Lecture 4:
The Principle of Criticality

[§5, §9, §11.3, and §∞.1 (+ §10.3) of
"The Principles of Deep Learning Theory (PDLT),"
arXiv:2106.10165]

The Principle of Sparsity for WIDE Neural Networks

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

statistics at *initialization* statistics *after training*

The Principle of Sparsity for WIDE Neural Networks

Dynamical Dan $z^\star\left(\hat{z}, \hat{H}, \widehat{\mathrm{d}H}, \ldots\right)$

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

statistics at *initialization* statistics *after training*

The Principle of Sparsity for WIDE Neural Networks

Statistical Sho

Dynamical Dan $z^\star\left(\widehat{z}, \widehat{H}, \widehat{\mathrm{d}H}, \ldots\right)$

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

statistics at *initialization*

statistics *after training*

The Principle of Sparsity for WIDE Neural Networks

Statistical Sho

Dynamical Dan $z^\star\left(\hat{z}, \hat{H}, \widehat{\mathrm{d}H}, \ldots\right)$

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

statistics at *initialization*

statistics *after training*

- Infinite width:

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

The Principle of Sparsity for WIDE Neural Networks

Statistical Sho

Dynamical Dan $z^\star\left(\hat{z}, \hat{H}, \widehat{\mathrm{d}H}, \ldots\right)$

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

statistics at *initialization* statistics *after training*

- Infinite width:

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

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

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

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

The Principle of Sparsity for WIDE Neural Networks

Statistical Sho

Dynamical Dan $z^\star\left(\hat{z}, \hat{H}, \widehat{\mathrm{d}H}, \ldots\right)$

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

statistics at *initialization* statistics *after training*

- Infinite width:

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

[*different notion from "sparsity" as in pruned networks]

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

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

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

The Principle of Sparsity for WIDE Neural Networks

Statistical Sho

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

statistics at *initialization* statistics *after training*

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

The Principle of Criticality for DEEP Neural Networks

Statistical Sho

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

statistics at *initialization* statistics *after training*

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

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

The Principle of Criticality for DEEP Neural Networks

Statistical Sho

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

statistics at *initialization* statistics *after training*

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

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

Some Practical Lessons:

ways to avoid exploding/vanishing gradient problems for deep neural networks

Review of Notations and Conventions

Initialization hyperparameters:

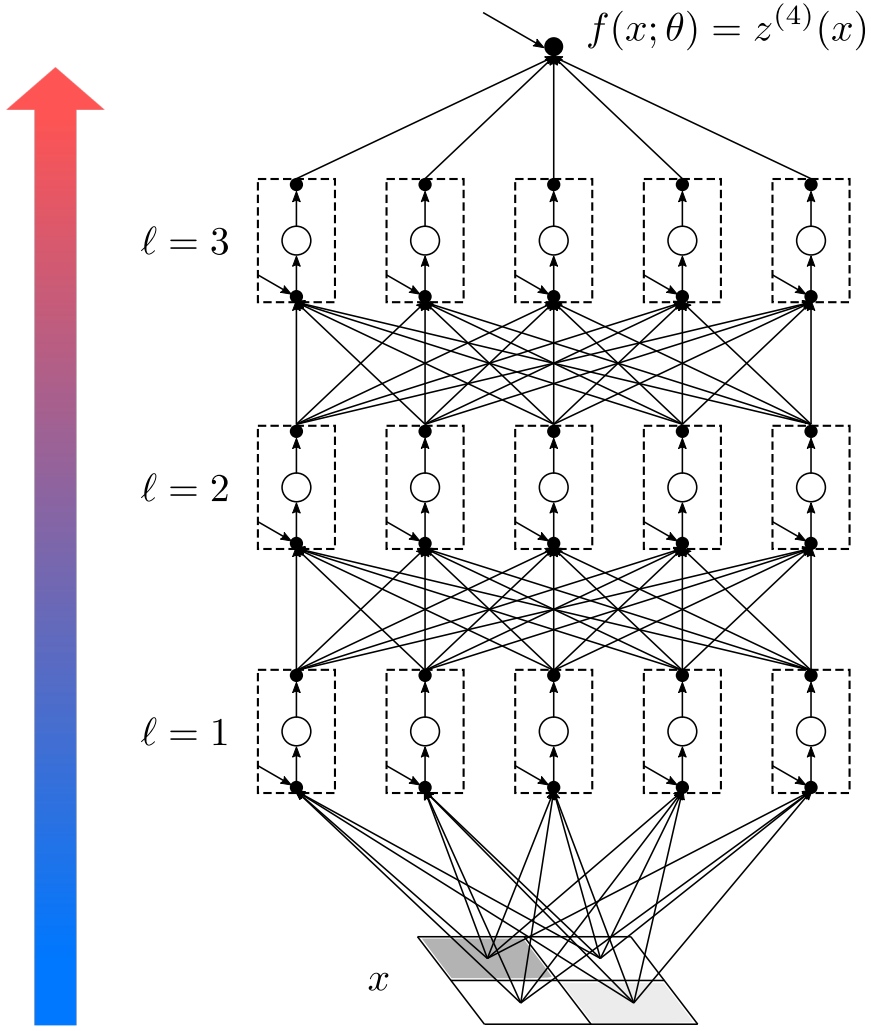
$$\mathbb{E}\left[b_{i_1}^{(\ell)} b_{i_2}^{(\ell)}\right] = \delta_{i_1 i_2} C_b \,, \quad \mathbb{E}\left[W_{i_1 j_1}^{(\ell)} W_{i_2 j_2}^{(\ell)}\right] = \delta_{i_1 i_2}\delta_{j_1 j_2}\frac{C_W}{n_{\ell-1}}$$

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

$$\lambda_{b_{i_1}^{(\ell)} b_{i_2}^{(\ell)}} = \delta_{i_1 i_2}\lambda_b \,, \quad \lambda_{W_{i_1 j_1}^{(\ell)} W_{i_2 j_2}^{(\ell)}} = \delta_{i_1 i_2}\delta_{j_1 j_2}\frac{\lambda_W}{n_{\ell-1}}$$

Two pedagogical simplifications: (i) a single input; (ii) layer-independent hyperparameters

Review of Strategy



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

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

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

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

Review of Some Recursions

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

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

Review of Some Recursions

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

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

$$\mathbb{E}[\widehat{H}_{i_1 i_2}^{(\ell)}] = \delta_{i_1 i_2} H^{(\ell)}$$

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

Review of Some Recursions

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

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

$$\mathbb{E}[\widehat{H}_{i_1 i_2}^{(\ell)}] = \delta_{i_1 i_2} H^{(\ell)}$$

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

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

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

One more thing...

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

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

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

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

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

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

Many more things...

$$D^{(\ell+1)} = \chi_\perp^{(\ell)}\chi_\parallel^{(\ell)} D^{(\ell)} + \left(\frac{\lambda_W^{(\ell+1)}}{C_W}\right)\left[C_W^2\left\langle\sigma(z)\sigma(z)\sigma(z)\sigma(z)\right\rangle_{K^{(\ell)}} - \left(C_W g^{(\ell)}\right)^2 + \left(\chi_\parallel^{(\ell)}\right)^2 V^{(\ell)}\right]$$
$$+ \Theta^{(\ell)}\left[C_W^2\left\langle\sigma(z)\sigma(z)\sigma'(z)\sigma'(z)\right\rangle_{K^{(\ell)}} - C_W g^{(\ell)}\chi_\perp^{(\ell)} + 2h^{(\ell)}\chi_\parallel^{(\ell)} V^{(\ell)}\right], \quad (9.13)$$

$$F^{(\ell+1)} = \left(\chi_\parallel^{(\ell)}\right)^2 F^{(\ell)} + C_W^2\left\langle\sigma(z)\sigma(z)\sigma'(z)\sigma'(z)\right\rangle_{K^{(\ell)}}\Theta^{(\ell)}, \quad (9.14)$$

$$B^{(\ell+1)} = \left(\chi_\perp^{(\ell)}\right)^2 B^{(\ell)} + C_W^2\left\langle\sigma'(z)\sigma'(z)\sigma'(z)\sigma'(z)\right\rangle_{K^{(\ell)}}\left(\Theta^{(\ell)}\right)^2, \quad (9.15)$$

