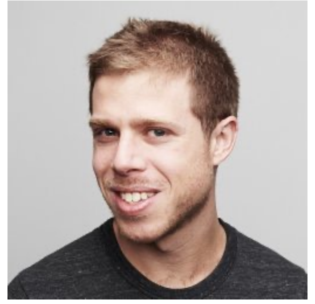
Effective Theory of Deep Neural Networks

Sho Yaida

Meta

Effective Theory of Deep Neural Networks



Dan Roberts, Sho Yaida, Boris Hanin [arXiv:2106.10165; Cambridge University Press]

Meta

Outline

1. Overview

2. Simplification at Large Width

3a. RG flow

3b. Criticality

1. Overview

Machine Learning in a Nutshell

Machine Learning in a Nutshell

$$f_{\text{target}}(x)$$

Machine Learning in a Nutshell

$$f(x; \theta^\star) \approx f_{\text{target}}(x)$$

Machine Learning in a Nutshell

$$\mathcal{L} = \left[f(x;\theta) - f_{\text{target}}(x)\right]^2$$

Machine Learning in a Nutshell

$$\mathcal{L} = \sum_{\alpha \in \mathcal{D}} \left[f(x_\alpha; \theta) - f_{\text{target}}(x_\alpha) \right]^2$$

Machine Learning in a Nutshell

$$\mathcal{L} = \sum_{\alpha \in \mathcal{D}} \left[f(x_\alpha; \theta) - f_{\text{target}}(x_\alpha) \right]^2$$

Machine Learning in a Nutshell

$$\mathcal{L} = \sum_{\alpha \in \mathcal{D}} \left[f(x_\alpha;\theta) - f_{\mathrm{target}}(x_\alpha) \right]^2$$

- Instantiate a model

$$f_{\mathrm{init}}(x) = f(x;\theta_{\mathrm{init}}) \quad \text{with} \quad \theta_{\mathrm{init}} \in p(\theta_{\mathrm{init}})$$

Machine Learning in a Nutshell

$$\mathcal{L} = \sum_{\alpha \in \mathcal{D}} \left[f(x_\alpha; \theta) - f_{\text{target}}(x_\alpha) \right]^2$$

- Instantiate a model

$$f_{\text{init}}(x) = f(x; \theta_{\text{init}}) \quad \text{with} \quad \theta_{\text{init}} \in p(\theta_{\text{init}})$$

- Train the model, e.g., by gradient descent

$$\theta_\mu(t+1) = \theta_\mu(t) - \eta \frac{\partial \mathcal{L}}{\partial \theta_\mu}\Big|_{\theta=\theta(t)}$$

Machine Learning in a Nutshell

$$\mathcal{L} = \sum_{\alpha \in \mathcal{D}} \left[f(x_\alpha; \theta) - f_{\text{target}}(x_\alpha) \right]^2$$

- Instantiate a model

$$f_{\text{init}}(x) = f(x; \theta_{\text{init}}) \quad \text{with} \quad \theta_{\text{init}} \in p(\theta_{\text{init}})$$

- Train the model, e.g., by gradient descent

$$\theta_\mu(t+1) = \theta_\mu(t) - \eta \frac{\partial \mathcal{L}}{\partial \theta_\mu}\bigg|_{\theta=\theta(t)}$$

- Use the trained model to make predictions

$$f_{\text{trained}}(x) = f(x; \theta_{\text{trained}})$$

Machine Learning in a Nutshell

$$\mathcal{L} = \sum_{\alpha \in \mathcal{D}} \left[f(x_\alpha;\theta) - f_{\text{target}}(x_\alpha) \right]^2$$

- Instantiate a model

$$f_{\text{init}}(x) = f(x;\theta_{\text{init}}) \quad \text{with} \quad \theta_{\text{init}} \in p(\theta_{\text{init}})$$

- Train the model, e.g., by gradient descent

$$\theta_\mu(t+1) = \theta_\mu(t) - \eta \frac{\partial \mathcal{L}}{\partial \theta_\mu}\bigg|_{\theta=\theta(t)}$$

- Use the trained model to make predictions

$$f_{\text{trained}}(x) = f(x;\theta_{\text{trained}})$$

Machine Learning in a Nutshell

$$\mathcal{L} = \sum_{\alpha \in \mathcal{D}} \left[f(x_\alpha;\theta) - f_{\text{target}}(x_\alpha) \right]^2$$

- Instantiate a model

$$f_{\text{init}}(x) = f(x;\theta_{\text{init}}) \quad \text{with} \quad \underline{\theta_{\text{init}} \in p(\theta_{\text{init}})}$$

- Train the model, e.g., by gradient descent

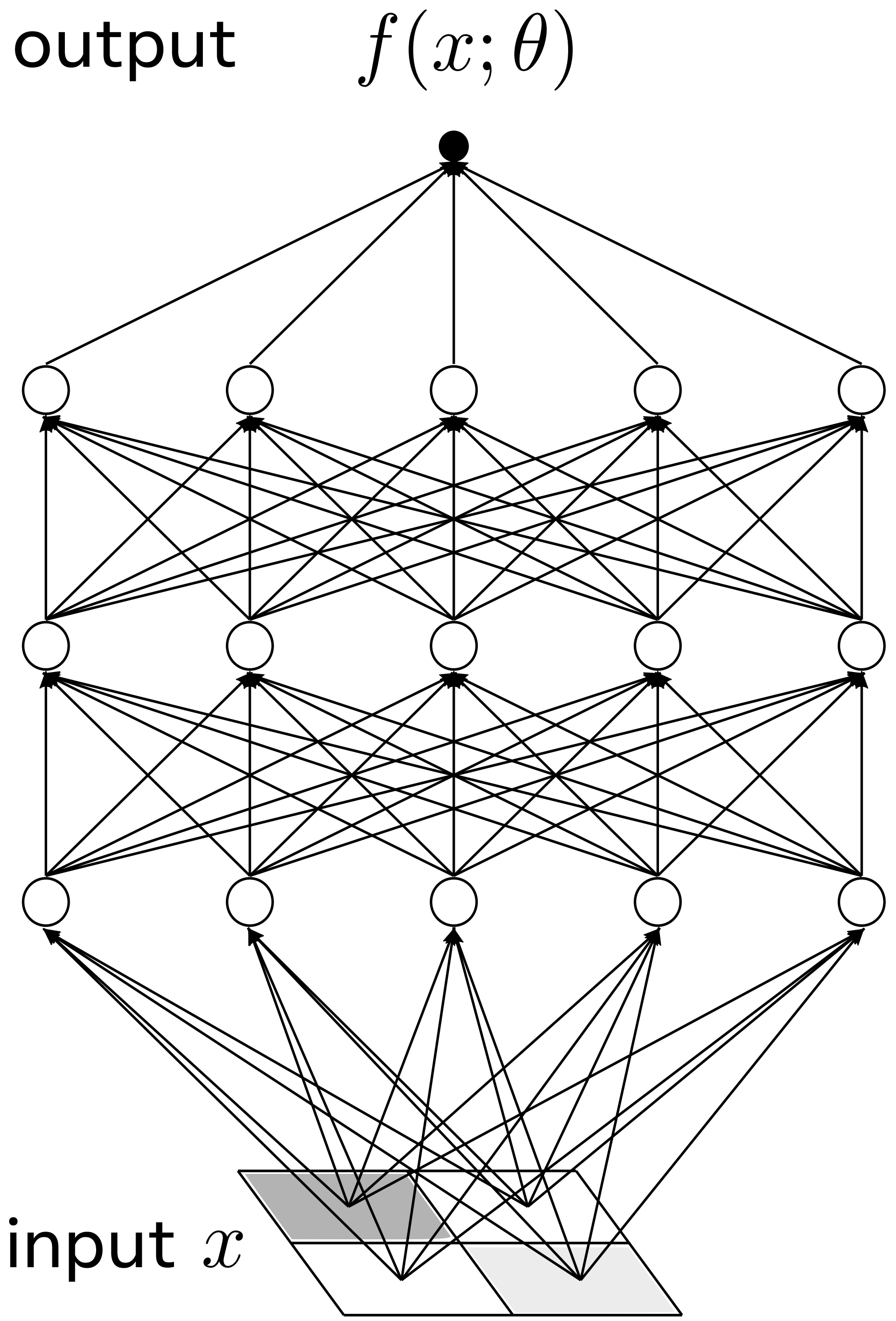
$$\theta_\mu(t+1) = \theta_\mu(t) - \eta \frac{\partial \mathcal{L}}{\partial \theta_\mu}\bigg|_{\theta=\theta(t)}$$

- Use the trained model to make predictions

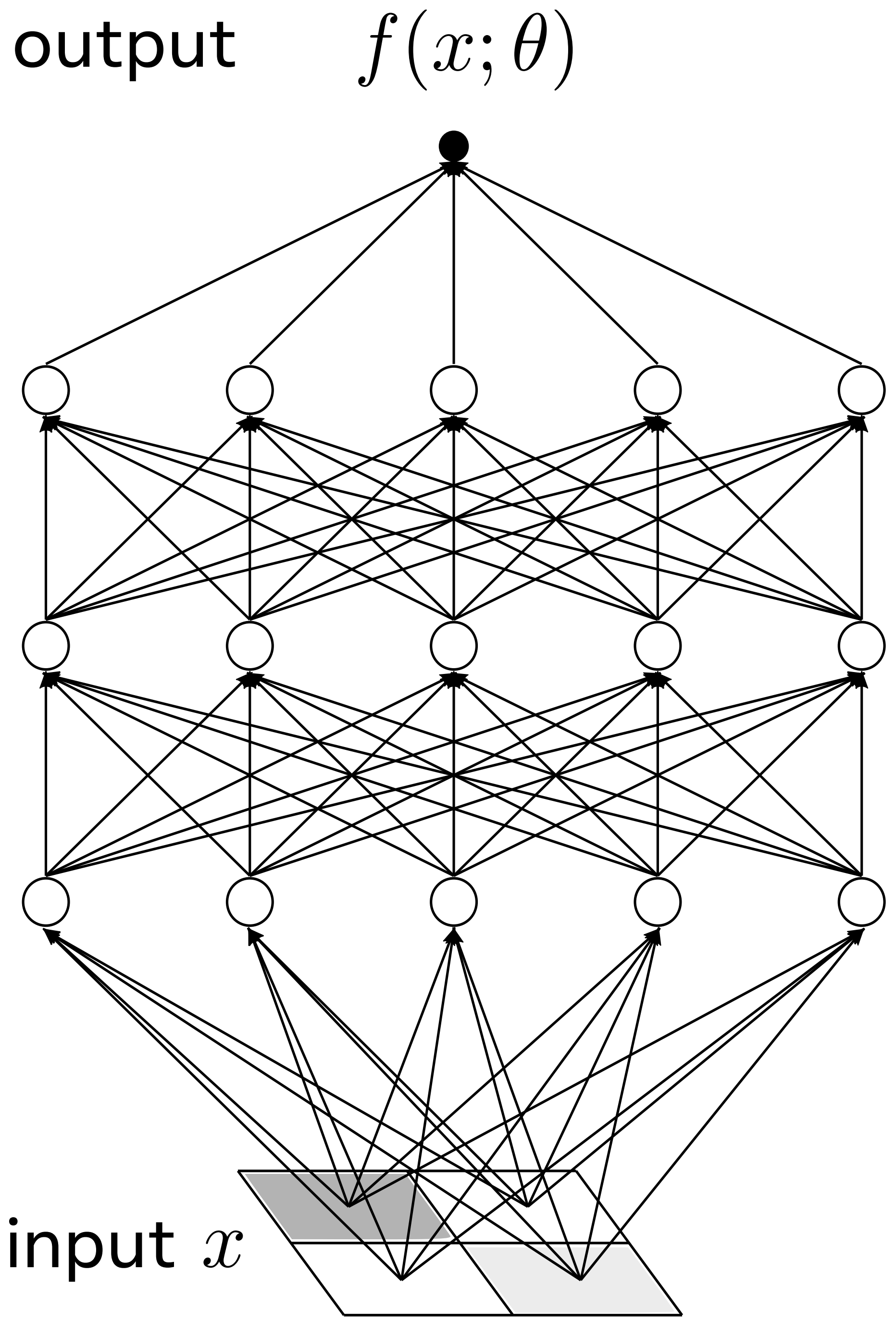
$$p(f_{\text{trained}})$$

mean, variance, etc.

