} = b_i^{(2)} + \sum_{j=1}^{n_1} W_{ij}^{(2)} \sigma\left(\widehat{z}_j^{(1)}\right)$

$$\mathbb{E}\left[\widehat{z}_{i_1}^{(2)}\widehat{z}_{i_2}^{(2)}\right] = \mathbb{E}\left[\left(b_{i_1}^{(2)} + \sum_{j_1=1}^{n_1} W_{i_1 j_1}^{(2)} \sigma\left(\widehat{z}_{j_1}^{(1)}\right)\right)\left(b_{i_2}^{(2)} + \sum_{j_2=1}^{n_1} W_{i_2 j_2}^{(2)} \sigma\left(\widehat{z}_{j_2}^{(1)}\right)\right)\right]$$

Wick $$= C_b \delta_{i_1 i_2} + \sum_{j_1, j_2=1}^{n_1} \frac{C_W}{n_1} \delta_{i_1 i_2} \delta_{j_1 j_2} \mathbb{E}\left[\sigma\left(\widehat{z}_{j_1}^{(1)}\right)\sigma\left(\widehat{z}_{j_2}^{(1)}\right)\right]$$

arrange $$= \delta_{i_1 i_2}\left[C_b + C_W\left(\frac{1}{n_1}\sum_{j=1}^{n_1} \mathbb{E}\left[\sigma\left(\widehat{z}_j^{(1)}\right)\sigma\left(\widehat{z}_j^{(1)}\right)\right]\right)\right]$$

$$\mathbb{E}\left[b_{i_1}^{(2)} b_{i_2}^{(2)}\right] = \delta_{i_1 i_2} C_b\,, \quad \mathbb{E}\left[W_{i_1 j_1}^{(2)} W_{i_2 j_2}^{(2)}\right] = \delta_{i_1 i_2}\delta_{j_1 j_2}\frac{C_W}{n_1}$$

Statistics of $\widehat{z}_i^{(2)} = b_i^{(2)} + \sum_{j=1}^{n_1} W_{ij}^{(2)} \sigma\left(\widehat{z}_j^{(1)}\right)$

$$
\begin{aligned}
\mathbb{E}\left[\widehat{z}_{i_1}^{(2)}\widehat{z}_{i_2}^{(2)}\right] =& \mathbb{E}\left[\left(b_{i_1}^{(2)} + \sum_{j_1=1}^{n_1} W_{i_1 j_1}^{(2)} \sigma\left(\widehat{z}_{j_1}^{(1)}\right)\right)\left(b_{i_2}^{(2)} + \sum_{j_2=1}^{n_1} W_{i_2 j_2}^{(2)} \sigma\left(\widehat{z}_{j_2}^{(1)}\right)\right)\right] \\
=& C_b \delta_{i_1 i_2} + \sum_{j_1, j_2=1}^{n_1} \frac{C_W}{n_1} \delta_{i_1 i_2} \delta_{j_1 j_2} \mathbb{E}\left[\sigma\left(\widehat{z}_{j_1}^{(1)}\right)\sigma\left(\widehat{z}_{j_2}^{(1)}\right)\right] \\
=& \delta_{i_1 i_2}\left[C_b + C_W\left(\frac{1}{n_1}\sum_{j=1}^{n_1} \mathbb{E}\left[\sigma\left(\widehat{z}_j^{(1)}\right)\sigma\left(\widehat{z}_j^{(1)}\right)\right]\right)\right] \\
=& \delta_{i_1 i_2}\left[C_b + C_W \langle \sigma(z)\sigma(z)\rangle_{G^{(1)}}\right] \equiv \delta_{i_1 i_2} G^{(2)}
\end{aligned}
$$

$$
\boxed{p\left(\widehat{z}^{(1)}\right) \propto \exp\left[-\frac{1}{2G^{(1)}}\sum_{i=1}^{n_1}\left(\widehat{z}_i^{(1)}\right)^2\right] = \prod_{i=1}^{n_1}\left\{\exp\left[-\frac{1}{2G^{(1)}}\left(\widehat{z}_i^{(1)}\right)^2\right]\right\}}
$$

$$
\langle f(z)\rangle_G \equiv \frac{1}{\sqrt{2\pi G}}\int dz\, f(z) e^{-\frac{z^2}{2G}}
$$

Statistics of $\widehat{z}_i^{(2)} = b_i^{(2)} + \sum_{j=1}^{n_1} W_{ij}^{(2)} \sigma\left(\widehat{z}_j^{(1)}\right)$

$$
\begin{aligned}
\mathbb{E}\left[\widehat{z}_{i_1}^{(2)}\widehat{z}_{i_2}^{(2)}\right] =&\mathbb{E}\left[\left(b_{i_1}^{(2)} + \sum_{j_1=1}^{n_1} W_{i_1 j_1}^{(2)} \sigma\left(\widehat{z}_{j_1}^{(1)}\right)\right)\left(b_{i_2}^{(2)} + \sum_{j_2=1}^{n_1} W_{i_2 j_2}^{(2)} \sigma\left(\widehat{z}_{j_2}^{(1)}\right)\right)\right] \\
=&C_b \delta_{i_1 i_2} + \sum_{j_1, j_2=1}^{n_1} \frac{C_W}{n_1} \delta_{i_1 i_2} \delta_{j_1 j_2} \mathbb{E}\left[\sigma\left(\widehat{z}_{j_1}^{(1)}\right)\sigma\left(\widehat{z}_{j_2}^{(1)}\right)\right] \\
=&\delta_{i_1 i_2}\left[C_b + C_W\left(\frac{1}{n_1}\sum_{j=1}^{n_1}\mathbb{E}\left[\sigma\left(\widehat{z}_j^{(1)}\right)\sigma\left(\widehat{z}_j^{(1)}\right)\right]\right)\right] \\
=&\delta_{i_1 i_2}\left[C_b + C_W \langle \sigma(z)\sigma(z)\rangle_{G^{(1)}}\right] \equiv \delta_{i_1 i_2} G^{(2)}
\end{aligned}
$$

- Recursive.
- $\mathbb{E}\left[W_{i_1 j_1}^{(2)} W_{i_2 j_2}^{(2)}\right] = \delta_{i_1 i_2}\delta_{j_1 j_2}\frac{C_W}{n_1}$ width-scaling was important.

Statistics of $\widehat{z}_i^{(2)} = b_i^{(2)} + \sum_{j=1}^{n_1} W_{ij}^{(2)} \sigma\left(\widehat{z}_j^{(1)}\right)$

$$
\begin{aligned}
&\mathbb{E}\left[\widehat{z}_{i_1}^{(2)}\widehat{z}_{i_2}^{(2)}\widehat{z}_{i_3}^{(2)}\widehat{z}_{i_4}^{(2)}\right] \\
=&\mathbb{E}\left[\left(b_{i_1}^{(2)} + \sum_{j_1=1}^{n_1} W_{i_1 j_1}^{(2)} \sigma\left(\widehat{z}_{j_1}^{(1)}\right)\right)\cdots\left(b_{i_4}^{(2)} + \sum_{j_4=1}^{n_1} W_{i_4 j_4}^{(2)} \sigma\left(\widehat{z}_{j_4}^{(1)}\right)\right)\right]
\end{aligned}
$$

Wick $= \left(\delta_{i_1 i_2}\delta_{i_3 i_4} + \delta_{i_1 i_3}\delta_{i_2 i_4} + \delta_{i_1 i_4}\delta_{i_2 i_3}\right)$

$$
\times\left\{C_b^2 + 2C_b C_W \frac{1}{n_1}\sum_{j=1}^{n_1}\mathbb{E}\left[\sigma\left(\widehat{z}_j^{(1)}\right)\sigma\left(\widehat{z}_j^{(1)}\right)\right] + C_W^2 \frac{1}{n_1^2}\sum_{j_1, j_2=1}^{n_1}\mathbb{E}\left[\sigma\left(\widehat{z}_{j_1}^{(1)}\right)\sigma\left(\widehat{z}_{j_1}^{(1)}\right)\sigma\left(\widehat{z}_{j_2}^{(1)}\right)\sigma\left(\widehat{z}_{j_2}^{(1)}\right)\right]\right\}
$$

$$
\mathbb{E}\left[b_{i_1}^{(2)} b_{i_2}^{(2)}\right] = \delta_{i_1 i_2} C_b\,, \quad \mathbb{E}\left[W_{i_1 j_1}^{(2)} W_{i_2 j_2}^{(2)}\right] = \delta_{i_1 i_2}\delta_{j_1 j_2}\frac{C_W}{n_1}
$$