$$A^{(\ell+1)} = \left(\chi_\perp^{(\ell)}\right)^2 A^{(\ell)} + \left(\frac{\lambda_W^{(\ell+1)}}{C_W}\right)^2\left[C_W^2\left\langle\sigma(z)\sigma(z)\sigma(z)\sigma(z)\right\rangle_{K^{(\ell)}} - \left(C_W g^{(\ell)}\right)^2 + \left(\chi_\parallel^{(\ell)}\right)^2 V^{(\ell)}\right]$$
$$+ 2\left(\frac{\lambda_W^{(\ell+1)}}{C_W}\right)\Theta^{(\ell)}\left[C_W^2\left\langle\sigma(z)\sigma(z)\sigma'(z)\sigma'(z)\right\rangle_{K^{(\ell)}} - C_W g^{(\ell)}\chi_\perp^{(\ell)} + 2h^{(\ell)}\chi_\parallel^{(\ell)} V^{(\ell)}\right]$$
$$+ 2\left(\frac{\lambda_W^{(\ell+1)}}{C_W}\right)\chi_\perp^{(\ell)}\chi_\parallel^{(\ell)} D^{(\ell)} + 4h^{(\ell)}\chi_\perp^{(\ell)}\Theta^{(\ell)} D^{(\ell)}$$
$$+ \left(\Theta^{(\ell)}\right)^2\left[C_W^2\left\langle\sigma'(z)\sigma'(z)\sigma'(z)\sigma'(z)\right\rangle_{K^{(\ell)}} - \left(\chi_\perp^{(\ell)}\right)^2 + \left(2h^{(\ell)}\right)^2 V^{(\ell)}\right]. \quad (9.16)$$

$$R^{(\ell+1)} = \left(\chi_\perp^{(\ell)}\right)^2 R^{(\ell)} \quad (\infty.14)$$
$$+ \lambda_W^{(\ell+1)} C_W\left\langle\sigma''\sigma'\sigma'\sigma\right\rangle_{K^{(\ell)}}\left(\Theta^{(\ell)}\right)^2 + C_W^2\left\langle\sigma'''\sigma'\sigma'\sigma'\right\rangle_{K^{(\ell)}}\left(\Theta^{(\ell)}\right)^3$$
$$+ \chi_\perp^{(\ell)}\left(\lambda_W^{(\ell+1)}\left\langle\sigma''\sigma\right\rangle_{K^{(\ell)}} + C_W\Theta^{(\ell)}\left\langle\sigma'''\sigma'\right\rangle_{K^{(\ell)}}\right)\left(B^{(\ell)} + P^{(\ell)}\right)$$
$$+ \chi_\perp^{(\ell)}\left(\lambda_W^{(\ell+1)}\left\langle\sigma'\sigma'\right\rangle_{K^{(\ell)}} + C_W\Theta^{(\ell)}\left\langle\sigma''\sigma''\right\rangle_{K^{(\ell)}}\right)P^{(\ell)},$$

$$S^{(\ell+1)} = \left(\chi_\perp^{(\ell)}\right)^2 S^{(\ell)} \quad (\infty.15)$$
$$+ C_W\lambda_W^{(\ell+1)}\left\langle\sigma'\sigma'\sigma'\sigma'\right\rangle_{K^{(\ell)}}\left(\Theta^{(\ell)}\right)^2 + C_W^2\left\langle\sigma''\sigma''\sigma'\sigma'\right\rangle_{K^{(\ell)}}\left(\Theta^{(\ell)}\right)^3$$
$$+ \chi_\perp^{(\ell)}\left[\lambda_W^{(\ell+1)}\left\langle\sigma'\sigma'\right\rangle_{K^{(\ell)}} + C_W\Theta^{(\ell)}\left\langle\sigma''\sigma''\right\rangle_{K^{(\ell)}}\right]B^{(\ell)},$$

$$T^{(\ell+1)} = \left(\chi_\perp^{(\ell)}\right)^2 T^{(\ell)} \quad (\infty.16)$$
$$+ 2C_W\lambda_W^{(\ell+1)}\left\langle\sigma''\sigma'\sigma'\sigma\right\rangle_{K^{(\ell)}}\left(\Theta^{(\ell)}\right)^2 + C_W^2\left\langle\sigma''\sigma''\sigma'\sigma'\right\rangle_{K^{(\ell)}}\left(\Theta^{(\ell)}\right)^3$$
$$+ \left(\lambda_W^{(\ell+1)}\right)^2\left\langle\sigma'\sigma'\sigma\sigma\right\rangle_{K^{(\ell)}}\Theta^{(\ell)}$$
$$+ \left(\lambda_W^{(\ell+1)}\left\langle z\sigma'\sigma\right\rangle_{K^{(\ell)}} + C_W\Theta^{(\ell)}\left\langle z\sigma''\sigma'\right\rangle_{K^{(\ell)}}\right)^2\frac{F^{(\ell)}}{\left(K^{(\ell)}\right)^2}$$
$$+ 2\chi_\perp^{(\ell)}\left[\lambda_W^{(\ell+1)}\left(\left\langle\sigma''\sigma\right\rangle_{K^{(\ell)}} + \left\langle\sigma'\sigma'\right\rangle_{K^{(\ell)}}\right) + C_W\Theta^{(\ell)}\left(\left\langle\sigma'''\sigma'\right\rangle_{K^{(\ell)}} + \left\langle\sigma''\sigma''\right\rangle_{K^{(\ell)}}\right)\right]Q^{(\ell)},$$

$$U^{(\ell+1)} = \left(\chi_\perp^{(\ell)}\right)^2 U^{(\ell)} + C_W^2\left\langle\sigma''\sigma''\sigma'\sigma'\right\rangle_{K^{(\ell)}}\left(\Theta^{(\ell)}\right)^3. \quad (\infty.17)$$

$$P^{(\ell+1)} = C_W^2\left\langle\sigma''\sigma'\sigma'\sigma\right\rangle_{K^{(\ell)}}\left(\Theta^{(\ell)}\right)^2 + C_W\chi_\perp^{(\ell)}\left\langle\sigma''\sigma\right\rangle_{K^{(\ell)}}B^{(\ell)}$$
$$+ \left[C_W\chi_\perp^{(\ell)}\left\langle\sigma''\sigma\right\rangle_{K^{(\ell)}} + \left(\chi_\perp^{(\ell)}\right)^2\right]P^{(\ell)}, \quad (11.55)$$

$$Q^{(\ell+1)} = C_W^2\left\langle\sigma''\sigma'\sigma'\sigma\right\rangle_{K^{(\ell)}}\left(\Theta^{(\ell)}\right)^2 + \frac{\lambda_W^{(\ell+1)}}{C_W}F^{(\ell+1)} + 2h^{(\ell)}\chi_\parallel^{(\ell)}\Theta^{(\ell)}F^{(\ell)}$$
$$+ \left[C_W\chi_\perp^{(\ell)}\left\langle\sigma''\sigma\right\rangle_{K^{(\ell)}} + \left(\chi_\perp^{(\ell)}\right)^2\right]Q^{(\ell)}, \quad (11.56)$$

Outline

1. Scale-Invariant Activation Functions: `linear`, `ReLU`, `leaky ReLU`, ...

2. More Generally
   - $K^* = 0$ universality class: `tanh`, `sin`, ...
   - Half-stable universality class: `GELU`, `SWISH`, ...
   - No criticality, no deep learning: `perceptron`, `sigmoid`, `softplus`, ...

3. Finite-Width Effects and *Deep* Learning

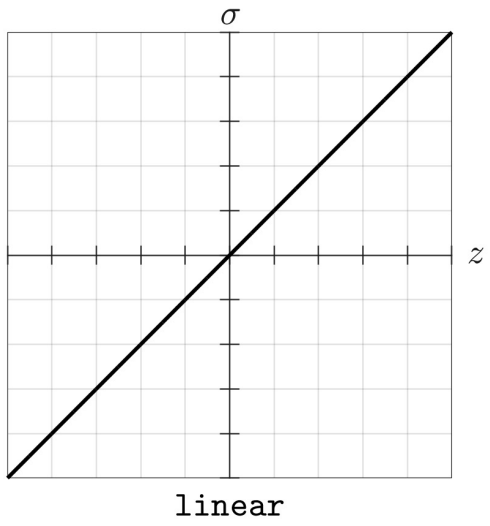
4. More on *Why* Criticality?