Neural Networks



Neural Networks



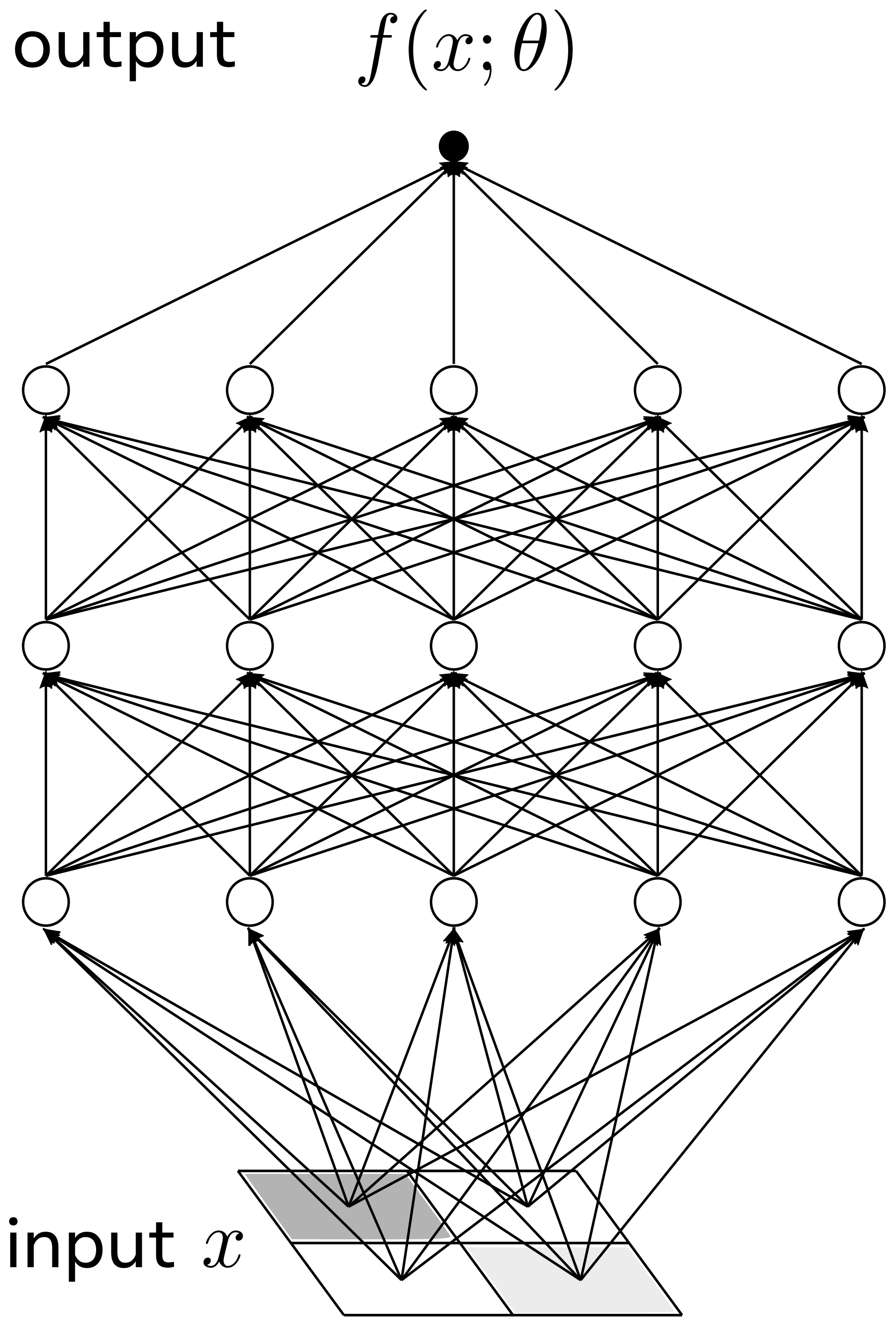
- Function:

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

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

$$f(x;\theta) = z^{(L)}(x)$$

Neural Networks



- Function:

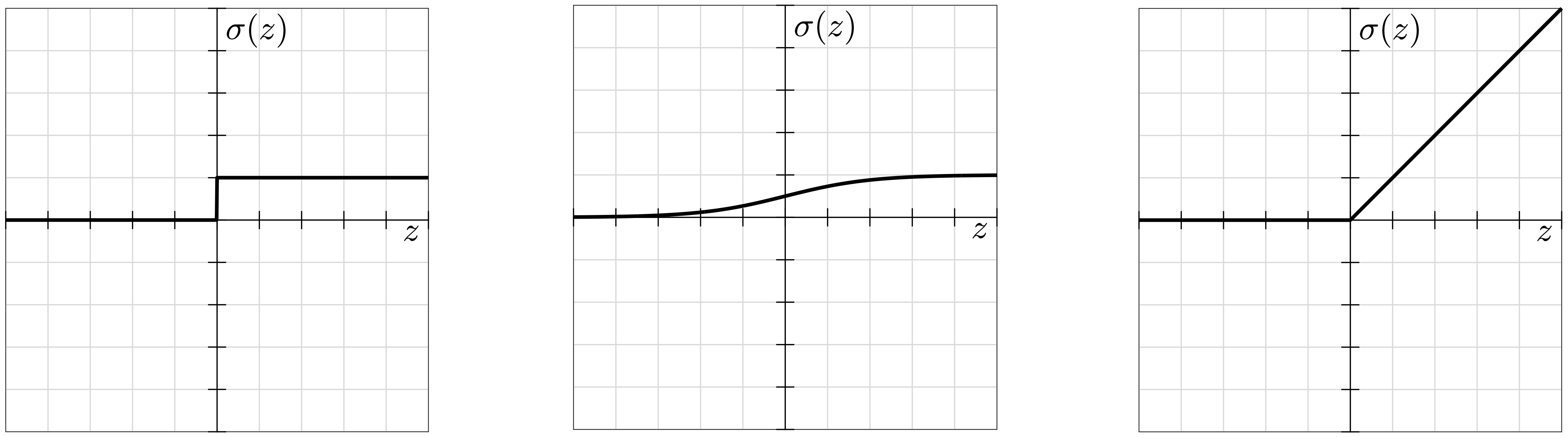
$$z_i^{(1)}(x) \equiv b_i^{(1)} + \sum_{j=1}^{n_0} W_{ij}^{(1)} x_j \quad \text{for} \quad i = 1, \dots, n_1 \,,$$