Statistics of $\widehat{z}_i^{(2)} = b_i^{(2)} + \sum_{j=1}^{n_1} W_{ij}^{(2)} \sigma\left(\widehat{z}_j^{(1)}\right)$

$$
\begin{aligned}
&\mathbb{E}\left[\widehat{z}_{i_1}^{(2)}\widehat{z}_{i_2}^{(2)}\widehat{z}_{i_3}^{(2)}\widehat{z}_{i_4}^{(2)}\right] \\
=&\mathbb{E}\left[\left(b_{i_1}^{(2)} + \sum_{j_1=1}^{n_1} W_{i_1 j_1}^{(2)} \sigma\left(\widehat{z}_{j_1}^{(1)}\right)\right)\cdots\left(b_{i_4}^{(2)} + \sum_{j_4=1}^{n_1} W_{i_4 j_4}^{(2)} \sigma\left(\widehat{z}_{j_4}^{(1)}\right)\right)\right] \\
\text{Wick}=&\left(\delta_{i_1 i_2}\delta_{i_3 i_4} + \delta_{i_1 i_3}\delta_{i_2 i_4} + \delta_{i_1 i_4}\delta_{i_2 i_3}\right) \\
&\times\left\{C_b^2 + 2C_b C_W \frac{1}{n_1}\sum_{j=1}^{n_1}\mathbb{E}\left[\sigma\left(\widehat{z}_j^{(1)}\right)\sigma\left(\widehat{z}_j^{(1)}\right)\right] + C_W^2 \frac{1}{n_1^2}\sum_{j_1,j_2=1}^{n_1}\mathbb{E}\left[\sigma\left(\widehat{z}_{j_1}^{(1)}\right)\sigma\left(\widehat{z}_{j_1}^{(1)}\right)\sigma\left(\widehat{z}_{j_2}^{(1)}\right)\sigma\left(\widehat{z}_{j_2}^{(1)}\right)\right]\right\} \\
=&\left(\delta_{i_1 i_2}\delta_{i_3 i_4} + \delta_{i_1 i_3}\delta_{i_2 i_4} + \delta_{i_1 i_4}\delta_{i_2 i_3}\right) \\
&\times\left\{C_b^2 + 2C_b C_W \langle\sigma(z)\sigma(z)\rangle_{G^{(1)}} + C_W^2\left[\frac{n_1^2 - n_1}{n_1^2}\langle\sigma(z)\sigma(z)\rangle_{G^{(1)}}\langle\sigma(z)\sigma(z)\rangle_{G^{(1)}} + \frac{n_1}{n_1^2}\langle\sigma(z)\sigma(z)\sigma(z)\sigma(z)\rangle_{G^{(1)}}\right]\right\}
\end{aligned}
$$

$j_1 \neq j_2$ $\qquad\qquad$ $j_1 = j_2$

Statistics of $\widehat{z}_i^{(2)} = b_i^{(2)} + \sum_{j=1}^{n_1} W_{ij}^{(2)} \sigma\left(\widehat{z}_j^{(1)}\right)$

$$\mathbb{E}\left[\widehat{z}_{i_1}^{(2)}\widehat{z}_{i_2}^{(2)}\widehat{z}_{i_3}^{(2)}\widehat{z}_{i_4}^{(2)}\right]$$

$$=\mathbb{E}\left[\left(b_{i_1}^{(2)} + \sum_{j_1=1}^{n_1} W_{i_1 j_1}^{(2)}\sigma\left(\widehat{z}_{j_1}^{(1)}\right)\right)\cdots\left(b_{i_4}^{(2)} + \sum_{j_4=1}^{n_1} W_{i_4 j_4}^{(2)}\sigma\left(\widehat{z}_{j_4}^{(1)}\right)\right)\right]$$

Wick $=(\delta_{i_1i_2}\delta_{i_3i_4} + \delta_{i_1i_3}\delta_{i_2i_4} + \delta_{i_1i_4}\delta_{i_2i_3})$

$$\times\left\{C_b^2 + 2C_bC_W\frac{1}{n_1}\sum_{j=1}^{n_1}\mathbb{E}\left[\sigma\left(\widehat{z}_j^{(1)}\right)\sigma\left(\widehat{z}_j^{(1)}\right)\right] + C_W^2\frac{1}{n_1^2}\sum_{j_1,j_2=1}^{n_1}\mathbb{E}\left[\sigma\left(\widehat{z}_{j_1}^{(1)}\right)\sigma\left(\widehat{z}_{j_1}^{(1)}\right)\sigma\left(\widehat{z}_{j_2}^{(1)}\right)\sigma\left(\widehat{z}_{j_2}^{(1)}\right)\right]\right\}$$

$$=(\delta_{i_1i_2}\delta_{i_3i_4} + \delta_{i_1i_3}\delta_{i_2i_4} + \delta_{i_1i_4}\delta_{i_2i_3})$$

$$\times\left\{C_b^2 + 2C_bC_W\langle\sigma(z)\sigma(z)\rangle_{G^{(1)}} + C_W^2\left[\frac{n_1^2-n_1}{n_1^2}\langle\sigma(z)\sigma(z)\rangle_{G^{(1)}}\langle\sigma(z)\sigma(z)\rangle_{G^{(1)}} + \frac{n_1}{n_1^2}\langle\sigma(z)\sigma(z)\sigma(z)\sigma(z)\rangle_{G^{(1)}}\right]\right\}$$

$$=(\delta_{i_1i_2}\delta_{i_3i_4} + \delta_{i_1i_3}\delta_{i_2i_4} + \delta_{i_1i_4}\delta_{i_2i_3})\left\{\left(G^{(2)}\right)^2 + \frac{1}{n_1}C_W^2\left[\langle\sigma(z)\sigma(z)\sigma(z)\sigma(z)\rangle_{G^{(1)}} - \langle\sigma(z)\sigma(z)\rangle_{G^{(1)}}^2\right]\right\}$$

Statistics of $\widehat{z}_i^{(2)} = b_i^{(2)} + \sum_{j=1}^{n_1} W_{ij}^{(2)} \sigma\left(\widehat{z}_j^{(1)}\right)$

$$
\begin{aligned}
&\mathbb{E}\left[\widehat{z}_{i_1}^{(2)}\widehat{z}_{i_2}^{(2)}\widehat{z}_{i_3}^{(2)}\widehat{z}_{i_4}^{(2)}\right] \\
=&\mathbb{E}\left[\left(b_{i_1}^{(2)} + \sum_{j_1=1}^{n_1} W_{i_1 j_1}^{(2)} \sigma\left(\widehat{z}_{j_1}^{(1)}\right)\right)\cdots\left(b_{i_4}^{(2)} + \sum_{j_4=1}^{n_1} W_{i_4 j_4}^{(2)} \sigma\left(\widehat{z}_{j_4}^{(1)}\right)\right)\right] \\
\text{Wick}=&\left(\delta_{i_1 i_2}\delta_{i_3 i_4} + \delta_{i_1 i_3}\delta_{i_2 i_4} + \delta_{i_1 i_4}\delta_{i_2 i_3}\right) \\
&\times\left\{C_b^2 + 2C_b C_W \frac{1}{n_1}\sum_{j=1}^{n_1}\mathbb{E}\left[\sigma\left(\widehat{z}_j^{(1)}\right)\sigma\left(\widehat{z}_j^{(1)}\right)\right] + C_W^2 \frac{1}{n_1^2}\sum_{j_1, j_2=1}^{n_1}\mathbb{E}\left[\sigma\left(\widehat{z}_{j_1}^{(1)}\right)\sigma\left(\widehat{z}_{j_1}^{(1)}\right)\sigma\left(\widehat{z}_{j_2}^{(1)}\right)\sigma\left(\widehat{z}_{j_2}^{(1)}\right)\right]\right\} \\
=&\left(\delta_{i_1 i_2}\delta_{i_3 i_4} + \delta_{i_1 i_3}\delta_{i_2 i_4} + \delta_{i_1 i_4}\delta_{i_2 i_3}\right) \\
&\times\left\{C_b^2 + 2C_b C_W \langle\sigma(z)\sigma(z)\rangle_{G^{(1)}} + C_W^2\left[\frac{n_1^2 - n_1}{n_1^2}\langle\sigma(z)\sigma(z)\rangle_{G^{(1)}}\langle\sigma(z)\sigma(z)\rangle_{G^{(1)}} + \frac{n_1}{n_1^2}\langle\sigma(z)\sigma(z)\sigma(z)\sigma(z)\rangle_{G^{(1)}}\right]\right\} \\
=&\left(\delta_{i_1 i_2}\delta_{i_3 i_4} + \delta_{i_1 i_3}\delta_{i_2 i_4} + \delta_{i_1 i_4}\delta_{i_2 i_3}\right)\left\{\left(G^{(2)}\right)^2 + \underline{\frac{1}{n_1}C_W^2\left[\langle\sigma(z)\sigma(z)\sigma(z)\sigma(z)\rangle_{G^{(1)}} - \langle\sigma(z)\sigma(z)\rangle_{G^{(1)}}^2\right]}\right\}
\end{aligned}
$$

Statistics of $\widehat{z}_i^{(2)} = b_i^{(2)} + \sum_{j=1}^{n_1} W_{ij}^{(2)} \sigma\left(\widehat{z}_j^{(1)}\right)$

$$
\begin{aligned}
&\mathbb{E}[\widehat{z}_{i_1}^{(2)}\widehat{z}_{i_2}^{(2)}\widehat{z}_{i_3}^{(2)}\widehat{z}_{i_4}^{(2)}]\big|_{\text{connected}} \\
\equiv&\mathbb{E}[\widehat{z}_{i_1}^{(2)}\widehat{z}_{i_2}^{(2)}\widehat{z}_{i_3}^{(2)}\widehat{z}_{i_4}^{(2)}] \\
&-\mathbb{E}[\widehat{z}_{i_1}^{(2)}\widehat{z}_{i_2}^{(2)}]\,\mathbb{E}[\widehat{z}_{i_3}^{(2)}\widehat{z}_{i_4}^{(2)}] - \mathbb{E}[\widehat{z}_{i_1}^{(2)}\widehat{z}_{i_3}^{(2)}]\,\mathbb{E}[\widehat{z}_{i_2}^{(2)}\widehat{z}_{i_4}^{(2)}] - \mathbb{E}[\widehat{z}_{i_1}^{(2)}\widehat{z}_{i_4}^{(2)}]\,\mathbb{E}[\widehat{z}_{i_2}^{(2)}\widehat{z}_{i_3}^{(2)}] \\
=&\frac{V^{(2)}}{n_1}(\delta_{i_1i_2}\delta_{i_3i_4} + \delta_{i_1i_3}\delta_{i_2i_4} + \delta_{i_1i_4}\delta_{i_2i_3})
\end{aligned}
$$

with $V^{(2)} = C_W^2 \left[\langle\sigma(z)\sigma(z)\sigma(z)\sigma(z)\rangle_{G^{(1)}} - \langle\sigma(z)\sigma(z)\rangle_{G^{(1)}}^2\right]$