1. Scale-Invariant Activation Functions

`linear`, `ReLU`, `leaky ReLU`, ...

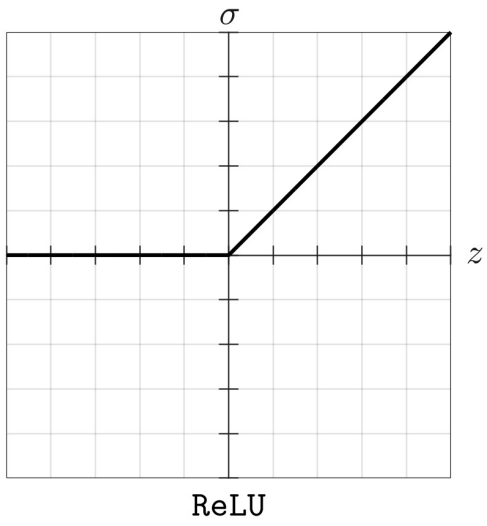
Scale-Invariant Activation Functions

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



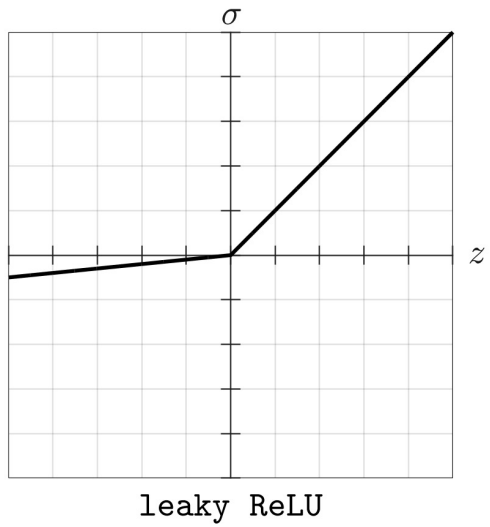
linear

$a_+ = 1,\ a_- = 1$



ReLU

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



leaky ReLU

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

Kernel Recursion

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

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

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

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

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

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

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

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

Kernel Recursion

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

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

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

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

Kernel Recursion

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

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

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

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

Kernel Recursion

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

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

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

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

Kernel Recursion

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

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

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

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

Kernel Recursion

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

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

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

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

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

Kernel Recursion

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

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

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

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

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

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

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

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

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

Kernel Recursion

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

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

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

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

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

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

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

Kernel Recursion

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

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

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

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

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

Kernel Recursion

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

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

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

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

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

- $\chi > 1$: exploding signal

Kernel Recursion

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

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

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

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

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

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

Kernel Recursion

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

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

$$K^{(\ell)} = \chi^{\ell-1} K^{(1)} \qquad \chi \equiv C_W A_2 = C_W \left(\tfrac{a_+^2 + a_-^2}{2} \right) \qquad K^{(1)} = C_b + C_W \left(\frac{1}{n_0} \sum_{j=1}^{n_0} x_j^2 \right)$$

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

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

Kernel Recursion

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

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

$$K^{(\ell)} = \chi^{\ell-1} K^{(1)} \qquad \chi \equiv C_W A_2 = C_W \left(\tfrac{a_+^2 + a_-^2}{2}\right) \qquad K^{(1)} = \cancel{C_b} + C_W \left(\frac{1}{n_0} \sum_{j=1}^{n_0} x_j^2 \right)$$

- $\chi > 1$: exploding signal
- $\chi < 1$: vanishing signal
- $\chi = 1$: critical signal propagation @ $C_W = \frac{1}{A_2} = \frac{2}{a_+^2 + a_-^2}$ Kaiming init. for ReLU

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

Kernel Recursion

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

[Exercise: $C_b \neq 0$ case]

NTK Mean Recursion

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

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

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

NTK Mean Recursion

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

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

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

Some integrals as before: $\left\langle \sigma(z)\sigma(z) \right\rangle_K = A_2 K\,, \quad \left\langle \sigma'(z)\sigma'(z) \right\rangle_K = A_2$

NTK Mean Recursion

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

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

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

Some integrals as before: $\left\langle \sigma(z)\sigma(z) \right\rangle_K = A_2 K\,, \quad \left\langle \sigma'(z)\sigma'(z) \right\rangle_K = A_2$

So we have: $\Theta^{(\ell+1)} = \lambda_b + \lambda_W A_2 K^{(\ell)} + \underbrace{C_W A_2 \Theta^{(\ell)}}_{= \chi}$

NTK Mean Recursion

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

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

$$\Theta^{(\ell+1)} = \lambda_b + \lambda_W A_2 K^{(\ell)} + \chi \Theta^{(\ell)}$$

NTK Mean Recursion

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

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

$$\Theta^{(\ell+1)} = \lambda_b + \lambda_W A_2 K^{(\ell)} + \chi \Theta^{(\ell)}$$

- $\chi > 1$: exploding *gradient*

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

NTK Mean Recursion

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

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

$$\Theta^{(\ell+1)} = \lambda_b + \lambda_W A_2 K^{(\ell)} + \chi \Theta^{(\ell)}$$

- $\chi > 1$: exploding *gradient*
- $\chi < 1$: vanishing *gradient* (for lower layers)

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

NTK Mean Recursion

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

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

$$\Theta^{(\ell+1)} = \lambda_b + \lambda_W A_2 K^{(\ell)} + \chi \Theta^{(\ell)}$$

- $\chi > 1$: exploding *gradient*
- $\chi < 1$: vanishing *gradient* (for lower layers)
- $\chi = 1$: critical *gradient* propagation @

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

$$K^{(\ell)} = K^* = \frac{1}{A_2} \left(\frac{1}{n_0} \sum_{i=1}^{n_0} x_i^2 \right)$$

$$\Theta^{(\ell)} = (\lambda_b + \lambda_W A_2 K^*) \times \ell$$

NTK Mean Recursion

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

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

$$\Theta^{(\ell)} = (\lambda_b + \lambda_W A_2 K^*) \times \ell$$

$$z^{(L)}_{i;\delta}(t+1) = z^{(L)}_{i;\delta}(t) - \sum \eta H^{(L)}_{ij;\delta\tilde{\alpha}}(t) \frac{\partial \mathcal{L}}{\partial z_{j;\tilde{\alpha}}} + \dots$$

Learning rate for deep networks:

$$\eta \lambda_{b,W} \sim \frac{1}{L}$$

Four-Point Recursion

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

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

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

Four-Point Recursion

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

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

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

After various integrals...:

$$\frac{1}{n_\ell} V^{(\ell+1)} = \frac{1}{n_\ell} C_W^2 (3A_4 - A_2^2)\left(K^{(\ell)}\right)^2 + \frac{1}{n_{\ell-1}} V^{(\ell)} \chi^2$$

$$A_2 \equiv \frac{a_+^2 + a_-^2}{2} \qquad A_4 \equiv \frac{a_+^4 + a_-^4}{2}$$

Four-Point Recursion

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

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

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

After various integrals...:

$$\frac{1}{n_\ell} V^{(\ell+1)} = \frac{1}{n_\ell} C_W^2 (3A_4 - A_2^2)\left(K^{(\ell)}\right)^2 + \frac{1}{n_{\ell-1}} V^{(\ell)} \chi^2$$

$$A_2 \equiv \frac{a_+^2 + a_-^2}{2} \qquad A_4 \equiv \frac{a_+^4 + a_-^4}{2}$$

At criticality $C_b = 0,\ C_W = 1/A_2$ **:**

$$\frac{1}{n_\ell} V^{(\ell+1)} = \frac{1}{n_\ell}\left(\frac{3A_4 - A_2^2}{A_2^2}\right)(K^*)^2 + \frac{1}{n_{\ell-1}} V^{(\ell)}$$