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

$$f(x;\theta) = z^{(L)}(x)$$

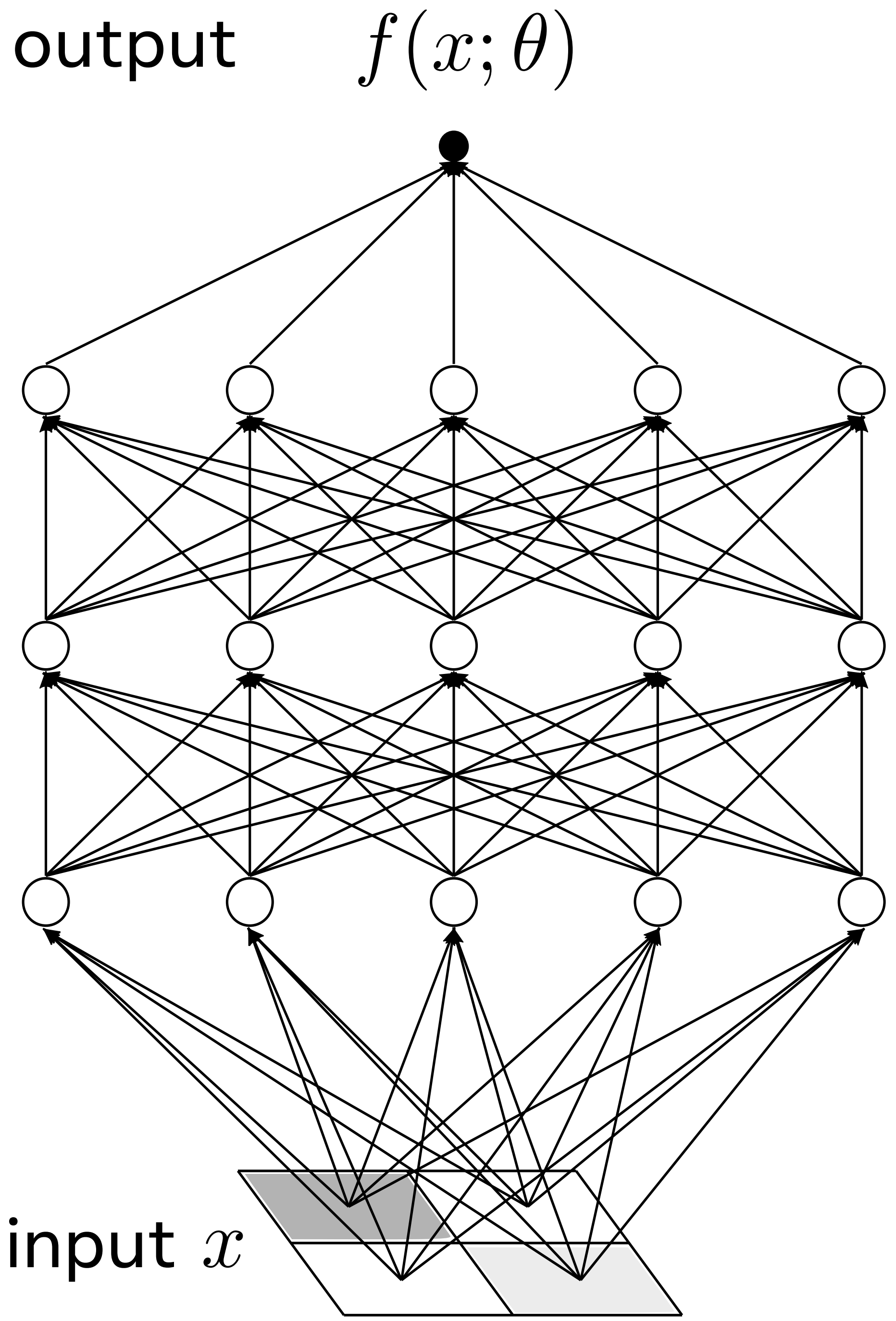
activation function

$$\sigma(z)$$



perceptron | sigmoid | ReLU

Neural Networks



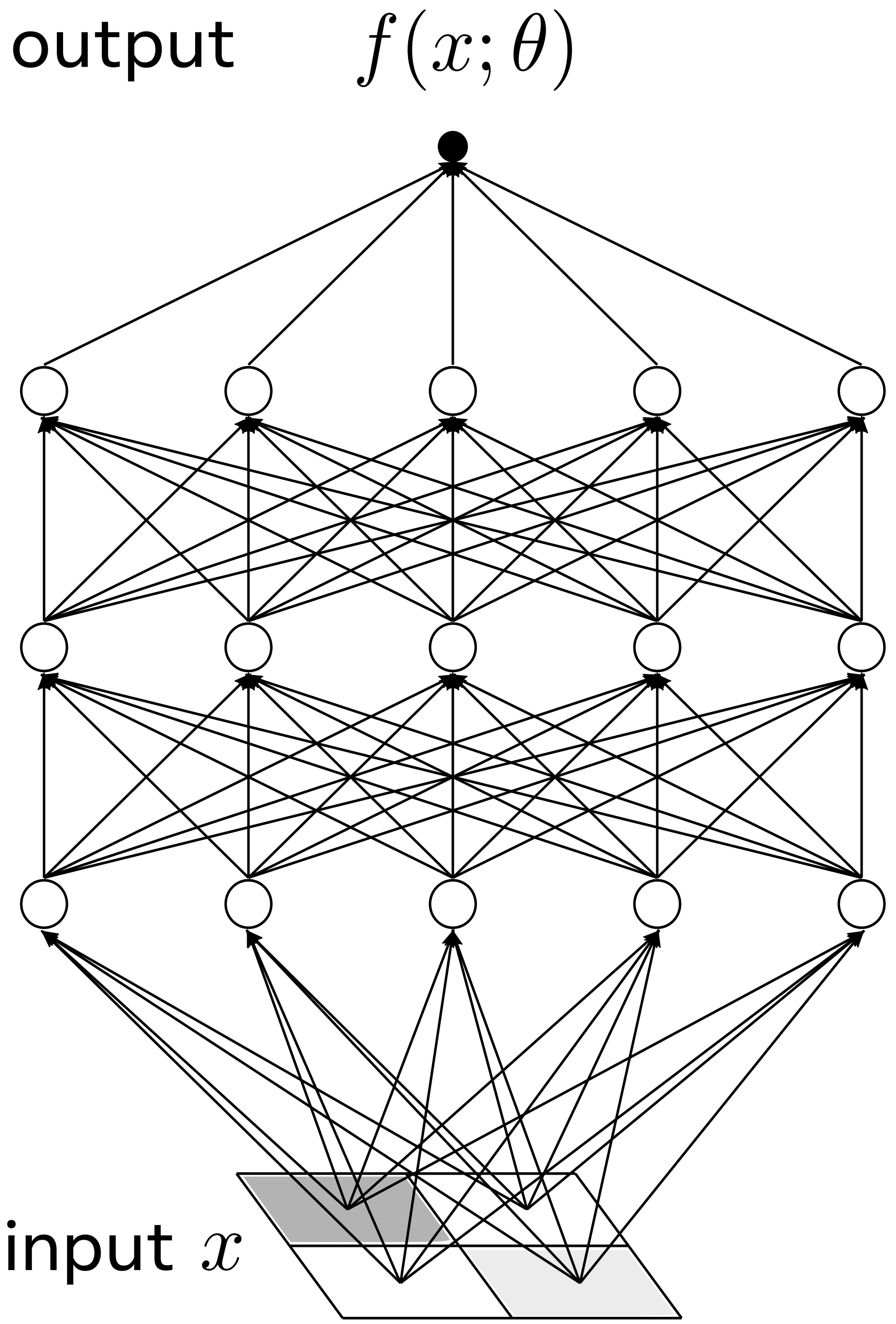
- Function:

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

$$z_i^{(\ell+1)}(x) \equiv b_i^{(\ell+1)} + \sum_{j=1}^{n_\ell} W_{ij}^{(\ell+1)} \sigma\Big(z_j^{(\ell)}(x)\Big) \quad \text{for} \quad i = 1, \ldots, n_{\ell+1}\,; \; \ell = 1, \ldots, L-1$$

$$f(x;\theta) = z^{(L)}(x)$$

Neural Networks



- Function:

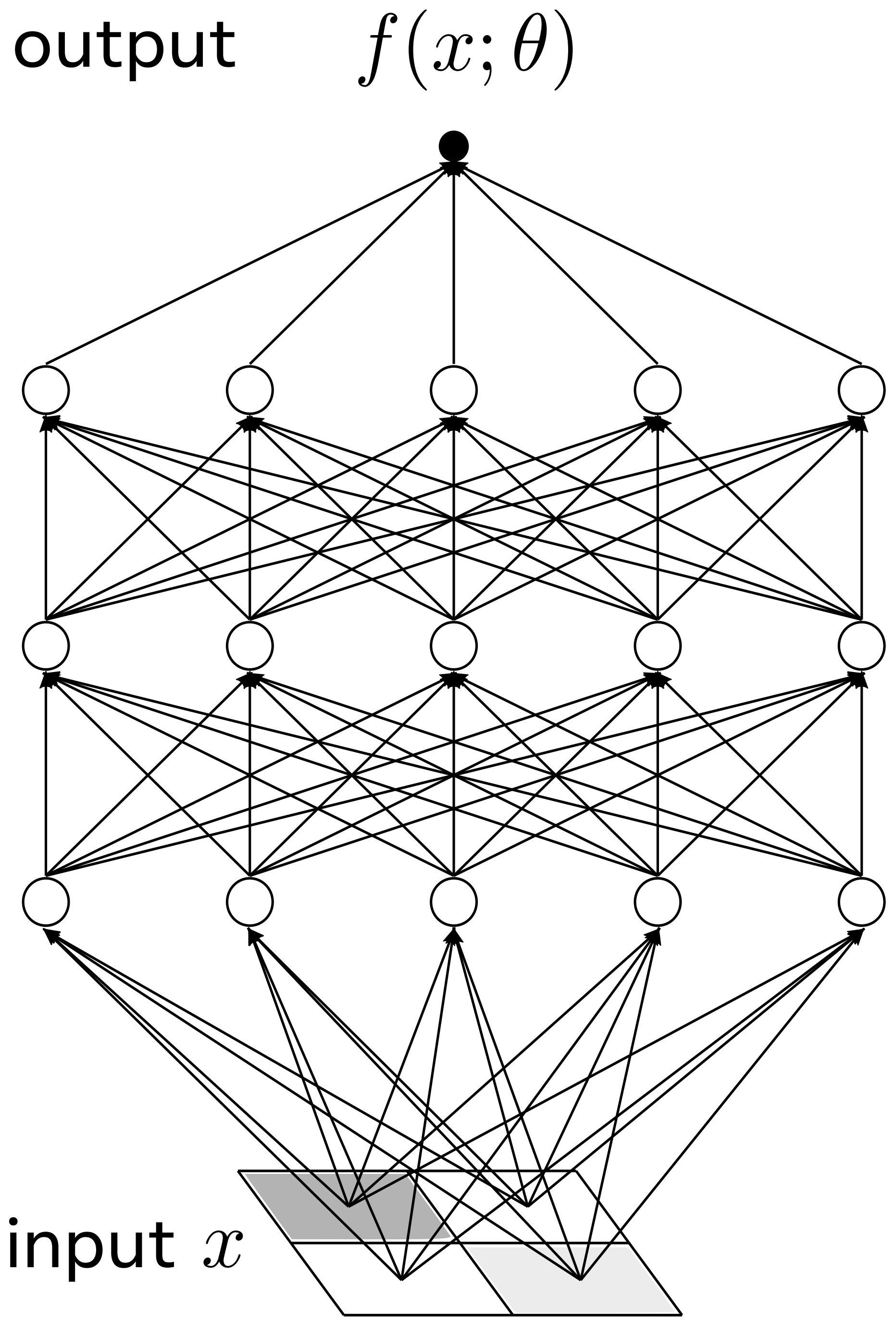
$$z_i^{(1)}(x) \equiv b_i^{(1)} + \sum_{j=1}^{n_0} W_{ij}^{(1)} x_j \quad \text{for} \quad i = 1, \ldots, n_1 \,,$$

$$z_i^{(\ell+1)}(x) \equiv b_i^{(\ell+1)} + \sum_{j=1}^{n_\ell} W_{ij}^{(\ell+1)} \sigma\Big(z_j^{(\ell)}(x)\Big) \quad \text{for} \quad i = 1, \ldots, n_{\ell+1}\,; \; \ell = 1, \ldots, L-1$$

$$f(x;\theta) = z^{(L)}(x)$$

- Model parameters: $\theta_{\mu=1,\ldots,P} = \Big\{ b_i^{(1)}, W_{ij}^{(1)}, b_i^{(2)}, W_{ij}^{(2)}, \ldots, b_i^{(L)}, W_{ij}^{(L)} \Big\}$

Neural Networks



- Function:

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

$$z_i^{(\ell+1)}(x) \equiv b_i^{(\ell+1)} + \sum_{j=1}^{n_\ell} W_{ij}^{(\ell+1)} \sigma\Big(z_j^{(\ell)}(x)\Big) \quad \text{for} \quad i = 1, \ldots, n_{\ell+1}\,; \; \ell = 1, \ldots, L-1$$