Statistics of $\widehat{z}_i^{(2)} = b_i^{(2)} + \sum_{j=1}^{n_1} W_{ij}^{(2)} \sigma\left(\widehat{z}_j^{(1)}\right)$

$$
\begin{aligned}
&\mathbb{E}[\widehat{z}_{i_1}^{(2)}\widehat{z}_{i_2}^{(2)}\widehat{z}_{i_3}^{(2)}\widehat{z}_{i_4}^{(2)}]\big|_{\text{connected}} \\
\equiv&\mathbb{E}[\widehat{z}_{i_1}^{(2)}\widehat{z}_{i_2}^{(2)}\widehat{z}_{i_3}^{(2)}\widehat{z}_{i_4}^{(2)}] \\
&-\mathbb{E}[\widehat{z}_{i_1}^{(2)}\widehat{z}_{i_2}^{(2)}]\,\mathbb{E}[\widehat{z}_{i_3}^{(2)}\widehat{z}_{i_4}^{(2)}]-\mathbb{E}[\widehat{z}_{i_1}^{(2)}\widehat{z}_{i_3}^{(2)}]\,\mathbb{E}[\widehat{z}_{i_2}^{(2)}\widehat{z}_{i_4}^{(2)}]-\mathbb{E}[\widehat{z}_{i_1}^{(2)}\widehat{z}_{i_4}^{(2)}]\,\mathbb{E}[\widehat{z}_{i_2}^{(2)}\widehat{z}_{i_3}^{(2)}] \\
=&\frac{V^{(2)}}{n_1}\left(\delta_{i_1i_2}\delta_{i_3i_4}+\delta_{i_1i_3}\delta_{i_2i_4}+\delta_{i_1i_4}\delta_{i_2i_3}\right)
\end{aligned}
$$

with $V^{(2)} = C_W^2\left[\langle\sigma(z)\sigma(z)\sigma(z)\sigma(z)\rangle_{G^{(1)}} - \langle\sigma(z)\sigma(z)\rangle_{G^{(1)}}^2\right]$

Nearly-Gaussian distribution for $n_1 \gg 1$

[Cf. Gaussian distribution in the first layer:

$$
\begin{aligned}
&\mathbb{E}[\widehat{z}_{i_1}^{(1)}\widehat{z}_{i_2}^{(1)}\widehat{z}_{i_3}^{(1)}\widehat{z}_{i_4}^{(1)}]\big|_{\text{connected}} \\
\equiv&\mathbb{E}[\widehat{z}_{i_1}^{(1)}\widehat{z}_{i_2}^{(1)}\widehat{z}_{i_3}^{(1)}\widehat{z}_{i_4}^{(1)}] \\
&-\mathbb{E}[\widehat{z}_{i_1}^{(1)}\widehat{z}_{i_2}^{(1)}]\mathbb{E}[\widehat{z}_{i_3}^{(1)}\widehat{z}_{i_4}^{(1)}]-\mathbb{E}[\widehat{z}_{i_1}^{(1)}\widehat{z}_{i_3}^{(1)}]\mathbb{E}[\widehat{z}_{i_2}^{(1)}\widehat{z}_{i_4}^{(1)}]-\mathbb{E}[\widehat{z}_{i_1}^{(1)}\widehat{z}_{i_4}^{(1)}]\mathbb{E}[\widehat{z}_{i_2}^{(1)}\widehat{z}_{i_3}^{(1)}] \\
=&0
\end{aligned}
$$

]

Statistics of $\widehat{z}_i^{(2)} = b_i^{(2)} + \sum_{j=1}^{n_1} W_{ij}^{(2)} \sigma\left(\widehat{z}_j^{(1)}\right)$

$$\boxed{\begin{aligned}
\mathbb{E}\left[\widehat{z}_{i_1}^{(2)}\widehat{z}_{i_2}^{(2)}\right] &= G^{(2)}\delta_{i_1 i_2} \\
\mathbb{E}\left[\widehat{z}_{i_1}^{(2)}\widehat{z}_{i_2}^{(2)}\widehat{z}_{i_3}^{(2)}\widehat{z}_{i_4}^{(2)}\right]\Big|_{\text{connected}} &= \frac{1}{n_1}V^{(2)}\left(\delta_{i_1 i_2}\delta_{i_3 i_4} + \delta_{i_1 i_3}\delta_{i_2 i_4} + \delta_{i_1 i_4}\delta_{i_2 i_3}\right) \\
\mathbb{E}\left[\widehat{z}_{i_1}^{(2)}\widehat{z}_{i_2}^{(2)}\widehat{z}_{i_3}^{(2)}\widehat{z}_{i_4}^{(2)}\widehat{z}_{i_5}^{(2)}\widehat{z}_{i_6}^{(2)}\right]\Big|_{\text{connected}} &= O\left(\frac{1}{n_1^2}\right)
\end{aligned}}$$

- Gaussian in the infinite-width limit, too simple; specified by one number (one matrix – kernel – more generally)
- Sparse description at $O\left(1/n\right)$; specified by two numbers (two tensors more generally, one of them having four sample indices)
- Interacting neurons at finite width.

Statistics of $\widehat{H}^{(2)}_{i_1 i_2}$

$$\widehat{H}^{(2)}_{i_1 i_2} \equiv \sum_{\mu,\nu} \lambda_{\mu\nu} \frac{d\widehat{z}^{(2)}_{i_1}}{d\theta_\mu} \frac{d\widehat{z}^{(2)}_{i_2}}{d\theta_\nu}$$

Statistics of $\widehat{H}^{(2)}_{i_1 i_2}$

$$\widehat{H}^{(2)}_{i_1 i_2} \equiv \sum_{\mu,\nu} \lambda_{\mu\nu} \frac{d\widehat{z}^{(2)}_{i_1}}{d\theta_\mu} \frac{d\widehat{z}^{(2)}_{i_2}}{d\theta_\nu}$$

$$= \lambda_b \sum_{j=1}^{n_2} \frac{d\widehat{z}^{(2)}_{i_1}}{db^{(2)}_j} \frac{d\widehat{z}^{(2)}_{i_2}}{db^{(2)}_j} + \frac{\lambda_W}{n_1} \sum_{j=1}^{n_2} \sum_{k=1}^{n_1} \frac{d\widehat{z}^{(2)}_{i_1}}{dW^{(2)}_{jk}} \frac{d\widehat{z}^{(2)}_{i_2}}{dW^{(2)}_{jk}}$$

$$+ \lambda_b \sum_{j=1}^{n_1} \frac{d\widehat{z}^{(2)}_{i_1}}{db^{(1)}_j} \frac{d\widehat{z}^{(2)}_{i_2}}{db^{(1)}_j} + \frac{\lambda_W}{n_0} \sum_{j=1}^{n_1} \sum_{k=1}^{n_0} \frac{d\widehat{z}^{(2)}_{i_1}}{dW^{(1)}_{jk}} \frac{d\widehat{z}^{(2)}_{i_2}}{dW^{(1)}_{jk}}$$

$$\lambda_{b^{(\ell)}_{i_1} b^{(\ell)}_{i_2}} = \delta_{i_1 i_2} \lambda_b \,, \qquad \lambda_{W^{(\ell)}_{i_1 j_1} W^{(\ell)}_{i_2 j_2}} = \delta_{i_1 i_2} \delta_{j_1 j_2} \frac{\lambda_W}{n_{\ell-1}}$$

Statistics of $\widehat{H}^{(2)}_{i_1 i_2}$

$$\widehat{H}^{(2)}_{i_1 i_2} \equiv \sum_{\mu,\nu} \lambda_{\mu\nu} \frac{d\widehat{z}^{(2)}_{i_1}}{d\theta_\mu} \frac{d\widehat{z}^{(2)}_{i_2}}{d\theta_\nu}$$

$$= \lambda_b \sum_{j=1}^{n_2} \frac{d\widehat{z}^{(2)}_{i_1}}{db^{(2)}_j} \frac{d\widehat{z}^{(2)}_{i_2}}{db^{(2)}_j} + \frac{\lambda_W}{n_1} \sum_{j=1}^{n_2} \sum_{k=1}^{n_1} \frac{d\widehat{z}^{(2)}_{i_1}}{dW^{(2)}_{jk}} \frac{d\widehat{z}^{(2)}_{i_2}}{dW^{(2)}_{jk}}$$

1st piece

$$+ \lambda_b \sum_{j=1}^{n_1} \frac{d\widehat{z}^{(2)}_{i_1}}{db^{(1)}_j} \frac{d\widehat{z}^{(2)}_{i_2}}{db^{(1)}_j} + \frac{\lambda_W}{n_0} \sum_{j=1}^{n_1} \sum_{k=1}^{n_0} \frac{d\widehat{z}^{(2)}_{i_1}}{dW^{(1)}_{jk}} \frac{d\widehat{z}^{(2)}_{i_2}}{dW^{(1)}_{jk}}$$