Four-Point Recursion

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

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

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

After various integrals...:

$$\frac{1}{n_\ell} V^{(\ell+1)} = \frac{1}{n_\ell} C_W^2 (3A_4 - A_2^2)\left(K^{(\ell)}\right)^2 + \frac{1}{n_{\ell-1}} V^{(\ell)} \chi^2 \qquad A_2 \equiv \frac{a_+^2 + a_-^2}{2} \qquad A_4 \equiv \frac{a_+^4 + a_-^4}{2}$$

At criticality $C_b = 0,\ C_W = 1/A_2$ **:**

$$\frac{1}{n_\ell} V^{(\ell+1)} = \frac{1}{n_\ell}\left(\frac{3A_4 - A_2^2}{A_2^2}\right)(K^*)^2 + \frac{1}{n_{\ell-1}} V^{(\ell)}$$

Solving that:

$$\frac{1}{n_{\ell-1}} V^{(\ell)} = \left(\sum_{\ell'=1}^{\ell-1} \frac{1}{n_{\ell'}}\right)\left[\left(\frac{3A_4 - A_2^2}{A_2^2}\right)(K^*)^2\right] = O\left(\frac{\ell}{n}\right)$$

Four-Point Recursion

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

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

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

At criticality $C_b = 0,\ C_W = 1/A_2$ **:**

$$\frac{1}{n_{\ell-1}} V^{(\ell)} = \left(\sum_{\ell'=1}^{\ell-1} \frac{1}{n_{\ell'}}\right) \left[\left(\frac{3A_4 - A_2^2}{A_2^2}\right) (K^*)^2\right] = O\left(\frac{\ell}{n}\right)$$

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.

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.

[* `linear` activation is super-degenerate.]

2. More Generally

`tanh`, `sin`, ...

`GELU`, `SWISH`, ...

`perceptron`, `sigmoid`, `softplus`, ...

The Principle of Criticality

Necessity of hyperparameter fine-tunings in order to avoid exponentially exploding/vanishing signal & gradient problems for deep neural networks

The Principle of Criticality

Necessity of hyperparameter fine-tunings in order to avoid exponentially exploding/vanishing signal & gradient problems for deep neural networks

For scale-invariant activation functions, criticality was attained by fine-tuning as

$$(C_b, C_W)^{\text{critical}} = \left(0, \frac{2}{a_+^2 + a_-^2}\right)$$

Criticality Condition for the Kernel

More generically, a kernel recursion

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

has a fixed point $K^* = K^*(C_b, C_W)$ satisfying

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

Criticality Condition for the Kernel

More generically, a kernel recursion

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

has a fixed point $K^* = K^*(C_b, C_W)$ satisfying

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

Expanding the recursion around this fixed point as $K^{(\ell)} = K^* + \Delta K^{(\ell)}$:

$$\Delta K^{(\ell+1)} = \chi_{\parallel}(K^*) \times \Delta K^{(\ell)} + O(\Delta^2)$$

with $$\chi_{\parallel}(K) \equiv \frac{d}{dK} \left(C_W \left\langle \sigma(z)\sigma(z) \right\rangle_K \right)$$

Criticality Condition for the Kernel

$$K^* = K^*(C_b, C_W) \quad \text{satisfying} \quad K^* = C_b + C_W \langle \sigma(z)\sigma(z) \rangle_{K^*}$$

$$\Delta K^{(\ell+1)} = \chi_{\parallel}(K^*) \times \Delta K^{(\ell)} + O(\Delta^2) \qquad [\, K^{(\ell)} = K^* + \Delta K^{(\ell)} \,]$$

$$\text{with} \quad \chi_{\parallel}(K) \equiv \frac{d}{dK}\left(C_W \langle \sigma(z)\sigma(z) \rangle_K\right) = \frac{C_W}{2K^2} \langle \sigma(z)\sigma(z)(z^2 - K) \rangle_K$$

Criticality Condition for the Kernel

$K^* = K^*(C_b, C_W)$ satisfying $K^* = C_b + C_W \langle \sigma(z)\sigma(z) \rangle_{K^*}$

$$\Delta K^{(\ell+1)} = \chi_{\parallel}(K^*) \times \Delta K^{(\ell)} + O(\Delta^2) \qquad [\, K^{(\ell)} = K^* + \Delta K^{(\ell)} \,]$$

with $$\chi_{\parallel}(K) \equiv \frac{d}{dK}\left(C_W \langle \sigma(z)\sigma(z) \rangle_K\right) = \frac{C_W}{2K^2} \langle \sigma(z)\sigma(z)(z^2 - K) \rangle_K$$

- $\chi_{\parallel}(K^*) > 1$: exploding away
- $\chi_{\parallel}(K^*) < 1$: collapsing signal
- $\chi_{\parallel}(K^*) = 1$: critical propagation

Criticality Condition for the Kernel

$K^* = K^*(C_b, C_W)$ satisfying $K^* = C_b + C_W \langle \sigma(z)\sigma(z) \rangle_{K^*}$

$$\Delta K^{(\ell+1)} = \chi_{\parallel}(K^*) \times \Delta K^{(\ell)} + O(\Delta^2) \qquad [\, K^{(\ell)} = K^* + \Delta K^{(\ell)} \,]$$

with $$\chi_{\parallel}(K) \equiv \frac{d}{dK}\left(C_W \langle \sigma(z)\sigma(z) \rangle_K\right) = \frac{C_W}{2K^2} \langle \sigma(z)\sigma(z)(z^2 - K) \rangle_K$$

- $\chi_{\parallel}(K^*) > 1$: exploding away
- $\chi_{\parallel}(K^*) < 1$: collapsing signal
- $\chi_{\parallel}(K^*) = 1$: critical propagation

$$\boxed{\chi_{\parallel}(K^*) = \chi_{\parallel}\Big(K^*(C_b, C_W)\Big) = 1}$$

Criticality Condition for the NTK mean

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

Criticality Condition for the NTK mean

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

$$\chi_\perp(K) \equiv C_W \left\langle \sigma'(z)\sigma'(z) \right\rangle_K$$

Criticality Condition for the NTK mean

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

$$\chi_\perp(K) \equiv C_W \left\langle \sigma'(z)\sigma'(z) \right\rangle_K$$

- $\chi_\perp(K^*) > 1$: exploding gradient
- $\chi_\perp(K^*) < 1$: vanishing gradient (for lower layers)
- $\chi_\perp(K^*) = 1$: critical gradient

$$\chi_\perp(K^*) = \chi_\perp\Big(K^*(C_b, C_W)\Big) = 1$$

Two Criticality Conditions

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

$$\boxed{\chi_{\parallel}(K^*) = \chi_{\parallel}\Big(K^*(C_b, C_W)\Big) = 1}$$

$$\chi_{\parallel}(K) \equiv \frac{d}{dK}\left(C_W \langle \sigma(z)\sigma(z) \rangle_K\right) = \frac{C_W}{2K^2}\langle \sigma(z)\sigma(z)(z^2 - K) \rangle_K$$

$$\boxed{\chi_{\perp}(K^*) = \chi_{\perp}\Big(K^*(C_b, C_W)\Big) = 1}$$

$$\chi_{\perp}(K) \equiv C_W \left\langle \sigma'(z)\sigma'(z) \right\rangle_K$$

Two Criticality Conditions

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

$$\boxed{\chi_{\parallel}(K^*) = \chi_{\parallel}\Big(K^*(C_b, C_W)\Big) = 1}$$

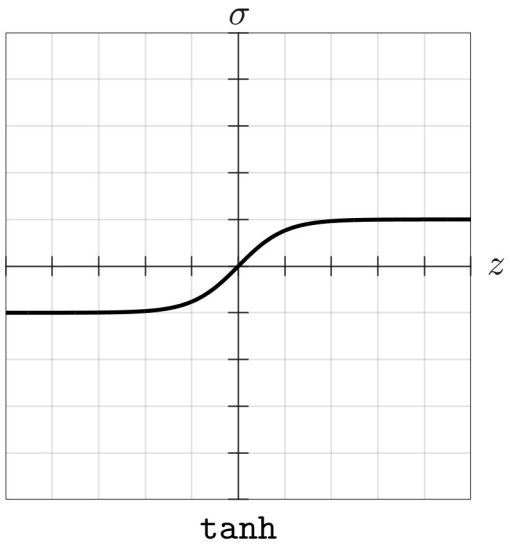
$$\chi_{\parallel}(K) \equiv \frac{d}{dK}\left(C_W \langle \sigma(z)\sigma(z) \rangle_K\right) = \frac{C_W}{2K^2}\langle \sigma(z)\sigma(z)(z^2 - K) \rangle_K$$