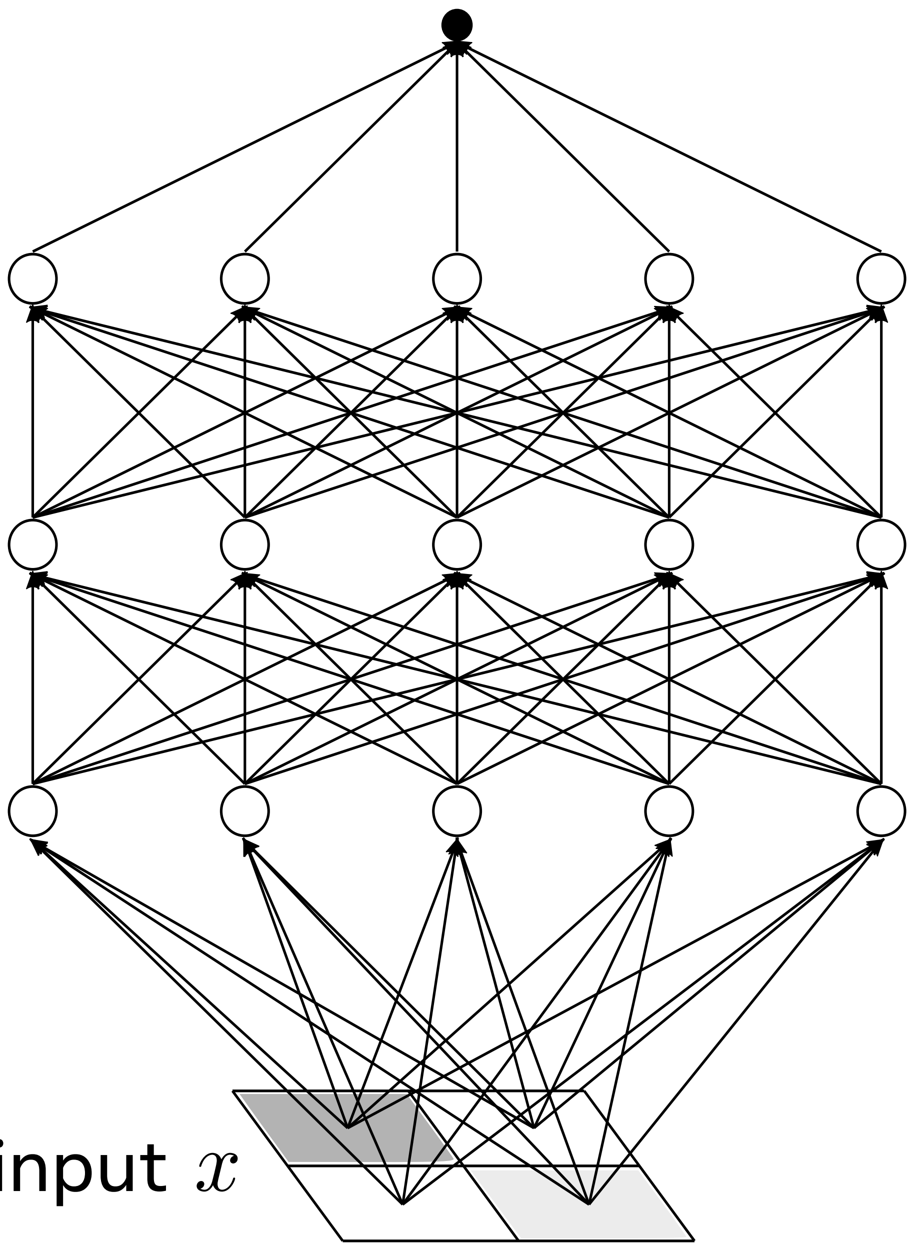
$$f(x;\theta) = z^{(L)}(x)$$

- Model parameters: $\theta_{\mu=1,\ldots,P} = \left\{ b_i^{(1)}, W_{ij}^{(1)}, b_i^{(2)}, W_{ij}^{(2)}, \ldots, b_i^{(L)}, W_{ij}^{(L)} \right\}$

- Initialization distribution: i.i.d. from mean-zero Gaussian with

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

Neural Networks



- Function:

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

$$z_i^{(\ell+1)}(x) \equiv b_i^{(\ell+1)} + \sum_{j=1}^{n_\ell} W_{ij}^{(\ell+1)} \sigma\Big(z_j^{(\ell)}(x)\Big) \quad \text{for} \quad i = 1, \dots, n_{\ell+1} \,; \; \ell = 1, \dots, L-1$$

$$f(x;\theta) = z^{(L)}(x)$$

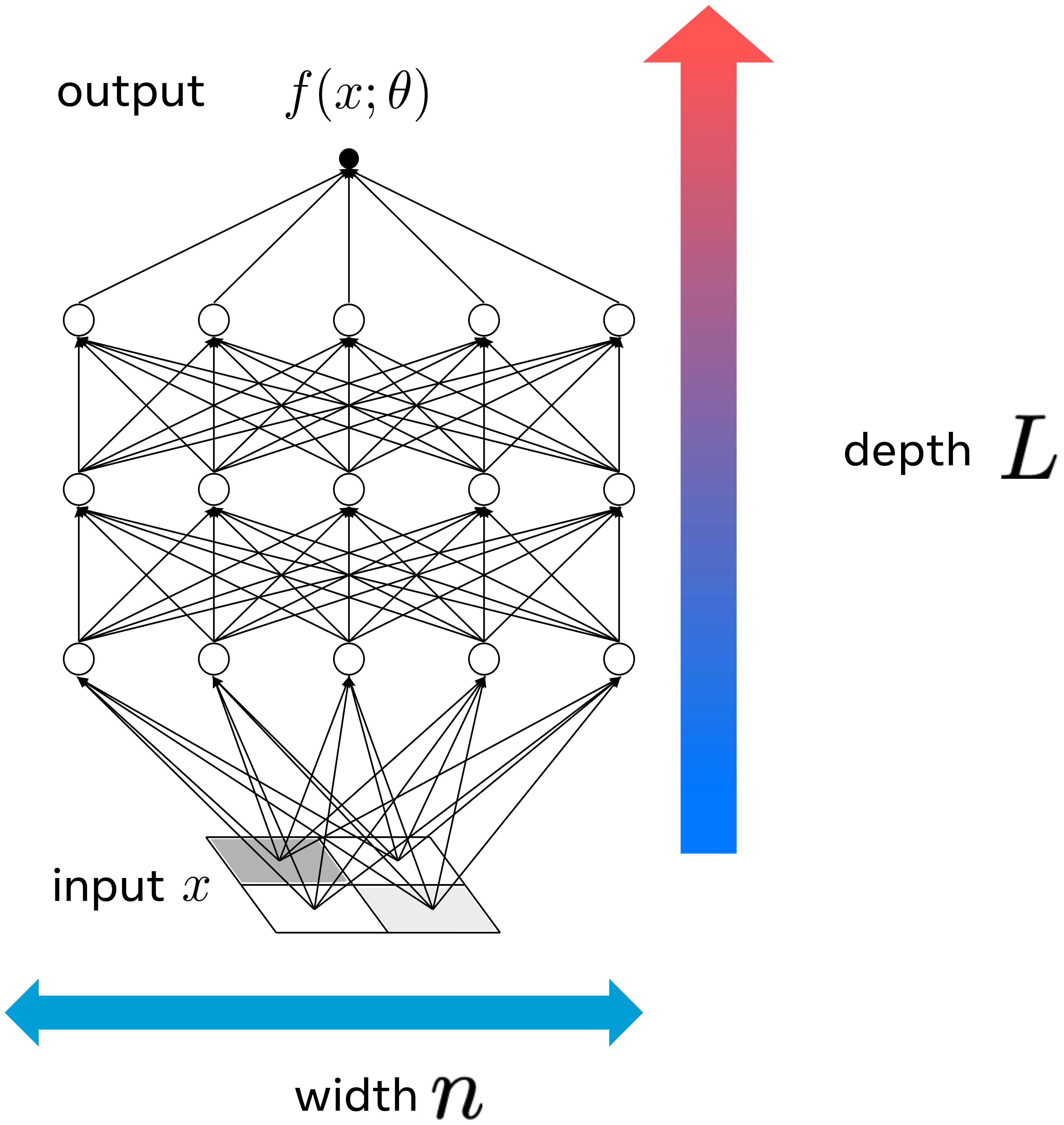
- Model parameters: $\theta_{\mu=1,\dots,P} = \left\{ b_i^{(1)}, W_{ij}^{(1)}, b_i^{(2)}, W_{ij}^{(2)}, \dots, b_i^{(L)}, W_{ij}^{(L)} \right\}$

- Initialization distribution: i.i.d. from mean-zero Gaussian with

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

good wide limit

Neural Networks



# of model parameters

$$P \sim n^2 L$$

Machine Learning in a Nutshell

- Instantiate a model

$$f_{\text{init}}(x) = f(x; \theta_{\text{init}}) \quad \text{with} \quad \underline{\theta_{\text{init}} \in p(\theta_{\text{init}})}$$

- Train the model, e.g., by gradient descent

$$\theta_\mu(t+1) = \theta_\mu(t) - \eta \frac{\partial \mathcal{L}}{\partial \theta_\mu}\Big|_{\theta=\theta(t)}$$

- Use the trained model to make predictions

$$\boxed{p(f_{\text{trained}})}$$

mean, variance, etc.

Problems 1, 2, & 3

Problems 1, 2, & 3

Trained function, Taylor-expanded around initialization:

$$f_{\text{trained}} = f_{\text{init}} + (\theta_{\text{trained}} - \theta_{\text{init}}) \frac{df}{d\theta}\Big|_{\text{init}} + \frac{1}{2}(\theta_{\text{trained}} - \theta_{\text{init}})^2 \frac{d^2 f}{d\theta^2}\Big|_{\text{init}} + \dots$$

Problems 1, 2, & 3

Trained function, Taylor-expanded around initialization:

$$f_{\text{trained}} = f_{\text{init}} + (\theta_{\text{trained}} - \theta_{\text{init}})\frac{df}{d\theta}\Big|_{\text{init}} + \frac{1}{2}(\theta_{\text{trained}} - \theta_{\text{init}})^2\frac{d^2 f}{d\theta^2}\Big|_{\text{init}} + \ldots$$

- Problem 1: too many terms in general

Problems 1, 2, & 3

Trained function, Taylor-expanded around initialization:

$$f_{\text{trained}} = f_{\text{init}} + (\theta_{\text{trained}} - \theta_{\text{init}}) \frac{df}{d\theta}\Big|_{\text{init}} + \frac{1}{2}(\theta_{\text{trained}} - \theta_{\text{init}})^2 \frac{d^2 f}{d\theta^2}\Big|_{\text{init}} + \ldots$$

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

$$p(\theta_{\text{init}}) \to p\left(\theta_{\text{init}}, f_{\text{init}}, \frac{df}{d\theta}\Big|_{\text{init}}, \frac{d^2 f}{d\theta^2}\Big|_{\text{init}}, \ldots\right)$$

statistics at *initialization*

Problems 1, 2, & 3

Trained function, Taylor-expanded around initialization:

$$f_{\text{trained}} = f_{\text{init}} + (\theta_{\text{trained}} - \theta_{\text{init}}) \frac{df}{d\theta}\Big|_{\text{init}} + \frac{1}{2}(\theta_{\text{trained}} - \theta_{\text{init}})^2 \frac{d^2 f}{d\theta^2}\Big|_{\text{init}} + \ldots$$

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

$$p(\theta_{\text{init}}) \to p\left(\theta_{\text{init}}, f_{\text{init}}, \frac{df}{d\theta}\Big|_{\text{init}}, \frac{d^2 f}{d\theta^2}\Big|_{\text{init}}, \ldots\right) \to \boxed{p(f_{\text{trained}})}$$

statistics at *initialization* statistics *after training*

Problems 1, 2, & 3

Trained function, Taylor-expanded around initialization:

$$f_{\text{trained}} = f_{\text{init}} + (\theta_{\text{trained}} - \theta_{\text{init}}) \frac{df}{d\theta}\Big|_{\text{init}} + \frac{1}{2}(\theta_{\text{trained}} - \theta_{\text{init}})^2 \frac{d^2 f}{d\theta^2}\Big|_{\text{init}} + \ldots$$

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

$$\theta_{\text{trained}} = [\theta_{\text{trained}}]\left(\theta_{\text{init}}, f_{\text{init}}, \frac{df}{d\theta}\Big|_{\text{init}}, \frac{d^2 f}{d\theta^2}\Big|_{\text{init}}, \ldots; \text{algorithm}; \text{data}\right)$$

$$p(\theta_{\text{init}}) \to p\left(\theta_{\text{init}}, f_{\text{init}}, \frac{df}{d\theta}\Big|_{\text{init}}, \frac{d^2 f}{d\theta^2}\Big|_{\text{init}}, \ldots\right) \to p(f_{\text{trained}})$$

statistics at *initialization* statistics *after training*

Despair & Hope

Despair & Hope

- Microscopic perspective (focusing on individuals):
  the more model parameters, the more complex. We are doomed…

Despair & Hope

- Microscopic perspective (focusing on individuals):
  the more model parameters, the more complex. We are doomed…

- Macroscopic perspective (focusing on averages):
  the more model parameters, the simpler. We can do this!