2nd piece

Statistics of $\widehat{H}^{(2)}_{i_1 i_2}$

1st piece, the same as before:

$$\lambda_b \sum_{j=1}^{n_2} \frac{d\widehat{z}^{(2)}_{i_1}}{db^{(2)}_j} \frac{d\widehat{z}^{(2)}_{i_2}}{db^{(2)}_j} + \frac{\lambda_W}{n_1} \sum_{j=1}^{n_2} \sum_{k=1}^{n_1} \frac{d\widehat{z}^{(2)}_{i_1}}{dW^{(2)}_{jk}} \frac{d\widehat{z}^{(2)}_{i_2}}{dW^{(2)}_{jk}}$$

$$= \lambda_b \sum_{j=1}^{n_2} \delta_{i_1 j}\delta_{i_2 j} + \frac{\lambda_W}{n_1} \sum_{j=1}^{n_2} \sum_{k=1}^{n_1} \delta_{i_1 j}\sigma\left(\widehat{z}^{(1)}_k\right)\delta_{i_2 j}\sigma\left(\widehat{z}^{(1)}_k\right)$$

$$= \delta_{i_1 i_2} \left\{ \lambda_b + \lambda_W \left[\frac{1}{n_1} \sum_{j=1}^{n_1} \left(\sigma\left(\widehat{z}^{(1)}_j\right)\right)^2 \right] \right\}$$

$$\widehat{z}^{(2)}_i = b^{(2)}_i + \sum_{j=1}^{n_1} W^{(2)}_{ij} \sigma\left(\widehat{z}^{(1)}_j\right)$$

Statistics of $\widehat{H}^{(2)}_{i_1 i_2}$

1st piece, the same as before:

$$
\begin{aligned}
&\lambda_b \sum_{j=1}^{n_2} \frac{d\widehat{z}^{(2)}_{i_1}}{db^{(2)}_j} \frac{d\widehat{z}^{(2)}_{i_2}}{db^{(2)}_j} + \frac{\lambda_W}{n_1} \sum_{j=1}^{n_2} \sum_{k=1}^{n_1} \frac{d\widehat{z}^{(2)}_{i_1}}{dW^{(2)}_{jk}} \frac{d\widehat{z}^{(2)}_{i_2}}{dW^{(2)}_{jk}} \\
=&\lambda_b \sum_{j=1}^{n_2} \delta_{i_1 j}\delta_{i_2 j} + \frac{\lambda_W}{n_1} \sum_{j=1}^{n_2} \sum_{k=1}^{n_1} \delta_{i_1 j}\sigma\left(\widehat{z}^{(1)}_k\right)\delta_{i_2 j}\sigma\left(\widehat{z}^{(1)}_k\right) \\
=&\delta_{i_1 i_2}\left\{\lambda_b + \lambda_W \left[\underline{\frac{1}{n_1} \sum_{j=1}^{n_1}} \left(\sigma\left(\widehat{z}^{(1)}_j\right)\right)^2\right]\right\}
\end{aligned}
$$

$$\lambda_{W^{(2)}_{i_1 j_1} W^{(2)}_{i_2 j_2}} = \delta_{i_1 i_2}\delta_{j_1 j_2}\frac{\lambda_W}{n_1}$$ width-scaling was important.

Statistics of $\widehat{H}^{(2)}_{i_1 i_2}$

2nd piece, chain rule:

$$\lambda_b \sum_{j=1}^{n_1} \frac{d\widehat{z}^{(2)}_{i_1}}{db^{(1)}_j} \frac{d\widehat{z}^{(2)}_{i_2}}{db^{(1)}_j} + \frac{\lambda_W}{n_0} \sum_{j=1}^{n_1} \sum_{k=1}^{n_0} \frac{d\widehat{z}^{(2)}_{i_1}}{dW^{(1)}_{jk}} \frac{d\widehat{z}^{(2)}_{i_2}}{dW^{(1)}_{jk}}$$

$$= \sum_{m_1, m_2 = 1}^{n_1} \frac{d\widehat{z}^{(2)}_{i_1}}{d\widehat{z}^{(1)}_{m_1}} \frac{d\widehat{z}^{(2)}_{i_2}}{d\widehat{z}^{(1)}_{m_2}} \left[\lambda_b \sum_{j=1}^{n_1} \frac{d\widehat{z}^{(1)}_{m_1}}{db^{(1)}_j} \frac{d\widehat{z}^{(1)}_{m_2}}{db^{(1)}_j} + \frac{\lambda_W}{n_0} \sum_{j=1}^{n_1} \sum_{k=1}^{n_0} \frac{d\widehat{z}^{(2)}_{m_1}}{dW^{(1)}_{jk}} \frac{d\widehat{z}^{(2)}_{m_2}}{dW^{(1)}_{jk}} \right]$$

$$\frac{d\widehat{z}^{(2)}_i}{d\theta^{(1)}_\mu} = \sum_{m=1}^{n_1} \frac{d\widehat{z}^{(2)}_i}{d\widehat{z}^{(1)}_m} \frac{d\widehat{z}^{(1)}_m}{d\theta^{(1)}_\mu}$$

Statistics of $\widehat{H}^{(2)}_{i_1 i_2}$

2nd piece, chain rule:

$$\lambda_b \sum_{j=1}^{n_1} \frac{d\widehat{z}^{(2)}_{i_1}}{db^{(1)}_j} \frac{d\widehat{z}^{(2)}_{i_2}}{db^{(1)}_j} + \frac{\lambda_W}{n_0} \sum_{j=1}^{n_1} \sum_{k=1}^{n_0} \frac{d\widehat{z}^{(2)}_{i_1}}{dW^{(1)}_{jk}} \frac{d\widehat{z}^{(2)}_{i_2}}{dW^{(1)}_{jk}}$$

$$= \sum_{m_1, m_2=1}^{n_1} \frac{d\widehat{z}^{(2)}_{i_1}}{d\widehat{z}^{(1)}_{m_1}} \frac{d\widehat{z}^{(2)}_{i_2}}{d\widehat{z}^{(1)}_{m_2}} \left[\lambda_b \sum_{j=1}^{n_1} \frac{d\widehat{z}^{(1)}_{m_1}}{db^{(1)}_j} \frac{d\widehat{z}^{(1)}_{m_2}}{db^{(1)}_j} + \frac{\lambda_W}{n_0} \sum_{j=1}^{n_1} \sum_{k=1}^{n_0} \frac{d\widehat{z}^{(2)}_{m_1}}{dW^{(1)}_{jk}} \frac{d\widehat{z}^{(2)}_{m_2}}{dW^{(1)}_{jk}} \right]$$

$$= \sum_{m_1, m_2=1}^{n_1} \frac{d\widehat{z}^{(2)}_{i_1}}{d\widehat{z}^{(1)}_{m_1}} \frac{d\widehat{z}^{(2)}_{i_2}}{d\widehat{z}^{(1)}_{m_2}} \delta_{m_1 m_2} H^{(1)}$$

Statistics of $\widehat{H}^{(2)}_{i_1 i_2}$

2nd piece, chain rule:

$$\lambda_b \sum_{j=1}^{n_1} \frac{d\widehat{z}^{(2)}_{i_1}}{db^{(1)}_j} \frac{d\widehat{z}^{(2)}_{i_2}}{db^{(1)}_j} + \frac{\lambda_W}{n_0} \sum_{j=1}^{n_1} \sum_{k=1}^{n_0} \frac{d\widehat{z}^{(2)}_{i_1}}{dW^{(1)}_{jk}} \frac{d\widehat{z}^{(2)}_{i_2}}{dW^{(1)}_{jk}}$$

$$= \sum_{m_1, m_2 = 1}^{n_1} \frac{d\widehat{z}^{(2)}_{i_1}}{d\widehat{z}^{(1)}_{m_1}} \frac{d\widehat{z}^{(2)}_{i_2}}{d\widehat{z}^{(1)}_{m_2}} \left[\lambda_b \sum_{j=1}^{n_1} \frac{d\widehat{z}^{(1)}_{m_1}}{db^{(1)}_j} \frac{d\widehat{z}^{(1)}_{m_2}}{db^{(1)}_j} + \frac{\lambda_W}{n_0} \sum_{j=1}^{n_1} \sum_{k=1}^{n_0} \frac{d\widehat{z}^{(2)}_{m_1}}{dW^{(1)}_{jk}} \frac{d\widehat{z}^{(2)}_{m_2}}{dW^{(1)}_{jk}} \right]$$

$$= \sum_{m_1, m_2 = 1}^{n_1} \frac{d\widehat{z}^{(2)}_{i_1}}{d\widehat{z}^{(1)}_{m_1}} \frac{d\widehat{z}^{(2)}_{i_2}}{d\widehat{z}^{(1)}_{m_2}} \delta_{m_1 m_2} H^{(1)}$$

$$= \sum_{m_1, m_2 = 1}^{n_1} W^{(2)}_{i_1 m_1} \sigma' \left(\widehat{z}^{(1)}_{m_1} \right) W^{(2)}_{i_2 m_2} \sigma' \left(\widehat{z}^{(1)}_{m_2} \right) \delta_{m_1 m_2} H^{(1)}$$

$$\widehat{z}^{(2)}_i = b^{(2)}_i + \sum_{j=1}^{n_1} W^{(2)}_{ij} \sigma \left(\widehat{z}^{(1)}_j \right)$$