$$\boxed{\chi_{\perp}(K^*) = \chi_{\perp}\Big(K^*(C_b, C_W)\Big) = 1}$$

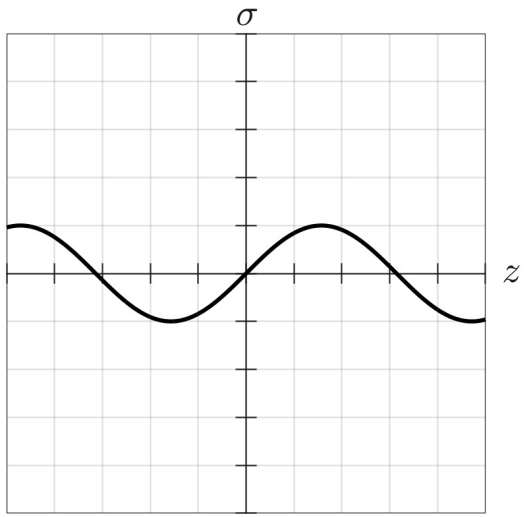
$$\chi_{\perp}(K) \equiv C_W \left\langle \sigma'(z)\sigma'(z) \right\rangle_K$$

[For scale-invariant case, $\chi_{\parallel}(K) = \chi_{\perp}(K) = C_W A_2 = \chi$]

$K^* = 0$ Universality Class: tanh, sin, ...



tanh



sin

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

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

$K^* = 0$ Universality Class: tanh, sin, ...

For smooth activation functions

$$\sigma(z) = \sigma_0 + \sigma_1 z + \frac{1}{2}\sigma_2 z^2 + \frac{1}{6}\sigma_3 z^3 + \dots$$

$$\chi_{\parallel}(K^*) = \chi_{\perp}(K^*) = 1 \quad \text{with} \quad K^* = 0$$

if and only if

$$\sigma_0 = 0 \quad \text{and} \quad \sigma_1 \neq 0$$

$K^* = 0$ Universality Class: `tanh`, `sin`, ...

Criticality attained at $(C_b, C_W)^{\text{critical}} = \left(0, \frac{1}{\sigma_1^2}\right)$

$K^* = 0$ Universality Class: `tanh, sin, ....`

Criticality attained at $$(C_b, C_W)^{\text{critical}} = \left(0, \frac{1}{\sigma_1^2}\right)$$

- Power-law decay: $$K^{(\ell)} \sim \frac{1}{\ell}, \quad \frac{V^{(\ell)}}{n_{\ell-1}} \sim \frac{1}{n\ell}$$

$K^* = 0$ Universality Class: tanh, sin, ...

Criticality attained at $$(C_b, C_W)^{\text{critical}} = \left(0, \frac{1}{\sigma_1^2}\right)$$

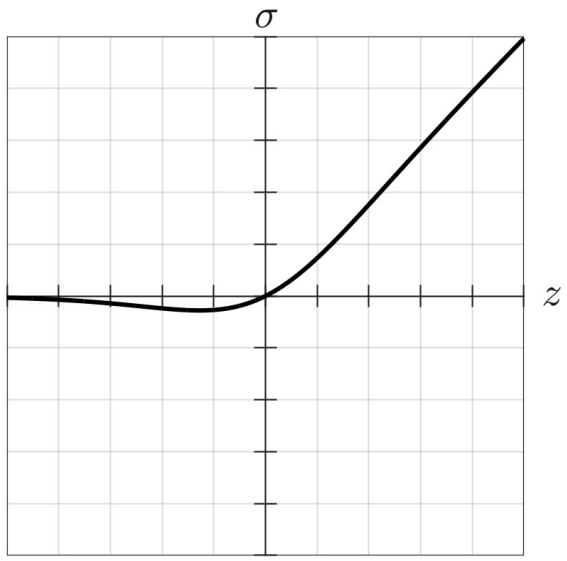
- Power-law decay: $$K^{(\ell)} \sim \frac{1}{\ell}, \quad \frac{V^{(\ell)}}{n_{\ell-1}} \sim \frac{1}{n\ell}$$

- "The Principle of Equivalence" to avoid the polynomial version of the exploding/vanishing gradient problem (i.e. to ensure equal contributions to NTK from all groups, §9.4) :

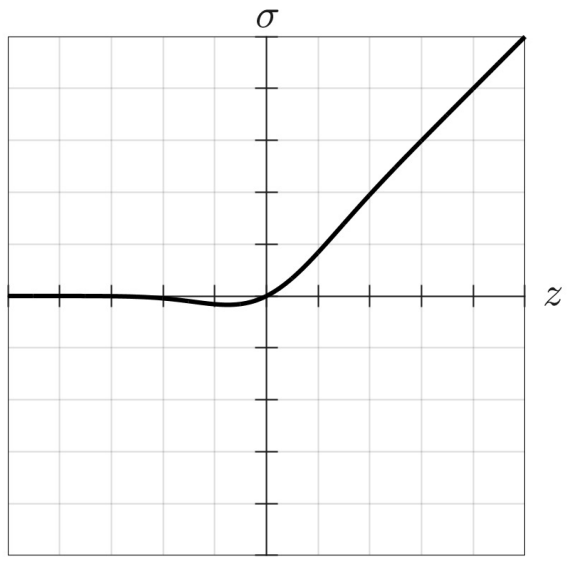
$$\lambda_b^{(\ell)} \propto \frac{1}{\ell}, \quad \frac{\lambda_W^{(\ell)}}{n_{\ell-1}} \propto \frac{1}{n_{\ell-1}}$$

(for odd smooth functions)

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



SWISH

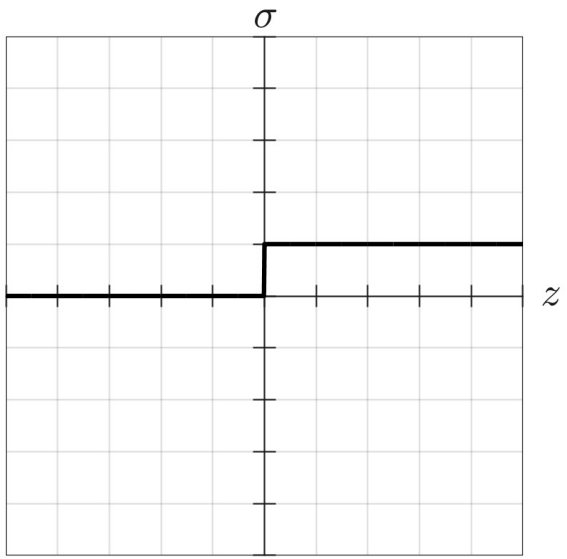


GELU

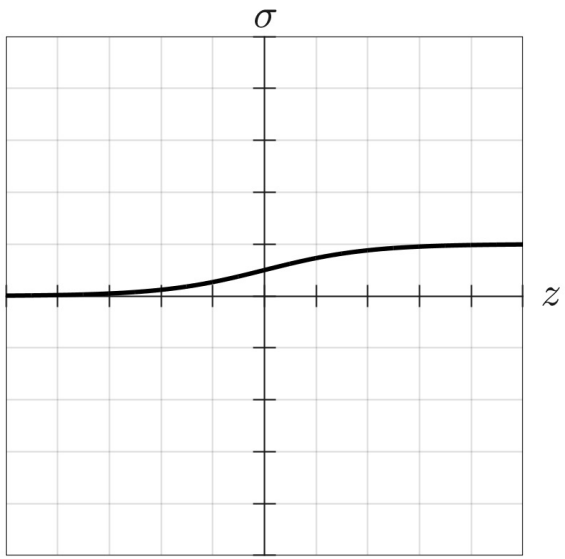
$$(C_b, C_W)^{\text{critical}} \approx (0.55514317, 1.98800468)$$

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

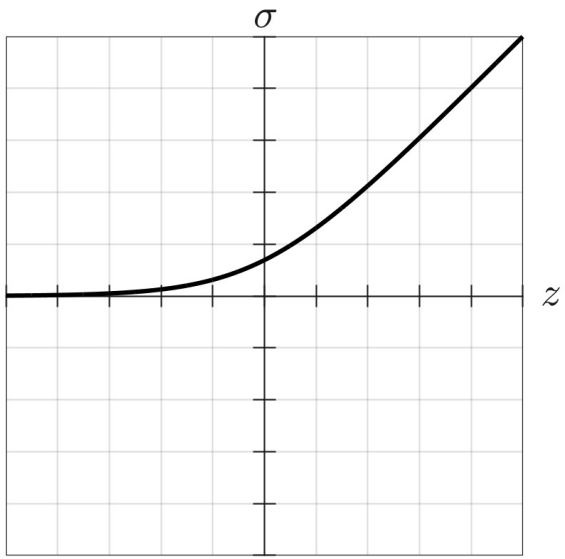
No Criticality, No Deep Learning: `perceptron`, `sigmoid`, `softplus`, ...