Despair & Hope

- Microscopic perspective (focusing on individuals):
  the more model parameters, the more complex. We are doomed…

- Macroscopic perspective (focusing on averages):
  the more model parameters, the simpler. We can do this!

Simplification when there are infinitely-many neurons in hidden layers
(a.k.a. **law of large numbers**; a *free theory* at $n = \infty$)
AND
systematically going beyond that idealized limit
(a.k.a. **perturbation theory**; a *weakly-interacting theory* when $n \gg L$)

Despair & Hope

- Microscopic perspective (focusing on individuals):
  the more model parameters, the more complex. We are doomed…

- Macroscopic perspective (focusing on averages):
  the more model parameters, the simpler. We can do this!

Simplification when there are infinitely-many neurons in hidden layers
(a.k.a. **law of large numbers**; a *free theory* at $n = \infty$)
AND
systematically going beyond that idealized limit
(a.k.a. **perturbation theory**; a *weakly-interacting theory* when $n \gg L$)

$$p(\theta_{\text{init}}) \to p\left(\theta_{\text{init}}, f_{\text{init}}, \frac{df}{d\theta}\bigg|_{\text{init}}, \frac{d^2 f}{d\theta^2}\bigg|_{\text{init}}, \ldots\right) \to p(f_{\text{trained}})$$

"Statistics become *sparse* & dynamics can be truncated"

2. Neural Networks at Large Width

Training Dynamics at Infinite Width

Gradient descent: $$\theta_\mu(t+1) = \theta_\mu(t) - \eta \frac{\partial \mathcal{L}}{\partial \theta_\mu}\Big|_{\theta=\theta(t)}$$

Training Dynamics at Infinite Width

Gradient descent: $$\theta_\mu(t+1) = \theta_\mu(t) - \eta \frac{\partial \mathcal{L}}{\partial \theta_\mu}\Big|_{\theta=\theta(t)}$$

Function evolution: $$f(t+1) = f(t) - \eta H(t) \frac{\partial \mathcal{L}}{\partial f} + \cancel{O\left(\frac{1}{n}\right)}$$

with Neural Tangent Kernel (NTK) $$H \sim \sum_\mu \left(\frac{\partial f}{\partial \theta_\mu}\right)^2$$

Training Dynamics at Infinite Width

Gradient descent: $$\theta_\mu(t+1) = \theta_\mu(t) - \eta \frac{\partial \mathcal{L}}{\partial \theta_\mu}\Big|_{\theta=\theta(t)}$$

Function evolution: $$f(t+1) = f(t) - \eta H(t) \frac{\partial \mathcal{L}}{\partial f} + O\left(\frac{1}{n}\right)$$

with Neural Tangent Kernel (NTK) $$H \sim \sum_\mu \left(\frac{\partial f}{\partial \theta_\mu}\right)^2$$

NTK evolution: $$H(t+1) = H(t) + O\left(\frac{1}{n}\right)$$

Training Dynamics at Infinite Width

Gradient descent:

$$\theta_\mu(t+1) = \theta_\mu(t) - \eta \frac{\partial \mathcal{L}}{\partial \theta_\mu}\bigg|_{\theta=\theta(t)}$$

Function evolution:

$$f(t+1) = f(t) - \eta H(t) \frac{\partial \mathcal{L}}{\partial f} + \cancel{O\left(\frac{1}{n}\right)}$$

with Neural Tangent Kernel (NTK) $H \sim \sum_\mu \left(\frac{\partial f}{\partial \theta_\mu}\right)^2$

NTK evolution:

$$H(t+1) = H(t) + \cancel{O\left(\frac{1}{n}\right)}$$

$\longrightarrow$ $H(t) = H_{\text{init}}$ “frozen” NTK

$\longrightarrow$ $f(t+1) = f(t) - \eta H_{\text{init}} \frac{\partial \mathcal{L}}{\partial f}$

Solving "Problem 3" (Dynamics) at Infinite Width

$$f(t+1) = f(t) - \eta H_{\text{init}} \frac{\partial \mathcal{L}}{\partial f}$$

Solving "Problem 3" (Dynamics) at Infinite Width

E.g., for $\mathcal{L} = \frac{1}{2}(f - y)^2$

$$f(t+1) = f(t) - \eta H_{\text{init}}[f(t) - y]$$

Solving “Problem 3” (Dynamics) at Infinite Width

E.g., for $\mathcal{L} = \frac{1}{2}(f - y)^2$

$$f(t+1) = f(t) - \eta H_{\text{init}}[f(t) - y]$$

(exponentially)

$$f_{\text{trained}} = f_{\text{init}} - \text{“}H_{\text{init}} * H_{\text{init}}^{-1}\text{”}[f_{\text{init}} - y]$$

Solving "Problem 3" (Dynamics) at Infinite Width

$$f_{\text{trained}} = f_{\text{init}} - \text{``}H_{\text{init}} * H_{\text{init}}^{-1}\text{''}[f_{\text{init}} - y]$$

$$p\left(\theta_{\text{init}}\right) \rightarrow p\left(f_{\text{init}}, H_{\text{init}}\right) \rightarrow p(f_{\text{trained}})$$

Solving "Problems 1 & 2" (Statistics) at Infinite Width

$$p\left(\theta_{\text{init}}\right) \rightarrow p\left(f_{\text{init}}, H_{\text{init}}\right) \rightarrow p(f_{\text{trained}})$$

Solutions to “Problems 1 & 2” (Statistics) at Infinite Width

- Gaussian distribution [R. Neal (1996), J. Lee+Y. Bahri et al. (ICLR 2018), A. Matthews et al. (ICLR2018)]

$$p\left(f_{\text{init}}\right) \propto \exp\left(-\frac{1}{2}\text{“}f_{\text{init}} K^{-1} f_{\text{init}}\text{”}\right)$$

K some (calculable) matrix

- Deterministic NTK [A. Jacot, F. Gabriel, & C. Hongler (NeurIPS 2018)]

$$p\left(H_{\text{init}}\right) = \delta\left(H_{\text{init}} - \Theta\right)$$

Θ some (calculable) matrix

$$p\left(\theta_{\text{init}}\right) \to p\left(f_{\text{init}}, H_{\text{init}}\right) \to p(f_{\text{trained}})$$

Solutions to “Problems 1 & 2” (Statistics) at Infinite Width

- Gaussian distribution [R. Neal (1996), J. Lee+Y. Bahri et al. (ICLR 2018), A. Matthews et al. (ICLR2018)]

$$p\left(f_{\text{init}}\right) \propto \exp\left(-\frac{1}{2} \text{“} f_{\text{init}} K^{-1} f_{\text{init}} \text{”}\right)$$

K some (calculable) matrix

- Deterministic NTK [A. Jacot, F. Gabriel, & C. Hongler (NeurIPS 2018)]

$$p\left(H_{\text{init}}\right) = \delta\left(H_{\text{init}} - \Theta\right)$$

Θ some (calculable) matrix

$$p\left(\theta_{\text{init}}\right) \rightarrow p\left(f_{\text{init}}, H_{\text{init}}\right) \rightarrow p(f_{\text{trained}})$$

These “calculations” involve RG flow (next section)

Training Dynamics at Finite Width

$$f(t+1) = f(t) - \eta H(t) \frac{\partial \mathcal{L}}{\partial f} + \frac{\eta^2}{2} \mathrm{d}H(t) \left(\frac{\partial \mathcal{L}}{\partial f}\right)^2 + \frac{\eta^3}{6} \mathrm{dd}H(t) \left(\frac{\partial \mathcal{L}}{\partial f}\right)^3 + O\left(\frac{1}{n^2}\right)$$

$O(1/n)$

$$f_{\text{trained}} = f_{\text{init}} - \text{``}H_{\text{init}} * H_{\text{init}}^{-1}\text{''} [f_{\text{init}} - y] + \text{despicable}(y, f_{\text{init}}, H_{\text{init}}, \mathrm{d}H_{\text{init}}, \mathrm{dd}H_{\text{init}}; \text{algorithm})$$

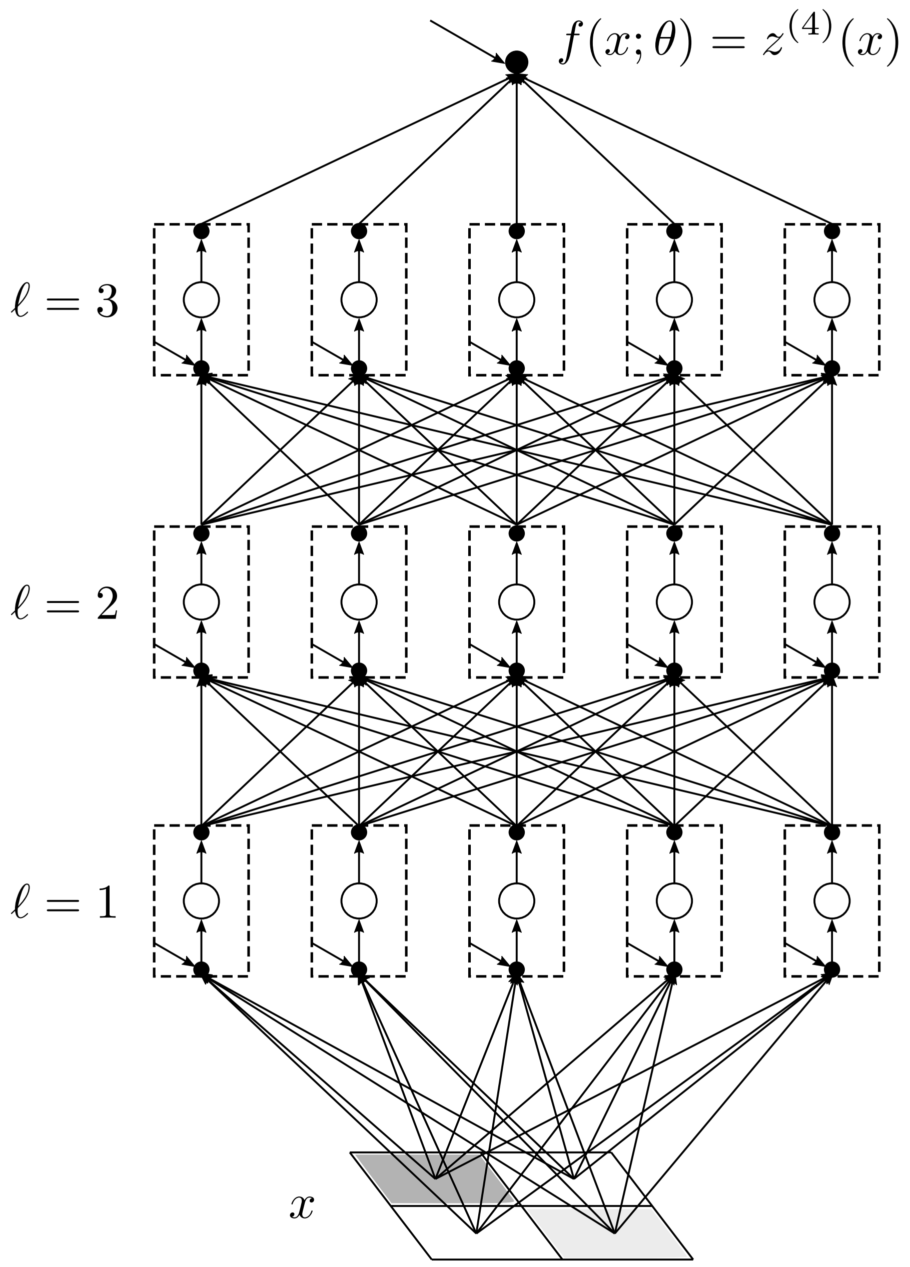
$$p(\theta_{\text{init}}) \to p(f_{\text{init}}, H_{\text{init}}, \mathrm{d}H_{\text{init}}, \mathrm{dd}H_{\text{init}}) \to p(f_{\text{trained}})$$