Statistics of $\widehat{H}^{(2)}_{i_1 i_2}$

2nd piece, chain rule:

$$\lambda_b \sum_{j=1}^{n_1} \frac{d\widehat{z}^{(2)}_{i_1}}{db^{(1)}_j} \frac{d\widehat{z}^{(2)}_{i_2}}{db^{(1)}_j} + \frac{\lambda_W}{n_0} \sum_{j=1}^{n_1} \sum_{k=1}^{n_0} \frac{d\widehat{z}^{(2)}_{i_1}}{dW^{(1)}_{jk}} \frac{d\widehat{z}^{(2)}_{i_2}}{dW^{(1)}_{jk}}$$

$$= \sum_{m_1, m_2 = 1}^{n_1} \frac{d\widehat{z}^{(2)}_{i_1}}{d\widehat{z}^{(1)}_{m_1}} \frac{d\widehat{z}^{(2)}_{i_2}}{d\widehat{z}^{(1)}_{m_2}} \left[\lambda_b \sum_{j=1}^{n_1} \frac{d\widehat{z}^{(1)}_{m_1}}{db^{(1)}_j} \frac{d\widehat{z}^{(1)}_{m_2}}{db^{(1)}_j} + \frac{\lambda_W}{n_0} \sum_{j=1}^{n_1} \sum_{k=1}^{n_0} \frac{d\widehat{z}^{(2)}_{m_1}}{dW^{(1)}_{jk}} \frac{d\widehat{z}^{(2)}_{m_2}}{dW^{(1)}_{jk}} \right]$$

$$= \sum_{m_1, m_2 = 1}^{n_1} \frac{d\widehat{z}^{(2)}_{i_1}}{d\widehat{z}^{(1)}_{m_1}} \frac{d\widehat{z}^{(2)}_{i_2}}{d\widehat{z}^{(1)}_{m_2}} \delta_{m_1 m_2} H^{(1)}$$

$$= \sum_{m_1, m_2 = 1}^{n_1} W^{(2)}_{i_1 m_1} \sigma' \left(\widehat{z}^{(1)}_{m_1} \right) W^{(2)}_{i_2 m_2} \sigma' \left(\widehat{z}^{(1)}_{m_2} \right) \delta_{m_1 m_2} H^{(1)}$$

$$= \sum_{m=1}^{n_1} W^{(2)}_{i_1 m} W^{(2)}_{i_2 m} \sigma' \left(\widehat{z}^{(1)}_{m} \right) \sigma' \left(\widehat{z}^{(1)}_{m} \right) H^{(1)}$$

Statistics of $\widehat{H}^{(2)}_{i_1 i_2}$

Putting things together, NTK *forward* equation:

$$\boxed{\widehat{H}^{(2)}_{i_1 i_2} = \delta_{i_1 i_2} \left\{ \lambda_b + \lambda_W \left[\frac{1}{n_1} \sum_{j=1}^{n_1} \left(\sigma\left(\widehat{z}^{(1)}_j\right)\right)^2 \right] \right\} + \sum_{m=1}^{n_1} W^{(2)}_{i_1 m} W^{(2)}_{i_2 m} \sigma' \left(\widehat{z}^{(1)}_m\right) \sigma' \left(\widehat{z}^{(1)}_m\right) H^{(1)}}$$

Statistics of $\widehat{H}^{(2)}_{i_1 i_2}$

Putting things together, NTK *forward* equation:

$$\boxed{\widehat{H}^{(2)}_{i_1 i_2} = \delta_{i_1 i_2}\left\{\lambda_b + \lambda_W\left[\frac{1}{n_1}\sum_{j=1}^{n_1}\left(\sigma\left(\widehat{z}^{(1)}_j\right)\right)^2\right]\right\} + \sum_{m=1}^{n_1} W^{(2)}_{i_1 m} W^{(2)}_{i_2 m}\sigma'\left(\widehat{z}^{(1)}_m\right)\sigma'\left(\widehat{z}^{(1)}_m\right) H^{(1)}}$$

- "Stochastic": it fluctuates from instantiation to instantiation.
- "Defrosted": it *can* evolve during training.

Statistics of $\widehat{H}^{(2)}_{i_1 i_2}$

Putting things together, NTK *forward* equation:

$$\boxed{\widehat{H}^{(2)}_{i_1 i_2} = \delta_{i_1 i_2}\left\{\lambda_b + \lambda_W\left[\frac{1}{n_1}\sum_{j=1}^{n_1}\left(\sigma\left(\widehat{z}^{(1)}_j\right)\right)^2\right]\right\} + \sum_{m=1}^{n_1} W^{(2)}_{i_1 m} W^{(2)}_{i_2 m}\sigma'\left(\widehat{z}^{(1)}_m\right)\sigma'\left(\widehat{z}^{(1)}_m\right) H^{(1)}}$$

Fun for the weekend (solutions in §8):

$$\mathbb{E}\left[\widehat{H}^{(2)}_{i_1 i_2}\right] = \delta_{i_1 i_2}\left[\lambda_b + \lambda_W\left\langle\sigma(z)\sigma(z)\right\rangle_{G^{(1)}} + C_W H^{(1)}\left\langle\sigma'(z)\sigma'(z)\right\rangle_{G^{(1)}}\right] \equiv \delta_{i_1 i_2} H^{(2)}$$

$$\mathbb{E}\left[\widehat{H}^{(2)}_{i_1 i_2}\widehat{H}^{(2)}_{i_3 i_4}\right] - \mathbb{E}\left[\widehat{H}^{(2)}_{i_1 i_2}\right]\mathbb{E}\left[\widehat{H}^{(2)}_{i_3 i_4}\right] = \frac{1}{n_1}\left[\delta_{i_1 i_2}\delta_{i_3 i_4}A^{(2)} + \left(\delta_{i_1 i_3}\delta_{i_2 i_4} + \delta_{i_1 i_4}\delta_{i_2 i_3}\right)B^{(2)}\right] = O(1/n)$$

$$\mathbb{E}\left[\widehat{H}^{(2)}_{i_1 i_2}\widehat{z}^{(2)}_{i_3}\widehat{z}^{(2)}_{i_4}\right] - \mathbb{E}\left[\widehat{H}^{(2)}_{i_1 i_2}\right]\mathbb{E}\left[\widehat{z}^{(2)}_{i_3}\widehat{z}^{(2)}_{i_4}\right] = \frac{1}{n_1}\left[\delta_{i_1 i_2}\delta_{i_3 i_4}D^{(2)} + \left(\delta_{i_1 i_3}\delta_{i_2 i_4} + \delta_{i_1 i_4}\delta_{i_2 i_3}\right)F^{(2)}\right] = O(1/n)$$

Statistics of $\widehat{H}^{(2)}_{i_1 i_2}$ and beyond

Putting things together, NTK *forward* equation:

$$\boxed{\widehat{H}^{(2)}_{i_1 i_2} = \delta_{i_1 i_2}\left\{\lambda_b + \lambda_W\left[\frac{1}{n_1}\sum_{j=1}^{n_1}\left(\sigma\left(\widehat{z}^{(1)}_j\right)\right)^2\right]\right\} + \sum_{m=1}^{n_1} W^{(2)}_{i_1 m} W^{(2)}_{i_2 m}\sigma'\left(\widehat{z}^{(1)}_m\right)\sigma'\left(\widehat{z}^{(1)}_m\right) H^{(1)}}$$

Fun for the weekend (solutions in §8, §11.2, and §∞.3):

$$\mathbb{E}\left[\widehat{H}^{(2)}_{i_1 i_2}\right] = \delta_{i_1 i_2}\left[\lambda_b + \lambda_W \langle\sigma(z)\sigma(z)\rangle_{G^{(1)}} + C_W H^{(1)}\langle\sigma'(z)\sigma'(z)\rangle_{G^{(1)}}\right] \equiv \delta_{i_1 i_2} H^{(2)}$$

$$\mathbb{E}\left[\widehat{H}^{(2)}_{i_1 i_2}\widehat{H}^{(2)}_{i_3 i_4}\right] - \mathbb{E}\left[\widehat{H}^{(2)}_{i_1 i_2}\right]\mathbb{E}\left[\widehat{H}^{(2)}_{i_3 i_4}\right] = \frac{1}{n_1}\left[\delta_{i_1 i_2}\delta_{i_3 i_4} A^{(2)} + \left(\delta_{i_1 i_3}\delta_{i_2 i_4} + \delta_{i_1 i_4}\delta_{i_2 i_3}\right) B^{(2)}\right] = O(1/n)$$

$$\mathbb{E}\left[\widehat{H}^{(2)}_{i_1 i_2}\widehat{z}^{(2)}_{i_3}\widehat{z}^{(2)}_{i_4}\right] - \mathbb{E}\left[\widehat{H}^{(2)}_{i_1 i_2}\right]\mathbb{E}\left[\widehat{z}^{(2)}_{i_3}\widehat{z}^{(2)}_{i_4}\right] = \frac{1}{n_1}\left[\delta_{i_1 i_2}\delta_{i_3 i_4} D^{(2)} + \left(\delta_{i_1 i_3}\delta_{i_2 i_4} + \delta_{i_1 i_4}\delta_{i_2 i_3}\right) F^{(2)}\right] = O(1/n)$$

$$\mathbb{E}\left[\widehat{\mathrm{d}H}^{(2)}_{i_0 i_1 i_2}\widehat{z}^{(2)}_{i_3}\right] = \frac{1}{n_1}\left[\delta_{i_0 i_3}\delta_{i_1 i_2} P^{(2)} + \left(\delta_{i_0 i_1}\delta_{i_2 i_3} + \delta_{i_0 i_2}\delta_{i_1 i_3}\right) Q^{(2)}\right] = O(1/n)$$

$$\mathbb{E}\left[\widehat{\mathrm{dd}H}^{(2)}\right] = O(1/n)$$

[*for smooth activation functions]