`perceptron`



`sigmoid`



`softplus`

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

Never again for deep learning

The Principle of Criticality

Necessity of hyperparameter fine-tunings in order to avoid exponentially exploding/vanishing signal & gradient problems for deep neural networks

We now have a principled way to identify critical initialization hyperparameters (and also to give no-go for some activation functions).

3. Finite-Width Effects and *Deep* Learning

What Really Matters

Scale-invariant universality class:

$$K^{(\ell)} \sim 1\,, \quad \frac{V^{(\ell)}}{n_{\ell-1}} \sim \frac{\ell}{n}$$

$K^* = 0$ universality class:

$$K^{(\ell)} \sim \frac{1}{\ell}, \quad \frac{V^{(\ell)}}{n_{\ell-1}} \sim \frac{1}{n\ell}$$

What Really Matters

Scale-invariant universality class:

$$K^{(\ell)} \sim 1\,, \qquad \frac{V^{(\ell)}}{n_{\ell-1}} \sim \frac{\ell}{n}$$

$K^* = 0$ universality class:

$$K^{(\ell)} \sim \frac{1}{\ell}, \qquad \frac{V^{(\ell)}}{n_{\ell-1}} \sim \frac{1}{n\ell}$$

$$[K] = [z^2],\ [V] = [z^4]$$

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

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

What Really Matters

Scale-invariant universality class:

$$K^{(\ell)} \sim 1\,, \qquad \frac{V^{(\ell)}}{n_{\ell-1}} \sim \frac{\ell}{n}$$

$K^* = 0$ universality class:

$$K^{(\ell)} \sim \frac{1}{\ell}, \qquad \frac{V^{(\ell)}}{n_{\ell-1}} \sim \frac{1}{n\ell}$$

$$[K] = [z^2],\ [V] = [z^4]$$

$$\boxed{\frac{1}{n_{\ell-1}} \frac{V^{(\ell)}}{\left(K^{(\ell)}\right)^2} \sim \frac{\ell}{n}}$$

Scaling Relations for Finite-Width Effects

Non-Gaussianity

$$\frac{1}{n}\frac{V^{(L)}}{\left(K^{(L)}\right)^2}$$

NTK fluctuations (§9)

$$\frac{1}{n}\frac{A^{(L)}}{\left(\Theta^{(L)}\right)^2}, \quad \frac{1}{n}\frac{B^{(L)}}{\left(\Theta^{(L)}\right)^2}, \quad \frac{1}{n}\frac{D^{(L)}}{K^{(L)}\Theta^{(L)}}, \quad \frac{1}{n}\frac{F^{(L)}}{K^{(L)}\Theta^{(L)}}$$

dNTK (§11.3)

$$\frac{1}{n}\frac{P^{(L)}}{\left(\Theta^{(L)}\right)^2}, \quad \frac{1}{n}\frac{Q^{(L)}}{\left(\Theta^{(L)}\right)^2}$$

ddNTK (§∞.2)

$$\frac{1}{n}\frac{R^{(L)}K^{(L)}}{\left(\Theta^{(L)}\right)^3}, \quad \frac{1}{n}\frac{S^{(L)}K^{(L)}}{\left(\Theta^{(L)}\right)^3}, \quad \frac{1}{n}\frac{T^{(L)}K^{(L)}}{\left(\Theta^{(L)}\right)^3}, \quad \frac{1}{n}\frac{U^{(L)}K^{(L)}}{\left(\Theta^{(L)}\right)^3}$$

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

Good

Non-Gaussianity

$$\frac{1}{n}\frac{V^{(L)}}{\left(K^{(L)}\right)^2}$$

NTK fluctuations (§9)

$$\frac{1}{n}\frac{A^{(L)}}{\left(\Theta^{(L)}\right)^2}, \quad \frac{1}{n}\frac{B^{(L)}}{\left(\Theta^{(L)}\right)^2}, \quad \frac{1}{n}\frac{D^{(L)}}{K^{(L)}\Theta^{(L)}}, \quad \frac{1}{n}\frac{F^{(L)}}{K^{(L)}\Theta^{(L)}}$$

dNTK (§11.3)

$$\frac{1}{n}\frac{P^{(L)}}{\left(\Theta^{(L)}\right)^2}, \quad \frac{1}{n}\frac{Q^{(L)}}{\left(\Theta^{(L)}\right)^2}$$

ddNTK (§∞.2)

$$\frac{1}{n}\frac{R^{(L)}K^{(L)}}{\left(\Theta^{(L)}\right)^3}, \quad \frac{1}{n}\frac{S^{(L)}K^{(L)}}{\left(\Theta^{(L)}\right)^3}, \quad \frac{1}{n}\frac{T^{(L)}K^{(L)}}{\left(\Theta^{(L)}\right)^3}, \quad \frac{1}{n}\frac{U^{(L)}K^{(L)}}{\left(\Theta^{(L)}\right)^3}$$

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

Bad

Non-Gaussianity

$$\frac{1}{n}\frac{V^{(L)}}{\left(K^{(L)}\right)^2}$$

NTK fluctuations (§9)

$$\frac{1}{n}\frac{A^{(L)}}{\left(\Theta^{(L)}\right)^2}, \quad \frac{1}{n}\frac{B^{(L)}}{\left(\Theta^{(L)}\right)^2}, \quad \frac{1}{n}\frac{D^{(L)}}{K^{(L)}\Theta^{(L)}}, \quad \frac{1}{n}\frac{F^{(L)}}{K^{(L)}\Theta^{(L)}}$$

dNTK (§11.3)

$$\frac{1}{n}\frac{P^{(L)}}{\left(\Theta^{(L)}\right)^2}, \quad \frac{1}{n}\frac{Q^{(L)}}{\left(\Theta^{(L)}\right)^2}$$

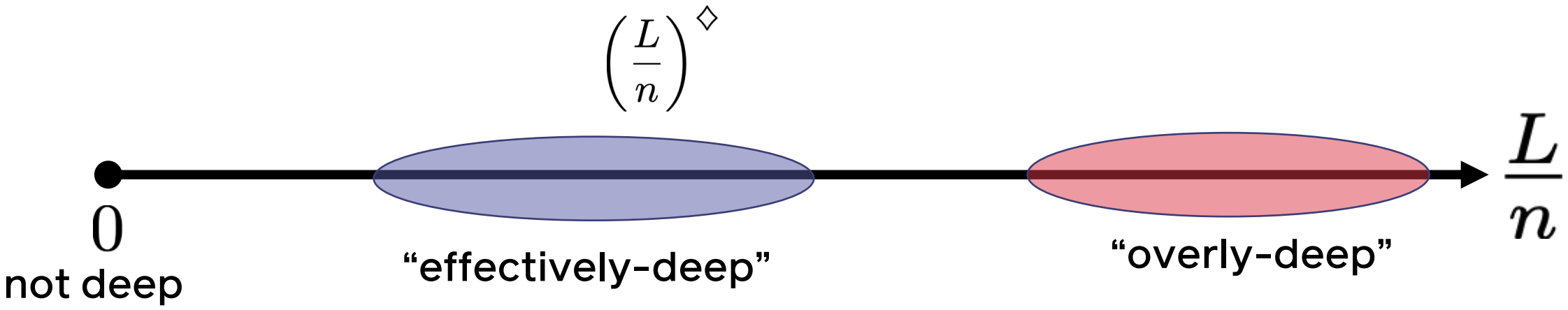
ddNTK (§∞.2)

$$\frac{1}{n}\frac{R^{(L)}K^{(L)}}{\left(\Theta^{(L)}\right)^3}, \quad \frac{1}{n}\frac{S^{(L)}K^{(L)}}{\left(\Theta^{(L)}\right)^3}, \quad \frac{1}{n}\frac{T^{(L)}K^{(L)}}{\left(\Theta^{(L)}\right)^3}, \quad \frac{1}{n}\frac{U^{(L)}K^{(L)}}{\left(\Theta^{(L)}\right)^3}$$

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

Sweet Spot

(a part of Lecture 5; also cf. Appendices A and B)



4. More on *Why* Criticality?

So Many Reasons!