Statistics at Finite Width

Nearly-Gaussian [§4, §8, § 11.2, & §∞.3 of arXiv:2106.10165]

$$p\left(\theta_{\text{init}}\right) \rightarrow p\left(f_{\text{init}}, H_{\text{init}}, \mathrm{d}H_{\text{init}}, \mathrm{dd}H_{\text{init}}\right) \rightarrow p(f_{\text{trained}})$$

Statistics at Finite Width



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

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

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

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

$$p\left(\theta_{\mathrm{init}}\right) \to p\left(f_{\mathrm{init}}, H_{\mathrm{init}}, \mathrm{d}H_{\mathrm{init}}, \mathrm{dd}H_{\mathrm{init}}\right) \to p\left(f_{\mathrm{trained}}\right)$$

3a. RG flow

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

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

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

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

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

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

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

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

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

$$\cdots$$

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

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

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

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

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

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

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

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

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

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

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

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 Other Stuffs

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:

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

NTK mean:

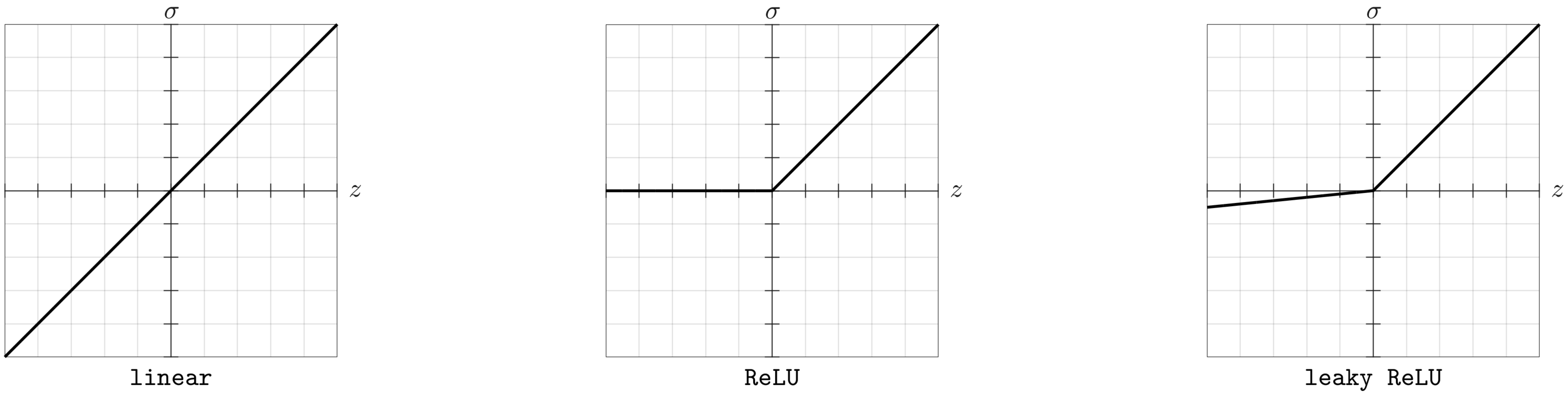
$$H^{(\ell+1)} = \lambda_b + \lambda_W \langle \sigma(z)\sigma(z) \rangle_{G^{(\ell)}} + C_W H^{(\ell)} \langle \sigma'(z)\sigma'(z) \rangle_{G^{(\ell)}} + O\left(\frac{1}{n}\right)$$

...

3b. Criticality

E.g., Scale-Invariant Activation Functions

$$\sigma(z) = \begin{cases} a_+ z\,, & z \geq 0\,, \\ a_- z\,, & z < 0\,. \end{cases}$$



$a_+ = 1,\ a_- = 1$ $\quad$ $a_+ = 1,\ a_- = 0$ $\quad$ $a_+ = 1,\ a_- = 0.1$

Kernel Recursion

$$\sigma(z) = \begin{cases} a_+ z\,, & z \geq 0\,, \\ a_- z\,, & z < 0\,. \end{cases}$$

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

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

Kernel Recursion

$$\sigma(z) = \begin{cases} a_+ z\,, & z \geq 0\,, \\ a_- z\,, & z < 0\,. \end{cases}$$

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

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

$$\left\langle \sigma(z)\sigma(z) \right\rangle_K \equiv \frac{1}{\sqrt{2\pi K}} \int_{-\infty}^{\infty} dz\ \sigma(z)\sigma(z) e^{-\frac{z^2}{2K}}$$

Kernel Recursion

$$\sigma(z) = \begin{cases} a_+ z\,, & z \geq 0\,, \\ a_- z\,, & z < 0\,. \end{cases}$$

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

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

$$\begin{aligned} \langle \sigma(z)\sigma(z) \rangle_K &\equiv \frac{1}{\sqrt{2\pi K}} \int_{-\infty}^{\infty} dz\ \sigma(z)\sigma(z) e^{-\frac{z^2}{2K}} \\ &= \frac{1}{\sqrt{2\pi K}} \left[a_+^2 \int_0^{\infty} dz\ z^2 e^{-\frac{z^2}{2K}} + a_-^2 \int_{-\infty}^{0} dz\ z^2 e^{-\frac{z^2}{2K}} \right] \end{aligned}$$

Kernel Recursion

$$\sigma(z) = \begin{cases} a_+ z\,, & z \geq 0\,, \\ a_- z\,, & z < 0\,. \end{cases}$$

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

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

$$\begin{aligned}
\left\langle \sigma(z)\sigma(z) \right\rangle_K &\equiv \frac{1}{\sqrt{2\pi K}} \int_{-\infty}^{\infty} dz\ \sigma(z)\sigma(z) e^{-\frac{z^2}{2K}} \\
&= \frac{1}{\sqrt{2\pi K}} \left[a_+^2 \int_0^{\infty} dz\ z^2 e^{-\frac{z^2}{2K}} + a_-^2 \int_{-\infty}^{0} dz\ z^2 e^{-\frac{z^2}{2K}} \right] \\
&= \frac{1}{\sqrt{2\pi K}} \left[\frac{a_+^2}{2} \int_{-\infty}^{\infty} dz\ z^2 e^{-\frac{z^2}{2K}} + \frac{a_-^2}{2} \int_{-\infty}^{\infty} dz\ z^2 e^{-\frac{z^2}{2K}} \right]
\end{aligned}$$

Kernel Recursion

$$\sigma(z) = \begin{cases} a_+ z\,, & z \geq 0\,, \\ a_- z\,, & z < 0\,. \end{cases}$$

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

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

$$\begin{aligned}
\langle \sigma(z)\sigma(z) \rangle_K &\equiv \frac{1}{\sqrt{2\pi K}} \int_{-\infty}^{\infty} dz\ \sigma(z)\sigma(z) e^{-\frac{z^2}{2K}} \\
&= \frac{1}{\sqrt{2\pi K}} \left[a_+^2 \int_0^{\infty} dz\ z^2 e^{-\frac{z^2}{2K}} + a_-^2 \int_{-\infty}^{0} dz\ z^2 e^{-\frac{z^2}{2K}} \right] \\
&= \frac{1}{\sqrt{2\pi K}} \left[\frac{a_+^2}{2} \int_{-\infty}^{\infty} dz\ z^2 e^{-\frac{z^2}{2K}} + \frac{a_-^2}{2} \int_{-\infty}^{\infty} dz\ z^2 e^{-\frac{z^2}{2K}} \right] \\
&= \left(\frac{a_+^2 + a_-^2}{2} \right) \frac{1}{\sqrt{2\pi K}} \int_{-\infty}^{\infty} dz\ z^2 e^{-\frac{z^2}{2K}}
\end{aligned}$$

Kernel Recursion

$$\sigma(z) = \begin{cases} a_+ z\,, & z \geq 0\,, \\ a_- z\,, & z < 0\,. \end{cases}$$

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

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

$$\begin{aligned}
\left\langle \sigma(z)\sigma(z) \right\rangle_K &\equiv \frac{1}{\sqrt{2\pi K}} \int_{-\infty}^{\infty} dz\ \sigma(z)\sigma(z) e^{-\frac{z^2}{2K}} \\
&= \frac{1}{\sqrt{2\pi K}} \left[a_+^2 \int_0^{\infty} dz\ z^2 e^{-\frac{z^2}{2K}} + a_-^2 \int_{-\infty}^{0} dz\ z^2 e^{-\frac{z^2}{2K}} \right] \\
&= \frac{1}{\sqrt{2\pi K}} \left[\frac{a_+^2}{2} \int_{-\infty}^{\infty} dz\ z^2 e^{-\frac{z^2}{2K}} + \frac{a_-^2}{2} \int_{-\infty}^{\infty} dz\ z^2 e^{-\frac{z^2}{2K}} \right] \\
&= \left(\frac{a_+^2 + a_-^2}{2} \right) \frac{1}{\sqrt{2\pi K}} \int_{-\infty}^{\infty} dz\ z^2 e^{-\frac{z^2}{2K}} \\
&= \left(\frac{a_+^2 + a_-^2}{2} \right) K
\end{aligned}$$

Kernel Recursion

$$\sigma(z) = \begin{cases} a_+ z\,, & z \geq 0\,, \\ a_- z\,, & z < 0\,. \end{cases}$$

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

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

$$\left\langle \sigma(z)\sigma(z) \right\rangle_K = A_2 K \quad \text{with} \quad A_2 \equiv \frac{a_+^2 + a_-^2}{2}$$

Kernel Recursion

$$\sigma(z) = \begin{cases} a_+ z\,, & z \geq 0\,, \\ a_- z\,, & z < 0\,. \end{cases}$$

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

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

$$\left\langle \sigma(z)\sigma(z) \right\rangle_K = A_2 K \quad \text{with} \quad A_2 \equiv \frac{a_+^2 + a_-^2}{2}$$

$$K^{(\ell+1)} = C_b + C_W A_2 K^{(\ell)}$$

Kernel Recursion

$$\sigma(z) = \begin{cases} a_+ z\,, & z \geq 0\,, \\ a_- z\,, & z < 0\,. \end{cases}$$