Statistics of $\widehat{H}^{(2)}_{i_1 i_2}$ and beyond

Putting things together, NTK *forward* equation:

$$\boxed{\widehat{H}^{(2)}_{i_1 i_2} = \delta_{i_1 i_2}\left\{\lambda_b + \lambda_W\left[\frac{1}{n_1}\sum_{j=1}^{n_1}\left(\sigma\left(\widehat{z}^{(1)}_j\right)\right)^2\right]\right\} + \sum_{m=1}^{n_1} W^{(2)}_{i_1 m} W^{(2)}_{i_2 m}\sigma'\left(\widehat{z}^{(1)}_m\right)\sigma'\left(\widehat{z}^{(1)}_m\right) H^{(1)}}$$

Fun for the weekend (solutions in §8, §11.2, and §∞.3):

$$\mathbb{E}\left[\widehat{H}^{(2)}_{i_1 i_2}\right] = \delta_{i_1 i_2}\left[\lambda_b + \lambda_W \langle\sigma(z)\sigma(z)\rangle_{G^{(1)}} + C_W H^{(1)}\langle\sigma'(z)\sigma'(z)\rangle_{G^{(1)}}\right] \equiv \delta_{i_1 i_2} H^{(2)}$$

$$\mathbb{E}\left[\widehat{H}^{(2)}_{i_1 i_2}\widehat{H}^{(2)}_{i_3 i_4}\right] - \mathbb{E}\left[\widehat{H}^{(2)}_{i_1 i_2}\right]\mathbb{E}\left[\widehat{H}^{(2)}_{i_3 i_4}\right] = \frac{1}{n_1}\left[\delta_{i_1 i_2}\delta_{i_3 i_4}A^{(2)} + \left(\delta_{i_1 i_3}\delta_{i_2 i_4} + \delta_{i_1 i_4}\delta_{i_2 i_3}\right)B^{(2)}\right] = O(1/n)$$

$$\mathbb{E}\left[\widehat{H}^{(2)}_{i_1 i_2}\widehat{z}^{(2)}_{i_3}\widehat{z}^{(2)}_{i_4}\right] - \mathbb{E}\left[\widehat{H}^{(2)}_{i_1 i_2}\right]\mathbb{E}\left[\widehat{z}^{(2)}_{i_3}\widehat{z}^{(2)}_{i_4}\right] = \frac{1}{n_1}\left[\delta_{i_1 i_2}\delta_{i_3 i_4}D^{(2)} + \left(\delta_{i_1 i_3}\delta_{i_2 i_4} + \delta_{i_1 i_4}\delta_{i_2 i_3}\right)F^{(2)}\right] = O(1/n)$$

$$\mathbb{E}\left[\widehat{\mathrm{d}H}^{(2)}_{i_0 i_1 i_2}\widehat{z}^{(2)}_{i_3}\right] = \frac{1}{n_1}\left[\delta_{i_0 i_3}\delta_{i_1 i_2}P^{(2)} + \left(\delta_{i_0 i_1}\delta_{i_2 i_3} + \delta_{i_0 i_2}\delta_{i_1 i_3}\right)Q^{(2)}\right] = O(1/n)$$

$$\mathbb{E}\left[\widehat{\mathrm{dd}H}^{(2)}\right] = O(1/n)$$

[*for smooth activation functions]

Representation Learning

Statistics of $\widehat{H}^{(2)}_{i_1 i_2}$ and beyond

$$z^{\star}_{i;\delta} = \widehat{z}_{i;\delta} - \sum \widehat{H}_{ij;\delta\tilde{\alpha}_1} \left(\widehat{H}^{-1}\right)^{jk;\tilde{\alpha}_1\tilde{\alpha}_2} \left[\widehat{z}_{k;\tilde{\alpha}} - y_{k;\tilde{\alpha}}\right]$$
$$+ \,\mathrm{despicable}(y, \widehat{z}, \widehat{H}, \widehat{\mathrm{d}H}, \widehat{\mathrm{dd}H}; \mathrm{algorithm})$$

$$H^{\star} \neq \widehat{H}$$

$$\mathbb{E}\left[\widehat{\mathrm{d}H}^{(2)}_{i_0 i_1 i_2} \widehat{z}^{(2)}_{i_3}\right] = \frac{1}{n_1}\left[\delta_{i_0 i_3}\delta_{i_1 i_2} P^{(2)} + \left(\delta_{i_0 i_1}\delta_{i_2 i_3} + \delta_{i_0 i_2}\delta_{i_1 i_3}\right) Q^{(2)}\right] = O(1/n)$$

$$\mathbb{E}\left[\widehat{\mathrm{dd}H}^{(2)}\right] = O(1/n)$$ [*for smooth activation functions]

Representation Learning

Statistics of Two-Layer Neural Networks

- Two interpretations:
  (i) outputs, NTK, … of a two-layer network; or
  (ii) preactivations, mid-layer NTK, … in the second layer of a deeper network.

- Neurons *do* talk to each other; they are statistically *dependent*.

- Yes representation learning (and yes algorithm dependence);
  they can now capture rich dynamics of real, finite-width, neural networks.

Statistics of Two-Layer Neural Networks

- Two interpretations:
  (i) outputs, NTK, … of a two-layer network; or
  (ii) preactivations, mid-layer NTK, … in the second layer of a deeper network.

- Neurons *do* talk to each other; they are statistically *dependent*.

- Yes representation learning (and yes algorithm dependence);
  they can now capture rich dynamics of real, finite-width, neural networks.

But what is being amplified by *deep* learning?

4. Deep Neural Networks

$$p\left(\widehat{z}^{(\ell)}, \widehat{H}^{(\ell)}, \widehat{\mathrm{d}H}^{(\ell)}, \ldots\right)$$

Statistics of $\widehat{z}_i^{(\ell+1)} = b_i^{(\ell+1)} + \sum_{j=1}^{n_1} W_{ij}^{(\ell+1)} \sigma\left(\widehat{z}_j^{(\ell)}\right)$

$$\mathbb{E}\left[\widehat{z}_{i_1}^{(\ell+1)}\widehat{z}_{i_2}^{(\ell+1)}\right] = \mathbb{E}\left[\left(b_{i_1}^{(\ell+1)} + \sum_{j_1=1}^{n_\ell} W_{i_1 j_1}^{(\ell+1)} \sigma\left(\widehat{z}_{j_1}^{(\ell)}\right)\right)\left(b_{i_2}^{(\ell+1)} + \sum_{j_2=1}^{n_\ell} W_{i_2 j_2}^{(\ell+1)} \sigma\left(\widehat{z}_{j_2}^{(\ell)}\right)\right)\right]$$

$$\mathbb{E}\left[b_{i_1}^{(\ell+1)} b_{i_2}^{(\ell+1)}\right] = \delta_{i_1 i_2} C_b\,, \quad \mathbb{E}\left[W_{i_1 j_1}^{(\ell+1)} W_{i_2 j_2}^{(\ell+1)}\right] = \delta_{i_1 i_2}\delta_{j_1 j_2}\frac{C_W}{n_\ell}$$

Statistics of $\widehat{z}_i^{(\ell+1)} = b_i^{(\ell+1)} + \sum_{j=1}^{n_1} W_{ij}^{(\ell+1)} \sigma\left(\widehat{z}_j^{(\ell)}\right)$

$$\mathbb{E}\left[\widehat{z}_{i_1}^{(\ell+1)}\widehat{z}_{i_2}^{(\ell+1)}\right] = \mathbb{E}\left[\left(b_{i_1}^{(\ell+1)} + \sum_{j_1=1}^{n_\ell} W_{i_1 j_1}^{(\ell+1)} \sigma\left(\widehat{z}_{j_1}^{(\ell)}\right)\right)\left(b_{i_2}^{(\ell+1)} + \sum_{j_2=1}^{n_\ell} W_{i_2 j_2}^{(\ell+1)} \sigma\left(\widehat{z}_{j_2}^{(\ell)}\right)\right)\right]$$

Wick $$= C_b \delta_{i_1 i_2} + \sum_{j_1, j_2 = 1}^{n_\ell} \frac{C_W}{n_\ell} \delta_{i_1 i_2} \delta_{j_1 j_2} \mathbb{E}\left[\sigma\left(\widehat{z}_{j_1}^{(\ell)}\right)\sigma\left(\widehat{z}_{j_2}^{(\ell)}\right)\right]$$

arrange $$= \delta_{i_1 i_2}\left[C_b + C_W\left(\frac{1}{n_\ell}\sum_{j=1}^{n_\ell}\mathbb{E}\left[\sigma\left(\widehat{z}_j^{(\ell)}\right)\sigma\left(\widehat{z}_j^{(\ell)}\right)\right]\right)\right]$$

$$= \delta_{i_1 i_2}\left[C_b + C_W \langle \sigma(z)\sigma(z)\rangle_{G^{(\ell)}} + O\left(\frac{1}{n}\right)\right] \equiv \delta_{i_1 i_2} G^{(\ell+1)}$$

$$\mathbb{E}\left[b_{i_1}^{(\ell+1)} b_{i_2}^{(\ell+1)}\right] = \delta_{i_1 i_2} C_b\,, \quad \mathbb{E}\left[W_{i_1 j_1}^{(\ell+1)} W_{i_2 j_2}^{(\ell+1)}\right] = \delta_{i_1 i_2}\delta_{j_1 j_2}\frac{C_W}{n_\ell}$$