- Taming exploding/vanishing kernel problem: today+§3 (DLN)+§5 (general)

So Many Reasons!

- Taming exploding/vanishing kernel problem: today+§3 (DLN)+§5 (general)
- Taming exploding/vanishing gradient problem: today+§9.4

So Many Reasons!

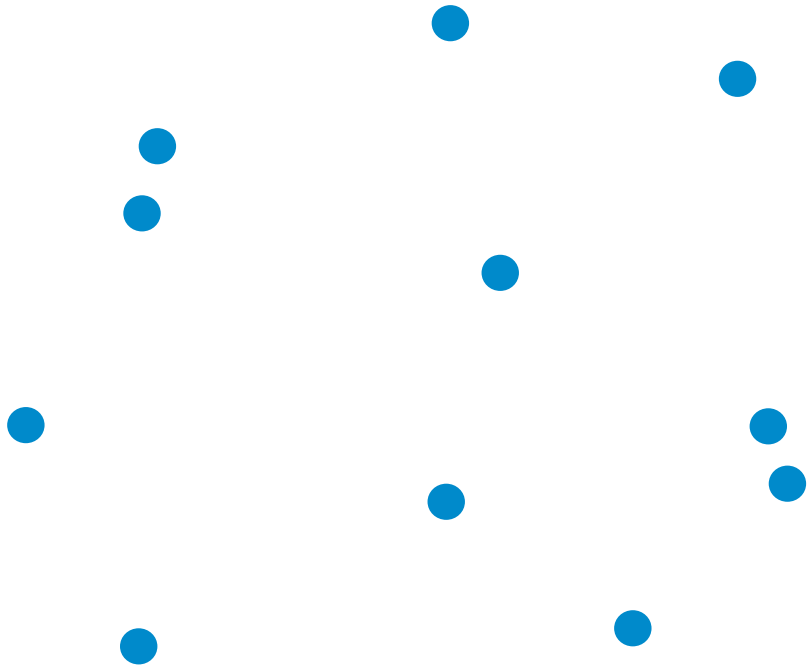
- Taming exploding/vanishing kernel problem: today+§3 (DLN)+§5 (general)
- Taming exploding/vanishing gradient problem: today+§9.4
- Bayesian evidence: §6.3.1

So Many Reasons!

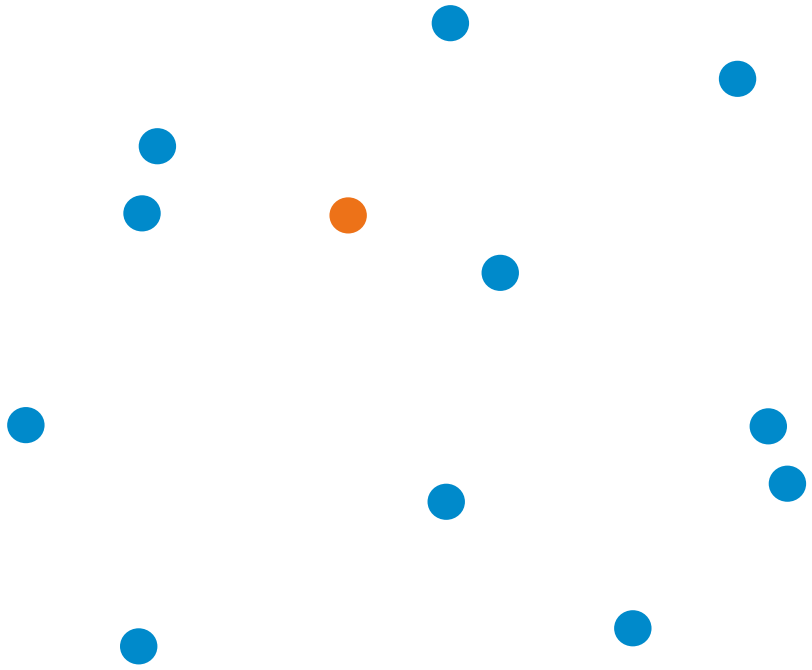
- Taming exploding/vanishing kernel problem: today+§3 (DLN)+§5 (general)
- Taming exploding/vanishing gradient problem: today+§9.4
- Bayesian evidence: §6.3.1
- Generalization error: §10.3

Generalized Bias-Variance Tradeoff

training inputs $\tilde{\alpha} \in \mathcal{A}$



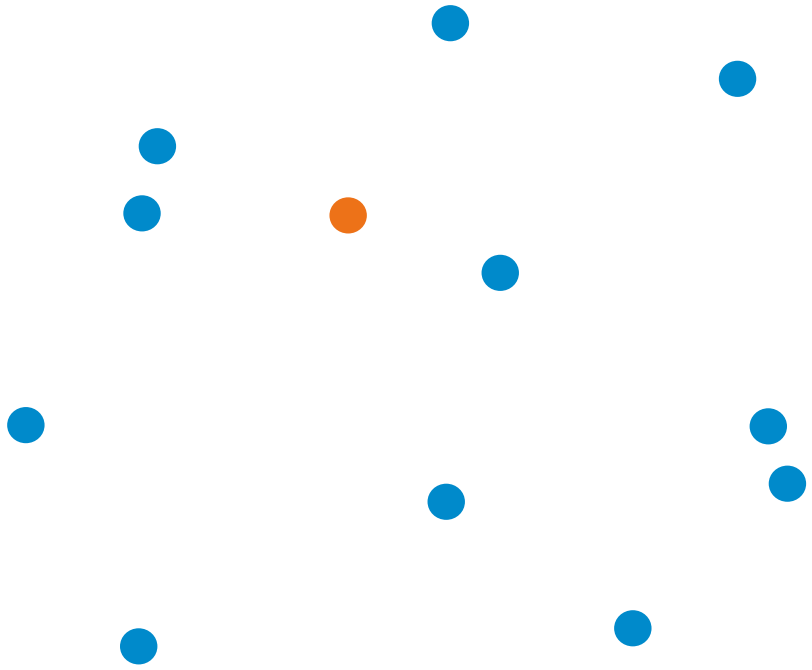
Generalized Bias-Variance Tradeoff



training inputs $\tilde{\alpha} \in \mathcal{A}$

test input $\dot{\beta} \in \mathcal{B}$

Generalized Bias-Variance Tradeoff



training inputs $\tilde{\alpha} \in \mathcal{A}$

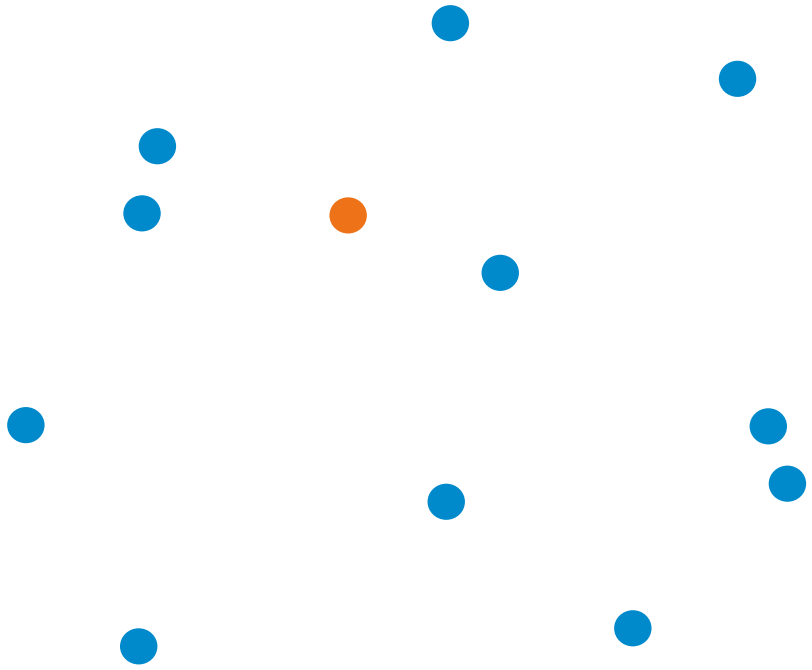
test input $\dot{\beta} \in \mathcal{B}$