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

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

Let me simplify further $C_b = 0,\ \chi \equiv C_W A_2$

$$K^{(\ell+1)} = \cancel{C_b} + \underbrace{C_W A_2}_{\equiv\, \chi} K^{(\ell)}$$

Kernel Recursion

$$\sigma(z) = \begin{cases} a_+ z\,, & z \geq 0\,, \\ a_- z\,, & z < 0\,. \end{cases}$$

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

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

Let me simplify further $C_b = 0,\ \chi \equiv C_W A_2$

$$K^{(\ell+1)} = \chi K^{(\ell)}$$

Kernel Recursion

$$\sigma(z) = \begin{cases} a_+ z\,, & z \geq 0\,, \\ a_- z\,, & z < 0\,. \end{cases}$$

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

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

Let me simplify further $C_b = 0,\ \chi \equiv C_W A_2$

$$K^{(\ell+1)} = \chi K^{(\ell)}$$

$$\longrightarrow \quad K^{(\ell)} = \chi^{\ell-1} K^{(1)}$$

$$K^{(1)} = \cancel{C_b} + C_W \left(\frac{1}{n_0} \sum_{j=1}^{n_0} x_j^2 \right)$$

Kernel Recursion

$$\sigma(z) = \begin{cases} a_+ z\,, & z \geq 0\,, \\ a_- z\,, & z < 0\,. \end{cases}$$

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

$$K^{(\ell)} = \chi^{\ell-1} K^{(1)}$$

$$\chi \equiv C_W A_2 = C_W \left(\frac{a_+^2 + a_-^2}{2}\right)$$

$$K^{(1)} = C_b + C_W \left(\frac{1}{n_0} \sum_{j=1}^{n_0} x_j^2 \right)$$

Kernel Recursion

$$\sigma(z) = \begin{cases} a_+ z\,, & z \geq 0\,, \\ a_- z\,, & z < 0\,. \end{cases}$$

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

$$K^{(\ell)} = \chi^{\ell-1} K^{(1)}$$

$$\chi \equiv C_W A_2 = C_W \left(\frac{a_+^2 + a_-^2}{2}\right)$$

$$K^{(1)} = \cancel{C_b} + C_W \left(\frac{1}{n_0} \sum_{j=1}^{n_0} x_j^2 \right)$$

- $\chi > 1$: exploding signal

Kernel Recursion

$$\sigma(z) = \begin{cases} a_+ z\,, & z \geq 0\,, \\ a_- z\,, & z < 0\,. \end{cases}$$

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

$$K^{(\ell)} = \chi^{\ell-1} K^{(1)}$$

$$\chi \equiv C_W A_2 = C_W \left(\frac{a_+^2 + a_-^2}{2}\right)$$

$$K^{(1)} = \cancel{C_b} + C_W \left(\frac{1}{n_0} \sum_{j=1}^{n_0} x_j^2 \right)$$

- $\chi > 1$: exploding signal
- $\chi < 1$: vanishing signal

Kernel Recursion

$$\sigma(z) = \begin{cases} a_+ z\,, & z \geq 0\,, \\ a_- z\,, & z < 0\,. \end{cases}$$

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

$$K^{(\ell)} = \chi^{\ell-1} K^{(1)}$$

$$\chi \equiv C_W A_2 = C_W \left(\frac{a_+^2 + a_-^2}{2} \right)$$

$$K^{(1)} = \cancel{C_b} + C_W \left(\frac{1}{n_0} \sum_{j=1}^{n_0} x_j^2 \right)$$

- $\chi > 1$: exploding signal
- $\chi < 1$: vanishing signal
- $\chi = 1$: critical signal propagation

$$K^{(\ell)} = K^{(1)} = \text{constant} = K^*$$

Kernel Recursion

$$\sigma(z) = \begin{cases} a_+ z\,, & z \geq 0\,, \\ a_- z\,, & z < 0\,. \end{cases}$$

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

$$K^{(\ell)} = \chi^{\ell-1} K^{(1)}$$

$$\chi \equiv C_W A_2 = C_W \left(\frac{a_+^2 + a_-^2}{2} \right)$$

$$K^{(1)} = \cancel{C_b} + C_W \left(\frac{1}{n_0} \sum_{j=1}^{n_0} x_j^2 \right)$$

- $\chi > 1$: exploding signal
- $\chi < 1$: vanishing signal
- $\chi = 1$: critical signal propagation @

$$\boxed{C_W = \frac{1}{A_2} = \frac{2}{a_+^2 + a_-^2}}$$

Kaiming init. for $\texttt{ReLU}$

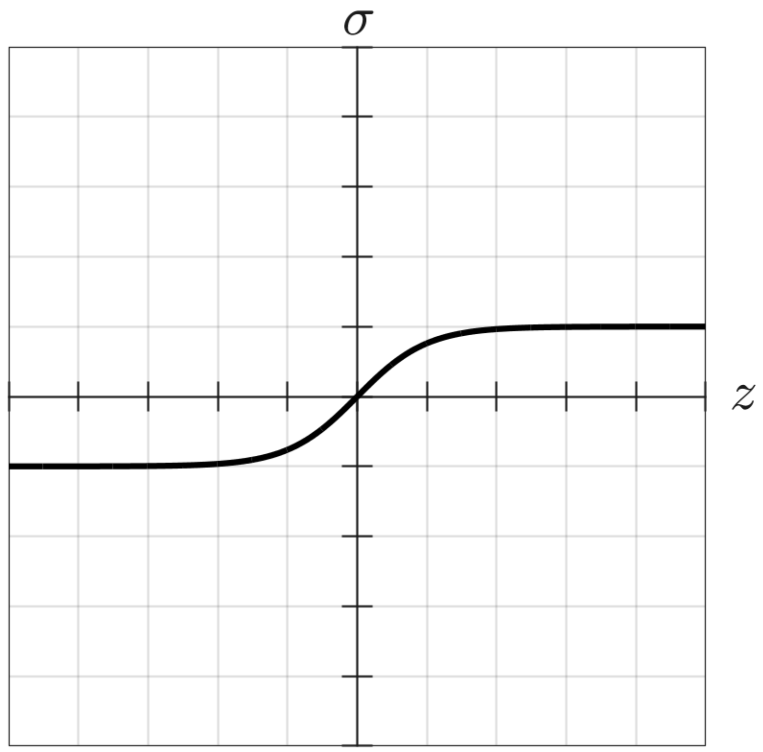
$$K^{(\ell)} = K^{(1)} = \text{constant} = K^*$$

Scale-Invariant Universality Class

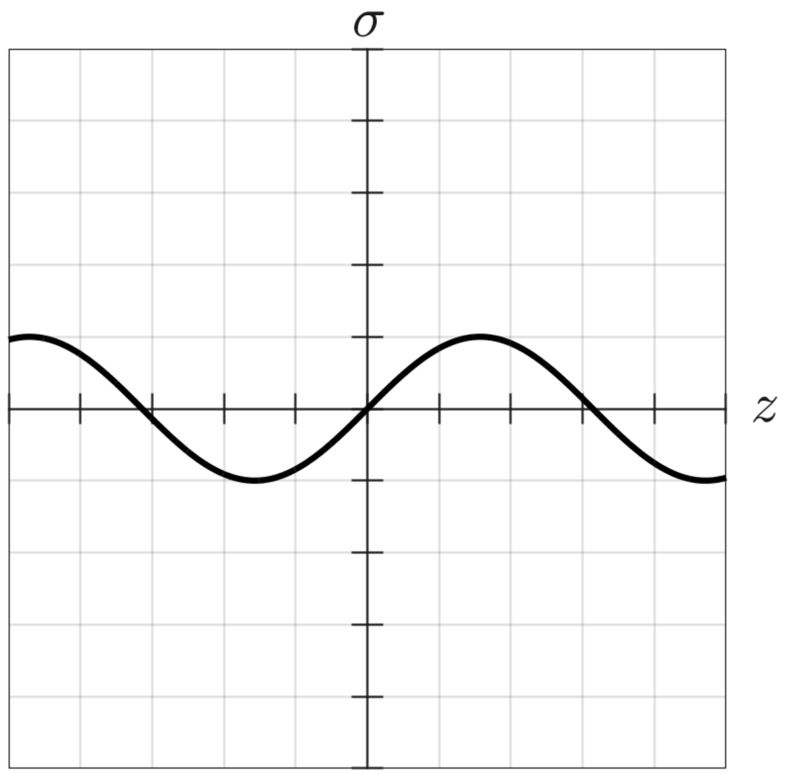
$$\sigma(z) = \begin{cases} a_+ z\,, & z \geq 0\,, \\ a_- z\,, & z < 0\,. \end{cases}$$

Aside from differences in order-one coefficients,
they all behave similarly when networks become deep.