Statistics of $\widehat{z}_i^{(\ell+1)} = b_i^{(\ell+1)} + \sum_{j=1}^{n_1} W_{ij}^{(\ell+1)} \sigma\left(\widehat{z}_j^{(\ell)}\right)$

$$
\begin{aligned}
\mathbb{E}\left[\widehat{z}_{i_1}^{(\ell+1)}\widehat{z}_{i_2}^{(\ell+1)}\right] =& \mathbb{E}\left[\left(b_{i_1}^{(\ell+1)} + \sum_{j_1=1}^{n_\ell} W_{i_1 j_1}^{(\ell+1)} \sigma\left(\widehat{z}_{j_1}^{(\ell)}\right)\right)\left(b_{i_2}^{(\ell+1)} + \sum_{j_2=1}^{n_\ell} W_{i_2 j_2}^{(\ell+1)} \sigma\left(\widehat{z}_{j_2}^{(\ell)}\right)\right)\right] \\
=& C_b \delta_{i_1 i_2} + \sum_{j_1, j_2=1}^{n_\ell} \frac{C_W}{n_\ell} \delta_{i_1 i_2} \delta_{j_1 j_2} \mathbb{E}\left[\sigma\left(\widehat{z}_{j_1}^{(\ell)}\right) \sigma\left(\widehat{z}_{j_2}^{(\ell)}\right)\right] \\
=& \delta_{i_1 i_2}\left[C_b + C_W\left(\frac{1}{n_\ell}\sum_{j=1}^{n_\ell} \mathbb{E}\left[\sigma\left(\widehat{z}_j^{(\ell)}\right)\sigma\left(\widehat{z}_j^{(\ell)}\right)\right]\right)\right] \\
\text{leading} \quad =& \delta_{i_1 i_2}\left[C_b + C_W \langle \sigma(z)\sigma(z) \rangle_{G^{(\ell)}} + O\left(\frac{1}{n}\right)\right] \equiv \delta_{i_1 i_2} G^{(\ell+1)}
\end{aligned}
$$

Statistics of $\widehat{z}_i^{(\ell+1)} = b_i^{(\ell+1)} + \sum_{j=1}^{n_1} W_{ij}^{(\ell+1)} \sigma\left(\widehat{z}_j^{(\ell)}\right)$

Two-point:

$$G^{(\ell+1)} = C_b + C_W \left\langle \sigma(z)\sigma(z) \right\rangle_{G^{(\ell)}} + O\left(\frac{1}{n}\right)$$

Statistics of $\widehat{z}_i^{(\ell+1)} = b_i^{(\ell+1)} + \sum_{j=1}^{n_1} W_{ij}^{(\ell+1)} \sigma\left(\widehat{z}_j^{(\ell)}\right)$

Two-point:

$$G^{(\ell+1)} = C_b + C_W \langle \sigma(z)\sigma(z) \rangle_{G^{(\ell)}} + O\left(\frac{1}{n}\right)$$

Four-point:

$$\begin{aligned}\frac{1}{n_\ell} V^{(\ell+1)} =& \frac{1}{n_\ell} C_W^2 \left[\langle \sigma(z)\sigma(z)\sigma(z)\sigma(z) \rangle_{G^{(\ell)}} - \langle \sigma(z)\sigma(z) \rangle_{G^{(\ell)}}^2 \right] \\ &+ \frac{C_W^2}{4n_{\ell-1}} \frac{V^{(\ell)}}{\left(G^{(\ell)}\right)^4} \left\langle \sigma(z)\sigma(z) \left(z^2 - G^{(\ell)}\right) \right\rangle_{G^{(\ell)}}^2 + O\left(\frac{1}{n^2}\right)\end{aligned}$$

Statistics of $\widehat{H}^{(\ell+1)}_{i_1 i_2}$

Two-point:

$$G^{(\ell+1)} = C_b + C_W \langle \sigma(z)\sigma(z)\rangle_{G^{(\ell)}} + O\left(\frac{1}{n}\right)$$

Four-point:

$$\frac{1}{n_\ell} V^{(\ell+1)} = \frac{1}{n_\ell} C_W^2 \left[\langle \sigma(z)\sigma(z)\sigma(z)\sigma(z)\rangle_{G^{(\ell)}} - \langle \sigma(z)\sigma(z)\rangle^2_{G^{(\ell)}}\right] + \frac{C_W^2}{4n_{\ell-1}} \frac{V^{(\ell)}}{\left(G^{(\ell)}\right)^4} \left\langle \sigma(z)\sigma(z)\left(z^2 - G^{(\ell)}\right)\right\rangle^2_{G^{(\ell)}} + O\left(\frac{1}{n^2}\right)$$

NTK mean:

$$H^{(\ell+1)} = \lambda_b + \lambda_W \langle \sigma(z)\sigma(z)\rangle_{G^{(\ell)}} + C_W H^{(\ell)} \langle \sigma'(z)\sigma'(z)\rangle_{G^{(\ell)}} + O\left(\frac{1}{n}\right)$$

Statistics of $\widehat{H}^{(\ell+1)}_{i_1 i_2}$

NTK *forward* equation:

$$\widehat{H}^{(\ell+1)}_{i_1 i_2} \equiv \sum_{\mu,\nu} \lambda_{\mu\nu} \frac{d\widehat{z}^{(\ell+1)}_{i_1}}{d\theta_\mu} \frac{d\widehat{z}^{(\ell+1)}_{i_2}}{d\theta_\nu}$$

$$= \lambda_b \sum_{j=1}^{n_{\ell+1}} \frac{d\widehat{z}^{(\ell+1)}_{i_1}}{db^{(\ell+1)}_j} \frac{d\widehat{z}^{(\ell+1)}_{i_2}}{db^{(\ell+1)}_j} + \frac{\lambda_W}{n_\ell} \sum_{j=1}^{n_{\ell+1}} \sum_{k=1}^{n_\ell} \frac{d\widehat{z}^{(\ell+1)}_{i_1}}{dW^{(\ell+1)}_{jk}} \frac{d\widehat{z}^{(\ell+1)}_{i_2}}{dW^{(\ell+1)}_{jk}}$$

1st trivial piece

$$+ \sum_{m_1,m_2=1}^{n_\ell} \frac{d\widehat{z}^{(\ell+1)}_{i_1}}{d\widehat{z}^{(\ell)}_{m_1}} \frac{d\widehat{z}^{(\ell+1)}_{i_2}}{d\widehat{z}^{(\ell)}_{m_2}} \widehat{H}^{(\ell)}_{m_1 m_2}$$

2nd chain-rule piece

Statistics of $\widehat{H}^{(\ell+1)}_{i_1 i_2}$

NTK *forward* equation:

$$\widehat{H}^{(\ell+1)}_{i_1 i_2} \equiv \sum_{\mu,\nu} \lambda_{\mu\nu} \frac{d\widehat{z}^{(\ell+1)}_{i_1}}{d\theta_\mu} \frac{d\widehat{z}^{(\ell+1)}_{i_2}}{d\theta_\nu}$$

$$= \delta_{i_1 i_2} \left\{ \lambda_b + \lambda_W \left[\frac{1}{n_\ell} \sum_{j=1}^{n_\ell} \left(\sigma\left(\widehat{z}^{(\ell)}_j\right)\right)^2 \right] \right\}$$

1st trivial piece

$$+ \sum_{m_1,m_2=1}^{n_\ell} W^{(\ell+1)}_{i_1 m_1} W^{(\ell+1)}_{i_2 m_2} \sigma'\left(\widehat{z}^{(\ell)}_{m_1}\right) \sigma'\left(\widehat{z}^{(\ell)}_{m_2}\right) \widehat{H}^{(\ell)}_{m_1 m_2}$$

2nd chain-rule piece

Statistics of $\widehat{H}^{(\ell+1)}_{i_1 i_2}$ and beyond

Two-point:

$$G^{(\ell+1)} = C_b + C_W \langle \sigma(z)\sigma(z) \rangle_{G^{(\ell)}} + O\left(\frac{1}{n}\right)$$

Four-point:

$$\frac{1}{n_\ell} V^{(\ell+1)} = \frac{1}{n_\ell} C_W^2 \left[\langle \sigma(z)\sigma(z)\sigma(z)\sigma(z) \rangle_{G^{(\ell)}} - \langle \sigma(z)\sigma(z) \rangle^2_{G^{(\ell)}} \right] + \frac{C_W^2}{4 n_{\ell-1}} \frac{V^{(\ell)}}{\left(G^{(\ell)}\right)^4} \left\langle \sigma(z)\sigma(z)\left(z^2 - G^{(\ell)}\right) \right\rangle^2_{G^{(\ell)}} + O\left(\frac{1}{n^2}\right)$$