$$\mathbb{E}\left[\left(z^{\star}_{i;\dot{\beta}} - y_{i;\dot{\beta}}\right)^2\right]$$

$$= \mathbb{E}\left[\left(z^{\star}_{i;\dot{\beta}} - \mathbb{E}[z^{\star}_{i;\dot{\beta}}] + \mathbb{E}[z^{\star}_{i;\dot{\beta}}] - y_{i;\dot{\beta}}\right)^2\right]$$

$$= \mathbb{E}\left[(z^{\star}_{i;\dot{\beta}})^2\right] - \left(\mathbb{E}[z^{\star}_{i;\dot{\beta}}]\right)^2 + \left(\mathbb{E}[z^{\star}_{i;\dot{\beta}}] - y_{i;\dot{\beta}}\right)^2$$

Generalized Bias-Variance Tradeoff



training inputs $\tilde{\alpha} \in \mathcal{A}$

test input $\dot{\beta} \in \mathcal{B}$

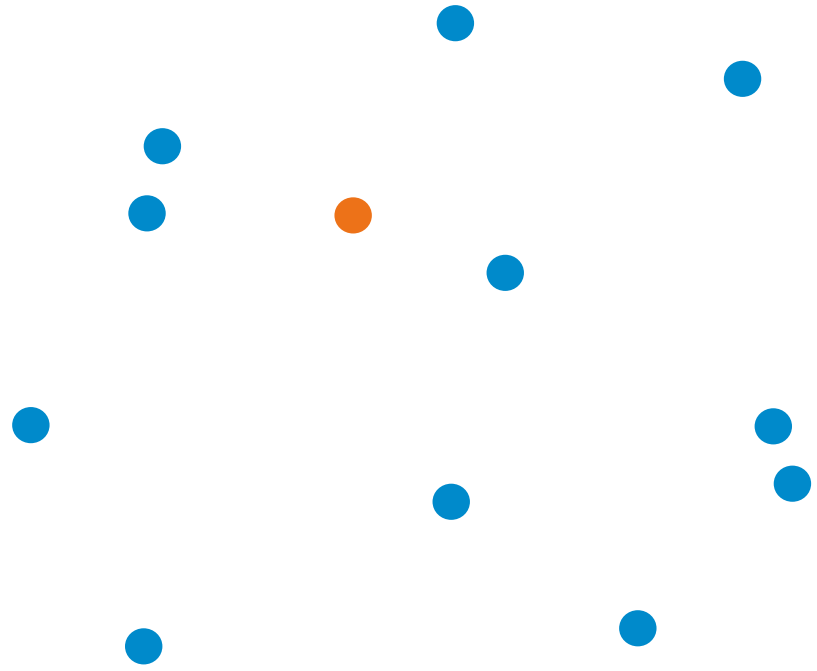
$$\mathbb{E}\left[\left(z^{\star}_{i;\dot{\beta}} - y_{i;\dot{\beta}}\right)^2\right]$$

$$= \mathbb{E}\left[\left(z^{\star}_{i;\dot{\beta}} - \mathbb{E}[z^{\star}_{i;\dot{\beta}}] + \mathbb{E}[z^{\star}_{i;\dot{\beta}}] - y_{i;\dot{\beta}}\right)^2\right]$$

$$= \mathbb{E}\left[(z^{\star}_{i;\dot{\beta}})^2\right] - \left(\mathbb{E}[z^{\star}_{i;\dot{\beta}}]\right)^2 + \left(\mathbb{E}[z^{\star}_{i;\dot{\beta}}] - y_{i;\dot{\beta}}\right)^2$$

variance bias

Generalized Bias-Variance Tradeoff



training inputs $\tilde{\alpha} \in \mathcal{A}$

test input $\dot{\beta} \in \mathcal{B}$

$$\mathbb{E}\left[\left(z^{\star}_{i;\dot{\beta}} - y_{i;\dot{\beta}}\right)^2\right]$$

$$= \mathbb{E}\left[\left(z^{\star}_{i;\dot{\beta}} - \mathbb{E}[z^{\star}_{i;\dot{\beta}}] + \mathbb{E}[z^{\star}_{i;\dot{\beta}}] - y_{i;\dot{\beta}}\right)^2\right]$$

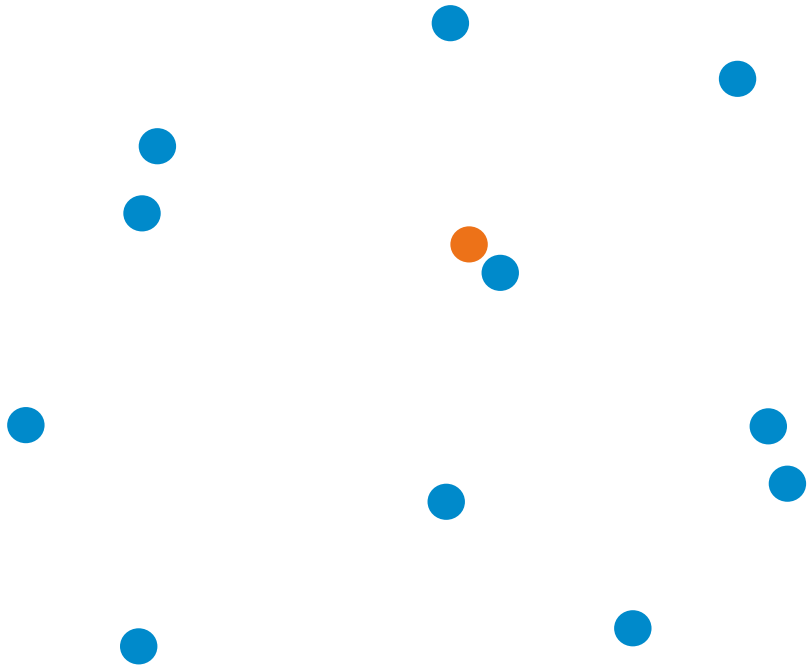
$$= \mathbb{E}\left[(z^{\star}_{i;\dot{\beta}})^2\right] - \left(\mathbb{E}[z^{\star}_{i;\dot{\beta}}]\right)^2 + \left(\mathbb{E}[z^{\star}_{i;\dot{\beta}}] - y_{i;\dot{\beta}}\right)^2$$

variance bias

!Here $\mathbb{E}$ is over instantiations of networks,
not over realizations of training samples!

Generalized Bias-Variance Tradeoff

An analytically tractable case I:
nearby test input (§10.3.1)



training inputs $\tilde{\alpha}_\sharp \in \mathcal{A}$

test input $\dot{\beta}_\sharp \in \mathcal{B}$

Deep networks with $\chi_\perp < 1$:
too inflexible/confident, too biased

Deep networks with $\chi_\perp > 1$:
too floppy/sensitive, too varied

The Principle of Sparsity for WIDE Neural Networks

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

- At infinite width: $p\left(\widehat{z}, \widehat{H}\right)$
- At $O\left(\frac{L}{n}\right)$: $p\left(\widehat{z}, \widehat{H}, \widehat{\mathrm{d}H}, \widehat{\mathrm{dd}H}\right)$

The Principle of Criticality for DEEP Neural Networks

Critical signal/gradient propagation @ $(C_b, C_W)^{\mathrm{critical}}$

Emergence of universality classes