$K^* = 0$ Universality Class: $\texttt{tanh}, \texttt{sin}, \ldots.$



tanh

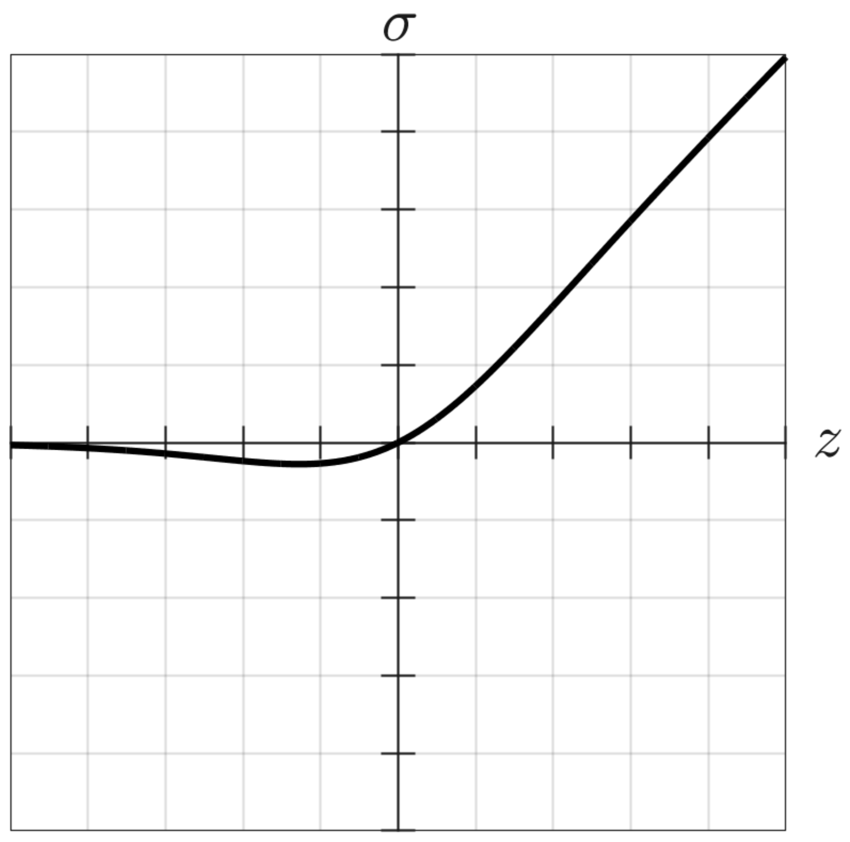


sin

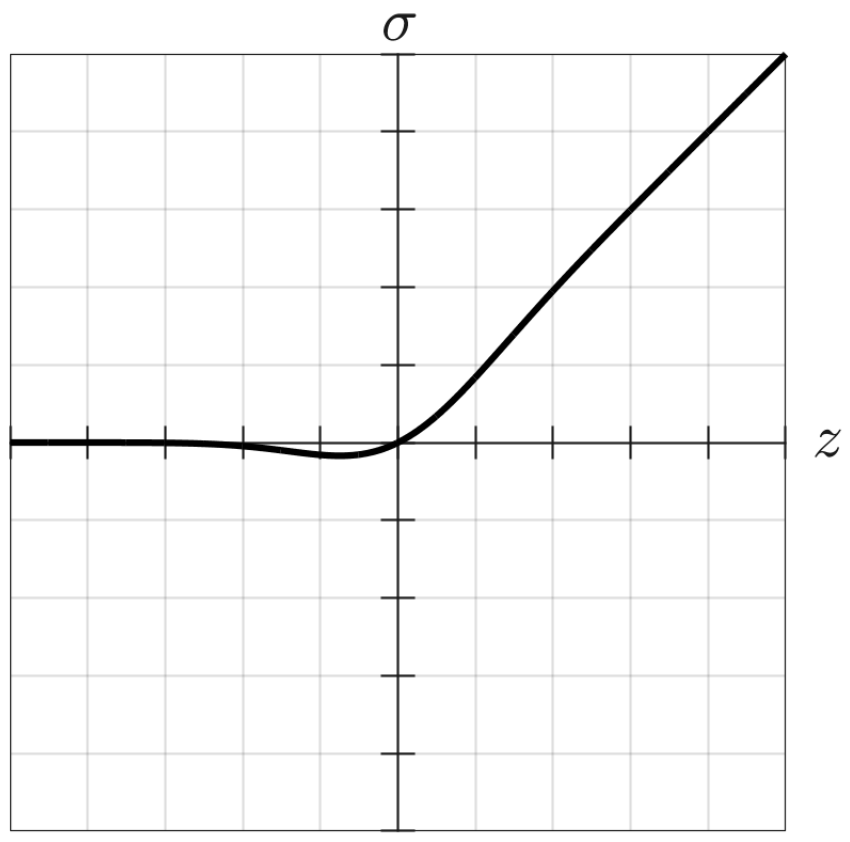
$$(C_b, C_W)^{\text{critical}} = (0, 1)$$

at which we have a fixed point $K^* = 0$ with $\chi_{\parallel}(K^*) = \chi_{\perp}(K^*) = 1$

Half-Stable Universality Class: GELU, SWISH, ...



SWISH

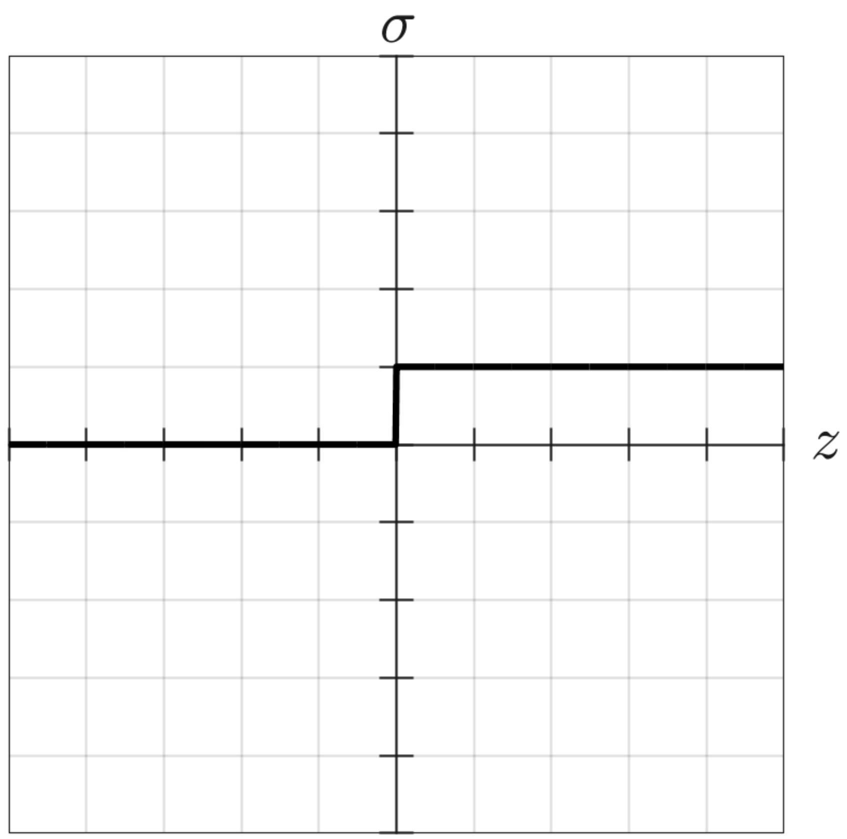


GELU

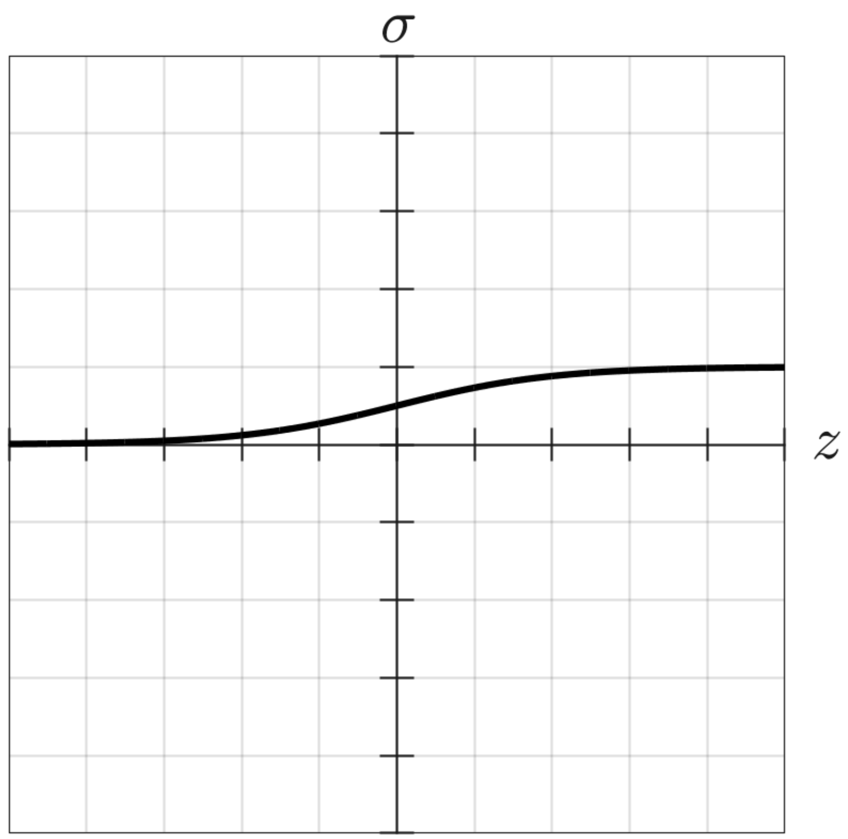
$$(C_b, C_W)^{\text{critical}} \approx (0.55514317, 1.98800468)$$

$$(C_b, C_W)^{\text{critical}} \approx (0.17292239, 1.98305826)$$

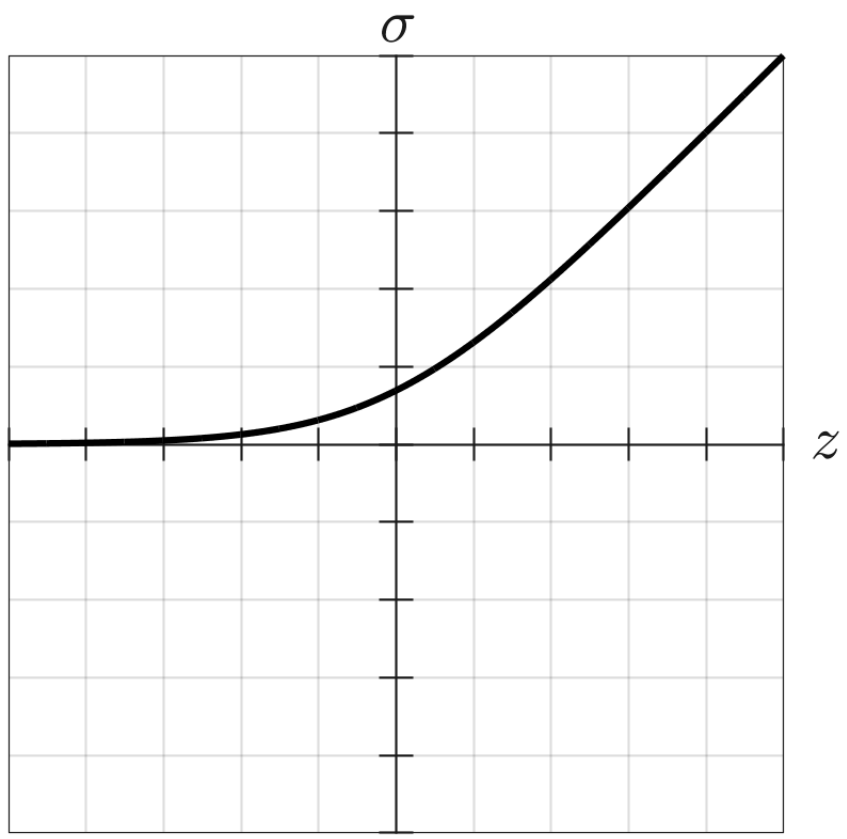
No Criticality, No Deep Learning: `perceptron`, `sigmoid`, `softplus`, ...



`perceptron`



`sigmoid`



`softplus`

$$\chi_{\parallel}(K^*) = \chi_{\perp}(K^*) = 1$$ unsatisfiable

Never again for deep learning

Four-Point Recursion

$$\text{finite-width effects} \propto \frac{\text{depth}}{\text{width}}$$

Two Endnotes

Preceding discussions assume the neural-tangent scaling of various hyperparameters;
for more general cases including the maximal-update scaling, see:
arXiv:2210.04909 (S. Yaida, "Meta-Principled Family of Hyperparameter Scaling Strategies")
interpolates the neural-tangent scaling @ $s = 0$ and the mean-field/maximal-update scaling @ $s = 1$

In order to keep representation-learning ability,

we should keep $\gamma = \frac{L}{n^{1-s}}$ fixed in scaling up the model

- depth~width for the neural-tangent scaling strategy
  (γ is a coupling constant)

- depth=fixed for the mean-field/maximal-update scaling strategy
  (intrinsically strongly-coupled)

Two Endnotes

[arXiv:2304.02034](https://arxiv.org/abs/2304.02034)
(Emily Dinan, S. Yaida, Susan Zhang, “Effective Theory of Transformers at Initialization”)