NTK mean:

$$H^{(\ell+1)} = \lambda_b + \lambda_W \langle \sigma(z)\sigma(z) \rangle_{G^{(\ell)}} + C_W H^{(\ell)} \langle \sigma'(z)\sigma'(z) \rangle_{G^{(\ell)}} + O\left(\frac{1}{n}\right)$$

Similarly for $\overbrace{A^{(\ell)}, B^{(\ell)}, D^{(\ell)}, F^{(\ell)}}^{\text{NTK fluctuations (§8)}}, \underbrace{P^{(\ell)}, Q^{(\ell)}}_{\text{dNTK (§11.2)}}, \overbrace{R^{(\ell)}, S^{(\ell)}, T^{(\ell)}, U^{(\ell)}}^{\text{ddNTK (§}\infty\text{.3)}}$

Statistics of $\widehat{H}^{(\ell+1)}_{i_1 i_2}$ and beyond

Two-point:

$$G^{(\ell+1)} = C_b + C_W \langle \sigma(z)\sigma(z) \rangle_{G^{(\ell)}} + O\left(\frac{1}{n}\right)$$

Four-point:

$$\begin{aligned}\frac{1}{n_\ell} V^{(\ell+1)} =& \frac{1}{n_\ell} C_W^2 \left[\langle \sigma(z)\sigma(z)\sigma(z)\sigma(z) \rangle_{G^{(\ell)}} - \langle \sigma(z)\sigma(z) \rangle^2_{G^{(\ell)}} \right] \\ &+ \frac{C_W^2}{4 n_{\ell-1}} \frac{V^{(\ell)}}{\left(G^{(\ell)}\right)^4} \left\langle \sigma(z)\sigma(z)\left(z^2 - G^{(\ell)}\right) \right\rangle^2_{G^{(\ell)}} + O\left(\frac{1}{n^2}\right)\end{aligned} \qquad O\left(\frac{1}{n}\right)$$

NTK mean:

$$H^{(\ell+1)} = \lambda_b + \lambda_W \langle \sigma(z)\sigma(z) \rangle_{G^{(\ell)}} + C_W H^{(\ell)} \langle \sigma'(z)\sigma'(z) \rangle_{G^{(\ell)}} + O\left(\frac{1}{n}\right)$$

Similarly for $\overbrace{A^{(\ell)}, B^{(\ell)}, D^{(\ell)}, F^{(\ell)}}^{\text{NTK fluctuations (§8)}}, \underbrace{P^{(\ell)}, Q^{(\ell)}}_{\text{dNTK (§11.2)}}, \overbrace{R^{(\ell)}, S^{(\ell)}, T^{(\ell)}, U^{(\ell)}}^{\text{ddNTK (§∞.3)}}$ $\qquad O\left(\frac{1}{n}\right)$

Statistics of $\widehat{H}^{(\ell+1)}_{i_1 i_2}$ and beyond

Two-point:

$$G^{(\ell+1)} = C_b + C_W \langle \sigma(z)\sigma(z)\rangle_{G^{(\ell)}} + O\left(\frac{1}{n}\right)$$

Four-point:

$$\frac{1}{n_\ell}V^{(\ell+1)} = \frac{1}{n_\ell}C_W^2\left[\langle \sigma(z)\sigma(z)\sigma(z)\sigma(z)\rangle_{G^{(\ell)}} - \langle \sigma(z)\sigma(z)\rangle^2_{G^{(\ell)}}\right] + \frac{C_W^2}{4n_{\ell-1}}\frac{V^{(\ell)}}{\left(G^{(\ell)}\right)^4}\left\langle \sigma(z)\sigma(z)\left(z^2 - G^{(\ell)}\right)\right\rangle^2_{G^{(\ell)}} + O\left(\frac{1}{n^2}\right)$$

$$O\left(\frac{L}{n}\right)$$

NTK mean:

$$H^{(\ell+1)} = \lambda_b + \lambda_W \langle \sigma(z)\sigma(z)\rangle_{G^{(\ell)}} + C_W H^{(\ell)} \langle \sigma'(z)\sigma'(z)\rangle_{G^{(\ell)}} + O\left(\frac{1}{n}\right)$$

Similarly for $\overbrace{A^{(\ell)}, B^{(\ell)}, D^{(\ell)}, F^{(\ell)}}^{\text{NTK fluctuations (§8)}}, \underbrace{P^{(\ell)}, Q^{(\ell)}}_{\text{dNTK (§11.2)}}, \overbrace{R^{(\ell)}, S^{(\ell)}, T^{(\ell)}, U^{(\ell)}}^{\text{ddNTK (§∞.3)}}$ $\quad O\left(\frac{L}{n}\right)$

The Principle of Sparsity for WIDE Neural Networks

Sho Dan

$$p(\theta) \rightarrow p\left(\widehat{z}, \widehat{H}, \widehat{\mathrm{d}H}, \ldots\right) \rightarrow p(z^\star)$$

statistics at *initialization* statistics *after training*

The Principle of Sparsity for WIDE Neural Networks

$$p(\theta) \to p\left(\widehat{z}, \widehat{H}, \widehat{\mathrm{d}H}, \dots\right) \to p(z^{\star})$$

statistics at *initialization* statistics *after training*

- Infinite width:

$p\left(\widehat{z}, \widehat{H}\right)$ specified by $G^{(L)}, H^{(L)}$

The Principle of Sparsity for WIDE Neural Networks

$$p(\theta) \to p\left(\widehat{z}, \widehat{H}, \widehat{\mathrm{d}H}, \ldots\right) \to p(z^{\star})$$

statistics at *initialization* statistics *after training*

- Infinite width:

$$p\left(\widehat{z}, \widehat{H}\right)$$ specified by $$G^{(L)}, H^{(L)}$$

- Large-but-finite width at $O\left(\frac{1}{n}\right)$ $[n_1, n_2, \ldots, n_{L-1} \gg L]$:

$$p\left(\widehat{z}, \widehat{H}, \widehat{\mathrm{d}H}, \widehat{\mathrm{dd}H}\right)$$ specified by

$$G^{(L)}, H^{(L)}, V^{(L)}, A^{(L)}, B^{(L)}, D^{(L)}, F^{(L)}, P^{(L)}, Q^{(L)}, R^{(L)}, S^{(L)}, T^{(L)}, U^{(L)}$$

All determined through recursion relations
(RG-flow interpretation: §4.6)

$$G^{(L)}, H^{(L)}, V^{(L)}, A^{(L)}, B^{(L)}, D^{(L)}, F^{(L)}, P^{(L)}, Q^{(L)}, R^{(L)}, S^{(L)}, T^{(L)}, U^{(L)}$$

All determined through recursion relations
(RG-flow interpretation: §4.6)

Next Lecture: Solving Recursions
"The Principle of Criticality"
for
DEEP Neural Networks

$$G^{(L)}, H^{(L)}, V^{(L)}, A^{(L)}, B^{(L)}, D^{(L)}, F^{(L)}, P^{(L)}, Q^{(L)}, R^{(L)}, S^{(L)}, T^{(L)}, U^{(L)}$$

One more thing...

$$\mathbb{E}[\widehat{z}_{i_1}^{(\ell)}\widehat{z}_{i_2}^{(\ell)}] = \delta_{i_1 i_2} G^{(\ell)} = \delta_{i_1 i_2}\left[K^{(\ell)} + O\left(\frac{1}{n}\right)\right]$$

$$\boxed{K^{(\ell+1)} = C_b + C_W \langle \sigma(z)\sigma(z)\rangle_{K^{(\ell)}}}$$

$$\mathbb{E}[\widehat{H}_{i_1 i_2}^{(\ell)}] = \delta_{i_1 i_2} H^{(\ell)} = \delta_{i_1 i_2}\left[\Theta^{(\ell)} + O\left(\frac{1}{n}\right)\right]$$

$$\boxed{\Theta^{(\ell+1)} = \lambda_b + \lambda_W \langle \sigma(z)\sigma(z)\rangle_{K^{(\ell)}} + C_W \Theta^{(\ell)} \langle \sigma'(z)\sigma'(z)\rangle_{K^{(\ell)}}}$$

$$\mathbb{E}\left[\widehat{z}_{i_1}^{(\ell)}\widehat{z}_{i_2}^{(\ell)}\widehat{z}_{i_3}^{(\ell)}\widehat{z}_{i_4}^{(\ell)}\right]\Big|_{\text{connected}} = \frac{1}{n_{\ell-1}} V^{(\ell)} \left(\delta_{i_1 i_2}\delta_{i_3 i_4} + \delta_{i_1 i_3}\delta_{i_2 i_4} + \delta_{i_1 i_4}\delta_{i_2 i_3}\right)$$

$$\boxed{\begin{aligned}\frac{1}{n_\ell} V^{(\ell+1)} =& \frac{1}{n_\ell} C_W^2 \left[\langle \sigma(z)\sigma(z)\sigma(z)\sigma(z)\rangle_{K^{(\ell)}} - \langle \sigma(z)\sigma(z)\rangle_{K^{(\ell)}}^2\right] \\ &+ \frac{C_W^2}{4 n_{\ell-1}} \frac{V^{(\ell)}}{\left(K^{(\ell)}\right)^4} \left\langle \sigma(z)\sigma(z)\left(z^2 - K^{(\ell)}\right)\right\rangle_{K^{(\ell)}}^2 + \cancel{O\left(\frac{1}{n^2}\right)}\end{aligned}}$$